Visualization and analysis of mobile phone location data

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

Matthew Kwan

BSc Hons, MBA (Melb)

School of Mathematical and Geospatial Sciences

College of Science, Engineering and Health

RMIT University

July 2012

Declaration

I certify that:

a) except where due acknowledgement has been made, the work is that of the author alone;

b) the work has not been submitted previously, in whole or in part, to qualify for any other academic award;

c) the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program;

d) any editorial work, paid or unpaid, carried out by a third party is acknowledged;

e) ethics procedures and guidelines have been followed.

_____ _____

Matthew Kwan Date

Acknowledgements

I would like to thank my primary and secondary supervisors, Colin Arrowsmith and William Cartwright, respectively. It all went very smoothly.

Table of Contents

List of Tables

List of Figures

List of Acronyms

3G	3rd Generation mobile telecommunications standards
3GPP	3rd Generation Partnership Project, the consortium responsible for developing and maintaining the 3G standards
ABS	Australian Bureau of Statistics
ACMA	Australian Communications and Media Authority
AGD66	Australian Geodetic Datum 1966, an old datum used in Australia
AOA	Angle of Arrival
ASU	Active Set Update, an integer value proportional to the received signal strength measured by a mobile phone
BSC	Base Station Controller
BSS	Base Station Subsystem
BSSAP	Base Station Subsystem Application Part. The part of the SS7 protocol responsible for Base Station Subsystem communications
BTS	Base Transceiver Station
CBD	Central Business District
CD-ROM	Compact Disc – Read Only Memory
CDMA	Code Division Multiple Access, a channel access method used by some mobile phone standards
CID	Cell Identifier
CPICH RSCP	Common Pilot Channel Received Signal Code Power
GDA94	Geocentric Datum of Australia 1994, the current datum used in Australia
GIS	Geographic Information System
GPRS	General Packet Radio Service, a packet-based mobile data communications standard used by 3G and GSM
GPS	Global Positioning System
GSM	Global System for Mobile, a second generation mobile phone standard
HLR	Home Location Register, a database containing details of mobile phone subscribers
IMSI	International Mobile Subscriber Identity, a unique identifier for a mobile phone subscriber
LA	Location Area
LAI	Location Area Identifier
LBS	Location-Based Services
MAC address	Media Access Control address, a unique address for network interfaces
MAP	Mobile Application Part. The part of the SS7 protocol used to provide mobile phone services

MCC	Mobile Country Code, identifies a country
MNC	Mobile Network Code, identifies a mobile phone operator in a country
MS	Mobile Station, a device capable of communicating with a mobile network
MSC	Mobile Switching Centre, part of a mobile network responsible for routing calls
MSIN	Mobile Subscriber Identification Number, identifies a subscriber within a mobile network
NSS	Network and Switching Subsystem, the part of a mobile network responsible for call routing and mobility management
OD	Origin-destination, usually referring to an OD matrix
OTV	Optus Traffic View, a traffic information service offered by Optus
PC	Personal Computer
PERT	Program Evaluation and Review Technique, a statistical tool used in project management
PSC	Primary Scrambling Code, used by 3G networks
PTU	Portable Tracking Unit, a waist-mounted GPS unit used in conjunction with a transmitter attached to an ankle. Use to monitor house arrest.
RMIT	Royal Melbourne Institute of Technology
RRL	Record of Radiocommunications Licences
RTT	Round Trip Time
SIM	Subscriber Identity Module, a circuit, usually part of a SIM card, storing a subscriber's IMSI
SMS	Short Message Service
SMS-CB	Short Message Service - Cell Broadcast, a version of SMS that broadcasts to all handsets within a cell
SMS-PP	Short Message Service – Point-to-point, same as SMS
SS7	Signalling System No. 7, a set of telephony signalling protocols
TA	Timing Advance, a length of time needed to compensate for speed-of-light delays on GSM networks
TMSI	Temporary Mobile Subscriber Identity, a randomly-generated identifier assigned to mobile subscribers to prevent them from being tracked by radio eavesdroppers
UMTS	Universal Mobile Telecommunications System, a third generation mobile phone standard
VLR	Visitor Location Register, a database containing details of the mobile phones currently operating in the jurisdiction of an MSC

Summary

This thesis investigates the use of passively-collected data from mobile phone networks to map population movements. In Australia, as in most other developed countries, nearly all teenagers and working-age adults carry a mobile phone. When these phones communicate with the network they reveal their location to be within the coverage area of the base station antenna that received their transmission. This location data, if it were collected, could be used to derive movement information for most of the population. Such information does not currently exist.

The thesis begins by investigating what information is available within a mobile phone network during normal operations. It looks at how difficult it is to extract this information, how frequently it is generated, and the spatial accuracy when it is used to locate a mobile handset. A new technique is described for estimating the location of a handset within the coverage area of a directional antenna.

The theoretical investigation is supplemented by the collection of field data with a GPS-equipped smart phone running custom software; by simulating the movement of Australia's mobile phones using census data and a database of base station antenna locations; and by analyzing the mobile phone billing records of an individual who elected to make his data public.

Having researched the accuracy and availability of mobile phone location data, the thesis then looks at the feasibility of using it for various applications. These applications include sending alerts to people in the path of a tsunami; predicting the utilization of a new public transport route; tracking the movements of fugitives and missing persons; measuring internal migration within Australia; identifying abnormal population concentrations in real-time; and measuring the population of a region throughout the day/year.

Finally, the thesis looks at techniques for visualizing the data. Existing techniques are explored, and a new one is proposed that makes use of clustered velocity vectors. This new approach can display the location, quantity, speed, and direction of large numbers of people at a point in time, and do so efficiently in terms of computational speed.

The thesis concludes by summarizing the potential applications of mobile phone location data and suggesting areas of further research.

1 Chapter One: Introduction

1.1 Introduction

This chapter will review the key aims of this research thesis and provide an overview of its structure.

1.2 Background

In his 1992 book “Mirror Worlds” (Gelernter 1992), David Gelernter described a vision of the future where every object in the real world is mirrored in the digital world. Whenever a real-world object moves or changes its status, its digital equivalent is immediately updated to reflect the change.

The rationale behind the vision was that once information becomes available in digital form it can be processed by computers. Anomalous behaviour can be detected and acted upon. Processes can be optimized. And patterns can be detected, extrapolated, and used to predict the future.

At present the “static” digital world – the representation of unmoving real-world objects such as terrain, roads, and buildings – is richly populated thanks to the digital mapping efforts of numerous organizations. But the *dynamic* mirror world, representing things that move and change, is still sparsely populated. And this is especially the case when it comes to representing people.

To dynamically populate a mirror world *sensors* are needed (Saffo 1997). These are devices that measure some aspect of the real world and make that information available in digital form. And when it comes to measuring the activities of people, mobile phones make excellent sensors. Not only do they tend to remain in close proximity to their owner, but they can also communicate their status in real-time.

According to the latest figures from the Australian Communications and Media Authority (ACMA), the rate of mobile phone ownership among working-age Australians is over ninety percent (Australian Communications and Media Authority 2009a). This ubiquity means that the location of Australia's handsets is a good proxy for the location of its adult population at any point in time. And unlike other movement tracking techniques, such as road sensors, public transport ticket tracking, and the filling out of surveys, the collection of mobile phone

location data provides widespread coverage of the population, in real-time, using existing physical infrastructure.

Although mobile phone networks were not designed with handset location in mind, there is a constant stream of location-related information flowing through them as a by-product of their normal operations, and much of this is stored for billing purposes. During regular use handsets register with the network, make and receive calls, move to different regions, and periodically update their status, each time sending their current *cell ID* across the network.

A cell ID corresponds to an antenna on a mobile phone tower, and this allows handsets to be located to within the antenna's coverage area, or *cell*. The size of a cell depends on the density of antennae in its vicinity, but they generally vary from a few hundred metres across in inner-city neighbourhoods to tens of kilometres in rural areas.

There are other techniques that can locate handsets more accurately, such as interacting with a handset's built-in GPS or using speed-of-light delay to determine a handset's distance from the antenna. However, these tend to be *active* techniques requiring the participation of the network infrastructure and the handsets, which places an additional load on both. The aim of this thesis is to investigate what can be achieved *passively* using existing infrastructure, so active techniques will be discussed only briefly.

Mobile networks can support hundreds of millions of subscribers, and passively collecting all of their location-related information would result in an enormous quantity of data. Analysis of the data would reveal population distributions at different times of the day, as well as the movement history of every handset on the network. Clearly, collecting and analyzing such volumes of data poses technical challenges, but it also raises privacy concerns.

Because the data reveals the movements of every handset, along with a unique identifier that can usually be traced back to an individual, a carrier releasing the data in its raw form would almost certainly be breaching privacy laws. On the other hand, redacting parts of the data to preserve privacy may reduce its usefulness. At the end of the day there will probably be a trade-off between privacy and effectiveness, as there is with most tracking techniques, but this is beyond the scope of the thesis, which will focus on the technical issues.

1.3 Research questions

The aim of this thesis is to develop a method for monitoring and visualizing the spatial behaviour of populations in a large urban environment using passively-collected mobile

phone location data. The following questions will be addressed -

- What location information is available in a mobile phone network, and how can it be extracted?
- How accurate is the data, spatially and temporally?
- Can the data be used to support *real-time* applications such as measuring crowds, finding missing persons, and managing emergency evacuations?
- Can the data be used for *historical* analysis, to investigate movement patterns and, for example, optimize the design of transport infrastructure and the urban environment?
- What is the best way to present this information visually?

1.4 Rationale

At first glance, the use of mobile phone location data would appear to have enormous potential to support the real-time and historic applications mentioned in the research questions above. The real-time applications can provide a birds-eye view of where people are currently located, which would be a valuable tool for emergency services. And the historic applications would provide previously-unavailable data about how populations move around, allowing planners to better design infrastructure and the urban environment.

A recent example of where historic data movement would be useful is in the design of electric cars (Simonite 2009). Because the choice of battery is critical to an electric car's range and performance, but is one of its most expensive components, it is important for car companies to select the cheapest battery that will meet the needs of most of its target market. Access to the accurate movement history of large populations would allow for designs that are optimized for actual usage, keeping costs down and thus stimulating demand.

With this in mind, a team designing electric cars at Carnegie Mellon University in Pittsburgh are appealing for GPS logs from commuters to use instead of the US's Urban Dynamometer Driving Standard, which is described as a “terrible model” of how people actually drive (Simonite 2009). While not being able to provide detailed information on driving habits, mobile phone location data could certainly be used to show distances travelled.

This thesis will investigate the technical details of using passively-scanned mobile phone data to monitor the location of populations. It will seek to determine its practicality, limitations and accuracy. And should the techniques prove feasible, prototype software will be developed to

demonstrate their capabilities. These prototypes may eventually evolve into real-world applications.

1.5 Methodology

This thesis will commence with a review of the literature in chapter two. It will begin by reviewing literature in the area of population tracking, with a particular focus on research involving tracking using mobile phones. The second half will cover literature relating to the visualization of movement data. Chapter three will investigate existing tracking techniques in more detail, and compare them using a number of criteria.

The thesis will then proceed using a simulation methodology similar to that described by Caceres *et al.* (2007), which moved virtual handsets along roads and generated the location update communications that would occur as they crossed cell boundaries. This will be extended to simulate mobile phone handsets for the entire Australian population as they move through Australia's cells.

The mobile phone data generated by the simulation will be used to derive movement information for the Australian population, which can then be compared with the movements used to drive the simulation in the first place. This will provide a measure of the accuracy of movement information which is derived from mobile phone data.

Additional data will be collected from a programmable mobile phone taken on journeys around Melbourne to help verify the accuracy of the simulated data. The phone will periodically collect cell and signal strength information along with the corresponding GPS coordinates.

Having generated the simulated data, a number of visualization algorithms will be applied to see if the data can be presented in ways that provide useful information to urban planners, emergency services, and other potential users.

The thesis will conclude with a summary of the findings and how they apply to the research questions, followed by recommendations for future research.

1.5.1 Data collection

The first step in collecting the data is to determine the geographic coverage area of Australia's mobile phone network. The Australian Communications and Media Authority (ACMA) publishes a CD-ROM called the *Record of Radiocommunications Licences* (Australian

Communications and Media Authority 2009b), a database of every antenna transmitting or receiving on a licensed frequency in Australia. The database contains information such an antenna's location, radio frequency, owner, whether it is directional, and, if so, what direction it points in.

By extracting the subset of antennae transmitting on frequencies used by mobile phones, this database can be used to estimate the coverage area of every mobile phone cell in Australia. These coverage areas can then be used to turn a cell ID into a geographic location, as discussed in chapter five.

Ideally the next step would be to obtain a large sample of real mobile phone data and analyze that. Some researchers have published studies using billing data that they obtained from places such as Estonia (Ahas *et al.* 2007), Boston, Massachusetts (Calabrese *et al.* 2010), and Harbin City, China (Yuan & Raubal 2010). On the other hand, others have mentioned their lack of success in accessing data in Rome, Italy (Reades *et al.* 2007) and the United Kingdom (White & Wells 2002).

It is not apparent from their papers what approach the successful researchers used to gain access to their data, but it turned out to not be available in Australia. Two carriers initially expressed an interest in providing data, but were unable to proceed. To the author's knowledge, no other researcher has succeeded in accessing Australian data either.

The only data that was available came from an individual in Germany who sued Deutsche Telekom to access six months of his own billing data, which he then made publicly available. This data is analyzed in chapter eight.

It would have been even better to obtain a sample of the data that flows through a mobile phone network when handsets carry out various activities. This network data could be captured by installing “probes” and reading the data as it passes. Depending on the equipment involved, these probes may be implemented as physical taps on a cable or, more likely, by enabling data logging on specific pieces of equipment. The mechanics of this process are covered in detail in chapter four.

However, the collection of real data requires the cooperation of a mobile phone carrier, who, among other things, must provide engineering resources to gain physical access to the data. Again, this turned out to not be possible, with no carrier able to provide the data.

With limited access to real mobile phone data, research will proceed using simulated data. Realistic handset “home” locations can be generated using mobile phone ownership rates

from ACMA surveys (Australian Communications and Media Authority 2008, Australian Communications and Media Authority 2009a) and population distributions from the Australian Bureau of Statistics (ABS) census data (Australian Bureau of Statistics 2008). Applying a movement algorithm to these virtual handsets and continually comparing their virtual position with cell coverage areas will generate a simulated stream of cell and handset IDs.

One benefit of these simulated location records is that the underlying virtual locations from which they were generated are known exactly. These can be compared to the locations derived from the simulated cell IDs, allowing the accuracy of cell IDs to be evaluated.

To check that the simulated data corresponds with the real world, a logging application will be developed to run on a mobile phone running the Android operating system (a “smart phone”) that will periodically record cell IDs, signal strengths, and GPS coordinates as the phone moves around. The phone will be taken on trips around Melbourne to collect sample data that can then be used to verify that cell IDs are being correctly generated for the virtual handsets in the simulation. This will be covered in chapter seven.

1.5.2 Data analysis

Having collected the location records, whether from a real network or from a simulation, the next step is to develop software tools to analyze them. Some thought will go into the storage format of the data, which must be compact, but allow tools to quickly extract relevant information without having to scan linearly through hundreds of millions of records. The storage format will probably be supported by a common interface library, ensuring that all tools can access the data in the same manner.

The analysis tools will have to be configurable and interactive, to allow various scenarios to be tested without having to write a new application for each. For example, a tool may allow a user to filter the data down to handsets located in the Melbourne CBD at 4pm, then run an animation showing where they go later in the day, and generate a report showing the average distance travelled. This will probably be achieved by linking the stored data to a high-level scripting language such as Python.

An important feature of the analysis will be visualization. Because of the enormous quantity of data, spread over space and time, effective visualization algorithms will be needed to make sense of it. A number of algorithms will be tested, and it is possible that different ones will be

selected for different applications. For example, the best way to display handset locations in a real-time scenario may not the best way to display commuting patterns extracted from from historic data.

The results of these scenario tests will be used to evaluate the feasibility of the various real-time and historic applications. This will be discussed in chapter nine.

1.6 Conclusions

This chapter has provided an outline of the aims and methodology of the thesis. The aim, broadly speaking, is to determine whether mobile phone location data collected from a network can provide information that is useful for various applications. This will be carried out by simulating the movement of mobile phones and by carrying out small-scale tests with a GPS-equipped programmable handset, then analyzing the resulting data.

The next chapter will commence the literature review.

2 Chapter Two: Literature review

2.1 Introduction

The aim of this chapter is to review the literature in the area of tracking and analyzing population movements. In particular it will focus on research relating to the use of mobile phone data to determine locations, and techniques relevant to the visualization of that data.

2.2 Mobile landscapes

The concept of Mobile Landscapes, proposed by Ratti *et al.* (2006), is about using aggregated mobile phone location data to visualize urban activities and show how they vary through space and time. In work pioneered by the Massachusetts Institute of Technology's SENSEable City Lab, real-time images were generated based on the number and duration of mobile phone calls made in each cell in a city, showing levels of mobile phone activity. This technique was used to generate maps of Graz, Austria (Ratti *et al.* 2005), Milan (Pulselli *et al.* 2005, Pulselli *et al.* 2006, Pulselli *et al.* 2008, Ratti *et al.* 2006) Rome (Calabrese & Ratti 2006, Reades *et al.* 2007, Rojas *et al.* 2007), and Pescara, Italy (Pulselli *et al.* 2008).

The Graz study (Ratti *et al.* 2005), carried out with the co-operation of A1/Mobilkom Austria, was designed as an art exhibit for the M-City exhibition in Graz, and focuses on the mapping and visualization of the collected data. In addition to displaying the call volumes in each cell, the exhibit showed handovers (handsets moving from one cell to another while on a call), and five-minute periodic location updates of actively-tracked handsets whose owners had opted in. The data was all processed within ArcGIS and exported for viewing in Flash. This was the first time mobile phone location data had been mapped this way, allowing a city to be visualized in almost real-time. As the authors say, “the results seem to open the way to a new paradigm in urban planning: that of the real-time city” (Ratti *et al.* 2005, p 2). However, at the time of publishing the results were preliminary, and no urban planning benefits were mentioned.

One of the first papers to discuss in detail the potential of mobile phone location data for urban planning was Ratti *et al.* (2006), which coined the term "Mobile Landscapes" to describe the new field. The paper not only provides a detailed overview of the visualization of mobile phone activity in Milan, but also discusses the history of location based services (LBS) in some depth. However, many of the potential applications they suggested - estimating flows

in/out of cities, emergency relief, commuting patterns weekday vs weekend vs holiday – could not be supported with call density data, which can't be used to track handset movements. In what turns out to be a common complaint in this field, the paper points out that the main factors holding back research are a lack of access to real handset data, followed by the need to develop custom software and systems in conjunction with the carriers.

The Milan data, covering a period of sixteen days, is used by Pulselli *et al.* (2005) and Pulselli *et al.* (2006) to demonstrate that cities behave like dynamic ecosystems, constantly changing and reacting to stimuli such as disasters, soccer matches, and street closures. According to the papers, the data "enables collective behaviours, perturbations and effects of fluctuations to be studied in complex urban systems with respect to [their] theoretical framework" (Pulselli *et al.* 2006, p 132). For example, it clearly showed that call activity in the morning is greatest in residential areas, moving to the central business district where it is especially busy around lunch time, then concentrated around transport hubs such as the main railway station between 5-6pm.

Like the Graz study, the Real Time Rome project (Calabrese & Ratti 2006) was designed as an art exhibit, in this case for the 10th International Architecture Exhibition in Venice, Italy. As with Graz and Milan it displayed Telecom Italia call volume information, but it also added GPS traffic data from buses and taxis, traffic noise from sensors placed around Rome, and more accurate call-in-progress data from selected GSM (Global System for Mobile) base station controllers. Using Telecom Italia's "Lochness" platform, the call-in-progress data combined the call's cell ID, angle of arrival, timing advance, and signal strength to provide a much more accurate location than could be obtained from just a cell ID. The call-in-progress data also included country-of-origin information, allowing "tourists" to be distinguished from locals.

The data collected for Rome was analyzed in a number of different ways. Initially Reades *et al.* (2007) used normalization of sites over space and time to identify shifts in the relative intensity of activity across Rome. The call intensity at each pixel was normalized against every other pixel in Rome at the same point in time, to produce values normalized in space. These values were then normalized against their average over time. These doubly-normalized results showed significant differences in activity by time of day and between weekdays and weekends. The paper then used a clustering algorithm to group together cells with similar patterns of activity. Working with eight different clusters they were able to divide Rome into regions that generally corresponded to distinct activities, e.g. urban core vs residential vs

entertainment districts.

