Localised Near Horizon Predictive Models of Cellular Load

Emmett Carolan

Maynooth University
National University of Ireland Maynooth

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: Prof. Ronan Farrell

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

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: 02/08/2017

Abstract:

As mobile technologies continue to mature network providers are experiencing ever increasing demands on network resources. This trend will continue for a range of reasons, from growing subscriber expectations to the network being viewed as an enabling technology for the Internet of Things. However, these changes pose significant challenges to network operators at a time when many are facing stagnant or falling Average Revenue per User (ARPU). To provide increased services with reduced costs, network operators are looking to improvements in technology such as Software Defined Networking (SDN) and Self Organising Networks (SON). Several of these techniques will become key components of future 5G networks. With growing network complexity and reduced revenue to hire staff, many of these advanced management techniques will benefit from detailed predictive models of network load to allow for the preallocation of network parameters and resources. This thesis uses anonymised Call Detail Records (CDR) from Meteor, a mobile network provider in the Republic of Ireland, to model network load and investigate how it can be serviced more efficiently. The Meteor network under investigation has over 1 million customers, which represents approximately a quarter of the state's 4.6 million inhabitants.

The main contributions of this thesis are

1. A novel methodology to predict near horizon traffic loads in practical spatially contiguous coverage regions.
2. A novel application of near horizon localised prediction models to the problem of self-organising green networks.
3. Empirically created foundational models of how the network experiences load.

4. A novel examination of causal influences on network load, spatial relationships, communication distances, load predictability, and load usage.
5. A range of novel algorithms and techniques from novel metrics for measuring load prediction performance to novel algorithms for estimating subscriber areas of interest, CDR feature extraction, CDR data cleaning, load visualistation etc.

Results from this thesis show that there is a significant underutilisation of network resources. It is demonstrated that sufficiently accurate predictive models of network load are attainable at useful levels of spatial aggregation. These models are applied to the problem of self-organising green networks and demonstrate that a substantial reduction of network resource underutilisation is possible.

Publications Arising From This Thesis To Date:

E. Carolan, R. Farrell, and S. McLoone, "Localised Near Horizon Predictive Models of Cellular Load," IEEE Transactions on Wireless Communications, *in preparation*.

E. Carolan, R. Farrell, and S. McLoone, "Green Cellular Networks: A predictive modelling approach," IEEE Transactions on Wireless Communications, *in preparation*.

E. Carolan, R. Farrell, and S. McLoone, "A Predictive Model for Minimising Power Usage in Radio Access Networks," in 7th EAI International Conference on Mobile Networks and Management, Santander, Spain, 2015.*

E. Carolan, S. C. McLoone, and R. Farrell, "Predictive modelling of cellular load," in Signals and Systems Conference (ISSC), 2015 26th Irish, 2015, pp. 1-6.

E. Carolan, S. McLoone, and R. Farrell, "Characterising Spatial Relationships in Base Station Resource Usage," in Proceedings of the 17th Research Colloquium on Communications and Radio Science into the 21st Century, 2014.

E. Carolan, S. C. McLoone, and R. Farrell, "Exploring spatial relationships and identifying influential nodes in cellular networks," in Irish Signals & Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communications Technologies (ISSC 2014/CIICT 2014). 25th IET, 2013, pp. 245-250.**

E. Carolan, S. C. McLoone, and R. Farrell, "Comparing and Contrasting Smartphone and Non-Smartphone Usage," in the proceedings of ISSC, LYIT, 2013.

E. Carolan, S. C. McLoone, S. F. McLoone, and R. Farrell, "Analysing Ireland's interurban communication network using call data records," in the proceedings of Signals and Systems Conference (ISSC 2012), IET Irish, 2012, pp. 1-6.

R. Farrell, E. Carolan, S. McLoone, C., and S. McLoone, F., "Towards a Quantitative Model of Mobile Phone Usage Ireland – a Preliminary Study," in the proceedings of ISSC, NUI Maynooth, Ireland, 2012.

* Winner of best paper award.

** Second place, best overall paper award.

Acknowledgements

I would like to acknowledge the staff and students of Maynooth University, particularly of the Electronic Engineering department for their help and support. I would like to thank Prof. Sean McLoone and Dr. John Doyle for their help and support in the early stages of my PhD. I would especially like to thank my supervisors Prof. Ronan Farrell and Dr. Seamus McLoone. This work has been supported through the SFI Centre for Telecommunications Research (SFI-CE-I1853). I gratefully acknowledge the support of Meteor for all their assistance with this project.

Finally, I would like to thank my parents Bridie and Terry, my sister Ruth, and my wife Rachael, for their help and support throughout my PhD.

Table of Contents

List of Figures:

List of Tables

List of Equations

List of Abbreviations

2G Second-Generation Wireless Telephone Technology

3G Third-Generation Wireless Telephone Technology

4G Forth-Generation Wireless Telephone Technology

ACPE Absolute Capacity Percentage Error

ACF Auto Correlation Function

ADF Augmented Dickey Fuller

AI Artificial Intelligence

AIC Akaike Information Criterion

AMR Adaptive multi-rate

ANN Artificial Neural Networks

APN Access Point Name

ARIMA Auto-Regressive Integrated Moving Average Models

ARMA Auto-Regressive Moving Average

ARPU Average Revenue per User

BBU Base Band Unit

BS Base Station

BSC Base Station Controllers.

BTS Base Transceiver Stations

CCDF Complementary Cumulative Distribution Function

CDF Cumulative Distribution Function

CDMA Code Division Multiple Access:

CDR Call Detail Records

CN Core Network

CPE Capacity Percentage Error

CSV Comma-Separated Values

EDGE Enhanced Data rates for GSM Evolution.

ETL Extract Transform Load

eUTRAN Evolved UMTS Terrestrial Radio Access Network

eNodeB Evolved Node

FDMA Frequency Division Multiple Access

GSM Global System for Mobile Communications

GRAN GSM Radio Access Network

HSDPA High Speed Data Packet Access

IMSI International Mobile Subscriber Identity

ICT Information and Communication Technology

IoT Internet of Things

KPI Key Performance Indicators

LM Levenberg-Marquardt

LTE Long Term Evolution

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

MLP Multilayer Perceptron

MME Mobility Management Entitles

MS Mobile Station

MSC Mobile Switching Centre

MSISDN Mobile Station International Subscriber Directory Number

NCHO Network Controlled Hand Off

LHS Left Hand Side

OCI Other Cell Interference

OFDMA Orthogonal Frequency Division Multiple Access

PACF Partial Auto Correlation Function

PMF Probability Mass Functions

PPMCC Pearson's Product Moment Correlation Coefficient

QoS Quality of Service

RA Routing Areas

RAN Radio Access Network

RAT Radio Access Technologies

RHS Right Hand Side

RMSE Root Mean Square Error

RNC Radio Network Controller

RRU Remote Radio Unit

SARIMA Seasonal Auto-Regressive Integrated Moving Average Models

SGSN Serving General Packet Radio Service

SMS Short Message Service

SO Self Organising

SON Self Organising Networks

TA Tracking Area

TED Total Equivalent Data

TDMA Time Division Multiple Access

UE User Equipment

WSN Wireless Sensor Networks

Chapter 1 Introduction

Cellular networks have evolved rapidly since their inception a few decades ago. As cellular technology has evolved, so too have the expectations placed upon it. This growth in expectations does not look set to abate anytime soon. Increasingly capable subscriber equipment has opened up whole new uses for cellular networks from on demand video streaming to online gaming. Coinciding with the evolution of cellular network technology, new industries and businesses are looking to cellular networks as an enabling technology for the growing Internet of Things (IoT). Therefore, it is not surprising that globally mobile data traffic has grown 18 fold over the past five years and is projected to grow sevenfold between 2016 and 2021 [1]. However, these changes pose significant challenges to network operators at a time when many are facing stagnant or falling Average Revenue per User (ARPU) [2]. Currently, the tuning of many network parameters is often carried out by network operators manually, using network planning tools or drive tests [3]. From the perspective of network operators, the manual configuration of an increasingly complex network incorporating multiple Radio Access Technologies (RATs) increases operational expenditure. The autonomous optimisation of network parameters which uses a minimum amount of overhead is thus an attractive proposition to network operators. Such autonomous configuration techniques are often referred to as Self Organisation (SO) methods by the network standardisation bodies [4]. SO is subdivided into three main classes: self-configuration, self-optimisation, and self-healing. These three enable the auto-configuration of basic system parameters, resource allocation, and recovery from node failure. A more detailed description of the various self-organising modes is provided in [5]; SO techniques have been widely studied for other communication networks such as Wireless Sensor Networks (WSNs) and ad-hoc networks. SO concepts are relatively new in cellular networks but have

already attracted an extensive body of research focusing on their implementation such as [6-8]. To provide increased services with reduced costs, network operators are looking to the incorporation of SO concepts into cellular networks yielding Self Organising Networks (SON) [9]. [10] highlights the need for SONs capable of reducing human intervention by showing the growth in complexity of the configuration of a typical network node. [10] finds that a typical 2G node has approximately 500 configurable parameters, a typical 3G node has approximately 1000, and a typical 4G node has approximately 1500. [11] projects even greater complexity for 5G networks with a typical node having 2000 or more configurable parameters.

The rollout of SON technologies and the subsequent removal of the need for the manual configuration of network parameters opens cellular networks up to new advanced management techniques such as: the secondary usage of valuable licenced spectrum [12], opportunistic traffic scheduling [13], the dynamic switching on and off of underutilised Base Stations (BSs) [14], etc. The need for these new advanced management techniques is highlighted by a number of studies which have found large scale underutilisation of network resources. [15] found that “10% of base stations carry 50-60% of the load” which indicated a significant spatial underutilisation of certain parts of the network and their servicing BSs. [16] found a dramatic difference between the peak and trough hours of load within BSs and wider regions. [16] suggests this represents a significant underutilisation of network resources in the temporal domain. This problem was found to be particularly acute during the early morning hours when the network was vastly overprovisioned for the demand it experienced. The utilisation of advanced management techniques to more efficiently use network resources via SONs is a key component of future 5G networks as discussed in [11].

The leveraging of techniques and concepts from Artificial Intelligence (AI) is a key requirement for the functioning of SONs and the advanced network management techniques that rely on them. Broadly speaking, future 5G SONs require AI to perform four main groups of tasks: Sensing, Mining, Optimisation, and Prediction [11].

- Sensing is concerned with the detection of network anomalies/events/states from large datasets from hybrid sources. For example, [17] utilises a variety of AI techniques to first learn what a functional cell's Key Performance Indicators (KPIs) are, and then use this information to identify aberrant cell behaviour.
- Mining in future 5G cellular networks is concerned with the classification of services according to their required provisioning mechanism (e.g. bandwidth, error rate, latency etc.) [11]. For example, [18] proposes the use of contextual information which can be mined from the application to optimise mobile connectivity for bandwidth-hungry but delay tolerant applications.
- Optimisation in future cellular networks is primarily concerned with the configuration of a series of parameters to maximise a performance metric. For example, [19] employs AI techniques to develop methods for finding optimal antenna tilt angles in BSs.
- Prediction in future cellular networks has many uses such as forecasting the mobility of User Equipment (UE) or predicting the traffic load ahead of time. For example, [20] employs user location information to predict their movement patterns and proactively anticipate traffic hotspots.

All four of the above areas are touched upon to varying degrees in this work. However, particular attention is given to prediction in cellular networks, specifically the prediction of the traffic load. Load modelling and prediction is a critical element in the performance, planning and evaluation of telecommunications networks and has

consequently attracted much attention. However, most of this research has focused on traditional wired broadband which has many different properties and needs in comparison to cellular networks. What work has been carried out on cellular networks is mostly focused on older voice-centric networks and datasets [12, 21-23]. Due to the increasing capabilities of devices connecting to the cellular network and the concomitant rise in data usage, cellular networks have shifted from being voice-centric to data centric networks [24, 25]. Other works such as [26] have access to both voice and cellular data but unfortunately only provide predictive results for the voice portion. Forecasting short term load on the macro cellular network scale is possible with a high degree of accuracy [27]. However, it is of limited practical value for many advanced management techniques such as green networks (networks with reduced energy consumption) [28] and spectrum sharing [12] which, due to cellular network subsidiarity, require more localised forecasts. For such applications, groupings with finer spatial resolution are required. [26] creates predictive models for voice calls on the network but cites the greater variance of cellular data at the individual cell level as prohibiting the creation of predictive models of data load. Knowing that accurate forecasting of cellular data load is possible at large spatial aggregations [27], raises the question of its possibility at lower aggregations. In the field of electricity load forecasting [29] the authors presented significant improvements in accuracy at relatively modest levels of aggregation. This raises the question, is cellular data load predictable on the network at useful levels of spatial aggregation? If predictive models of cellular load can be created at sufficiently small aggregation levels, then these models can be incorporated into and used to improve advanced network management techniques. For example, one such technique that would benefit greatly from the inclusion of these predictive models of cell load is cell on-off switching for green networking. Much work has gone into algorithms and techniques to dynamically switch

on/off cells or BSs [28, 30, 31]. However, most work in the area simply uses historical static load profiles or assumes that switching decisions can be made instantaneously. However, real world measurement results such as those presented in [16] show that switching can take up to 30 minutes due to the heating systems. Thus, predictions of the need to perform a switch ahead of time are important. This thesis will use anonymised Call Detail Records (CDR) from Meteor, a mobile network provider in the Republic of Ireland, to model network load and investigate the practicality of localised near horizon predictive models of cellular load on the target network. The Meteor network under investigation has over 1 million customers, which represents approximately a quarter of the state's 4.6 million inhabitants.

The main contributions of this thesis are:

1. A novel methodology to predict near horizon traffic loads in practical spatially contiguous coverage regions.
2. A novel application of near horizon localised prediction models to the problem of self-organising green networks.
3. Empirically created foundational models of how the network experiences load.
4. A novel examination of causal influences on network load, spatial relationships, communication distances, load predictability, and load usage.
5. A range of novel algorithms and techniques from novel metrics for measuring load prediction performance to novel algorithms for estimating subscriber areas of interest, CDR feature extraction, CDR data cleaning, load visualistation, etc.
6. A large scale measurement study of a nationwide cellular network.

Results from this thesis show that there is a significant underutilisation of network resources. It is demonstrated that predictive models of network load are attainable at useful levels of spatial aggregation and sufficient accuracy to allow for their practical application to advanced management techniques. These models are applied to the problem of self-organising green networks and demonstrate that a substantial reduction of network resource underutilisation is possible.

The rest of the thesis is laid out as follows:

- Chapter 2 provides a technical background to cellular networks and their operation. The dataset used in this thesis is also presented and the methods used to store and process it at are provided.
- Chapter 3 provides a large scale nationwide study of a cellular network. Analysis focuses on identifying trends and possible opportunities for resource rationalization. This chapter then provides empirically created foundational models of how the network experiences load i.e. models of arrival rates, connection durations and data consumption. These models are provided at a fine-grained level broken down by connecting device type and contract type.
- Chapter 4 focuses on the creation of a spatial representation of the entire network to allow for the association of load with defined spatial areas. A novel procedure is introduced to clean inaccuracies in the spatial coordinates of BSs. A method to visualise how the load is distributed spatially across the network both as a whole and across various services is provided. A novel algorithm to discover who lives and works within BSs/cells is created and examined. Chapter 4 also provides a novel exploration of the presence/lack of causal influence that exists between neighbouring BSs.

- Chapter 5 provides a novel examination of how different levels of load, service type, temporal aggregation, and spatial aggregation affect traffic load predictability. Chapter 5 then goes on to create and explore the predictability of practical real world spatially contiguous aggregations of network coverage regions.
- Chapter 6 defines and implements a novel and practical forecasting method for use in advanced management techniques incorporating predictive models. Two novel methods for the automatic modelling of large amounts of individual cells and their many possible permutations in different spatial aggregations are proposed, used and tested.
- Chapter 7 introduces a regional study of power usage on the study network. The use of near horizon predictive models of cellular load is validated via their incorporation into a novel and practical energy savings scheme which is tested on real world data across multiple regions.
- Chapter 8 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 Background

2.1 Introduction

This work exploits a large dataset provided by the Meteor mobile phone network, which is a nationwide network operating in the Republic of Ireland. This chapter has three main contributions:

1) It provides a general introduction to the technologies used on the network on which the dataset is generated.
2) It presents specific information on the network at the time of data collection including its topography, subscriber base and data collection procedures/format.
3) It provides an overview of the ETL (Extract, Transform, Load) process carried out on the raw data to prepare it for further analysis.

The rest of this chapter is laid out as follows: sections 2.2, 2.3, 2.4, and 2.5 introduce the fundamental technologies/concepts required to understand cellular networks. Section 2.6 provides specific information on the meteor network at the time of data collection and also details how the data was prepared for analysis. Finally, section 2.8 concludes the chapter.

2.2 Cellular Networks

A cellular network is a spatially distributed radio network which enables voice, text, or data communications between two or more devices [32]. Typically, a compatible communications device is connected via a wireless connection to a transceiver at a fixed location known as a tower. Each tower covers a spatial area which is known as a cell. A cell can range from several square kilometres in sparsely populated rural areas down to

a scale of hundreds of meters in densely populated urban environments. Each communication flow between devices, including intra-cell communications, passes from the initiating device through its connected transceiver. The flow is then routed through a hierarchical network of elements which facilitate information flow to a destination cell which services the spatial area the receiving device is located in. Finally, the destination cell communicates the information flow to the connected device via the appropriate transceiver as illustrated in Figure 2.1.

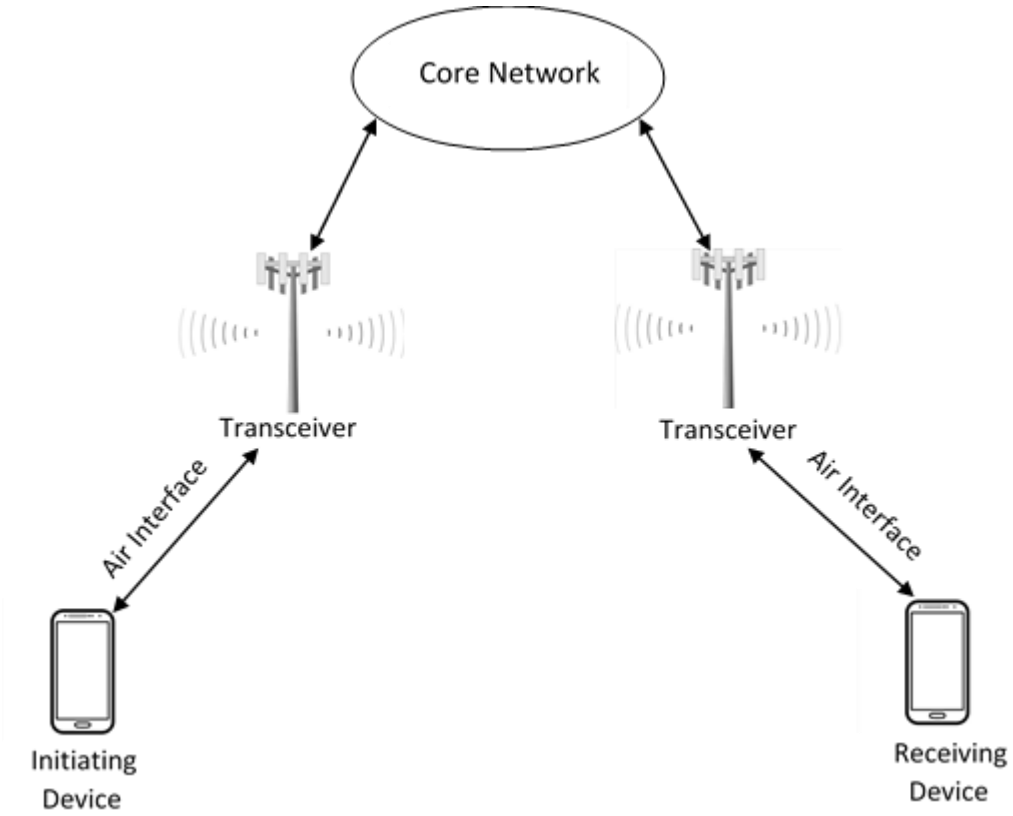


Figure 2.1: Overview of a simplified inter device communications flow in a cellular network

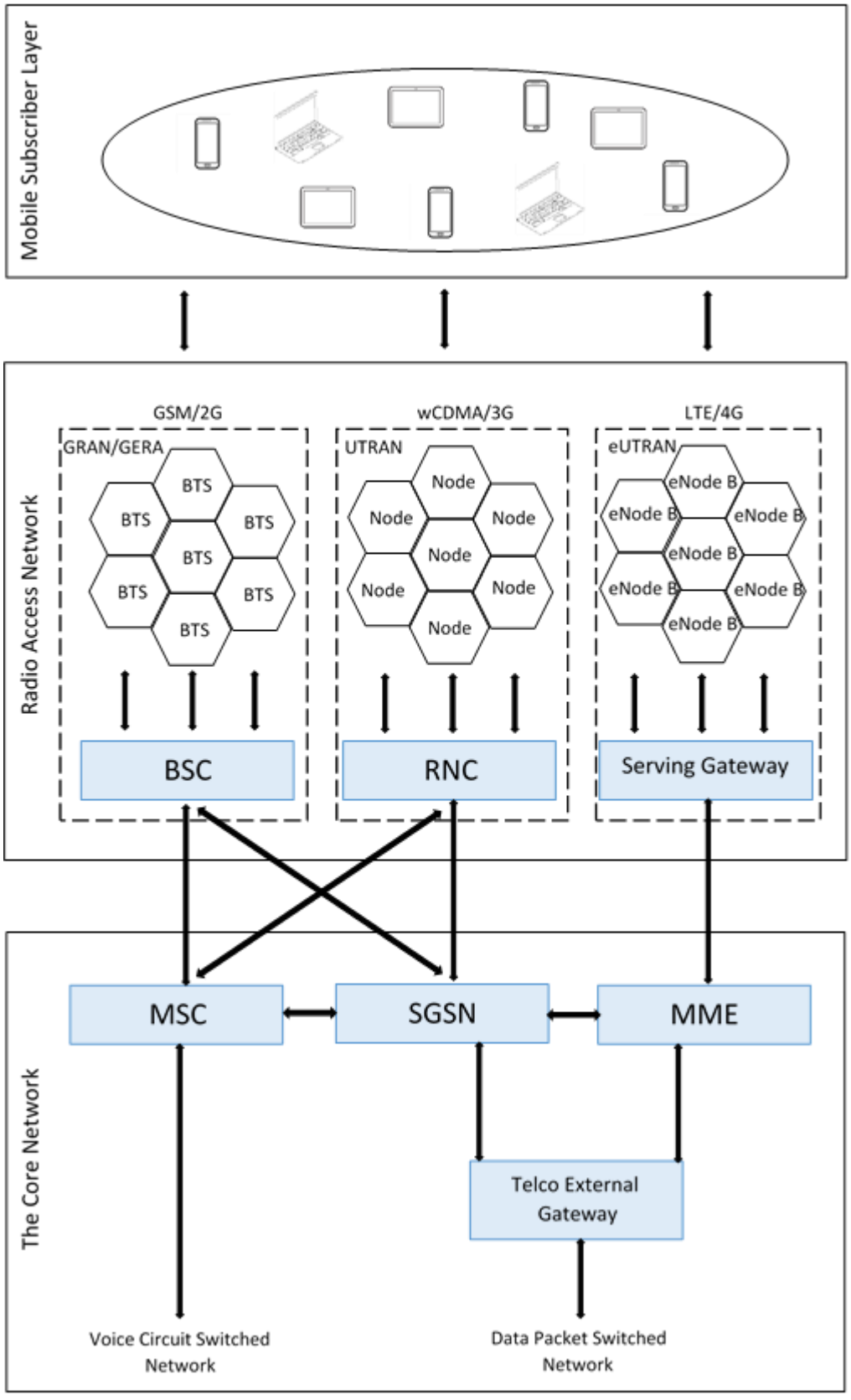


Figure 2.2: Simplified structure of a cellular network

Typically, cellular networks consist of a heterogeneous collection of technologies including those classed as second, third, and fourth generation wireless telephone technology. Figure 2.2 illustrates the simplified hierarchical layout of a heterogeneous cellular network. For simplicity, a cellular network may be divided into three primary sections: the mobile subscriber layer, the Radio Access Network (RAN), and the core network. The mobile subscriber layer consists of the mobile telephony enabled access devices or Mobile Stations (MS) which connect to the network. The RAN comprises the radio transceivers which are used to transfer data from the MS to the core network. The core network is the central part of the cellular network which provides services enabling communication, billing, and mobility. The RAN will vary depending on the communication standard employed between the 2G, 3G, and 4G versions. A GSM Radio Access Network (GRAN) is comprised 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 is made up of evolved Node B (eNode B) and serving gateways. The core network comprises elements of 2G, 3G, and 4G standards including Mobile Switching Centres (MSC), Serving GPRS Support Nodes (SGSN), and Mobility Management Entitles (MME). For a more detailed exposition of all the above network components see [33].

2.3 **Access Techniques**

Cellular networks enable simultaneous reception and transmission between communication devices within a certain amount of radio spectrum. This is carried out by a variety of access techniques which are primarily designed to allow transmitters to communicate with receivers with minimum interference [34]. Thus, the spectral efficiency is increased as more information is successfully transmitted and received over

limited spectrum. The access strategies used varies depending on the generational standard and are:

- Frequency Division Multiple Access (FDMA): Individual channels (unique frequency bands or spectrum slices as shown in Figure 2.3) are assigned to each MS on demand. During this time no other MS may use the channel.
- Time Division Multiple Access (TDMA): TDMA divides the radio channel up into time slots. Similarly to FDMA, each slot is assigned to an MS on demand and is allocated to the MS for the entire transmission as illustrated in Figure 2.3.
- Code Division Multiple Access (CDMA): CDMA is an example of multiple access, allowing several transmitters to send information simultaneously over a single communication channel. To facilitate multiple access without debilitating interference, CDMA employs spread spectrum technology with a coding scheme. CDMA multiples the narrowband message signal by a wideband signal known as the spreading signal. The spreading signal is a pseudo-noise code sequence with a chip rate orders of magnitude greater than the message signal's data rate [33]. Each MS is assigned a spreading code which is orthogonal to all other codes, and may transmit simultaneously using the same carrier. To recover the originally transmitted information, the receiver must decode the spreading code applied to it. Decoding is carried out using a time correlation operation with all the other code words appearing as noise due to decorrelation [33].
- Orthogonal Frequency Division Multiple Access (OFDMA): OFDMA uses time sharing coupled with dynamically assigned orthogonal subcarriers to provide multiple access to MS. MS that require high data rates may be assigned a higher number of subcarriers than those with lower data rate requirements.

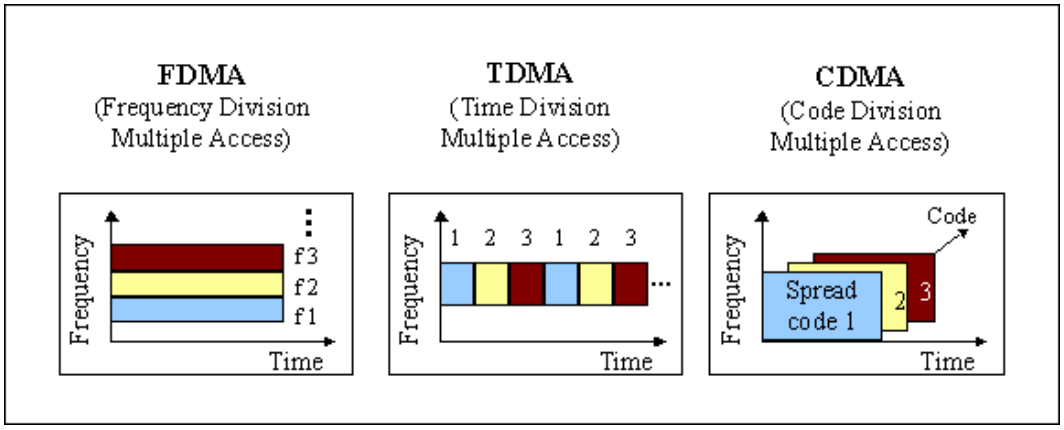


Figure 2.3: FDMA v TDMA v CDMA

For a more detailed exposition of all the above access techniques see [32-34].

2.4 **Coverage**

Transmission between RAN elements communicating wirelessly with connected devices in the subscriber layer suffer from path loss. Path loss is the energy lost between the transmission and reception of a signal. A transmission from an isotropic antenna will expand over a spherical wavefront, with the received energy a distance *d* away being inversely proportional to the sphere's surface area, $4\pi d^2$ [32]. The free space path loss is given by the Friis Formula:

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

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

As a result of the path loss, reliable communication is only possible over a limited distance for a defined maximum transmit power. Therefore, transmitters may operate using the same frequencies, at the same time if spatially isolated. Thus, the spatial area serviced by a cellular network is subdivided into smaller spatial regions. These smaller spatial regions are known as cells and contain a single Base Station (BS). To minimise

interference between adjacent cells, the transmit power of each transceiver should be configured to ensure that the signal strength is just strong enough at the cell boundaries. The same frequency channels may be reused in different, spatially isolated cells which greatly increases the available bandwidth. Thus, one way to increase the available bandwidth is to reduce the cell sizes (via reducing the transmit power) while increasing the number of cells. This results in many small densely packed cells in areas of high demand such as cities, as discussed further in Chapter 4. In practise, however, it is not possible to eliminate interference by selecting a transmit power that leads to perfect isolation between proximate cells. Thus, the amount of frequency reuse is selected to keep interference between cells below an acceptable threshold [33]. This intercell interference is referred to as Other Cell Interference (OCI) and negatively impacts performance. A commonly used technique to reduce OCI is to sectorise cells, where the sectorisation is carried out via directional antennas [32].

A typical cell layout is presented in Figure 2.4; the hexagonal shapes presented on the LHS of Figure 2.4 represent the idealised version of cell coverage. However, in practise, this does not accurately reflect real cell boundaries. The RHS of Figure 2.4 is a truer reflection of a real-world scenario where the geometrically irregular shape leaves some areas lacking coverage for a variety of reasons ranging from interference to obstructed signal propagation etc. [34]. To further increase the network's coverage and capacity in a region a network operator may also use a hierarchical cell structure as depicted in Figure 2.5. In such scenarios, a large macro cell may provide coverage to a spatial area as a whole while small cells service demand in smaller areas of particularly high demand within the larger area.

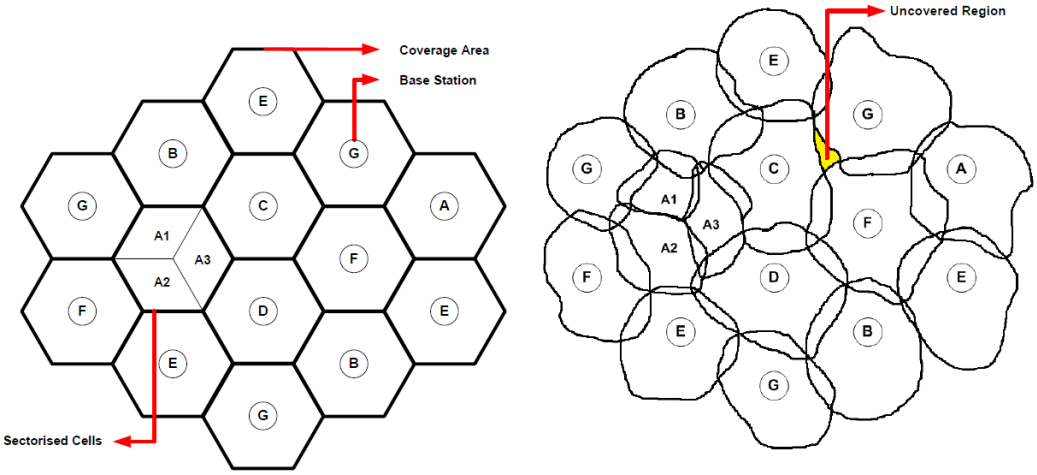


Figure 2.4: Typical cell layout; LHS shows the idealised version while the RHS shows the practical reality. A-G are the frequency channels used by each base station.

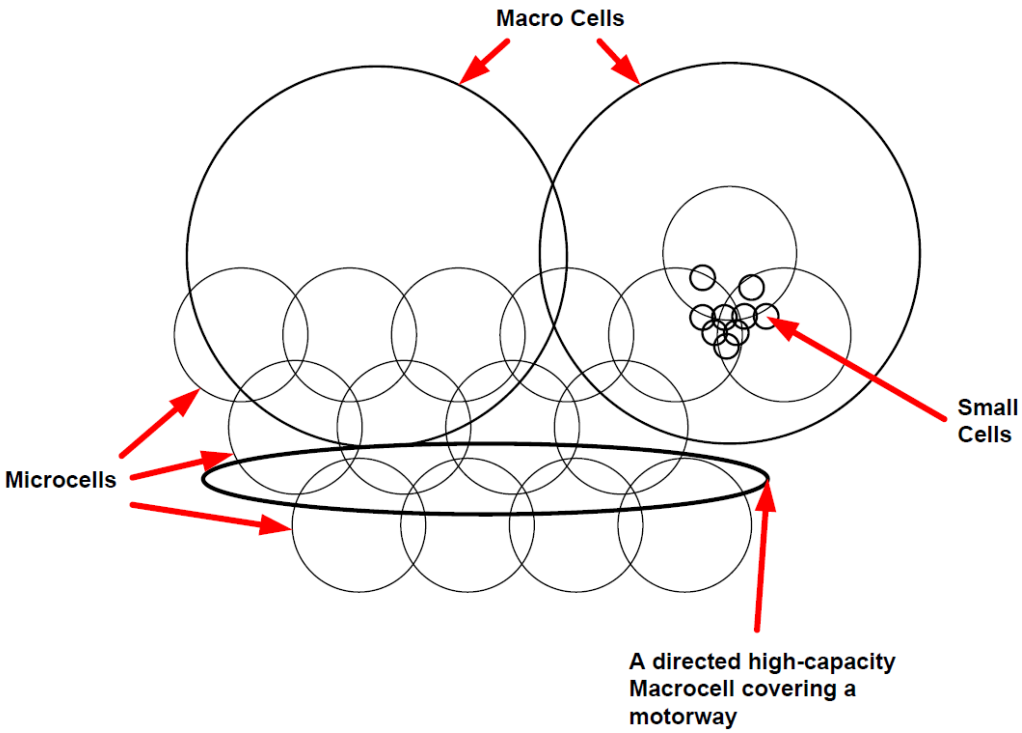


Figure 2.5: Simplified hierarchical cell structure

As each cellular network standard operates on different frequency ranges within the radio spectrum, network planners design each standard's network coverage layout independently. Thus, a BTS, Node-B, and eNode-B may all broadcast from the same tower and service overlapping spatial areas.

2.5 **Mobility Management**

As the network needs to be able to forward incoming communications, the location of the subscriber's device must be known to the network. When a mobile device is switched on it registers with the network. Thus, the network is made aware of the current location of the device. However, this location can change at any time as the user moves through the network's coverage area. If the subscriber's device moves into an area covered by a different cell, it may need to report this change to the network. To reduce the signalling load on the network, several cells are grouped into a larger location area. When a mobile device connects to a new cell, the network informs the mobile device of a new cell's ID and the Location Area Code (LAC) [33]. The mobile device will then only report its location if the new cell belongs to a different locating area from the previous cell (see Figure 2.6). One disadvantage of this method is that the network operator is only aware of the current location area of a mobile device and not the exact cell. Thus, the network must search for the mobile device in all cells of a location area for an incoming call or SMS. This searching procedure is known as paging. If the location areas are very large, there will be many mobile devices operating simultaneously within the area. This will result in a large amount of paging traffic, as every paging request must be broadcast to every BS in the location area. This wastes both bandwidth and also power in the mobile device by requiring it to listen to too many broadcast messages. However, if the location areas are too small, the mobile device must contact the network more frequently for location changes, which can also drain the device's battery. The size of the location area can be configured by the network operator and is typically 20-30 cells.

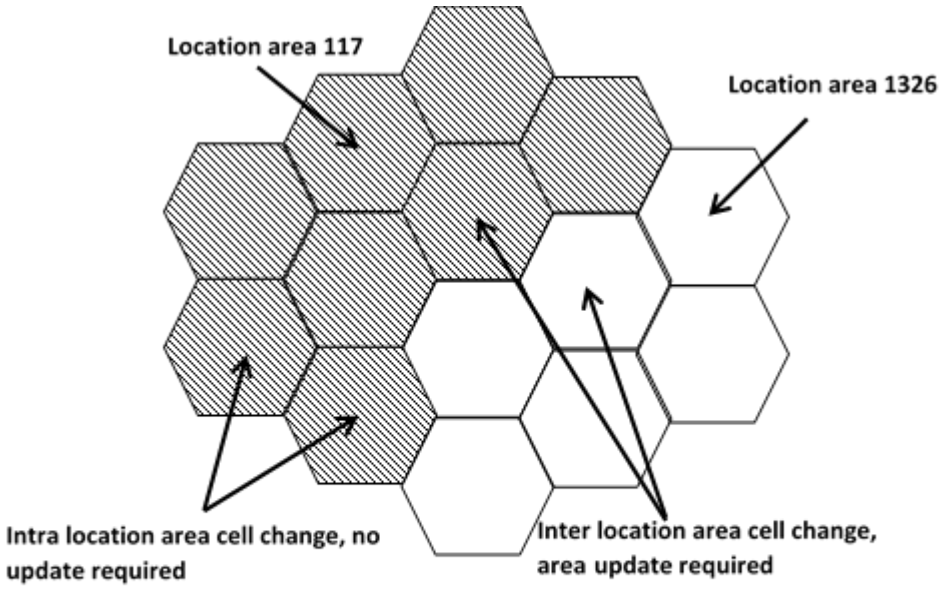


Figure 2.6: Cells in different location areas

For the packet-switched portion of the network, cells are aggregated into Routing Areas (RA). An RA is a subset of a location area, however, most network operators only use a single RA per location area [33]. A change from one RA to another (known as a "Routing Area Update") is almost identical to the procedure of changing from one location area to another. The primary difference is that due to the involvement of packet-switched data, the Serving GPRS Support Node (SGSN) is used. For newer networks supporting LTE, the equivalent of the location area and RA is the Tracking Area (TA). Again, the basic concepts behind the TA are very similar to those of the location area and RA. The network element involved in this case is the Mobility Management Entity (MME).

2.6 **Data Source**

The network under investigation in this work is the Meteor mobile phone network, which is a nationwide network operating in the Republic of Ireland. The network has over one million subscribers, which represents approximately one quarter of the

country's 4.6 million inhabitants. At the time of data collection the network operated both 2G and 3G services. The primary data source is Call Detail Records (CDR); CDR are primarily used as billing records for telecommunications transactions passing through the network. CDR are collected at the MSC and SGSN and contain records of all data transfers, voice calls and Short Message Service (SMS). The available dataset consists of approximately four months of data collected in 2011. The BS information provided includes geo-spatial coordinates in the Irish Grid Coordinate Reference System [35]. This coordinate system is the default system used, unless otherwise stated throughout this work. This coordinate system uses Easting and Northing projections which are defined in meter units from an origin point located at a latitude of 53°30'00 N and longitude 8°00'00 W. Other information about the BS includes technology type, the RNC or MSC serving the BS, and the transmitter azimuth. For more information on the topography of the network see section 4.2. The information provided for each subscriber contains their anonymised Mobile Station International Subscriber Directory Number (MSISDN), their subscription type (prepay/bill), year of birth, place of residence (town level), what their previous network was (if any), and how many upgrades they have availed of.

Records of SMS and voice calls are divided into originating and terminating files with data logs provided on cellular data sessions. The originating and terminating log files for voice calls provide information on both the caller and callee's anonymised MSIDSN, the time and duration of the call, the sectorised cell of both parties to the call when the call starts and also the respective cells when the call terminates. Note, the sectorised cell information is only available for Meteor subscribers. Similar information is provided for SMS in both the SMS originating and terminating log files. The cellular data log contains information on each data connection including: information on the anonymised MSIDSN, Access Point Name (APN), session start time, duration, amount of data

uploaded and downloaded, connected cell at the start and end of the connection, and the servicing SGSN.

The CDR data is processed via a repository server and three SFTP servers. The data is received in raw format as a CSV file from the Meteor server to the repository server. The repository server holds all the unprocessed data while the SFTP servers are used for data analysis. The data is transferred, pre-processed, and then loaded into MySQL databases on the relevant servers where each table is suitably optimised to allow for parameter extraction. A database table is a set of data elements (values) using a model of vertical columns (identified by name) and horizontal rows, the cell being the unit where a row and column intersect. A table has a specified number of columns, but can have any number of rows. The data can then be accessed directly on the processing server or remotely. An overview of the system architecture and some data examples are provided in Figure 2.7, while the table structures are displayed in Figure 2.8-Figure 2.12.

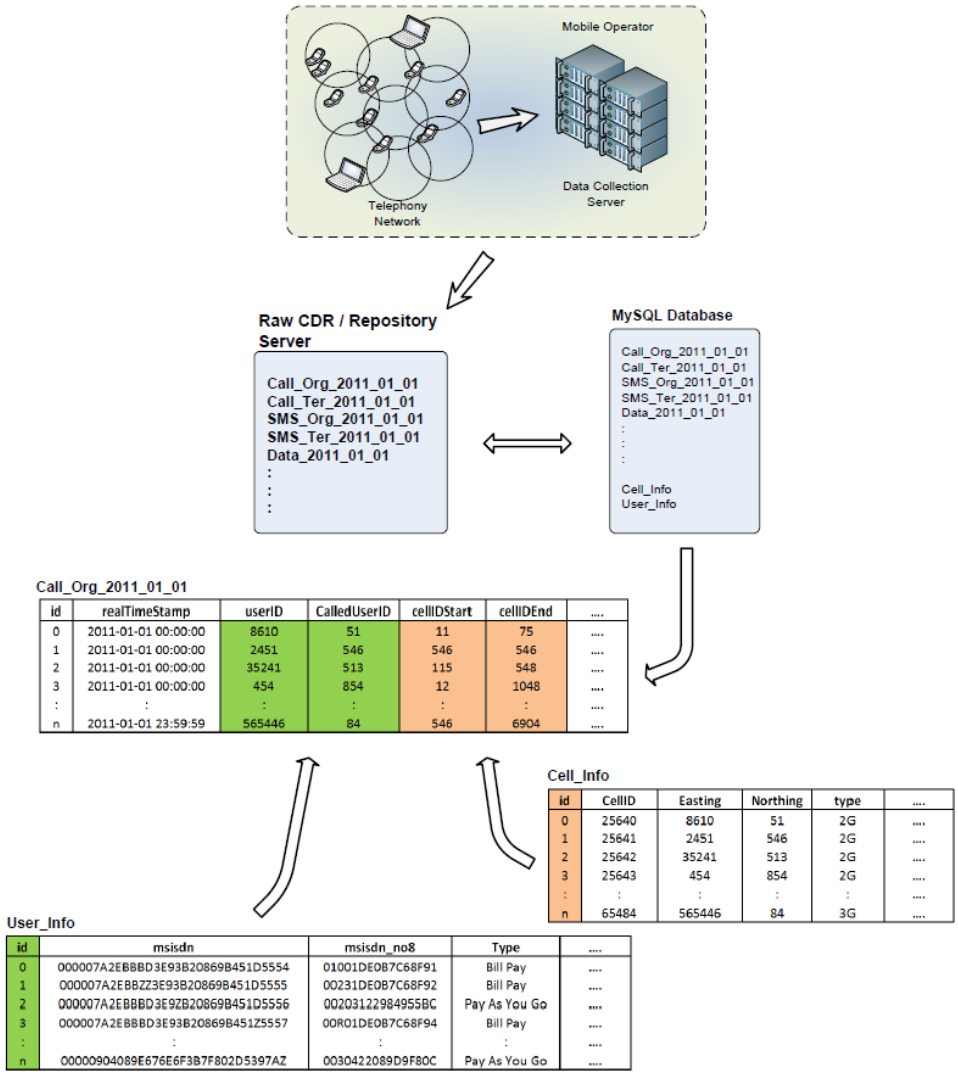


Figure 2.7: CDR processing architecture overview with some example table relationships

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

Figure 2.8: CDR call 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

Figure 2.9: CDR SMS 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 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

Figure 2.10: CDR call 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

Figure 2.11: CDR SMS terminating 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

Figure 2.12: CDR Data Session Table Structure

2.7 **Privacy**

The anonymity of subscribers is addressed by a hashing of the subscribers' unique MSISDN code. A MSISDN is a uniquely identifiable code which links to a person's subscription on a mobile cellular network. This hashing guarantees that a user's identity is not directly observable.

2.8 **Conclusion**

This chapter provided a general introduction to some of the technologies used on the network where the dataset was generated. Specific information on the network at the

time of data collection was also provided. The steps taken to extract, transform, and load the dataset to facilitate analysis were also provided. As in any research endeavour, the type and scope of the dataset does impose some restrictions on the type of research that it can be effectively applied to. For example, as discussed in section 2.6, the dataset provides details on the start and end cell of each communication event. However, it does not provide location details of devices/subscribers in-between communication events. Therefore, it only provides a sample of a device's/subscriber's location with a sampling rate determined by how often the device/subscriber communicates. As discussed in 2.5, a fuller dataset of a devices/subscribers location while not communicating is available to the network operator but unfortunately is difficult to obtain from network operators due to lack of incentive for long term storage. In contrast, activity based call detail records such as those used in this work are stored for longer and with greater care as they are required for legal compliance and billing [36]. Research areas which require detailed knowledge of a device's/subscriber's location at all times, such as modelling the instantaneous signalling load in a specific cell/area, while still possible with this dataset may benefit form additional data. This dataset also does not provide IP packet headers which could be used to identify the specific application/website being used. This precludes research that requires a detailed analysis of these features such as in [37] (however, a broader categorisation of application usage is possible and introduced in Chapter 3). While bearing these shortcomings in mind, the dataset described in this chapter is one of the largest and most complete (an entire nationwide network) ever used for a work of this kind.

Chapter 3 Analysing Cellular Network Load

3.1 Introduction

In the past two decades, mobile phones and devices utilising the mobile phone network have become ubiquitous in modern society. Mobile phone penetration has approached and, in some nations exceeds 100% [38]. Cellular networks are continuing to experience a large and sustained increase in demand for network resources [39]. As operators move to add capacity, a detailed understanding of the underlying dynamics of resource usage is increasingly important. Some previous works have attempted to provide this understanding, such as large a scale study of voice calls [12] or the study of user dynamics [15]. However, the practical usefulness of these studies is limited by several important factors. For example, [12] focuses on voice calls over the network which, as will be demonstrated in this chapter, are already a small fraction of network load and are projected to diminish further in the coming years [40]. Although [15] focuses on the data service, the dataset employed predates the widespread adoption of smartphones on the network and, thus, is of limited modern relevance.

This chapter has three main contributions:

1) The primary aim of this chapter is to provide empirically created foundational models of how the network experiences load i.e. models of arrival rates, connection durations and data consumption. These models are provided at a fine grained level broken down by connecting device type and contract type. The models presented in this work allow an interested third party to create their

own models of the most important factors of how the network experiences load at a fined grained level.

2) To provide an empirical measurement of network load and its constituent parts both at the network level and the level of the individual base station/cell.
3) To use quantitative and qualitative analysis of the network including both its load and topography to identify trends and possible opportunities for resource rationalization.

The primary novel feature of this chapter is the provision of empirical models of the fundamental network usage metrics. The first novel feature of these models is the scope and relevance of the dataset used to create these models. The dataset used comprises the entirety of a nationwide network and spans several months after the widespread adoption of smartphones. This is crucial, as previous work focused on the creation of empirical models in the pre-smartphone era [12]. The second novel feature of the models provided is the degree to which they are broken down by device type and contract type. All necessary parameters are provided to allow an interested party to recreate the source distributions. Thus, they will allow for the creation of more accurate models of network usage which will respond to changes in the mix of both device and contract types. The final novel feature of this chapter is the resolution to which the network load is quantified and qualified, both in spatial and temporal terms.

Section 3.2 provides an empirical examination of the total network load with a particular focus on the cellular data load. 3.3 provides an empirical examination of how the network load is serviced locally at the level of individual base stations/cells. 3.4 provides empirically derived models of network usage broken down by device type, time of day and contract type.

3.2 Total Network Load

3.2.1 Introduction

This section provides an empirical measurement of the network load and its constituent parts at the network level. Section 3.2 is broken down as follows: firstly, subsection 3.2.2 proposes and defines a metric to allow for the accurate comparisons of data volumes and load across disparate services and technologies. Next, 3.2.3 presents and discusses the aggregated network wide daily load curve. 3.2.4 discusses how the network load breaks down between the three main constituent parts i.e. voice, data and SMS. Subsection 3.2.5 implements a classification system to categorise data connections into one of several different types. 3.2.6 provides a concluding discussion of total network load in keeping with aim 2 presented in the chapter introduction "*identify trends and possible opportunities for resource rationalization*".

3.2.2 Total Equivalent Data (TED)

For the purposes of this work voice and SMS are expressed in terms of equivalent data services – as they are treated as such in a pure packet-switched network, for example Long Term Evolution (LTE). Voice is encoded in mobile phone networks using adaptive multi-rate (AMR) codecs. In GSM and wCDMA, a narrowband AMR scheme is used with a typical data rate of 12.2 kbps [41]. A higher quality wideband AMR is used in LTE and offers superior quality at a data range of 12.5 kbps [41, 42]. Higher and lower data rates are possible, but for this work a rate of 12.5 kbps will be used in converting voice channels to an equivalent data session. Text messages will be treated as a 200 byte message with 1 second duration. Multimedia messaging has not been included as it is negligible since the advent of 3G networks.