The approach used by Sevtsuk & Ratti (2010) was to apply a Fast Fourier Transform to the data to measure any periodic behaviour. As would be expected, the strongest frequency was the 24-hour cycle indicating daily patterns, although the strength of that peak varied strongly from cell to cell. The second strongest frequency was the 3.5-day cycle, which distinguishes weekdays from weekends.

The focus in Rojas *et al.* (2007) was on the strategies applied to visualizing the data. Six screens were used in the Real Time Rome exhibit, each displaying a map with data overlays addressing a different question - “Where in Rome do people converge over the course of a day?”, “Is public transport where the people are?”, “Where is traffic moving?”, “Which landmarks in Rome attract more people”, “Where do tourists congregate?”, and “What does Rome look like during special events?”. The paper found that different visualization techniques were needed depending on who the end user was and what question they were asking, and that in some cases text, pictures, and graphs may be better than maps for presenting the data.

In Pulselli *et al.* (2008), data from Pescara, on the Adriatic coast of Italy, is compared to the data from Milan. They find that there is much more seasonal variation in calling patterns in Pescara, probably owing to the large influx of tourists every summer.

2.3 Cell accuracy

Only a couple of studies could be found that experimentally measured the spatial accuracy of handset locations derived from cell IDs. Raja *et al.* (2004) used a Nokia 7210 running software called NetMonitor to explicitly connect to three different cells while stationary in Edinburgh, Scotland, and used a technique called Timing Advance to estimate the handset's distance from each antenna. With this method, they were able to use triangulation to determine its position with an accuracy of 310 metres. However, this information was only available to the handset, not the network, so the technique would be of no use for population tracking.

The paper does, however, provide cell ID accuracies, but without explaining how they were obtained. Rural accuracies were given as 1-35km, typically 15km; suburban cells 1-10km, typically 5km; urban 0.5-5km, typically 2km; micro-cells 50-500m, typically 200m; and indoor pico-cells 50-500m. It is not clear whether these values are specific to Edinburgh or

apply everywhere.

A more comprehensive experiment was carried out by Trevisani & Vitaletti (2004), who collected 8915 cell ID and GPS samples around Rome and New York. These samples were collected at two second intervals, and around 7300 provided useful data. With the cooperation of local phone companies they were able to associate the cell IDs with the latitude and longitude of the cell's antenna, and by comparing this with the GPS data were able to estimate the accuracy of cell ID positions. They found the average accuracy across Rome was around 500 metres, and 800 metres across New York. Their numbers are broken down by urban, suburban, and highway regions in table 2.1.

Region	**Error (km)**
Italy, urban	0.48
Italy, suburban	0.75
Italy, highway	1
US, urban	0.79
US, suburban	0.49
US, highway	2.91

Table 2.1: Average cell ID accuracy, by region (Trevisani & Vitaletti 2004).

In addition, they found that handsets used the cell of the closest antenna only 63 percent of the time, decreasing to 57 percent in urban Italy. In fact, the handset used antennae that were on average 150 metres further away than the the closest one. The reasons given for this were multipath propagation (i.e. reflections off obstacles), differences in transmitting power between antennae, and cell selection algorithms that may delay changing to new, stronger, cells to reduce signalling overheads.

2.4 Tracking using billing records

A number of population tracking studies have been carried out using the billing records collected by mobile phone network providers (usually referred to as “carriers”). Because a carrier needs the ability to carry out distance-based billing, every call and SMS sent or received generates a record containing the cell ID where it took place, allowing the phone's approximate location to be determined.

Candia *et al.* (2008) analysed the call activity patterns of six million handsets from an unnamed country using a month's worth of call records (consisting of time, cell ID, and an anonymized handset identifier). The analysis found a correlation between between number of calls made and distance travelled, with peaks in both occurring around noon and early evening, and both declining at night. It also showed how call behaviour follows fairly predictable patterns, allowing anomalous events – such as emergency situation – to be automatically identified by their variation from the mean behaviour.

In Ahas *et al.* (2007) the researchers used mobile phone call event data (call in/out, SMS in/out, internet access) to monitor tourist numbers in different parts of Estonia, and show how these vary based on the time of year and the visitor's country of origin. They had access to a database of records providing a time stamp, a unique anonymous handset identifier, the handset's country of origin, and the latitude and longitude of a cell antenna. With over 9.2 million records, generated by 720,000 different handsets, they were able to generate a comprehensive report on tourist destinations throughout Estonia over a twelve month period. When this call activity, broken down by country of origin, was compared with official accommodation statistics, they found a 0.97 correlation coefficient, indicating a very high accuracy. The only drawback was the reliance on call event data, so visitors who didn't use their phone while in Estonia were not recorded.

The Estonian tourist mobile phone location data was later made available through a commercial web-based product called *Positium Barometer*, described in Ahas *et al.* (2008). This paper concentrated on the practical uses of the data: who needs it, why they need it, and how the data can be presented to meet those needs. In terms of *why*, they identified five applications where the data could be used: Strategic planning; Business and investment plans, funding applications; Marketing and advertising; Real-time business management; and Public administration. In terms of *how* it could be presented they identified four categories: Statistics (including changes in numbers); Space-time movement analyses; Event modelling; and Real-time monitoring. In a 2-by-2 matrix the authors show the relevance of each category of data to each application, and this information was used to help design the web-based product. The product itself seems well thought out, with report generation tools capable of generating bar charts, pie charts, and colour-coded maps. It is, however, limited to using data from foreign handsets, perhaps due to privacy concerns. The authors do not speculate on the ways that domestic handset data could be used, but it seems likely that the five applications they identified would be relevant.

An interesting use of billing records was described by Song *et al.* (2010), where they were used to measure the predictability of people's movements. Pointing out that understanding movement patterns is important for applications such as city planning and understanding the spread of infectious diseases, the study found that 93 percent of human movement could be predicted based on previous movement patterns, regardless of the distances travelled by the person.

Although the study had access to the billing records of roughly 10 million users, the need to extract accurate journey information meant that only users who made at least one call every two hours were used. This reduced the number of candidate users to only 50,000, which highlights one of the main drawbacks of billing data – the limited number of location records available for infrequently-used handsets. Not only does this result in incomplete data for a large part of the population, it could also generate biased results if frequent callers have different movement patterns to infrequent callers.

A real-world example was described by Bengtsson *et al.* (2010), where mobile phone billing data was used to monitor population movements in the aftermath of the Haiti earthquake on 12 January 2010. Using anonymized data from two million Digicel subscribers over the period 1 January to 11 March 2010, the researchers were able to estimate the number of people leaving Port-au-Prince in the weeks following the earthquake (22 percent of the population), and where they went. They were then able to estimate the number returning over the following month (41 percent by 11 March).

Although the Haiti results have not been validated by other studies, the authors believe that their estimates based on billing data are the most accurate numbers available. However, they do acknowledge that their results are based on the assumption that mobile phone subscribers have the same movement patterns as non-users. This is still an open question and an area for further research.

2.5 Active tracking

While passive tracking makes use of data that is already being collected for other purposes (e.g. billing records), active tracking uses signals that are explicitly sent for the purpose of tracking handsets. Dufkova *et al.* (2008) go into some detail on the mechanics of this when discussing their SS7Tracker platform. They describe the SS7 Mobile Application Part (MAP) messages that are sent and the responses that are monitored by their equipment, and how these

can be used to retrieve the current cell ID of any phone on the network. Their main intent is to detect the parts of the network where roaming handsets tend to drop out, since roaming handsets are apparently very profitable for carriers and they want to keep them on their network for as long as possible, but their technique will work for other applications as well.

In their discussion of the benefits of active versus passive tracking, they point out that active tracking can provide sample data at an arbitrary predetermined rate, whereas with passive tracking obtains samples somewhat randomly. On the other hand, active tracking places extra traffic load on the network and drains power from the handset. This load was not an issue in the case study, which only tracked 247 roaming handsets in the Czech Republic over seven hours, but it is difficult to see how it could scale up to track all subscribers on a network. Note that in this paper the definition of passive tracking is billing record analysis, which only generates samples when a call or SMS is sent/received. The authors did not consider the option of passively monitoring location update messages, which are sent from handsets to the network on a regular basis.

2.6 Road traffic monitoring

There have been a number of studies on the use of mobile phones to monitor road traffic, many of which are reviewed in Rose (2006). According to this paper, the type of information that traffic engineers would like to collect falls into three broad categories: real-time traffic volumes and speeds; journey origin-destination information; and real-time identification of traffic jams and accidents. The paper compares data collected from mobile phones to that captured by fixed traffic sensors such as induction loops, cameras, and tollway sensors, a finds a number of benefits. It finds that mobile phones provide nearly universal coverage of a region's vehicles at a low fixed cost, whereas fixed sensors are an additional cost that scales with the number of measurement locations. On the other hand, mobile phones are less spatially accurate than fixed sensors, and can sometimes provide incorrect numbers because they are switched off by their owners when driving or there is more than one phone in the vehicle. Also, the most spatially accurate mobile tracking techniques tend to use active querying, where the location of a particular handset is queried, which places additional load on the network and handsets. But in general the paper is unable to provide much additional detail on the techniques used or their accuracy due to commercial confidentiality affecting much of the research in this area.

White & Wells (2002) and Caceres *et al.* (2007) discuss the use of mobile phone location data

to derive origin-destination (OD) matrices for vehicular traffic. Both papers conclude that the technique has enormous potential in terms of lower cost and higher accuracy versus other methods, such as household surveys and roadside monitoring, but were hampered by a lack of suitable location data from a carrier. White & Wells (2002) showed that cell IDs from billing records could partially reproduce a known OD matrix (obtained from a roadside survey in Kent), but that several days worth of data may be needed to complete the picture due to the infrequent rate of phone calls. More accurate OD information could probably be retrieved from the passive scanning of handset location updates, but that data was not available from the carrier involved.

Similarly, Caceres *et al.* (2007) were unable to access real passively-scanned location update data, so they used a simulated GSM network to generate it, in particular a stretch of highway between Seville and Huelva in Spain. The study found that the traffic count values calculated from the simulated data were accurate to within a few percent, with the main source of error being the estimate of the number of handsets per vehicle subscribing to a particular carrier. The authors estimated this number based on the mobile phone penetration rate, the carrier's market share, and the average number of occupants per vehicle. Their simulation generated random handset data with a distribution based on this estimate, and the same number was then used to scale up the data to convert the handset count back into a vehicle count. The errors in this case were due to the deliberate random variations of the simulated data, and as the simulations grew larger the average error grew smaller. However, with real-world data, another source of error would be the handset-per-vehicle number itself, which may be incorrect or vary by location or time of day.

A press release for Optus Traffic View (OTV), a service which has been available since June 2009, indicates that mobile phone data can be used to generate “high quality travel time and congestion information on every major highway and freeway” (Withers 2009). Using Cellular Floating Vehicle Data to cover 70,000km of Australian roads, OTV is not as accurate as GPS data, but “the larger number of mobile phones on the road can compensate for the relative inaccuracy of cell-based location once knowledge of the road network and the right statistical techniques are applied”. These statements suggest that their system uses knowledge about where cells intersect with roads to generate locations with more accuracy than cell boundaries alone, but it is unclear whether the service uses billing data, active polling of handsets, or some other technique.

2.7 Visualization

The GeoPKDD project (Geographic Privacy-aware Knowledge Discovery and Delivery) was a three-year European Commission initiative to “develop theory, techniques and systems for geographic knowledge discovery, based on new privacy-preserving methods for extracting knowledge from large amounts of raw data referenced in space and time” (GeoPKDD 2010). Running from December 2005 to November 2008, the project investigated the state of the art in the areas of collecting mobility data, mobility data mining (discovering useful information in large repositories of mobility data), and preserving privacy.

The area of mobility data mining was broken into two areas: knowledge discovery and knowledge delivery. When it came to knowledge delivery the project found that “Extracted patterns are seldom geographic knowledge *prêt-à-porter*: It is necessary to reason on patterns and on pertinent background knowledge, evaluate patterns' interestingness, refer them to geographic information and find out appropriate presentations and visualizations” (Gianotti & Pedreschi 2008, p 7). In other words, the results of knowledge discovery are, on their own, rarely sufficient for most uses, and need to be presented in a form suitable for evaluation and possible re-generation.

To this end, the GeoPKDD project focused the field of Visual Analytics, “the science of analytical reasoning facilitated by interactive visual interfaces” (Thomas & Cook 2005, p 28). It is worth noting that visual analytics is not the same as visualization, rather it is an interactive process that *makes use of* visualization, “synergistically” with a human analyst.

Andrienko *et al.* (2008) are of the opinion that visualization on its own, although playing an important role in analysis, is not sufficient, and that

> The challenge is to build analytical tools and environments where the power of computational methods is synergistically combined with man's background knowledge, flexible thinking, imagination, and capacity for insight.

If such a view is correct, it implies the need for a toolkit of visualization algorithms, in particular fast algorithms that can rapidly display different aspects of the data. However, even if this view is not correct and there is a single visualization algorithm that meets all requirements, such an algorithm has yet to be found. In the meantime it is worth reviewing the available techniques.

2.8 Visualizing static data

Although mobile phone location data occasionally includes demographic information such as age and gender (Yuan & Raubal 2010), in general it provides only an anonymous unique identifier. Under these circumstances, visualizing the location data at a point in time is akin to displaying population counts, with a spatial resolution the size of mobile phone cells.

When it comes to displaying static population counts – and other static population attributes – there are a number of well-established techniques in the field of cartography. Researchers have applied many of these to the display of data collected from mobile phone networks.

Choropleth maps use colour, shading, or texture to indicate the value of an attribute that varies spatially, such as population density. For example, figure 2.1 shows call activity levels across Milan using a colour spectrum. Choropleth maps are limited in the number of distinct values they can display because of the inability of the human eye to distinguish between slight changes in colour and intensity (Gregory & Ell 2007, Muehrcke *et al.* 1998), but they do provide an effective overview of an attribute's spatial distribution.

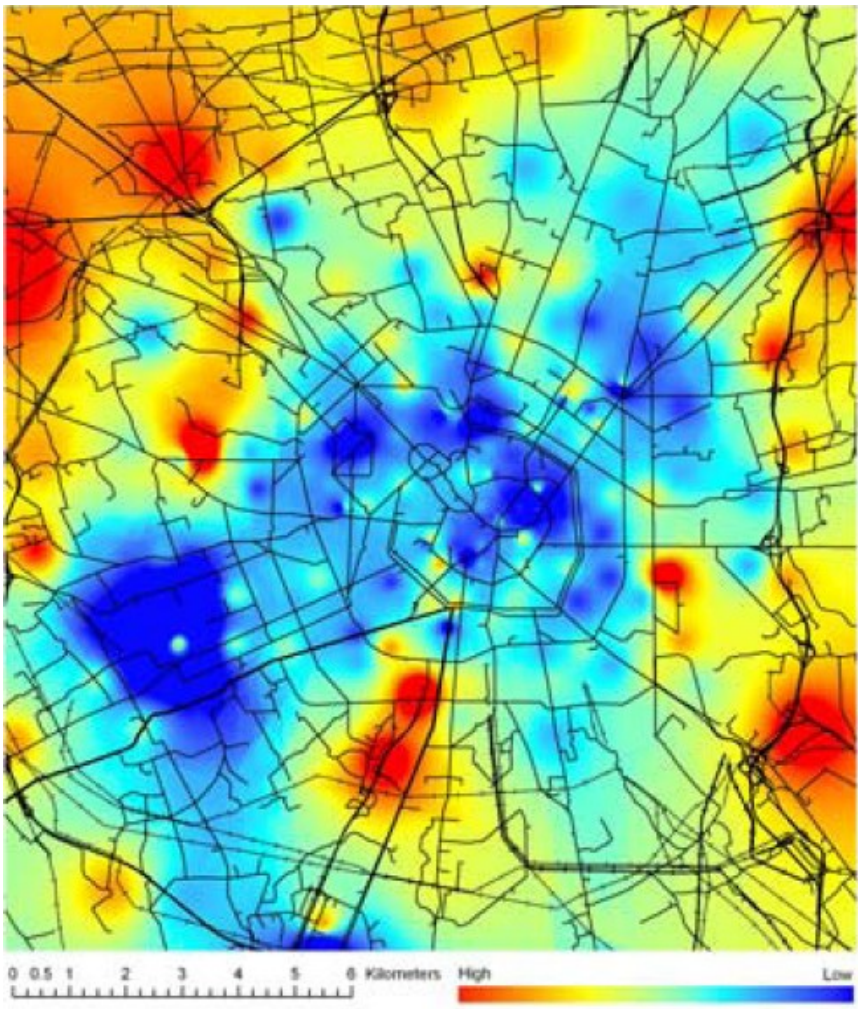

Figure 2.1: Choropleth map of call activity in Milan (Ratti *et al.* 2006).

A greater number of distinct attribute values can be displayed with *isopleth maps* (Muehrcke *et al.* 1998), which use contour lines to indicate the boundaries between regions with different values. An improvement of the basic isopleth map is the *shaded isopleth map*, which colours or shades the regions between the contour lines just like a choropleth map. Muehrcke *et al.* (1998) find that this combines the benefits of choropleth and isopleth maps, providing both a good overview of the attribute's distribution and an accurate representation of its value at each location. However, no examples could be found of isopleth maps being used to display mobile phone data.

Another way that spatial attributes can be displayed is through the use of *area cartograms*, essentially maps that distort the size of regions to represent the values of an attribute. However, for cartograms to be effective, Muehrcke *et al.* (1998) and Gregory & Ell (2007)

point out that the viewer must first be familiar with the original size and shape of the regions being distorted. While this is usually the case for regions such as countries and states, it is unlikely that any viewers will be familiar with mobile phone cell boundaries, so cartograms are probably not suitable for displaying mobile phone data.

Three dimensional views are another popular way to display statistical values. Typically they involve an oblique view of the 2D map, with either a stereoscopic (true) perspective or a parallel (orthogonal) one (Muehrcke *et al.* 1998). The statistical value is then represented by the "height" of a symbol or surface drawn over the map. Stereoscopic perspectives have traditionally been used because they look natural, but orthogonal perspectives have the advantage of displaying identical statistical values at the same height, regardless of location, allowing for easier comparison between values in an image (Shepherd 2008). Thus the choice of perspective depends on the application.

Figure 2.2 (Ratti *et al.* 2006) uses a 3D view to show the spatial distribution of cells across Milan and their call activity at a point in time. The vertical lines show the location of each individual cell, and their lengths indicate the level of call activity. This use of vertical lines has the benefit of concisely displaying the data, which consists of point locations and a value at each. With careful analysis, both the locations and the values can be accurately determined from the image.

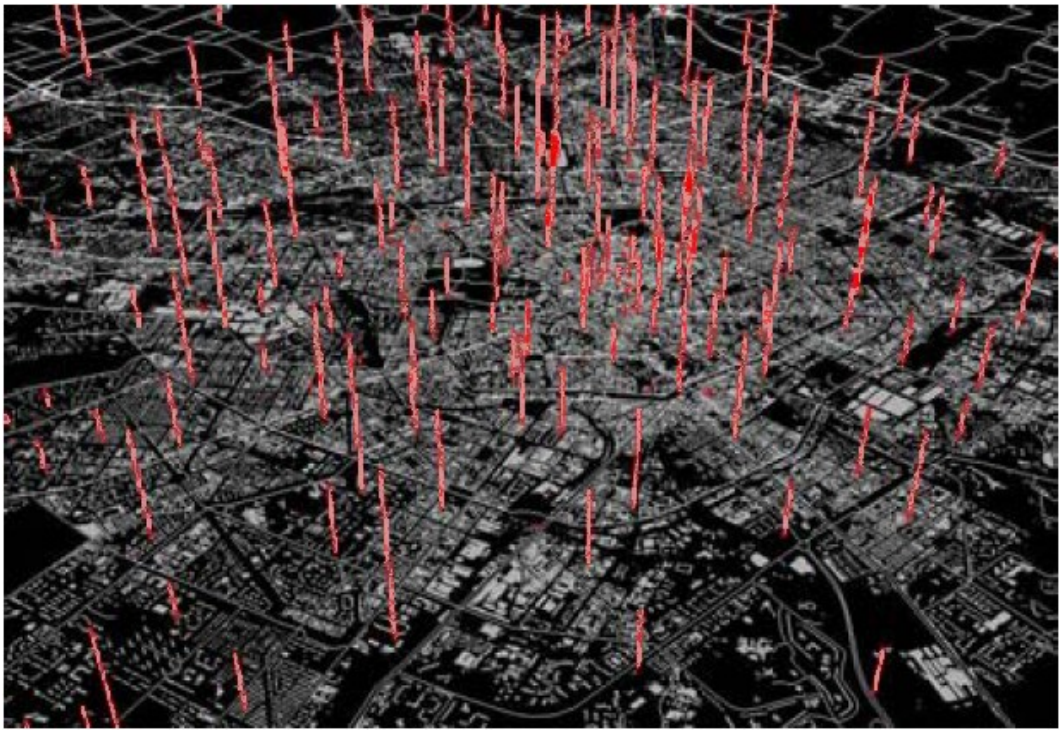

Figure 2.2: Cells in Milan and their call activity at a point in time (Ratti *et al.* 2006).

Another benefit of displaying symbols – be they vertical lines, cylinders, or some other vertical structure – is that they can represent multiple variables. For example, Shepherd (2008) provides numerous examples of stacked 3D symbols, where a multi-colour or multi-width cylinder displays the values of multiple variables. In a few cases red-green cylinders are used to show the number of occupants at numerous location, and the cylinders are split vertically into male (red) and green (female) to show the numbers of each. Such a technique could be used when mobile phone data contains gender information, e.g. Yuan & Raubal (2010), although it may not be suitable for representing age information.

Figure 2.3 (Ratti *et al.* 2006) shows the same data as figure 2.2, but interpolates between the cell centres to create a continuous statistical surface. Statistical surfaces have been found to be visually impressive, but they are not always suitable for detailed analysis. For example, Muehrcke *et al.* (1998) find that "although highs and lows are apparent, the exact height of the surface at a given location is difficult to determine" (Muehrcke *et al.* 1998, p 147). Of course, it should be noted that Muehrcke was referring to black and white images, and a coloured statistical surface may be no more difficult to interpret than a 2D choropleth map.

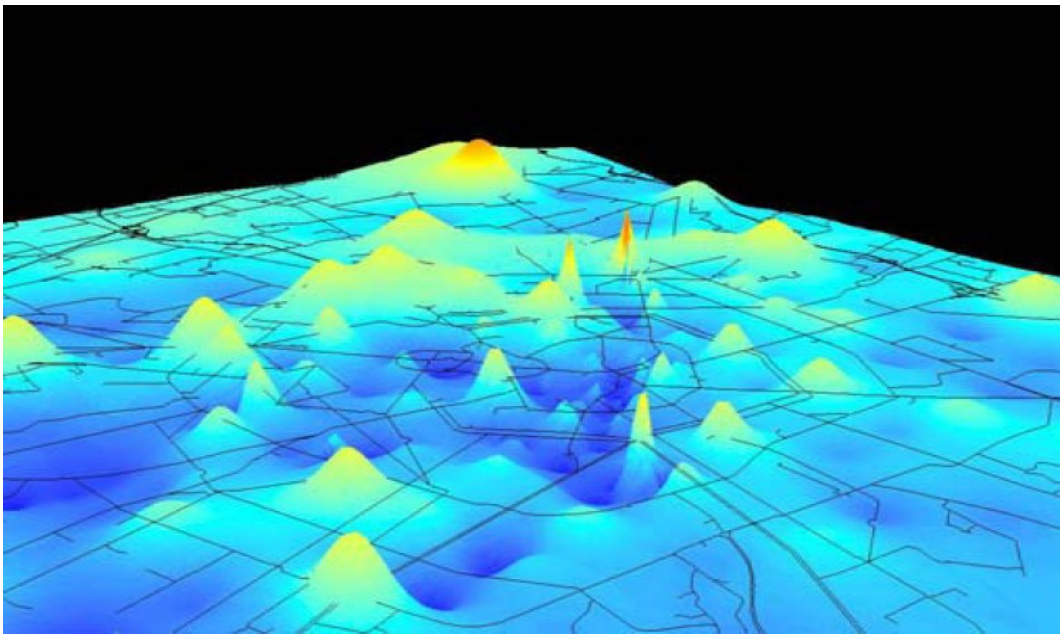

Figure 2.3: Call activity in Milan, interpolated between cell centres (Ratti *et al.* 2006).

When a suitable computer is available, Shepherd (2008) has found *interactive* 3D displays to be a more effective way of visualizing spatially-distributed values than static printed images. The ability to dynamically change perspective improves the viewer's ability to perceive depth and allows them to look around objects that are concealing information behind them. However, such a computer may not always be available, nor may the skills needed to interact with a 3D image, so an interactive approach should only be used when those things are known to be available.

2.9 Visualizing spatio-temporal data

Mobile phone location data falls in the category of *spatio-temporal* data, or spatial data that varies with time. A number of techniques exist for mapping this type of data, with Langran (1992) identifying four major classes -

- Time sequences, i.e. multiple images representing different times.
- Animations, i.e. time sequence images displayed one after another.
- Static maps with thematic symbols of a temporal theme, e.g. symbols depicting dates, rates, paths, or order of occurrence.
- Change data, i.e. text, graphic, or digital amendments to a base representation.

However, Langran concluded that the field was not yet mature, and that “while powerful in their raw form, the full potential of static and dynamic temporal maps has not yet been explored systematically by cartographers” (Langran 1992m p 24). This view was echoed by Muehrcke *et al.* (1998), who found that cartographers have focused primarily on the representation of static phenomena, “as if time were absent” (Muehrcke *et al.* 1998, p 160).

In contrast, the non-cartographic world has a number of well-established techniques for visualizing variables that change with time. For example, in figure 2.4 (Ratti *et al.* 2006) a cartesian graph is used to display call intensity (measured in Erlang, or the average number of concurrent calls) versus time-of-day for 14 different cells across the Milan metropolitan area, using 14 lines of different colours. At a glance, it demonstrates how call volumes drop to very low levels between midnight and 6am across the Milan metropolitan area.

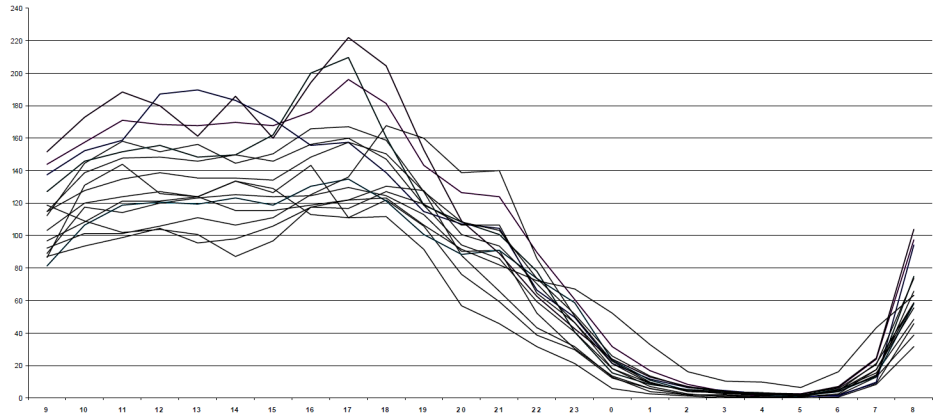

Figure 2.4: Milan call activity, in Erlang, throughout the day (Ratti *et al.* 2006).

As well as variants of cartesian graphs, other techniques include Gantt charts, flow charts, PERT charts, time lines, 3D cartesian graphs, graphical timetables, and bespoke visualizations such as Minard's representation of Napolean's Russian campaign of 1812 (Harris 1999, Tufte 1990).

Of these, Gantt charts and graphical timetables offer the most potential for representing mobile phone data for use in visual analytics. Not only can they be generated automatically, but they also display some spatial information. In both cases, space is collapsed to a single dimension along the Y axis, with a row allocated for each distinct location, while time is shown along the X axis. However, each location/row needs a distinct label, which may not always be available with mobile data.

2.10 Time series and animations

A simple way to show how spatial data varies with time is to produce a number of static maps, each showing the data at a point in time, known as a *time series*. These static maps could be choropleth, isopleth, 3D, or whatever representation is suitable for the data. For example, figure 2.1 is actually the first in a sequence of six images showing how mobile phone call density varies across Milan between 9am and 1pm (Ratti *et al.* 2006).

Such a representation is useful for the analysis of dynamic phenomena, but space often limits the number of images that can be presented. There are also limits to the ability of a viewer to detect differences in images that are laid side-by-side (Muehrcke *et al.* 1998).

When the differences of interest involve changes in magnitude rather than changes in location, Muehrcke *et al.* (1998) suggest the use of *quantitative change maps*, which show the direction and magnitude of change at a point in time. These are simply static maps that display the derivative of the data with respect to time, showing positive values when the magnitude is increasing and negative when it is decreasing. Care needs to be taken when choosing the time range, and whether to display relative or absolute changes, but these maps could provide a useful view of mobile phone data.

When a time series of static maps is displayed rapidly one after another, the result is an *animation*. The changes between individual images must be slight to produce the illusion of continuous movement (Muehrcke *et al.* 1998), but given the ability of the human eye to process moving images, animation can reveal patterns and insights are not noticed in static representations (Harrower & Fabrikant 2008).

However, animation is not suitable for all spatio-temporal data. Harrower & Fabrikant (2008) found that it is not suitable for displaying instantaneous events, such as house sales, because the information flickers in and out of existence too quickly to be understood.

Gregory & Ell (2007) identified a few other problems with animation. First, because each individual image is displayed so briefly, the data must simplified so the viewer can understand it quickly. This limits the number of variables that can be displayed. Second, animations can only be published electronically, preventing them from appearing in books and journals, and limiting their availability. And finally, there is a tendency for people who produce animations to focus on what looks good rather than what provides useful information, and this needs to be consciously resisted.

An example of animating mobile phone data can be seen in “Wake Up San Diego” (Airsage 2009b), which uses a sequence of choropleth maps to show mobile phone use across San Diego. A more sophisticated animation is “Obama Inauguration: The Power of Location Data” (Airsage 2009a), a sequence of annotated 3D maps with a musical soundtrack showing cell activity in the Washington DC area on the day of Barack Obama's inauguration (see figure 2.5 for a sample image).

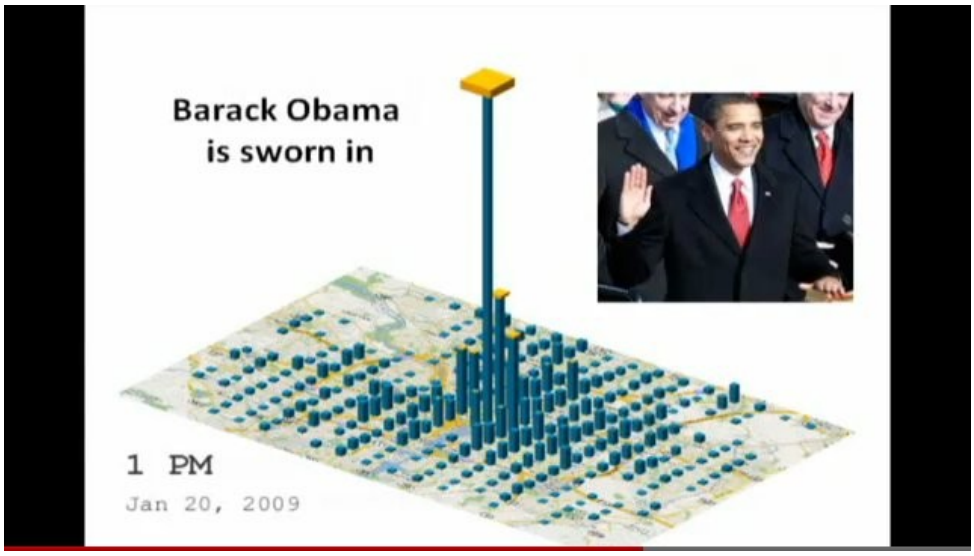

Figure 2.5: Animation of cell activity in Washington DC (Airsage 2009a).

It is interesting to note that the vertical bars used in the Obama animation are laid out in a grid pattern, which almost certainly does not correlate with the layout of cells in Washington DC. This implies that the bars are not showing actual activity levels, but values calculated by interpolating between cells. The decision to use interpolation in this case may have resulted in some loss of accuracy in the representation of activity levels, but the trade-off is an image that is easier to interpret because the bars are evenly-spaced and represent equal-sized areas.

2.11 Visualizing movement – routes and trajectories

Although animation is an effective way to view changes in spatially-distributed values such as handset density, simply displaying a sequence of static activity maps does not provide any information about the movement of individual handsets. For example, an animation may show an increase in handsets in a city centre during a workday, but it doesn't show where the handsets are coming from or how fast they are moving.

In some cases, such as the Milan study by Ratti *et al.* (2006), individual handset information

wasn't available, so a sequence of call activity images was an appropriate format for visualization. But in many cases mobile phone data contains handset identifiers, which is enough information to construct a *trajectory* for each handset, where a trajectory is defined as "a sequence of time-stamped locations" (Rinzivillo *et al.* 2008, p 225). Similar to a trajectory, mobile phone data can also be used to construct a *route*, a sequence of locations but without the timestamps.

One way to visualize trajectories is with a Hägerstrand space-time cube (Hägerstrand 1970) (also known as a space-time prism or space-time aquarium). A space-time cube displays movement using a line drawn in three-dimensional space, as shown in figure 2.6, with the X and Y axes representing location and the Z axis representing time. The image displays location and time information for an individual, which can also be used to estimate speed and direction. Sometimes the line itself is colour-coded to show additional information about the individual that changes with time (Kwan 2003).

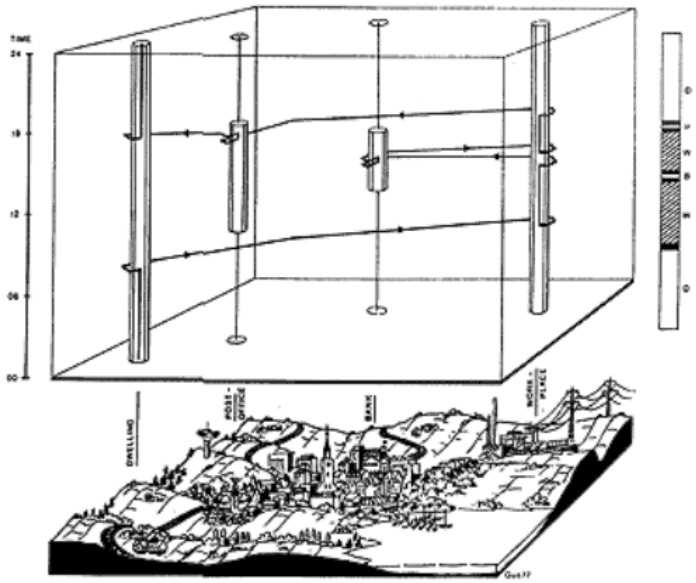

Figure 2.6: An individual's path in a space-time cube (Neumann 2005).

The use of a line in three dimensions concisely depicts an individual's movements through space and time, but it is often difficult to relate a point on the line to a corresponding position on the underlying map. This can be overcome somewhat through the use of vertical structures rising from important landmarks, or by projecting a "shadow" or "footprint" of the line onto the map (Kraak 2008).

Being a three-dimensional structure, space-time cubes are easier to interpret with an interactive computer interface than with a static two-dimensional representation. The ability to rotate and zoom the structure allows the user to better understand the shape and select the best view (Kraak 2008), but as with animation this requires electronic distribution, which limits availability.

Although effective at visualizing the movements of an individual, the main problem with using space-time prisms for mobile phone data is that they doesn't scale well. Kwan finds that "Although the 'aquarium' is a valuable representational device, interpretation of patterns becomes difficult as the number of paths increases with the number of individuals examined" (Kwan 2000, p 197). In other words, when the paths of more than a handful of individuals are displayed, the resulting image becomes too difficult to interpret. For a city with paths for hundreds of thousands, or even millions, of individuals, a space-time cube is clearly not suitable.

2.11.1 Route clustering

To overcome the problem of drawing large numbers of routes or trajectories, the approach taken by a number of researchers (Andrienko *et al.* 2007, Andrienko & Andrienko 2008, Rinzivillo *et al.* 2008) was to reduce their numbers by clustering similar trajectories and displaying them with thicker lines. Any lines below a certain thickness were treated as "noise" and deleted.

The first step in the clustering process was to define a *distance function* that measures the similarity between two trajectories. Different distance functions were used to explore different aspects of the trajectories, allowing an analyst to gain a better understanding of the data.

However, the clustering process was very computationally intensive. A dataset of 176,000 trajectories was obtained from 17,241 GPS-tracked vehicles over a week in Milan (Andrienko & Andrienko 2008, Rinzivillo *et al.* 2008), but this was too big to load and process in main memory. As a result, only 6187 trajectories from a four-hour period were used. Without further optimization, such a process would not be able to deal with the volumes of data generated by mobile phone data.

Trajectory clustering also ignores a lot of information. In the Milan data, regardless of the distance function used there was very little commonality among the trajectories. Between 77 and 90 percent of the trajectories had no frequent routes among them and were discarded as

"noise" (Rinzivillo *et al.* 2008).

Finally, the trajectory clustering studies all made use of frequently-updated GPS data, with sample intervals of less than a minute. This was accurate enough to determine transport routes, including individual roads. Mobile phone data will usually not be that accurate.

2.12 Visualizing movement - flow maps

For individuals moving between *N* pre-defined regions, Tobler (1987) describes a form of visualization known as a *flow map* (see figure 2.7 for an example). Several variations of flow maps are described, using United States internal migration data as an example.

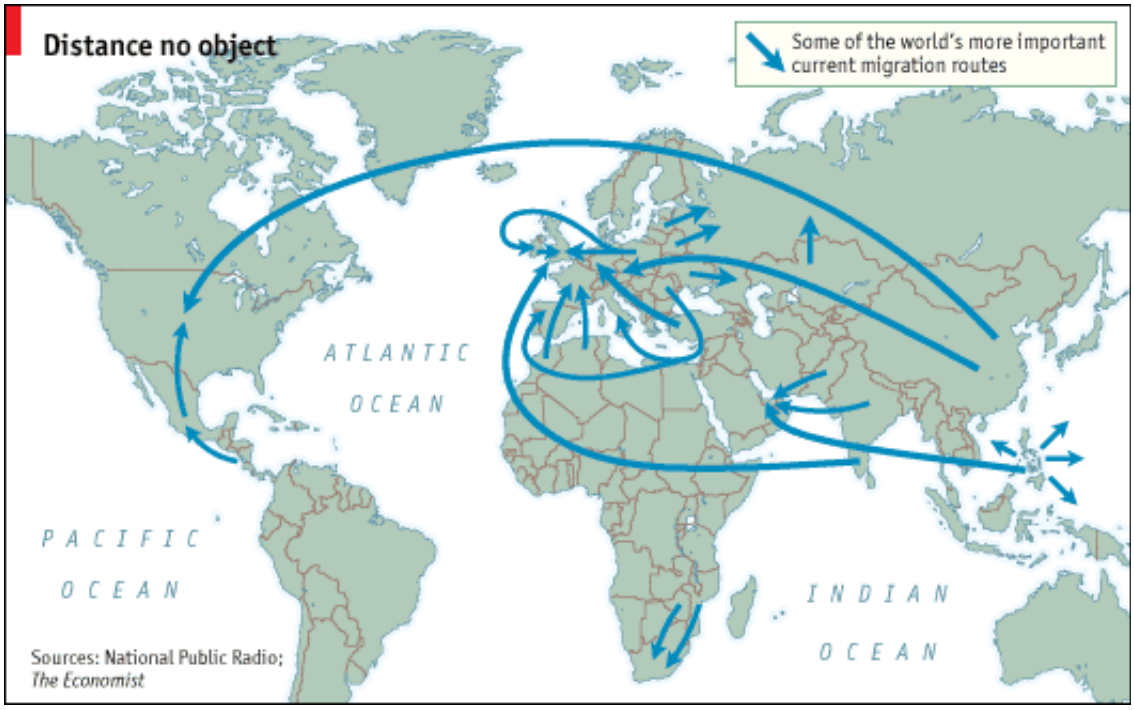

Figure 2.7: Visualization of migration routes using a flow map.

The problem with flow maps is that mobile phone users generally do not move en-mass between a small number of clearly defined regions, so when a population is spread throughout a city's suburbs it is unclear where the arrows would be drawn. There is also the difficulty of representing the many individuals whose journeys are round-trip, starting and ending at home.

Discrete flow maps indicate movement using arrows between region centroids, often with the width of the line indicating the number of individuals. By default, Tobler (1987) recommends that line widths are scaled so that the maximum width equals the distance between the two closest centroids, with all other widths scaled down accordingly.