3.2.3 Total Network Load

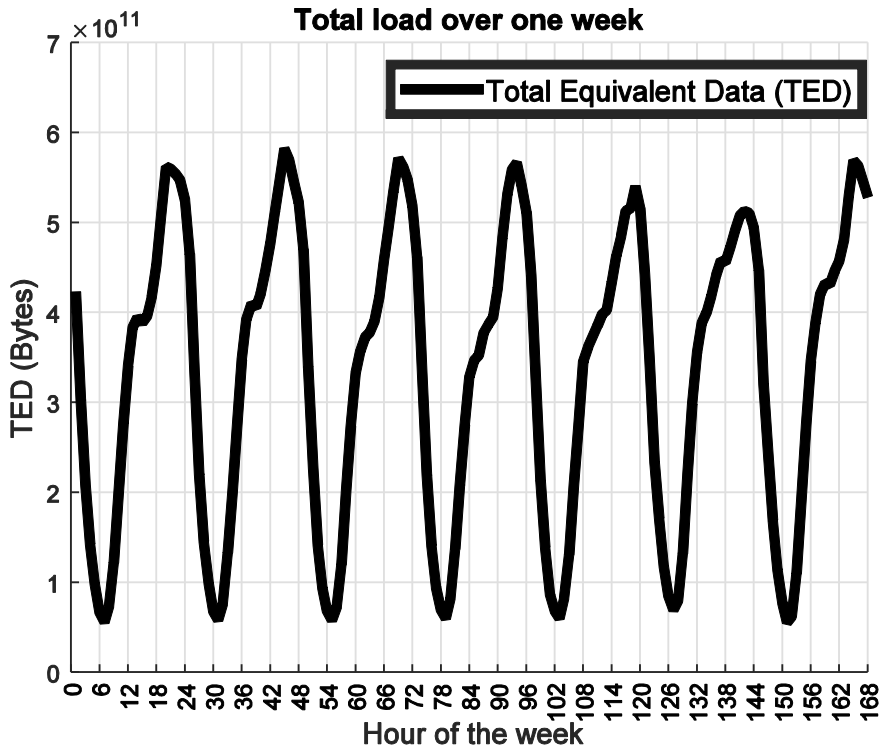


Figure 3.1: Total network load expressed as Total Equivalent Data (TED) in bytes over the course of one representative week. Note that hour zero is 0:00 on Monday morning.

Figure 3.1 illustrates the total aggregated network load across the entire network over one representative week. The first and most striking feature of note in Figure 3.1 is the rhythmic diurnal pattern of the load. Each day the load follows a similar trend with the peak occurring during the evening/night-time and the trough falling in the early morning hours. The peak network load is consistently an order of magnitude greater each day than the minimum load on the network. This highlights the classical peaking problem in resource distribution and shows that for much of the day large amounts of resources (spectrum, power etc.) are going to waste.

Interestingly, the peak hour of load shifts as the days of the week progress. On Mondays the peak load occurs between 6-7 p.m. and shifts slightly later each successive day until

it occurs between 10-11 p.m. on Friday and Saturday before moving to 7-8 p.m. on Sunday. Intuitively this would appear to match up with people going to bed early on work nights and staying out late when the next day is a day off. This is also seen when public holidays are considered. For example the day before a public holiday generally resembles a Friday while the public holiday resembles a Sunday (providing the next day is a working day). Interestingly, despite the shifting hour of peak load, the temporal location of the lowest load remains constant throughout the week, occurring between 6 and 7 a.m. each day. Historically, load forecasting in the electrical network has received more attention than data load and is consequently more advanced [43]. The two fields however share some similarities derived from the diurnal pattern of human activity. [27] uses a similar approach to electrical load forecasting to model and forecast the aggregated network data load for an entire US state. As in electrical load forecasting the authors of [27] proposed the use of two separate models, one for weekdays and one for weekend days. Examining Figure 3.1, the data suggests that on this network when modeling the total aggregated network load a better approach is to individually create a Monday-Thursday model, a Friday model, and a weekend model. Further investigation suggests that public holidays should be modeled as a weekend day. This will allow for greater nuance in the created model to capture different daily patterns.

3.2.4 Total Network Load by Service Type

Figure 3.2 (a) shows the number of usage events broken down into the three main services provided by the network operator: voice calls, SMS, and mobile data. The respective totals are: 63% of communication events on the network are SMS, 20% of events are cellular data usage while the remaining 17% are traditional voice calls. Figure 3.2 (a) clearly shows the predominance of SMS events on the network. However, Figure 3.2 (b) plots the distribution of load attributed to each service type and gives a very

different picture. Figure 3.2 (b) clearly demonstrates that from the perspective of data volumes transferred across the network that cellular data is the dominant service type. Despite cellular data connections accounting for only 20% of all communication events on the network, they are responsible for over 90% of the data volume on the network. Conversely SMS accounts for 63% of the connection events on the network but transfers less than 1% of the data on the network. This is in keeping with projections such as [44] which shows the network moving away from SMS and voice towards a more data centric paradigm.

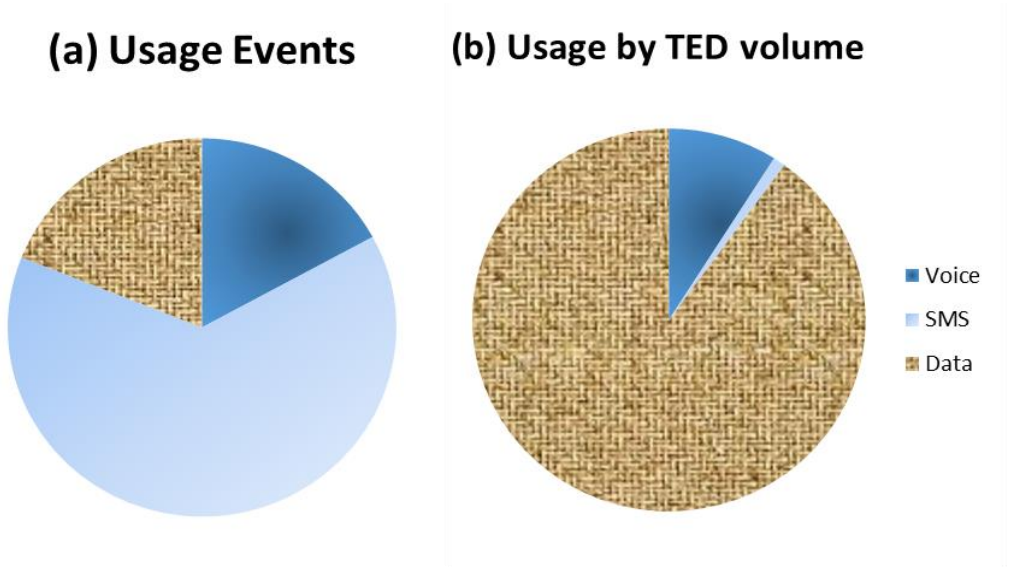


Figure 3.2: (a) The number of usage events broken down by service type over a typical day. (b) The total volume of data transferred over the whole network expressed as TED broken down by service type.

As the dataset employed in this work is mainly from 2011 it spans a time when smartphones were becoming widespread on the network. The results of this investigation will quantify the trend of smartphone users moving away from voice/SMS services towards alternative communication methods. This change from a voice/SMS centric network to a data centric network is forcing service providers to shift pricing

models from being call/SMS centric to data centric [45]. From the service providers perspective this can be partially blamed for reduced Average Revenue Per User (ARPU) but is good news from a consumer perspective as the price per byte transferred is greatly reduced [46].

Figure 3.3 shows how the total load on the network varies by service type over the course of a typical Monday. As in the aggregated usage mode case presented in Figure 3.1, the general trend is for traffic to be light during the early morning hours and then peak in the 8 p.m. to midnight period. This trend is driven by the predominance of data traffic on the network but interestingly masks a difference between voice/SMS and data. The peak hours of the former generally occur earlier than for the network as a whole, specifically during the 4 p.m.-8 p.m. slot. However, the call/SMS volume is relatively stable in the preceding and succeeding hours. Interestingly, this earlier peak hour better matches works such as [12] which relied on older datasets before the predominance of data services. 3.2.3 discussed the "peaking problem" in the network i.e. how the network is resourced for performance at peak hours of load and is thus consequently underutilised during the rest of the day. Figure 3.3 suggests that this problem is exacerbated further by the move towards cellular data. For example, for mobile data the ratio between usage during the midday-4 p.m. period and the 8 p.m.-midnight period is 1:1.55 while for voice it is almost 1:1. Thus, as the network becomes ever more data centric it is reasonable to assume that the peaking problem and the commensurate underutilisation of resources will become more acute. This is in keeping with findings produced in [40] which suggest the peaking problem is being exacerbated in both fixed line and mobile contexts due to the growth of data usage, particularly video applications. As will be demonstrated in 3.2.5, 63% of mobile data usage on this network is related to video applications. Tackling this problem will require more

advanced models of load and more active/automatic network management practices, many of which are developed in later chapters.

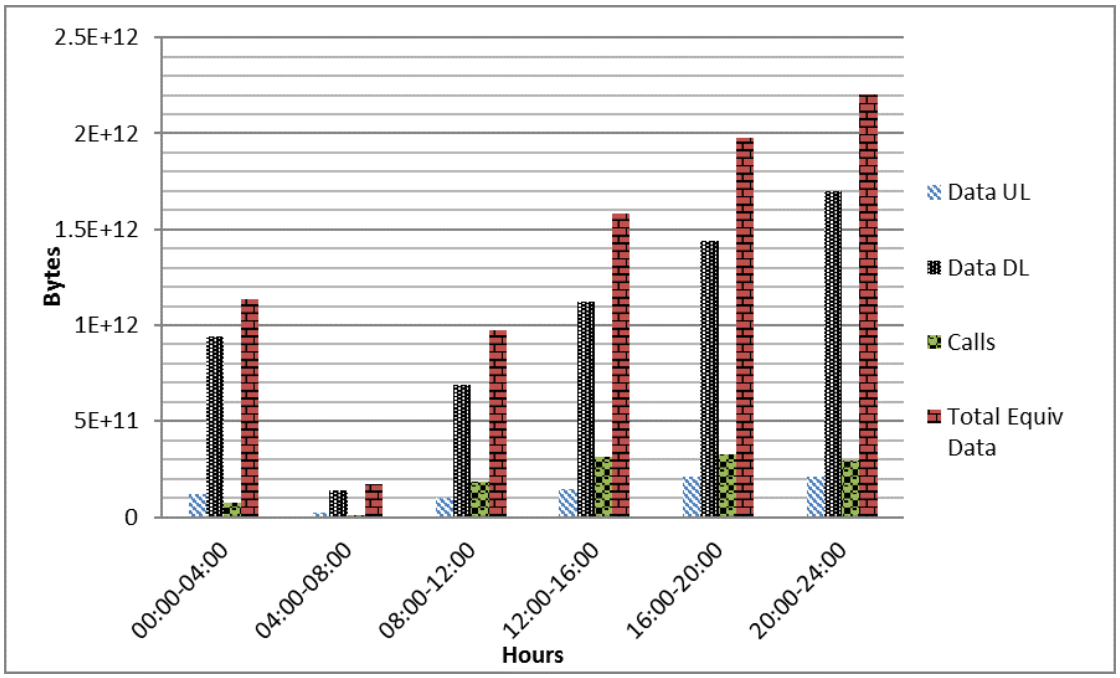


Figure 3.3: The total load on the network for a typical Monday broken down by traffic type and four hour period.

3.2.5 Qualifying and Quantifying Cellular Data Usage

Subsection 3.2.4 demonstrated the crucial role of cellular data usage when examining the total load present on the network. This subsection introduces a method to qualify cellular data usage into its constituent parts and then proceeds to quantify the contribution of each part. The dataset employed in this work is limited to CDR as discussed in Chapter 2 and, thus, does not directly contain information on what the purpose of each data session was. If packet header information was available, such as in [15], the purpose of each data session would be clearer. To overcome this limitation in the available data a classification step is required to classify the data usage into broad constituent parts. Figure 3.4 displays the clustering of activities according to: data volume, duration, and download to upload ratio. The plot suggests that there are a

number of clusters representing usage modes. From Figure 3.4 the following categories of usage can be identified:

- Short rapid communications (Apps): These activities correspond to small quantities of data used over short durations, generally less than 10 kB and for less than 60 seconds. This is representative of GPS updates, app interactions, advertising updates etc.
- Long duration, large volumes, mostly downloads (traditional): This consist of connections where large quantities of data are transferred asymmetrically (several Mbytes with large download to upload ratio) over an extended period of time. This is the traditional asymmetric usage mode of downloading webpages and other media consumption.
- Similar download/upload ratios, significant data volumes, less than 20 minutes (P2P Video/Voice): This suggests 1:1 communication with roughly equal data upload and downloaded. The average data rate for this category is 120 kbps. Alternatively, it could be file sharing, however, in that case the download to upload ratio would normally favour downloads.
- Fast, high data rates, mostly download, medium duration (Video): These sessions are classified by short bursts of high speed data usage with a large download to upload ratio.
- Long-time connections, low data volumes, similar upload/download ratios (Instant Messaging (IM)): In these sessions, the download to upload relationship is more symmetric with the connection not regularly timing out. This is indicative of two users communicating with one another but with insufficient data rates for voice or video which suggests text based instant messaging.

The bulk of the mobile data used can be broadly separated into the five categories which are quantified in Table 3.1. The categories outlined above and in Table 3.1 represent 84% of data connection events (Figure 3.5) and 89% of data volumes (Figure 3.6).

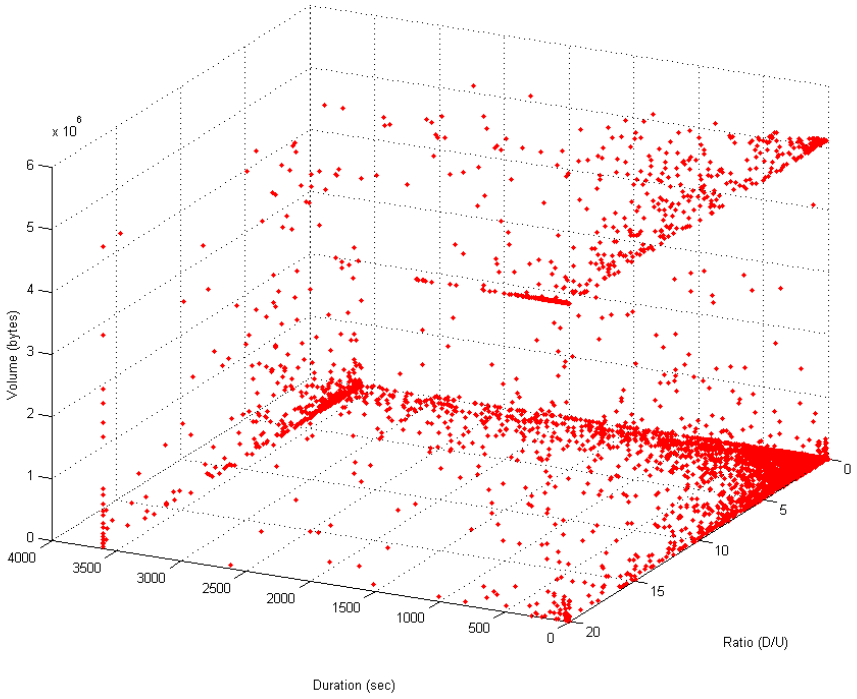


Figure 3.4: Plot of the data usage characteristics for one day (sampled 1:1000).

Table 3.1: Data usage categories

	Time	D/U Ratio	Volume
Apps	< 120 s	---	< 256kB
Traditional	> 300 s	>5	> 1 MB
Peer2Peer	<1200 s	>0.5, <1.5	> 256 kB
Video	< 300 s	>5	> 4 MB
Instant Messaging (IM)	> 600 s	< 2	< 256 kB

Figure 3.5 shows that the majority of connections to the network are still SMS, followed by data connections and then voice. However, as shown in [47] SMS is projected to shrink while data connections grow in importance. The data portion of the connections is further broken down into the categories listed in Table 3.1. Interestingly, app connections account for a plurality of data connection events, followed by video, IM and traditional browsing. Figure 3.6 displays the breakdown of cellular data usage volume into the categories presented in Table 3.1. Figure 3.6 clearly shows the predominance of video on network load; video accounts for the majority of data used on the network at 63%. Interestingly, despite accounting for a majority of data volumes video only represents 17% of data connections. When one considers not just mobile data connections but connections regardless of service type (only 20% of which are mobile data (Figure 3.2 (a))), videos proportion of all connections falls to 3.4%. From the perspective of total network load, including all service types, video accounts for 63% of the 90% that is mobile data (Figure 3.2 (b)). Thus video accounts for 56.7% of total network load regardless of service type while only being 3.4% of connections. This compares to a global average of 53% reported by [39] in 2013. Video clearly places a largely disproportionate load on network resources and managing it is a key task for network operators. Upgrading the network to newer technology such as LTE is one step although, as discussed in the following sections, when users get more capable devices they tend to consume more. Other options to curtail demand are available to operators such as pay per MB, usage caps, fair usage policies, etc. These features are already common on networks and all are employed on the network studied in this work. Tweaking these pricing instruments to balance quality of service while remaining competitive is key to an operator's viability. Another possible option for network operators is differential pricing bands for highly demanding video applications, receiving fees from preferred video content providers, throttling certain services, etc. However,

net neutrality regulations would currently prevent many of these options from being implemented [48].

Comparing Figure 3.5 and Figure 3.6 one sees that although app connections account for a plurality of data connections on the network (45%), they account for less than 1% of the total data volume. An established connection between the User Equipment (UE) and the network consumes a larger amount of energy in the UE than when the UE is not connected while also consuming network resources. Thus, after a period of inactivity from the UE the network ends a connection; this amount of time is usually a few seconds and is specified by the network's inactivity timer [33]. From the network operator's perspective, each change between connected and disconnected states causes a signalling load in the network. This load, if great enough, can cause network disruptions as discussed in [49]. These app connections disproportionately affect the signalling load on the network by constantly sending keep-alive messages, polling for data, etc. As discussed in [49] network operators can alter network parameters to ameliorate the deleterious effect of these repeated app connections. Of course a balance must be found between managing the signalling load on the network and a possible resultant deterioration in user experience [49]. App creators could also help by being mindful of the implications of their design decisions on the wireless network resource. For example, in 2013 Facebook released a software update to its Android and iOS app which single-handedly drove up signalling load and airtime consumption on some networks by 5-10% [50]. Better app design would benefit network operators through lower capital expenditure, users through better battery life, the environment through lower energy consumption from both the UE and network equipment and the app designer by making their apps more attractive to end users [49].

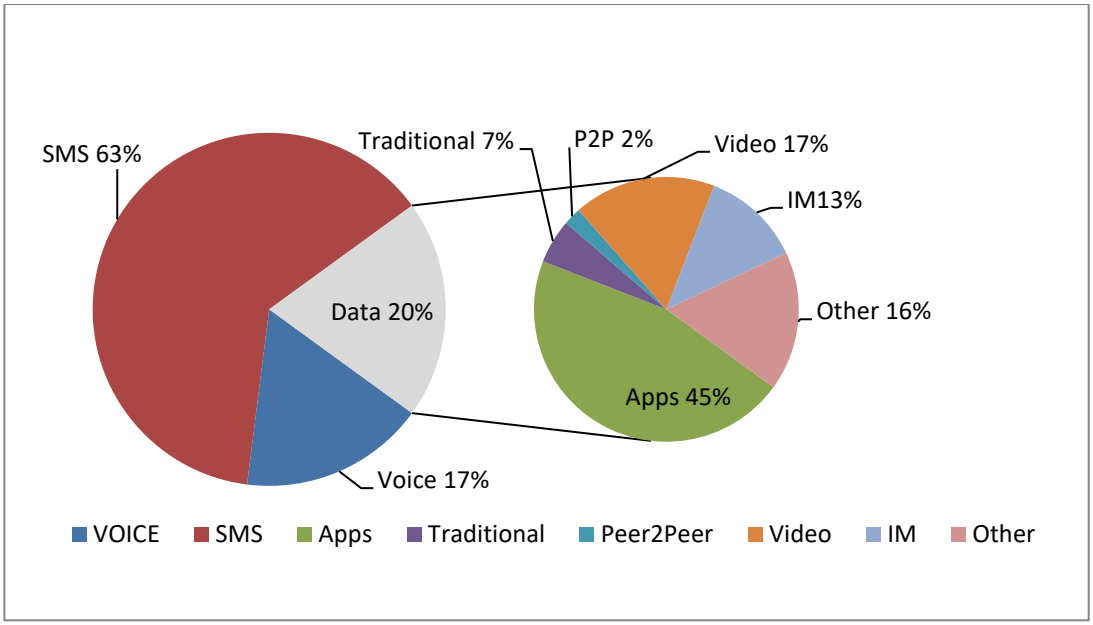


Figure 3.5: The pie chart on the left shows the total proportion of usage events by service type on a typical day. The pie chart on the right shows the breakdown of the cellular data segment into its constituent parts.

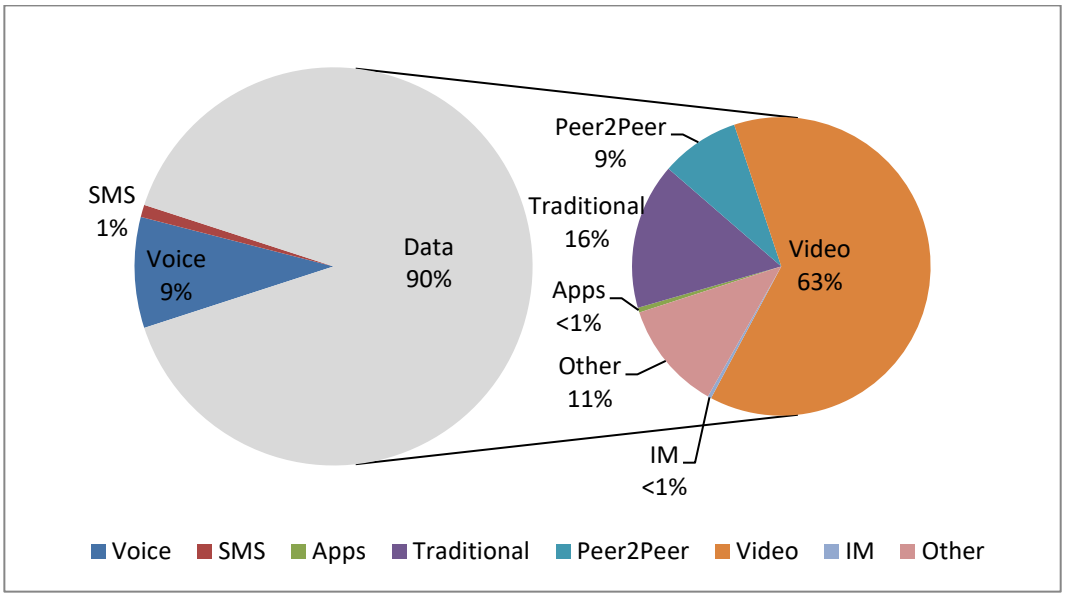


Figure 3.6: The pie chart on the left shows the total volume of data transferred over the whole network expressed as TED broken down by usage mode on a typical day. The pie chart on the right shows the breakdown of the cellular data segment into its constituent parts.

3.2.6 Conclusion

This section provided an empirical measurement of the network load and its constituent parts at the network level. Subsection 3.2.2 proposed and defined a novel metric to allow for the accurate comparisons of data volumes and load across disparate services and technologies. Subsection 3.2.3 presented and discussed the aggregated network wide daily load curve. Subsection 3.2.4 discussed how the network load breaks down between the three main constituent parts i.e. voice, data and SMS. Subsection 3.2.5 implemented a classification system to categorise data connections into one of several different types.

This section also identified and quantified some of the main trends and opportunities related to overall network load. For example, 3.2.3 identified and quantified the peaking problem on this network which is the source of much of the networks underutilisation of resources. Subsection 3.2.4 identified the trend that the problem is likely to be further exacerbated by more data usage in the future. 3.2.5 identified the predominance of video data on the network and some of the challenges it poses. Subsection 3.2.5 also identified the vastly disproportionate signaling load placed on the network by apps and discussed some ways to ameliorate this problem.

3.3 Local Load Distribution

3.3.1 Introduction

A list of three main contributions was provided in this chapter's introduction, the second of which was "*provide an empirical measurement of network load and its constituent parts both at the network level and the level of the individual base station/cell*". This section completes this objective (which was started in 3.2) by

providing an empirical measurement of the network load and its constituent parts at the level of individual base stations/cell. This chapter's third main contribution "*To use quantitative and qualitative analysis of the network including both its load and topography to identify trends and possible opportunities for resource rationalization*" is also completed in this section by identifying and quantifying some of the main trends and opportunities related to the topography of the network and localized load demands.

3.3.2 Local Load Distribution

Figure 3.7 (a) shows the distribution of daily loads (TED) serviced by base stations across the network broken down by day while Figure 3.7 (c) presents the same information broken down by base station sector (cell). The amount of load serviced varies by several orders of magnitude from a few megabytes up to tens of gigabytes. The key parameters of the distributions are presented in Table 3.2. These highlight the great variability in load serviced by different portions of the network; the busiest base station handles 2000 times the load of the least used base station. Comparing the base station loads presented in Figure 3.7 (a) with the data presented in [15] highlights the massive growth in data usage in the intervening years (the dataset in [15] is from 2007, predating the widespread adaption of smartphones). The median load on a base station in the network presented in [15] is approximately 15MB or one hundred times less than the median base station load of 1.5 GB as outlined in Table 3.2. The distribution of load has a positive skew of 2.25 at the base station level and 2.9 at the individual cell level. Skew is defined as the difference of a distributions mean and median divided by the distribution's standard devation. A positive skew means that the right tail of the distribution is longer i.e. there are more base stations/cells with below average loads and a smaller amount with much larger loads. This is a common feature of cellular

networks and a natural consequence of the network topography. Smaller cells (capacity cells) cover densely populated urban regions with high data demands while larger cells (coverage cells) provide mainly call and SMS coverage in sparsely populated rural areas (see Chapter 4 for a more detailed discussion). The individual cell level has a higher skew value than the base station level with a higher coefficient of variation c_v. The coefficient of variation is defined as:

$$c_v = \frac{\sigma}{\mu} \tag{3.1}$$

where σ is that standard deviation of base station/cell load and μ is the mean base station/cell load. Working out c_v for the base stations gives c_{vBS} = 2.8/2.5 = 1.12, while the equivalent value for cells c_{vCells} is c_{vCells} = 1.2/0.85 = 1.41. Thus the cells have a higher variability relative to their mean than base stations. This makes their load harder to predict and will be discussed in further detail in Chapter 5.

Table 3.2: Descriptive statistics of BS and sectorised cell load for typical weekday.

	Min	Max	Median ($\tilde{x}$)	Mean (μ)	SD (σ)	Skewness (γ)
Base Stations	14 MB	29 GB	1.5 GB	2.5 GB	2.8 GB	2.25
Cells	0.5 KB	14 GB	400 MB	850 MB	1.2 GB	2.9

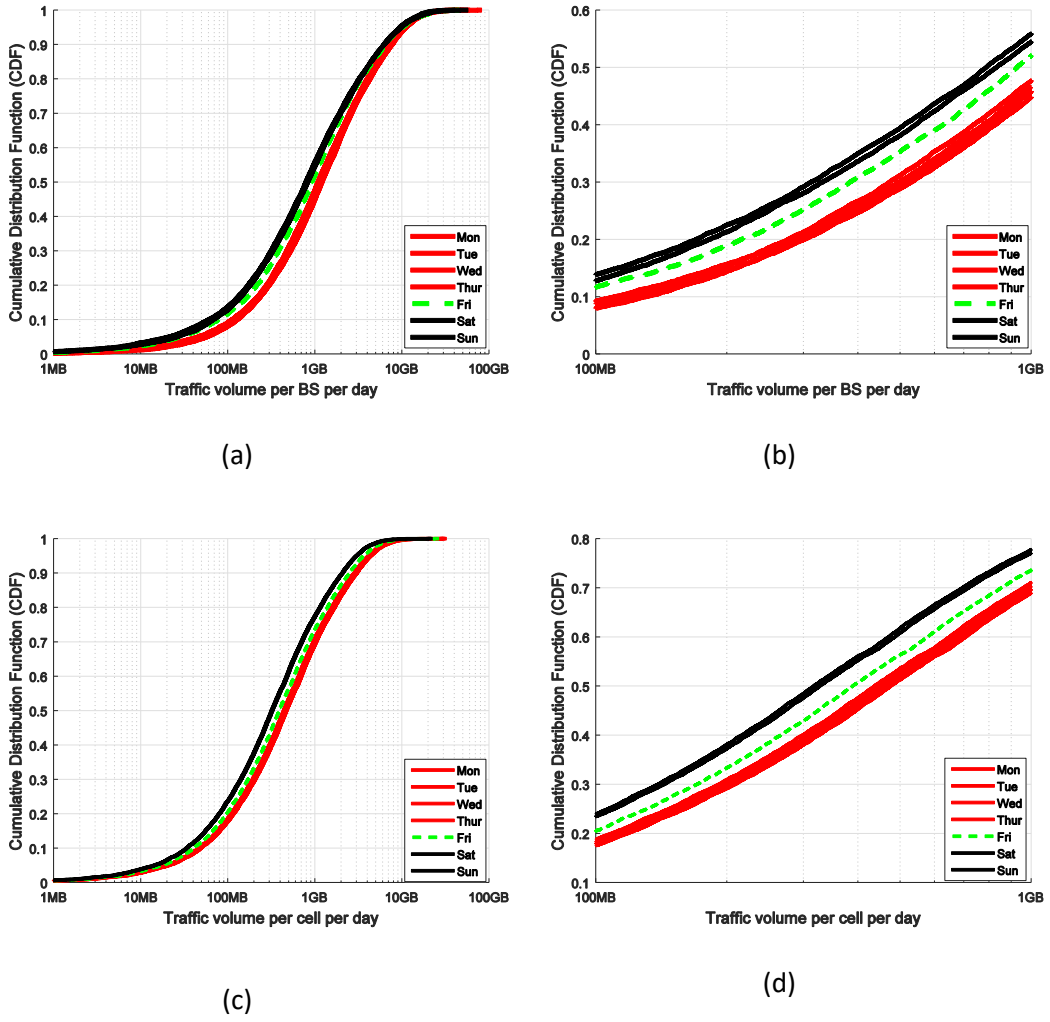


Figure 3.7: (a) CDF of the daily traffic (both uplink and downlink incl. cellular data, SMS and voice calls) per base station broken down by day of the week. (b) Zoomed in version of (a). (c) CDF of the daily traffic per cell broken down by day of the week. (d) Zoomed in version of (c). Note the similarity between Mon-Thur on all figures

Figure 3.7 (a) and Figure 3.7 (c) show the large variation in the daily traffic load serviced by individual base stations and individual cells on the network. Figure 3.8 & Figure 3.9 further demonstrate this by presenting the percentage of total network load serviced by a given percentage of the base stations/cells. Figure 3.8 shows that the most heavily loaded 1% of base stations service 12% of all network load. This is less than the equivalent figure of 20% from a 2007 dataset reported in [15] but larger than projected values in the future [51]. It appears that as total network load increases the load on the

network begins to spread between base stations and cells more evenly. This makes intuitive sense – due to economic factors, as the network grows the more densely populated areas receive the newest and most capable technology first. The less profitable areas are left with older less capable technology, discouraging or stifling use (e.g. areas with only GPRS/EDGE for data as opposed to HSDPA/LTE etc.). In time however, the networks capability to handle data spreads more evenly and the disparity begins to reduce. The imbalance is more acute at the level of specific sectorised cells as seen when comparing Figure 3.8 and Figure 3.9. For example 20% of network traffic is serviced by 1% of cells while at the base station level the top 1% of base stations service only 12% of the network load. Examining the usage patterns of individual subscribers reveals that a relatively small number of subscribers are responsible for a disproportionately large portion of the overall network traffic. In Chapter 4 the home and work locations of these subscribers are derived from a novel analysis of the data set. Doing so reveals that the presence of these heavy users in certain cells is an important factor in the disparity of cell loads.

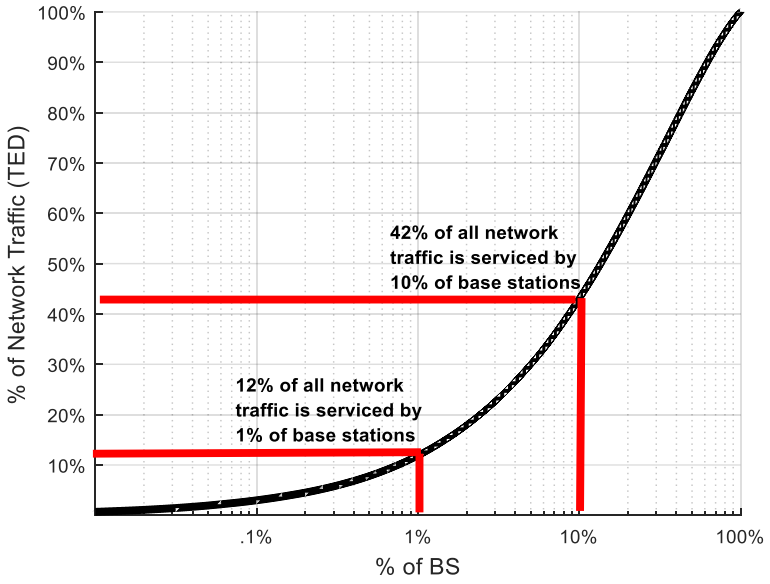
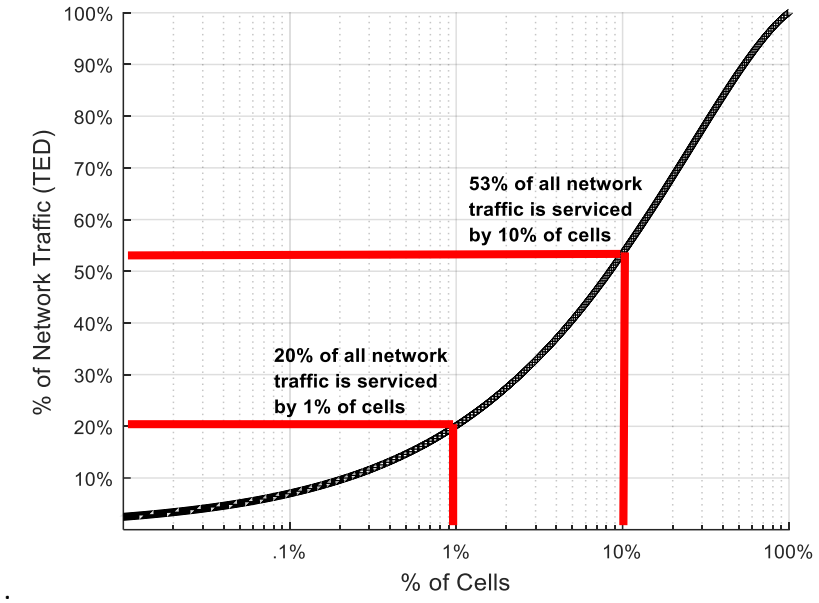


Figure 3.8: The percentage of total network traffic (TED) serviced by a given percentage of base stations



.

Figure 3.9: The percentage of total network traffic (TED) serviced by a given percentage of cells.

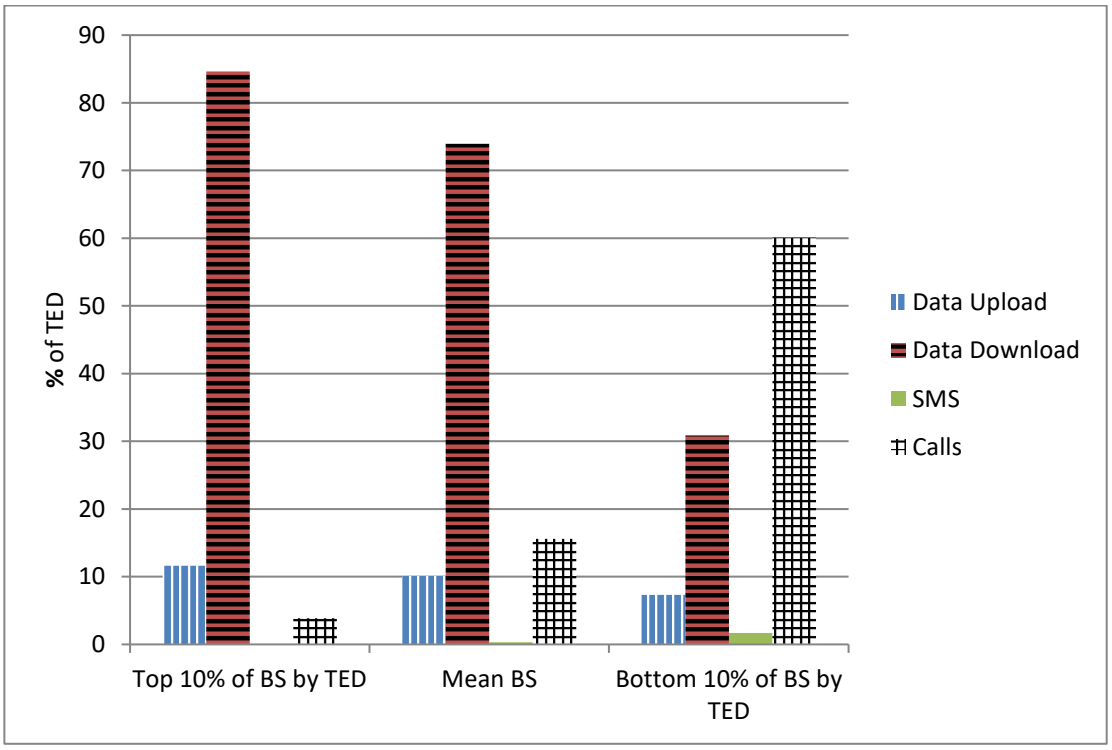


Figure 3.10: The load broken down by traffic type for three groups of BS as a percentage of overall traffic volume TED.

Figure 3.10 shows the load broken down by traffic type for: the top 10% of heavily loaded base stations, the bottom 10% of base stations by load and finally the load on the mean base stations. Figure 3.10 clearly shows that the highly loaded cells are almost exclusively loaded with cellular data while in the least loaded cells voice calls predominate. This is in keeping with the distinction between coverage and capacity cells as discussed previously.

3.3.3 Conclusion

This section provided an empirical measurement of the network load and its constituent parts at the level of individual base stations/cells. This section also identified and quantified some of the main trends and opportunities related to the topography of the network and localized load demands. A great disparity in network load was identified at the individual base station and cell level. For example, the base station with the heaviest load handles approximately two thousand times the traffic of the base station with the lowest load. On the network level 12% of the network's traffic is serviced by just 1% of the base stations. This result is even more extreme at the level of individual cells where 1% service 20% of the total network load. This disparity between cells coupled with the temporal peaking problem identified in 3.2 make clear the potential for greater resource rationalisation. Several methods of achieving this are possible, ranging from dynamic spectrum access, where valuable spectrum is shared between licensed primary and unlicensed secondary users [12], to the dynamic switching off of equipment to conserve energy as will be discussed in Chapter 7.

3.4 Models of Network Load

3.4.1 Introduction

Given the importance to overall network load of data sessions (Figure 3.2) this section provides empirically created models for the three most important aspects of data sessions:

1. Interarrival times of data sessions
2. Data session durations
3. Mean data session throughputs.

With these empirically created models the data usage on this network can be modelled by interested parties. This section also provides a novel breakdown of the models both by access device class and contract type. Voice and SMS have been modelled in previous works and the results produced on this network are similar, so to save space and avoid replication they are omitted. For empirical models of voice and SMS usage see [12].

3.4.2 Modelling Interarrival Time

Models of the interarrival times/arrival rates are important for creating accurate usage scenarios of how subscribers request network resources. The arrival rate is the number of arrivals per unit of time while the interarrival time is the time between each arrival into the system and the next. When modelling time series data an important consideration is the timescale over which the data to be modelled is stationary i.e. the timescale over which the model parameters such as mean and variance do not change. However, when modelling one also wishes to aggregate over timescales that are as large as possible to reduce the standard error (this becomes more of a problem when examining individual base stations with low arrival rates). To aid in the choice of an aggregation timescale Figure 3.11 shows how the network wide normalised average

data session arrival rate varies over four days, two representative weekdays and two representative weekend days.

Figure 3.11 demonstrates some key aspects of the network's arrival rate.

1. There are two distinct periods which approximate day and night times. The daytime period has high arrival rates in comparison to the night time period.
2. The greatest change in arrival rates occur during the latenight/early morning hours and the late morning hours. These intervals coincide with the transition from the day to night period and vice versa.
3. Apart from the transitional periods, the mean arrival rates appear (relatively) stationary over the course of 30 minutes.
4. Weekdays and weekends appear to show different trends in arrival rates over the course of the day. This is to be expected due to the change in many subscribers' schedules between weekdays and weekends as discussed in 3.2.3.

Taking the aforementioned points into consideration an aggregation of 30 minutes approximates the stationary behaviour desired.

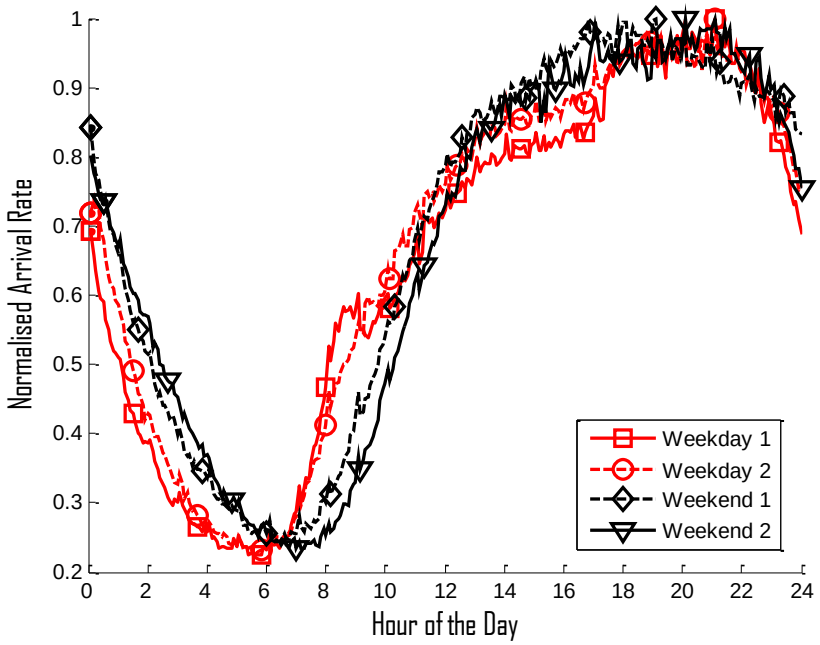


Figure 3.11: Normalised arrival rate by time of day

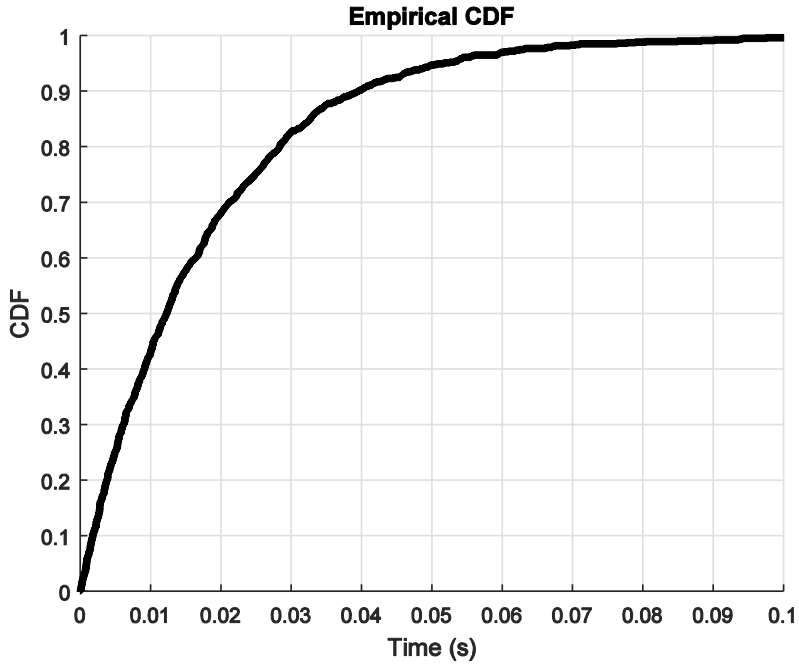


Figure 3.12: CDF of interarrival time over whole day

Figure 3.12 shows the empirical CDF of data session interarrival times on the network for an entire day. The CDF of a real-valued stochastic variable *X* is the function given by:

$$F_X(x) = P(X \leq x), \tag{3.2}$$

where the right hand side gives the probability that the stochastic variable *X* has a value less than or equal to *x*. The empirical CDF F_n for *n* independent identically distributed (iid) observations X_i is defined as:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x} \tag{3.3}$$

where $I_{Xi \leq x}$ is the indicator function which equals 1 if $X_i \leq x$ or 0 if $X_i > x$

The Interarrival time in cellular networks has traditionally been modelled as an exponential distribution [12] such as:

$$F_x(x) = 1 - e^{-\Phi x} \tag{3.4}$$

where *x* is the inter arrival time and Φ is the adjustable weight parameter. However, these models were primarily for calls and SMS, predating the widespread adoption of smartphones and the move to a more data centric network [24]. Figure 3.13 plots the interarrival time for two different periods of the day with their respective exponential fits (via non-linear least squares) of the form given in (3.4). There is a large difference in the interarrival time distributions between these distinct periods as would be expected given their differing arrival rates as plotted in Figure 3.11. Visually the fits are quite accurate with low respective RMSE as shown in Table 3.3, suggesting that the interarrival process for data can be modelled in a similar fashion to calls and SMS. These empirically crated models of the interarrival times are important for creating accurate usage scenarios of how subscribers request network resources. They will allow the interested reader to recreate the data connection request process without access to the original dataset.

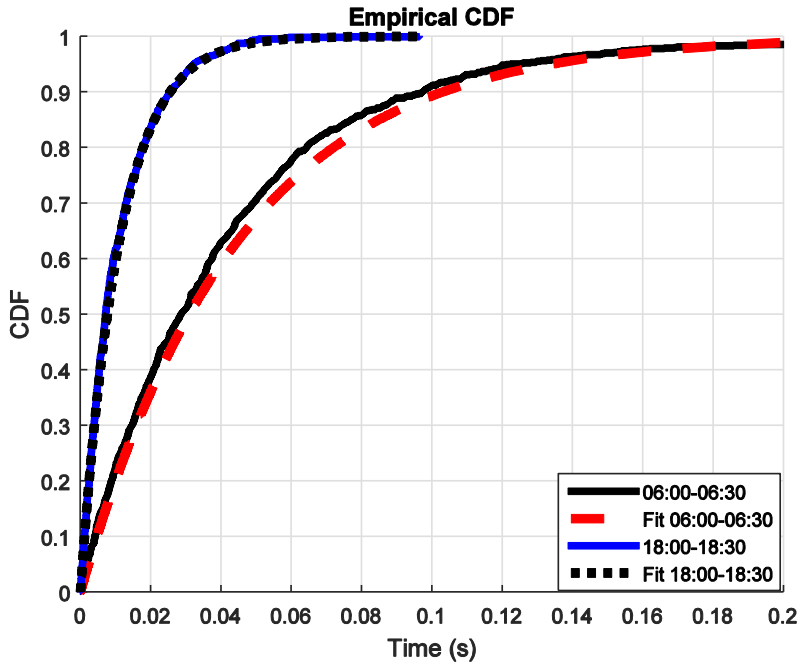


Figure 3.13: CDF of interarrival time for a period of low usage (06:00-06:30) and a period of high usage (18:00-18:30) with their respective exponential fits of the form given in (3.4) and with the parameters provided in Table 3.3.

Table 3.3: Interarrival time fit parameters by time period

Time Period	Φ	RMSE
00:00-00:30	72.49	.0060
06:00-06:30	22.35	.0099
12:00-12:30	75.99	.0051
18:00-18:30	90.23	.0047

3.4.3 Modelling Connection Duration

This subsection examines the distribution of data session durations and how they can be modelled. Initially results are presented for a general model, then, more detailed models broken down by the device/contract type are provided.

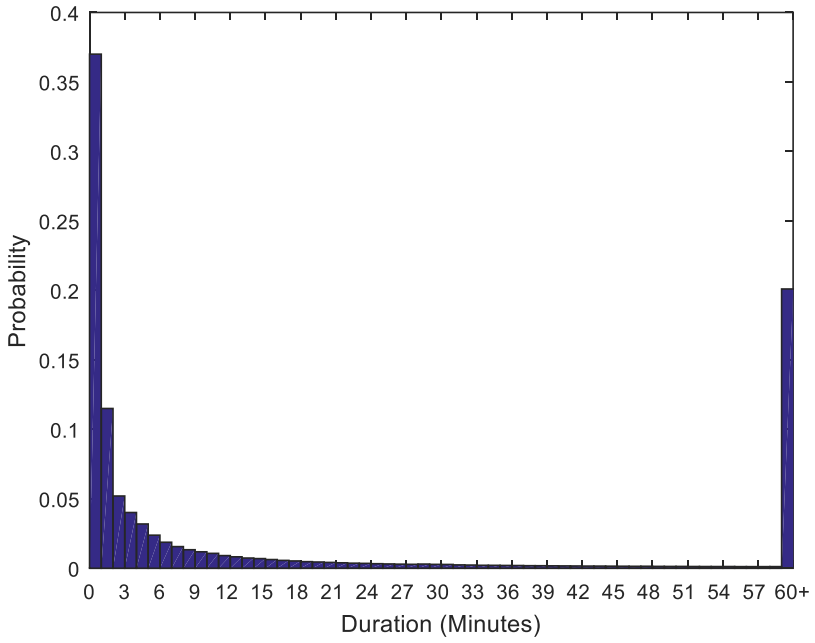


Figure 3.14: Histogram of data session durations. Each bin represents one minute, except for the final bin representing all durations ≥ one hour.

Figure 3.14 plots the histogram of data session durations with each bin representing one minute (except for the final bin which represents all times greater than one hour). Figure 3.14 illustrates that short duration data sessions (≤ 3 minutes) dominate accounting for approximately 50% of all connections. The predominance of these short connections is no surprise given that short app connections form a plurality of data connections as demonstrated in Figure 3.5. Longer durations are much scarcer with only approximately 20% of data sessions lasting one hour or more.

These values are broken down further in Figure 3.15 which shows how the CDF of duration varies by the connecting device type. For example, Figure 3.15 shows that a disproportionate amount of the connection durations over 60 minutes long are from mobile internet USB dongles. Conversely Figure 3.15 reveals that a disproportionate number of the short connections come from feature phones.