When this default width doesn't work, Tobler suggests a process of trial-and-error using manually-specified maximum widths. However, given the enormous increase in computing power since the paper was originally written, iterations of a trial-and-error process would now occur in a fraction of a second, so there is no reason why this should not be the standard process, perhaps using arrow widths based on the closest regions as a starting point only.

Other ways of indicating flow volumes were considered, such as varying the length, colour, pattern, and shading intensity of arrows. But based on the author's admittedly unscientific impressions, “not supported by real evidence” (Tobler 1987, p 157), these methods are interpreted less accurately than varying arrow widths proportional to flow volumes, and were not used.

Similarly, Muehrcke *et al.* (1998) find that variable-width flow lines are more effective. Constant-width lines that vary in texture, colour, or intensity need to be reasonably wide before a viewer can detect small changes, and this can clutter up the map. Another option is to draw multiple constant-width thin lines along a flow, each representing a quantum of volume. However, this was found to take up as much space as a single variable-width line, so doesn't offer a clear advantage.

The main drawback to variable-width flow lines was found to be the inability of the human eye to distinguish slight differences in width (Muehrcke *et al.* 1998). A proposed solution was to divide the volumes into a set of distinct ranges, each represented by a clearly-distinguishable line width.

2.12.1 Reducing visual clutter

When dealing with an area divided into N regions, there are NxN possible migration flows, or Nx(N-1) if only flows to different regions are considered. For moderately large values of N, representing all these flows as arrows produces an unacceptable level of clutter, so Tobler (1987) discusses some techniques for reducing them, using inter-state population movements within the United States as a case study.

He found that deleting all flows whose volumes fall below the mean volume was a very effective rule, deleting over 75 percent of the arrows while removing less than 25 percent of the total volume. This was successful because inter-state migration volumes within the United States were found to follow a power law distribution, with a small number of routes accounting for the majority of the flow volumes. However, it is not clear if this is true for all

population movements, so the rule should be applied with caution.

A second approach to reducing the number of arrows was to merge adjacent regions, effectively reducing the value of *N*. The problem with this approach was that the United States has a large number of small, densely-populated states in the north-east, and merging them conceals some important flows between them. However, it may be appropriate in situations where region sizes and population densities are more homogeneous.

The final approach suggested by Tobler (1987) for reducing arrow clutter was to only display flows between adjacent regions. For example, when displaying migration between Australian states, the flow between Victoria and Western Australia (two non-adjacent states) could be shown as two separate flows between Victoria and South Australia, and between South Australia and Western Australia.

Such an approach was applied to the display of GPS location data from Milan (Andrienko & Andrienko 2008, Rinzivillo *et al.* 2008), where "aggregate moves" between adjacent "significant places" were displayed with arrows of varying widths. These "significant places" tended to be road intersections, and although there was an attempt to locate these places automatically, it was largely unsuccessful and further work is needed.

Although the approach of displaying flows between adjacent regions discards origin and destination information, it does have the benefit of scaling up to large values of *N*. In fact, Tobler describes the case where *N* approaches infinity, whereupon a map becomes a *continuous flow map*, as shown in figure 2.8, with a "depiction more analogous to those used in fluid dynamics" (Tobler 1987, p 156).

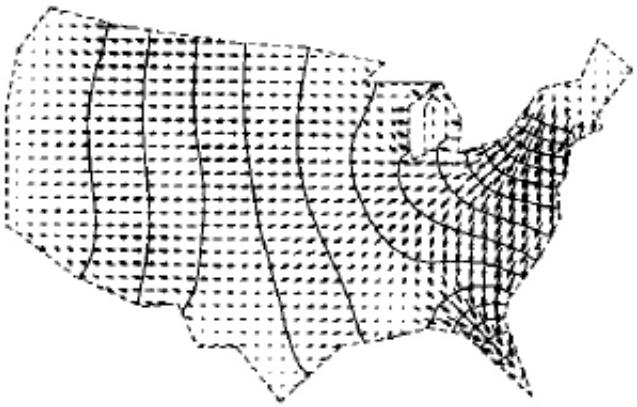

Figure 2.8: A continuous flow map (Tobler 1987).

However, a limitation of the map shown in figure 2.8 is that it only displays *net* flows. While this may be suitable for planning, say, infrastructure based on changes in population, it discards information that could be useful for transport planning since strong bi-directional flows would cancel out.

In fact Tobler (1987) struggles to deal with the representation of bi-directional flows. Two options were considered -

- Drawing the arrows on top of each other, with the thinner arrow on top.
- Drawing “half-barbed arrows”, as shown in figure 2.9.

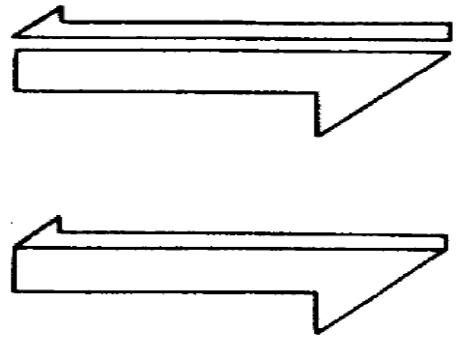

Figure 2.9: "Half-barbed" arrows representing bi-directional flows (Tobler 1987).

These representations were found to “not seem very effective visually” (Tobler 1987, p 157), although half-barbed arrows separated by a small distance (figure 2.9, top) have been used elsewhere with acceptable results (Andrienko *et al.* 2008). However, it is not clear why Tobler did not consider the option of simply drawing two full arrows side-by-side, which would appear to be a very simple and effective solution to the problem.

Rather than displaying *net* flows on a continuous flow map, Andrienko & Andrienko (2008) use *directional bar diagrams* to show the flow volume in eight directions, similar to a “wind rose” used by meteorologists to show cumulative wind direction. Displaying vehicle movements in Milan, the map contains a grid of these diagrams, each consisting of up to eight bars. Each bar points in a direction of travel, and is also colour coded according to the direction. The length of each bar is proportional to the number of cars moving in that direction, and the diagram is overlaid with a grey circle whose radius is proportional to the

number of stationary vehicles.

A similar representation of multiple flows at a single location is shown in figure 2.10 (Andrienko *et al.* 2008). Like a directional bar diagram, this figure is designed to show the numbers of entities moving in eight different directions, and, through the internal circle, the number of stationary entities. Although not mentioned by the author, this could potentially be adapted to show numbers *and* velocities by adjusting both the width and length of the eight points.

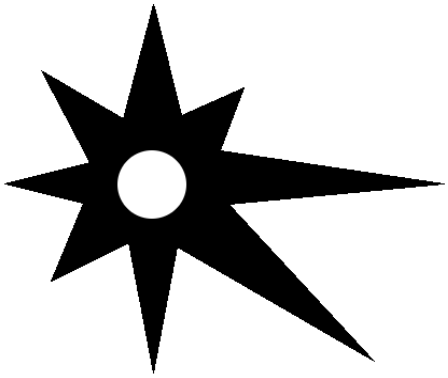

Figure 2.10: Representation of flows in eight directions (Andrienko *et al.* 2008).

Andrienko *et al.* (2008) explain that there is a trade-off between the number of these figures placed on a map and the size of each, where small sizes might affect legibility but large sizes might obscure the map. However, directional bar diagrams were successfully laid out on a 20x20 grid over a map of Milan (Andrienko & Andrienko 2008), and it is possible that a higher density could be achieved through the use of translucent figures.

2.13 Visualizing movement – other techniques

An interesting technique for visualizing movement is the *time cartogram*, which distorts a map so that the distance between points is proportional to the travel time between them. Such maps are usually an approximation, since travel times between two points are rarely the same in both directions (Muehrcke *et al.* 1998) and it is often topologically impossible to place all the points the correct distance from each other on a two-dimensional map. Time cartograms also ignore volume information, since the map distortions are based purely on travel time, and hence velocity.

Like most cartograms, time cartograms also rely on the viewer being familiar with the original map so that they will notice the distortions, which will largely depend on the map being used. No examples could be found of time cartograms being generated from mobile data, so it is not clear how useful they would be. Their main application to date has been in the display of public transport routes, where travel times are generally the same in both directions, the topology is simple, and users are more interested in travel times than distances (Muehrcke *et al.* 1998).

2.14 Conclusions

This chapter has reviewed literature in the areas of tracking, analyzing, and visualizing population movements using mobile phone data. This is a relatively new field of research, with nearly all research published in the past ten years, so there are a number of areas still to be investigated.

The research has found that the use of mobile phone data offers enormous potential, but until recently this has been largely unfulfilled due to a lack of available data from mobile carriers, with privacy concerns being the main reason cited. Some work has been done to simulate the data, but only using simplified models and not using the configurations of real mobile phone networks.

In the past few years a number of researchers have obtained access to (usually) anonymous billing data, which they have used to study the movements of earthquake evacuees, measure the similarity of people's day-to-day travels, and track the seasonal movements of tourists broken down by nationality. However, so far no research has been found that uses billing data for applications such as infrastructure planning, for example planning public transport routes. This would appear to be an area with great potential, and still unexplored.

When it comes to the visualization of mobile phone location data, the representation of spatio-temporal data is still a relatively immature area. Although the display of static spatial data is a well-established area within cartography, the incorporation of time information is relatively new and there is room for improvement.

Changes in mobile phone distributions have been displayed effectively with time series or animations of static maps, but techniques for the displaying of routes, velocities, directions, and volumes are still limited. In particular, the visualization of unconstrained movements within a region, as opposed to movements between pre-defined regions or along pre-defined

routes, has not been adequately explored. To some extent, this is because suitable data has only become available in recent years.

One technique that has the potential to visualize the enormous volume of data available from mobile phone billing data is the continuous flow map. In the past these have mostly been used to display net movements, but combined with multi-directional figures similar to those described by Andrienko *et al.* (2008) they have the potential to show the locations, speeds, directions, and volumes of large numbers of individuals.

Most research has also focused on the visualization of the journeys taken by individuals, attempting to show at least their beginning and end points. The problem with this approach is that it requires an analyst to specify the start and end times that define a journey, which precludes the fully automatic generation of images. The automated detection of the endpoints of journeys from sequences of spatio-temporal points may be an area for further research, but in the meantime the visualization of trajectories, routes, and flows is at least a partly manual process.

To get around the problem of finding journey endpoints, one possibility may be to focus on displaying instantaneous speeds and directions rather than journeys. Continuous flow maps have been shown to be capable of displaying that type of data, and other techniques such as time cartograms may also be applicable. To some extent it depends on the requirements of the end user, but there appears to be room for improvement over the current techniques.

The next chapter provides an overview of different techniques used for tracking individuals, including the use of mobile phones.

3 Chapter Three: Tracking techniques

3.1 Introduction

This chapter aims to provide an overview of techniques used to track population movements. Each will be described in some detail and evaluated using a number of criteria.

The scope of this chapter will be limited to techniques that track people or motor vehicles, so aviation and marine tracking techniques will not be discussed.

3.2 Benefits

There are many reasons for monitoring the movement of people and vehicles, whether simply counting numbers at fixed locations or tracking entire journeys.

Counting numbers at fixed locations can be used to determine the utilization of fixed assets such as roads, parks, shopping centres, and tourist attractions. This information is useful when determining appropriate levels of amenities and services to support those assets, or when to upgrade the asset itself. For example, knowing the number of people and vehicles that visit a national park on a day-by-day and hour-by-hour basis would allow staffing levels to be set appropriately, and amenities such as parking spaces to be added as needed.

Being able to track entire journeys and generate origin-destination (OD) matrices, even anonymous ones, is valuable when planning transport infrastructure. OD matrices allow planners to predict the utilization of new roads, new public transport routes, congestion charging zones, and even the usage patterns of new vehicles such as electric cars (Simonite 2009), resulting in more efficient designs.

Detailed OD matrices may also have commercial applications. Knowledge of a person's travel patterns could identify them as a particular demographic, e.g. city worker, stay-at-home, music festival fans etc. Marketing efforts could then be focused on neighbourhoods where a favourable demographic is concentrated.

Where location data is available in real time, the ability to track populations would be valuable for emergency service organizations. Knowing who is present in a given area would be useful for applications such as crowd control, emergency evacuations, and search-and-rescue.

Finally, being able to track the movements of individuals is a useful tool for security agencies. Real time covert tracking would allow the monitoring of potentially dangerous individuals, alerting police to any suspicious movements as they occur. Overt tracking on the other hand, using for example GPS ankle bracelets, can be used to discreetly monitor the movements of people under house arrest or whose movements are restricted by law (for example convicted child molesters who must stay away from schools). Individual tracking data can also be used to verify alibis, track fugitives and find missing persons.

3.3 Evaluation criteria

There are a number of techniques that are currently used for tracking populations movements. They will be evaluated using the following criteria.

- **Spatial accuracy**. When a subject's position is measured, how closely does it match their true latitude, longitude, and altitude? This is expressed as an error term, in metres.
- **Temporal accuracy**. This is a measure of the accuracy of the time stamp associated with each position measurement. Sometimes the time is known exactly, but sometimes it's approximate, which can impact the accuracy of velocity calculations.
- **Numeric accuracy**. How accurately does the technique record the number of people when taking a sample? For example, techniques that monitor vehicle movements may not generate accurate numbers of people, or a mobile phone may not always correspond to an individual.
- **Area range**. Are there limitations on the areas where this technique can be used? For example, are the subject's positions recorded at fixed points (e.g. as they pass a sensor), or can they be recorded at arbitrary locations (e.g. via GPS)? And if the points are fixed, how extensive is their coverage?
- **Sampling rate**. How often are measurements obtained? Is it periodic, on-demand, or vary with the subject's movements?
- **Demographic info / privacy**. How much additional demographic information can be attached to the spatial and temporal measurements? Can the subjects opt in to provide more information?
- **Ability to track journeys**. Can a subject be uniquely identified when recording their

positions, allowing their entire journey to be tracked?

- **Cost**. How much does it cost to implement this technique? How does the cost scale with the number of subjects and the number of sample locations (if using fixed locations)?
- **Population coverage**. What proportion of the subjects in a region will be measured by this technique? If the proportion is small, is there any selection bias in the subjects who are recorded, and if so, does it affect the results?

3.4 Existing techniques

There are many techniques for tracking the movements of a population. A number of them are described here in detail, and are evaluated against the criteria defined in section 3.3. A summary of these evaluations will be laid out in a table at the end.

3.4.1 Direct observation

Perhaps that simplest method for tracking people is **direct observation**, otherwise known as following them around (Hill 1984). The spatial and temporal accuracy of this technique depends on the accuracy of the equipment being carried by the observer, but in theory it can be arbitrarily precise. Direct observation can be used anywhere, is numerically accurate, can track journeys, and the sampling rate is up to the observer. Additional demographic information can sometimes be obtained with the cooperation of the subjects.

The obvious drawback of this technique, however, is cost, in particular the labour cost of the observers. Since an observer can probably only observe a single individual or group, the high cost of observation precludes widespread coverage, and may even result in sample sizes that are too small for valid extrapolation. There is also some selection bias, since direct observation can usually only record the movements of people who agree to be followed.

3.4.2 Self-reported surveys and interviews

To get around the high labour costs of direct observation, **self-reported surveys** and **interviews** are sometimes used. But while the results can be as accurate as direct observation, that largely depends on the diligence and memory of the participants (Hill 1984). This technique also adds some selection bias, since it only records the movements of people who have the time and motivation to fill in a survey or take part in an interview (Monheim 1998).

3.4.3 Fixed sensors

The limitations of using humans to record movements can be largely overcome by using automated equipment. In their simplest form, **fixed sensors** can count the number of people and vehicles passing a given point. This can be as primitive as a turnstile with a counter whose value is checked periodically, or as complex as inductive loop sensors embedded in roads that feed real-time signals to a central computer as cars pass over them (Klein 2001). Other technologies include pressure pads (Melville & Ruohonen 2004) and infra-red and ultrasonic sensors (Greene-Roesel *et al.* 2008).

However, although these types of sensors are adequate for counting subjects, they are unable to record a unique identifier for each. This inability to match readings to individual subjects prevents them from compiling journey information.

The spatial accuracy of fixed sensors is usually excellent, often less than a metre, if their fixed location is known precisely, and their temporal accuracy is also very good if the recording device to which they are attached stores accurate time values. Where they really excel, however, is in their numerical accuracy and total population coverage. Because they are usually designed so that people and vehicles cannot pass without triggering them, they provide an excellent count of the number of subjects passing a point. If multiple sensors are laid out to cover all the entrances and exits to an area, then they will also provide an accurate count of the number of people in that area at any given time. The main drawbacks with fixed sensors, however, are their restricted spatial coverage and costs that scale with the number of sensor locations. They also rely on subjects physically moving past them, so there is no control over the rate at which samples are taken.

3.4.4 Identifying fixed sensors

To compile journey information **identifying fixed sensors** can be used, which record a unique identifier for each subject that passes by them. Examples of these types of sensors include toll road scanners such as Melbourne CityLink's e-Tag system (Holmes 2000), public transport smart card tickets (Zhao *et al.* 2007), and licence plate cameras (Castilloa *et al.* 2008). Since these sensors need to be in place to collect road tolls, ensure payment on public transport networks, etc., the additional cost of using them to track journeys is minimal. Because the physical infrastructure is already in place and the raw data is already being stored, all that's needed to extract journey information is additional computer processing.

In many cases unique identifiers can be traced back to an individual, although this often requires a court order in order to overcome privacy laws. Toll roads sensors usually either scan a licence plate number (which can be linked to the vehicle's owner via a vehicle registration database) or query a transponder such as an e-Tag, which will be linked to an account in the owner's name, usually containing a billing address, credit card details, or a bank account number. Also, many public transport tickets are paid for by credit card, which can usually be traced back to an individual. This ability to link a unique identifier to an individual means that a lot of additional demographic information is theoretically available, although in practice privacy laws and the need to combine data from different organizations may make it impractical.

Apart from their ability to compile journey information and sometimes link that to demographic data, identifying sensors in most other respects have the same benefits and drawbacks as regular fixed sensors. Their spatial, temporal, and numerical accuracy are usually excellent, while their spatial coverage is restricted and there is little control over the rate at which samples are collected.

3.4.5 GPS tracking devices

Another way to collect journey information is for the subjects to carry **tracking devices**. These devices can either transmit their location in real time or store it in memory for later retrieval. Most use the Global Positioning System (GPS) to obtain their spatial coordinates, which is accurate to a few metres and works on most of the earth's surface (McNamara 2008). GPS does, however, require line-of-sight access to four satellites simultaneously, so it is often unable to determine its location while indoors or surrounded by solid obstacles. Also, because of the way GPS satellites broadcast their location data, it can take up to 30 seconds to obtain an initial reading.

GPS devices broadly fall into three categories – dedicated navigation devices, dedicated tracking devices, and programmable devices. The most common example of a navigation device is an in-car GPS unit, which guides drivers to their destination. Other devices on the market are designed for aviation, cycling and marine applications. Most are capable of *track logging*, where their location is periodically recorded to a log file which later be downloaded to a PC (McNamara 2008). If track logging is enabled, these logs provide journey information with accurate spatial data taken at a high (and often configurable) sample rate. Note, however, that only the movements of the vehicle are tracked, which are not necessarily the same as the

movements of an individual, and the number of occupants is not recorded.

Dedicated tracking devices usually combine a GPS receiver with some kind of transmitter, and are attached to an object of interest. This object might be a vehicle in a corporate fleet, a shipping container with a valuable cargo, or a criminal offender under house arrest (Kucharson 2006). On a periodic basis they transmit their location to a central computer.

For example, with criminals under house arrest, a tamper-proof battery-operated transmitter is placed around their ankle. The signal from this transmitter is monitored by a portable tracking unit (PTU), worn around the offender's waist. The PTU consists of a GPS receiver, mobile phone circuitry, and a simple computer. The GPS receiver continually records the location of the offender, while the mobile phone periodically uploads the location data to a central computer where it is stored.

The most common programmable GPS devices are mobile “smart” phones with built-in GPS receivers, for example the Apple iPhone. Because of their ability to run applications, they can be programmed to periodically takes GPS readings and either store them locally or upload them to a central computer (Miluzzo *et al.* 2008). And since these devices tend to be carried by their owners anyway, their movements can be recorded at little or no additional cost. From the owner's perspective, the main drawbacks are the time and effort needed to install the software, the additional drain on the handset's battery due to continual use of the GPS receiver, and possibly higher data charges caused by the uploading of the location data to a central computer.

GPS-equipped smart phones are usually more expensive than non-programmable non-GPS equivalents, so their market share is low (but growing fast) and their ownership is possibly skewed towards early adopters and wealthier individuals. However, the main factor limiting population coverage for all types of GPS devices is likely to be difficulty of convincing subjects to allow their locations to be recorded (except for criminal offenders, whose alternative is usually prison).

3.4.6 Mobile phones

One device that most people carry with them is their mobile phone. As shown in figure 3.1, in June 2008 most of the Australian population owned one, and among working-age adults the number was around 90 percent (Australian Communications and Media Authority 2008, Australian Communications and Media Authority 2009a). Multiplying the ownership rates by

the population in each age group yields a total of 15.29 million mobile phone owners in Australia, out of a total population of 21.43 million (Australian Bureau of Statistics 2008). This implies an overall ownership rate of 71.4 percent. The actual percentage may be slightly higher, since the ownership rate for children aged 7 and under was not available and was assumed to be zero, but in reality may be higher.

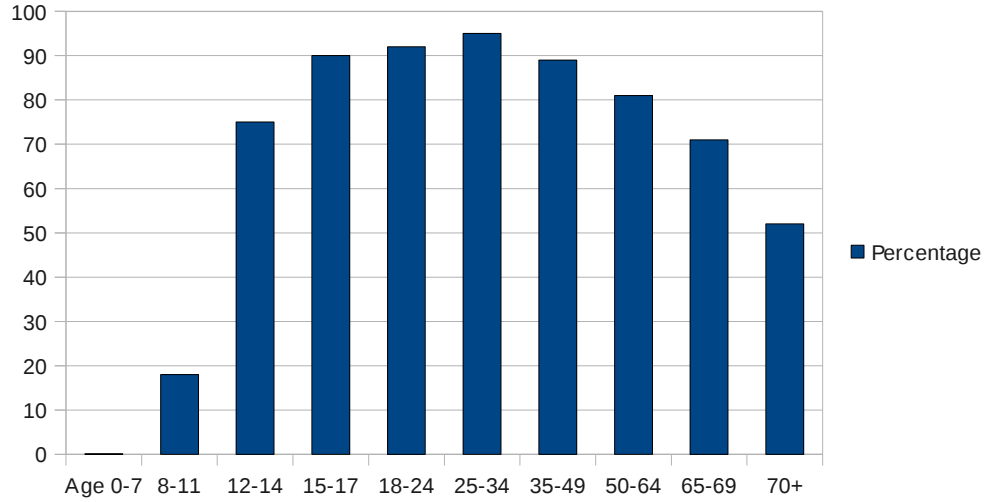

Figure 3.1: Australian mobile phone ownership rates, by age, June 2008 (Australian Communications and Media Authority 2008).

It should be noted, however, that not all mobile phones are uniquely associated with an individual. Although phones were owned by 15.29 million Australians in June 2008, at the time there were actually 22.12 million mobile phone subscriptions (Australian Communications and Media Authority 2008), implying that many people own more than one phone. In practice, some people carry more than one handset, but many subscriptions are also used for wireless internet access, embedding in remote sensors, or used in bulk SMS messaging equipment, and are not necessarily carried by a person. As a result, it may be necessary to empirically determine a scaling factor to covert a handset count to a head count in the real world.

Although not all phones are equipped with GPS, they can still be located with varying degrees of accuracy by their radio communications with a mobile phone tower, since the locations of towers are fixed and known accurately. For example, by triangulating using the relative signal strengths of all the towers detectable by a handset, Chen *et al.* (2006) were able to estimate locations to a median accuracy of 94-196 metres in metropolitan Seattle using the towers of a

single carrier, and to 65-135 metres using the towers of all three available carriers.

Similarly, using a pre-recorded database of the signal strength “fingerprints” (the strengths of all detectable cells at a particular location) Varshavsky *et al.* (2006) were able to estimate locations in Seattle to an accuracy of 75 metres outdoors and 5 metres indoors. However, while very accurate, this method requires a fingerprint database that can only be built by physically traversing the entire area with a computer linked to a GPS and mobile phone. This is a time-consuming and expensive process.

However, the main problem with these two techniques is that the location is only known by the handset, and because of the difficulty of passing that information to a central collection point it is of limited use for tracking large numbers of people.

More useful are mobile phone tracking techniques that can be carried out by the network rather than the handsets themselves. Data can be collected at a small number of fixed locations, and the population doesn't need to change their behaviour or use new hardware or software.

Perhaps the easiest way to track mobile phones using network data is to measure **cell activity**, or the number of calls in progress at each cell throughout the day, and to use that to estimate the population within a cell's coverage area (Ratti *et al.* 2005, Ratti *et al.* 2006). An advantage of cell activity data is that it is recorded as a matter course by carriers to track the utilization of their equipment.

The spatial accuracy of a cell ranges from a few hundred metres in a city to over a hundred kilometres in rural areas (Trevisani & Vitaletti 2004), so measuring cell activity is not as accurate as the handset-based methods described above. Also, because it only measures calls in progress, extrapolating that to an actual population count has a wide margin of error, since it needs to estimate what proportion of the population in that area is making a call at that time of day.

Another limitation is that the data can't be used to generate journey or demographic information because handset identifiers aren't recorded. But on the other hand, it works anywhere there is mobile phone coverage, and requires no additional hardware – and possibly very little additional software.

3.4.7 Active querying of handsets

Because cell activity measurements only counts handsets that are on a call or

sending/receiving an SMS, attempts at tracking populations will be biased by frequent users. This can be overcome by the **active querying** of individual handsets, where the location of a handset can be polled as needed.

Querying the location of a handset will at least retrieve its current cell ID, but a number of more sophisticated techniques are described in Raja *et al.* (2004) that can return more accurate locations. Many of these require handsets that co-operate with the network in non-standard ways, but some can be implemented with normal handsets and upgraded network infrastructure.

Because radio waves travel through the atmosphere at roughly the speed of light, measuring the round trip delay in communications allows the distance between a handset and cell antenna to be calculated. On GSM networks, radio frequencies are divided into eight time-separated channels, and each channel must transmit during its allocated time slots, each of which lasts 0.577 milliseconds. In order to synchronize with these slots the handset much compensate for speed-of-light delays, using a technique known as *timing advance* (TA) (Lin & Chlamtac 2001).

A TA duration is calculated by the cell base station and sent to the handset, which delays its signal accordingly. A TA value in the range 0 – 233 microseconds is encoded in six bits (possible values 0 - 63), yielding a resolution of 233/63 = 3.7 microseconds. Multiplying 3.7 microseconds by the speed of light gives a round-trip distance resolution of approximately 1100 metres, or a one-way distance resolution of half that. In other words, on a GSM network the distance between the antenna and the handset can be estimated with an accuracy of about 550 metres using timing advance (Lin & Chlamtac 2001).

3G (Third Generation) networks don't use synchronized time slots, but they still calculate a *round trip time* (RTT) value used for other purposes (Wigren & Wennervirta 2009). The time resolution of an RTT value is called a *chip*, equal to 1/3840000 seconds (Lin & Chlamtac 2001). During this time light travels approximately 78 metres, so the one-way distance between a 3G antenna and a handset can be estimated with an accuracy of 39 metres.

Using round trip delays to locate a handset will place it somewhere on a “donut” surrounding omnidirectional antennae or on an arc facing directional antennae. The thickness of the donut or arc will be 550 metres on GSM networks and 39 metres on 3G networks.

On a 3G network it is theoretically possible to use RTT values from multiple antennae and locate a handset where the donuts or arcs intersect. However, this can only be calculated when

a handset is engaged in a *soft handover* between antennae, when it is communicating with multiple antennae simultaneously, and field tests indicate that handsets are engaged with two antennae only 20% of the time, and with three or more antennae less than 10% of the time (Wigren & Wennervirta 2009).

Another technique that can be used with standard handsets is called Angle of Arrival (AOA). If a cell is equipped with a directional antenna it can calculate the direction from which a handset's signal arrives. This information, when combined with a distance estimate derived from the round trip delay, can locate the handset to within a small arc. In practice however, this technique relies on having a line-of-sight connection between the handset and antenna, because reflected signals will provide incorrect angles (Raja *et al.* 2004). Since line-of-sight cannot be guaranteed in most environments, this technique is rarely used.

However, the main problem with active querying is that generates extra traffic on a mobile phone network above and beyond that used for normal operations. Since networks are generally designed to handle normal traffic loads, there is probably insufficient capacity to support the constant active querying of all handsets, especially during peak times (Dufkova *et al.* 2008). Possible options are to query a limited subset of handsets, limit active querying to off-peak periods, or the upgrade the network – at considerable expense – to support around-the-clock active querying of all handsets. It should also be pointed out that active querying also drains the batteries of handsets slightly faster than normal, since they have to communicate with the network more often.

The use of active querying to calculate population numbers also faces the problem that it is necessary to know the identity of every handset that *might* be in the area in order to query it. Fortunately, as will be discussed in chapter four, carriers maintain databases called Visitor Location Registries (VLRs) that maintain a list of all handsets that are known to be in a region called a Location Area (LA), whether they belong to the same network or are roaming from elsewhere (Lin & Chlamtac 2001). Thus querying all the handsets that appear in a VLR will guarantee that every handset within the LA will be polled, which should generate fairly accurate population numbers.

3.4.8 Passive querying of handsets

Finally, it should be possible to track the location of handsets on a mobile phone network with **passive scanning**. This involves recording details of the time, handset ID, and cell ID sent

when a handset communicates with the network as part of normal operations.

With existing infrastructure this data is already available in the form of **billing records**. Because carriers need to maintain the option of distance-based call tariffs, all “billable events” - phone calls, SMSes, and internet access - generate a billing record containing the cell ID of both the caller and (where relevant) the recipient (Candia *et al.* 2008, White & Wells 2002). These cell IDs can be used to calculate the distance of the call for billing purposes, but as a side-effect can also be used to track the movements of the caller and recipient.

As with cell activity measurements, billing records are only accurate to the nearest cell and are only created when a handset is active. On the other hand, they can be linked to an individual, providing journey information and additional demographic data. And because this billing information is collected automatically and stored for at least a month (or however frequently billing occurs), it can provide location information for all handsets upon request. In fact, billing records have proved to be a valuable source of evidence in police investigations, since they allow alibis to be easily confirmed or refuted (Schmitz *et al.* 2000).

To find out just how much billing data was being recorded by carriers, in 2010 German Green party politician Malte Spitz filed suit against Deutsche Telekom to release all the records they held relating to his account (Cohen 2011). The resulting data, consisting of 35,830 records over the six month period September 2009 to February 2010, was then made publicly available on the Zeit Online website (Biermann 2011).

Deutsche Telekom was recording just under 200 records per day, of which 168 contained spatial information in the form of a latitude and longitude, as well as the cell ID and the direction of the cell's antenna (if directional). Roughly 24 percent of the records were generated by SMS, 13 percent by voice calls, and 54 percent by GPRS internet access. The remaining records were not identified.

However, as a full-time politician Spitz may have used his mobile phone significantly more often than the average user. In another study, data from almost a million subscribers in the Boston area over the period 30 July 2009 – 12 September 2009 consisted of 130 million records (Calabrese *et al.* 2010), implying an average of only 2.9 records per subscriber per day in that dataset, barely one sixtieth the rate of Spitz's data. However, this data was “aggregated and anonymous” so it is not clear if it contains all the original data.

In addition to billable events, a handset will also communicate with the network when it is first switched on, when it changes Location Area, and periodically at a rate specified by the

network (typically every hour). These messages, known as **signalling data**, are a superset of billing data, and by passively scanning all of them it should be possible to track the location of every handset on the network at a greater frequency than with billing data. Apart from the equipment that does the scanning, no additional hardware would be needed.

Passive scanning would be accurate to the nearest cell ID, could track journeys, and would work wherever there is mobile phone coverage. The sampling rate would depend on handset usage, but would at least be as frequent as the network's requested periodic update rate (used to check whether a handset is still switched on). This rate is typically around once per hour, and can easily be increased by the carrier at the cost of additional network traffic. It should also be pointed out that passive scanning could be used in conjunction with active querying, in cases where there are handsets whose location is needed more frequently and/or more accurately.

3.4.9 Location service providers

When a mobile phone needs to know its location and cannot obtain an accurate GPS reading, it often resorts to querying a **location service provider**. This involves the phone sending information about its current environment, such as the cell IDs and signal strengths of visible mobile phone cells or the MAC (Media Access Control) addresses and signal strengths of nearby WiFi antennae. The provider looks up this information in a database and uses it to calculate an estimated latitude and longitude for the device.

By logging all these queries – especially if they contain a unique identifier for each querying device – location service providers can track the movements of large numbers of people. One of the largest providers, Skyhook Wireless, who supply location data for iPhones and other devices, offers access to this information via their SpotRank service, shown in figure 3.2. To quote their 25 March 2010 media release (Skyhook Wireless 2010),

> Skyhook launched a new data intelligence service last week called SpotRank. SpotRank predicts the density of people in predefined urban square-block areas worldwide at any hour, any day of the week.
>
> SpotRank is based on data from Skyhook’s Core Engine, which powers the location on tens of millions of devices. With this reach, the Core Engine network processes 300 million location requests daily. Skyhook has developed SpotRank by continually mining this anonymous location data to predict

human behaviour.

> Skyhook provides four types of SpotRanks for 500 million square-block Spots in 770 metro areas around the world. Each Spot is assigned an average for every hour and day of the week.

Skyhook claims location accuracies of 10 - 20 metres, using a combination of WiFi addresses, cell IDs, and GPS data (possibly using WiFi data to improve the accuracy of GPS readings). WiFi addresses are more accurate than cell IDs, but are not always available outside of cities, so actual accuracies may vary considerably between suburban and rural areas.

Although Skyhook assigns a unique random ID to all devices that access their service, this information is not stored. To quote their privacy policy (Skyhook Wireless 2009),

> At no time do we store or retain the random ID; this is used only to pair a location request with a location requestor to facilitate a specific, one-time location transaction and for no other purpose. No information about a particular user or device is retained on the system.

Notwithstanding their privacy policy, in theory Skyhook's system could track the movements of “tens of millions” of people to an accuracy of 20 metres. Given their claim of 300 million location requests per day, and assuming that “tens of millions” means around 30 million, that translates to around 10 requests per device per day, implying a temporal resolution of around two hours.

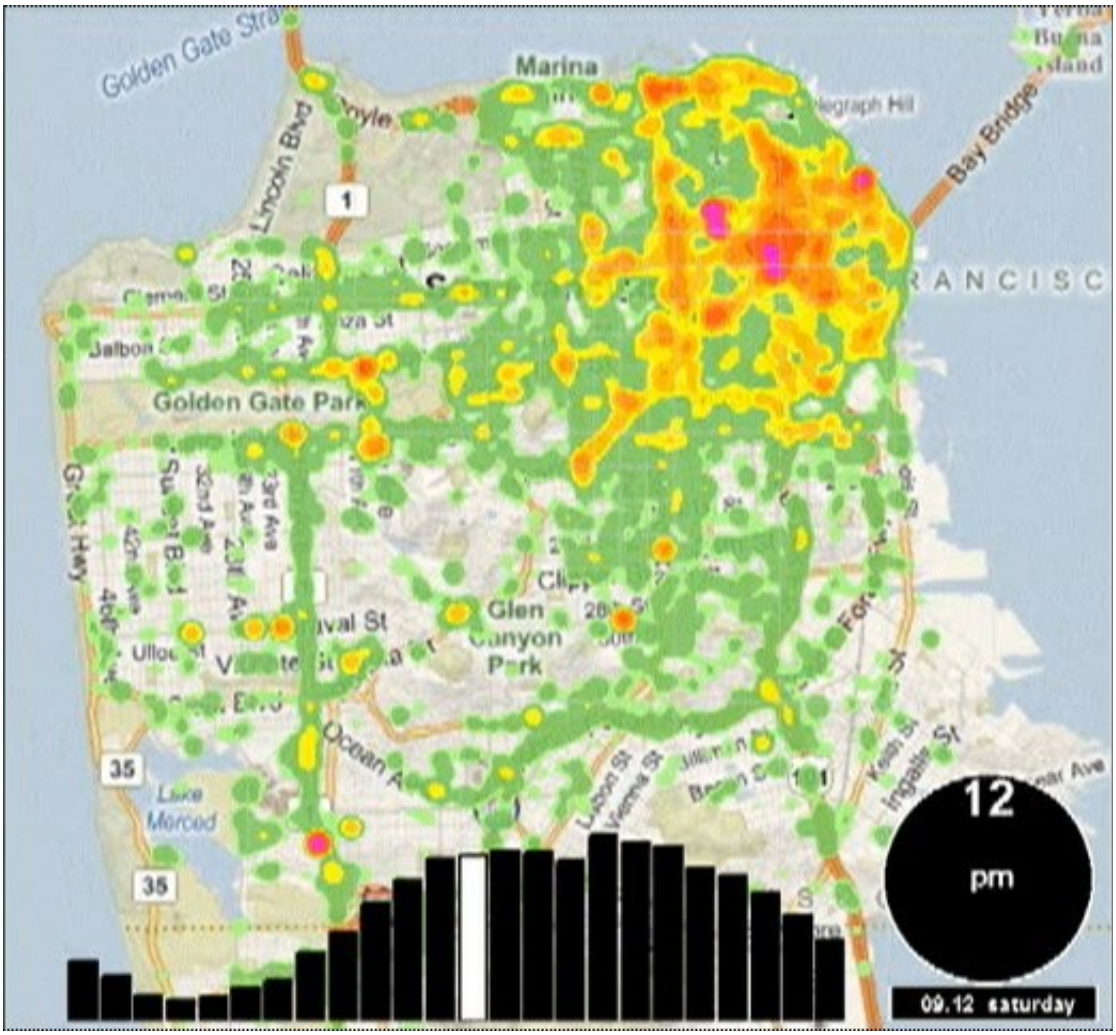

Figure 3.2: SpotRank view of San Francisco, midday Saturday 12 September 2009 (Skyhook Wireless 2010).

In terms of population coverage, Skyhook's map in figure 3.3 shows extensive coverage in the industrialized parts of the world, such as North America, Western Europe, Japan, South Korea, Australia, and New Zealand, plus partial coverage in developing places such as China, India, and South America. Assuming these parts of the world contain around a billion people, and that “tens of millions” are using Skyhook's service regularly, this implies coverage of a few percent of the population.

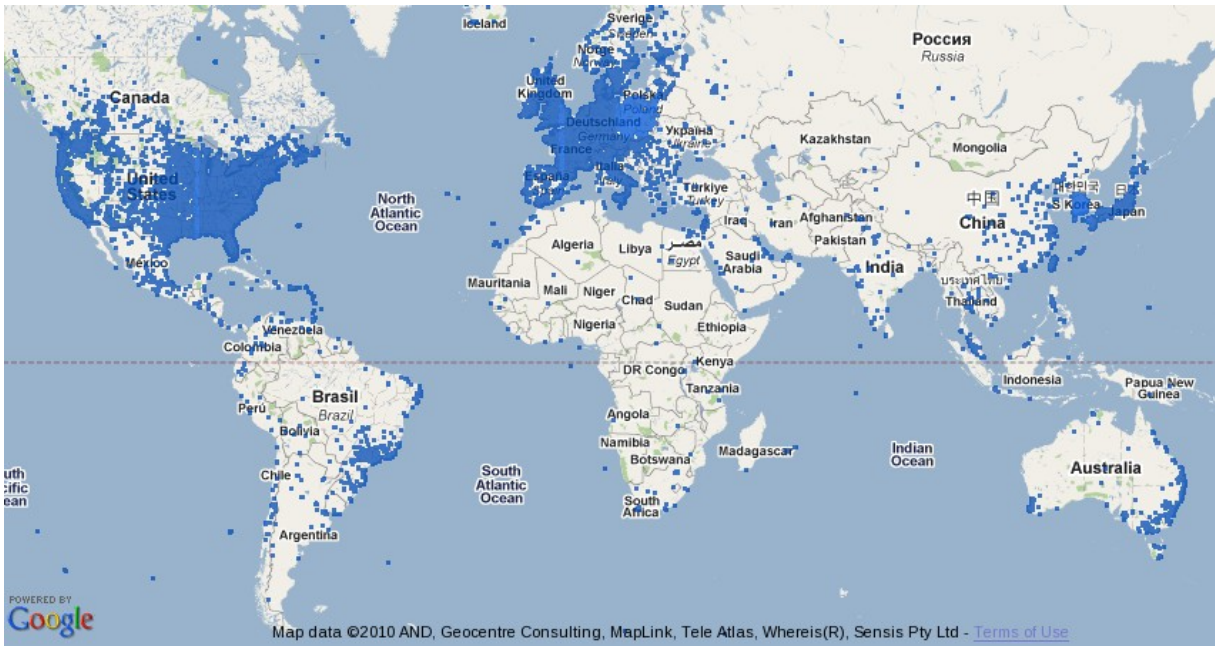

Figure 3.3: Areas covered by Skyhook Wireless (Skyhook Wireless 2010).