Comparing smartphones with feature phones shows that feature phones connect to the network for much shorter periods. Their median connection time (160s) is less than half that of prepay smartphones (350s) and under a third of bill pay smartphones (550s). Interestingly at the time of data collection feature phones were much more likely to be on prepay price plans than smartphones. This coupled with a poorer interface and experience could partially explain the difference.

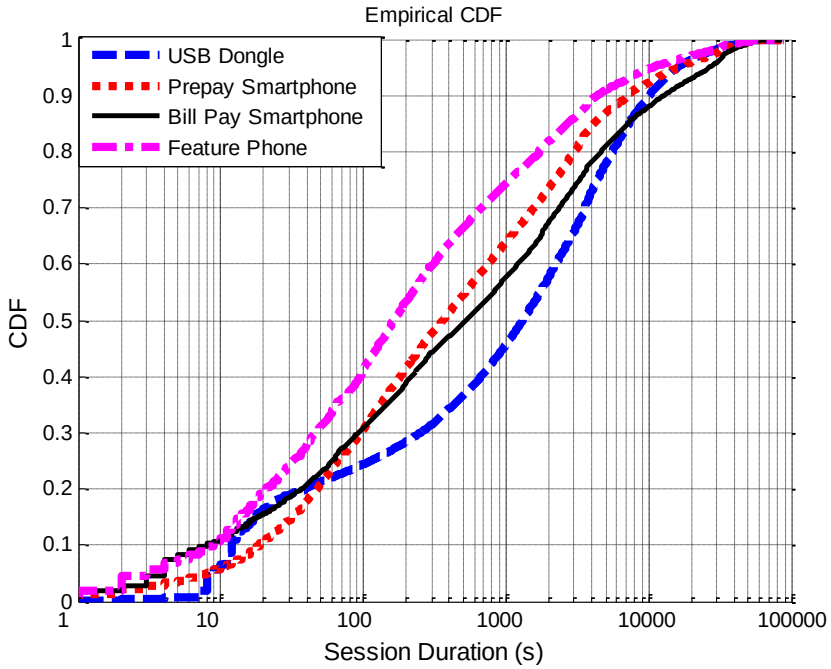


Figure 3.15: CDF of data session durations broken down by connecting device type.

As seen in Figure 3.1 there appears to be approximately two distinct periods of usage during the day – a early morning period and a daytime/night-time period. Also [12]

reported that there are two distinct call duration distributions, one for the night time and one for the daytime. To investigate if this is also the case for the durations of data sessions, the variation in the empirical CDF of the hour-wise durations is compared to the overall daily empirical CDF as was suggested for call durations in [12]. To do this the Kolmogorov-Smirnov statistic [52] is computed. This is the maximum difference between the overall empirical CDF and the hourly empirical CDF. The Kolmogorov-Smirnov statistic for two samples is defined as:

$$D_{n,n'} = \mathrm{Max}\left|F_{1,n}(x) - F_{2,n'}(x)\right| \tag{3.5}$$

where $F_{1,n}$ and $F_{2,n'}$ are the empirical CDFs (see equation(3.3)) of the first and second samples respectively while sup is the supremum function.

However, on repeating the methodology of [12] and comparing the variation in the empirical CDF of the hour-wise durations to the overall daily empirical CDF no significant distinct daily periods of data session durations were found. Thus, it appears that unlike call durations, the distribution of data durations is not broken into distinct daily periods.

Figure 3.16 shows the empirical CDF of data session duration distribution. Note that the duration value for a particular session is assigned to the time period in which it was initiated. The duration distributions resemble a lognormal distribution and are modelled as such in Figure 3.16. The Probability Density Function (PDF) of the lognormal distribution of the data session durations can be reproduced via:

$$f_x(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} \tag{3.6}$$

where x is the data session durations, μ are the data sessions' mean duration and σ is the standard deviation of the data sessions durations [54]. The fit applied to the data session durations in Figure 3.16 can be reproduced from the PDF described in equation (3.6) using the input parameters in Table 3.4 and the method of CDF calculation used in

(3.2) and (3.3). Visual inspection of the goodness of fit in Figure 3.16 coupled with the small RMSE reported in Table 3.4 supports the efficacy of log normal fits for cellular data session durations.

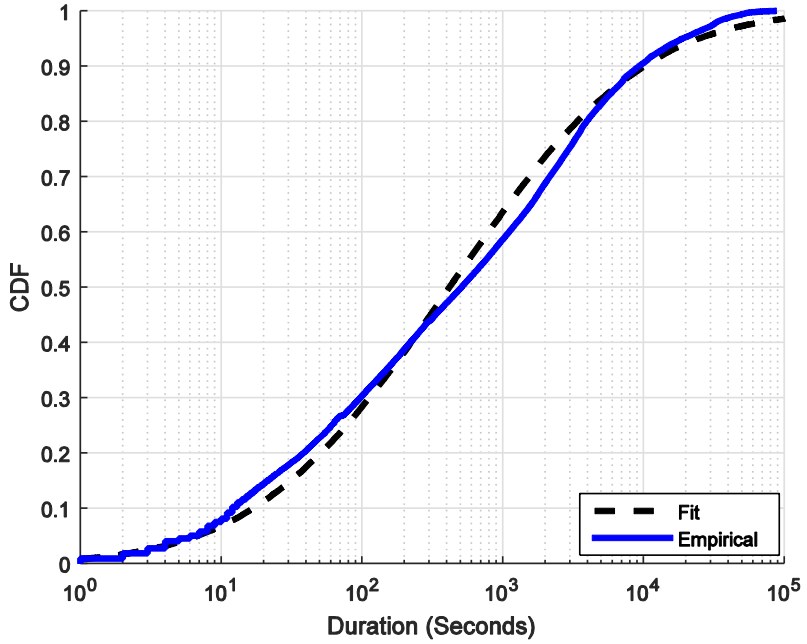


Figure 3.16: Data session duration distribution and lognormal fit.

Table 3.4: Parameters for lognormal model of data session duration distributions.

Distribution	μ	σ	RMSE
Data Session Durations	6.01894	2.49531	.0396

Figure 3.15 demonstrated that the distribution of connection durations is highly dependent on the type of device connecting to the network and to a lesser extent the type of contract the user has with the network (bill pay v prepay). Thus, Figure 3.17 illustrates some of the results for modelling the distribution of data session durations broken down by connection type and contract type. The complete list of parameters used to produce fitted models similar to those in Figure 3.17 for all the distinct device/contract type identified in Figure 3.15 are presented in Table 3.5.

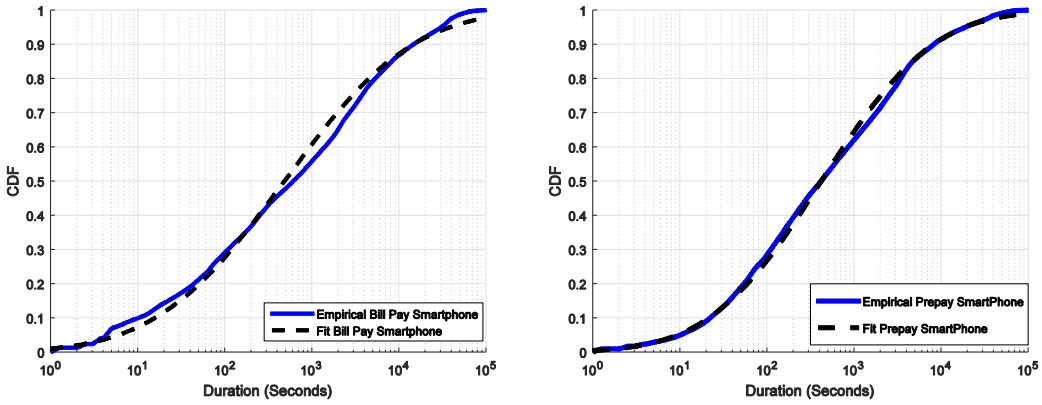


Figure 3.17: Data session duration distributions for: (left) bill pay smartphone connections and their lognormal fits, (right) prepay smartphone connections and their lognormal fits

Table 3.5: Parameters for lognormal models of data session duration distributions.

Distribution	μ	σ	RMSE
Bill Pay Smartphone	6.17992	2.65923	.0451
Prepay Smartphone	5.87245	2.30635	.0236
Feature Phone	5.19	2.39028	.0373
USB Dongle	6.64844	2.44977	.0874

3.4.4 Modelling Mean Throughput

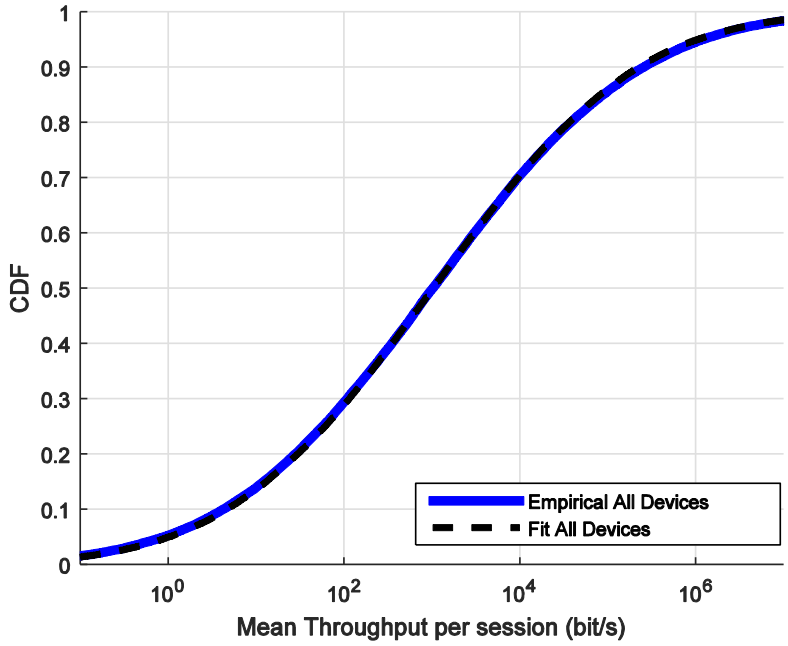


Figure 3.18: Empirical Mean Throughput per session (bit/s) & lognormal fit

Figure 3.18 shows the empirical CDF of the throughput per session distribution; the throughput distribution resembles a lognormal distribution and is modelled as such in Figure 3.18 in a similar fashion to the models in 3.4.3. The distributions may be reproduced as lognormal distributions with the following input parameters:

Table 3.6: Parameters for lognormal model of mean throughput per data session

Distribution	μ	σ	RMSE
Daytime/Night-time	6.92205	4.19348	.0034

Figure 3.15 illustrated that device type played an important role in determining the mean duration of a data session connection. A natural follow-on from this is to explore the role played by device type in the mean throughput. Figure 3.19 demonstrates the great disparity that exists in mean throughput between the different connecting

devices. USB dongles have a much larger mean throughput than the other devices connecting to the network with a median value of 75kbps versus just 10bps for prepay smartphones. Interestingly, the mean throughput of feature phones is on par with bill pay smartphones and greater than that of prepay smartphones. It is worth reiterating here that this refers to mean throughput for each session and not instantaneous throughput. This gives users who stay connected for a long period (while functionally inactive or passively consuming tiny amounts of data through small app updates etc.) a greatly diminished mean throughput. The large disparity between mean throughput for bill pay and prepay is also a striking feature of the results with bill pay users having a median throughput ten times greater than their prepay counterparts. One possible explanation for this disparity in mean throughput is that bill pay customers may be more likely to use data intensive applications such as video streaming given they have a set amount of cellular data allocation each month. Prepay customers on the other hand pay per byte and thus may be more likely to restrict data intensive high usage applications such as video streaming or offload this to WIFI networks. This disparity underlines the importance of also considering contract type when producing models of usage. The parameters of the lognormal fits by device type and contract type are provided in Table 3.7 allowing the interested reader to reproduce the distributions.

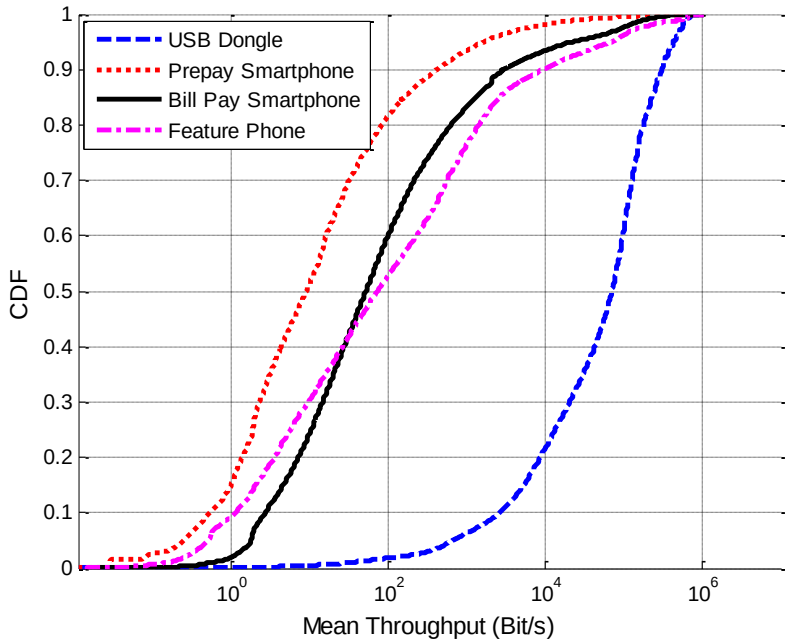


Figure 3.19: Mean throughput per data session broken down by device

Table 3.7: Parameters for lognormal models of mean throughput per session

Distribution	μ	σ	RMSE
Bill Pay Smartphone	4.32686	2.8086	.0027
Prepay Smartphone	2.43485	2.69727	.0053
Feature Phone	4.33676	3.48404	.0089
USB Dongle	10.5527	2.06099	.0269

3.4.5 Models of Network Load Conclusion

This section provided foundational, empirically created models of how the network experiences load. The three fundamental aspects of data sessions from a network operator perspective were modelled:

1. Interarrival times of data sessions.
2. Data session durations

3. Mean data session throughputs

These empirically created models of data usage on this network will allow other interested parties to recreate these models for their own use. This subsection also provided a novel breakdown of the models both by access device class and contract type. It was shown that short connections (<3 minutes) predominate on the network accounting for approximately 50% of all connections (primarily comprised of app interactions as discussed in 3.2.5). A difference in median connection time was discovered between smartphones depending on the nature of the contract with the median prepay smartphone connection lasting approx. 400 seconds compared to approx. 500 seconds for bill pay smartphones. However, the longest connections by far came from USB dongles with a median connection time of approx. 1500 seconds. Unlike call durations, the time of day was not found to have an impact on data session durations. Empirically created models were provided for all possible permutations of connecting device type and contract type. Finally, the mean throughput of all data connections was modelled and then this was further broken down by connecting device type and contract type. Interestingly, contract type was found to be of crucial importance when considering mean throughput with bill pay smartphone connections having a median mean throughput ten times greater than prepay smartphone connections. These empirically created models will allow for the accurate recreation and modelling of these key network features, not only at the general level but crucially at the device and contract specific level.

3.5 **Conclusion**

The introduction of this chapter identified its three main contributions - each of these aims was accomplished in the succeeding sections. For example, the primary aim of this chapter was to provide empirically created foundational models of how the network

experiences load i.e. models of arrival rates, connection durations, and data consumption. These models were to be provided at a fine-grained level broken down by connection time, connecting device type and, finally, contract type. 3.4 achieved this objective by providing empirically created models for the three most important aspects of data sessions: (i) Interarrival times of data sessions, (ii) data session durations and (iii) mean data session throughputs. This section also provided a novel breakdown of the models by access device class and contract type. These empirically created models will allow for the accurate recreation and modelling of these key network features, not only at the general level but crucially at the device type and contract specific level.

The second contribution of this chapter was "*To provide an empirical measurement of network load and its constituent parts both at the network level and the level of the individual base station/cell*". This was achieved at the network level in 3.2 and at the level of the individual base stations/cell in 3.3. 3.2 provided a network wide examination of network load and introduced a classification system for CDR to allow for a detailed breakdown of data usage. 3.3 provided a more fine-grained approach to examining network load and focused on the local disparities between individual base stations/cells.

The final contribution of this chapter was "*To use quantitative and qualitative analysis of the network including both its load and topography to identify trends and possible opportunities for resource rationalisation*". Firstly, a metric for comparing load across service type was introduced. Then the peaking problem on the network was introduced and discussed. This is where peak time loads are an order of magnitude higher than trough time loads. This peaking problem was found to be getting relatively worse as more and more mobile data was being used on the network. When the mobile data connections were further analysed and classified it was found that the primary driver of mobile data usage on this network was video streaming. However, despite the

importance of video to the total volume of data transferred, when considering signalling overhead app connections were found to be having a vastly disproportionate impact. Some of the problems caused by this and possible solutions to this were discussed and identified. 3.3 identified the great disparity in load at the local level with the most highly loaded base stations having a load two thousand times greater than the least loaded base stations. On the network level 12% of the network's traffic is serviced by just 1% of the base stations. This result is even more extreme at the level of individual cells where 1% service 20% of the total network load. This disparity between cells coupled with the temporal peaking problem identified in 3.2 make clear the potential for greater resource rationalisation. Several methods of achieving this are possible, ranging from for example dynamic spectrum access where valuable spectrum is shared between licensed primary and unlicensed secondary users [12] to the dynamic switching off of equipment to conserve energy as discussed in Chapter 7.

Chapter 4 Spatial Usage in Cellular Networks

4.1 Introduction

The preceding chapter explored the network's load dynamics from a network wide perspective. Although that exploration was important and useful, it did not examine the highly *localised* nature of cellular networks; any examination of cellular networks is not complete without reference to their defining characteristic, spatial subsidiarity. For this thesis to complete its task of providing and examining practical Near Horizon Localised Load Forecasting models for cellular networks then a strong understanding of network spatiality is crucial. To that end this chapter focuses on the spatial properties and causal relationships present in the network. The primary contributions of this chapter are:

1. *The creation of a spatial representation of the entire network to allow for the association of load with defined spatial areas*. These defined coverage areas for both base stations and sectorised cells are the spatial building blocks of the network. In later chapters they will be modelled and their load predicted both individually and in larger spatial amalgamations.
2. *A novel procedure is introduced to clean inaccuracies in the spatial coordinates of cell towers*. Due to the importance of the spatial locations of base stations and sectorised cells in the following chapters, it is imperative that every effort is made to identify and exclude inaccuracies.
3. *A method to visualise how the load is distributed spatially across the network both as a whole and across various services*. This provides an important network

wide view of the load distribution which is a crucial element in understanding how the load varies spatially.

4. *The provision of a novel method to discover who lives and works within the defined spatial coverage areas introduced in point 1 and how they interact with other network users spatially.* It is axiomatic that the number of subscribers in an area will influence the load of that area. It is therefore useful from a network operator's perspective to understand how their subscribers are distributed throughout the network. However, cell phones are also known as *mobile* phones for a reason, and thus it is not enough to simply know where subscribers live. It is also important to understand where they spend large amounts of their time such as where they work.
5. *An examination of the degree, or lack thereof of spatial correlation in load across the network.* The previous chapter already highlighted that there is a large disparity in load across the network at the level of the individual base station/cell. Using the coverage regions introduced in this chapter the degree or lack thereof spatial correlation in load across these coverage regions is explored.
6. *A novel exploration of the presence/lack of causal influence between neighbouring cells within the network* i.e. an examination of whether a cell's load has any influence on neighbouring cells. The causality present in the network can be used to aid localised prediction of load, the identification of key cells/base stations whose failure would be particularly deleterious to user experience, travel mode discovery (paths taken by subscribers as they move throughout the network) etc.

The above contributions are valuable to network providers and relevant to many advanced network management techniques. They are particularly important to those

techniques which rely on a strong spatial understanding such as dynamic spectrum allocation [12], reduced sampling techniques [55], fault detection, and spatially influenced power saving schemes [56] such as the one presented in Chapter 7. The remainder of this chapter is laid out as follows:

- 4.2 examines the spatial representation of the network. 4.2.2 explains how the dataset presented in Chapter 2 can be represented by spatial coverage regions. 4.2.3 provides a novel algorithm to identify out-dated spatial information in the dataset and exclude it from further study. 4.2.4 presents a method of network wide load visualisation based on the spatial regions presented in the preceding sections.
- 4.3 provides an examination of how users communicate over spatial distance. 4.3.2 introduces a novel algorithm to calculate the home and work populations of each cell in the network. 4.3.3, 4.3.4, and 4.3.5 respectively introduce, implement, and test one possible model of spatial communication distance.
- 4.4 examines the spatial relationships and dependencies present within the network structure. 4.4.2 explores the spatial correlations present in the network's load. 4.4.3 - 4.4.7 explores the causal structure of the network's load.
- 4.5 provides a concluding discussion of the chapter's results.

4.2 Spatial Representation of the Network

4.2.1 Introduction

This section focuses on the creation of spatial representations of the network firstly at the localised base station and sectorised cell coverage level, then the aggregated network level. The spatial coverage region representations introduced in 4.2.2 are the foundational step in beginning to examine the network spatially. Much of the later work

and many of the techniques introduced later require the use of these spatial coverage regions. Given their importance 4.2.3 introduces a novel method to identify and remove errors in their positioning. 4.2.4 provides a method to visualise the spatial distribution of cellular load across the network. The techniques employed in 4.2.4 could be generalised to not only represent load distribution but also other properties of interest such as connection events, subscriber distribution etc.

4.2.2 Base Station and Cell Coverage Regions

As discussed in Chapter 2 BTS, Node-B and eNode-B may all be mounted on the same tower, with each servicing various spatially overlapping geographical regions. It is possible to approximate idealised cell site coverage areas via Voronoi tessellation [57] by using the geo-spatial coordinates and the network type of each cell, where each centre represents a base station site location. A Voronoi tessellation is a partitioning of a plane into regions based on distance to points in a specific subset of the plane [57]. Figure 4.1 depicts the base station site Voronoi tessellations areas for the 2G and 3G base stations on the network under examination (note Figure 4.1 - Figure 4.3 are placed together at the end of this subsection to facilitate their comparison). It is important to note that the accuracy of the tessellation in approximating base station coverage areas is affected by channel characteristics, topography of the area and physical layer parameters which include transmitter frequency, tilt, height, and transmission power [34]. The collection of this information is prohibitively expensive and, as such, is not factored into this analysis. Thus, it should be noted that the estimation technique applied does introduce some approximation error at a local level.

Figure 4.1 was created with the MATLAB plotting function. The MATLAB function VORONOI was used to create the Voronoi tessellations using the site locations as inputs. A polygon is returned for each unique site location, thus base stations with matching

site locations on the same network share the same site polygon. The county geographical regions polygons presented are sourced from Ordinance Survey Ireland [58]. Note that the coordinate system used in Figure 4.1 - Figure 4.3 is the Irish Grid Coordinate Reference System [35]. This coordinate system uses the projections of Easting and Northing, which are in units of meters from an origin point located at latitude 53°30'00 N and longitude of 8°00'00 W.

The polygons presented in Figure 4.1 are a reasonable approximation of inland coverage regions and coverage regions not adjacent to the border with Northern Ireland; however, the absence of a limiting threshold for polygon size means that coverage regions along the coast are less accurately approximated. Thus, to improve costal accuracy a maximum site radius, S_{max}, of 20 km and 15 km is introduced for 2G and 3G networks respectively. These limits reflect the realistic limits of communication within each standard given the network topology [33]. The site radius S_r for each site is calculated by

$$S_r = \min\left(\sqrt{\frac{S_a}{\pi}}, S_{max}\right) \tag{4.1}$$

where S_a is the coverage area of the base station's site defined as:

$$S_a = \frac{1}{2}\sum_{i=0}^{N-1}(x_i y_{i+1} - x_{i+1} y_i) \tag{4.2}$$

where N is the number of points in the coverage polygon and (x,y) are the spatial coordinates of each point.

Figure 4.2 shows the effect of introducing the base station coverage radius limit. The difference is particularly evident along the coast and border with Northern Ireland. Along these areas in Figure 4.1 the coverage regions stretched to infinity but are now

more realistically limited in Figure 4.2. The limited base station boundary is found by extracting the polygon of the spatial intersection of the idealised site coverage polygon with the circle of the maximum site size. This intersection is carried out via the POLYBOOL function from the mapping toolbox of MATLAB. Sectorised coverage regions of the larger base stations may be extracted by using the transmitter azimuth angle information in the tessellation. These sectorised coverage regions are shown in Figure 4.3; the restricted coverage regions displayed in Figure 4.2 are now subdivided into individual sectorised cells in Figure 4.3. To generate the coverage regions in Figure 4.3 each coverage polygon in Figure 4.2 is subdivided by the unique transmitter azimuth angles of cells associated with the site. It is important to note that sectorised cells at the same site sharing the same azimuth angle will share the same cell coverage polygon, C_p. The Cell radius (C_r) and Cell area (C_a) is calculated via equation (4.1) and equation (4.2) respectively. An individual cell's centroid Easting and Northing location, (C_x, C_y), is calculated by equations (4.3) and (4.4), respectively.

$$C_x = \frac{1}{6C_a}\sum_{i=0}^{N-1}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1}y_i) \tag{4.3}$$

$$C_y = \frac{1}{6C_a}\sum_{i=0}^{N-1}(y_i + y_{i+1})(x_i y_{i+1} - x_{i+1}y_i) \tag{4.4}$$

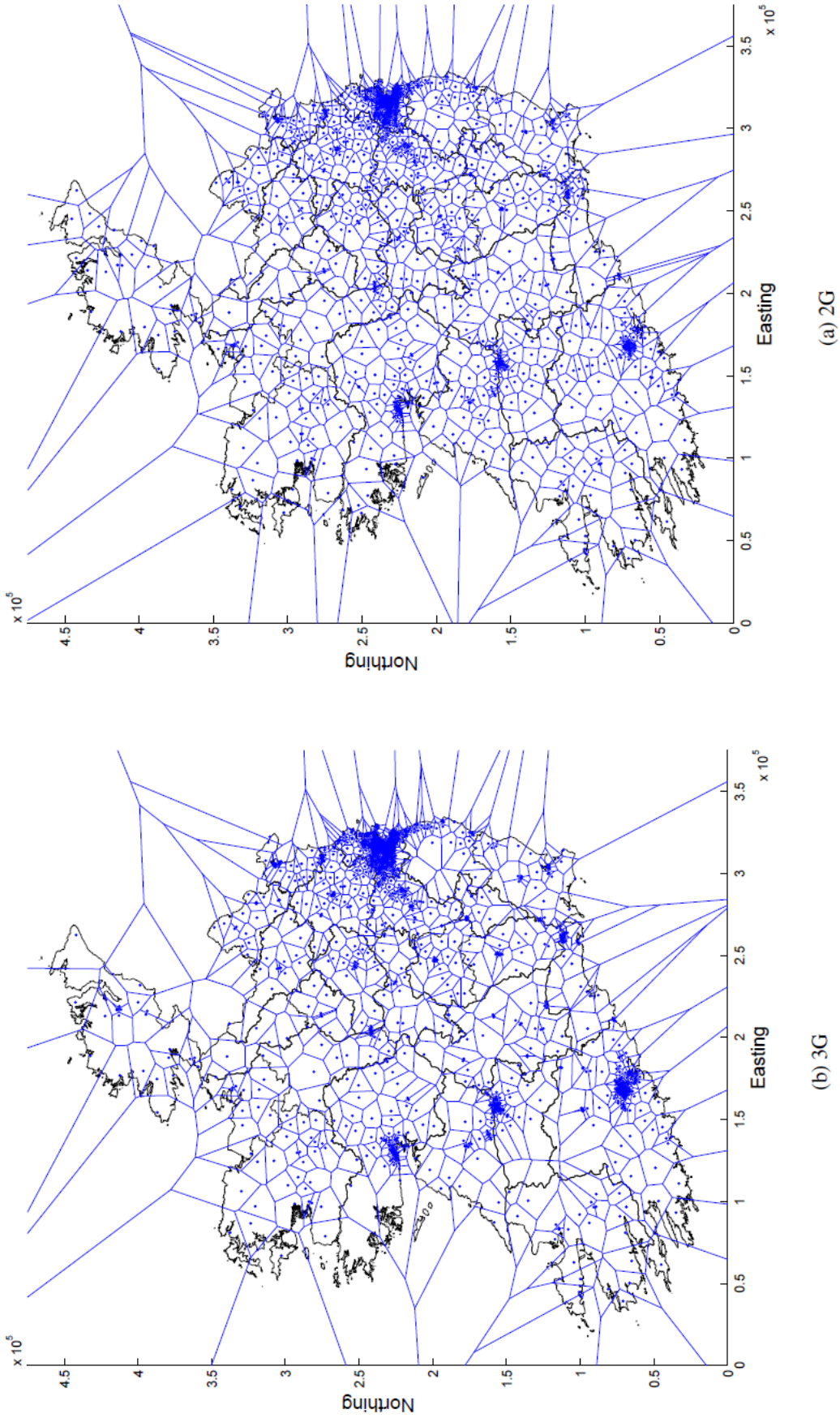


Figure 4.1: Voronoi diagram of 2G (top) and 3G (bottom) cell site coverage regions.

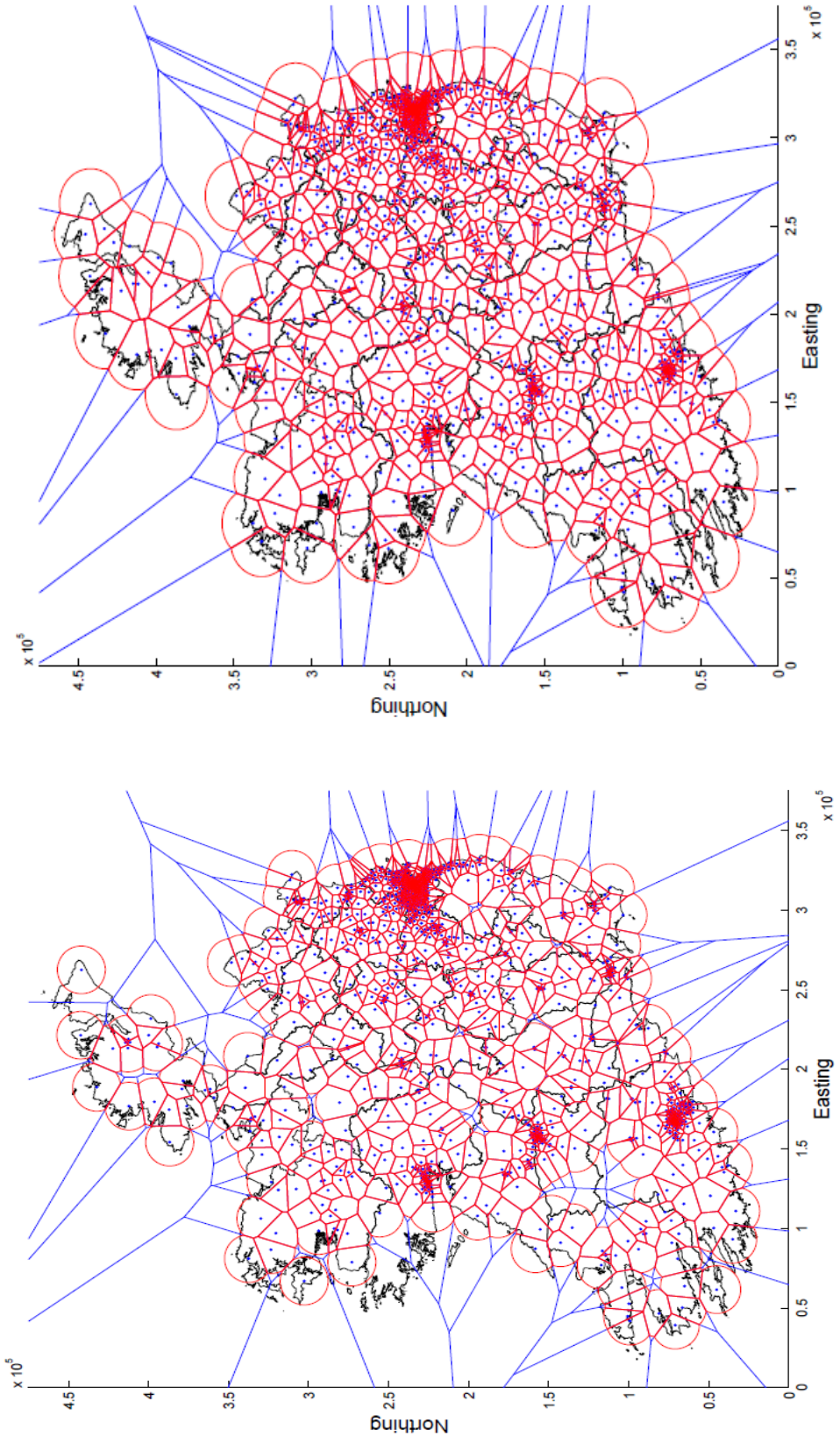


Figure 4.2: Restricted 2G (top) and 3G (bottom) cell site coverage regions.

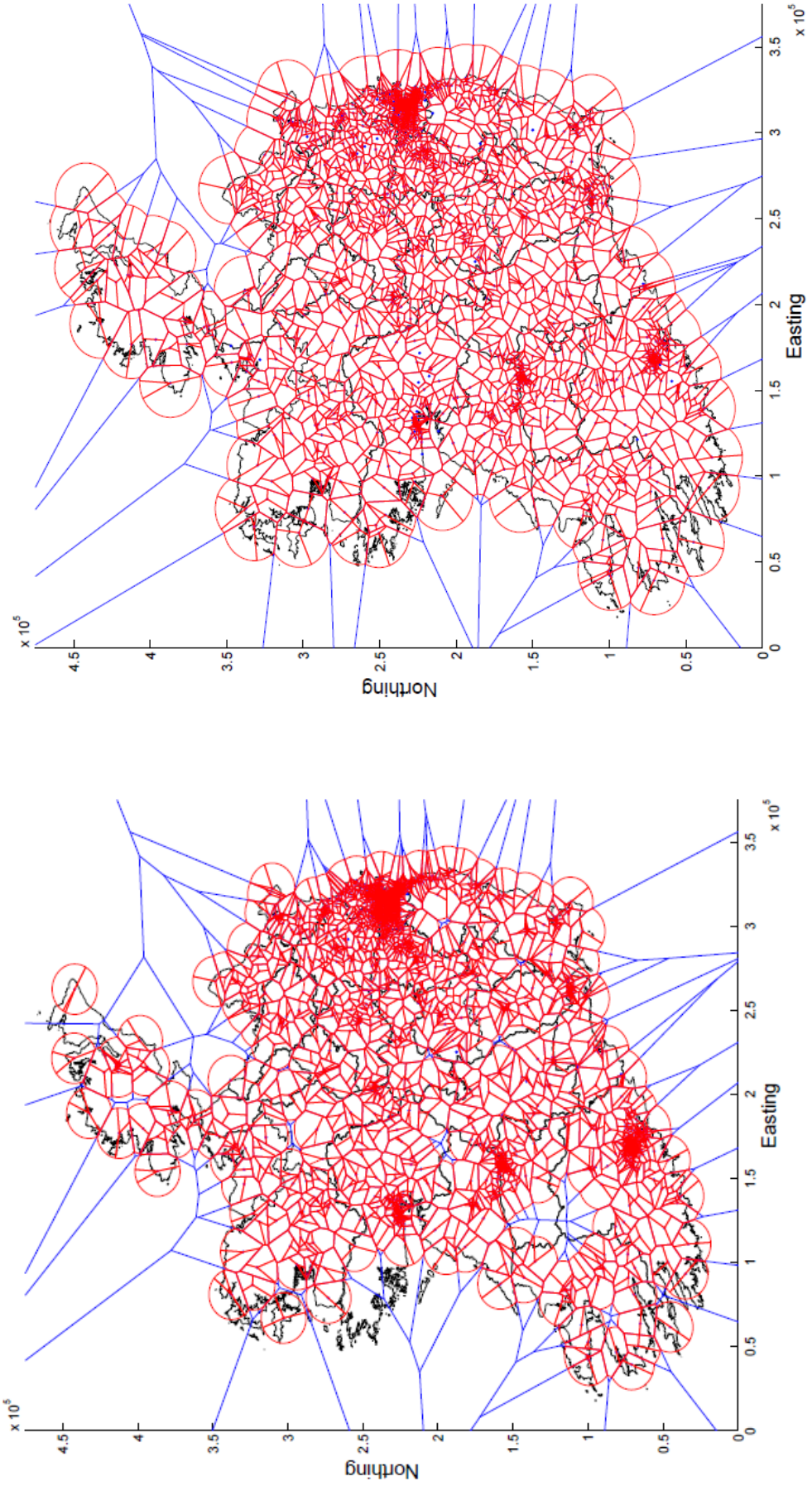


Figure 4.3: Sectored 2G (top) and 3G (bottom) cell coverage regions.

4.2.3 Data Cleaning

As part of the routine operation of a large scale cellular network, operators sometimes relocate hardware around their network. Consequently, through time if the network operator does not keep up to date records on the movement of all hardware within the network the spatial locations of hardware such as cells may become outdated. This can introduce errors in analysis where the spatial locations of cells are important (such as in localised load forecasting schemes as introduced in Chapter 7). For example, the switching technique introduced in Chapter 7 relies on the spatial redundancy between proximate cells. If these cells were not in fact proximate then this would invalidate the entire switching process. Thus, validating the hardware location information received from the network operator as discussed in Chapter 2 is a vital step in any spatial analysis of the network. Some errors can be clearly seen when examining the distance travelled between certain cells in a given time. A subscriber serviced by cell C_x at time t_x and who is subsequently observed in cell C_y at time t_y is assumed to have travelled from the coverage polygon of cell C_x, C_{px}, to cell C_y's coverage polygon C_{py}. The upper bound on the journey time between the two cell's coverage regions is given by $t_y - t_x$. The actual distance travelled by the subscriber will depend on the particular size of the cell coverage polygons involved ranging from d_{xy}^{min} to d_{xy}^{max}. Figure 4.4 illustrates the maximum possible distance travelled d_{xy}^{max}, the average distance d_{xy}, and finally the minimum distance d_{xy}^{min}. As illustrated in Figure 4.4 the maximum distance in any two cell coverage polygons will be the distance between two vertices giving

$$d_{xy}^{max} = max_{ab} \left\| C_{pxa} - C_{pyb} \right\| \tag{4.5}$$

where C_{pij} is the j^{th} vertex of cell i's coverage polygon. The number of vertices in the coverage polygon C_{px} is denoted $a = [1 \rightarrow A_x]$, where A_x is the total number of vertices used to define the cell coverage polygon C_{px}. Similarly, the number of vertices in the

coverage polygon C_{py} is denoted b = [1 → B_y], where B_y is the total number of vertices used to define the cell coverage polygon C_{py}. However, as demonstrated in Figure 4.4 the minimum distance between two coverage polygons can be between a vertex and a side. This, in theory, makes the calculation of d_{xy}^{min} more complicated as every point in every coverage polygon must be compared with all points in every other polygon (unless the polygons are found to be overlapping or adjacent). Also, as the coverage polygons are defined by their vertices locations, it necessitates the interpolation of the points between each vertex at an arbitrary granularity. However, in practice as coverage polygons are only an approximation of actual cell coverage regions which vary due to topography, load etc. this is needlessly complex. A simpler solution is to use a heuristic that the minimum possible distance between two non-adjacent/non-overlapping polygons is

$$d_{xy}^{min} = d_{xy} - (C_{pxr} + C_{pyr}) \tag{4.6}$$

where d_{xy} is the Euclidian distance between centroids of cell coverage polygons C_{px} and C_{py}; C_{pxr} and C_{pyr} denote the maximum distance between the coverage polygon centroids and their respective farthest vertex.

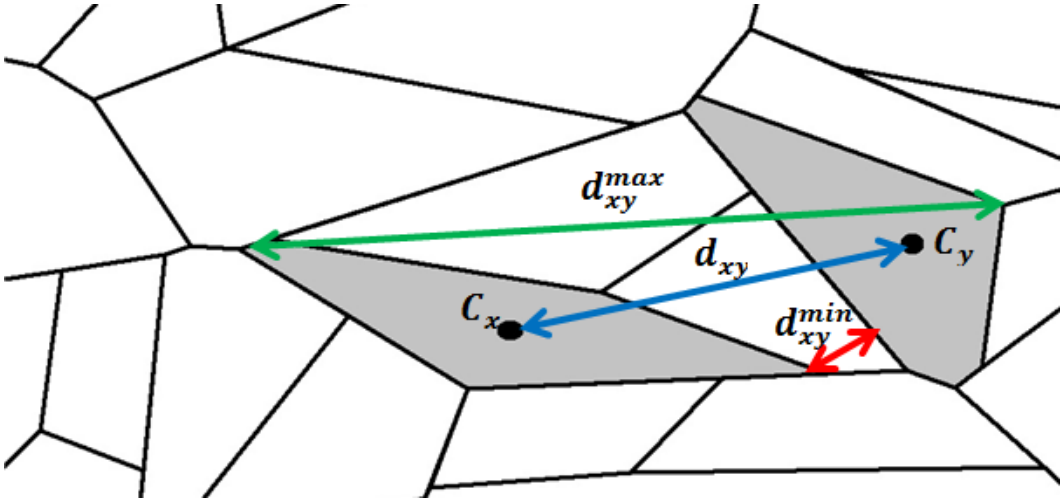


Figure 4.4: The range of possible distances travelled in a transition from cell C_x to C_y in time t_x to t_y. The distance,d_{xy}, is given by the distance between the centroids of the two cell coverage polygons. The maximum distance is given by d_{xy}^{max} with the minimum distance being d_{xy}^{min}

When the transitions between coverage regions are examined, a small proportion are found to occur in impracticably small-time periods given the supposed distance between the coverage regions. Thus algorithm 4.1 is used to identify out of date cell coverage regions. Upon completion of the geographic data cleaning, 5% of cell's geographic locations were found to be out of date and excluded from further examination in this work.

Algorithm 4.1:

The geographic data cleaning algorithm.

1: Let D = (d_{ij}) be a two dimensional distance array (*m*x*m*) where m is the number of cells in the network and d_{ij} is minimum distance between cell polygon *i* and *j* i.e. $d_{ij} = d_{xy}^{min}$.

2: Let T = (t_{ij}) be a two dimensional transition time array (*m*x*m*) where m is the number of cells in the network and t_{ij} is minimum observed transition time between cells *i* and *j*.

3: Let F = (f_{ij}) be a two dimensional flag array (*m*x*m*) where m is the number of cells in the network and f_{ij} is = 1 if the transition between cells *i* and *j* is flagged as infeasible and 0 otherwise.

4: Iterate through each column of D, d_{i1},d_{i2},..d_{im} (i.e. the distance between cell *i* and all other cells). For each element check the corresponding element in T, t_{ij}. If $d_{ij} > 0$ but $t_{ij} = 0$ flag the cell pair in F as $F(f_{ij}) = 1$. If $d_{ij} / t_{ij} > 120$ kph (the motorway speed limit is used as an upper bound on expected transition speed) flag the cell pair in F at $F(f_{ij}) = 1$. Otherwise set $F(f_{ij}) = 0$

5: Calculate the sum for each row in F (giving the number of infeasible pairs the cell is involved in).

6: Iterate through the flag array F. For each transition pair flagged as infeasible, mark the cell with the most infeasible transitions calculated in the previous step as out of date. Decrease the infeasible value for the other member of the pair and continue until all cells are assigned as either up to date or out of date geographically.

4.2.4 Usage Visualisation

Temporal variation in cell load was explored in the previous chapter; however, this examination made no reference to the spatial structure of the network. As demonstrated in 4.2.2 each cell's spatial structure can be visualised to build up a map of the entire network. Thus, a spatio-temporal load map can be constructed for the entire network by combining the spatial structure visualised in 4.2.2 with time series data representing cell load as presented in the previous chapter. A spatial smoothing function is required to enable the visualisation of a spatio-temporal load map for a network comprising many overlapping cell coverage regions of various sizes and shapes. The spatial smoothing function utilises an individual Gaussian function for each cell centred on the cell's coverage region centroid as discussed in 4.2.3. Each Gaussian function's spreading factor is a function of cell radius and spreads each cell's load, C_a, over a spatial lattice, $\delta(x,y)$. The weighted spreading function for a cell is given by:

$$\delta(x,y) = \alpha C_a \exp\left(-\frac{(x - C_x)^2}{2C_r^2} - \frac{(y - C_y)^2}{2C_r^2}\right) \tag{4.7}$$

where C_r is the cell radius, (C_x,C_y) are the coordinates of the cell's centroid, (x,y) are coordinates of points in the spatial lattice, and α denotes the scaling weight which ensures the combined weights in $\delta(x,y)$ sum to C_a. Each lattice point may extend to a temporal horizon t by incorporating the parameter t representing the desired time sample. The resultant lattice $\delta(x,y,t)$ can then be combined with other lattices to view the spatial distribution of activities in a desired area for time sample t. The combined weighted lattice, $\theta(x,y,t)$, is given by:

$$\theta(x,y,t) = \sum_{C=1}^{N_C} \delta_C(x,y,t) \tag{4.8}$$

where $\delta_c(x,y,t)$ is the lattice representing cell C and N_c is the total number of cells in the spatial region of interest.

Figure 4.5, Figure 4.6 and Figure 4.7 illustrate representative sample cell load maps for data load, call load, and SMS load respectively across the network for both peak and minimum load. To create these visualisations the spatial extent of the network was divided into 200x200 meter squares indicating an individual pixel. Each pixel was assigned a load intensity via equations (4.7) and (4.8) with a temporal bin of 300 seconds. The visualisations were completed using the built in MATLAB plotting functions. To smooth out high frequency temporal variations in load, a temporal smoothing function was employed. The function is defined as:

$$\theta(x,y,t) = \frac{1}{5}\sum_{i=t-2}^{t+2} \theta(x,y,i) \tag{4.9}$$

This is a moving average filter which averages the current temporal sample over five temporal samples. Interestingly, the plots show the strong spatial unevenness in the distribution of load across the network. The relationship between population density and load is evident across all service types and for hours of maximum and minimum load. For example, compare the densely populated greater Dublin region with the more sparsely populated and hence lower usage North West of Ireland in Figure 4.5 - Figure 4.7. Figure 4.5, Figure 4.6 and Figure 4.7 also indicate spatial correlation between the loads on the three different service types.

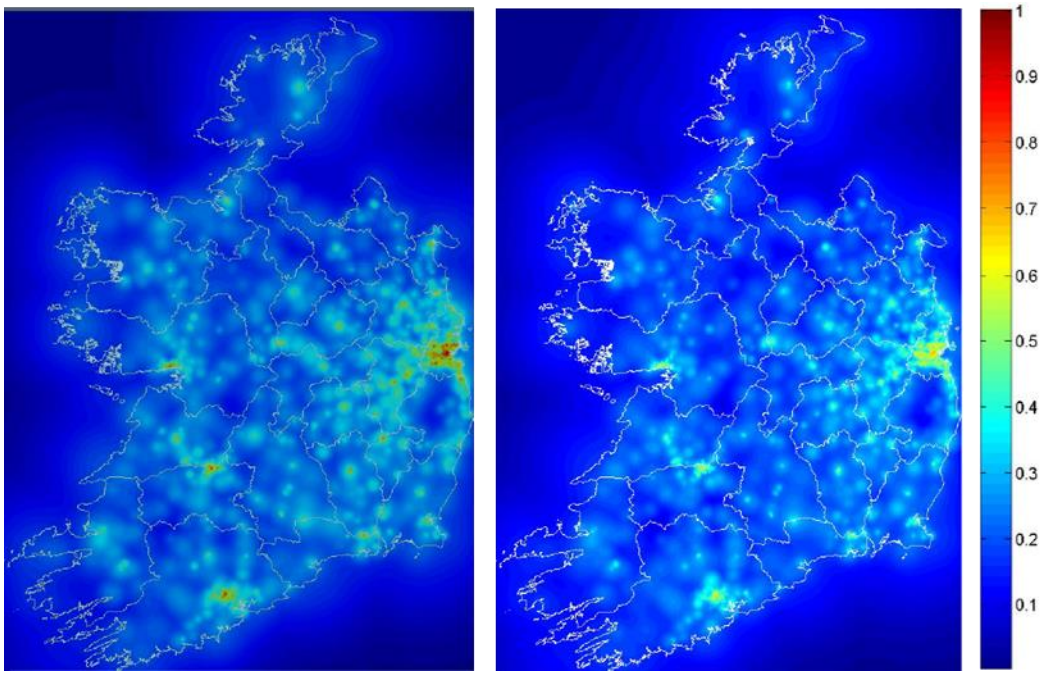


Figure 4.5: Visualisation of data load on the network. On the Left the load at its daily maximum and on the right the load at its daily minimum.

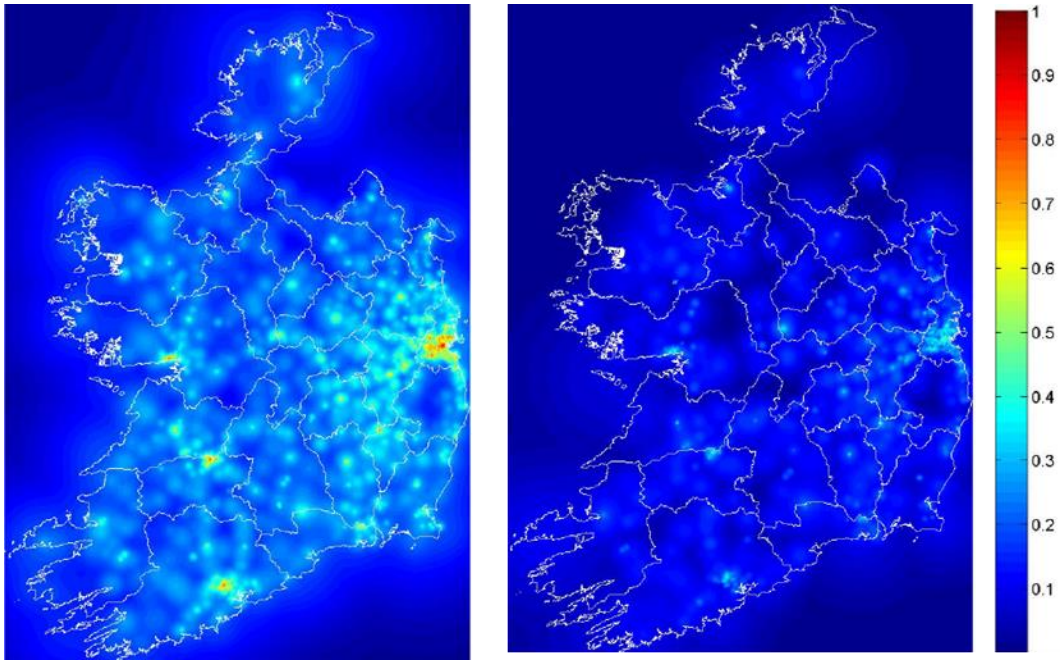


Figure 4.6: Visualisation of voice call load on the network. On the Left the load at its daily maximum and on the right the load at its daily minimum.