Another location service provider is NAVTEQ, who provides mapping data for Nokia mobile devices, among others. According to NAVTEQ's privacy policy (NAVTEQ 2010),

> If you access NAVTEQ Products and Services from a mobile device, NAVTEQ may collect limited information from your device, including location data, but will not store this information in a manner that identifies you individually unless you consent for us to do so.

Since 2008, NAVTEQ has been generating road traffic information from the location data collected from GPS-equipped mobile devices (NAVTEQ 2008). Because traffic information requires location data from uniquely identified devices, it is presumably only generated from users who have consented to allow this. Their pilot study in San Francisco Bay Area enlisted "more than 10,000 handset owners who have agreed to provide their mobile phone GPS positioning data".

As well as NAVTEQ, a number of other firms are using GPS data from in-car navigation systems and mobile phones to provide real-time traffic data (Strategy Analytics 2010). These include Google, TomTom, TeleCommunications Systems, TeleNav, and Inrix.

3.5 Summary of tracking techniques

Table 3.1 below summarizes the various tracking techniques discussed in this chapter and

how well they satisfy the evaluation criteria.

Technique	Spatial accuracy	Temporal accuracy	Numeric accuracy	Area range	Sample rate	Demographic info	Track journey	Cost	Pop'n coverage
Direct observation	< 100m	Minutes	Very good	Very good	As needed	Very good	Yes	High	Poor
Interviews / surveys	Varies	Hours	Very good	Very good	As needed	Very good	Yes	High	Poor
Fixed sensors	Usually < 1m	< 1 sec - daily	Medium / good	Low / med	Varies	None	No	High	Very good
Identifying fixed sensors	Usually < 1m	< 1 sec	Medium / Very good	Low / med	Varies	Poor – very good	Yes	Very high	Very good
GPS tracking devices	5m - 500m	< 30 sec	Good	Very good	Up to user	Poor – v good	Yes	Very high	Poor
Cell activity	100m - 100km	< 1 sec	Good	Good	As needed	None	No	Low	Medium
Mobile billing records	100m - 100km	< 1 sec	Good	Good	Varies	Very good	Yes	Low	Good
Active mobile query	100m - 100km	< 1 sec	Good	Good	As needed	Very good	Yes	Low Med	Low / Medium
Passive mobile scanning	100m - 100km	< 1 sec	Good	Good	> 1 per hour	Very good	Yes	Low	Good
Location service providers	5m - 500m	< 1 sec	Medium /Good	Good / Very good	Every 2 hours or so	Poor	In theory	Low	A few percent

Table 3.1: Summary of tracking techniques.

3.6 Conclusions

This chapter described existing techniques for tracking population movements, along with their relative benefits and drawbacks. It found that the passive scanning of mobile phone networks scores well on most evaluation criteria, and among techniques that can cover entire populations it is generally the best. If passively scanned signalling data is not available, then mobile phone billing records are almost as good, although they have a lower sample rate.

The *active* querying of mobile phones allows for on-demand sample rates and provides greater spatial accuracy than passive techniques (using distance-from-antenna estimates), but the additional network traffic it generates makes it impractical for tracking entire populations.

The next chapter will go into more technical detail on how population tracking can be carried out by passively scanning a mobile phone network.

4 Chapter Four: How a mobile phone network operates

4.1 Introduction

The aim of this chapter is to provide an overview of how mobile phone networks operate. It looks at the type of information that flows through a network during normal operations, and how this information could be used to determine the location of mobile phones on a more frequent basis than using billing data, and do so in real-time.

4.2 Network structure

The purpose of a mobile phone network is route voice calls, text messages, and, increasingly, internet data to and from mobile phones. Given the limited range and reliability of radio waves, the mobile nature of mobile phones, and the sheer number of phones in use, this is a very complex task involving the interactions of several types of equipment (Lin & Chlamtac 2001).

The most common piece of equipment is the handset, or mobile phone, known technically as a *Mobile Station* (**MS**), the device used by subscribers to make and receive calls (Lin & Chlamtac 2001). For a handset to work with a network it must first be fitted with a *Subscriber Identity Module* (**SIM**) card, and the SIM must be registered to a valid account recognized by the network.

Each SIM is uniquely identified by a 15-digit code called an *International Mobile Subscriber Identity* (IMSI) which is made up of three components – a 3-digit *Mobile Country Code* (MCC), a 2- or 3-digit *Mobile Network Code* (MNC), and a 10-digit or less *Mobile Subscriber Identification Number* (MSIN) (Lin & Chlamtac 2001). For example, an IMSI on the Vodafone Australia network starts with 50503 and is followed by another ten digits, e.g. 505030005765039.

A handset communicates by radio with an antenna on a mobile phone tower, technically known as a *Base Transceiver Station* (BTS). Each antenna on the BTS provides coverage for an area known as a *cell*, which has a unique *cell ID* (CID). On a GSM (Global System for Mobile) network cell IDs are 16-bit values in the range 0 - 65535, while on 3G networks they are 32-bit values in the range 0 – 4294967295 (Lin & Chlamtac 2001).

A BTS is typically configured with either a single omnidirectional antenna, or with three

directional antennae each covering 120-degree sectors. One or more BTSs are then connected to a *Base Station Controller* (**BSC**) to form a *Base Station Subsystem* (**BSS**), as shown in figure 4.1. The link connecting the BTSs to the BSC is known as the A_{bis} *Interface* (Lin & Chlamtac 2001).

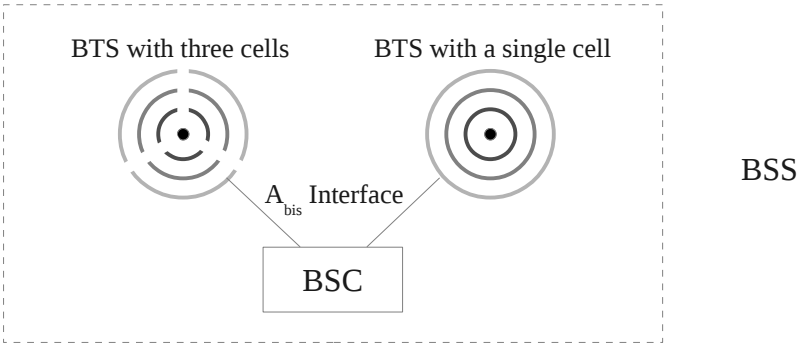

Figure 4.1: Base Station Subsystem.

One or more BSSs are then connected to a *Mobile Switching Centre* (**MSC**), which is essentially a telephone exchange for mobile networks. The link connecting the BSSs to the MSC is known as the *A Interface* (Lin & Chlamtac 2001).

The BSSs connected to an MSC are grouped into one or more *Location Areas* (**LA)** (see figure 4.2), each identified by a unique *Location Area Identifier* (**LAI**). On both GSM and 3G networks an LAI is a 16-bit value in the range 0 – 65535 (Lin & Chlamtac 2001).

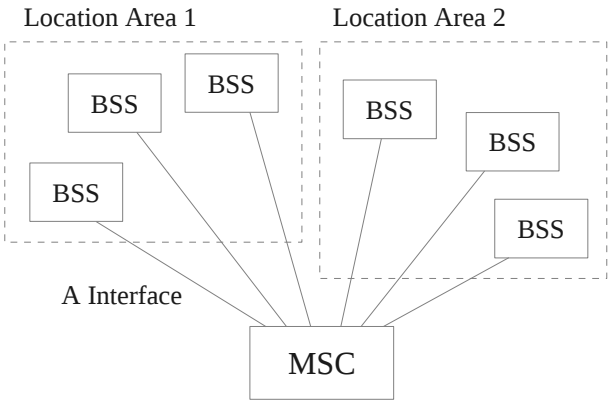

Figure 4.2: Mobile Switching Centre.

One or more MSCs are in turn served by a database called a *Visitor Location Register* (**VLR**). The VLR is responsible for storing the current LAI of all the handsets registered with the MSCs. In practice, a VLR usually serves a single MSC, and the MSC and VLR are often

combined into a single piece of equipment. It is also fairly common for an MSC to in turn only serve a single Location Area (Lin & Chlamtac 2001).

The final component in the network is the *Home Location Register* (**HLR**). This is a database maintained by a handset's home carrier, and contains information such as the handset's phone number, its IMSI, its remaining credits, and the VLR where it is currently registered (Lin & Chlamtac 2001).

The HLR, VLR, and MSC components of the network are collectively known as the *Network and Switching Subsystem* (**NSS**), as shown in figure 4.3.

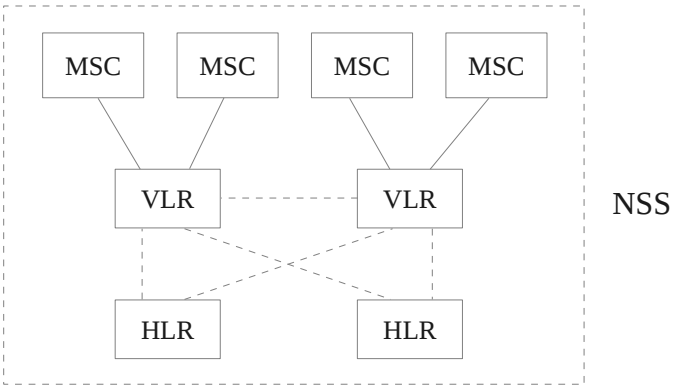

Figure 4.3: Network and Switching Subsystem.

The various components of the network communicate with a protocol called Signalling System No. 7 (**SS7**). SS7 is in turn used to encapsulate messages from higher-level protocols such as Base Station Subsystem Application Part (**BSSAP**) and Mobile Application Part (**MAP**). BSSAP is used to communicate between a BSS and an MSC over the A Interface, while MAP is used between MSCs, VLRs, and HLRs (Lin & Chlamtac 2001).

4.3 Active tracking

As discussed in section 3.4.7, a handset is considered to be actively tracked when its position is queried on demand. This is a brief description of the technologies involved in active tracking.

At any time a handset can be in one of three states -

1. *Off-network*. Switched off or out of range;
2. *Idle*. Receiving broadcasts from the network, but not transmitting; or
3. *Active*. Transmitting and receiving.

It is only when a phone is active that it sends its current CID, thus making its location known to the network. While it is in the idle state it is aware of its current CID, but this information is not shared with the network, which only knows the phone's current LAI. In general, handsets remain in the idle state as much as possible to save power (Lin & Chlamtac 2001).

The active tracking of mobile phones refers to techniques where they are explicitly put into an active state so their location can be queried. This can be done either by the network “pinging” the handset, or with the cooperation of the handset itself.

Pinging involves sending a signal to the handset that triggers a response, usually the SS7 “MAP Any Time Interrogation” service (Dialogic 2010), as described in 3GPP TS 09.02 (3GPP 2004). The response from the handset contains, among other things, the handset's current CID. Advanced networks can also estimate the handset's distance from the antenna based on signal delays (Raja *et al.* 2004).

The advantage of active tracking is that locations can be queried as needed. For example, a handset's location could be requested more frequently when it is in motion than when it is stationary, thus reducing the average location error. But the drawbacks of active tracking are the additional traffic load the pings and responses place on the network, and the drain on handset batteries caused by replying to all the pings (Dufkova *et al.* 2008).

Another tracking option is to take advantage of features built in to the handset. Modern smart-phones often come equipped with GPS, or at the very least can identify their nearest cells and their relative signal strengths, allowing for triangulation. If suitable software is installed, and if the handset's owner agrees to it, the handset can cooperate with the network to share its location information.

This technique has the potential to provide the most accurate location data, but it has the same drawbacks as pinging in terms of additional network load and drain on handset batteries – probably even more so if GPS is used. But more importantly, handset cooperation will not work with all phones (only smart-phones have the necessary capabilities), and it requires the approval of handset owners.

Because of these drawbacks, active tracking is only suitable for locating a small subset of handsets on a network. To track all the handsets on a network passive techniques are needed.

4.4 Passive tracking

As discussed in section 3.4.8, the passive tracking of handsets involves analyzing data that is

already stored or transmitted through a mobile phone network as part of its normal operations. As such, it requires no additional traffic to be sent across the network, nor should it place any extra load on the existing network components such as the VLR.

When a mobile phone number is dialled, the first step is to query the number's HLR. The identity of the HLR is determined by the number's carrier prefix (e.g. in Australia the prefix “0410” is Vodafone) plus a query to the Mobile Number Portability database to check for ported numbers.

The HLR returns the IMSI corresponding to the phone number and the identity of the VLR where the handset is currently registered. The IMSI is then passed to the VLR, which returns the current LA of the handset (Lin & Chlamtac 2001).

The MSC handling that LA then broadcasts a *paging* signal containing the IMSI through all the cells within the LA. If the handset receives that signal it will send a response which includes the CID of its current cell. If the recipient then accepts the call, the network will route the voice data to that cell (Lin & Chlamtac 2001).

Note that the network does not generally store a handset's current CID. The most accurate location it records is the handset's LAI, which is maintained by the VLR. However an LA could cover an area containing hundreds of cells, so it is not accurate enough for most applications.

On the other hand, a handset is aware of its current cell at all times, as well as its neighbouring cells and their relative signal strength. As the handset moves around it constantly updates this information internally, but while it is in idle mode (which is most of the time) it does not communicate this information back to the network. To do so would greatly increase network traffic and reduce its battery life.

Instead, the handset only informs the network of its current location under the following circumstances -

1. The handset is switched on and successfully registers with the network.
2. The handset moves to a new LA.
3. Its Location Update timer expires (typically every hour or so).
4. The handset makes or receives a call or SMS, or accesses the internet.

In cases 1 - 3 it sends a MAP_UPDATE_LOCATION_AREA message to the VLR, which

updates the record containing the handset's LAI (if the LAI has changed), and in case 4 it sends a MAP_SEND_INFO_FOR_OUTGOING_CALL or MAP_PROCESS_ACCESS_REQUEST message (Lin & Chlamtac 2001). In all these cases the message contains the handset's CID as well as its LAI, but only in case 4 is the CID stored by the network (in a billing record).

A conceptually simple way to capture CIDs for tracking purposes would be to modify the VLR to store the CID along with the LAI whenever it receives a location update. However, this has some practical problems. First, a VLR is typically an off-the-shelf solution purchased from a telecommunications vendor, and is not easily modified by an end-user. Second, querying a VLR to extract location data would place additional load on the equipment. And finally, this solution would only store *current* location data, and not generate the stream of *historical* location data needed by many applications.

One way around these problems is to install a *probe* in the network that monitors traffic and extracts information from the data as it passes. The ideal location for such a probe would be between the MSC and VLR, where it could intercept the location update messages. However, since the MSC and VLR are usually implemented as a single piece of equipment for performance reasons, it may not be possible to monitor the traffic since it will be sent internally.

It is possible that an MSC/VLR will provide a diagnostic mode that allows internal traffic to be captured, or perhaps the vendor can be persuaded to implement such a mode. But if not, it may instead be necessary to scan the location data by monitoring the A Interface between the BSS and MSC.

There is, however, a complication when monitoring the A Interface. The ideal probe would produce a stream of records containing the IMSI, LAI, and CID of handsets as they update their locations. But for security reasons, the IMSI of a handset is sent over the airwaves as rarely as possible, since it could allow individuals to be tracked with a radio scanner (Lin & Chlamtac 2001). Instead, the VLR issues a *Temporary Mobile Subscriber Identifier* (**TMSI**) that it uses to identify the handset, and this is used instead of an IMSI on the A Interface. This identifier is unique to the LA, so it changes when the handset changes LA. The VLR can also change the TMSI at any time, and will do so periodically.

Internally, the VLR maintains an IMSI/TMSI look-up table, which it adds to / deletes from whenever a handset enters / leaves, and which it updates when it issues a new TMSI (Lin &

Chlamtac 2001). Because a probe won't have access to this look-up table, it will have to maintain its own, probably by reading the original IMSI registration message when a phone is switched on, plus all the TMSI update messages sent by the VLR.

However, monitoring the A Interface on its own will not handle the situation where a handset changes VLR. In that case, the handset's IMSI is sent to the new VLR via the inter-VLR link, and won't be seen by a probe on the A Interface. The probe will only see the new handset's TMSI, and will have no way to determine its corresponding IMSI. This is sufficient to track the movements of a handset while it remains within a Location Area, but once it departs its identity will be lost.

To maintain a valid IMSI/TMSI look-up table it is necessary to probe both the A Interface and the inter-VLR link simultaneously, to some extent replicating the functionality of a VLR. The data arriving on the inter-VLR link will associate a handset's IMSI with its current TMSI, while data being sent over the A Interface will identify any changes to that TMSI.

To summarize, a CID can be captured and associated with an IMSI with passive scanning under the following circumstances -

1. The MSC and VLR are separate pieces of equipment and traffic between them can be probed; or
2. The MSC/VLR is implemented as a single unit which provides a diagnostic interface that allows internal traffic to be monitored. In particular, traffic containing handset IMSI and CID values; or
3. The A Interface and inter-VLR link of the VLR can both be probed simultaneously to allow an IMSI/TMSI lookup table to be maintained.

If the MSC/VLR is a single unit and only the A Interface can be probed, CIDs can still be captured, but can only be associated with a TMSI rather than an IMSI. This means that a handset's movements cannot be tracked beyond that LA.

4.4.1 Real world examples

In the United States, a company called AirSage Inc (http://www.airsage.com) provides location data to its customers based on information retrieved from mobile phone networks. It is not entirely clear what information they use, but according to one of their patents (Smith *et al.* 2002)

> In the exemplary embodiment of the present invention shown in FIG. 1, the Traffic Information System may receive data from a variety of locations in the wireless network. These locations include the BSC and its interface, through the A_{bis} Interface, with the BTS, MSC, the HLR, and the MPS.

In other words AirSage *may* be probing any of the interfaces described above, but it is not clear which ones they *actually* use. However, according to an MIT study using AirSage data (Phithakkitnukoon *et al.* 2010, p 3), the handset locations are

> recorded when the users are engaged in communication via the cellular network. Specifically, the locations are estimated at the beginning and the end of each voice call placed or received, when a short message is sent or received, and while internet is connected.

This suggest that, for that study at least, the AirSage data didn't include records generated by periodic updates or changes in Location Area. Their dataset contained 129 million records from roughly one million handsets collected over 45 days from 30 July to 12 September 2009. This implies a sampling rate of 2.87 records per handset per day, or one every 8.4 hours. Since this is significantly less frequent than a handset's periodic updates, it indicates that only billing data was used.

However, another MIT study (Wang *et al.* 2010) states that their AirSage dataset consisted of "829 millions of anonymous location estimations – latitude and longitude – from close to 1 million devices in 1 month, which are generated each time the device connects to the cellular network". Assuming this means a million handsets over 30 days, it implies a sampling rate of 28 records per day, or 1.15 per hour.

It is not clear why the sampling rates of the records are so different, since both studies appear to be using the same dataset from Boston in 2009. Based on the frequency of its records, the second study *could* be using periodic update and Location Area change records, but it also mentions that Call Data Records (CDRs) are being used, which are "already produced by the charging system of the telecom infrastructure when users make phone calls, send/receive messages/emails or browse web pages." So, unless the authors were mistaken, this confirms that AirSage was only providing billing records.

Another study (Qiu & Cheng 2007), based in Shanghai, used data from a region covering 9 million mobile phone subscribers. According to the authors, "Real-time cell phone data from A interface of the GSM wireless network have been archived from March of 2005 to now",

and they describe how location updates – including periodic updates and changes of Location Area – can be retrieved from the A Interface. Thus it appears that the method described in this chapter, or one very similar to it, may actually be deployed in Shanghai. However, the paper does not provide any details about who collected the data, how many records were available, or how they dealt with the IMSI/TMSI conversion problem, so there is still some uncertainty about what method was used.

4.4.2 Sample rate

One disadvantage of passive tracking versus active tracking is that there is less control over the rate at which location updates are received. This can lead to inaccurate results if handsets are moving rapidly, since the most recently recorded location can be some distance from the handset's current position.

However, the rate at which location updates are received can be influenced by the configuration of the network. When a handset selects its current cell it also listens to that cell's Broadcast Channel. This channel sends information such as the cell's CID, LAI, and carrier, but also parameters such as the location update mode and update period (Lin & Chlamtac 2001).

The update mode within a Location Area can be one of “always update”, “periodically update”, or “never update” (Naor & Levy 1998), but in practice mobile phone networks are always set to “periodically update”, typically with a period between thirty minutes and two hours. By reducing this period, updates will be sent more frequently by handsets, thereby improving the timeliness of the data.

Another option is to use the “always update” mode. This tells a handset to send a location update every time it changes cell. From a location tracking point of view, this is the ideal setting, since it guarantees that a handset's recorded CID will always reflect its current position. And since the updates occur at the point of cell change, it can be assumed that the handset's location at the time is on the boundary of two cells, potentially allowing for greater location accuracy.

The drawback with this mode, and also with higher frequency updates, is the additional traffic it generates on the network. Data collected for this thesis detected 17 distinct LAIs in the Melbourne area but 738 different CIDs, indicating that cell changes would occur over 40 times more frequently than LA changes.

Networks are configured so that during busy hours the load on a BSC is roughly divided into 20-25% on call activities, 10-15% on paging and SMS, 20-25% on hand-off and location updates, and 15-20% on system monitoring, with a target of 80% utilization during peak times (Lin & Chlamtac 2001). Since location updates already account for 20-25% of network traffic, increasing them 40-fold would require a significant upgrade of network capacity, probably at great expense. As a result, use of the “always update” mode is not likely to be feasible.

4.5 Conclusions

This chapter provided an overview of how a mobile phone network operates, and showed how the data flowing through it could be used to track mobile phone locations. It showed that the passive scanning of network traffic is theoretically possible, and is more practical if the VLR and MSC are separate pieces of equipment or provide a diagnostic interface. Such scanning would provide a steady stream of IMSI/LAI/CID values, typically at least one per IMSI per hour, without placing additional load on the network.

The next chapter describes a case study where this data is simulated using the locations of Australia's mobile phone cells and recent population data.

5 Chapter Five: Case study – Simulation of mobile phone data

5.1 Introduction

This chapter describes a case study that simulates the movement of Australia's mobile-phone-owning population and generates the passively-scanned location data that would result. The chapter covers the methodology and limitations of the study, with the analysis and evaluation of the data covered in chapter six.

5.2 Background

Because real-world mobile phone data for a population was not available for analysis, it was decided to simulate the data instead. Although a simulation may differ from what happens in the real world, it does have the benefit of providing the exact location of all the simulated handsets, which is not available with real data. And knowing these exact positions allows the accuracy of cell-based positioning to be measured all across Australia.

A simpler simulation of this kind was carried out by Caceres *et al.* (2007), who simulated vehicle movements along a stretch of road between the Spanish cities of Huelva and Seville, and generated the resulting mobile phone signals. Traffic speeds and volumes were determined by historical traffic data, and all handsets were assumed to belong to the same GSM network. It is not clear whether they used actual GSM antenna locations, or created their own.

A similar approach was taken here, except the simulation covered Australia's multiple overlapping mobile phone networks, both GSM and 3G, and the handset movements were not limited to roads. The simulation may have been more accurate if the handsets *had* been limited to roads, but simulating 24 hours of realistic traffic movements along every road in Australia was not feasible.

5.3 Mobile phone cell locations

In order to simulate the movements of mobile phones in Australia it was first necessary to build a database of all mobile phone cells, and each record in the database must contain the following information -

- The spatial coordinates of the cell's antenna.
- The maximum range of the antenna.
- Whether the antenna is directional, and if so, what direction does it point in and what is its angle of coverage.
- The cell ID (CID) and Location Area Identifier (LAI) of the cell
- Does the cell use the GSM or 3G protocol?
- Which mobile phone network (i.e. carrier) does the cell belong to?

Using this data it is then possible to estimate which cell a handset resides in, given the handset's location, carrier, and band (GSM or 3G). The assumption is that a handset will select a cell from those with the same carrier and a compatible band, and use the one with the closest antenna. Although Trevisani & Vitaletti (2004) have demonstrated that handsets will only use the closest antenna 63 percent of the time, no better method could be found for selecting a cell's handset.

The source of much of the cell data was the Australian Communications and Media Authority (ACMA) Radiocommunications Records of Licences (RRL) Database on CD-ROM (Australian Communications and Media Authority 2009b). This database provides a registry of all radio antenna licences in Australia, covering every antenna that transmits on licensed frequencies. Each licence contains information such as the spatial coordinates of the antenna (latitude and longitude), its operating frequency, its owner (licensee), and the direction it points in (if it's directional).

The RRL database contains 411,219 entries, including television, AM/FM radio, and emergency services, so the first step was to extract only those entries that correspond to mobile phone antennae, and also to remove entries that appear to be duplicates.

Mobile phone antennae in the RRL database were identified based on the assigned frequency of their licences. In Australia GSM towers transmit in the frequency ranges 935-960 MHz and 1805-1880 MHz (Wikipedia 2009b), while 3G towers transmit in the ranges 869-894 MHz and 2110-2170 MHz (Wikipedia 2009a). Antennae were selected that fell within these ranges and, based on their frequency, were flagged as being either GSM or 3G. The validity of these selections were verified by checking that their mode in the RRL database was *transmit*, not *receive*, which proved to be accurate in all cases. In all there were 61,664 entries that met the criteria.

A number of these mobile phone antennae provided overlapping coverage from the same tower, perhaps to deliver a strong directional signal in areas with unfavourable terrain. But since these overlapping entries result in ambiguous cell IDs where they occur, they were discarded, leaving 45,715 entries.

Before saving the selected RRL entries to the cell database, it was first necessary to change the coordinate system of the antenna locations. The RRL database expressed all locations in the AGD66 coordinate system, which is Australia-specific and largely obsolete. To ensure compatibility with other geospatial data such as GPS readings and Google Maps, these locations were converted to the GDA94 coordinate system before being stored in the cell database.

5.4 Mobile phone carriers

For most mobile phone antennae it was possible to identify their carrier based on the name of the licensee in the RRL database. However, there were often several different licensees with similar names, so some assumptions were made. The following licensees were assumed to be Telstra -

- Telstra Corporation Limited
- Telstra Corporation Ltd attn R Preston
- Telstra Corporation Ltd/VIC SMR
- Telstra 3G Spectrum Holdings Pty Ltd - Mr Noel Eldridge
- Telstra Corporation Ltd

These were assumed to be Optus -

- Singtel Optus Pty Limited
- Singtel Optus Pty Limited Attn Andrew Smith

And the following three were assumed to be Vodafone -

- Vodafone Network Pty Ltd
- Vodafone Pacific Pty Limited
- Vodafone Holdings Australia Pty Ltd

For the “3” network, also known as “Hutchison”, there was a single licensee “Vodafone Hutchison Australia Pty Limited”. This licensee was the result of a May 2009 joint venture between Vodafone and Hutchison to share infrastructure. Although it will eventually merge operationally with Vodafone, at the time of writing it was still a distinct network with a distinct carrier code, and will thus be treated as a separate carrier known as “Hutchison”.

There were also 66 licensees that fell into the category of “other”, covering 792 mobile phone antennae. Because those antennae could not be linked to any of the four national carriers, they were discarded from the dataset. For reference, the names of the discarded licensees are listed in Appendix A.

5.5 Cell coverage areas

Given the spatial coordinates of each antenna, the next step was to determine their coverage area, which in this chapter refers to the region where their signal can be picked up by a handset. Note that this is different to an antenna's cell area, which is the region where the antenna has the strongest signal amongst all the antennae that a handset is able to communicate with.

When calculating an antenna's coverage area, the important parameters are the antenna's location, its angle of coverage, and its range. The location was provided by the RRL database, but in order to determine angle of coverage and range some heuristics were needed.

Most studies that have discussed antenna coverage areas have limited themselves to GSM networks, where all the antennae are omnidirectional and assumed to be of equal power (Caceres *et al.* 2007, Candia *et al.* 2008, Rose 2006, White & Wells 2002). These studies have also only discussed regions where mobile phone coverage is continuous, so cell boundaries always occur along lines separating neighbouring antennae, resulting in a map of Thiessen polygons each containing a single GSM antenna. As a result there was no need to know the maximum range of each antenna.

The basic assumption of these studies is that a handset will communicate with the nearest GSM antenna, and that was the approach used in the simulation, even though Trevisani & Vitaletti (2004) showed it only occurs 63 percent of the time. However, unlike those studies, the simulation also needed to cover regions where mobile coverage ends, so antenna ranges had to be established, as described later.

Calculating the coverage areas of the directional antennae used by 3G networks required more

effort, because as well as their range, their angle of coverage needed to be established.

Only one example could be found of research discussing the coverage areas of directional antennae (Ratti *et al.* 2006), and it assumed a fixed 120 degree coverage arc centred on the antenna's azimuth, and a range of 400-500 metres. While this assumption was suitable for displaying the location of handsets around Milan, it was not suitable for the simulation. In the RRL database antennae are not spaced at exact 120 degree intervals on a tower, so there would be gaps in coverage. Also, a range of 400-500 metres is inadequate in parts of Australia where towers are separated by tens of kilometres.

As a result, some assumptions had to be made about the angle of coverage for directional antennae, making use of the azimuth information provided by the RRL database.

- Directional antennae are usually designed to cover 120 degrees, so their coverage arc is set to be *at least* sixty degrees left and right of the azimuth.
- Directional antennae at the same site, owned the same carrier and on the same frequency band, are assumed to provide 360 degree coverage around the tower. Thus the left and right angle of coverage of each antenna must reach at least half way to its left and right neighbours, respectively, up to a maximum of 90 degrees (on the assumption that directional antennae can't transmit backwards).

Calculating the range of each antenna was more complicated. Values aren't provided by the RRL database, and the range of a mobile phone antenna depends on many factors, so their range was estimated using the following constraints -

- The GSM protocol works by dividing each frequency into eight time slices, and because of speed-of-light delays mobile phones must delay their signals to synchronize with their allocated slot. The maximum delay allowed by the GSM protocol is 233 microseconds, and during this time a radio signal travels 70km. Thus the maximum range of a GSM antenna is half this round trip distance, or 35km (Lin & Chlamtac 2001).
- The maximum range of a 3G antenna is really only constrained by the transmitting power of the receiving handset (since two-way communications are needed). Typically, given line-of-sight and clear weather, this is about 80km (Johnson 2008).
- GSM antennae are usually omnidirectional, and in urban areas are laid out in a hexagonal pattern. Each cell and its six neighbours use a different subset of their

carrier's allocated spectrum, allowing patterns of seven cells to be repeated without cell interfering with their immediate neighbours. Thus the distance between a GSM antenna and its sixth-closest neighbour forms an upper bound on the radius *r* of the hexagon. In a hexagonal layout, the nearest pair of antennae using the same frequencies will be separated by 2.6*r*, so setting a range limit of *r* will ensure full coverage of the area with no interference. Thus the maximum range of a GSM antenna is assumed to be the distance to its sixth-closest neighbour.

- For Telstra's 3G rural antennae, the RRL database provides their actual operating frequency. To ensure that an antenna will not interfere with a neighbouring antenna, its range is constrained to 70% of the distance to the nearest antenna on the same frequency in the direction that it's pointing.
- Urban 3G antennae are constrained using the same criteria as GSM antennae, i.e. the distance to their sixth-nearest neighbour.

5.6 Cell identifiers

In the real world, a mobile phone cell has four numeric identifiers, which are broadcast to all handsets within range. These are the identifiers used -

- Mobile Country Code (MCC), set to 505 in Australia
- Mobile Network Code (MNC), 1 for Telstra, 2 for Optus, 3 for Vodafone, 6 for Hutchison “3”
- Location Area Identifier (LAI)
- Cell ID (CID)

Since the simulation uses Australian data, the MCC of all cells in the simulation is set to 505. And because a cell's carrier can be determined from its licensee, their MNC is set accordingly. But the LAI and CID of the cells are not specified in the RRL database, nor are they published by the carriers themselves.

Although it was not strictly necessary to use accurate cell IDs in the simulation, it would allow results to be directly compared with data sampled in the field with smart phones. And having accurate LAIs is very important, since handsets send a location update when they move to a cell with a different LAI, and it is these location updates that are being simulated.

Fortunately, the CID and LAI of some of Australia's cell have been collected by the OpenCellID project (OpenCellID 2009). This is an open source project whose goal is to crowd-source cell identifiers worldwide. Volunteers install special software on their GPS-equipped mobile phones that periodically logs the latitude, longitude, and four numeric identifiers of their current cell. These values are then uploaded to the OpenCellID website where they are made publicly available. As of 19 September 2009 the OpenCellID database contained 1,466,186 measurement from Australian networks, covering 11,259 distinct cells.

Assigning these OpenCellID values uniquely to antennae from the RRL database is not a straightforward task, since a particular spatial coordinate may sit within the coverage areas of more than one antenna from the same carrier. That means that the cell ID measured at that coordinate could belong to one of several antennae.

For example, consider the configuration in figure 5.1, where a mobile phone tower contains one omnidirectional GSM antenna “A” and three directional 3G antennae “B”, “C” and “D”, all owned by the same carrier. The coloured areas indicate the coverage area of each antenna and the numbers indicate cell IDs measured at different locations. Note that 3G handsets can also operate in GSM cells if there is no 3G reception.

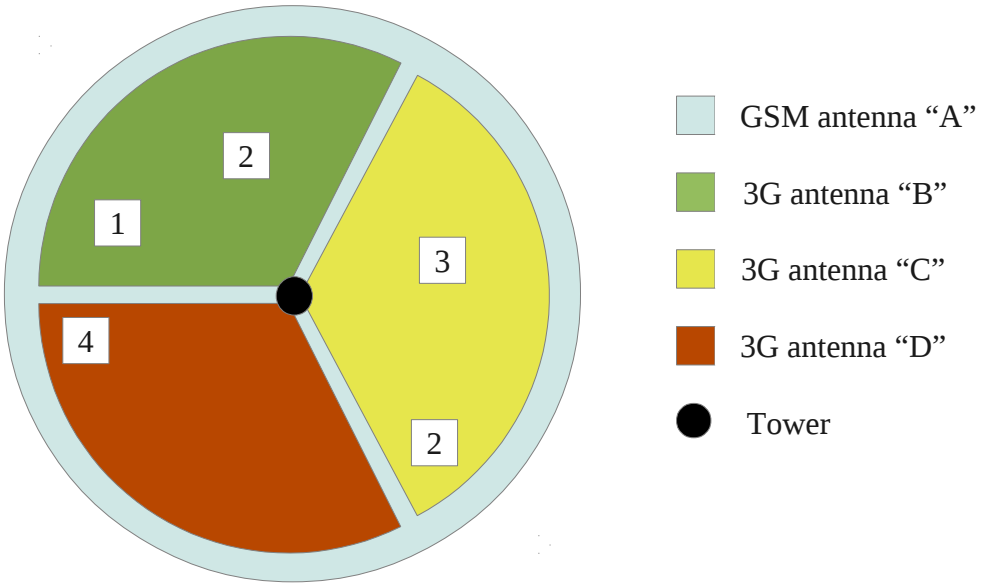

Figure 5.1: Example tower configuration.

First, note that the cell ID readings all fall within the overlapping coverage areas of two antennae (antenna “A” and one other), so it is not immediately obvious which cell ID belongs to which antenna. In reality, it's likely to be even more complicated, since reading **4** above would probably fall within the overlap zone between antennae “B” and “D”, meaning there

would be three antennae to choose from.

Second, this diagram doesn't show the coverage areas of antennae on nearby towers. Once they are taken into account, each reading could fall within a few more coverage areas.

To help allocate the OpenCellID values to antennae, scores were assigned indicating the likelihood that a particular ID belongs to a particular antenna. These scores were calculated by "the fraction of a cell's readings that fall within the antenna's coverage area" multiplied by "the fraction of readings within the antenna's coverage area that belong to the cell" divided by "the average distance of the cell's readings from the antenna".

Each cell is then assigned to the antenna with the highest score that doesn't already have a cell assigned. In the example above, cell **1** would be assigned to antenna "B", cell **2** to antenna "A", cell **3** to "C", and cell **4** to "D".

Using this algorithm, 9,823 of the 11,259 cells in the OpenCellID database were successfully assigned to antennae. The remaining 35,892 antennae were assigned random cell IDs and the same LAI as the nearest antenna with a known LAI that was owned by the same carrier.

5.7 Virtual handsets

The purpose of the simulation was to move virtual handsets around a simulated Australian mobile phone network and log the location updates that would be generated. Having creating the simulated cells, the next step was to generate the handsets at their starting location, and assign them a carrier and whether they were GSM or 3G.

In order to make the starting locations of the virtual handsets match the actual locations of Australia's handsets as closely as possible, they were generated based on 2008 Australian census data (Australian Bureau of Statistics 2008). The finest-granularity census data available was based on *Collection Districts*, regions containing approximately two hundred households. Each of the 38,704 Collection Districts in Australia had a boundary described by one or more polygons and a population estimate for June 30th 2008.

Based on Australian mobile phone ownership rates, 71.4 percent of each collection district's population was assumed to own a mobile phone. These phones were generated at random locations evenly distributed throughout the collection district's boundary and assigned a carrier and band based on the market shares in table 5.1. Also, working on the assumption that people will only use a phone that has reception at home, handsets were only assigned carriers and bands that provided coverage at their starting location.

There were 32,501 handsets (0.23% of the total) generated in locations that didn't have coverage, and these were assigned to Telstra 3G on the assumption that they were located in remote rural areas, where Telstra provides the best coverage.

Market share	GSM	3G
Telstra	22.3%	18.8%
Optus	23.8%	8.6%
Vodafone	14.7%	3.2%
Hutchison "3"	0.0%	8.6%

Table 5.1: Market share by carrier and band (Australian Communications and Media Authority 2008).

The end result was a table of 14.1 million virtual handsets spread throughout Australia, as plotted on Google Maps images in figures 5.2 and 5.3. The details of these handsets were saved to a database for later use.

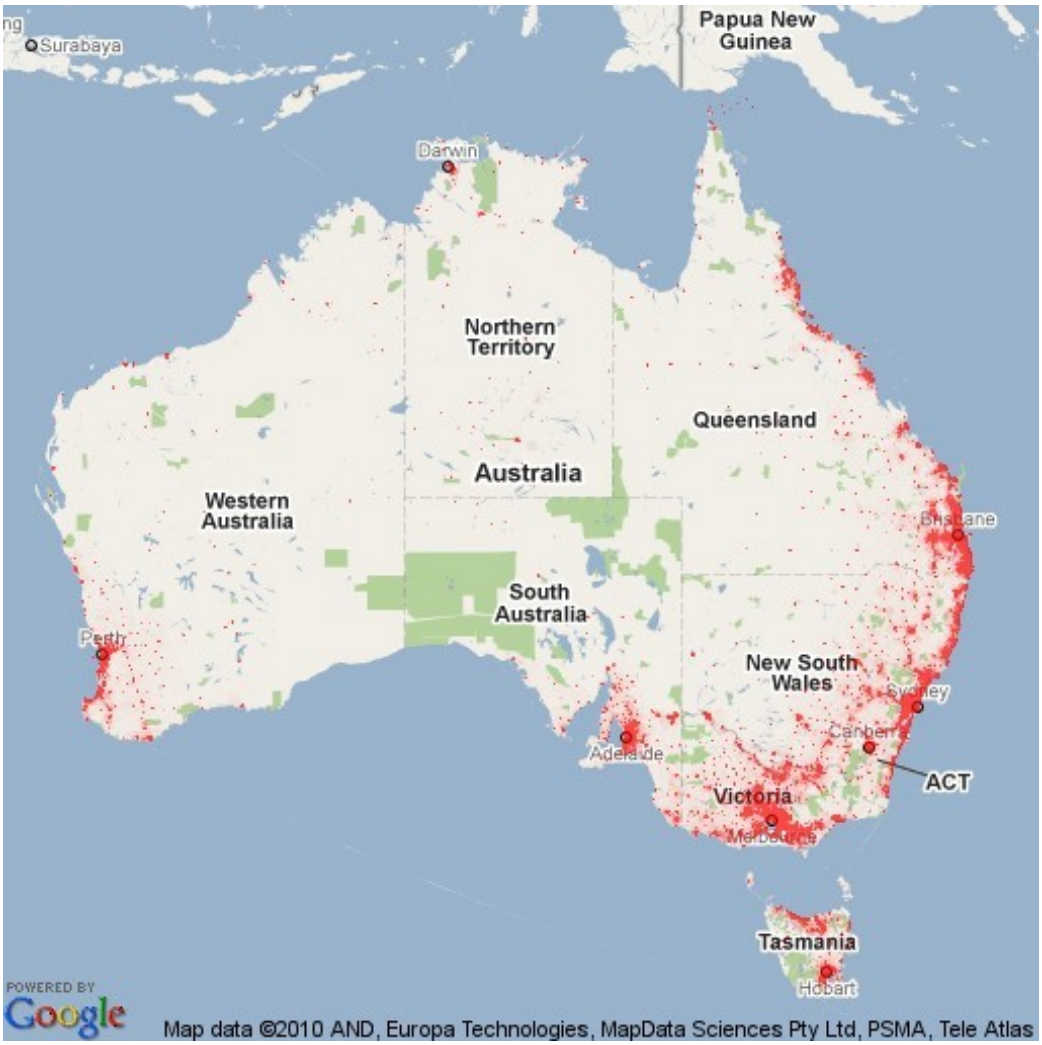

Figure 5.2: Starting locations of the virtual handsets, Australia-wide.