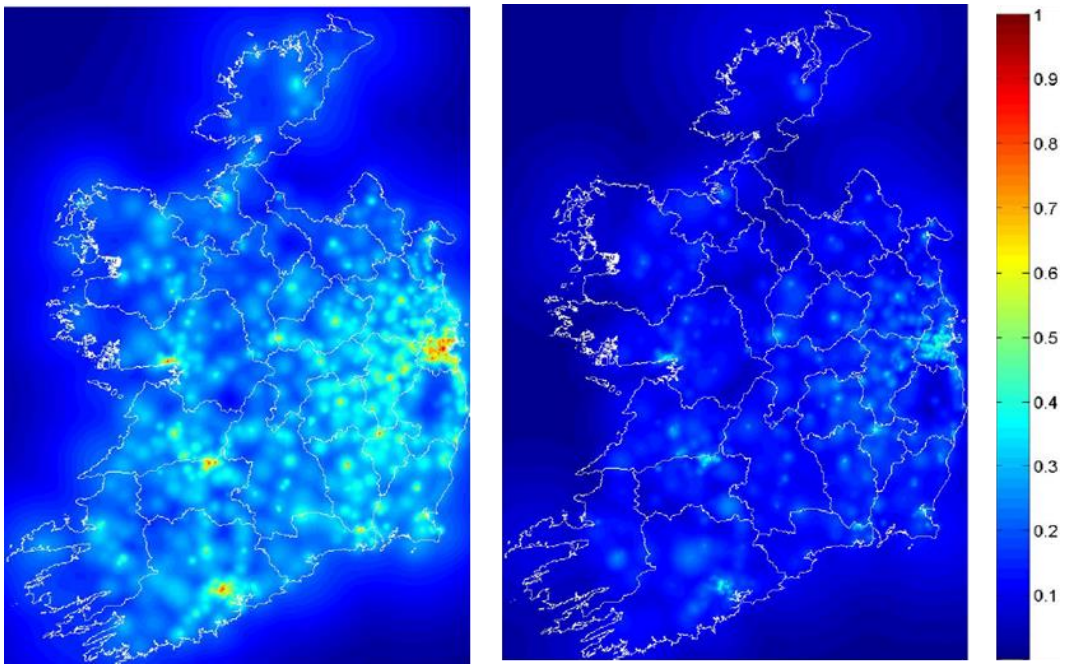


Figure 4.7: Visualisation of SMS load on the network. On the Left the load at its daily maximum and on the right the load at its daily minimum.

4.2.5 Conclusion

This section focused on the creation of spatial representations of the network firstly at the localised base station and sectorised cell coverage level, then the aggregated network level. The spatial coverage region representations introduced in 4.2.2 are the foundational step in beginning to examine the network spatially. Much of the later work and many of the techniques introduced later require the use of these spatial coverage regions. Given their importance 4.2.3 introduced a novel method to identify and remove errors in their positioning. Subsection 4.2.4 provided a method to visualise the spatial distribution of cellular load across the network. The techniques employed in 4.2.4 are generalizable to not only represent load distribution but also other properties of interest such as connection events, subscriber distribution etc.

4.3 Communication Distance

4.3.1 Introduction

This section provides an examination of how users communicate over spatial distance. Subsection 4.3.2 introduces a novel algorithm to calculate the home and work populations of each cell in the network. This algorithm allows for the creation of accurate maps of the network's subscriber base for different classes of cells. 4.3.3 - 4.3.5 examines and models how subscribers communicate with one another spatially. The classic gravity model of spatial communication distance is applied in a novel manner to cellular networks utilising the cellular coverage regions identified in 4.2.2 and the novel cellular population estimation techniques presented in 4.3.2.

4.3.2 Cell Populations

One of the defining features of a cellular network is the population density of the spatial region that the network services. Two popular methods of estimating population density when examining a network are the use of census records or the address information provided by the customer upon signing up to the network [59]. Both methods have their shortcomings. In the case of census information, it cannot be assumed that the network of interest has equal penetration across all areas studied. A large drawback of using address information provided by the subscriber is its lack of accuracy. Subscribers often provide unreliable information to service providers. This is self-evident in the customer data provided for prepay customers (see Chapter 2). Many of these prepay customers have blank address information or simple placeholders such as "zzz" etc. Customers with a bill phone are obliged to submit correct home address details but there is no such guarantee with pre-pay users. This is particularly challenging due to the growth in popularity of pre-pay plans [60]. Bill-pay customers currently

account for just approximately 10% of the users on the network under investigation. Apart from the lack of accuracy of home locations inherent in both methods, neither takes account of the daily movement of people throughout the network. For much of the day a large proportion of the people living in a certain area will not be there - further reducing the usefulness of address or census information. A more useful dataset would include for example, the home and work/study locations of the subscriber base without recourse to self-reported address or census information. Such a data set was created with the use of four months of CDRs as outlined in Algorithm 4.2 & Algorithm 4.3.

Algorithm 4.2:

The home location estimation algorithm used.

1: Extract all events over the study time period and group them by day of the week. Exclude any events that occur on Friday, Saturday or Sunday.

2: For a day in the study period extract all events which occur at "home times" i.e. 8pm -6am and group them by user id.

3: Load list of cell towers in the area of interest.

4: For each subscriber count how many events occur within each cell's coverage polygon region (see §4.2.2)

5: Iterate through all subscribers and determine the most frequent cell for each subscriber for the day of interest.

6: Assign the subscriber to the cell found in step 5 for that particular day.

7: Repeat steps 2 to 6 for each day of interest and find the cell the subscriber is assigned to for the largest amount of days. Set this cell as the subscriber's home location. If a subscriber is associated with two or more cells for the same amount of days, pick one at random.

8: Sum all the subscribers assigned to each cell tower and set the result as each respective cell's home population.

Algorithm 4.3:
The work location estimation algorithm used.

1: Extract all events over the study time period and group them by day of the week. Exclude any events that occur on Saturday or Sunday.

2: For a day in the study period extract all events which occur at "work times" i.e. 9am - 4pm and group them by user id.

3: Load list of cell towers in the area of interest.

4: For each subscriber count how many events occur within each cell's coverage polygon region (see §4.2.2)

5: Iterate through all subscribers and determine the most frequent cell for each subscriber for the day of interest.

6: Assign the subscriber to the cell found in step 5 for that particular day.

7: Repeat steps 2 to 6 for each day of interest and find the cell the subscriber is assigned to for the largest amount of days. Set this cell as the subscriber's work location. If a subscriber is associated with two or more cells for the same amount of days, pick one at random.

8: Sum all the subscribers assigned to each cell tower and set the result as each respective cell's work population.

Both algorithm 4.2 and 4.3 were implemented in a mix of C and SQL for the entire network. Figure 4.8 shows the CDF of the home and work populations calculated for each cell on the network as calculated using Algorithms 4.2 and 4.3 respectively. A wide range of both home and work populations are evident in each cell ranging from a minimum of 1 to a maximum of 1000. The median home population is 38 while the median work population is 40. However, the mean home and work populations are

more than double their respective medians at 84 and 91 respectively. This indicates that there are many cells with low home and work populations while a disproportionate amount of subscribers live/work in a relatively small cohort of heavily loaded cells. This is consistent with both the findings presented in the previous chapter and subsection 4.2.4. The CDF of home and work populations look very similar but it bears remembering that they are not necessarily for the same cells (see the comments in Figure 4.9). For example, a cell covering an industrial park may have a large working population with much a smaller residential population. The home population to work population ratio for each cell is displayed in Figure 4.9. Generally, the two are similar with the home population ranging from half to twice the work population for 85% of cells. However, in some cases the home population can be one tenth the work population at one extreme or ten times greater than the work population at the other extreme.

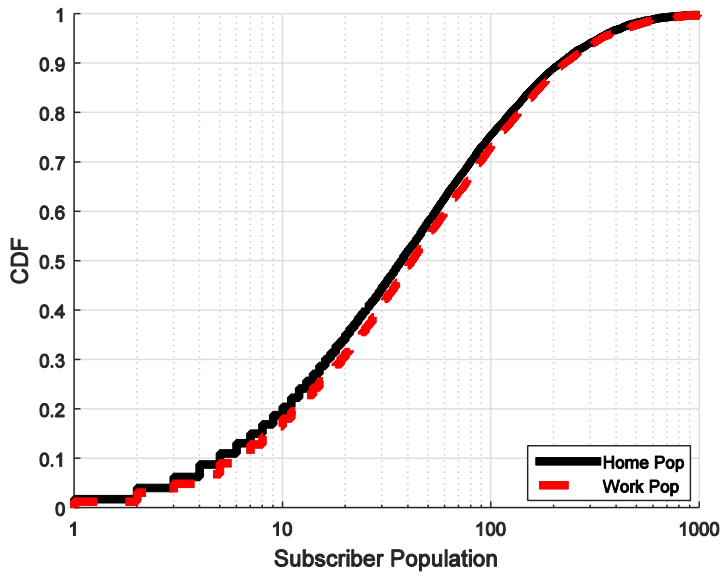


Figure 4.8: CDF of the home population and work population for each cell on the networks

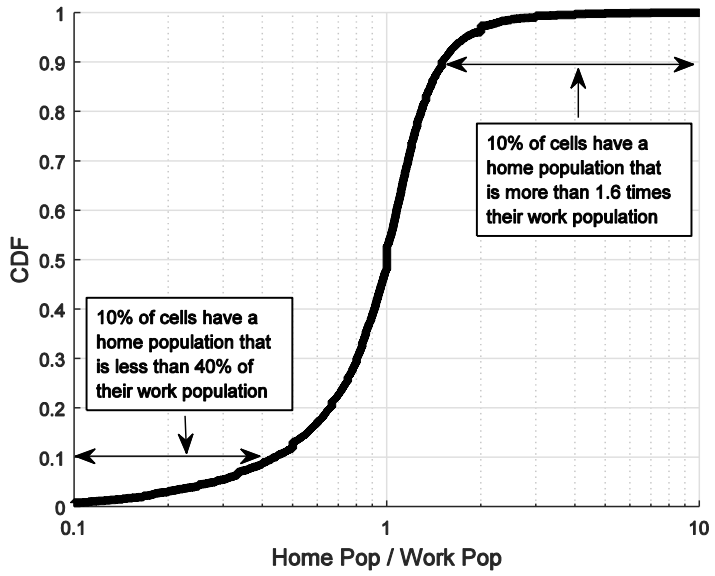


Figure 4.9: CDF of the home population of each cell divided by each cell's respective work population.

4.3.3 The Gravity Model

Interestingly, having accurate home and work locations for mobile subscribers allows for, in some respects, the treatment of cellular networks like old fixed line connections. This permits the revalidation of fundamental laws of fixed line communications such as the gravity law in a cellular network context. It has been previously demonstrated that various systems can be represented as a network of nodes, connected by weighted or unweighted links [61]. It is a common technique to represent social networks as a network where each node represents a person and links between the nodes indicated social interactions. [62] utilises a dataset similar to CDRs to highlight the importance of weak ties to the propagation of information through a communication network. Several other authors have made use of large recently available phone and email datasets to study human connections and behaviours [63-66]. Geographical information allows for a more detailed and interesting exploration of group and individual interactions. For

example, [67] uses a mobile phone dataset to show that the probability of a call between two people decreases by the square of their distance.

Interurban connections such as passenger flows and phone messages and their dependence on separation distance have been studied for a considerable amount of time [68, 69]. In various economic and social networks, interactions between actors such as regions and countries has led to models similar to Newton's Gravity law, where the size of the actor plays the role of mass [70]. These Gravity models take the following form:

$$W_{ij} = K\frac{M_i M_j}{d_{ij}^n} \tag{4.10}$$

where W_{ij} is the weight of the link between node i and node j, d_{ij} is the distance between nodes M_i and M_j, n is the exponent of the distance, and K is a constant.

Studies have also been carried out on road and airline networks between cities [9, 10]. In the case of road networks it appears that the gravity model holds for the strength of interactions. [71] analysis a CDR dataset but unlike [62] it associates users with locations and aggregate links between users to links between locations. [71] explores how the strength of the links between locations varies relative to separation distance and population. It finds that the strength of the link between locations is proportional to the populations at the locations and inversely proportional to the distance between the locations. Hence, [71] concludes that the inter-city communication intensity is characterised by a gravity model.

4.3.4 Estimating population size and communication links

One limitation of [71] is that it relies on the billing address Zip code provided by the subscribers to the network operator. All users in a specific Zip code are aggregated and Zip codes are aggregated to form cities. However, this introduces a potential source of

error as users often provide unreliable information to service providers as discussed in 4.3.2.

As can be seen from equation (4.10) it is important that an accurate estimate of the population of the two cities/areas be made. The population *M* of city *i* is calculated as follows:

$$M(i) = \sum_{c=1}^{n_c} M(c) \tag{4.11}$$

where *c* is a cell with all or part of its coverage region contained within city *i*'s boundary, and n_c is the number of cells with all or part of their coverage regions contained within city *i*'s boundary. The city boundaries are defined as the boundaries employed by the Central Statistics Office for the 2011 Irish census [72].

Equation (4.11) provides an accurate estimation of the subscriber population of cities i.e. M_i and M_j in equation (4.10). However, verification of equation (4.10) also requires values for the link weight *W* between cities. To generate the interurban communications network link weight the total communications originating and terminating in a city are aggregated together. The weight of the link (*W*) between two cities *α* and *β* can thus be defined as:

$$W_{\alpha\beta} = \sum_{i\in\alpha, j\in\beta} w_{ij} \tag{4.12}$$

where w_{ij} is a link between individual users in the respective cities. The weight of the links between twenty-five cities/towns is thus calculated for each of the seven days in a week including workdays and weekend days. The weight of the links between the cities/towns is also calculated for two times of interest during each day – work times (9am-4pm Monday-Thursday and 9am-3pm Friday) and home times (8pm-7am Monday – Thursday). Additionally, the weight of the links is calculated for daytime weekend

(7am Saturday & Sunday to 8pm Saturday and Sunday) and night time weekend (8pm Saturday and Sunday to 7am Sunday and Monday).

All the above calculations are performed for three different metrics of link weight – number of calls between cities, total call time in seconds between cities and number of SMS between cities.

4.3.5 Testing the gravity model

[73] performed a high level investigation of the gravity model on Ireland's communication network. [73] found that the gravity model approximates the actual data under their specific aggregations. [73] aggregated results over the period of one week and were dealing with much larger regional aggregations. Equation (4.10) can be rearranged as follows:

$$d_{ij}^{n} = \frac{K\ M_i M_j}{W_{ij}} \tag{4.13}$$

Using equation (4.13) the results obtained can be tested for degrees of compliance with the gravity model using linear regression. Two different measures of distance were used when testing the gravity model. The first was the spatial travel distance between the centres of two cities and the second was the travel time by road between two cities. Figure 4.10 compares one week of data plotted for both travel distance by spatial distance and by travel time. Figure 4.10 illustrates that the gravity model performs better when distances are measured in spatial distance. This result is repeated for all the cities examined in the study with the agreement between the gravity model and the results being on average 15% less when travel time is used.

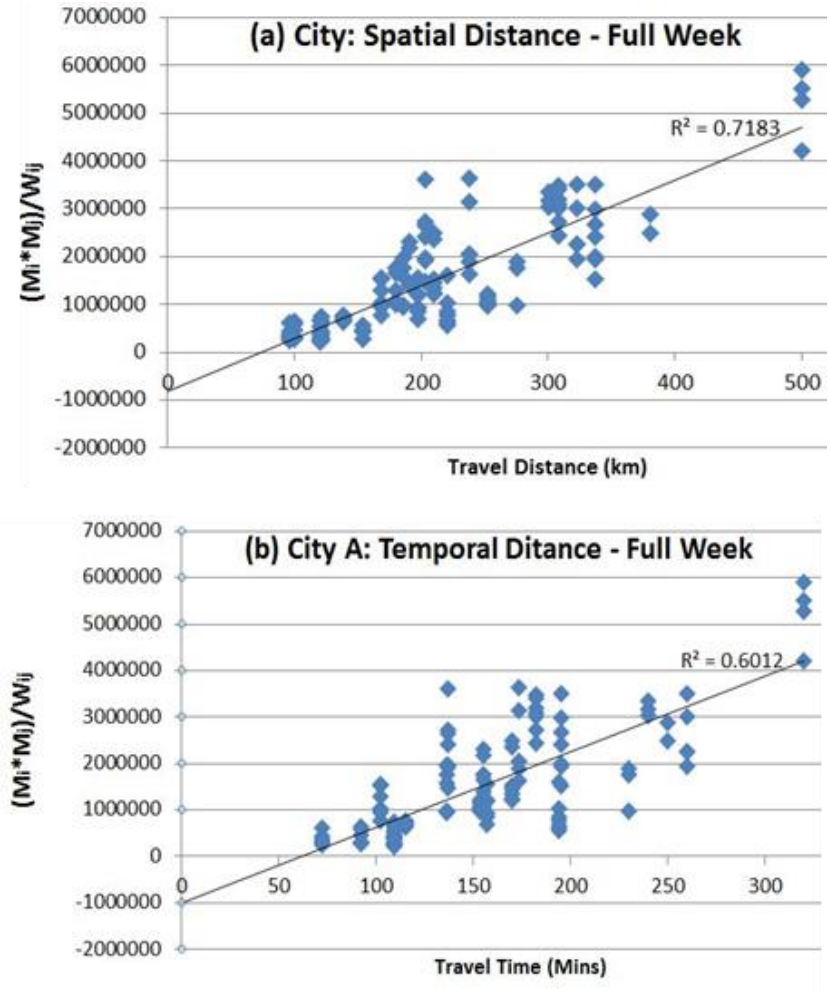


Figure 4.10: One full week of data plotted with two different measures of distance.

The model was tested for three different types of communications links – total number of call connections made, total connection time of all calls and total number of SMS sent and received. The greatest agreement with the gravity model was found when total number of SMS was used. This result is repeated for all the cities examined in the study with the agreement between the gravity model and the results being on average 17% less when total number of connections or total call time is used. It is not immediately clear why this is; it could represent an underlying difference in communication behaviour between calls and SMS. It could, however, also be a result of users sending on average over 4 times more text messages than making calls. As shown in Figure 4.11, smaller town to smaller city/town i.e. communications with few links disproportionality

affects the results. This smaller city/town to smaller city/town effect is reduced when dealing with SMS as the number of links is greater.

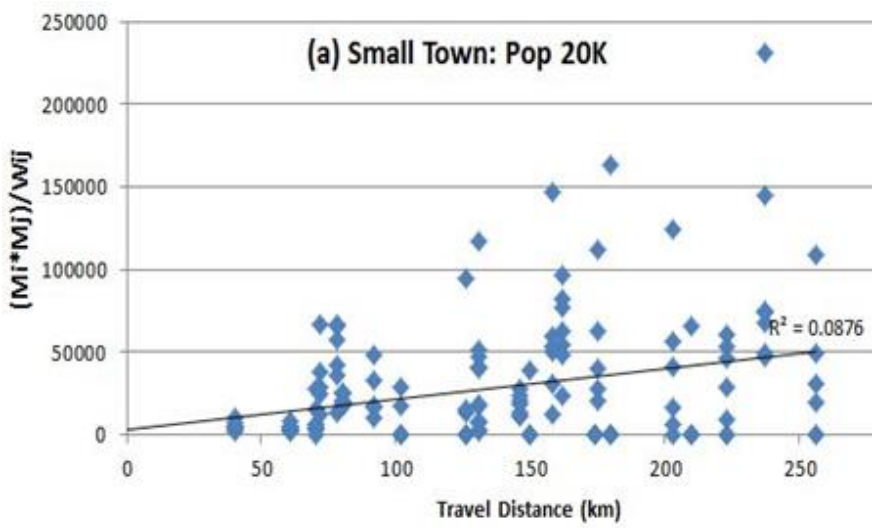


Figure 4.11: Small town to small town communication over one week

The agreement between the results and the model vary both with the day of the week and the time of the day. Figure 4.12 shows how the results change between the working week and the weekend. On average, the gravity model performs worse for cities during the weekend (on average approximately 10% less agreement between observation and the model) when compared with the working week. One possible explanation is the large amount of Irish people who work/study in the cities during the week and move back to the small towns/rural areas where they grew up on the weekends. There is also a small change in the agreement with the model based on the time of day. During the daytime/evening there is a slightly larger agreement between the gravity model and the results than at night. The effect is smaller than the weekday/weekend shift and is probably a result of non-residents being present in the city during daytime hours on weekdays and returning home outside the city at night.

Figure 4.10 and Figure 4.12 seem to indicate a value of K=1 in equation (4.13). There are several possible reasons for this. For example, the Republic of Ireland's urban areas are separated by relatively small distances. This allows people to work/study in one part of the country while maintaining strong links with their relatively close places of birth. This large degree of mobility between urban areas would not be possible in a larger country.

The relevance of the model greatly depends on having at least one large population centre on either end of the communication link. There are two main interurban communication scenarios considered. The first is when a large population is present on either side of the link (large population communicating with small, small to large and large to large). This always provides the best fit with the gravity model (see Figure 4.10 (a)) even when taking into account variations due to the time of the week (Figure 4.12) or time of the day.

The second population scenario is where there is no large population centre on either side of the link (smaller town to smaller town). This primarily affects the smaller towns with populations of less than 50,000 inhabitants (Figure 4.11). This scenario is prevalent in Ireland due to many of Ireland's urban areas being relatively small by international standards. The Republic of Ireland only has five cities with a population greater than 50,000 inhabitants. Thus, for the remainder of the Republic's urban areas the gravity model is a poor choice for modelling interurban communication.

This is a key difference between this study and that of [73] which shows an approximate national agreement with the gravity model. The conclusion of [73] states that "this work has focused on county-level interaction". Out of the twenty six counties of the Republic Of Ireland covered in their study only two have a population of less than 50,000 with most having significantly more [74]. Thus, the gravity model is only relevant

when dealing with sufficiently large populations, either concentrated in a large urban area or more widely spread out over a larger region.

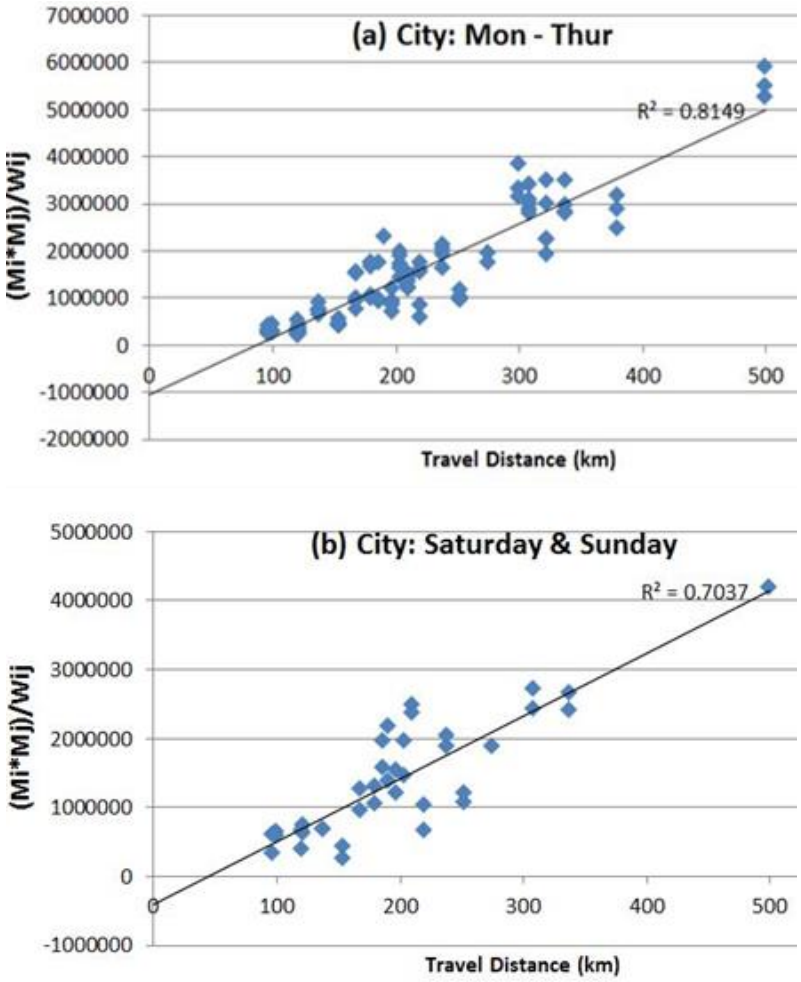


Figure 4.12: Change in communication patterns (a) Mon-Thurs (working days) (b) Saturday and Sunday (weekend)

4.3.6 Conclusion

4.3.2 saw the introduction of novel techniques to ascertain the network subscriber specific home and work populations for each cell in the network. These techniques allow for the creation of accurate maps of a networks subscriber base for different classes of cells. 4.3.2 focused on home and work cells but the techniques introduced could easily be generalised to build up maps of different cells e.g. socialising cells etc.

4.3.3 - 4.3.5 examined how subscribers communicate with one another spatially. The gravity model was tested as one possible model for communication distance in cellular networks. The performance of the model was found to vary largely based on the type of link chosen, the time of the week, and to a lesser extent the time of day. The value of *K* in equation (4.13) was also found to be 1 indicating a linear relationship. The gravity model may be more suited to static landlines than mobile phones. The simplicity of the model does not consider the highly mobile nature of the Irish population. This is probably exacerbated by the relatively small size of the Republic. This small size facilitates people working/studying in one area during the week while maintaining strong links to their place of origin. The gravity model was found to be only helpful when dealing with large population centres of more than 50,000 inhabitants. As the Republic of Ireland only has five cities with a population of 50,000 inhabitants or more the gravity model is a poor choice for modelling interurban communication between the country's smaller urban centres. In future interurban work smaller population centres should be amalgamated into larger groups or a more sophisticated model should be employed.

4.4 Spatial Relationships

4.4.1 Introduction

In the previous chapter, 3.3 examined how network load varied between individual base stations and sectorised cells. A large disparity in load was identified with some base stations and cells servicing several orders of magnitude more load than others. Concomitant with those findings, 4.3.2 identified a large variation in the amount of people living and working in cells and the relevant ratios of both. Thus, it is already known that there is a great diversity of cells present on the network. The question this section explores is how do these differences manifest spatially? 4.4.2 explores the loads

serviced by cells spatially correlated while 4.4.3 - 4.4.7 explores the causal structure of the network's load.

4.4.2 Spatial Correlation

This subsection examines how spatially correlated load is across the entire network. The load of each cell is now used to investigate the extent of the spatial correlation on the network by cross-correlating pairs of base stations' time series with one another. Cross-correlation is a widely used statistical method of measuring the similarity (the degree of correlation) between two time series [75]. Figure 4.13 shows the cross-correlation calculated at zero lag for all cells on the network and also for cells based on certain distance ranges over two weeks of data at a granularity of one hour. Similar results were also obtained for the 15-minute interval but are omitted due to their similarity. The cross-correlation between cells was found to be quite high with the one-hour interval displaying slightly higher values than the 15-minute interval. The median cross-correlation was approximately 0.65 for the one-hour interval and 0.5 for the 15-minute interval. 80% of cells had a cross-correlation greater than or equal to 0.5 for the one-hour interval. Cross-correlation was also found to be dependent on the distance between the cells as shown by the groups in Figure 4.13. For example, the median cross-correlation between cells within 2km of each other was 0.8 falling to 0.7 for all cells within 20km.

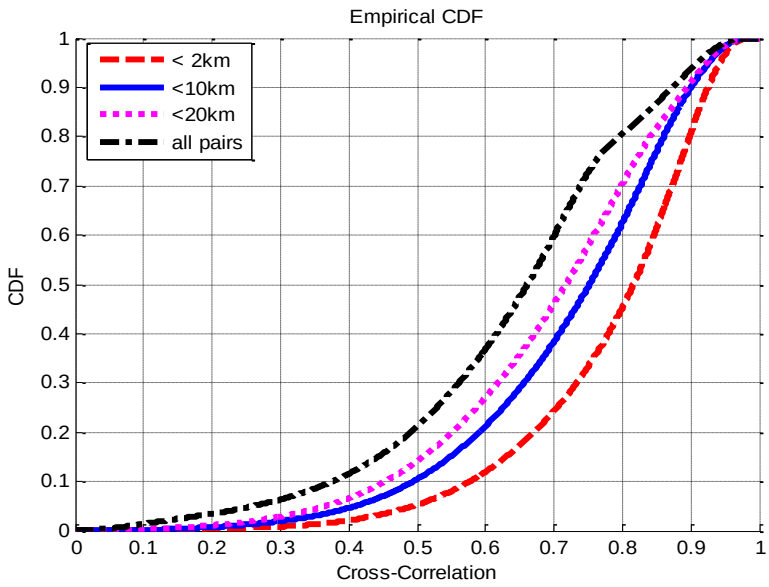


Figure 4.13: CDF of the cross-correlation between all pairs of cells and also within certain distance bands based on hourly load. The distance is defined as $\boldsymbol{d_{xy}^{max}}$as in 4.2.3

To further examine the degree of spatial correlation identified in Figure 4.13, a different metric known as Moran's I statistic is employed [76]. Moran's I statistic is a regularly employed measure of spatial autocorrelation. It quantifies the correlation between different measurements or observations based on their spatial location. Geographic distance is used to indicate proximity and is employed as a weight in the formula. Moran's I statistic is defined as:

$$I = round\left(\frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right) \quad (4.14)$$

where x is the random variable being studied, $\bar{x}$ represents the sample mean, x_i's are the observations, w_{ij} is the weight associated with each pair (x_i,x_j) and N is the number of observations. In this situation, the random variable x being studied is the hourly load on a cell. Similarly, to other Moran's I studies binary weights $w_{ij} = 1$ are employed when the cells are in close proximity (d_{xy} = 2km), otherwise $w_{ij} = 0$. The value of Moran's I is then plotted in Figure 4.14 for each hour of the week. Figure 4.14 shows that Moran's I

statistic varies from a low of approximately 0.1 to a high of approximately 0.4 indicating a varying degree of spatial correlation within the network. Interestingly, the periodic pattern displayed in Figure 4.14 is reminiscent of the diurnal archetype for cellular load identified in 3.2.3. This suggests that the degree of spatial correlation is greatest when the network's load is itself at its greatest. Thus, indicating a general tendency for the load of proximate cells to be more correlated when their loads are higher. This intuitively makes sense, as discussed in Chapter 5 - when the load on a cell or group of cells is very low, for example in the early morning hours, one subscriber connecting to a cell using a data intensive application may greatly increase the load on one cell in percentage terms when compared to its barely used neighbours. During hours of peak load however, the percentage increase will be diminished and also given the finite nature of cellular spectrum the new heavy subscriber's bandwidth will be much more limited reducing his/her distortive capacity.

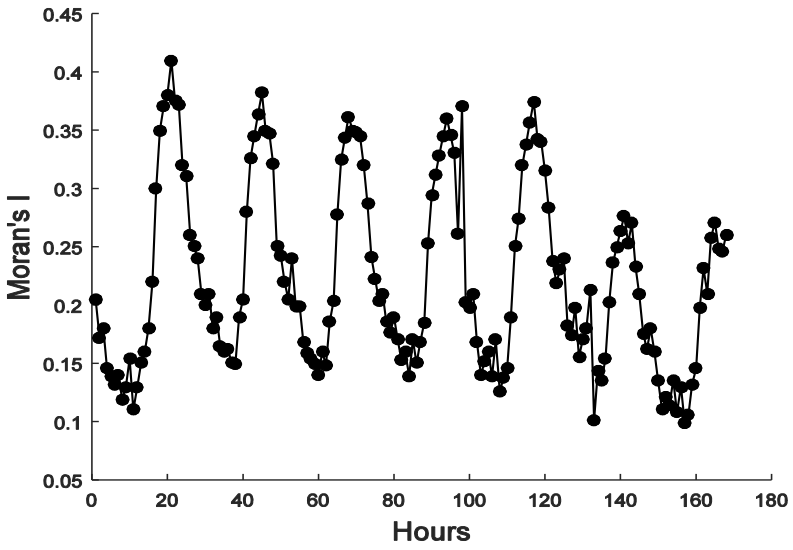


Figure 4.14 Moran's I for each hour of the week for all cells on the network. The plot has been smoothed to remove noise by using sliding window averaging with the window size = 4 hours.

4.4.3 Causal Structure

The previous subsection examined correlations in the spatial extent of network load; this subsection goes beyond spatial correlation and examines the functional influence present in the network. A key metric to understand the underlying functional connectivity present in the network is the causal influence between cells. The causal relationships present in the network have many uses, including load prediction [77], travel mode discovery [78], and identifying influential nodes to reduce load sampling overhead [13]. This section uses one popular measure of causality known as Granger Causality [79] which is a statistical framework for measuring causality between time series.

4.4.4 Granger Causality

Granger causality establishes if one time series improves the forecasting of another time series. One stochastic variable, X_2, Granger causes another stochastic variable X_1 if information in the past of X_2 helps predict the future of X_1 with a better accuracy than is possible with only the information in the past of X_1 alone [79]. Thus, Granger causality is present in the direction X_2 to X_1, provided that the inclusion of X_2 in the model improves the prediction of X_1 by a statistically significant amount. However, this relationship is not necessarily symmetrical and thus 'X_2 Granger-causes X_1' does not imply that 'X_1 Granger-causes X_2' [77]. For example, suppose there are two time series $X_1(t)$ and $X_2(t)$, both having a length of T. As in [80] the two time series can be described using a bivariate autoregressive model:

$$X_1(t) = \sum_{i=1}^{p} A_{11,i} X_1(t-1) + \sum_{i=1}^{p} A_{12,i} X_2(t-1) + \varepsilon_1(t). \quad (4.15)$$

$$X_2(t) = \sum_{i=1}^{p} A_{21,i} X_1(t-1) + \sum_{i=1}^{p} A_{22,i} X_2(t-1) + \varepsilon_2(t). \quad (4.16)$$

where $p < T$ is the model order i.e. the maximum number of lagged observations of X_2 used to predict the current value of X_1 or vice versa at time (t). The matrix A contains the model coefficients while ε_1 & ε_2 are the residuals of the autoregressive model. X_2 Granger causes X_1 if all the coefficients of A_{12} are non-zero i.e. if the residuals are reduced by the inclusion of the second time series in the model. In practice, a threshold is set to determine if the relationship is statistically significant. One such method is the F-test; to be considered statistically significant the F-value should be greater than a desired significance threshold ranging from 0 to 1 [80]. The closer the significance threshold is to zero the greater the significance of the result. The Akaike Information Criterion (AIC) was used to estimate the model order [81].

Using the methods of [80] the model order was found using the AIC as illustrated in Figure 4.15. The time series X_1 and X_2 in equations (4.15) and (4.16) are the cell loads on pairs of cells with neighbouring or overlapping coverage grids, as defined in 4.2.2, aggregated over 10 minute intervals. The model order is generally quite low with about 80% of pairings having an order of 8 or less. This suggests that in most cases only a small number of previous samples from causally connected neighbours are required. For the F-test of significance the significance threshold level was set to the commonly used 0.05. The causality is tested for every pair of neighbouring cells in both directions. On this network 38% of cell pairs were found to have a statistically significant causal relationship in at least one direction at a granularity of 10 minutes.

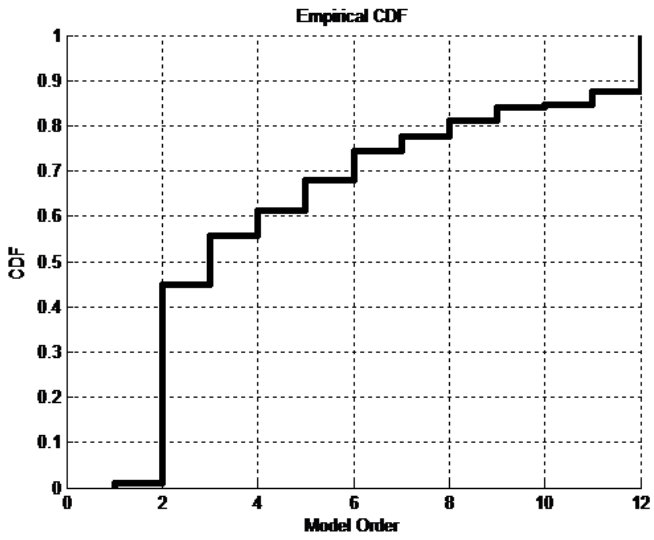


Figure 4.15: CDF of the model order for each pair of neighbouring base stations using the Akaike Information Criterion with a granularity of one hour.

To examine the network as a whole a causality graph is created using the pair-wise causal relationships [75]. The resulting graph of Granger causality interactions is a directed graph (a graph that is set of vertices connected by edges, where edges have a direction associated with them) $G = (V, E)$ where V is the set of vertices, E is the set of edges. Thus, each cell becomes a node on the graph and there is an edge from node a to b (i.e. $(a,b) \in E$) if there is a significant Granger causality interaction between them and they are neighbours in terms of coverage grid. This causal graph allows for the exploration and quantification of some causal properties useful in identifying influential nodes [80]. These properties are outlined in the following subsection.

4.4.5 Causal Density

Causal density is a global measure of the causal interactivity in a dynamic system; causal density shows the mean causality over the entire network. A high value of causal density indicates that the constituent parts of the network are coordinated in their activity [80]. It is the average G-causality over all the pairs of cells examined. Causal

density can take on a value between 0 and 1 and gives the average amount of significant Granger causality interactions over the entire network. Granger causality is defined using the causality graph:

$$Causal\ Density = \frac{\sum_{a \in V} \sum_{b \in V-(a)} I[(b,a) \in E]}{\sum_{a \in V} |N_a|} \tag{4.17}$$

where N_a is the set of neighbours of the cell corresponding to node a and I is the indicator function. On this network the causal density was found to be 0.38 indicating the presence of statistically significant Granger causality in the network.

4.4.6 Causal Flow

The causal graph representation enables the examination of which cells are the influencers and which are the influenced i.e. which cells have a causal influence on their neighbours and which exhibit the results of this influence. Using the causal graph representation, the influence emanating from node a is its out-degree (the number of edges going from node a). The influence node a experiences from its neighbours is given by node a's in-degree (the number of edges going into node a). Figure 4.16 illustrates the out and in degree of every node on the network. Note that some nodes have a very strong influence on their surroundings, for example, the top 5% of nodes have an out-degree of 15 or greater.

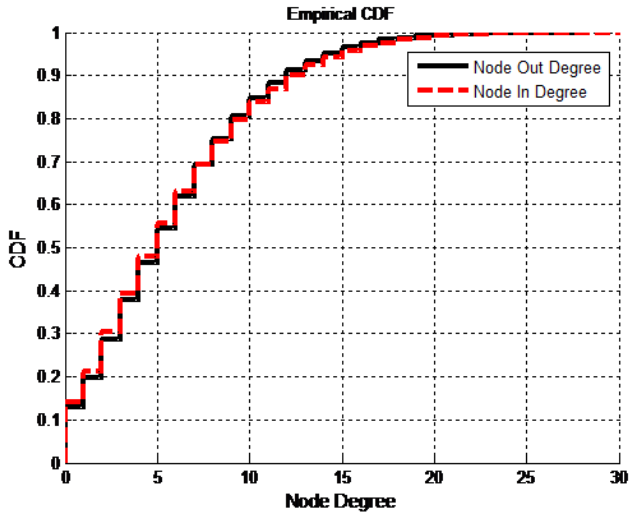


Figure 4.16: CDF of the out and in degree of every node on the network.

To get a more holistic view of the influence of a node while taking into account the influence it experiences, a metric known as the causal flow is employed. The causal flow of a node (base station/cell) is the difference between the causal interaction it exerts on its neighbours and the causal interaction its neighbours, in turn, exert on it. Thus, on the causality graph, the causal flow is the difference between the node's out degree and its in-degree. Nodes with positive causal flows are causal sources while nodes with negative causal flows are causal sinks. The more positive or negative the flow is, the stronger the source or sink is respectively. Figure 4.17 shows the CDF of the causal flow for each cell on the network. The information presented in Figure 4.17 can be used to identify causal sources and sinks in the network. For example, 10% of cells are causal sources with causal flows greater than or equal to five. Conversely, 10% of cells are causal sinks with flows less than or equal to negative five. The strong sources and sinks identified in Figure 4.17 will be further examined in the following subsection.

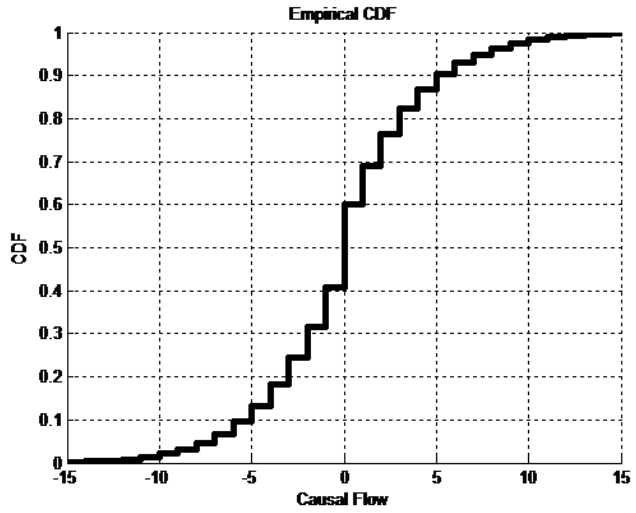


Figure 4.17: CDF of the causal flow of each cell on the network.

Another interesting causal metric to explore is the causal path lengths present in the network. These causal paths represent how the causal influence propagates or flows through the nodes in the network. This indicates the spatial paths throughout the network in which information can be gleaned from previous network states. Causal paths are defined as continuously traversable paths from vertex to vertex via connecting edges in the network graph *G* as defined in 4.4.4. Figure 4.18 displays the CDF of the causal path lengths present in the network and indicates the existence of a wide range of causal path lengths present in the network. The median causal path length in the network was found to be 15 with a 90^{th} percentile path length of approximately 50. Preliminary investigations of these long causal path lengths indicate that when plotted spatially many of them follow major transport infrastructure such as busy motorways etc. In future work it would be interesting to more thoroughly investigate this and examine if there is a relationship between any other geographical features and causal paths present in the network.

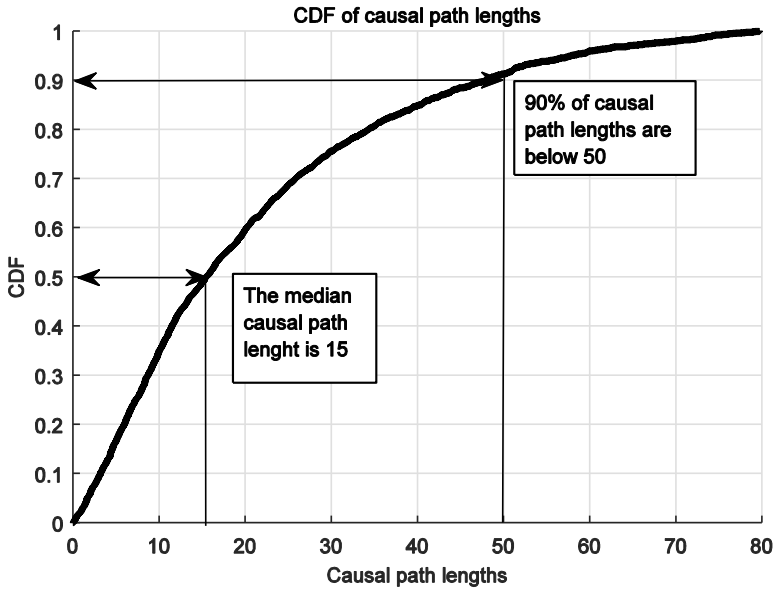


Figure 4.18: CDF of the causal path lengths found in the network

4.4.7 Sources and Sinks

In the previous subsection cells that exert/experience influence on/from their neighbours were identified. These cells were known as sources and sinks respectively. In this subsection these sources and sinks are examined and compared with each other and the general network to see if they have any special properties that stand out. Figure 4.19 shows the CDF of each cell's total equivalent data usage grouped by their causal flow. The three groupings are strong sources (top 10% of cells ranked by causal flow), all cells, and strong sinks (bottom 10% of cells ranked by causal flow). It is readily apparent that the strong sources experience much higher usage than the other two groups. For example, the median total equivalent data usage of a strong source cell is approximately 4.5 times that of the median for all cells on the network.

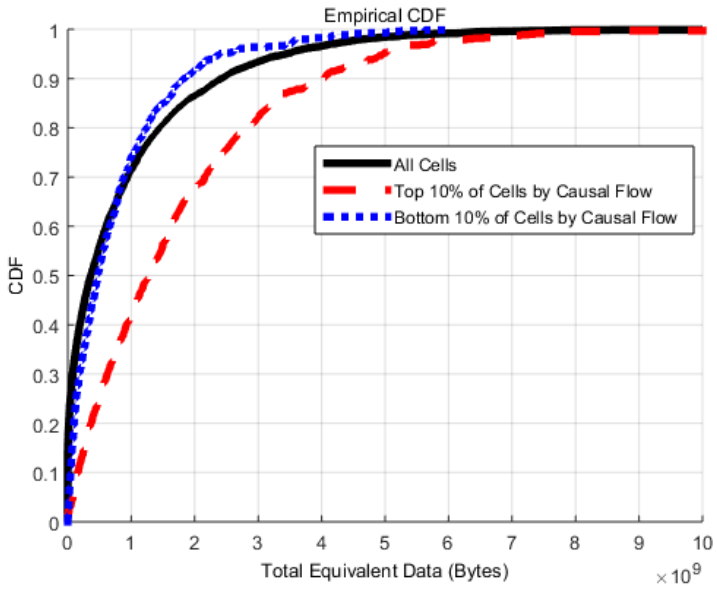


Figure 4.19: CDF of the Total Equivalent Data used per cell ranked by their Causal Flow. The top 10% represent strong sources while the bottom 10% represents strong sinks.

Figure 4.20 shows the CDF of the total number of connections (i.e. data connections, voice or SMS) made per cell over one day as ranked by their causal flow. The top 10% represents strong sources while the bottom 10% represents strong sinks. Figure 4.20 illustrates that strong sources have a much larger amount of connections per day than the other groups. The median strong source cell has approximately 2.5 times the number of connections per day as the median of all cells. Thus, strong source cells generally use the most data and have the largest number of connections in a day.

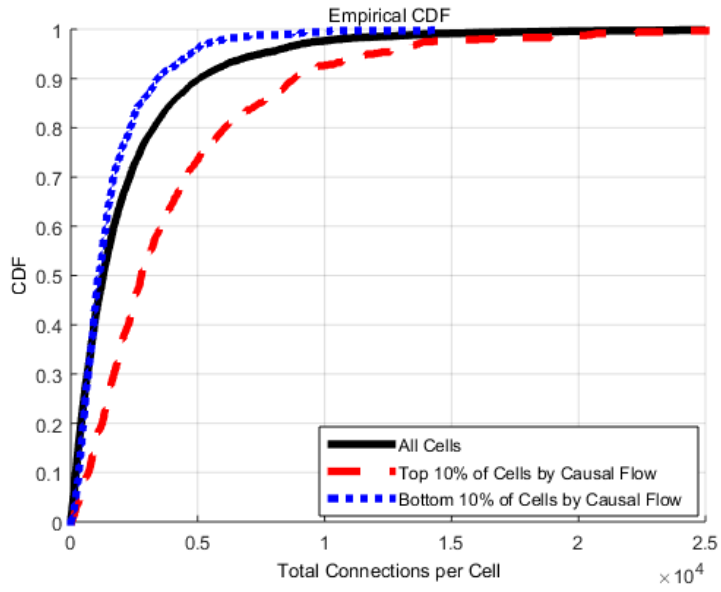


Figure 4.20: CDF of the total number of connections made per cell over one day ranked by their causal flow. The top 10% represent strong sources while the bottom 10% represent strong sinks.

4.4.8 Conclusion

Subsection 4.4.2 found that there is a significant amount of spatial correlation between cell coverage regions in close proximity, decreasing as the separation distance increases. Interestingly, it was found that these correlations vary throughout the day in a similar diurnal pattern to that identified for load in the previous chapter. Spatial correlation increases during times of high load and decreases during times of low load. 4.4.3 - 4.4.7 went beyond spatial correlation by examining the functional influence present in the network. The methodology of Granger causality was employed to identify and understand the underlying functional connectivity present in the network. Causal influences were found to be common in the network with 38% of neighbouring cell pairs experiencing statistically significant influence in either one or both directions. Long chained paths of causal influence were found to flow throughout the network. Anecdotally these paths appear to follow significant transport networks. In future work

a more rigorous examination of these causal flows and their spatial extent would be interesting. Highly influential/influenced cells in the network were also identified and examined. The main difference between these cells and cells with less extreme degrees of influence appears to be how much load/many connections they service. This could again indicate the presence of transport hubs, busy street intersections etc.

4.5 Discussion and Conclusion

The introduction to this chapter identified the importance of understanding the network from a spatial perspective given the larger goal of creating near horizon localised load forecasting techniques. This chapter started out with the creation of spatial representations of base station and sectorised cell coverage regions in 4.2. These spatial coverage region representations are the foundational step in beginning to examine the network spatially. Much of the later work and many of the techniques introduced later require the use of these spatial coverage regions. Given their importance 4.2.3 introduced a novel method to identify and remove errors in their positioning. Subsection 4.2.4 provided a method to visualise the spatial distribution of cellular load across the network. The techniques employed in 4.2.4 could be generalised to not only represent load distribution but also other properties of interest such as connection events, subscriber distribution etc. 4.3 saw the introduction of novel techniques to ascertain the network subscriber specific home and work populations for each cell in the network. These techniques allow for the creation of accurate maps of a network's subscriber base for different classes of cells. 4.3.2 focused on home and work cells but the techniques introduced could easily be generalised to build up maps of different cell e.g. socialising cells etc.

4.3.3 - 4.3.5 examined how subscribers communicate with one another spatially. To explore this the classic gravity model of spatial communication distance was applied in a

novel manner to cellular networks utilising the cellular coverage regions identified in section 4.2.2 and the novel cellular population estimation techniques presented in section 4.3.2. The performance of the model was found to vary largely based on the type of link chosen/the time of the week and to a lesser extent the time of day. The gravity model was found to be only helpful when dealing with large population centres of more than 50,000 inhabitants. As the Republic of Ireland only has five cities with a population of 50,000 inhabitants or more, the gravity model is a poor choice for modelling interurban communication between the country's smaller urban centres. In future interurban work smaller population centres should be amalgamated into larger groups or a more sophisticated model should be employed.

4.4 found that there is a significant amount of spatial correlation between cell coverage regions in close proximity, decreasing as the separation distance increases. Interestingly, it was found that these correlations vary throughout the day in a similar diurnal pattern to that identified for load in the previous chapter. Spatial correlation increases during times of high load and decreases during times of low load. This intuitively makes sense, when the load on a cell or group of cells is very low, for example in the early morning hours, one subscriber connecting to a cell using a data intensive application may greatly increase the load on one cell in percentage terms when compared to its barely used neighbours. During hours of peak load however, the percentage increase will be diminished and also given the finite nature of cellular spectrum the new subscriber's bandwidth will be much more limited reducing his distortive capacity. Significant spatial correlation indicates that for monitoring purposes it may only be necessary to monitor a subset of base stations. 4.4.3 went beyond spatial correlation by examining the functional influence present in the network. The methodology of Granger causality was employed to identify and understand the underlying functional connectivity present in the network. Causal influences were found to be common in the network with 38% of

neighbouring cell pairs experiencing statistically significant influence in either one or both directions. Long chained paths of causal influence were found to flow throughout the network. Anecdotally these paths appear to follow significant transport networks. In future work a more rigorous examination of these causal flows and their spatial extent would be interesting. Highly influential/influenced cells in the network were also identified and examined. The main difference between these cells and cells with less extreme degrees of influence appears to be how much load/many connections they service. This could again indicate the presence of transport hubs, busy street intersections etc.

The above contributions are valuable to network providers and relevant to many advanced network management techniques. They are particularly important to those techniques which rely on a strong spatial understanding such as dynamic spectrum allocation [12], reduced sensing techniques [55], fault detection, and spatially influenced power saving schemes [56] such as the one presented in Chapter 7.

Chapter 5 Local Traffic Load Predictability

5.1 Introduction

Traffic modelling and prediction is a critical element in the performance, planning and evaluation of telecommunications networks and has consequently attracted much attention. However, most of this research has focused on traditional wired broadband which has many different properties and needs in comparison to cellular networks. What work has been carried out on cellular networks is mostly focused on older voice-centric networks and datasets [12, 21]. However, due to the increasing capabilities of devices connecting to the cellular network and the concomitant rise in data usage, cellular networks have shifted from being voice-centric to data centric networks [24, 25]. This shift has resulted in an on-going explosion of traffic on cellular networks at the same time as Average Revenue Per User (ARPU) stagnates or falls [47]. This fundamental challenge has inspired research into new ways to more efficiently use limited network resources such as spectrum [12] or power [16] while still meeting growing user Quality of Service (QoS) expectations. Much of the promising work in this area involves Self Organising Networks (SON) that can dynamically manage their resource usage [12, 16, 82]. An important facet of many of these SON scenarios is the accurate modelling and prediction of traffic load in locally contiguous spatial areas. Up until now, much of the focus on traffic load predictability has been concerned with macro scale network wide predictions of load such as in [27, 83]. However, macro scale predictions are of limited practical value for many SON applications such as green networks [28] and spectrum sharing [12]. For such applications, groupings with finer spatial resolution are required. Thus, the central aim of this chapter is to identify smaller subsets of the network that provide sufficient predictability to allow for their use in SON techniques. The subsets must be sufficiently small and spatiality continuous so as to be

useful for SON techniques. These subsets provide network operators with new ways of viewing their network as opposed to the more traditional macro whole network view or the individual BS view [33]. To that end, this chapter aims to examine the predictability of network load and also defines and examines the predictability of three possible spatially contiguous coverage region aggregations of the network. In Chapter 6 these coverage region aggregations will be used to create localised predictive models of cellular load. Chapter 7 will take these localised predictive models and apply them to a real world SON application. The main contributions of this chapter are:

1) A novel examination of how different levels of load, service type, temporal aggregation, and spatial aggregation affect traffic load predictability.
2) The creation and examination of practical real world spatially contiguous aggregations of network coverage regions.

The remainder of this chapter's sections are laid out as follows:

- 5.2 introduces concepts from information theory and applies these to the traffic load across the various service types. This provides a framework for understanding the relative predictability of the various service types, how this varies between cells for the same service type, and an understanding of how predictability changes with time of day and load.
- 5.3 introduces some of the most practically useful levels of spatial aggregation and examines how they influence predictability.
- 5.4 provides a concluding discussion to the chapter.

5.2 Traffic Predictability

5.2.1 Predictability

This section will use concepts from entropy theory to examine the predictability of network load. Subsection 5.2.2 discusses entropy theory from which concepts are taken to quantify the predictability of data load while 5.2.3 introduces the methodology used to apply it to this dataset outlined in Chapter 2. 5.2.4 examines how entropy varies across the various service types; 5.2.5 explores the relationship between predictability and load.

5.2.2 Entropy

In recent years, frameworks and tools from information theory [84] have been applied to disparate fields of study from human mobility [84] to the predictability of market returns [85]. Information theory originated from the study of the digital transmission of random variables [86]. The objective was to find the most efficient method/coding for the transmission of these variables. It was found that the greater the uncertainty of a random variable, the longer the most efficient possible transmission code would be. This can be precisely quantified, and thus, provides a universal measure of the uncertainty of a random variable [86]. This universal measure of the uncertainty of a random variable is called entropy. Entropy is employed in this work as it provides a precise definition of the informational content of predictions via the appropriate Probability Mass Functions (PMFs). (Note that PMFs are employed as opposed to Probability Density Functions due to the data being quantised into discrete levels). Entropy also proves to be a generally applicable concept as it makes no assumptions about the underlying model. Thus, entropy is used in this work to provide a metric for traffic predictability across disparate BSs/cells and utilising a variety of different prior

and/or auxiliary information. For the interested reader [87, 88] provide a more detailed discussion of the applicability of entropy as a predictability metric in different application domains.

5.2.3 Quantifying Predictability

The dataset discussed in Chapter 2 was processed with the traffic of the three services (voice, data and SMS) sorted by time and cell ID. The traffic during a certain time period *i* within a given cell is quantized into *Q* quantisation levels. The quantisation level of the traffic at time *i*, *QuantLevel(i),* is given by equation (5.1):

$$QuantLevel(i) = \begin{cases} ceil\left(Q \times \dfrac{ObsTraf(i)}{Capacity}\right), if\ ObsTraf(i) \geq 1 \\ 1, \qquad\qquad otherwise \end{cases} \tag{5.1}$$

where *ObsTraf(i)* is the traffic observed at time *i,* ceil is the ceiling function which maps a real number to the least succeeding integer [89], and *Capacity* is the traffic capacity of a given cell. The capacity of a cell varies depending on the technology used (GPRS, EDGE, HSDPA etc.), the number of transceivers employed, etc. Approximation is required as it is not possible to give an exact figure for the capacity of a cell; capacity varies from cell to cell and throughout the day depending on local conditions such as interference, the modulation scheme used, etc. [33]. Thus, for convenience the traffic load in every cell is quantised into Q = 10 levels over the target period. From this, the corresponding traffic distributions are obtained. For example, Figure 5.1 depicts the PMF in one cell derived from the quantised levels for the three services. For the data service depicted in Figure 5.1 the cell under investigation spends approximately 22.5% of its time with a load in the lowest decile, approximately 2% of time in the highest decile, etc. This indicates better than uniform predictability i.e. the cell spends a disproportionate amount of time in the lowest quantisation level meaning its

quantisation level is easier to predict. If the cell spent an equal amount of time in every quantisation level then it would be much harder to predict its quantisation level at any given time. This is a common pattern across the networks with most cells spending a large majority of the time in the lowest quantisation levels as depicted in Figure 5.2. Thus, the presence of this identifiable pattern indicates that useful load predictions can be made for many cells on the network.

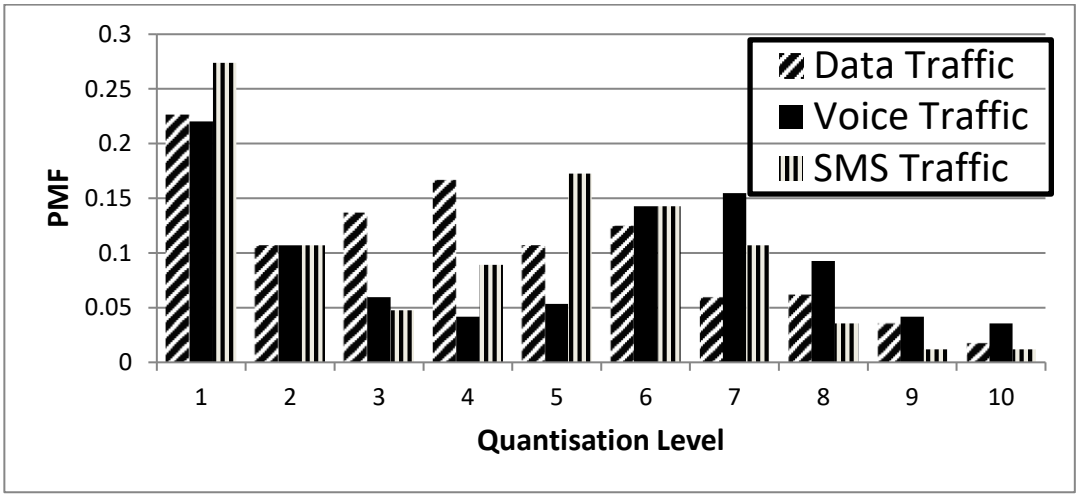


Figure 5.1: The Probability Mass Function of a representative cell

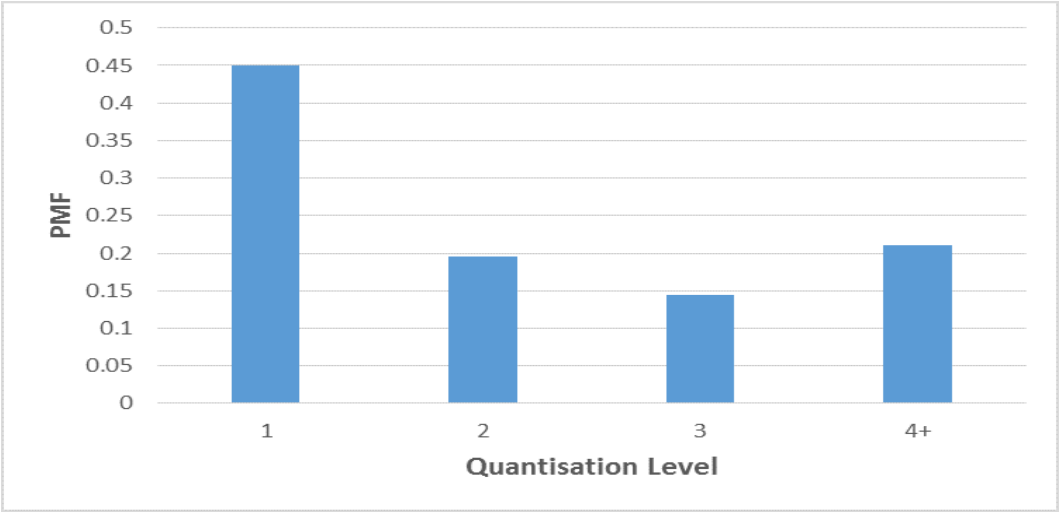


Figure 5.2: The mean PMF of the quantisation level on all cells over one week.

Entropy which is used to quantify the uncertainty of events [88] is defined by a discrete random variable *X* with *n* possible values [x_1 , … x_n] and the corresponding PMF *P(X)*:

$$H(X) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i) \tag{5.2}$$

As discussed in previous chapters, cells will have specific location (e.g. suburban v city centre), technological (e.g. EDGE v HDSPA) and management dependant (e.g. coverage v capacity) characteristics leading to different traffic distributions and hence distinct entropies. Thus, entropy provides a metric to quantify the predictability of traffic in Radio Access Networks (RAN). For example, Figure 5.3 illustrates a weekly traffic trace broken down by service type for two typical cells. Cell 1 has data entropy of 2.19, voice entropy of 2.03 and SMS entropy of 2.26. On the other hand, Cell 2 has data entropy of 2.25, voice entropy of 1.67 and SMS entropy of 1.5. Thus, it can be seen that, for instance, the voice and SMS traffic is more predictable (lower entropy) in Cell 2 than in Cell 1 while the data traffic has similar predictability across both cells. To put these figures into context, an entropy value *H(X)* = 0 would indicate that the traffic load on a given cell was perfectly predictable, with the load staying constant in one of the *Q* quantisation levels for the entire time. This could result from for example, a malfunctioning cell showing no load, an extremely saturated cell constantly in the highest quantisation level of load, etc. Given that there are *Q* = 10 quantisation levels as depicted in Figure 5.1 the maximum possible value of the entropy is given by $H_{max}(X)$ = $\log_2(10)$ = 3.32. This would indicate that each quantisation level has the same probability at each timeslot.

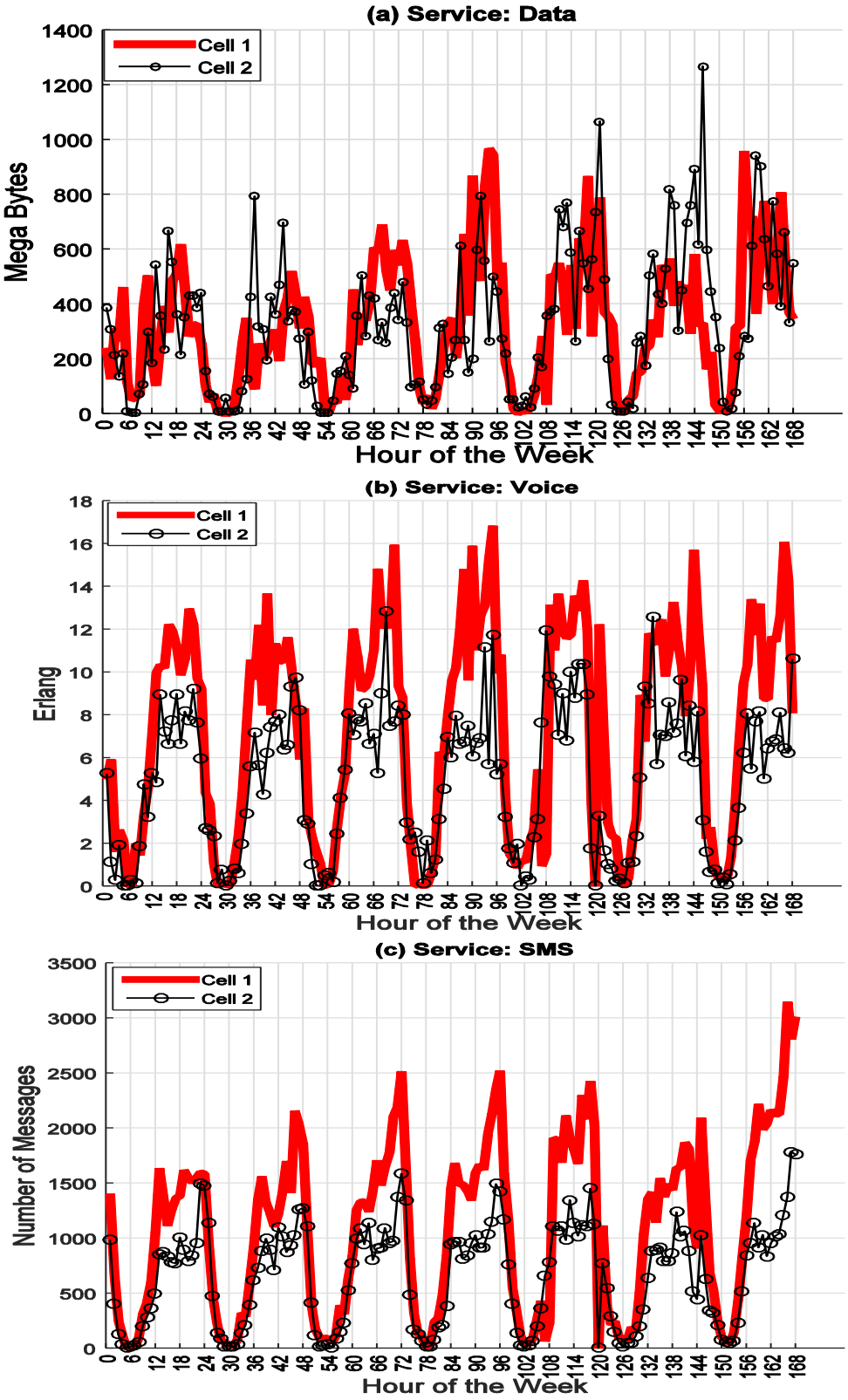


Figure 5.3: The traffic load over one week for two typical cells with different entropies. The entropies for data, voice and SMS for Cell 1 is 2.19, 2.03, 2.26 while the equivalent values for Cell 2 are 2.25, 1.67 and 1.5.

5.2.4 Entropy and Service Type

Much of the previous predictive work carried out on cellular networks has focused on voice-centric networks and datasets [12, 21-23]. However, as shown in Chapter 3, from a network load perspective, cellular data is by far the most important service type offered on the network [24, 25]. Thus, this subsection explores how the predictability of the network is affected by the service type offered. Figure 5.4 is the Cumulative Distribution Function (CDF) of the entropies of each cell broken down by service type. Figure 5.4 illustrates the general pattern that data traffic has the highest entropy (least predictable), SMS has an intermediate entropy, and voice has the lowest entropy (most predictable). This is also borne out by the results presented in Table 5.1 where data not only has the highest mean entropy but also the largest standard deviation at 0.92 bits followed by 0.74 bits for SMS and 0.68 bits for voice traffic.

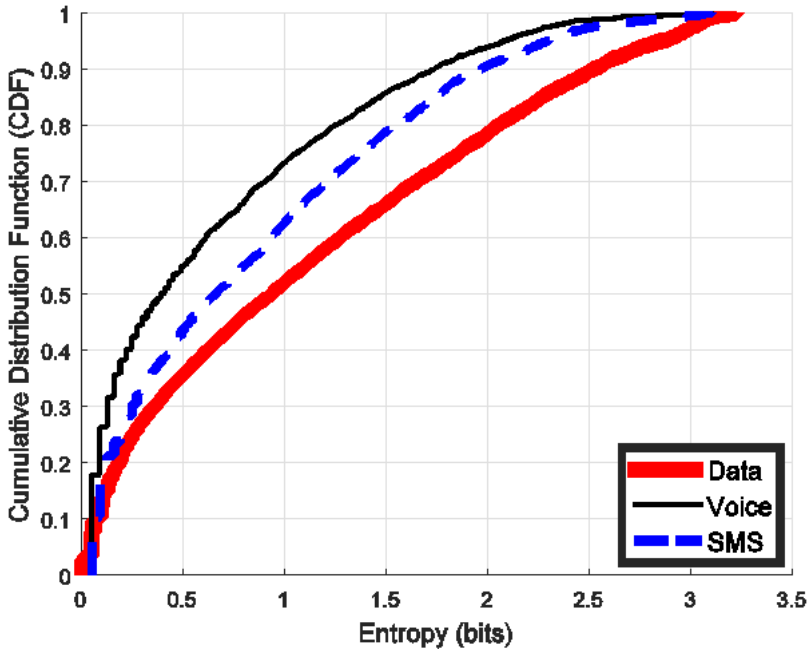


Figure 5.4: CDF of the entropies of all cells broken down by service type. The maximum possible entropy given 10 quantisation levels is $H_{max}(X) = \log_2(10) = 3.32$ bits.

Table 5.1: Entropy Values by Service Type

	Entropy (bits)			
Service	**Mean**	**Std Dev (σ)**	**Min**	**Max**
Data	1.12	0.92	0.01	3.225
Voice	0.66	0.68	0.05	3.11
SMS	0.86	0.74	0.05	3.10

Given the ever growing predominance of data on the network (Chapter 3) it is unfortunate that it is also the least predictable service. However, the predictability of the data service can be improved if only certain times of the day are considered. For example, Figure 5.5 shows the low usage of the network during large parts of the day, particularly the early morning hours. These are the very hours that many resource rationalisation strategies are best suited for [28]. If only the early morning hours are examined from for example 2am to 7am then the mean entropy of the data service drops from 1.12 bits to 0.57 bits. The mean PMF for all cells over all hours, and all cells only over the early morning hours is plotted in Figure 5.6. It illustrates that the reason for the lower entropy value during the early morning hours is the disproportionately large amount of time spent in the bottom quantisation level (89% v 64%). This is a result of the much lower load demand on the network during these times; if the load rarely ever moves beyond the first quantisation level it is much more predictable.

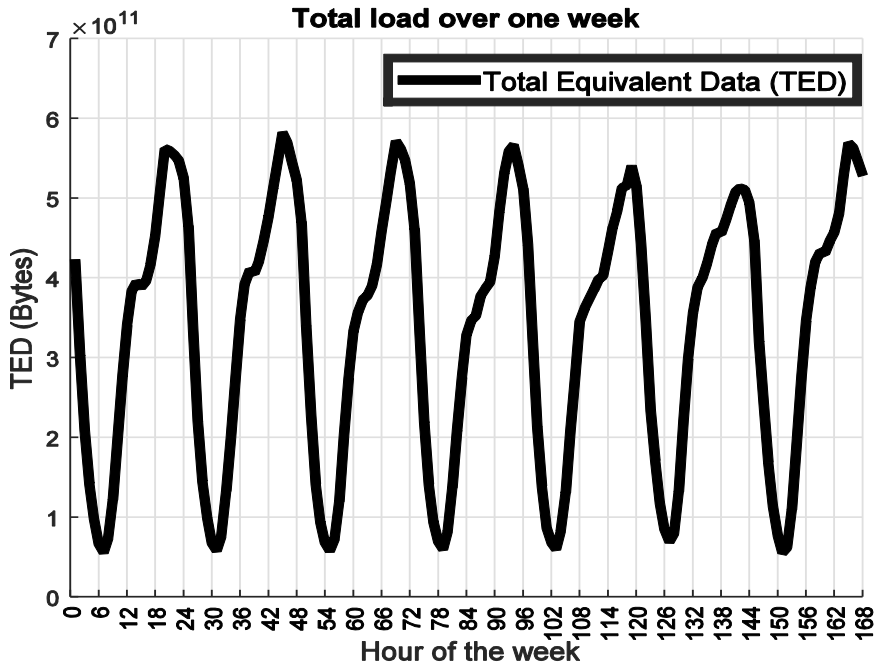


Figure 5.5: Total network load expressed as Total Equivalent Data (TED) in bytes over the course of one representative week. Note that hour zero is 0:00 a.m. on Monday morning. Note this figure was originally presented as Figure 3.1 and is reproduced here for the reader's convenience

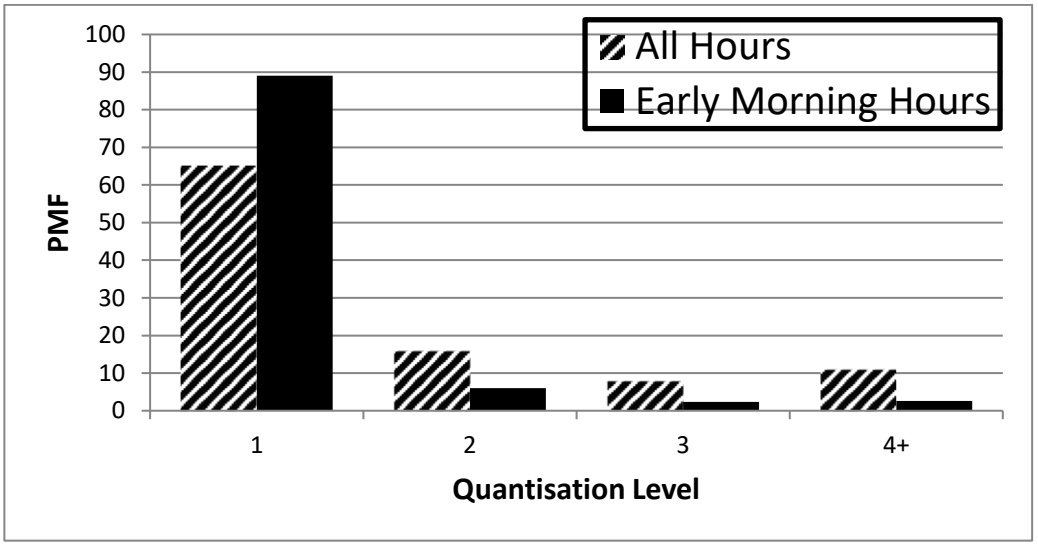


Figure 5.6: The mean PMF of all cells for the data load broken down by all hours and just the early morning hours 2am-7am.

5.2.5 Entropy V Cell Load

From Table 5.1 one sees that there is a large variation in the entropies of cells even within the same service type. As discussed in Chapter 3 the load on cells varies

considerably across the network; a small amount of cells account for a disproportionately large amount of the traffic. These cells are referred to as capacity cells as they are added to the network to increase capacity in high load areas (as opposed to coverage cells which provide basic coverage with limited capacity over large sparse areas as discussed in Chapter 3). As discussed in Chapter 7 these capacity cells provide the greatest opportunities for power savings due to their density and have the most valuable spectrum due to their dense urban locations. Thus, the relationship between cell load and predictability is especially important.

Interestingly, all three services show a strongly positive linear correlation between the weekly load experienced on a cell and the cells entropy. To quantify this correlation, Pearson's Product Moment Correlation Coefficient (PPMCC) [90], denoted r (a measure of the strength of the linear relationship between two variables), was calculated for the relationship between the load of each of the three service types (data, voice, SMS) and entropy. The relationship between data usage and entropy was found to have r = 0.98, voice usage and entropy has r = 0.93, and SMS usage and entropy has an r value of 0.91 indicating a strong correlation. (r can range between 0 and 1, with values close to 1 indicating a strong relationship). In Figure 5.7 the relationship between weekly cell traffic and entropy for the three services is presented. All three services show a strong positive relationship between entropy/unpredictability and load. All three also show that at very high loads the rate of increase in entropy with data slows or levels off. The small amount of extremely heavily loaded cells spend much of their time at saturation and thus their entropy does not have much scope to increase further. The correlation between load and predictability means that the least used cells are the most predictable.

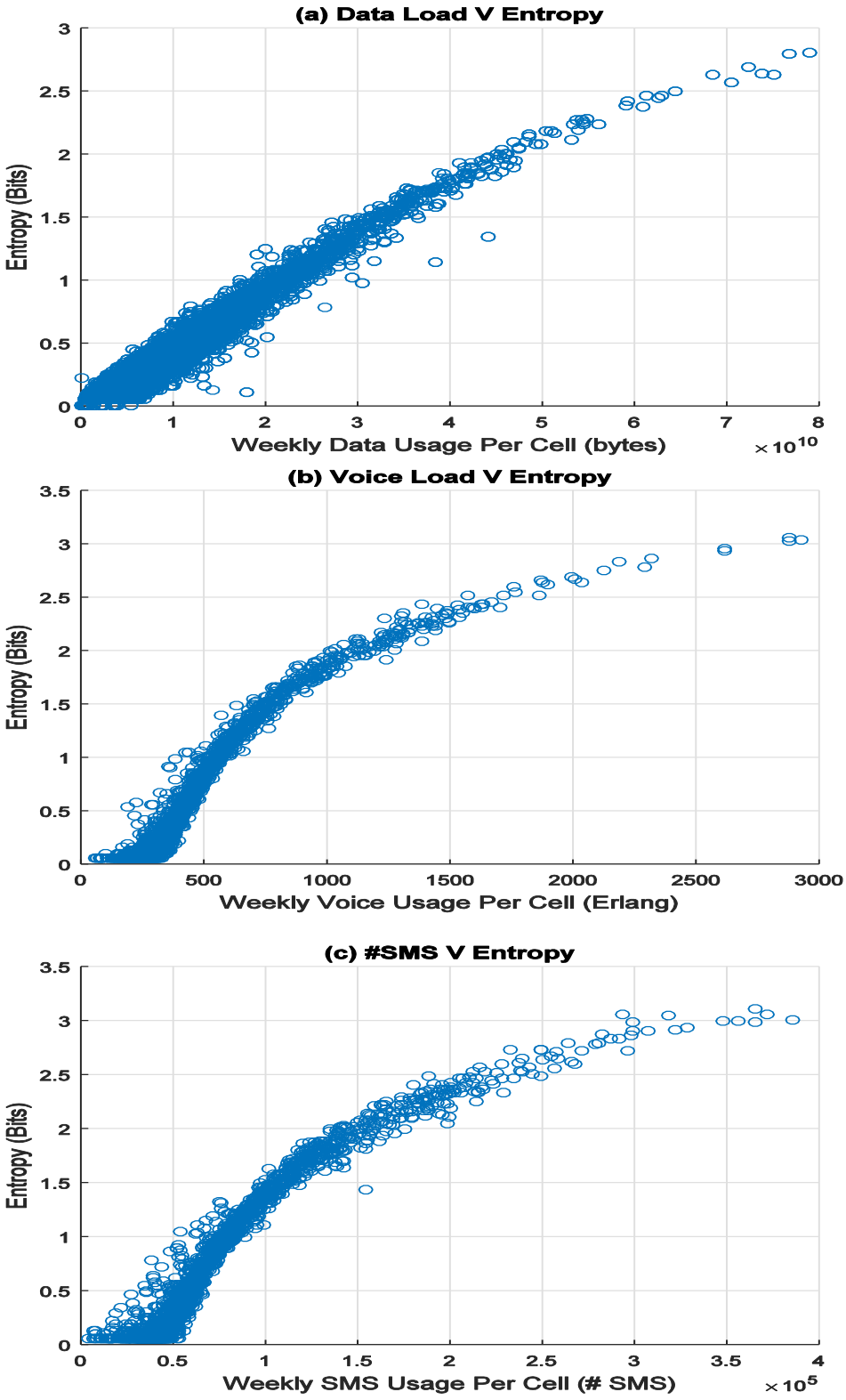


Figure 5.7: Relationship between cell traffic and entropy for data, voice and SMS respectively

5.2.6 Predictability Conclusion

The predictability of all three service types was found to vary largely over all the cells studied. It was found that voice is the most predictable service, followed by SMS and then finally data. Given the growing predominance of data it was unfortunate that it was the least predictable. However, it was found that during particular times of interest e.g. the early morning hours data load became much more predictable. This was due to the lower loads experienced during these hours resulting in the load staying in the lower quantisation levels for a disproportionate amount of time. This better predictability during hours of low usage is particularly useful given that these are the hours most likely to benefit from advanced resource management techniques. 5.2.5 explored the relationship between predictability and load; it was found that cells with lower loads were more predictable. This is an encouraging result as it is these cells in particular which are the greatest source of the large underutilisation of network resources identified in Chapter 7. Thus, they are also some of the cells that could benefit the most from predicative models informing resource utilisation improvement techniques.

5.3 Levels of Spatial Aggregation

5.3.1 Introduction: Levels of Spatial Aggregation

Forecasting short term load on the macro network scale is possible with a high degree of accuracy [27, 83]; however, it is of limited practical value for many applications such as green networks [28] and spectrum sharing [12]. For such applications, groupings with finer spatial resolution are required. In the following subsections, three useful levels of spatial aggregation are introduced: no aggregation i.e. individual cells, overlapping cells, and coverage grids. To aid in the comparisons between the different levels of spatial aggregation, a comparison subsection, 5.3.5 is included at the end of this section. To

simplify the presentation of results, and in keeping with Long Term Evolution (LTE) standards, all three services are aggregated together to give one total figure for load Total Equivalent Data (TED) as described in 3.2.2.

5.3.2 Individual Cells

The individual cell level is the finest grain of spatial resolution available in the dataset and also the most difficult to make predictions for due to its relatively higher entropy as shown in Figure 5.16 which is located at the end of subsection 5.3.4.

Although the traffic load fluctuates over time, the traffic at the same time on consecutive days or during consecutive hours is relatively stable. The short/medium term traffic stability at the cell level is assessed by calculating the medium term traffic variation $V(i, h_{current})$ at hour $h_{current}$ for cell i as:

$$V(i, h_{current}) = \left| \frac{T(i, h_{current}) - T(i, h_{previous})}{Capacity(i)} \right| \quad (5.3)$$

where $T(i, h_{current})$ is the traffic load on cell i at the current hour $h_{current}$ and $T(i, h_{previous})$ is the traffic load on cell i during a previous hour (h-1,h-24 etc) $h_{previous}$. Capacity(i) is the maximum capacity of cell i defined in the same way as in equation (5.1).

Figure 5.8 plots the CDF of the short term traffic variation $V(i,h)$ for all cells for each pair of hours h and h+1 over one month. Figure 5.8 shows that the median cell has a mean hour to hour traffic variation of approximately 3% of the cells' total capacity over one month. Figure 5.8 also shows that 90% of cells have a mean hour to hour variation of less than 9% of their overall capacity. In addition to the mean for each cell Figure 5.8 plots: 1) the 95th percentile of hour to hour load variation for each cell over a month as a percentage of that cells' total capacity and 2) the maximum hour to hour traffic

variation for each cell as a percentage of that cells' capacity. Figure 5.8 illustrates that 95% of the time the median cell has an hour to hour variation of less than 9% and a maximum hour to hour variation of approximately 10% of the cells' capacity. At the upper end of the distribution of cells presented in Figure 5.8, 90% of cells have a 95th percentile deviation of less than 25% and a maximum hour to hour deviation of less than 36% of their capacity.

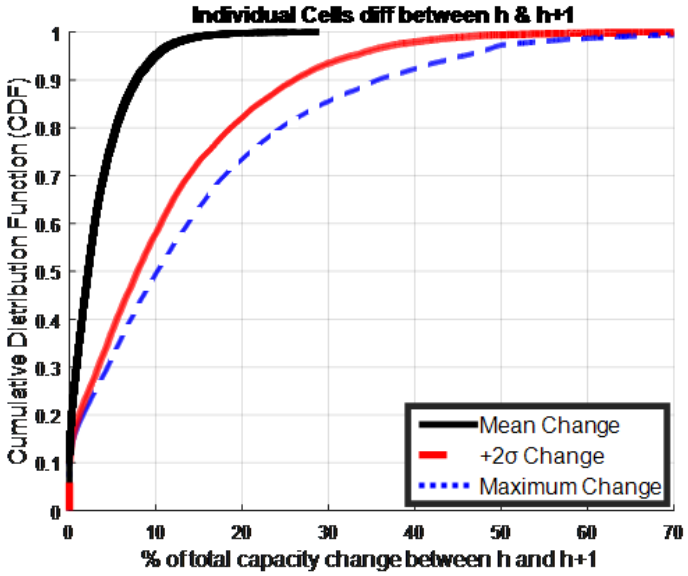


Figure 5.8: CDF of % of total capacity usage change for individual cells when comparing hour h to hour h+1.

Figure 5.9 plots the CDF of the medium term traffic variation $V(i,h)$ for all cells for each pair of hours h and h+24 over one month. Figure 5.9 shows that the median cell has a mean day to day traffic variation of approximately 4% of the cells' total capacity over one month. Thus, this is higher than the equivalent figure for the hour to hour variation shown in Figure 5.8. This stronger autocorrelation of load with the previous/next hour than the same hour yesterday/tomorrow is present across all the different aggregation levels examined in this section. 95% of the time the median cell has a day to day variation of 10% and a maximum day to day variation of approximately 14% of the cells' capacity. At the upper end of the distribution of cells presented in Figure 5.9, 90% of cells have a 95th percentile deviation of less than 30% and a maximum day to day

deviation of less than 45% of their capacity. Considering that these results are for individual cells which will be shown in 5.3.5 to be the most volatile, these results highlight the possibility of useful load prediction. For example, 3.3 demonstrated that that a large majority of cells are only ever using a small percentage of their total capacity. Also, as already shown (Chapter 3) even heavily loaded cells suffer from the classical resource provision peaking problem. That is, during most of the day the vast majority of cells are severely underutilised with peak to trough ratios in excess of 10:1 being common. Cells roughly follow a set daily pattern of load depending on demographics, topography, etc (see Chapter 3). Thus, it is generally possible to predict these periods of peaks and troughs. Therefore, if the load on a cell is currently 10% of capacity, and it is known from Figure 5.8 that the cells' maximum hour to hour variation is less than 10% then it is reasonable to predict the load will not exceed 20% of capacity. If the operator wished to err on the side of caution they could double this value to 40% and still free up carrier frequencies for secondary usage or turn off transceivers to reduce operating expenditure.

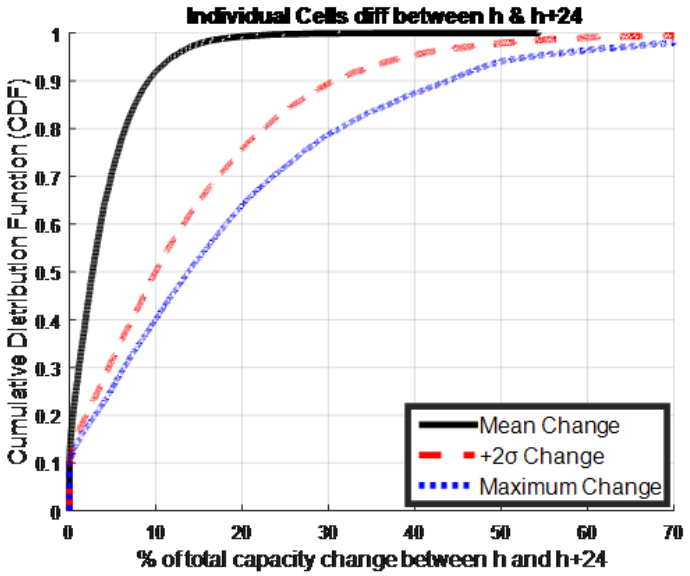


Figure 5.9: CDF of % of total capacity usage change for individual cells when comparing hour h to hour h+24.

5.3.3 Overlapping Cells

Operators normally think of cells at the BS or individual cell level, however, there are other possible aggregations to consider. For example, Figure 5.10 shows the cell coverage zone for Dublin city (as defined by the Central Statistics Office [72]). As shown in Figure 5.10 the operator has many cells operating at different frequencies in the same coverage areas to increase capacity as discussed in Chapter 2 and Chapter 4. Figure 5.10 shows that in almost all cases there are at least two cells covering an area and in densely populated areas often many more. This level of spatial aggregation is useful because it conforms to pre-existing areas which are defined by the network operator and already in use. Thus, it is already known that these are realistic coverage zones and that overlapping cells are capable of covering each other's area.

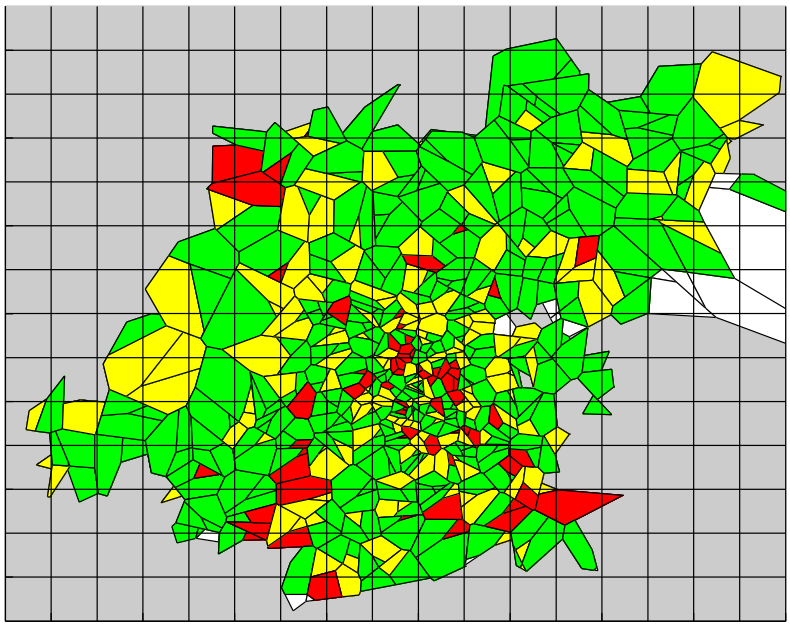


Figure 5.10: Cell coverage zones for Dublin city [72]. Each square corresponds to 1km^2. White zones are covered by one cell, green by two, yellow by three, and red by four or more.

Subsection 5.3.2 introduced the concept of short/medium term traffic variation for individual cells and defined the variation metric in equation (5.3). This section again examines the short/medium term traffic variation but instead of individual cells it examines groups of overlapping cells. In equation (5.3) the denominator was simply the capacity of the individual cell. Equation (5.3) is now modified for the general case of more than one cell of varying capacity:

$$V(i,h) = \left| \frac{T(i,h_{current}) - T(i,h_{previous})}{\sum_{q=1}^{n} Capacity(q)} \right| \tag{5.4}$$

where $V(i,h)$ is the percentage of variation of total aggregate capacity between the hours under investigation for overlapping cell group i at hour h. Each cell group i is comprised of n individual cells denoted by q. $T(i,h_{current})$ is the traffic load on overlapping cell group i at the current hour $h_{current}$ and $T(i,h_{previous})$ is the traffic load on the overlapping cell group i during a previous hour (h-1,h-24 etc) $h_{previous}$. Capacity(q) is the maximum capacity of cell $n = q$ and is defined in the same way as in equation (5.1).

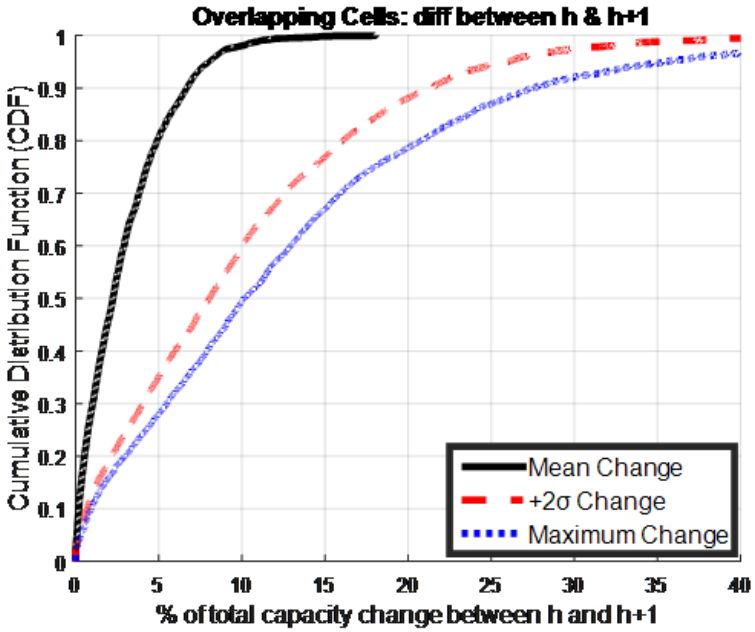


Figure 5.11: CDF of % of total capacity usage change for overlapping cells when comparing hour h to hour h+1.

Figure 5.11 plots the CDF of the short term traffic variation $V(i,h)$ for all groups of overlapping cells for each pair of hours h and $h+1$ over one month. Figure 5.11 illustrates that the median overlapping group has a mean hour to hour traffic variation of approximately 3% of the group's total capacity over one month (compared with 4% for individual cells). Figure 5.11 also shows that 90% of groups have a mean hour to hour variation of less than 7% of their overall capacity (compared to 9% for individual cells). In addition to the mean for each group Figure 5.11 plots: 1) the 95th percentile of hour to hour load variation for each group over a month as a percentage of that group's total capacity 2) the maximum hour to hour traffic variation for each cell as a percentage of that group's capacity. Figure 5.11 shows that 95% of the time the median group has an hour to hour variation of 8% and a maximum hour to hour variation of approximately 12% of the group's capacity. Figure 5.11 demonstrates that 90% of groups have a 95th percentile deviation of less than 20% and a maximum hour to hour deviation of less than 25% of their capacity (compared with 25% and 36% respectively for individual cells).

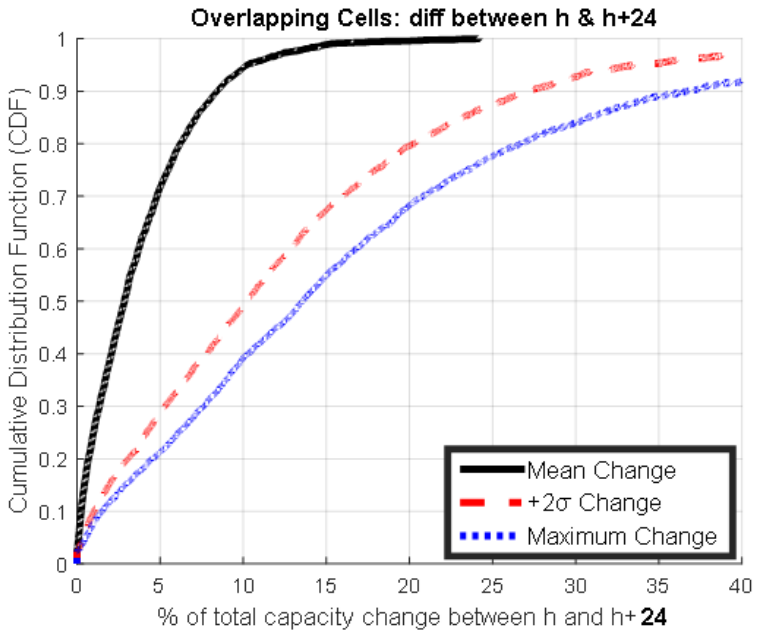


Figure 5.12: CDF of % of total capacity usage change for overlapping cells when comparing hour h to hour h+24.

Figure 5.12 plots the CDF of the medium term traffic variation *V*(*i,h*) for all overlapping groups for each pair of hours *h* and *h*+24 over one month. As in the individual cell case, the variation is greater between the same hours on consecutive days than it is for consecutive hours. The next subsection will examine a larger aggregation of cells/BSs known as coverage grids.

5.3.4 Coverage Grids

Subsection 5.3.2 explored the predictability and short/medium term traffic variation of individual cells. 5.3.3 extended this to examine the predictability and short/medium term traffic variation of aggregated overlapping cell groups. This subsection will explore similar features of a larger useful spatial aggregation of BSs (usually several sectorised cells per BS) denoted as coverage grids. Coverage grids are groups of BSs that are within a certain transmission distance of each other and can provide coverage to other members of the group if a BS is disabled for whatever reason. These coverage grids differ from the overlapping cell zones discussed in 5.3.3 as they do not merely consist of already overlapping coverage areas. That is, they use techniques from SON (Self Organising Networks) to dynamically alter their group membership depending on certain limitations and desired outcomes. The size of the transmission distance and how groups are formed depend on many factors such as the local topography, user density, SNR etc. Coverage grids are partitioned so that each BS in each coverage grid is equivalent. BSs are equivalent if they can replace each other while communicating with subscribers. Specifically, if the distance between two BSs *i* and *j* is *d(i,j)*, then BSs *i* and *j* are equivalent if:

$$r_i + d(i,j) \leq R_j \tag{5.5}$$

$$r_j + d(i,j) \leq R_i \tag{5.6}$$

where r_i and r_j are the normal communication ranges, R_i and R_j are the maximum possible communication ranges of i and j respectively. The location information discussed in Chapter 4 is used in conjunction with the transmission range of each BS to decide whether proximate BSs are equivalent or not. The transmission range of a BS may vary from 200m to 2km in cities and from 1km to 20km in rural areas depending on several features including topography, population/building density, etc. [33]. The local population of a BS was calculated in 4.3.2, this combined with the BS coverage area [4.2] is used as a proxy for population/building density. The maximum transmission range of each BS is decided on a sliding partition of population per coverage area. The maximum transmission range for each BS is partitioned as in Table 5.2. For example, the 91st to 100th percentiles (most densely populated BSs) have a maximum transmission range assigned of 250m, the bottom decile (least densely populated BSs) have a maximum transmission range of 20km. The grid formation algorithm used is outlined in algorithm 5.1.

Table 5.2: Maximum transmission range assignment

Maximum Transmission	Percentiles of BS by ordered
250	91-100
500	81-90
1000	71-80
2000	61-70
3000	51-60
4000	41-50
5000	31-40
10000	21-30
15000	11-20
20000	1-10

Algorithm 5.1:

BS switching algorithm

1: Let L = (l_i) be a one dimensional BS location array (*m*x1) where m is the number of BS in the network and l_{i1} is the geographic location of BS *i* (2.6).

2: Let D = (d_{ij}) be a two dimensional distance array (*m*x*m*) where m is the number of BS in the network and d_{ij} is distance between BS *i* and *j*

3: Let T = (t_i) be a one dimensional transmission distance array (*m*x1) where m is the number of BS in the network and t_{i1} is the maximum transmission distance of BS *i*.

3: Let R = (r_i) be a one dimensional normal transmission range array (*m*x1) where m is the number of BS in the network and r_{i1} is the current transmission distance of BS *i* (calculated in 4.2).

4: Let F = (f_i) be a one dimensional flag array (*m*x*1*) where m is the number of BS in the network and f_{i1} is = 0 if the BS does not currently belong to a grid and 1 otherwise.

5: From L, select the BS, *x*, with the most north westerly location in the network and set F(f_x) = 1. BS *x*, is now the sole member of a new grid g_1. Subsequent grids will be denoted g_2,… g_m where *m* is the number of grids that will eventually be formed from the network (unknown until all BS are assigned to a grid).

6: Iterate through each column of D, d_{x1},d_{x2},..d_{xm} (i.e. the distance between BS *x* and all other BSs). For each element check the corresponding element in F, f_y. If f_y = 1 (the BS already belongs to a grid) discard the BS and move onto the next column of D, d_{xy+1}.

7: If f_y = 0 check if BSs *x* and *y* are equivalent i.e. $r_x + d_{xy} \leq T_x$ and $r_y + d_{xy} \leq T_y$

8: If *x* and *y* are equivalent then *y* is added to grid g_1 and f_y = 1. If *x* and *y* are not equivalent then continue iterating through each column of D, d_{x1},d_{x2},..d_{xm} and checking relevant BS for equivalence and grid formation availability as in steps 6 and 7. However, to be added to

a grid, the BS under investigation must be equivalent to all current members of the grid.

9: When all BS have been checked for equivalence with grid g_1, select the most north westerly *unassigned* BS designated q, with $F(f_q) = 0$ and designate it as grid g_2. Build up g_2 from the unassigned BS and once g_2 is completed continue creating grids until all BS are flagged as belonging to a grid in F.

An example of a simple coverage gird generation procedure is illustrated in Figure 5.13. Here BS_1 is equivalent to BS_2 and BS_3, but is not equivalent to BS_4, BS_5 or BS_6. A coverage grid is formed when all the BSs in it are equivalent as in equations (5.5) and (5.6). There are many possible permutations of coverage grids based on the origin point chosen. For this work, the top left corner of a spatial region is defined as the origin point for grid formation. That is, the first BS considered for grid formation is the one closest to the top left corner. Then, all equivalent BSs are clustered from top down and left to right. A new coverage grid is generated when a BS is found to be not equivalent to at least one BS in the current coverage grid. This process is repeated until the bottom right corner of the spatial region is reached and all BSs in the spatial area are assigned a coverage grid. In the example case presented in Figure 5.13, there are three coverage grids formed. A different choice of origin, e.g. the bottom right would result in a different coverage grid formation. However, this will not greatly affect the overall result as it does not alter the inherent proximity relationships in the network. For example, choosing the bottom right as opposed to the top left as the origin would still result in three coverage grids comprising (BS_6,BS_5), (BS_4,BS_3), and (BS_2,BS_1). The average overall BS density of the coverage grids would still remain the same at 2 under both scenarios, however, local capacity would just differ slightly at different spots.

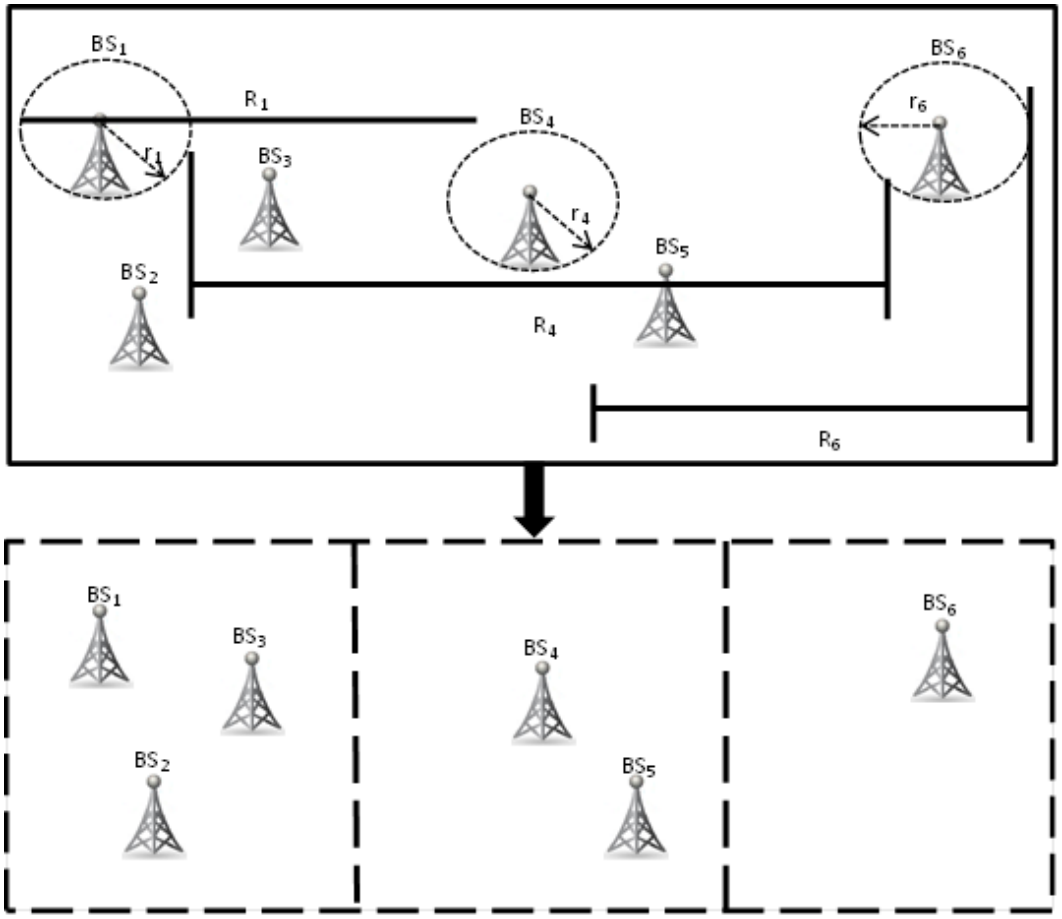


Figure 5.13: Example of coverage grid formation. Top: The spatial locations and transmission distances of 6 BS. Bottom: The coverage grid divisions.

To quantify the short/medium term variation of the coverage grids equation (5.4) is employed where the groups are now coverage grids of cells that do not necessarily overlap.

Figure 5.14 plots the CDF of the short term traffic variation $V(i,h)$ for all coverage grids for each pair of hours h and $h+1$ over one month. It shows for instance that the median coverage grid has a mean hour to hour traffic variation of approximately 2.2% of the group's total capacity over one month (compared with 4% for individual cells and 3% for overlapping groups). Figure 5.14 also shows that 90% of coverage grids have a mean hour to hour variation of less than 4% of their overall capacity (compared to 9% for individual cells and 7% for overlapping groups of cells). In addition to the mean for each

coverage grid Figure 5.14 plots: 1) the 95th percentile of hour to hour load variation for each coverage gird over a month as a percentage of that grid's total capacity and 2) the maximum hour to hour traffic variation for each grid as a percentage of that grid's capacity. 95% of the time the median coverage grid has an hour to hour variation of less than 5% and a maximum hour to hour variation of less than 7% of the grid's capacity (much lower than individual cells which have values of 9% and 10% for the same metrics). At the upper end of the distribution of coverage grids presented in Figure 5.14 illustrate that 90% of grids have a 95th percentile deviation of less than 10% and a maximum hour to hour deviation of less than 15% of their capacity (compared with 25% and 36% for individual cells).

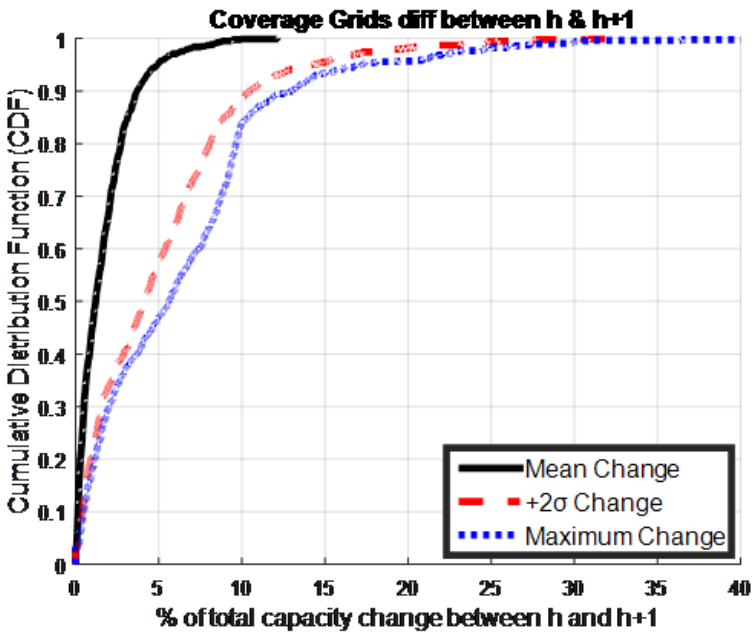


Figure 5.14: CDF of % of total capacity usage change for overlapping cells when comparing hour h to hour h+24.

Figure 5.15 plots the CDF of the medium term traffic variation $V(i,h)$ for all coverage groups for each pair of hours h and h+24 over one month. As in the individual and overlapping cell cases, the variation is greater between the same hours on consecutive days than it is for consecutive hours.

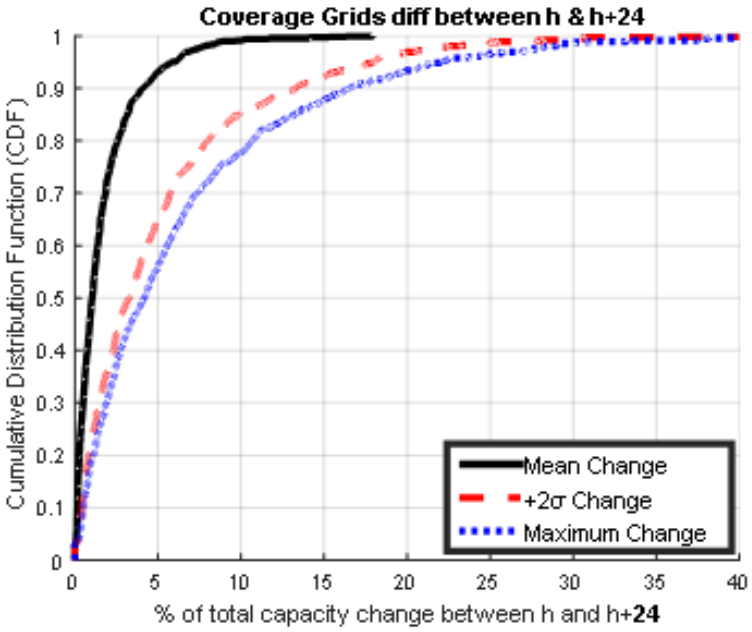


Figure 5.15: CDF of % of total capacity usage change for coverage grids when comparing hour h to hour h+24.

5.3.5 Comparing Levels of Spatial Aggregation

This section has presented the three most practically useful levels of possible spatial aggregation and has illustrated that all three provide promising results for load prediction. The variation in the load between consecutive hours and also the variation between the same hours on consecutive days was examined. The results obtained for the three levels of spatial aggregation are summarised in Table 5.3 and Table 5.4. For all three levels of spatial aggregation it was found that consecutive hours had a smaller variation than the same hours on consecutive days. Table 5.3 shows that 95% of the time, the median coverage grid (in terms of data load) has an hour to hour variation of 5% of its capacity compared to 8% for the median overlapping cell group and 9% for the median individual cell. Thus, this means that the load is more stable hour to hour at larger aggregations making it easier to predict at these aggregations. Table 5.4 shows similar results for the 90th percentile cell/spatial aggregations (in terms of data load). 95% of the time the 90th percentile coverage grid has an hour to hour variation of 10% of

its capacity compared to 20% for the 90th percentile overlapping cell group and 25% for the 90th percentile individual cell.

Table 5.3: The % of capacity change across aggregation levels for the median cell/aggregation

Time Period	Aggregation Level					
	Individual Cells		Overlapping Cells		Coverage Grid	
	Mean of median cell	+2σ	Mean of median cell	+2σ	Mean of median cell	+2σ
Inter Hour	4	9	3	8	2.2	5
Inter Day	4	10	3	8	3	4

Table 5.4: The % of capacity change across aggregation levels for the 90th percentile cell/aggregation

Time Period	Aggregation Level					
	Individual Cells		Overlapping Cells		Coverage Grid	
	Mean of 90th percentile cell	+2σ	Mean of 90th percentile cell	+2σ	Mean of 90th percentile cell	+2σ
Inter Hour	9	25	7	20	4	10
Inter Day	9	30	8	24	4.5	13

The increased stability and predictability of load is further borne out by Figure 5.16 where the largest level of aggregation i.e. the coverage grids have the smallest entropies followed by the overlapping cells while the non-aggregated individual cells have much higher entropies. This is an important result; normally operators examine the network at the level of an individual cell/BS. However, other possible aggregations exist such as the overlapping cells and coverage grids presented here. In fact, the results from this work indicate that if an operator or modeller is concerned with the predictability/stability of the time series, the larger spatial aggregation levels are superior to the individual cell level.

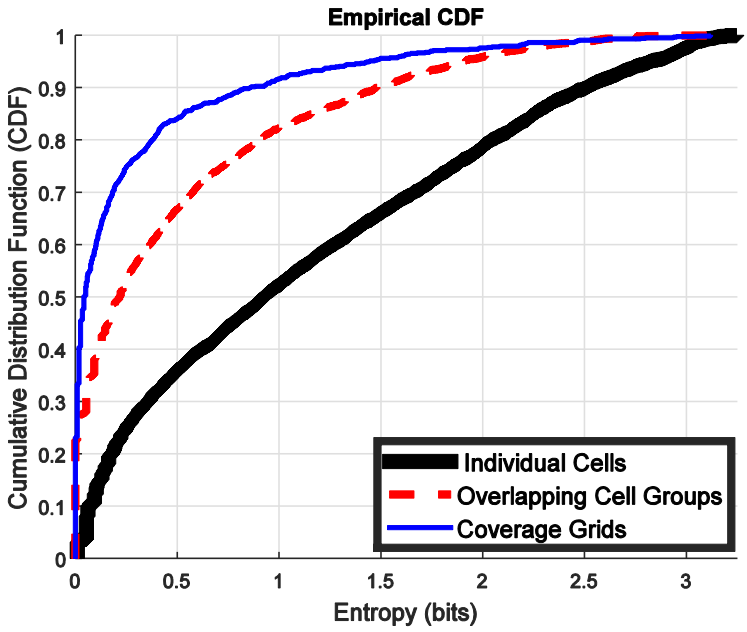


Figure 5.16: CDF of the entropy for different aggregation levels

5.4 Discussion and Conclusion

As discussed in the introduction, much of the work on predictability and modelling of cellular load excluded the primary driver of cellular load, cellular data. Thus, to better understand how cellular data affected predictability relative to the more widely studied service types of voice and SMS, 5.2 examined how entropy varies by service type and time of day. It was found that voice is the most predictable service, followed by SMS and then finally data. The predictability of all three service types was found to vary largely over all the cells studied. Given the growing predominance of cellular data it was unfortunate that it was the least predictable. However, it was found that during particular times of interest e.g. the early morning hours data load became relatively much more predictable. This better predictability during hours of low usage is particularly useful given that these are the hours most likely to benefit from advanced resource management techniques. 5.2 also explored the relationship between predictability and load. It was found that cells with lower loads were more predictable. This is an encouraging result as it is these cells in particular which are the greatest

source of the large underutilisation of network resources (as will be discussed in greater detail in Chapter 7). Thus, they are also some of the cells that could benefit the most from predicative models informing resource rationalisation techniques.

Previous work has shown that forecasting short term load on the macro network scale is possible with a high degree of accuracy [27, 83]; however, it is of limited practical value for many applications such as green networks [28] and spectrum sharing [12]. For such applications, groupings with finer spatial resolution are required. Thus, the central aim of this chapter was to identify smaller subsets of the network that provide sufficient predictability to allow for their use in SON techniques. These subsets had to be sufficiently small and spatiality continuous as to be useful for SON techniques. Two novel subsets (spatial aggregations) meeting these requirements were proposed and compared with the smallest available spatial aggregation, the individual cell level. Individual cells benefit from already being defined and operational in the network; overlapping cells are a summation of individual cells covering the same areas at different frequencies. Coverage grids are more complicated amalgamations of BSs into larger (potentially) mutually coverable regions. The simplicity of both individual cells and overlapping cells is a distinct advantage in their use as predictive regions. However, the greater capacity redundancy of coverage grids make them more applicable to advanced management techniques such as green networks [28] and spectrum sharing [12]. Unfortunately, due to the inherent localised reuse nature that defines cellular networks (coupled with maximum coverage range limitations) larger spatial aggregations above coverage grid are less useful for SONs. Thus, there is a trade-off to be made between predictability and spatial aggregation. Comparing the three spatial aggregation levels in 5.3.5 demonstrated that load is more stable and predictable at larger spatial aggregations. Given coverage grid's greater predictability, coupled with their mutual redundancy, coverage grids appear to occupy the optimal position in the

trade-off between predictability and practicality. Another benefit of a coverage grid centric view of the network is that their inherent spatial redundancy makes them an ideal default grouping for mutual coverage in the event of localised equipment failure. Although not explored further here, the ability to dynamically alter coverage with confidence within the grid aggregation level would also increase the overall networks redundancy. Traditionally, the network is examined and modelled at the individual cell or individual BS level. However, the results of this chapter indicate that if predictability is an important factor in the analysis/model/network management technique, then higher levels of spatial aggregation are more suitable. It is hoped that these aggregations provide network operators with new ways of viewing their network as opposed to the more traditional macro whole network view or the individual BS view.

While this chapter has focused on the inherent predictability of the three levels of spatial aggregation, the next chapter will use these spatial aggregations to create actual predictive models of cellular load.

Chapter 6 Localised Load Forecasting in Cellular Networks

6.1 Introduction

Predictive modelling of network load is an area of growing importance due to the rise of Self Organising Networks (SON) that can dynamically manage their resource usage. The previous chapter discussed the inapplicability of much of the previous network load forecasting work when applied to some promising advanced SON techniques (namely unsuitable spatial scales, voice-centric data sets, etc.). These issues were addressed in the previous chapter by using a nationwide dataset (where the vast majority of network load comes from the cellular data service) to construct and examine novel spatial aggregations relevant to many advanced SON techniques. Thus, this chapter builds on the contributions of the previous chapter by taking the local, spatially contiguous regions defined in it and creating near horizon predictive models of their load. The central argument of this chapter is that network load can be predicted for spatial aggregations defined in the previous chapter with sufficient accuracy to allow for their use in advanced SON techniques.

Section 6.2 discusses the many shortcomings of most traditional methods of evaluating network models and predictability when dealing with cellular networks. Cognisant of these shortcomings, 6.2 proposes a novel metric for evaluating the predictability of network models for cellular network load. Next, 6.3 introduces two methods of predicting local network load and an automatic procedure to build the large amount of models required. 6.4 presents the results obtained across the various levels of spatial aggregation while 6.5 provides a concluding discussion.

In summary the main contributions of this chapter are as follows:

1) The drawbacks of many common forecasting metrics when applied to cellular network forecasting are identified and discussed. Thus, in this chapter a novel and practical metric is proposed and implemented: Absolute Capacity Percentage Error (ACPE).
2) Two methods of load forecasting are applied in a novel manner to cells of different levels of spatial aggregation with a discussion of their results.
3) Two novel methods for the automatic modelling of large amounts of individual cells and their many possible permutations in different spatial aggregations are proposed and used.

6.2 Evaluating Forecast Accuracy

6.2.1 Evaluating Forecast Accuracy Introduction

An important facet of quantifying the accuracy of a forecasting approach and the practicality of applying it to real situations is the forecast metric used. Generally, there are three broad categories of measures employed in evaluating the accuracy of time series forecasts: scale-dependent measures, percentage error measures, and scaled error measures. This section will discuss the three main categories and their advantages and disadvantages with regards to the practical application of local cellular network load forecasting. A novel measure is proposed and justified as the most useful for practical application. In the following sections y_i denotes the *i*th observation of y; $\hat{y}_i$ denotes a forecast of y_i. When discussing time-series data a common term that arises is seasonality. Seasonality is the presence of variations that occur at specific regular intervals, such as quarterly, monthly, weekly etc. In the context of cellular network load

the most obvious "*seasonal*" trend is the daily diurnal pattern as discussed in Chapter 3. Thus, unless otherwise stated a season is defined as one day.

6.2.2 Scale-Dependent Errors

The forecast error is defined as $e_i = y_i - \hat{y}_i$, which has the same scale as the dataset. Thus, accuracy measures that are based on e_i are scale-dependent and cannot be used to compare results obtained from data with different scales. Two of the commonly used scale-dependent measures are based on the absolute error or the square of the errors:

$$Mean\ Absolute\ Error: MAE = mean(|e_i|), \tag{6.1}$$

$$Root\ Mean\ Square\ Error: RMSE = \sqrt{mean(e_i^2)}. \tag{6.2}$$

The use of the absolute value by MAE and the squared value in RMSE avoids negative and positive errors from offsetting each other. The scale dependant nature of these methods renders them unsuitable for their practical application to cellular network load forecasting for the following reasons:

a) As demonstrated in 3.2.2 network load exhibits a strong diurnal pattern with the scale of the load being largely dependent on the hour of the day and, to a lesser extent, the day of the week.
b) As discussed in 3.3 there is a disparity between the scale of the load experienced by different cells in the network.
c) As discussed in Chapter 5 the network will be examined at different levels of network aggregation; this naturally leads to datasets with different scales.

A common technique to overcome some of these problems is the use of normalisation, however, it suffers from many of the same drawbacks as will be discussed in 6.2.4.

6.2.3 Scaled Errors

In [91] the authors suggest the use of scaled error metrics as an alternative to percentage error techniques when working with data on different scales. They propose the scaling of errors based on the training MAE from a simplistic forecasting method. The authors argue that this allows for a more meaningful comparison between models, particularly on disparate data sets across a variety of scales. In the non-seasonal case using the simplistic method, a one-step-ahead forecast is computed from each data point in a given sample. Thus, a scaled error q_j is defined as:

$$q_j = \frac{e_t}{\frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|} \tag{6.3}$$

where y_t denotes the observation of the load y at time t; $\hat{y}_t$ denotes a forecast of y_t. T is the number of observations (samples) of the load y over which the error is to be scaled; the forecast error is defined as $e_t = y_t - \hat{y}_t$.

As both the numerator and denominator include values on the scale of the original data, the result is independent of the data's scale. A scaled error of less than one results when the forecast is better than the mean simplistic forecast of the training data. A value greater than one indicates that the forecast was worse than the simplistic forecast calculated from the training set. As discussed in 3.2.2 network load exhibits a strong diurnal i.e. seasonal pattern which must be accounted for in the simplistic forecast component.

In the case of seasonal data the authors of [91] suggest defining the scaled error by employing a seasonal simplistic forecast:

$$q_j = \frac{e_t}{\frac{m}{T}\sum_{t=m+1}^{T}|y_t - y_{t-m}|} \tag{6.4}$$

where *m* is the seasonality component of the data. For example, setting *m* = 24 uses the value of the load 24 hours ago as a naïve forecast of the load now.

The Mean Absolute Scaled Error (MASE) is thus defined as:

$$MASE = mean(|q_j|) \tag{6.5}$$

Scaled error metrics such as MASE overcome many of the problems identified in 6.2.2. These scaled error metrics are also particularly good for quantifying the usefulness of modelling techniques as they provide, by their definition, an immediate comparison with simplistic forecasting techniques. However, if comparing modelling techniques is not the sole focus scaled error metrics provide results which are not readily understandable for practical application. An ideal metric would be intuitive and easily applied to practical applications such as resource rationalisation where the key questions to be asked are usually of the form "*what percentage of resource x's capacity is/isn't being used*…" Thus, the next subsection introduces another popular class of metrics, percentage errors.

6.2.4 Percentage Errors

Percentage error is defined as:

$$p_i = 100\left(\frac{e_i}{y_i}\right) \tag{6.6}$$

Percentage errors are scale independent and are thus often used to compare forecast performance between different data sets or at different levels of aggregation such as in electricity load forecasting [29]. The most commonly used measure is:

$$Mean\ Absolute\ Percentage\ Error: MAPE = mean(|p_i|) \quad (6.7)$$

Although percentage error techniques solve some of the problems outlined in 6.2.2 they suffer from the following problems:

a) If the load y_i falls to zero there will be a division by zero. y_i can occasionally reach zero, particularly when examining individual low demand cells during the early morning hours.
b) Percentage errors can be extremely large when y_i is very low.
c) Due to the diurnal pattern of load a cell could have for example an hourly load of 200MB at 6AM and 1GB at 9PM with predicted loads of 100MB and 900MB respectively. Using equation (6.6) would yield an error of 50% for 6AM and 6.67% for 9PM with both situations being only 100MB from the true answer. Despite having a much larger error the 6AM prediction may be much more useful; a prediction of a low relative load might indicate the possibility of turning off equipment to save power or share spectrum.

6.2.5 Absolute Capacity Percentage Errors

6.2.4 discussed percentage errors and some of their drawbacks with regard to forecasting the load on cellular networks. This subsection introduces a novel metric referred to as the Capacity Percentage Error (CPE) which is defined as follows:

$$CPE = \left(\frac{e_i}{\sum_{c=1}^{n} Capacity(c)}\right) \times 100 \quad (6.8)$$

where *c* is the index of the cell belonging to a group of *n* such cells. The capacity for each cell is determined as in Chapter 5 and summed over the entire set of cells forming a group or *n=1* for an individual cell. The CPE leads to the ACPE (Absolute Capacity Percentage Error) which is defined as:

$$ACPE = |CPE| \tag{6.9}$$

The ACPE has the following advantageous properties:

1. It adequately handles situations where the load is 0, very small or negative.
2. It allows for the comparison between data sets on different scales which is important when dealing with different Radio Access Technologies (RAT) and levels of spatial aggregation.
3. It is intuitive and easily applied to practical applications such as resource rationalisation where the key questions to be asked are usually of the form "*what percentage of capacity is/isn't being used*…"

Thus, in the following section ACPE will primarily be employed.

6.2.6 Evaluating Forecast Accuracy Conclusion

This section introduced and discussed the benefits and drawbacks of some of the most common metrics used to evaluate predictive models. Due to the problems identified with these metrics, a novel metric ACPE was defined. Throughout the rest of this work, ACPE will be primarily employed.

6.3 Prediction Methods

6.3.1 Prediction Methods Introduction

The results from Chapter 5 have demonstrated the feasibility of cellular load prediction on a nationwide network and laid the foundation for predictions at three different practical levels of spatial aggregation. As shown in [24, 92] several factors can affect the traffic load: time of the day, day of the week, location, special events, etc. Thus, a useful prediction method must be capable of learning the relationships between these factors

and load. There are several possible methods available such as Auto-Regressive Moving Average (ARMA) models [78], Seasonal ARMA models (SARMA) [93], Auto-Regressive Integrated Moving Average Models (ARIMA) [93], Artificial Neural Networks (ANN) [94], wavelet based methods [95], compressed sensing based prediction methods, etc. With due consideration to the accuracy and the computational complexity of traffic prediction both SARIMA and ANN models are employed as recommended by [94, 96].

6.3.2 SARIMA Models

Let y_t: $t = 0,1,2\ldots$ be a non-stationary time series containing seasonality i.e. a seasonal periodic component repeats itself after every s observations, y_t then depends on past values such as y_{t-1s}, y_{t-2s}, *etc.* as well as y_{t-1}, y_{t-2}, *etc.*

Let $w_t = \nabla^d y_t$ where ∇ is a differencing operator and d is the order of non-seasonal differencing. B is the backshift operator such that $Bw_t = w_{t-1}$, $B^2 w_t = w_{t-2}$ … $B^n w_t = w_{t-n}$. Thus w_t is an Autoregressive Moving Average (ARMA(p,q)) process:

$$\varphi_p(B) w_t = c + \theta_q(B) e_t \tag{6.10}$$

where

$$\varphi_{\mathrm{p}}(B) = 1 - \varphi_1 \mathrm{B} - \cdots - \varphi_p B^p \tag{6.11}$$

and

$$\theta_q(B) = 1 - \theta_1 \mathrm{B} - \cdots - \theta_q B^q \tag{6.12}$$

where φ_{p} are the autoregressive polynomials and ϑ_{q} are the moving average polynomials of order p and q respectively. c is the constant or bias and e_t represents the errors which are assumed to be independent and identically distributed (*i.i.d.).* Equation (6.10) is an ARMA model of w_t which is itself a differenced version of y_t. Hence equation (6.10) is an Autoregressive Integrated Moving Average (ARIMA(p,d,q)) model of the original non-

differenced time series y_t with autoregressive order *p*, non-seasonal differencing order *d* and moving average order *q.*

If $w_t = \nabla^d \nabla^D y_t$, where D is the order of seasonal differencing then the model becomes a Seasonal Autoregressive Integrated Moving Average model (SARIMA(p,d,q)(P,D,Q)) given by:

$$\varphi_p(B^d)\Phi_P(B^D)w_t = c + \theta_q(B^d)\Theta_Q(B^D)e_t \tag{6.13}$$

where,

$$\Phi_{\mathrm{P}}(B^D) = 1 - \Phi_1 \mathrm{B} - \cdots - \Phi_p B^D \tag{6.14}$$

and

$$\Theta_{\mathrm{Q}}(B^D) = 1 - \Theta_1 \mathrm{B} - \cdots - \Theta_Q B^D \tag{6.15}$$

where Φ_{P} are the seasonal autoregressive polynomials and Θ_{Q} are the moving average polynomials of order *P* and *Q* respectively.

6.3.3 SARIMA Model Selection

To obtain the forecasts from the SARIMA model the Box-Jenkins methodology set out in [83] is adopted. The Box-Jenkins method uses a three-stage approach to select an appropriate model for the purpose of modelling and forecasting a time-series. The three steps are 1) model identification 2) the estimation of the parameters and 3) diagnostic checks. However, given the aim of modelling the behaviour of over ten thousand cells plus several thousand more conglomerations of cells at different levels of spatial aggregations an automated approach is required. Once a time series is found to be stationary (see section 3.4), or made stationary through differencing, the usual Box-Jenkins approach involves a manual examination of a time series' Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF). The manner in which these

decay or fall below statistical significance provides the modeller with information about the need for and order of the autoregressive and moving average components in the model. However due to the large number of models required the modelling process shown in Figure 6.1 is employed.

First, load a time series comprising three months of hourly traffic loads either for a single cell or a group of cells summed together in larger spatial aggregations as discussed in Chapter 5. Next the Augmented Dickey Fuller (ADF) test is performed to check if the time series is stationary [97]. If the time series is found to be non-stationary successive differencing is applied until the series is found to be stationary or the differencing list is exhausted. Next iterate through a predefined list of possible candidate models and apply them to two months of data (the training set). These models were compiled by manually performing the usual Box-Jenkins model selection approach on a smaller subset of cells. The Akaike Information Criterion (AIC) is used to select a model that strikes a balance between accuracy and parsimony [98]. The selected model is used to generate the forecasts and compute the error metric on a separate month of testing data for each cell/conglomeration of cells.

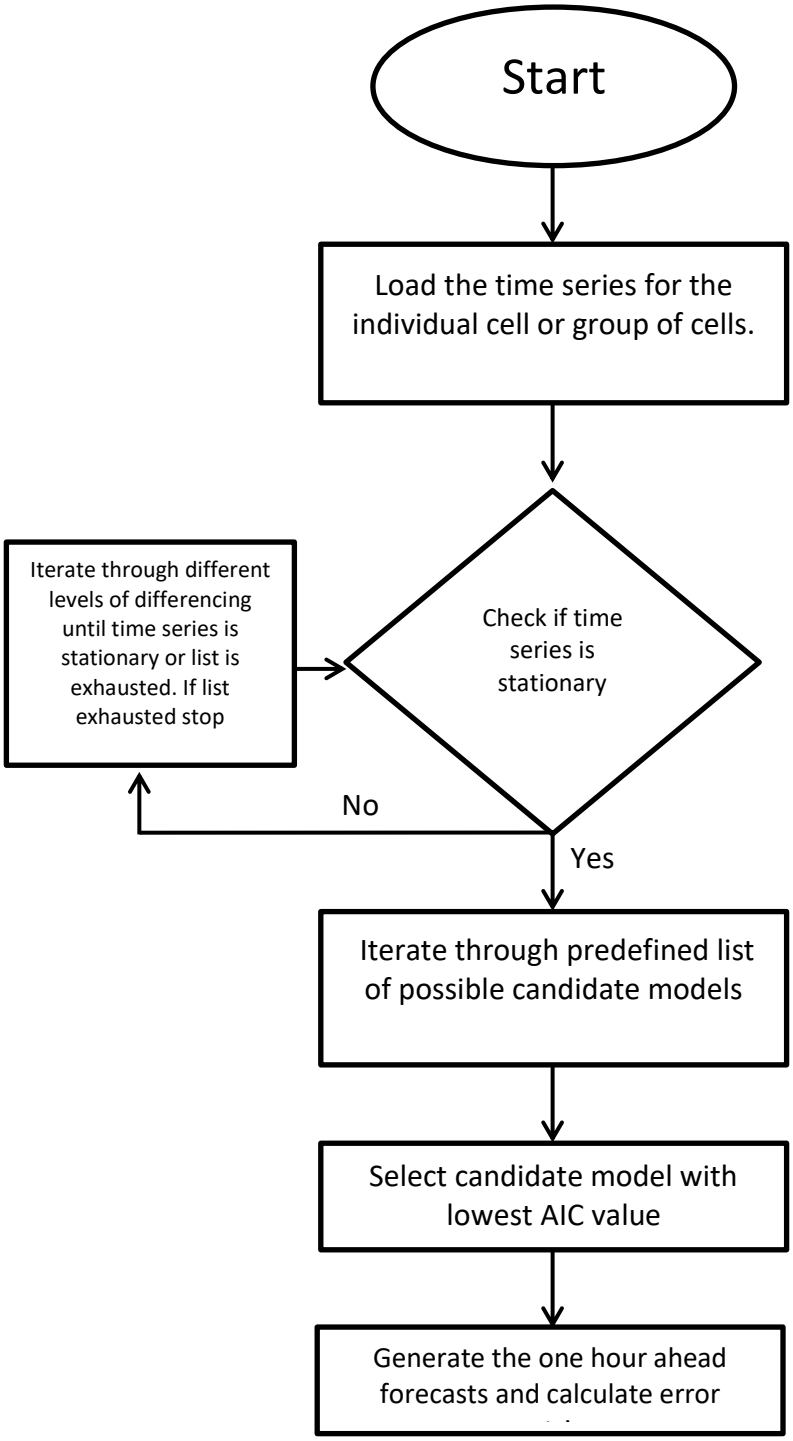


Figure 6.1: Automated Modelling Process SARIMA

6.3.4 ANN and SANN Models

Artificial Neural Networks (ANN) are a class of flexible computing frameworks which attempt to mimic the working of the brain to solve a broad range of nonlinear problems. ANN attempt to identify relationships and patterns in the input data, learn from experience and then provide generalised results based on what they have learned. ANNs have been applied to a wide range of problems and have the following features which make them suited to this forecasting application:

1. ANNs are data-driven and self-adaptive [91, 94]. There is no need to specify a particular model or to make *a priori* assumptions about the underlying statistical distribution of a dataset; the model is adaptively formed based on the features present in the data. This is especially useful in scenarios where there is no/poor theoretical understanding of the data generation process. This is also practically useful where there are too many models required to manually check the structure of each dataset as in this works application.
2. ANNs are non-linear which allows them to be more complex and accurate when modelling complex systems when compared to traditional linear approaches such as ARIMA/SARIMA models [91, 94, 99].
3. ANNs are universal function approximators [100] meaning they can approximate any continuous function to any desired level of accuracy [91, 100]. ANNs can also deal with situations where the input data is incomplete, erroneous or fuzzy [94].

The most widely used ANN for time series forecasting is the Multilayer Perceptron (MLP) [91, 94] which has a feedforward architecture of an input layer, one or more hidden layers and finally an output layer. Each layer contains a number of nodes which

are connected to those in the following layer by acyclical links (i.e. there are no cyclical paths present) [101].

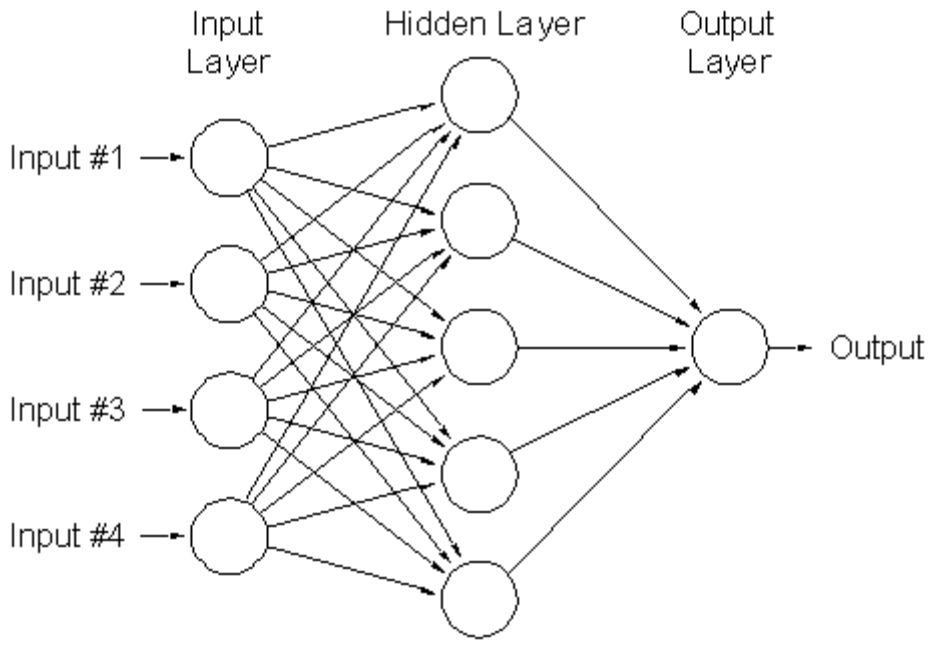


Figure 6.2: Architecture of a MLP ANN model with four inputs, one hidden layer and a single output.

Figure 6.2 shows an example of a MLP ANN with four inputs and one hidden layer. In a MLP with *p* inputs, *m* hidden nodes, and a single output the relationship between the inputs y_{t-i} where *(i=1,2,..,p)* and the output y_t is given by the formula [102]:

$$y_t = \alpha_0 + \sum_{j=1}^{m} \alpha_j g\left(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} y_{t-i}\right) + \varepsilon_t, \qquad \forall t \tag{6.16}$$

where α_j (j = 0,1,2,…,m) and β_{ij} (i = 0,1,2,…, p; j = 0,1,2,…,m) are the connection weights and ε_t is the random shock; α_0 and β_0 are the bias terms if present. The logical sigmoid function is commonly used as the nonlinear activation function where:

$$g(x) = \frac{1}{1 + e^{-x}} \tag{6.17}$$

The MLP model in (6.16) performs a non-linear functional mapping from the past observations of the time series to the future values , i.e. $y_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-p}, \mathbf{w}) + \varepsilon_t$, where **w** is a vector of all parameters and f is a function which is determined by a combination of the network structure and connection weights [91, 99]. The connection weights and parameters such as bias are denoted $\Psi = (\mathbf{w}, \alpha_0 , \beta_0)$ and estimated via non-linear least squares based on the minimisation of the error function [103]:

$$F(\Psi) = \sum_t e_t^2 = \sum_t (y_t - \hat{y}_t)^2 \tag{6.18}$$

The optimisation techniques used to minimise the error function in (6.18) are known as learning rules. The most commonly used learning-rule is error backpropagation [104] or back error propagation which is also known as the generalised delta rule. The MLP given by (6.16) is commonly known as a (*p,m,*1) ANN model which performs one-step ahead forecasting. Similarly, a (*p,m,r*) ANN model can be used for *r*-step ahead forecasting where *r* is the number of output nodes.

Since their inception ANNs have been used numerous times for various applications of seasonal time series forecasting. However, some people question their efficiency when compared to more traditional methods [96, 105, 106]. More recently [107] has proposed the Seasonal ANN (SANN) which is shown to be particularly effective when there is strong seasonality in the data. SANN does not require any pre-processing of the raw data; SANN can learn the seasonal pattern in a seasonal time series without removing it (SARIMA models for instance first apply seasonal differencing before modelling begins). In SANN models, the number of input and output neurons is represented by the seasonal period *s* of the time series as shown in Figure 6.3. Thus, the SANN model is similar to a (*s, m, s*) ANN where *m* is the number of hidden nodes.

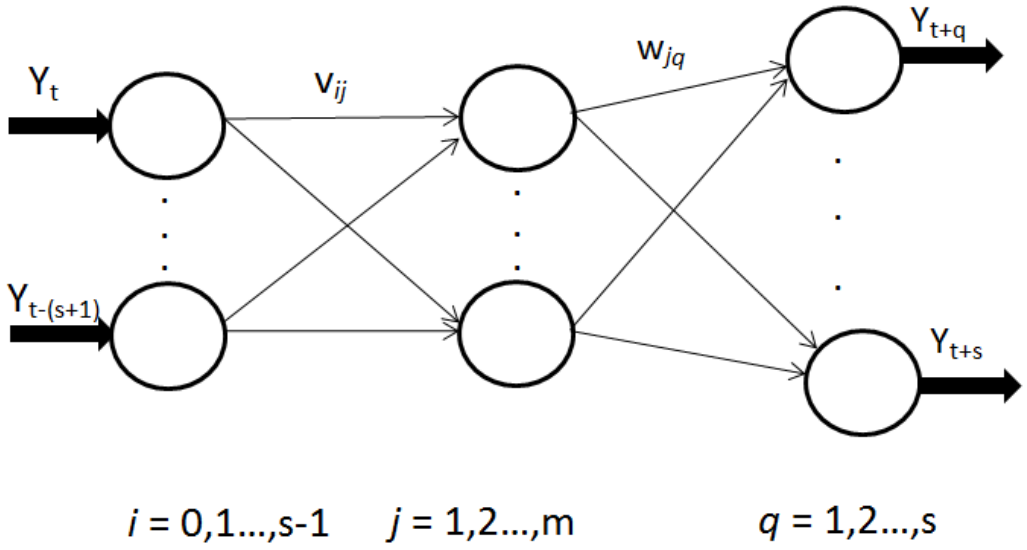


Figure 6.3: SANN Configuration for seasonal time series

The output of the SANN model can be expressed as [107]:

$$Y_{t+q} = \alpha_q + \sum_{j=1}^{m} w_{jq} g\left(\theta_j + \sum_{i=0}^{s-1} v_{ij} Y_{t-i}\right) \quad \forall t; q = 1,2,..,s \tag{6.19}$$

where Y_{t+q} *(q = 1, 2, …, s)* are the predictions for the future *s* periods and Y_{t-i} (*i = 0, 1, 2, …, s-1)* are the observations for the previous *s* periods; *m* is the number of hidden nodes; v_{ij}*(i = 0, 1, 2, …, s-1; j = 1, 2, …, m)* are the weights of the connections from input nodes to hidden nodes and w_{jq}*(j = 1, 2, …, m; l = 1, 2, …, s)* are the weights of the connections from the hidden nodes to the output nodes. Additionally, α_q (*q* = 1,2,…,*s*) and ϑ_j (*j* = 1,2,…, *m)* are the weights of the bias connections and *g* is the activation function. Thus, while forecasting, the number of input neurons should be 24 for diurnal data, 12 for monthly, 4 for quarterly etc. The number of hidden nodes to be used can be selected by performing suitable experiments on the training data set.

6.3.5 SANN Model Selection

As discussed in [107] when using a (*s,m,s*) SANN model the number of both input nodes is specified in terms of the seasonal period of the data *s*. Thus, only the appropriate

number of hidden nodes *m* needs to be determined. For this purpose, the AIC can be used in a similar fashion as in 6.3.3 to select a model structure that performs well while encouraging parsimony. For the set of seasonal cell load time series with approximate periods *s*=24, the maximum number of hidden nodes is equal to *s/*4.

The dataset employed (Chapter 2) comprises four months of time series data representing the hourly traffic loads either for a single cell or a group of cells summed together in larger spatial aggregations as discussed in Chapter 5. The models are trained on a training set denoted $\mathbf{Y}_{train}$ comprising two months of data (the same two months used as training data in 6.3.3). To avoid over-fitting, a separate month is used as validation denoted $\mathbf{Y}_{val.}$ Finally, the fourth month of data is used as a test set and is denoted $\mathbf{Y}_{test}$.

For each member of the set of time series, the (*s,m,s*) SANN model is successively trained on the training set by varying the number of hidden nodes *m* from 1 up to the maximum number of hidden nodes. The number of hidden nodes *m* for each member of the set of time series is then set to whatever value of *m* gives the minimal AIC. The activation function employed is the popular logical sigmoid function as discussed in 0 and for the optimisation of weights the Levenberg-Marquardt (LM) algorithm [108] is used. As in 6.3.3 there are too many time series to model manually and thus the models must be trained, validated and tested automatically. The algorithm to achieve this is outlined in Figure 6.4.

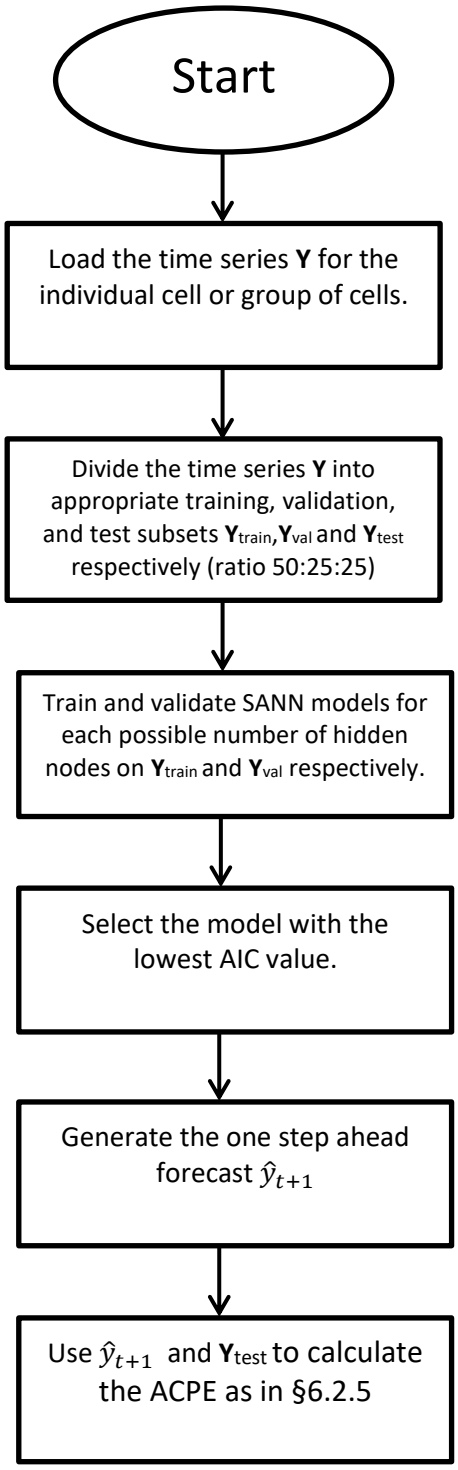


Figure 6.4: Automated Modelling Process SANN

6.3.6 Prediction Methods Conclusion

This section introduced and discussed two commonly used predictive models suitable for load localised near horizon load forecasting in cellular networks. The two methods employed were SARIMA models and SANN models discussed in 6.3.2 and 6.3.5 respectively. Due to the large number of models to be created an automated modelling approach was required for both methods. These automated approaches were introduced and discussed in 6.3.3 and 6.3.5.

6.4 Results

6.4.1 Results Introduction

Section 6.2 introduced a novel metric for evaluating the accuracy of predictive models in cellular networks. 6.3 introduced two different methods of cellular load prediction with automated modelling algorithms. This section will present the results obtained from utilising these methods on the spatial aggregations defined in Chapter 5. Subsection 6.4.2 begins the section by presenting some example results from a few representative cells/cell aggregations. 6.4.3 discusses the effect that the forecasting methods has on the perception of the results. Subsection 6.4.4 presents the complete network wide results for both methods and all the various levels of spatial aggregation introduced in Chapter 5. Subsection 6.4.5 concludes this section.

6.4.2 Example Results

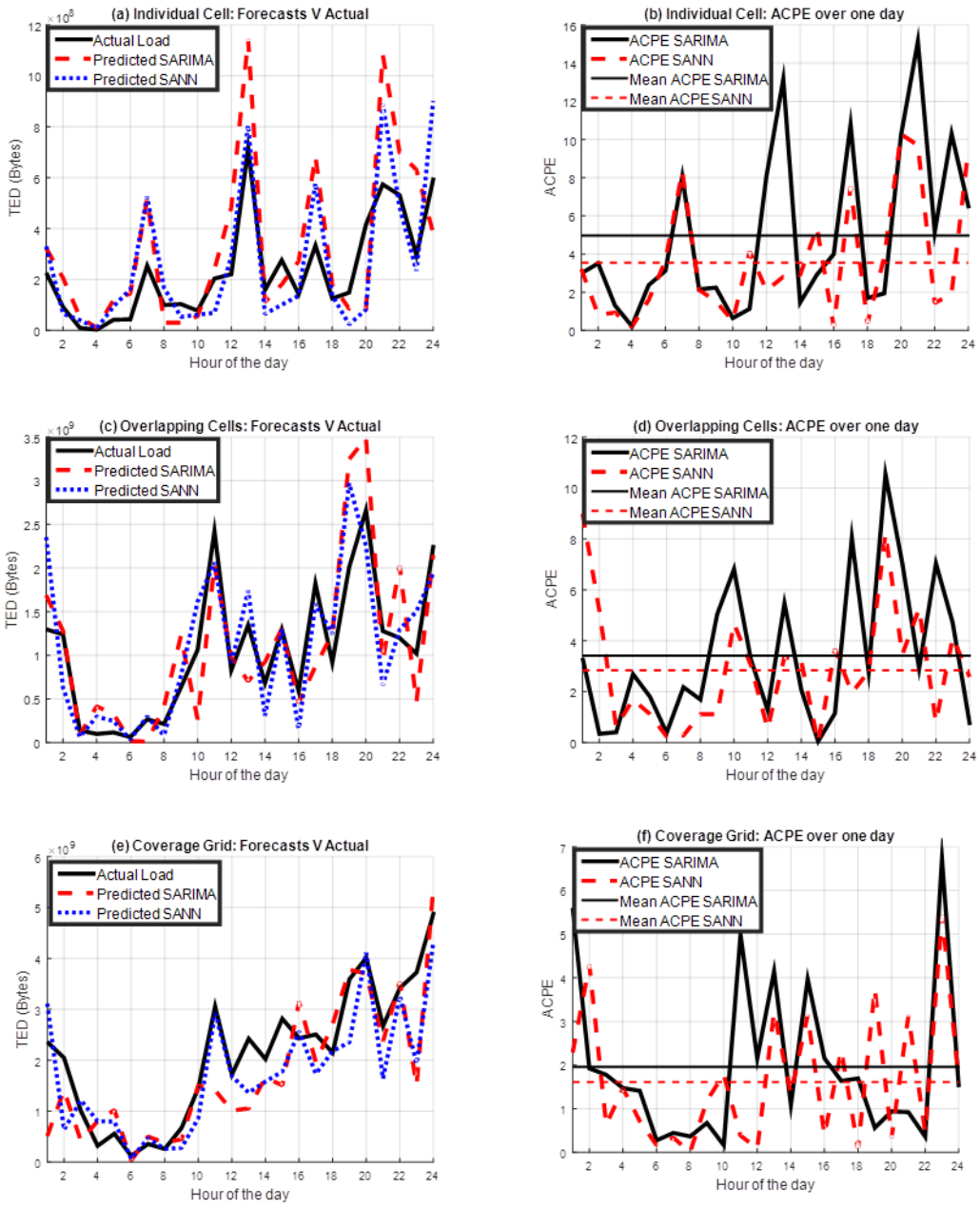


Figure 6.5: Example results for three different levels of spatial aggregation over one day. (a) Forecasted load V actual for individual cell. (b) ACPE for both forecasting methods for individual cell. (c) Forecasted load V actual for 3 overlapping cells. (d) ACPE for both forecasting methods for 3 overlapping cells. (e) Forecasted load V actual for a coverage grid of 9 cells. (f) ACPE for both forecasting methods for a coverage grid of 9 cells.

Figure 6.5 (a), (c) and (e) show an example of the forecasting results against the actual load (test data) for the three different levels of spatial aggregation outlined in Chapter 5. These show an important pattern that is repeated in other groups of cell aggregations: as the spatial aggregation level increases the hour to hour saw-like oscillations of load experienced in individual cells gradually gives way to a smoother load curve. With increasing spatial aggregation this load curve begins to more closely resemble the total daily aggregate load curve for the entire network as presented in Chapter 3. At lower levels of aggregation and particularly at hours of low load the presence of an individual heavy user can significantly affect the usage in an individual cell. At higher levels of aggregation these peaks and troughs in individual cells gradually start to even each other out. Thus, generally at higher levels of spatial aggregation smoother load curves are found (as demonstrated in Chapter 5 using entropy) which are more easily modelled accurately.

The corresponding set of figures (b), (d) and (f) show the performance of the two different forecasting methods as outlined in 6.3 against the ACPE metric introduced in 6.2.5. Generally, the SANN method performs slightly better than the SARIMA method over the three spatial aggregation levels achieving a lower daily mean ACPE in all three examples.

6.4.3 Effect of forecasting metric on result perception

Interestingly the performance metric employed can greatly influence the perception of the modelling outcome. For example, in Figure 6.5 (b), (d) and (f) the period of the day with the consistently smallest ACPE is the early morning hours. However, if the metric used to examine the accuracy of the model is changed a different impression can be given. For example, Figure 6.6 replots the SARIMA modelling ACPE results from Figure 6.5 (f) along with the more popular APE and MAPE metrics introduced in 6.2.4. When

using ACPE to take into consideration overall capacity the error is consistently lowest during the early morning hours approximately 2-6 AM. This is because the total load during those hours is much smaller than at other times in the day as shown in Figure 6.5 (e). Thus, even if the forecasted values are out by a large percentage of the actual value, in practical terms as a percentage of the total available capacity the forecast can still be very useful. For instance, at 5 AM the forecasted value has an APE of 150%. This compares poorly with the daily MAPE for Figure 6.6 of 40%. However, the actual forecast is only out by approximately 1.6% of the total grid of 9 cells capacity, below the daily average ACPE of 2%.

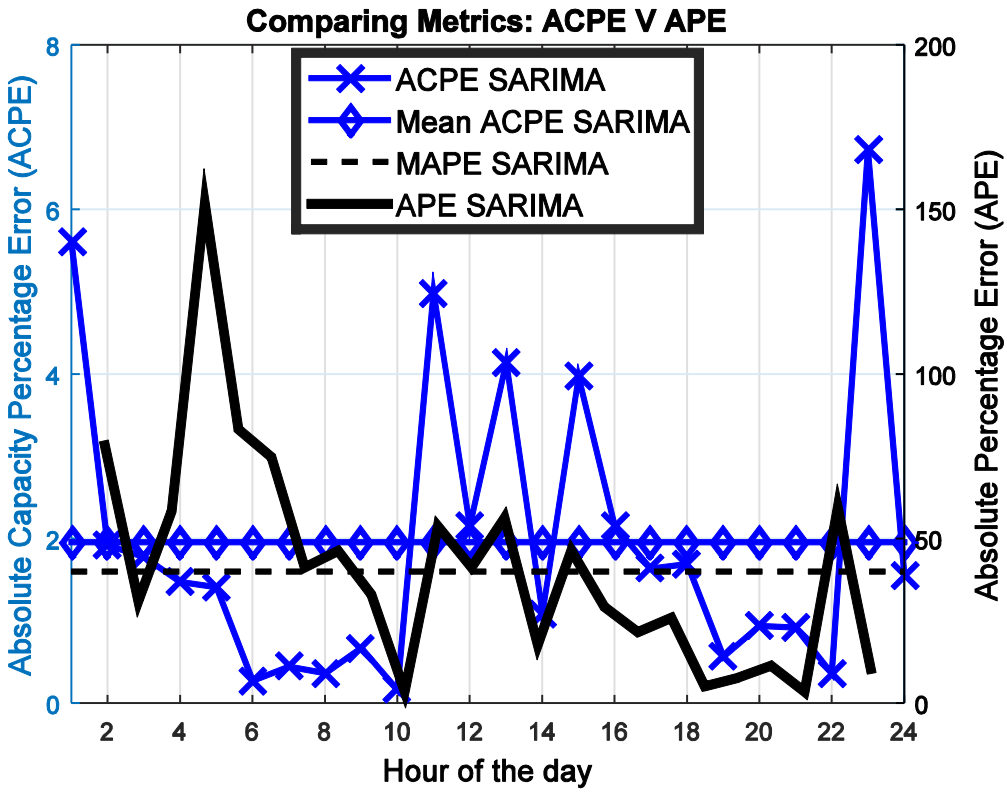


Figure 6.6: Comparing Metric: The ACPE and the mean ACPE of a SARIMA model are plotted on the left vertical axis. The APE and the MAPE are plotted on the right vertical.

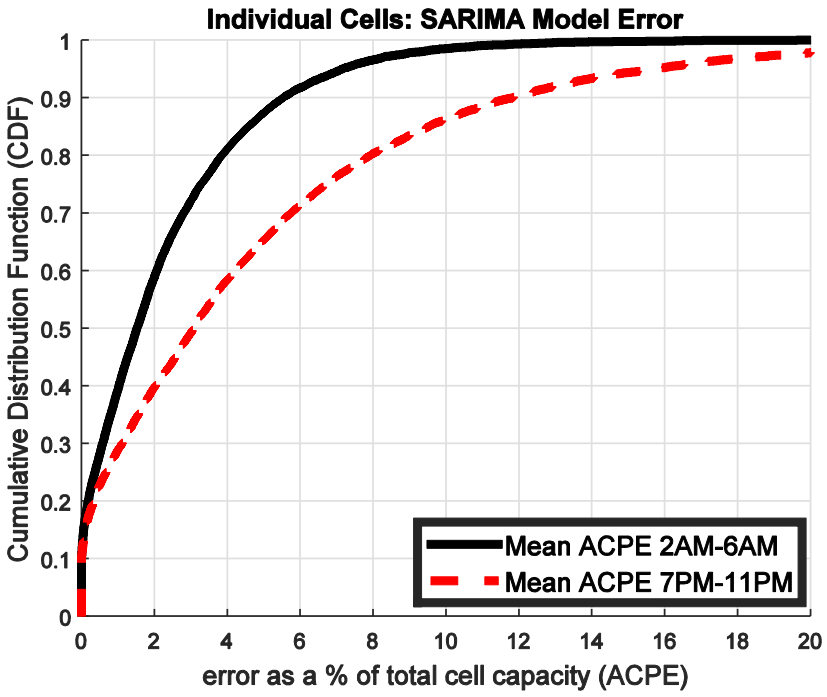


Figure 6.7: CDF of the mean ACPE for the hours 2 AM – 6 AM (inclusive) and 7 PM - 11 PM (inclusive) for all cells over one month of test data for the SARIMA model.

The minimisation of the error during the early morning hours is illustrated in Figure 6.7. Figure 6.7 presents the difference in mean ACPE between the early morning hours and late night hours for all cells on the network over the test month. The difference suggests that generally the easiest time to make predictions for cell load is during the early morning hours. This presents many opportunities to better use resources with one such application discussed in the following chapter.

6.4.4 Network wide results

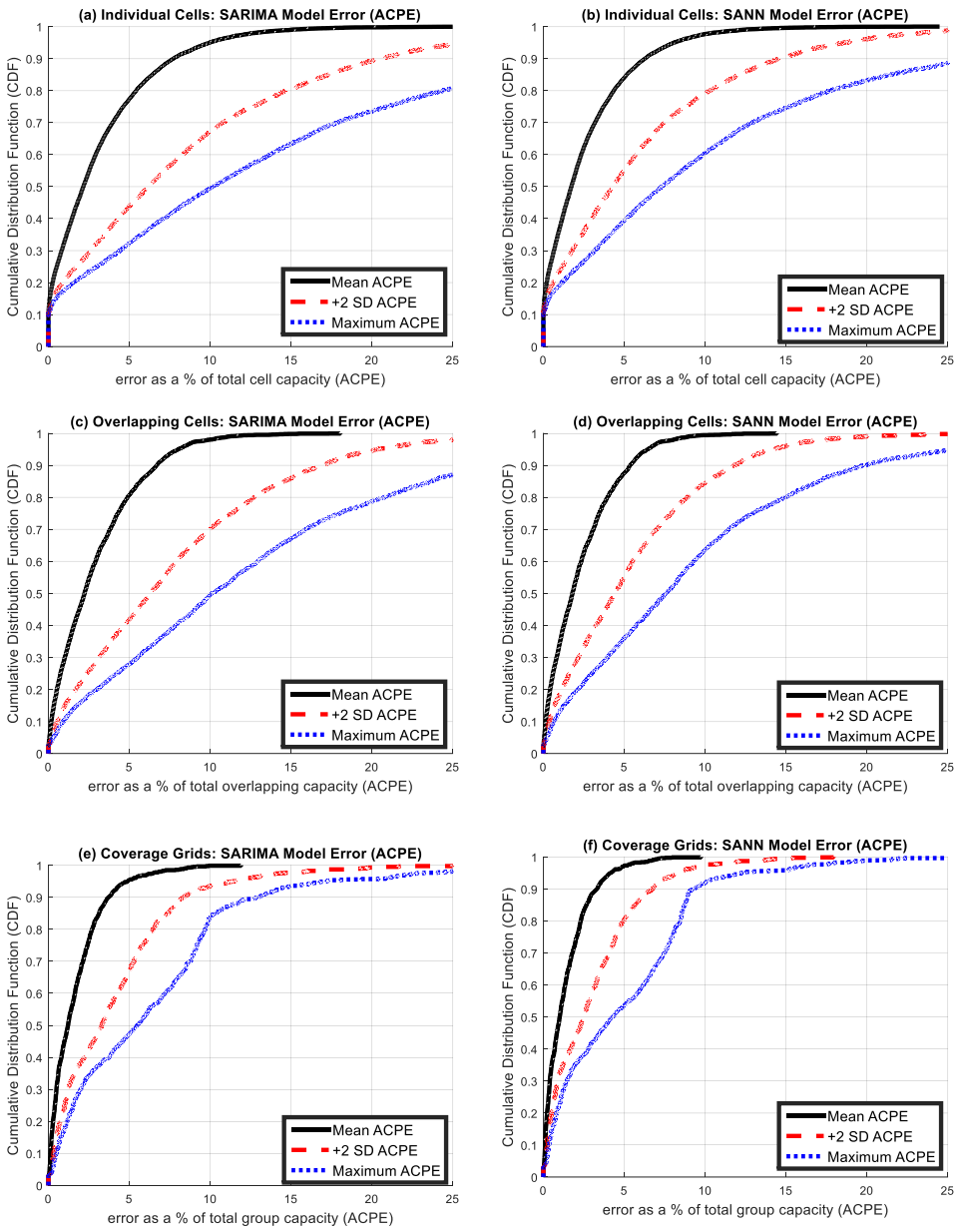


Figure 6.8: CDF of % ACPE for SARIMA models for individual cells. (b) CDF of % ACPE for SANN models for individual cells (c) CDF of % ACPE for SARIMA models for overlapping cells. (d) CDF of % ACPE for SANN models for overlapping cells (e) CDF of % ACPE for SARIMA

Figure 6.8 shows the CDF for the ACPE obtained after carrying out the forecasting process outlined in 6.3.3 and 6.3.5 for the three different aggregation levels outlined in Chapter 5. Comparing Figure 6.8 with the results presented in Chapter 5 it is immediately obvious that across all aggregation levels, both model types improve the predictability when compared with the between hour and between day variations (these can be thought of as simple AR processes where $y_t = y_{t-1}$ for between hour variation and $y_t = y_{t-24}$ for between day variations). Table 6.1 shows the ACPE of the four models for the median cell while Table 6.2 shows the ACPE for the cell at the 90th percentile of ACPE. Both tables again show that the SANN and SARIMA models perform best across all aggregation levels. Table 6.1 shows that 95% of the time the median individual cell has an ACPE below 4% and, promisingly, 95% of the time the median coverage grid has an ACPE of less than 3%. This means that for 50% of the coverage grids, their load can be predicted to within 3% of their maximum capacity 95% of the time.

Generally, for all the models the ACPE decreases as the aggregation level improves. For example, Table 6.2 shows that for the SANN models, 90% of individual cells have a mean ACPE of less than 6%, 90% of overlapping cells have a mean ACPE of less than 5.5% and 90% of coverage grids have a mean ACPE of less than 3.2%. Table 6.2 again, promisingly, shows that 95% of the time, 90% of individual cells have an ACPE below 15% and 95% of the time median coverage grid has an ACPE of less than 7%. This means that for 90% of the coverage grids, their load can be predicted to within 7% of their maximum capacity 95% of the time.

Generally, the SANN models perform slightly better than the SARIMA models. This is probably due to the SANN models being more suited to automatic model creation as outlined in [107].

Table 6.1: Model ACPE across aggregation levels for the median cell

Forecasting Method	Aggregation Level					
	Individual Cells		Overlapping Cells		Coverage Grid	
	Mean of median cell	+2σ	Mean of median cell	+2σ	Mean of median cell	+2σ
Inter Hour	4	9	3	8	2.25	5
Inter Day	4	10	3	8	3	4
SARIMA	2.5	6.5	2	6	1	3.5
SANN	2	4	2	4	1	3

Table 6.2: Model ACPE across aggregation levels for the 90th percentile cell

Forecasting Method	Aggregation Level					
	Individual Cells		Overlapping Cells		Coverage Grid	
	Mean of 90th percentile cell	+2σ	Mean of 90th percentile cell	+2σ	Mean of 90th percentile cell	+2σ
Inter Hour	9	25	7	20	4	10
Inter Day	9	30	8	24	5	13
SARIMA	7.5	20	6.5	16	3.5	7
SANN	6	15	5.5	13	3.2	6.5

6.4.5 Results Conclusion

This section presented the results obtained from utilising the prediction methods outlined in section 6.3 on the spatial aggregations defined in section 5.3. Subsection 6.4.2 began the section by presenting some example results from a few representative cells/cell aggregations. 6.4.3 discussed how the metric used to evaluate the predictive models can influence perception of the outcome. It was found that the novel ACPE metric introduced in 5.2.6 produced the most intuitive results. Using it, the early morning hours were identified as the hours most suited to modelling which other popular metrics, discussed in 5.2.6 failed to do. Subsection 6.4.4 presented the complete network wide results obtained for both methods for all the various levels of spatial aggregation. Generally, it was found that the SANN models performed better

across all aggregation levels. As expected from Chapter 5, generally the error reduced as the spatial aggregation size increased.

6.5 Discussion and Conclusion

Predicting traffic in cellular networks is becoming increasingly important as the explosion in demand for radio access coupled with falling or stagnating ARPU drives new research into SONs. The previous chapter, Chapter 5, explored the predictability of three different levels of spatial aggregations. This chapter went further by creating predictive models for the spatial aggregation regions defined in section 5.3. Firstly, section 6.2 discussed the three main categories of predictive model metrics used in the literature. Particular attention was given to their advantages and disadvantages with regards to the practical application of local cellular network load forecasting. Due to the problems identified with these metrics, a novel metric ACPE was defined. Section 6.3 introduced and discussed two possible predictive methods suitable for localised near horizon load forecasting in cellular networks. The two methods employed were SARIMA models and SANN models discussed in 6.3.2 and 6.3.4 respectively. Due to the large number of models to be created an automated modelling approach was required for both methods. These automated approaches were introduced and discussed in 6.3.3 and 6.3.5. 6.4 presented the results obtained from utilising the prediction methods outlined in 6.3 on the spatial aggregations defined in Chapter 5. Subsection 6.4.2 began the section by presenting some example results from a few representative cells/cell aggregations. 6.4.3 discussed how the metric used to evaluate the predictive models can influence perception of the outcome. It was found that the novel ACPE metric introduced in 6.2 produced the most intuitive results. Using it, the early morning hours were identified as the hours most suited to modelling, which other popular metrics discussed in 6.2 failed to do. Subsection 6.4.4 presented the complete network wide

results obtained for both methods and all the various levels of spatial aggregation. Generally, it was found that the SANN model performed better across all aggregation levels. As expected from Chapter 5, the error reduced as the spatial aggregation size increased. This indicates that to improve the performance of predictive models of network load, the largest practicable cellular aggregation should be used as the basis for the predictions. Table 6.2 showed, promisingly, that 95% of the time, 90% of individual cells have an ACPE below 15% and 95% of the time 90% of coverage grids have an ACPE of less than 7%. Thus, meaning that for 90% of the coverage grids, their load can be predicted to within 7% of their maximum capacity 95% of the time.

Direct comparisons of these results with other works is challenging for several reasons 1) the paucity of comparable work 2) the ambiguity in the scales of the datasets used 3) the ambiguity in spatial & temporal aggregations 4) the use of unsuitable metrics as discussed in 6.2, and a general lack of visibility of actual results. For example, taking the first reason outlined above, much of the work carried out on cellular network load prediction is focused on older voice-centric networks and datasets [12, 21-23]. Although voice is still an important function of cellular networks, it is an ever decreasing component of overall cellular load as discussed in Chapter 3. This work, in keeping with Long Term Evolution (LTE) standards, has aggregated voice, data, and SMS together to give one total figure for load Total Equivalent Data (TED) as described in Chapter 3. Thus, although voice load is correlated with total network load (Chapter 3), it behaves differently and is more predictable than cellular data as discussed in Chapter 5. Addressing point 2 and 3 above, works such as [27] show that forecasting short term load on the macro network scale is possible with a high degree of accuracy. However, as discussed in Chapter 5, this is of limited practical value for many applications such as green networks [28] and spectrum sharing [12]. For such applications, groupings with finer spatial resolution are required which motivated this work focusing on small

spatially contiguous portions of the network. Other works such as [26] have access to both voice and cellular data but unfortunately only provide predictive results for the voice portion. [26] cites the greater variance in cellular data load when compared with the voice traffic as a reason for not producing modelling results. Perhaps the APE (see 6.2.4) metric used in [26] discouraged the authors from further work. While the APE can be very high for cellular data compared to voice given cellular data's greater variance, if the variance is normalised by the cell's actual capacity (as in 6.2.5) it appears to be a much more manageable problem.

Although there is a lack of suitable external results with which to judge the effectiveness of the predictive models, they can be compared with the results obtained in 5.3.5 which are summarised as the inter hour/day rows in Table 6.1 and Table 6.2. These can be thought of as simple forecasts, for example an inter hour value of 5 ACPE means that using hour *h*'s load as the predicted load for hour *h+1,* results in an ACPE of 5. As previously discussed in Chapter 5, the inter hour load routinely gives a slightly better prediction than the inter day load and thus is used for further comparisons. With the inter hour load as a benchmark, both models improve the accuracy of the predictions over the simplistic model. For example, Table 6.2 shows that 90% of coverage grids, 95% of the time have an ACPE of less than 10% when using the simplistic inter hour model. The use of the predictive models outlined in 6.3 reduce the ACPE by 30% for the SARIMA models, and 35% for the SANN models. As shown in Table 6.1 and Table 6.2 these improvements over the simplistic inter hour case are present across all levels of spatial aggregations. Thus, the models introduced here do improve on simple self-similarity methods used in works such as [16]. Generally, the SANN models perform slightly better than the SARIMA models. This is probably due to the SANN models being more suited to automatic model creation as outlined in section 6.3.

This chapter has demonstrated that automatic, localised near horizon load forecasting is feasible, particularly at higher levels of spatial aggregation. The next chapter will use the models generated here to demonstrate the real and practical possibilities presented by localised near horizon predictive models of cellular load.

Chapter 7 Utilising Predictive Models for the Minimisation of Power Usage in Radio Access Networks

7.1 Introduction

Accompanying the growth of cellular usage discussed in Chapter 3 there has also been a large increase in the energy used by cellular networks [99]. It is estimated that cellular networks account for approximately 10% of the total carbon emitted by the Information and Communication Technology (ICT) sector with this expected to increase further in the future [101]. In addition to the environmental concerns there are real economic benefits for network operators to minimise power consumption [102].

It is currently estimated that 80% of the total infrastructure power consumption takes places in the Radio Access Network (RAN), particularly Base Stations (BSs) [109]. Despite significant temporal and spatial variations in demand [25, 72, 78], networks are currently optimised for peak throughput at peak demand. As shown in [93], large underutilisation of RAN resources are present and particularly pronounced at the BS level. Unfortunately, the infrastructure of currently deployed networks is largely load invariant, meaning largely underutilised BSs stay active despite a lack of demand. This is a costly inefficiency in terms of power consumption but it also underutilises valuable licensed spectrum which could be made available for secondary usage [83].

Accurate short and medium term predictive models of load (primary usage) at the local level (cell, BS, coverage grid etc.) are critical if Self-Organising Networks (SON) are to ameliorate the network's inefficient usage of power and spectrum. For example, if it can

be predicted that traffic in a particular BS or group of BSs falls below a certain threshold at certain times then SON algorithms can use this information to alter the network to save energy [28, 30, 31]. Also, if low demand by primary users of valuable licensed spectrum can be predicted in certain cells/areas at, for example, off-peak times this can provide opportunities for secondary usage in these frequency bands [12].

Much work has gone into algorithms and techniques to dynamically switch on/off cells or BSs [28, 30, 31]. However, most work in the area simply uses historical static load profiles or assumes that switching decisions can be made instantaneously. However, real world measurement results such as those presented in [16] show that switching can take up to 30 minutes due to the heating systems. Thus, predictions of the need to perform a switch ahead of time are important.

The previous chapters introduced novel predictive models of load in spatially contiguous aggregations of BSs. Chapter 5 demonstrated that as the spatial aggregations increase the predictability of the load increases. However, for many advanced network management techniques such as green networks and spectrum sharing, as the spatial aggregation increases the usefulness of the predictions decreases. Thus, there is a fundamental tradeoff to be made between the accuracy of the prediction and its usefulness. The point at which the tradeoff is made is application dependent and is a crucial step in designing a worthwhile solution. For example, the provisioning of new backhaul infrastructure may require network wide predictions of load several years in advance. However, in the green networking application under investigation here the maximum useful spatial aggregation is the coverage grid as introduced in the previous chapters. This is due to the requirement of mutual redundancy, that any one member of the grid can service another member's load if it is powered off. If this fundamental

requirement is not satisfied then the predictions are not useable for their intended purpose.

The primary contributions of this chapter are:

- A novel and practical energy savings scheme tested on real world data across multiple regions.
- A validation of the usefulness of near horizon localised predictive models of cellular load in an advanced management technique.
- A large scale examination and identification of the underutilisation present in an Irish network and how it varies by time of day, and crucially, by region.

7.2 introduces four regions representing different examples of areas typically found in a cellular network. The section then goes on to examine the underutilisation present in the network both at the network, regional, and BS level. 7.3 introduces a novel energy saving scheme utilising the previously created coverage grids and predictive models. 7.4 presents the results of testing the energy savings scheme on real world traffic data. Finally, 7.5 provides a concluding discussion for this chapter.

7.2 Network Underutilisation

7.2.1 Network Underutilisation Introduction

As discussed in Chapter 3 the cellular network under investigation suffers from the classic peaking problem of resource allocation. In effect, this means that the network is provisioned to deal with peak loads during late evening/early nighttime hours. Thus, for large portions of the day the network is largely overprovisioned for the load experienced. This problem, as identified in Chapter 3 is particularly acute during the early morning hours. Also, as discussed in Chapter 3 and Chapter 4 there is a great

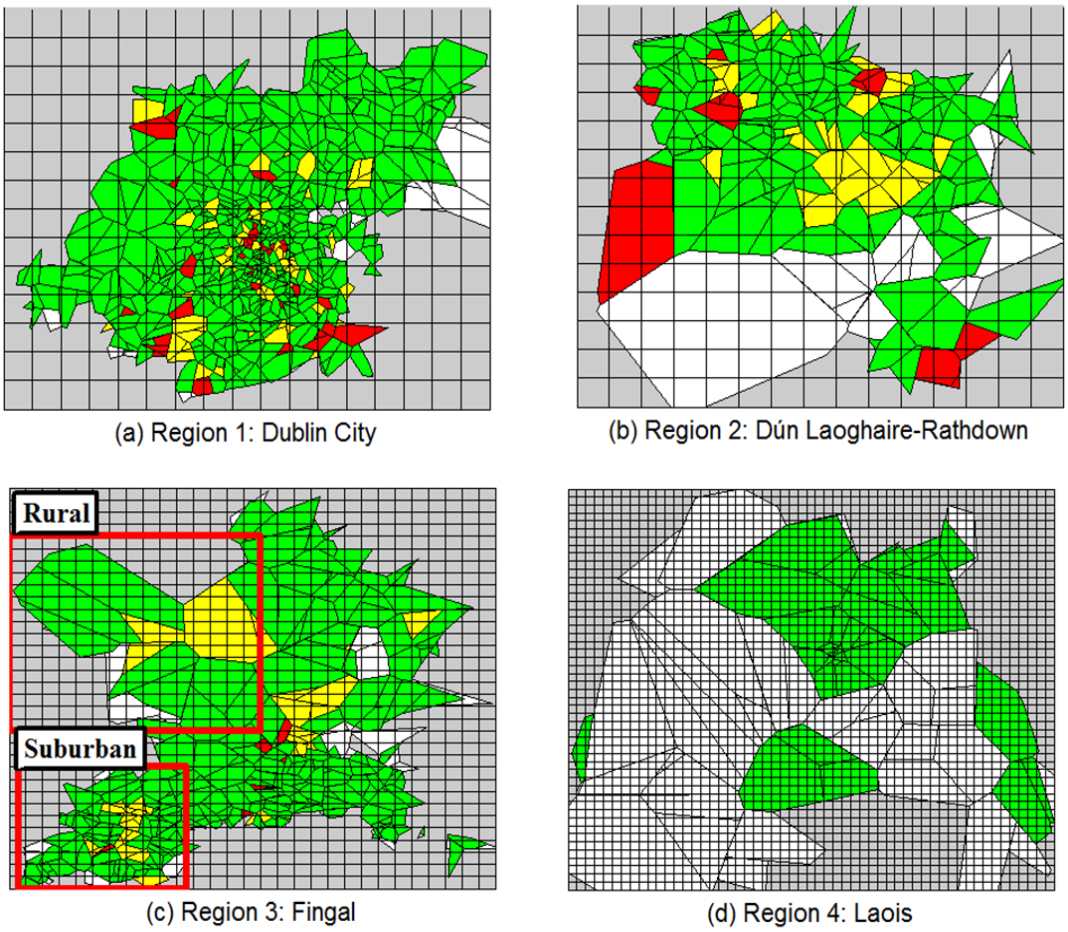
spatial disparity in network usage which also lends itself to large scale underutilisation. This section presents network measurements to quantify and qualify the diversity of traffic load in both time and space across the network. Key features relevant to the reduction in power usage on the network are identified with both the challenges and opportunities they present discussed. Data collected from four diverse regions representing many different typical area types are used in this chapter. 7.2.2 begins by presenting the four disparate regions selected which collectively represent the diversity of the Irish network's topography. Subsection 7.2.3 then examines the temporal diversity of the four regions and the extent of the peaking problem within each of the regions. 7.2.4 explores the extent of the underutilisation both at the regional and the local level.

7.2.2 Region Selection

The planning and organisation of cellular networks varies greatly by population density, topography, etc. Thus, it is more instructive to examine sub networks within the whole that are representative of particular planning features such as population density, etc. To this end, the four regions selected are outlined in Table 7.1 and illustrated in Figure 7.1. Note: unless otherwise stated all city/county boundary information is taken from [72] while demographic information is taken from the 2011 Irish census [74].

Figure 7.1 shows the approximate cell coverage areas of the four different regions. Note, that for simplicity and due to lack of available data this work restricts itself to the examination of only the 3G network. Region 1 consists of Dublin city which is the most densely populated region with a population density of 4471 people/km^2. Apart from having a large residential population it also has a broad mix of commercial/industrial and cultural sights which result in a large inflow of daily commuters. Region 2 is the administrative county of Dún Laoghaire-Rathdown which is a mainly suburban area to

the south east of region 1 with a population density of 1624/km^2. Region 3 is the administrative county of Fingal which is a suburban and semi-rural area on the northern border of region 1 with a population density of 598/km^2. Region 4 is the landlocked county of Laois which is a mainly rural county in the midlands of Ireland with a population density of just 46.8/km^2, below the national average of 65/km^2 [74].



(a) Region 1: Dublin City

(b) Region 2: Dún Laoghaire-Rathdown

(c) Region 3: Fingal

(d) Region 4: Laois

Figure 7.1: Cell coverage zones in the four regions. Each square corresponds to 1km^2. White coverage zones have one cell covering that area, green have two, yellow have three and red have four or more cells covering that zone. Region 3 is further subdivided into a suburban area around Blanchardstown and a rural area to the north west of the county.

Table 7.1: Information on the four regions under investigation

	Region 1	Region 2	Region 3	Region 4
Area (km^2)	118	127	458	1720
Number of BS	525	192	116	59
Population	527612	206261	273991	80559
Pop/km^2	4471	1624	598	46.8
Classification	Urban	Suburban	Suburban - Semi Rural	Rural

7.2.3 Temporal Diversity

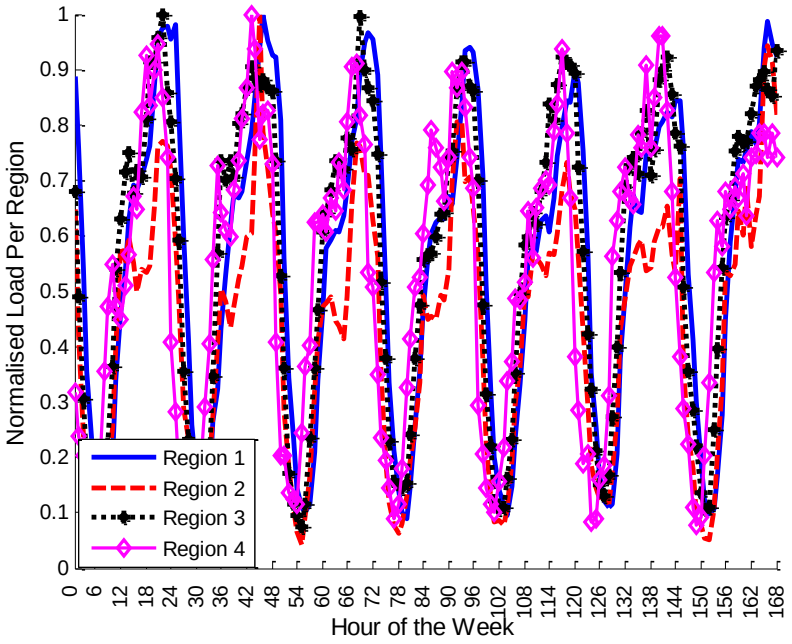


Figure 7.2: One week of total traffic in each region starting at 00:00 on Monday running to 23:59 on Sunday

Figure 7.2 plots the normalised aggregate traffic load for each of the four regions. Similarly to the network wide results presented in 3.2, a strong diurnal pattern is evident in each of the four regions with a large gap between peaks and troughs. Interestingly, different regions tend to exhibit somewhat different patterns. For example, the peaks and troughs in region 1 (city) seem to occur at different times to

those observed in the rural region (4) etc. Also, the rural region for example tends to deviate from the urban and suburban regions to the greatest extent on Sundays.

To quantify the diurnal temporal traffic variation the mean ratio of the maximum to minimum traffic load is computed for each BS in the four regions. For each BS the aggregate load is calculated for every hour of the day resulting in 24 hourly loads per day. The maximum load of a BS is defined as the load on the BS during the hour h_{max} when load was highest (between 6 PM and 1 AM in about 70% of BSs). The minimum load of a BS is defined as the load on the BS during the hour h_{min} when load was smallest (between 1 AM and 7 AM in over 90% of all BSs).

Figure 7.3 presents the Complementary Cumulative Distribution Function (CCDF) of the maximum to minimum traffic ratio for all BSs in the four regions. Interestingly the maximum to minimum ratio is greater than ten for more than 80% of BSs in all regions. This indicates that there is a high degree of temporal diversity in almost every BS. Such strong temporal diversity indicates large underutilisation of both network infrastructure and spectrum in the time domain. This network infrastructure inefficiency indicates that there are large savings to be made from a move towards networks where the power consumption is dependent on traffic load. Also, the inefficient use of spectrum shows the real possibility for large scale secondary usage of licensed spectrum with minimum impact on primary usage.

Unsurprisingly, the sparsely populated rural region 4 deviates from the more densely populated urban regions. On visual inspection, it appears that the four regions maximum to minimum load ratios are positively related to population density. This is further borne out by examining smaller sub regions such as those identified in region 3 as shown in Figure 7.1. This is interesting as it indicates that the wastage due to peak provisioning is greatest in densely populated urban environments. These densely

populated urban environments with higher maximum to minimum load ratios are where spectrum is most limited and also most valuable.

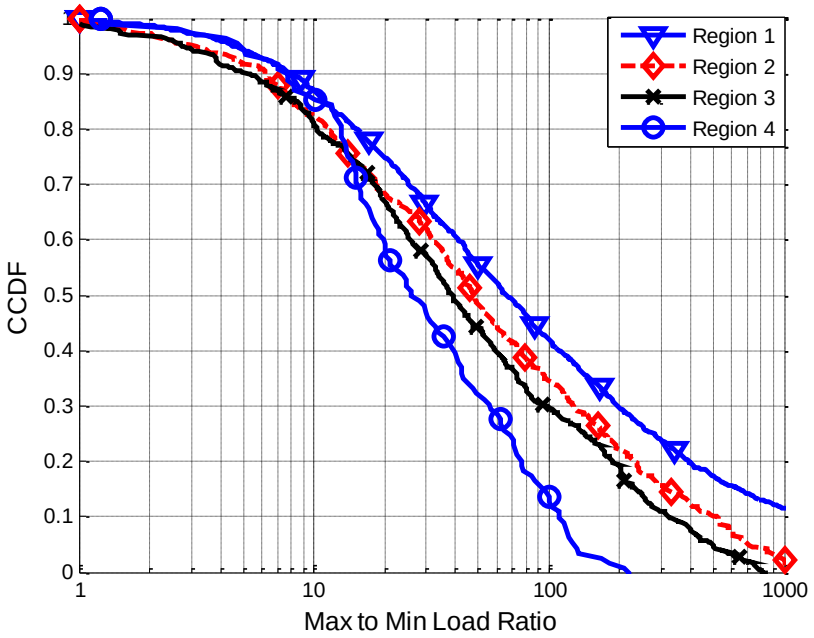


Figure 7.3: Maximum to Minimum Traffic Load Ratio

7.2.4 Regional and Local Underutilisation

The aggregate traffic load in a given region is found to be only a fraction of the aggregate traffic capacity in that region. To verify this, the percentage of the total regional capacity *X(h)* used during each hour *(h)* is calculated. *X(h)* is defined as the ratio of the aggregate traffic load during hour *h* in a region to the sum of the peak observed load in each BS in that region over the entire period (a lower bound estimate of the cells capacity):

$$X(h) = \frac{\sum_{i=1}^{n} L(i,h)}{\sum_{i=1}^{n} L(i,h_{max})} \tag{7.1}$$

where *n* is the number of BSs in a region, *L(i,h)* is the traffic load of BS *i* during hour *h* and $L(i,h_{max})$ is the largest load observed on BS *i* during the observation period. The

percentage of regional traffic being used during each hour of one week is presented in Figure 7.4. This shows that at no point during the week in any region is the aggregate traffic demand greater than 45% of the total regional capacity. Furthermore, in the urban/suburban regions where spectrum and space are most limited, the regional utilisation peaks at approximately 20-25% and drops into single digits for approximately the first 12 hours of each day.

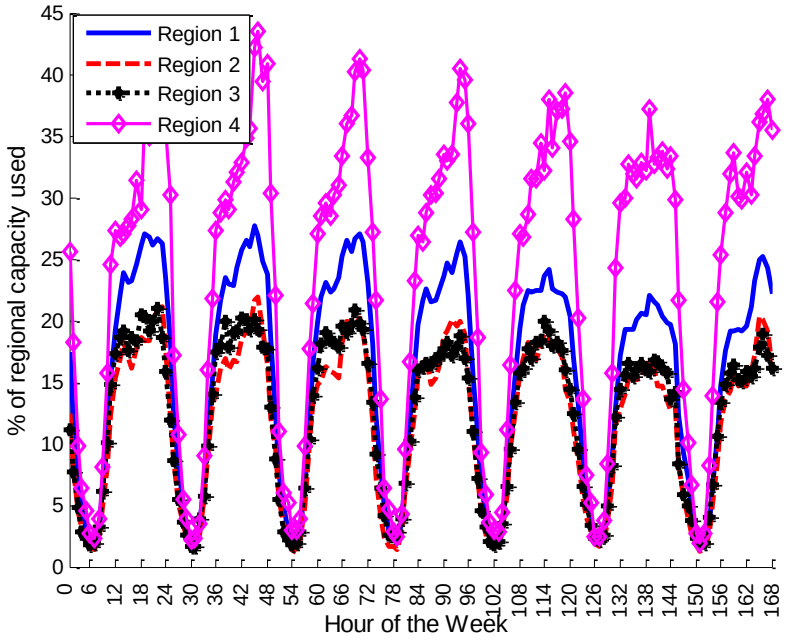


Figure 7.4: The percentage of total regional capacity being used over the course of one week.

One reason for the large underutilisation is that all the cells in a region do not peak simultaneously. The operator deploys infrastructure to service peak demand at each location even though this only lasts for a small period of the day. As this peak hour is location dependant the aggregate deployed capacity (the sum of all BS capacity in an area) is much greater than the actual traffic demand at any given time. The degree to which peak hour varies within a region influences the amount of underutilisation. For example, region 4 is a rural area where most BSs have similar profiles and peak at the

same time. In contrast regions 1-3 are a more complex mixture of residential/business/semi-rural areas with diverse profiles.

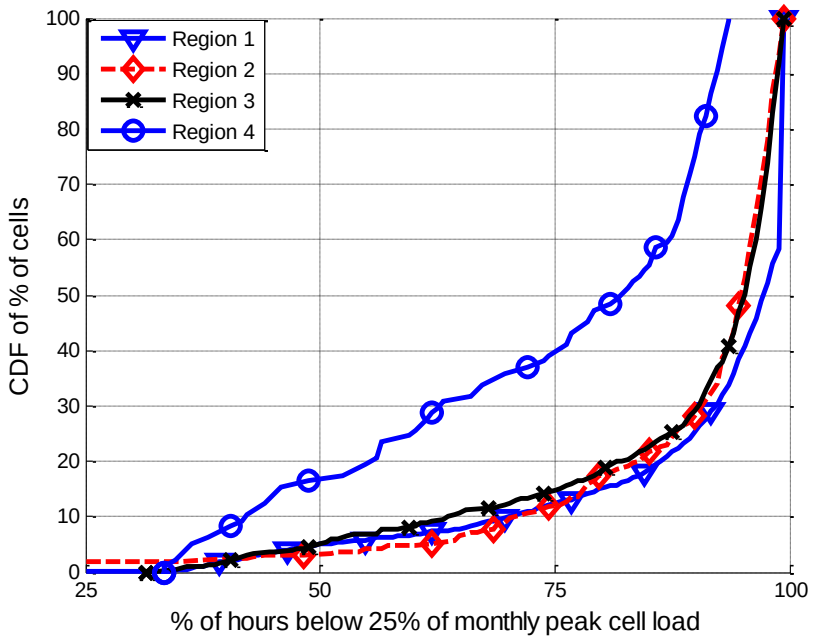


Figure 7.5: The percentage of hours in a month where each cells load falls below 25% of the maximum hourly load observed in that cell during the month.

To further examine temporal dynamics at the individual cell level, the percentage of hours in a month where the traffic load in a cell falls below 25% of that cell's maximum observed load is plotted in Figure 7.5. Here the maximum load is defined as the maximum hourly load observed in the cell during the month under investigation (i.e. the lower bound on the cell's actual capacity as the cells are overprovisioned to deal with future network growth). Figure 7.5 helps underline the large underutilisation of both network hardware and licensed spectrum discussed above. Figure 7.5 shows that for the three Urban/Suburban regions (Regions 1-3) 66% of cells spend at least 90% of the hours in a month with a load of less than 25% of their maximum observed hourly load. Figure 7.5 again like Figure 7.3 and Figure 7.4 shows a difference between the more densely populated regions and the sparser region 4. The greater underutilisation and

consequent opportunity is again present in the more densely populated regions. This is possibly, a result of larger daily flows in and out of these urban regions resulting in a larger peaking problem for network planners compared to the more static nature of rural areas. However, even in the rural region, the median cell spends 80% of their time with a load below 25% of their maximum observed load.

Figure 7.6 illustrates the normalised relative distribution of hours of the day where traffic falls below 25% of capacity for all cells in all regions. Figure 7.6 shows that the hours between 2 AM and 8 AM are approximately twice as likely to see the load fall below 25% of capacity compared to the hours from 6 PM to 11 PM. Of course the hours of low load may vary by location; as discussed in 7.2.4 the local distribution depends on the local profile (urban, suburban, rural, commercial etc.). These hours of low load are particularly suited to modelling due to their larger near term traffic stability as outlined in Chapter 6.

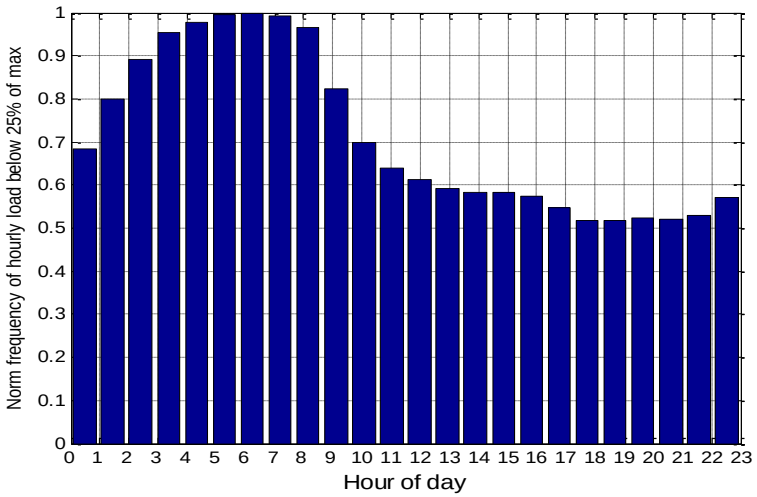


Figure 7.6: Normalised Frequency of hours with a load below 25% of max

7.3 Traffic Prediction Based Energy Savings Scheme

Chapter 6 demonstrated the feasibility of cellular load prediction on small, spatially contiguous groups of mutually interchangeable BSs known as coverage grids. 7.2 identified a large underutilisation of network resources. This underutilisation can be ameliorated by the use of load prediction in these coverage grids to switch off underutilised BS ahead of time. On/off algorithms are a well-studied area of green cellular networks [16, 28], however, many of these works react to instantaneous traffic demands. This may not be entirely realistic as BS switching is not instantaneous, requiring cooling, many complex parameter updates to alter coverage etc. Thus, an element of localised near horizon load prediction is an important enabling step in practical applications. To that end, 7.3.1 gives some background on power consumption in a BS and how it can be modelled. Subsection 7.3.2 introduces a novel BS switching algorithm incorporating near horizon localised predictive models of cell load. 7.3.3 discusses how the switching procedure outlined in 7.3.2 can be practically integrated with current technology. 7.3.4 outlines and justifies the selection of the parameters used in the simulation.

7.3.1 Modelling Power Consumption

The infrastructure of the 3G network is comprised of two main parts: the RAN and the Core Network (CN). The RAN is comprised of the User Equipment (UE), the Radio Network Controller (RNC), and the BS which can be further subdivided into cells. Each RNC manages many BSs which are split into cells and service subscribers through their air interface with the UE [33].

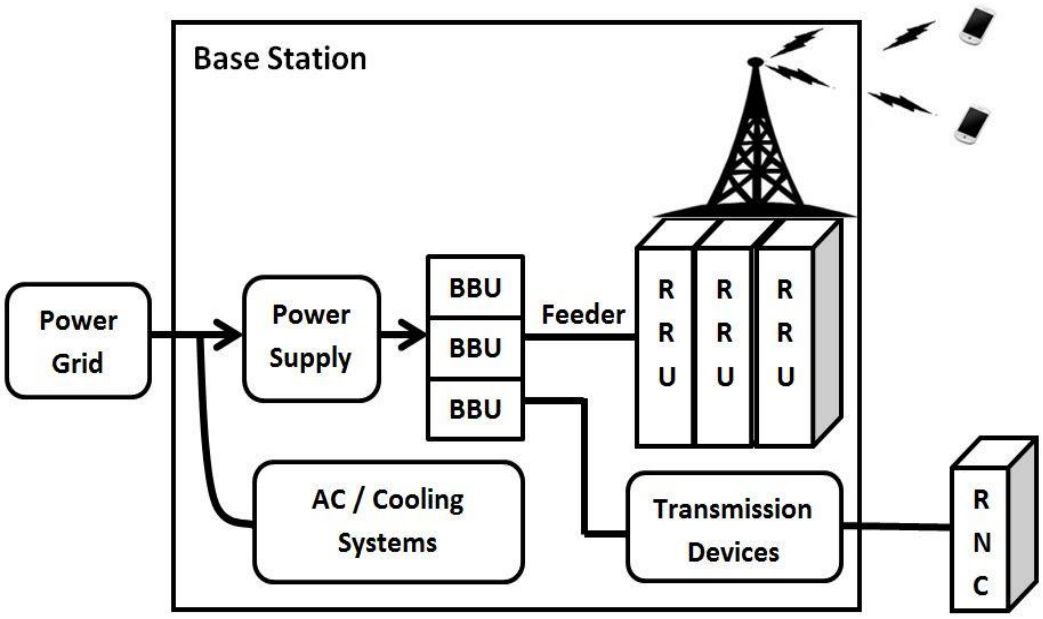


Figure 7.7: A typical BS in a 3G Network

There are two primary subsystems: the communications subsystem and the support subsystem. The communications subsystem is comprised of the Remote Radio Unit (RRU), the Feeder, and the Base Band Unit (BBU). The RRU provides the radio hardware for each sector of the base station. Each BS may have several RRUs near the antennas to allow for varying coverage and capacity [16]. The BBU is responsible for all the other communication functions such as control, Iub interfaces to the RNC, base band, scrambling, link quality measurements, soft handovers, etc. [33]. Each BS may also have several BBUs. The feeder is a fiber optic pair cable connecting the RRUs to the BBUs. The supporting subsystem is comprised of the cooling subsystem and supporting devices. The cooling subsystem maintains an appropriate operating temperature at the BS.

The cooling subsystem coupled with some of the transmission modules are responsible for the consumption of a significant amount of the power in a BS (over 50% [16]) but are load invariant i.e. their power consumption does not proportionately scale down with low demand. Thus, the RAN can conserve large amounts of power by powering down

certain BSs under low load conditions. For BSs, the power consumption models outlined in [16, 110] are applied where the total power consumption *P* at a given BS is given by:

$$P = P_{tx} + P_{misc} \tag{7.2}$$

where, P_{tx} accounts for the power used to provide network access to subscribers' UE. This includes power consumed by the RRUs, the BBUs, the feeder, and the RNC transmissions. P_{misc} is the power consumed by cooling, monitoring and the auxiliary power supply.

P_{tx} can be linearly approximated as:

$$P_{tx}(L) = P_{\alpha} \cdot L + P_{\beta} \tag{7.3}$$

where *L* is the traffic load factor on a BS. P_{tx} varies as a result of both the RRU and BBU. For example, during periods of high traffic the RRU consumes more power servicing more active links. Thus, the power consumption varies with traffic load. Conversely, the BBU carries out base band processing for all frequencies used by the BS. Its power consumption is mainly determined by the number of frequency carriers and not the number of active links. Also, other operations such as signaling over control channels use energy even under low loads. The coefficient P_{α} depends on the transmission distance of the BS as greater power is consumed communicating over a greater distance [16]. P_{misc} as outlined in [16] is mainly a function of external conditions such as temperature. Due to its dependence on temperature, P_{misc} changes constantly, however, it is largely invariant with load.

7.3.2 Switching procedure

Chapter 5 defined spatial regions known as coverage grids where each BS in the region was equivalent i.e. each BS could cover the others area. The near horizon cellular load

was then predicted for these coverage grids in Chapter 6. Now, the next step is to select the correct set of BSs to leave on/switch off. A good switching procedure will:

1. Meet both capacity and coverage requirements while reducing power consumption.
2. Predict the load ahead of time to allow the network to adjust (cooling etc. [16])
3. Minimise the number of on/off switches in a grid.

To meet these requirements a novel switching algorithm was developed which is outlined in Algorithm 7.1. The primary novel feature of this algorithm is the inclusion of the localised prediction models generated in the previous chapters. Firstly, each of the four regions is divided up into coverage grids as discussed in 5.3. Next, an order of switching merit is created. This order of merit determines the order in which BS are switched on and off. The order of merit is created by first sorting the BS within a coverage grid by their capacity. The highest capacity BS should be the first to be turned on and the last to be turned off. Next, BSs are sorted by their power consumption. If two BS have the same capacity but different power consumption then the BS with the lowest power consumption should be the first to be turned on and the last to be turned off. Finally, the BSs are sorted by their distance from the centre of the coverage grid. If two BS have the same capacity, the same power consumption, the one closest to the centre should be the first to be turned on and the last to be turned off. If two or more BSs are equivalent on all of the above, their order of merit can be assigned randomly. BSs closest to the centre of the coverage grids are preferred as this minimises the distance between the User Equipment (UE) and the BS, and thus, minimises the power consumption in the UE [33]. A greater distance between the BS and UE can result in the UE increasing its transmit power when transmitting uplink data. However, it is noteworthy that the uplink to downlink ratio is approximately 1:10 in the network as

shown in Figure 7.8. This means that the BS and not the UE, bears the increased power consumption for 90% of the communication volume. Increased power consumption in the BS due to a greater transmission range is already factored into the BS power model. Also, the extended ranges of the BSs in switching scenarios is still within the limits regularly deployed in rural regions with which the UE are designed to work. Using the order of merit has two primary benefits; firstly, it reduces the number of BS switches required. Minimising BS switching is important because every time a BS switches on/off, handover procedures are initiated generating signalling load in the network. Also, frequent switching is wasteful from an energy perspective as a BS newly switched on may require extra cooling while an already on BS will be currently cooled, etc. Switching decisions are made ahead of time via the predictive SANN models generated in Chapter 6. It is important to assure that the local aggregated capacity assigned for the predicted hour $C_{agg}(h)$ is greater than or equal to the predicted load for that hour $L_{pred}(h)$ plus a certain margin of error γ. The selection of an appropriate margin of error is an important consideration as it a trade-off between maintaining spare capacity to ensure quality of service and minimising power consumption. The results presented in Chapter 6 are used as a guide in this regard. For example, for the SANN predictive model over one month, the maximum error (ACPE) in the prediction for 99% of coverage grids was found to be 20% of capacity. Thus, as a conservative estimate the margin of error γ is set to 20% of a coverage grid's aggregate capacity.

Algorithm 7.1:
BS switching algorithm

1: Create coverage grids for each region as outlined in 5.3.4

2: Create an order of merit for each coverage grid.

I. Sort each BS in the grid by capacity (Decreasing)
II. Sort BS with same capacity by power consumption (Increasing)
III. Sort each BS with same capacity and same power consumption by distance from centre of grid, favouring BS closest to the centre.

3: Use the predictive SANN models outlined in Chapter 6 to predict the next hours load, hour (h).

4: Turn on enough BSs to meet the predicted load ($L_{pred}(h)$) + a margin of error (γ). Select these BS from the order of merit, starting with the highest ranked BS until the aggregate capacity for hour h is $C_{agg}(h) \geq L_{pred}(h) + \gamma$

5: During hour h, predict the load ($L_{pred}(h+1)$) for hour $h+1$.

6: There are three possible scenarios

I. ($C_{agg}(h)$ = ($L_{pred}(h+1)$ + γ), leave BS configuration as is.
II. ($C_{agg}(h)$ < ($L_{pred}(h+1)$ + γ), turn on BSs until $C_{agg}(h) \geq$ ($L_{pred}(h+1)$ + γ). Select BSs to turn on via order of merit (highest currently off first etc.).
III. ($C_{agg}(h)$ > ($L_{pred}(h+1)$ + γ), turn off BSs until doing so would cause ($C_{agg}(h)$ < ($L_{pred}(h+1)$ + γ). Select BSs to turn off via order of merit (lowest currently on is turned off first etc.)

7: Repeat the cycle from point **5** for all subsequent hours.

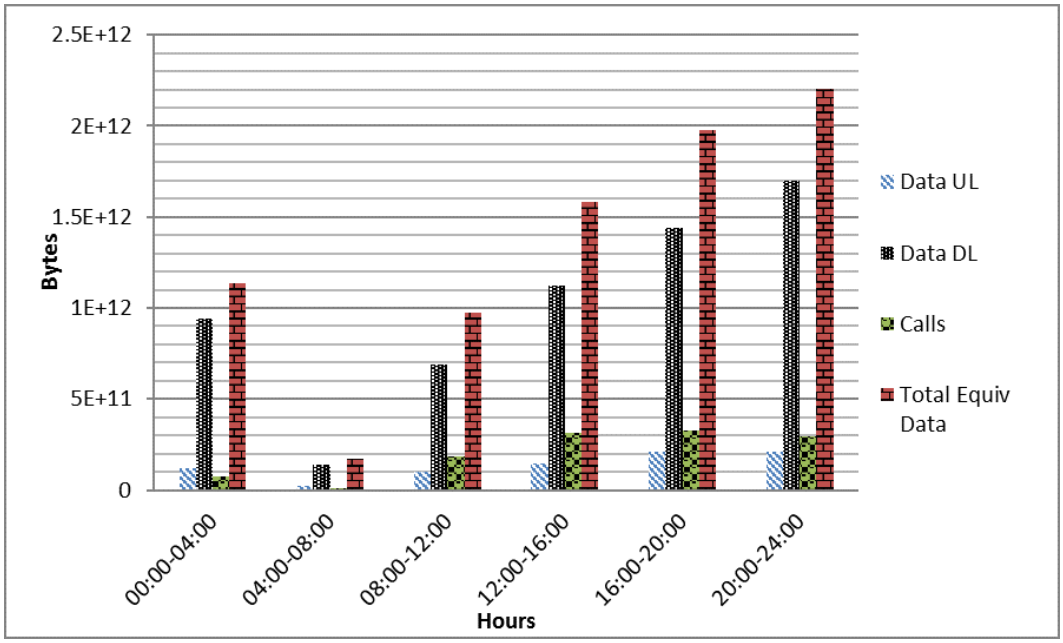


Figure 7.8: The total load on the network for a typical Monday broken down by traffic type and four hour period. This figure originally appeared in Chapter 3 but is reproduced here for the convenience of the reader.

7.3.3 Implementation within Standards

The BS switching algorithm outlined in 7.3.2 raises a number of practical problems that must be overcome within the confines of the current technology. The primary issues that need to be addressed are:

1. How can a BS dynamically change its coverage area to service an area that was previously covered by a BS that has been switched off?
2. When a BS is being switched off, how are its connected clients migrated to another active BS?
3. How can load information be shared between the BSs comprising a coverage grid to enable the real time updating of the predictive models and subsequent switching decisions?

4. How can a switched off BS be brought back online when required and how long will this take?

These four issues can be answered as follows:

Altering Coverage Area: A fundamental component of the switching procedure is the ability to dynamically alter a BS's coverage area to include that of its sleeping neighbours. This can be achieved via the use of cell breathing [111] which is currently used to adjust cell coverage boundaries within 3G networks. Normally, cell breathing is a mechanism which allows overloaded cells to offload subscriber traffic to neighbouring cells by changing the geographic size of their coverage area. Thus, heavily loaded cells decrease in size while neighbouring cells increase their size to compensate. Therefore, a portion of the traffic is distributed from the overloaded cell to neighbouring cells which helps balance the load. In a similar fashion to its primary function as load balancing mechanism, it could also be used for the switching procedure outlined in 7.3.2. In a power saving implementation of cell breathing, a cell would increase in size not when its neighbour's traffic is high but when its neighbour's traffic is low. If the newly expanded cell could meet the traffic needs of both its current and neighbouring service area, then the neighbouring cell could be switched off.

Subscriber Migration: When a BS is switched off to save power, the users currently connected to it will need to be transferred to a new replacement BS. Mobility management of subscribers moving between BS coverage areas is already a fundamental component of 3G networks. Thus, the current Network Controlled HandOff (NCHO) procedure is utilised. The subscriber migration procedure is depicted in Figure 7.9 for a subscriber connected to a BS being switched off. In Figure 7.9 the UE is connected to BS_1 which is being switched off, BS_2 will now service the UE once it is handed over to it. The following handover procedure is carried out for all UE connected

to BS_1: (1) Upon being commanded to switch off, BS_1 sends a handoff request to a neighbouring active BS within the same coverage grid via the RNC; (2) BS_2 sends an acknowledgement of the handoff request to BS_1 via the RNC, BS_2 provisions resources for the migrating UE; (3) BS_1 sends the UE a handoff command; (4) The UE completes the handoff and transmits this through BS_2 which it is now connected to. If the handover procedure fails, BS_1 repeats it with other BSs within the same coverage grid. BS_1 waits until all of the UE connected to it are transferred before switching off. Note that the presence of the handover procedure already in the 3G standard greatly reduces the barrier to implementation of the switching procedure.

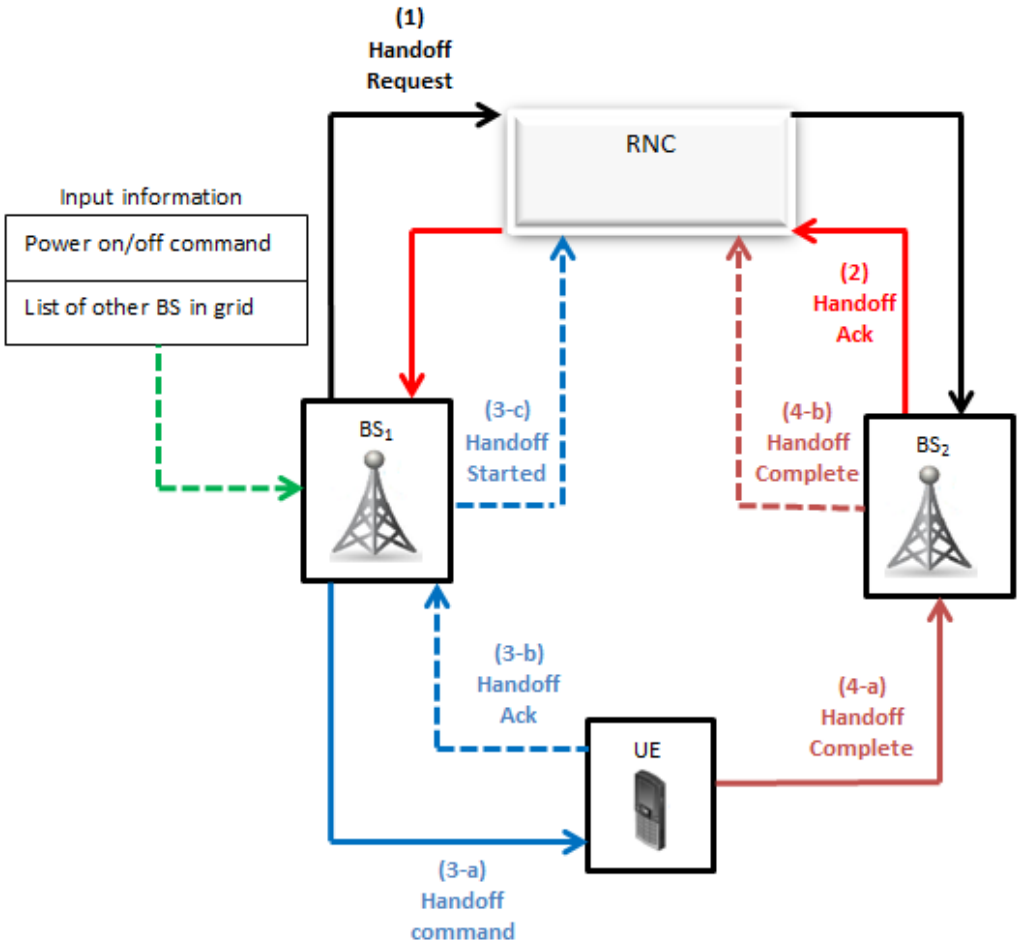


Figure 7.9: Subscriber migration procedure (3G)

Coordination: The managing of the switching procedure necessitates the exchange of load information between the BSs in a coverage grid and a centralised controller capable of implementing the predictive models of cell load and making the required switching decisions. An obvious place for this information sharing and command decision making to take place is at the RNC. Normally, all the BS's within a coverage grid will belong to the same RNC. If on the rare occasion they do not, the local grids can be reconstructed to make sure that all members of a respective grid share an RNC. This will have a marginally negative impact on grid formation but by eliminating inter RNC signalling, it greatly reduced the complexity and the signalling overhead.

Switching On: Many of the subsystems in a BS are designed to operate within certain temperature conditions, hence the AC/cooling system discussed in 7.3.1. When a BS is powered off, the ambient temperature may fluctuate outside of the desired operating range. Therefore, before the BS can be switched on again, the AC/cooling subsystem needs to be turned on and operational in advance. According to real world measurements provided by [16], it can take on average 30 minutes for the AC/cooling system to bring the BS's machine room temperatures into the desired range. This is a problem for many green cellular network procedures which react to the current traffic demand without predicting the future state [28, 110]. Thus, the primary novel feature in this work is the integration of predictive models into the switching procedure. This gives a BS's machine room the required time to reach an optimal temperature and for the BS's subsystems to self-configure.

7.3.4 Parameter Selection

The first step in evaluating the switching procedure outlined in 7.3.2 is the definition of the parameters to be used. Firstly, the margin of error term γ is set to 20% as discussed in 7.3.2. As shown in equation (7.2), the power consumed by a BS comprises P_{tx} and

P_{misc}; P_{tx}, or the transmit power accounts for the power used to provide network access to subscribers' UE. This includes power consumed by the RRUs, the BBUs, the feeder, and the RNC transmissions. P_{misc}, the miscellaneous power, is the power consumed by cooling, monitoring, and the auxiliary power supply. The value of these elements is taken from real world measurements provided in [16]. Equation (7.2) becomes P_{tx} = 6L + 600W at a BS's normal transmission range and P_{tx} = 12L + 600W when expanded to its maximum transmission range. P_{misc} as outlined in [16] is mainly a function of external conditions such as temperature. Due to its dependence on temperature, P_{misc} changes constantly, however, it is largely invariant with load. Thus, for simplicity the assumption is made that the supporting subsystem power consumption stays constant at 1500w as suggested by [16]. The predictive models of cellular load are trained and validated for the coverage grids as in Chapter 6. Then, the switching procedure is tested on a separate independent week of traffic data.

7.4 **Results**

This section will present the results of the simulation of the switching procedure outlined in 7.3.2 to real world independent cellular traffic.

Table 7.2: Power savings broken down by region

	Region 1	Region 2	Region 3	Region 4
$E_{Current}$ (MWh)	30.21	10.23	6.47	3.31
$E_{optimised}$ (MWh)	17.21	6.44	4.91	2.94
$E_{savings}$ (%)	43%	37%	24%	11%
% hours missed	0.0023%	0.0014%	0.017%	0.011%

Table 7.3: Power savings by region by time period

Region	Early Morning $E_{savings}$ (%)	Late Evening/Early Night $E_{savings}$ (%)
Region 1	62.02%	31.12%
Region 2	54.14%	22.01%
Region 3	36.84%	16.23%
Region 4	16.73%	6.30%

The results of the simulation broken down by region, on one independent week of data using the parameters outlined in 7.3.4 are displayed in Table 7.2. The first row, $E_{Current}$, presents the current total daily energy consumption in each region. The second row, $E_{optimised}$, presents the simulated daily energy consumption in each region after using the switching procedure outlined in 7.3.2. The third row, $E_{savings}$, shows the % savings achieved by adapting the switching procedure in the simulation. Finally, the last row % hours missed, shows the percentage of hours over all grids in a region that the aggregate capacity allocated via the prediction step was insufficient for the demand. Table 7.3 shows $E_{savings}$, broken down by period of the day. The two periods are chosen due to them representing opposite ends (peak and trough) of each region's load profile as demonstrated in Figure 7.4. The following observations can be made from the results obtained:

I. As shown in the $E_{savings}$ row of Table 7.2, it is possible to make significant power savings while also preallocating the network resources ahead of time via predictive models. The predictive switching procedure manages to save energy in each of the four disparate study regions while maintaining a negligible % of hours missed.

II. The power savings achievable appear to be a function of the BS density. The more dense the BS deployment in a region the higher the potential power saving as outlined in Table 7.2. For example, comparing the density classifications in Table 7.1 with the potential power savings in Table 7.2 shows that the % savings are: 43% in the most dense urban region 1, 37% in the second most dense suburban region 2, 24% in the third most dense suburban with some semi/rural region 3, and finally 11% in the least dense rural region 4. This is to be expected as the deployment density determines the degree of capacity redundancy available in a coverage grid which influences on/off switching decisions. In sparsely populated regions, such as region 1, a coverage grid may only comprise one BS. Thus, no matter how low the predicted load is at a given time the BS will never be turned off.

III. The potential energy saving is largely dependent on the traffic load. Underutilisation of network resources is disproportionately large at hours of low load. Figure 7.6 shows that these hours are much more likely to occur during the early morning period. Thus, it is no surprise that as outlined in Table 7.3, the largest power savings are made during the lightly loaded early morning hours. For example, in region 1, 62% energy savings are possible during the early morning hours while the corresponding figure for the hours of peak load is only 31%. Across all regions, the energy saving is between 2 and 3 times greater during the early morning hours than the peak Late Evening/Early Night hours.

IV. Energy savings are currently possible even at peak hours, particularly for the regions with the largest deployment density (31% possible in region 1 during peak hours). This may be a result of overprovisioning of network resources to meet future growth in traffic demands, particularly in more profitable densely populated urban regions. However, it may also be symptomatic of a network

that is currently planned and optimised on an ad hoc basis at the BS level. Perhaps, a more holistic view of network provisioning centered on larger spatial aggregation regions with mutual internal redundancy such as coverage grids would lead to a more efficient use of network resources. This becomes even more important as network heterogeneity increases. Fortunately, advances in Software Defined Networking and Self Organising Networks will make the planning and management of these larger coverage areas comprising heterogeneous technologies more feasible.

7.5 Conclusion and Discussion

The implementation of energy efficient cellular networks has been a topic that is widely studied due to both the operational and environmental costs associated with excessive power consumption. Broadly speaking, this work and others like it seek to build energy proportional 3G networks with non-energy proportional 3G BSs. However, much of the work focused on 3G infrastructure relies on idealised simulated traffic traces and complicated unrealistic centralised optimisation models without regard to practical implementation. This chapter has provided a novel, local, distributed, practical approach to BS switching which employs localised near horizon load prediction to give the network infrastructure the required time to react to changes in traffic demand. The energy savings procedure implemented in this chapter gives savings of:

- 43% for densely packed urban areas
- 37% for suburban areas
- 11% for sparsely populated rural areas.

This is less than in other works such as [28] which estimates savings of 50%-80% in densely populated urban areas while [112] suggests theoretical network wide savings of

60%. However, unlike other works, the energy savings procedure presented here (1) does not assume instantaneous BS switching, (2) does not assume perfect knowledge of the load (3) builds in a set margin of error to maintain QOS (4) is computationally simpler (5) mainly relies on techniques and procedures already operational in the network.

The results of the regional analysis show that the degree to which power savings are achievable is a function of the BS deployment density. The denser the BS deployment in a region, the higher the potential power saving. In current and future networks, to overcome capacity constraints cells are getting smaller and more densely deployed (particularly with the widespread adoption of femto and pico cells). It is theorised that these smaller cells will improve the actual power *efficiency* of the network by offloading small high demand areas from the traditional larger macro BSs [113]. This could reduce the need for additional macro BSs, but despite gains in *efficiency*, more cells will overall result in a larger aggregate power consumption. Thus, given the relationship between power saving potential and deployment density, these particularly densely packed high load regions will be particularly suited to the novel power savings scheme outlined here. Widespread underutilisation of the network was found at all times, even in densely populated urban regions at peak hours. This underutilisation is less apparent when viewed from the perspective of an individual BS or sectorised cell. It is also more understandable that individual cells and BSs would be overprovisioned by network operators given their smaller aggregate capacity and greater traffic unpredictability. In future deployment scenarios, it may be more suitable for network operators to focus more on planning the network as a collection of mutually redundant coverage areas (such as the coverage grids presented in this work) as opposed to individual cells/BSs. Of course this is easier said than done, and given the sometimes ad-hoc nature of equipment deployment it would require careful planning and execution to insure

continued redundancy within a grid. However, advances in Software Defined Networking and Self Organising Networks will make the planning and management of these larger coverage areas comprising heterogeneous technologies more feasible.

The primary aim of this chapter was to validate the applicability of predictive models of network load to practical network management strategies. Given more time and data a number of enhancements could be made to the problem of minimising power consumption which are left for future work

(1) This work focused on the 3G network which was the most advanced equipment deployed at the time of data collection. The power saving procedure outlined in this chapter is designed for a heterogeneous network where the various technologies have different power consumption and aggregate capacity characteristics. However, future work could examine the network further with the inclusion of data from more modern radio access technologies.

(2) Possible alternatives to cell breathing for coverage range extension could be considered. For example, BS's containing multiple BBU/RRU subsystems. Many of the components of these subsystems such as power amplifiers are designed to give peak performance under certain transmission conditions. For example, a BS could come equipped with a dual BBU/RRU subsystem, one designed to operate in "urban mode" i.e. over small range and one designed to operate in "rural mode" i.e. long range. When the BS wishes to extend its range, instead of cell breathing it could switch its subsystems from "urban mode" to "rural mode". Another alternative to cell breathing is for a BS to use lower frequency bands if it wishes to extend its range. However, the availability of these bands and managing possible interference would increase the complexity of the solution.

(3) This work was carried out on the network of a single operator. However, typically there are multiple operators servicing the same region each with similar coverage profiles. BSs from different operators are often co-located to minimize rent, capital expenditure, planning issues, and to fully utilise sites that are naturally conducive to broadcasting such as high ground overlooking a town etc. Given that the equipment is co-located, it is possible that the BSs would be capable of providing inter network redundancy. This could increase both the predictability of the now larger multi network coverage grids and also increase the possibilities for energy savings given the (now accessible) greater deployment density. However, for this to be possible the switching procedure outlined in this work would have to be altered and made more complex. For example, coverage grids comprising BSs spanning multiple RNCs are already excluded due to the added complexity and signaling load of the inter-RNC handover process. Internetwork handovers, although possible increase the signaling load further and would push the information sharing point further back from a common RNC with the ensuing complexity. However, in the face of stagnant revenue and a move towards network operator consolidation, it would be an interesting and timely extension to this work.

Chapter 8 Concluding Summary and Future Work

8.1 Concluding Summary

This thesis used anonymised Call Detail Records (CDR) from a mobile network provider in the Republic of Ireland, to model network load and investigate the practicality of localised near horizon predictive models of cellular load on the target network. The Meteor network under investigation had over 1 million customers, representing approximately one quarter of the state's 4.6 million inhabitants when the data was collected.

This thesis began by providing a technical background of cellular networks and their operation in Chapter 2. The dataset used in this thesis was presented and the methods used to store, process, and analysis the dataset were provided. Once the dataset was prepared for analyses, a large scale nationwide study of the cellular network was carried out in Chapter 3. Analysis focused on identifying trends and possible opportunities for resource rationalization. Great spatial disparities in network load were identified within the network with 1% of cells servicing 20% of the total network load. Also, intracell temporal disparities in load were identified with peak hour loads an order of magnitude higher than trough time loads for most cells. This peaking problem was found to be getting relatively worse as more and more cellular data was being used on the network. The spatial disparity between cell loads coupled with the temporal peaking problem made clear the potential for greater resource rationalisation, which would later be implemented in Chapter 7. Chapter 3 also provided empirically created foundational models of how the network experiences load i.e. models of arrival rates, connection

durations and data consumption. These empirically created models will allow for the accurate recreation and modelling of these key network features, not only at the general level but crucially at the device type and contract specific level.

Chapter 4 focused on the creation of a spatial representation of the entire network to allow for the association of load with defined spatial areas. These spatial coverage region representations are a foundational step in beginning to examine the network spatially. Much of the subsequent work and many of the techniques introduced later required the use of these spatial coverage regions. Given their importance, a novel procedure was introduced to clean inaccuracies in the spatial coordinates of cells. A method to visualise how the load is distributed spatially across the network both as a whole and across various services was provided. This method could be generalised to not only represent the load distribution but also other properties of interest such as connection events, subscriber distribution, etc. A novel algorithm to discover who lives and works within BSs/cells was also created and examined. These techniques allowed for the creation of accurate maps of the network's subscriber base and also an examination of how distance affects communication likelihood between areas. Chapter 4 concluded by providing a novel exploration of the presence/lack of causal influence that exists between neighbouring cells. It was found that there is a significant amount of spatial correlation between cell coverage regions in close proximity, decreasing as the separation distance increases. Significant spatial correlation indicates that for monitoring purposes it may only be necessary to monitor a subset of cells. Causal influences were found to be common in the network with 38% of neighbouring cell pairs experiencing statistically significant influence in either one or both directions. Long chained paths of causal influence were found to flow throughout the network. Anecdotally, it appears that these pathways follow significant transport networks. A strong understanding of the spatial aspects of network load are valuable to network

providers and relevant to many advanced network management techniques. They are particularly important to those techniques which rely on a strong spatial understanding such as dynamic spectrum allocation [12], reduced sensing techniques [55], fault detection, and spatially influenced power saving schemes [56] such as the one presented in Chapter 7.

Chapter 5 provided a novel examination of how different levels of load, service type, temporal aggregation, and spatial aggregation affect traffic load predictability. As discussed in Chapter 5, much of the work on predictability and modelling of cellular load excluded the primary driver of cellular load, cellular data. To better understand how cellular data affected predictability relative to the more widely studied service types of voice and SMS, an examination of how predictability varies by service type and time of day was carried out. It was found that voice is the most predictable service, followed by SMS and then finally data. The predictability of all three service types was found to vary largely over all the cells studied. Given the growing predominance of cellular data it was unfortunate that it was the least predictable. However, it was found that during particular times of interest e.g. the early morning hours data load became relatively much more predictable. This better predictability during hours of low usage is particularly useful given that these are the hours most likely to benefit from advanced resource management techniques. Chapter 5 then went on to identify small subsets of the network that provide sufficient predictability to allow for their use in SON techniques. These subsets had to be sufficiently small and spatiality continuous as to be useful for SON techniques. Two novel subsets/spatial aggregation (overlapping cells and coverage grids) meeting these requirements were proposed and compared with the smallest available spatial aggregation, the individual cell level. These comparisons demonstrated that load is more stable and predictable at larger spatial aggregations. Given coverage grid's greater predictability, coupled with their mutual redundancy,

coverage grids appear to occupy the optimal position in the trade-off between predictability and practicality. Traditionally, the network is examined and modelled at the individual cell or individual BS level. However, the results of Chapter 5 indicate that if predictability is an important factor in the analysis/model/network management technique, then higher levels of spatial aggregation are more suitable. It is hoped that the aggregations provided give network operators new ways of viewing their network as opposed to the more traditional macro whole network view or the individual BS view.

Chapter 6 built on the work of Chapter 5 by creating predictive models for the spatial aggregation regions it defined. Two novel methods for the automatic modelling of large amounts of individual cells and their many possible permutations in different spatial aggregations were proposed, used and tested. One of these was based on SARIMA predictive models while the other was based on SANN predictive models. The influence of the metric used to evaluate the predictive models on the perception of the outcome was discussed and led to the creation of a novel metric (ACPE) for this work. Generally, it was found that the SANN model performed better across all aggregation levels. As expected from Chapter 5, the error reduced as the spatial aggregation size increased. This indicated that to improve the performance of predictive models of network load, the largest practicable cellular aggregation should be used as the basis for the predictions. Chapter 6 showed, promisingly, that 95% of the time, 90% of individual cells have an ACPE below 15% and 95% of the time 90% of coverage grids have an ACPE of less than 7%. This means that for 90% of the coverage grids, their load can be predicted to within 7% of their maximum capacity 95% of the time. This demonstrates that automatic, localised near horizon load forecasting is feasible, particularly at higher levels of spatial aggregation.

Chapter 7 introduced a regional study of power usage on the network. The use of near horizon predictive models of cellular load was validated via their incorporation into a novel and practical energy savings scheme which was tested on real world data across multiple regions. The novel energy saving scheme presented is a local, distributed, practical approach to BS switching which employs localised near horizon load prediction to give the network infrastructure the required time to react to changes in traffic demand. The energy savings procedure gave savings on the order of 43% for densely packed urban areas, 37% for suburban areas and 11% for sparsely populated rural areas. This is less than in some other works, however, as discussed further in Chapter 7, the comparable energy saving schemes rely on perfect knowledge of future load, leave no room for error, and require instantaneous BS switching – all three of which are unrealistic real-world requirements. As discussed in greater detail in Chapter 7, the energy savings scheme presented here solves these problems via the incorporation of load prediction. Therefore, the energy saving scheme introduced here is more feasible for practical introduction into real world network deployments. In future deployment scenarios, it may be more suitable for network operators to focus more on planning the network as a collection of mutually redundant coverage areas (such as the coverage grids presented in this work) as opposed to individual cells/BSs. This would be more complicated than traditional network provisioning and given the sometimes ad-hoc nature of equipment deployment it would require careful planning and execution to insure continued redundancy within a grid. However, advances in Software Defined Networking and SONs will make the planning and management of these larger coverage areas comprising heterogeneous technologies more feasible. As discussed in Chapter 1, these technological advances will facilitate the utilisation of advanced management techniques to more efficiently use network resources via SONs as a key component of future 5G networks. The work presented in this thesis demonstrates the feasibility of

creating sufficiently accurate predictive models of network load at useful levels of spatial aggregation. The authors hope that these predictive models can be incorporated into future SONs and the advanced network management techniques that rely on them.

8.2 Future Work

This thesis has addressed many of the central issues related to the creation and incorporation of localised near horizon predictive models of cellular load into SONs. However, there is more work that could be done to update and extend the contributions from this thesis. For example, networks and their underlying technologies, the devices accessing them, and subscriber behaviour are constantly evolving and changing. For example, the data source used for this work was collected after the widespread adoption of smartphones but before the network operator's nationwide rollout of 4G services. Thus, any future work in this area would benefit from updating to include 4G services.

As in any research endeavour, the type and scope of the dataset does impose some restrictions on the research that it can be effectively applied to. For example, the dataset provides details on the start and end cell of each communication event. However, it does not provide location details of devices/subscribers in-between communication events. Therefore, it only provides a sample of a device's/subscriber's location with a sampling rate determined by how often the device/subscriber communicates. Research areas which require detailed knowledge of a device's/subscriber's location at all times, such as modelling the instantaneous signalling load in a specific cell/area, would benefit from additional data. This dataset also does not provide IP packet headers which could be used to identify the specific application/website being used. This precludes research that requires a detailed analysis of these features.

Preliminary investigations of the long causal paths identified in Chapter 4 indicate that when plotted spatially many of the paths follow major transport infrastructure such as busy motorways etc. In future work, it would be interesting to more thoroughly investigate this and examine if there is a relationship between any other geographic features and causal load relationships. Causal relationships indicate that a cells past and present loads can improve the predictability of its causal neighbour's future loads. Quantifying the improvements in predictability from causal neighbour's incorporation in predictive models would be an interesting extension to this work.

Coverage grids' inherent spatial redundancy makes them an ideal default grouping for mutual coverage in the event of localised equipment failure. Although not explored further here, the ability to dynamically alter coverage with confidence within the grid aggregation level suggests the potential for increased equipment redundancy. A quantification and exploration of the increased redundancy from a grid based planning approach to network planning would be interesting. Another interesting avenue of further research would be the incorporation of the predictive models outlined in this work to spectrum sharing schemes. For example, reliable predictions of low loads in a coverage gird with a large load redundancy could indicate the possibility of temporarily freeing up spectrum for secondary usage. The power minimisation technique presented in this work could be improved and updated by incorporating small cell and alternatives to cell breathing such as dual BBU/RRU subsystems.

One major and valuable extension to this work would be the incorporation of data from other overlapping network operators. While this work was carried out on the network of a single operator, typically, there are multiple operators servicing the same region each with similar coverage profiles. Equipment from different operators is often co-located to minimize rent, capital expenditure, planning issues, and to fully utilise sites

that are naturally conducive to broadcasting such as high ground overlooking a town etc. Given that the equipment is co-located, it is possible that the BSs would be capable of providing inter network redundancy. This could increase both the predictability of the now larger multi network coverage grids and also increase the possibilities for energy savings given the (now accessible) greater deployment density. Of course, the considerable benefits of this would have to be compared with the technical and business challenges its implementation would entail.

References:

[1] CISCO. (2017, Feb). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 White Paper*. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[2] Commission for Communications Regulation. (2016, Feb). *Irish Communications Market Quarterly Key Data Report*. Available: https://www.comreg.ie/media/dlm_uploads/2016/09/ComReg-1676r.pdf

[3] J. M. Graybeal and K. Sridhar, "The evolution of SON to extended SON," *Bell Labs Technical Journal,* vol. 15, pp. 5-18, 2010.

[4] E. U. T. R. A. Network, "Self-configuring and self-optimizing network (SON) use cases and solutions," *Third Generation Partnership Project (3GPP) specification TR,* vol. 36, 2010.

[5] S. Feng and E. Seidel, "Self-organizing networks (SON) in 3GPP long term evolution," *Novel Mobile Radio Research,* 2008.

[6] J. Huang, R. A. Berry, and M. L. Honig, "Distributed interference compensation for wireless networks," *IEEE Journal on Selected Areas in Communications,* vol. 24, pp. 1074-1084, 2006.

[7] A. L. Stolyar and H. Viswanathan, "Self-organizing dynamic fractional frequency reuse in OFDMA systems," in *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*, 2008, pp. 691-699.

[8] B. Rengarajan, A. L. Stolyar, and H. Viswanathan, "Self-organizing dynamic fractional frequency reuse on the uplink of OFDMA systems," in *Information Sciences and Systems (CISS), 2010 44th Annual Conference on*, 2010, pp. 1-6.

[9] S. Latif, F. Pervez, M. Usama, and J. Qadir, "Artificial Intelligence as an Enabler for Cognitive Self-Organizing Future Networks," *arXiv preprint arXiv:1702.02823,* 2017.

[10] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: how to empower SON with big data for enabling 5G," *IEEE Network,* vol. 28, pp. 27-33, 2014.

[11] R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, and H. Zhang, "Intelligent 5G: When Cellular Networks Meet Artificial Intelligence," 2017.

[12] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary users in cellular networks: A large-scale measurement study," 2008, pp. 1-11.

[13] U. Paul, M. M. Buddhikot, and S. R. Das, "Opportunistic Traffic Scheduling in Cellular Data Networks," 2012.

[14] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE transactions on wireless communications,* vol. 12, pp. 2126-2136, 2013.

[15] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *INFOCOM*, 2011, pp. 882-890.

[16] C. Peng, S. B. Lee, S. Lu, H. Luo, and H. Li, "Traffic-driven power saving in operational 3G cellular networks," 2011, pp. 121-132.

[17] G. F. Ciocarlie, U. Lindqvist, S. Nováczki, and H. Sanneck, "Detecting anomalies in cellular networks using an ensemble method," in *Network and service management (CNSM), 2013 9th international conference on*, 2013, pp. 171-174.

[18] O. N. Yilmaz, C. Wijting, P. Lundén, and J. Hämäläinen, "Optimized mobile connectivity for bandwidth-hungry, delay-tolerant cloud services toward 5G," in *Wireless Communications Systems (ISWCS), 2014 11th International Symposium on*, 2014, pp. 6-10.

[19] B. Partov, D. J. Leith, and R. Razavi, "Utility fair optimization of antenna tilt angles in LTE networks," *IEEE/ACM Transactions on Networking,* vol. 23, pp. 175-185, 2015.

[20] S. Gondor, A. Uzun, T. Rohrmann, J. Tan, and R. Henniges, "Predicting User Mobility in Mobile Radio Networks to Proactively Anticipate Traffic Hotspots," in *MOBILe Wireless MiddleWARE, Operating Systems and Applications (Mobilware), 2013 International Conference on*, 2013, pp. 29-38.

[21] D. Thilakawardana and K. Moessner, "Traffic modelling and forecasting using genetic algorithms for next-generation cognitive radio applications," *annals of telecommunications-annales des télécommunications,* vol. 64, pp. 535-543, 2009.

[22] M. F. A. YASIN YUR, FATIH ABUT, "Short term voice traffic forecast in 3g/umts networks using machine learning and statistical methods," in *Proceedings of Academics World 41st International Conference,* , Barcelona,, 2016.

[23] G. Pandey, K. M. Siddiqui, and A. Choudhary, "Telecom Voice Traffic Prediction for GSM using Feed Forward Neural Network," *International Journal of Engineering Science and Technology,* vol. 5, p. 505, 2013.

[24] E. Carolan, S. C. McLoone, and R. Farrell, "Comparing and Contrasting Smartphone and Non-Smartphone Usage," presented at the ISSC, LYIT, 2013.

[25] R. Farrell, E. Carolan, S. McLoone, C., and S. McLoone, F., "Towards a Quantitative Model of Mobile Phone Usage Ireland – a Preliminary Study," presented at the ISSC, NUI Maynooth, Ireland, 2012.

[26] K. Kumar, A. Gupta, R. Shah, A. Karandikar, and P. Chaporkar, "On analyzing Indian cellular traffic characteristics for energy efficient network operation," in *Communications (NCC), 2015 Twenty First National Conference on*, 2015, pp. 1-6.

[27] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," 2011, pp. 305-316.

[28] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *Wireless Communications, IEEE Transactions on,* vol. 12, pp. 2126-2136, 2013.

[29] R. Sevlian and R. Rajagopal, "Short Term Electricity Load Forecasting on Varying Levels of Aggregation," *arXiv preprint arXiv:1404.0058,* 2014.

[30] L. Saker, S.-E. Elayoubi, and T. Chahed, "Minimizing energy consumption via sleep mode in green base station," in *Wireless Communications and Networking Conference (WCNC), 2010 IEEE*, 2010, pp. 1-6.

[31] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: A survey, some research issues and challenges," *Communications Surveys & Tutorials, IEEE,* vol. 13, pp. 524-540, 2011.

[32] A. Ghosh, J. Zhang, J. G. Andrews, and R. Muhamed, *Fundamentals of LTE*: Pearson Education, 2010.

[33] M. Sauter, *From GSM to LTE-Advanced: An Introduction to Mobile Networks and Mobile Broadband, Revised Second Edition*: Wiley, 2014.

[34] A. R. Mishra, *Fundamentals of cellular network planning and optimisation: 2G/2.5 G/3G... evolution to 4G*: John Wiley & Sons, 2004.

[35] Ordnance Survey Ireland. (2016, 20/09/2016). *Irish Grid Reference System*. Available: https://www.osi.ie/resources/reference-information-2/irish-grid-reference-system/

[36] F. Bignami, "Privacy and law enforcement in the European union: the data retention directive," *Chi. J. Int'l L.,* vol. 8, p. 233, 2007.

[37] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing Geospatial Dynamics of Application Usage in a 3G Cellular Data Network," 2012.

[38] Y. F. Chuang, "Pull-and-suck effects in Taiwan mobile phone subscribers switching intentions," *Telecommunications Policy,* vol. 35, pp. 128-140, 2011.

[39] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018," 2013.

[40] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020," 2016.

[41] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *Speech and Audio Processing, IEEE Transactions on,* vol. 10, pp. 620-636, 2002.

[42] H. Taddei, I. Varga, L. Gros, C. Quinquis, J. Y. Monfort, F. Mertz, and T. Clevorn, "Evaluation of AMR-NB and AMR-WB in packet switched conversational communications," in *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, 2004, pp. 2003-2006.

[43] S. A.-h. Soliman and A. M. Al-Kandari, *Electrical load forecasting: modeling and model construction*: Elsevier, 2010.

[44] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2011-2016. White Paper, 2012.," 2011.

[45] M. G. Calum Dewar, "Rebalancing the value from voice and SMS to data," *GSMA Intelligence,* 2014.

[46] A. Lucent, "The declining profitability trend of mobile data:what can be done? ," http://www3.alcatel-lucent.com/belllabs/advisory-services/documents/2011.

[47] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011–2016," http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html2012.

[48] E. Parliament, "REGULATION (EU) 2015/2120 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL," *Official Journal of the European Union,* 2015.

[49] T. H. f. Christian Schwartz, Frank Lehrieder, and Phuoc Tran-Gia, "Angry Apps: The Impact of Network Timer Selection on Power Consumption, Signalling Load, and Web QoE," *Journal of Computer Networks and Communications,* 2013.

[50] J. Loudiadis. (2014). *The mobile app impact rankings: Where apps meet network, battery and data plan*.

[51] F. Economics, "Modelling the network cost savings to mobile network operators of a change of use of the 700 MHz band," Comreg2015.

[52] F. J. Massey Jr, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American statistical Association,* vol. 46, pp. 68-78, 1951.

[53] W. Rudin, "Principles of Mathematical Analysis (International Series in Pure & Applied Mathematics)," 1976.

[54] N. Johnson, S. Kotz, and N. Balakrishnan, "Lognormal distributions. Continuous Univariate Distributions (Vol. 1)," ed: John Wiley & Sons, 1994.

[55] U. Paul, L. Ortiz, S. R. Das, G. Fusco, and M. M. Buddhikot, "Learning probabilistic models of cellular network traffic with applications to resource management," in *Dynamic Spectrum Access Networks (DYSPAN), 2014 IEEE International Symposium on*, 2014, pp. 82-91.

[56] E. Carolan, R. Farrell, and S. McLoone, "A Predictive Model for Minimising Power Usage in Radio Access Networks," in *7th EAI International Conference on Mobile Networks and Management*, Santander, Spain, 2015.

[57] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu, *Spatial tessellations: concepts and applications of Voronoi diagrams* vol. 501: John Wiley & Sons, 2009.

[58] Ordnance Survey of Ireland. (21/02/2016). Available: http://www.osi.ie/Home.aspx.

[59] 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,* vol. 2009, p. L07003, 2009.

[60] E. Carolan, S. C. McLoone, S. F. McLoone, and R. Farrell, "Analysing Ireland's interurban communication network using call data records," in *Signals and Systems Conference (ISSC 2012), IET Irish*, 2012, pp. 1-6.

[61] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science,* vol. 286, pp. 509-512, 1999.

[62] J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A. L. Barabási, "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sciences,* vol. 104, p. 7332, 2007.

[63] P. S. Dodds, R. Muhamad, and D. J. Watts, "An experimental study of search in global social networks," *science,* vol. 301, pp. 827-829, 2003.

[64] M. C. Gonzalez, C. A. Hidalgo, and A. L. Barabási, "Understanding individual human mobility patterns," *Nature,* vol. 453, pp. 779-782, 2008.

[65] J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. Argollo de Menezes, K. Kaski, A. L. 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.

[66] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove, "Mobile call graphs: beyond power-law and lognormal distributions," 2008, pp. 596-604.

[67] R. Lambiotte, V. D. 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, pp. 5317-5325, 2008.

[68] G. K. Zipf, *Human behavior and the principle of least effort: An introduction to human ecology*: Ravenio Books, 1949.

[69] W. K. Davies, "Urban connectivity in Montana," *The Annals of Regional Science,* vol. 13, pp. 29-46, 1979.

[70] G. A. Carrothers, "An historical bedew of the gravity and potential concepts of human interaction," *Journal of the American Institute of Planners,* vol. 22, pp. 94-102, 1956.

[71] 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,* vol. 2009, p. L07003, 2009.

[72] C. S. Office. (2011). *Census 2011 Boundry Files*. Available: http://www.cso.ie/en/census/census2011boundaryfiles

[73] D. Kelly, J. Doyle, and R. Farrell, "Analysing Ireland's Social and Transport Networks using Sparse Cellular Network Data," presented at the ISSC, Trinity Col Dublin, 2011.

[74] CSO. (2011, 01/02/2016). *Census 2011 Reports*. Available: http://www.cso.ie/en/census/census2011reports/

[75] Y. Kim, R. Balani, H. Zhao, and M. B. Srivastava, "Granger causality analysis on ip traffic and circuit-level energy monitoring," in *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, 2010, pp. 43-48.

[76] P. A. P. Moran, "Notes on continuous stochastic phenomena," *Biometrika,* pp. 17-23, 1950.

[77] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding spatial relationships in resource usage in cellular data networks," 2012, pp. 244-249.

[78] E. Carolan, S. McLoone, and R. Farrell, "Characterising Spatial Relationships in Base Station Resource Usage," in *Proceedings of the 17th Research Colloquium on Communications and Radio Science into the 21st Century*, 2014.

[79] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society,* pp. 424-438, 1969.

[80] A. K. Seth, "A MATLAB toolbox for Granger causal connectivity analysis," *Journal of neuroscience methods,* vol. 186, pp. 262-273, 2010.

[81] H. Akaike, "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on,* vol. 19, pp. 716-723, 1974.

[82] D. J. L. R. R. Bahar Partov, "Energy-Aware Configuration of the Small Cell

Networks," *Unpublished Work as of Sep 2013,* 2013.

[83] G. E. Box and G. M. Jenkins, *Time series analysis: forecasting and control, revised ed*: Holden-Day, 1976.

[84] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science,* vol. 327, pp. 1018-1021, 2010.

[85] J. Viebig, "Are Emerging Market Investors Overly Pessimistic in Extreme Risk-off Periods?," *Journal of Behavioral Finance,* vol. 16, pp. 163-172, 2015.

[86] S. Kullback, *Information theory and statistics*: Courier Corporation, 1997.

[87] T. DelSole, "Predictability and information theory. Part I: Measures of predictability," *Journal of the atmospheric sciences,* vol. 61, pp. 2425-2440, 2004.

[88] T. Schneider and S. M. Griffies, "A conceptual framework for predictability studies," *Journal of climate,* vol. 12, pp. 3133-3155, 1999.

[89] D. E. Knuth, R. L. Graham, and O. Patashnik, "Concrete mathematics," *Adison Wesley,* 1989.

[90] R. A. Fisher, "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," *Biometrika,* vol. 10, pp. 507-521, 1915.

[91] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks:: The state of the art," *International journal of forecasting,* vol. 14, pp. 35-62, 1998.

[92] E. Carolan, S. McLoone, S. McLoone, and R. Farrell, "Analysing Ireland's Interurban Communication Network using Call Data Records," presented at the ISSC, NUI Maynooth, 2012.

[93] E. Carolan, S. C. McLoone, and R. Farrell, "Predictive modelling of cellular load," in *Signals and Systems Conference (ISSC), 2015 26th Irish*, 2015, pp. 1-6.

[94] J. Kamruzzaman, *Artificial Neural Networks in Finance and Manufacturing*: IGI Global, 2006.

[95] Y. Chen, B. Yang, and J. Dong, "Time-series prediction using a local linear wavelet neural network," *Neurocomputing,* vol. 69, pp. 449-465, 2006.

[96] G. P. Zhang and M. Qi, "Neural network forecasting for seasonal and trend time series," *European journal of operational research,* vol. 160, pp. 501-514, 2005.

[97] W. H. Greene, "The econometric approach to efficiency analysis," *The measurement of productive efficiency and productivity growth,* pp. 92-250, 2008.

[98] D. R. Anderson, *Model based inference in the life sciences: a primer on evidence*: Springer Science & Business Media, 2007.

[99] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing,* vol. 50, pp. 159-175, 2003.

[100] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks,* vol. 2, pp. 359-366, 1989.

[101] R. Adhikari and R. Agrawal, "Forecasting strong seasonal time series with artificial neural networks," *Journal of Scientific and Industrial Research,* vol. 71, p. 657, 2012.

[102] G. P. Zhang, "A neural network ensemble method with jittered training data for time series forecasting," *Information Sciences,* vol. 177, pp. 5329-5346, 2007.

[103] J. Kihoro, R. Otieno, and C. Wafula, "Seasonal time series forecasting: a comparative study of ARIMA and ANN models," *African Journal of Science and Technology,* 2004.

[104] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," DTIC Document1985.

[105] T. Hill, M. O'Connor, and W. Remus, "Neural network models for time series forecasts," *Management science,* vol. 42, pp. 1082-1092, 1996.

[106] I. Alon, M. Qi, and R. J. Sadowski, "Forecasting aggregate retail sales:: a comparison of artificial neural networks and traditional methods," *Journal of Retailing and Consumer Services,* vol. 8, pp. 147-156, 2001.

[107] C. Hamzaçebi, "Improving artificial neural networks' performance in seasonal time series forecasting," *Information Sciences,* vol. 178, pp. 4550-4559, 2008.

[108] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the Marquardt algorithm," *Neural Networks, IEEE Transactions on,* vol. 5, pp. 989-993, 1994.

[109] G. Fettweis and E. Zimmermann, "ICT energy consumption-trends and challenges," in *Proceedings of the 11th International Symposium on Wireless Personal Multimedia Communications*, 2008, p. 6.

[110] C. Peng, S.-B. Lee, S. Lu, H. Luo, and H. Li, "Traffic-driven power saving in operational 3G cellular networks," in *Proceedings of the 17th annual international conference on Mobile computing and networking*, 2011, pp. 121-132.

[111] G. Americas, "The benefits of SON in LTE: self-optimizing and self-organizing networks," *White paper,* 2009.

[112] V. Pritika, G. P. Venkatesan, and N. Angayarkanni, "A Base Station Switching Scheme for Green Cellular Networks."

[113] L. Saker, G. Micallef, S.-E. Elayoubi, and H. O. Scheck, "Impact of picocells on the capacity and energy efficiency of mobile networks," *annals of telecommunications-annales des télécommunications,* vol. 67, pp. 133-146, 2012.