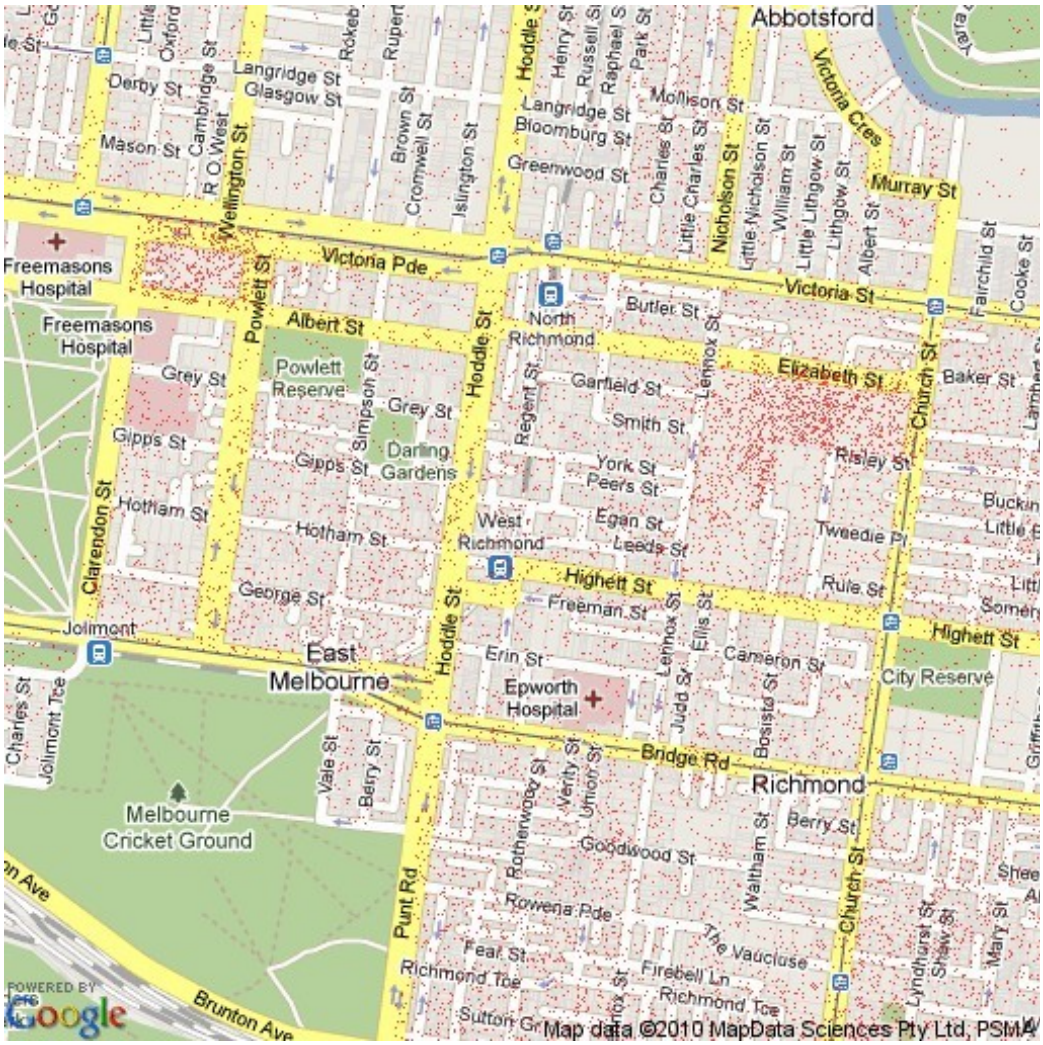

Figure 5.3: Starting locations of the virtual handsets, inner-city Melbourne.

In the map of inner-city Melbourne in figure 5.3, the locations of the handsets appear to be a reasonable approximation of reality, with high density in the multi-storey apartments on Albert and Elizabeth Streets, and low density in the light-industrial area of Collingwood north of Victoria Parade. Note, however, that choosing random locations within Collection Districts resulted in a number of handsets being located on streets and within parks.

5.8 Handset movements

When simulating handset movements there were a numbers of goals. First, the simulated movements should be similar to real-world movements. Second, the movements need to be controlled by equations that allow a handset's position to be determined exactly at an arbitrary time *t*, so that actual positions could be compared with cell-derived positions at any time. This

precluded algorithms where movements are calculated moment-by-moment based on a handset's immediate surroundings.

Finally, there should be a diverse range of movements - starting at different times, and moving in different directions at different speeds - with the goal of testing the effectiveness of mobile phone location data across a wide variety of circumstances.

The algorithm chosen was for the movement of each handset to be an out-and-back journey driven by a set of parameters derived from the handset's unique ID. These handset IDs were allocated sequentially, starting at zero. The reason for using parameters based on a unique ID was that the handset's actual movements could be easily reconstructed from that ID at a later date and compared with movements derived from a sequence of cell IDs.

The following five parameters were used -

- N_1 (range 0 ... 11). The direction of movement, in degrees clockwise from north, was set to $30N_1$. This direction was reversed on the return leg of the journey. This parameter ensures that handset are moving in all directions.
- N_2 (range 0 ... 14). The speed of movement was set to 1.67 (N_2 + 1) m/sec. This results in speeds ranging from walking pace (6 km/h) to highway speeds (90 km/h), corresponding to movement speeds that might be seen in the real world..
- N_3 (range 0 ... 17). The time of day that the journey starts was set to 21600 + $600N_3$ seconds past midnight. In other words, some time between 6 and 8:50am, corresponding to the period when people usually depart for work or school.
- N_4 (range 0 ... 5). The duration of the outbound and return legs was set to 900 (N_4 + 1) seconds, giving a value between 15 minutes and 1.5 hours, representing the time taken to get to and from work or school.
- N_5 (range 0 ... 11). The duration to wait between completing the outbound leg and commencing the return leg was set to 21600 + $1800N_5$ sec, i.e. somewhere in the range 6 to 11.5 hours, representing the time spent at work or school.

All possible values of these parameters can be uniquely encoded in the numbers 0 ... 233279 using the equation $N_1 + 12N_2 + 180N_3 + 3240N_4 + 19440N_5$. The parameters for a particular handset can thus be determined from the handset's unique ID modulo 233,280.

Using these parameters, the aim was to simulate the movements of 14.1 million virtual

handsets for a virtual day. Unfortunately, this proved to be too computationally intensive to run to completion, running out of memory before finishing. Even if the memory issue could be resolved, the running time was likely to exceed a week.

As a result, it was decided to run the simulation on a randomly-selected 10 percent subset of the handsets. This subset of 1,414,710 handsets contained at least six examples of each of the 233,280 possible movements, so it was decided that the handset movements would have the same statistical behaviour as a full simulation, and would be sufficient for the purposes of this study.

During the simulation a location update was generated whenever a handset changed its Location Area, or whenever an hour passed (starting from a time randomly-generated for each handset between midnight and 1am). The result was a database of just over 40 million location updates. This amounted to 28.9 records per handset, in line with what would be expected from a day's worth of hourly updates plus occasional changes of Location Area. The data is analyzed in more detail in the next chapter.

5.9 Limitations of the simulation

Although attempts were made to make the simulation as realistic as possible, it does vary from reality in a number of ways. These differences should not materially affect the results of the simulation, but they have to be acknowledged.

It only simulates periodic updates and change of Location Area. The simulation generates a new record whenever a virtual handset changes Location Area, or periodically every hour. In the real world, however, records would also be generated whenever a handset is switched on, makes or receives a call, sends or receives an SMS, or accesses the internet. In theory these events could be simulated, but without access to accurate data about how often they occur, it was decided that they would not add to the accuracy of the simulation.

The Location Area Identifiers may not be correct. Cell IDs and LAIs were assigned based on OpenCellID data where possible, but this only covered 20 percent of the antennae in Australia. The rest of the antennae were assigned an LAI interpolated from surrounding antennae whose LAI was known. As a result, the Location Area boundaries for large parts of Australia are probably incorrect.

The simulation updates in one-minute increments. The simulation works by updating the position of each handset every minute, and generates a new record if that movement caused

the handset to change Location Area. In reality, a handset would be constantly measuring the signal strength of surrounding cells and could switch Location Area at any time, not just on a one-minute boundary. These updates could theoretically be used to position a handset more accurately, since the handset would be somewhere on the Location Area boundary at that exact time.

Handset starting positions are randomly distributed within Collection Districts. The starting positions of the handsets are generated at random locations within the census Collection Districts, resulting in a uniform distribution within each district. An average, this is fairly realistic in urban Collection Districts (although many handsets end up on streets and in parks), but might be less so in large rural districts where populations are more clustered.

Handset movements are unrealistic. The handsets in the simulation move in a there-and-back pattern at a constant speed in a random direction, regardless of terrain. This is easy to simulate and analyze, but in reality handsets don't move like that. Rather, they tend to follow roads or public transport routes and move in a stop-start manner. Also, major transport routes are generally well served by mobile phone towers, so by not following those routes the simulation may move handsets further away from antennae than they would normally go.

Cell coverage areas are approximate. In the real world, handsets usually select their cell based on the antenna with the strongest signal they can detect, but many factors affect signal strength, such as the transmitting power of the antenna, the antenna's height, and intervening terrain. The simulation assumes that all antennae transmit at the same power in all directions, over homogeneous terrain, and as a result, when the coverage areas of two antennae overlap, the boundary between the two cells is half way between them. As discussed in section 2.3, in reality the boundary is likely to be different, varying with terrain, the relative power of the antennae, and the cell selection algorithm used by the handset (Trevisani & Vitaletti 2004).

Cell ranges are approximate. The simulation assumes that each antenna has a fixed range in all directions, and that all handsets within that range will receive a signal while those outside won't. However, the fixed range is based on rules of thumb that may not always be accurate, and in the real world actual range is affected by factors such as terrain and the power of the receiving handset (Trevisani & Vitaletti 2004).

The simulation also assumes that antennae transmit with the same power in all directions, resulting in cells with circular or circular arc boundaries when there is no overlap with other cells. Although this is generally true for omnidirectional antennae, directional antennae tend

to transmit with more power in the direction they're facing (Lin & Chlamtac 2001). For example, a directional antenna transmitting in a 120 degree arc may produce several dB more power straight ahead than it does 60 degrees to the side, resulting in a cell shaped like figure 5.4. When three of these antennae are positioned on a tower, spaced 120 degrees apart, the resulting coverage area would be not quite circular.

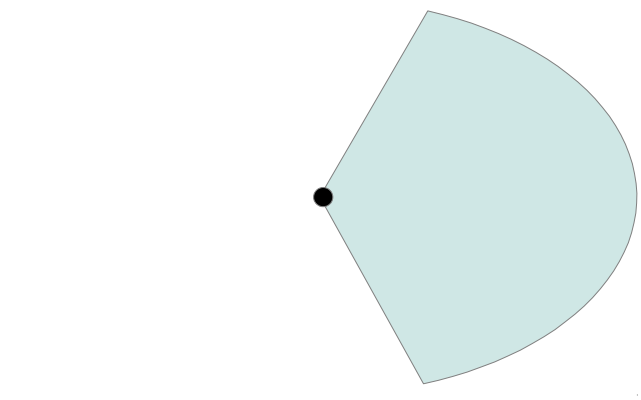

Figure 5.4: Cell shape for a directional antenna.

5.10 Conclusions

This chapter described a case study that simulated the generation of mobile phone location data from the movements of virtual handsets. It was based on Australian census data, the location of Australia's mobile phone antennae, and a simple population movement algorithm.

The next chapter looks at the accuracy of this simulated data.

6 Chapter Six: Case study – Analysis of simulated data

6.1 Introduction

The aim of this chapter is to analyze the simulated data from the previous chapter to see how accurately handsets can be located using cell IDs. The positions derived from the simulated cell IDs will be compared with the known handset positions that drove the simulation to provide an estimate of location error.

6.2 Handsets location estimates

The locations of 14.1 million handsets were generated, and, for a random 10 percent subset, their movements were simulated over a 24 hour period, generating 40 million location update records. Each location update record contained a time stamp, a handset identifier, and the ID of the cell the handset was in. These records can be used to estimate the position of a handset at a point in time from the location of its last known cell.

In theory, a handset's location could be determined more accurately by interpolating between multiple location records, or by exploiting the fact that change-of-Location-Area records must occur on the boundary of the two Location Areas. But for the sake of simplicity, in this study each handset's location is derived only from its last known cell.

When it comes to deriving a handset's location from a cell ID, there are, broadly speaking, two methods – using *cell geometry* and using *experimentally measured locations*. Both have their benefits and drawbacks.

6.3 Cell geometry

With cell geometry techniques, the position of a handset is calculated using the shape of the cell's coverage area. The coverage area is itself calculated using the position and other characteristics of the cell's antenna, the position and other characteristics of neighbouring antennae, and possibly details of the surrounding terrain.

The easiest, if not the most accurate, way to estimate a handset's location is to simply use the location of the cell's antenna. This is a fairly sensible approach for cells with omnidirectional antennae – which are used by default on GSM networks – but is less accurate for directional antennae, since it results in handsets being located at the vertex of a wedge-shaped cell.

A more sophisticated technique is to estimate the shape of a cell's coverage area and use the geometric centre of the shape as the location. The shape is usually based on the location of neighbouring antennae, but can also take into account the transmitting power and height of the antennae, as well as the terrain. The handset location is then usually assumed to be at the coverage area's centre of gravity, or *centroid*.

This technique is used by Telstra's location management interface (Telstra Corporation Limited 2006). This runs on an Ericsson platform that, for some cells, can also estimate a handset's distance from the antenna. In response to a query for the location of handset, the interface will return both the shape of the region containing the handset and the “centre” of the shape. Depending on the type of antenna serving the cell, the shape could be a polygon, a circle, or a circular arc, as shown in figure 6.1.

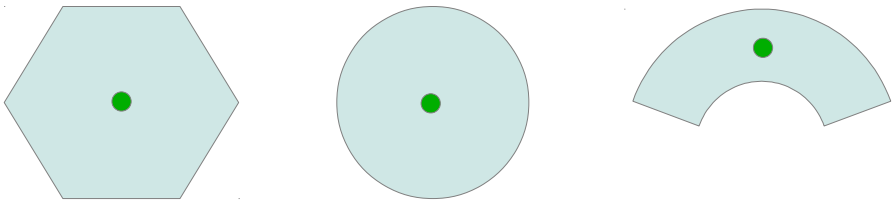

Figure 6.1: Polygonal, circular, and arc regions, and the “centres”.

For polygons and circles, the Telstra interface uses the shape's centroid to define its “centre”. However, for circular arcs the centre is specified as being half way between the inside and outside radius, on the centre angle line. This is not actually the mathematical centroid of the shape, but it ensures that the point - which is used as the best-guess location of the handset – actually falls within the shape's boundary.

The geometric centre method has the benefit of being straightforward to calculate, but when the handset's distance from the antenna is not known (as is the case when using passively-scanned location data) it requires an accurate estimate of the shape of the cell's boundary. Cell shapes can be estimated based on antenna locations and relative transmission power, but there will usually be errors due to terrain (Trevisani & Vitaletti 2004).

Using cell data from the simulation, cell boundary polygons were calculated using an algorithm that assumed that each point was covered by the cell with the closest antenna pointing in its direction. The algorithm was designed and implemented by the author, and written in the C++ programming language. It worked as follows -

1. A hexagonal polygon was created for each cell, centred on the antenna and with a radius equal to the antenna's range. Ideally this starting shape should have been a circle, but given the need for polygonal boundaries, a hexagon was a compromise between accuracy and simplicity.
2. For directional antennae, a wedge was cut out of the hexagon corresponding to the antenna's angle of coverage. This wedge of the polygon was retained.
3. In the neighbourhood of the cell, for each antenna which was owned by the same carrier and on the same band, a line was drawn that ran equidistant between it and the cell's antenna. In other words, a line was drawn that ran perpendicular to a line connecting the two antennae, intersecting at the half-way point. If this equidistant line intersected with the cell's polygon, then the polygon was cut in two along the line, and the piece containing the cell's antenna was retained.
4. The polygon remaining at the end of this process was used as the cell boundary.

Using the centroids of these polygons to estimate the starting positions of the 14.1 million handsets resulted in an average error of 3,372 metres, significantly higher than the 2,545 metres using antenna locations alone. However, using a mixture of antenna locations for omnidirectional antennae and polygon centroids for directional antennae reduced this to 2,483 metres (see the red line in figures 6.4 and 6.5 for details). It is unclear why polygon centroids are so inaccurate for finding the centre of omnidirectional cells, but they are clearly less useful than simply using antenna locations.

6.4 Experimentally measured cell locations

Apart from requiring prior knowledge of a network's antenna layout, the main drawback to using cell geometry is that it assumes an even population distribution within a cell's coverage area, when in fact the population may be clustered. For example, imagine a cell covering a lake containing a populated island. The geometric centre of the cell is probably in the lake itself, but the actual location of a handset in that cell is almost certainly on the island.

An alternative to using cell geometry is to experimentally measure the average position of handsets within each cell, taking repeated samples of latitude/longitude/CID with GPS-equipped mobile phones. This was the approach taken by OpenCellID (2009), a crowd-sourced project involving thousands of volunteers around the world.

Once samples have been collected, they are grouped by cell ID, with the set of each cell's

latitude/longitude readings describing its approximate extents. The average position of the samples can be used as the cell's centre, and the standard deviation of the samples can be used as an error estimate.

The benefit of this method is that it requires no prior knowledge of a network's physical layout or cell IDs, which is often commercially-sensitive information. It also uses locations where handsets *actually* existed rather than where they w*ould have* existed if they were uniformly distributed throughout a cell, so estimates of handset locations are more likely to match reality.

The main drawback is that this technique relies heavily on the quality and quantity of the samples. If there are no samples for a particular cell ID, for example because no volunteer has passed that way while subscribed to the cell's carrier, then no location information can be provided for that cell. In the case of the OpenCellID data, only 11,259 of Australia's 61,664 cells have been sampled.

There is also the possibility of sample bias, since location estimates are based on the movements of the volunteers who provide the data, which may be different to the movements of the people whose locations are being queried. For example, many cells may be sampled only when volunteers drive through them, resulting in samples that follow main roads. But the majority of the location requests may come from people who actually live within the cell, and who may be located in residential back streets. The resulting location estimates will thus be near the main roads, which is the average position of the samples, whereas the average position of the handsets is actually in the residential areas.

When handset locations were calculated from the simulation data using this technique, the average error for the starting point of the handsets was only 2,010 metres. Note that this is not a realistic result, since the handsets' positions were being used to estimate their own positions, and cells containing a single handset would produce error-free results. However, it does provide a rough lower bound of how accurate cell ID locations *could* be if a comprehensive set of location samples were available. See figures 6.4 and 6.5 for more details, in particular the green line.

6.5 A hybrid approach

A mixture of the two methods, geometric and experimental, was also tested using the simulated data. As before, the location of a cell with an omnidirectional antenna was taken to

be the antenna itself, since it broadcasts equally in all directions to form a roughly circular or hexagonal cell, and is thus close enough to the cell's geometric centre.

With a directional antenna, however, the antenna is on the edge of the cell, and the geometric centre is somewhere within the coverage area. As shown in figure 6.2, the hybrid approach assumes that the centre of a cell with a directional antenna lies along the antenna's directional vector, *d* metres in front of it. The technique, which will be referred to as “directional offset”, is a hybrid in the sense that it uses the antenna's known position plus experimentally measured locations to find the value of *d* that yields the best results.

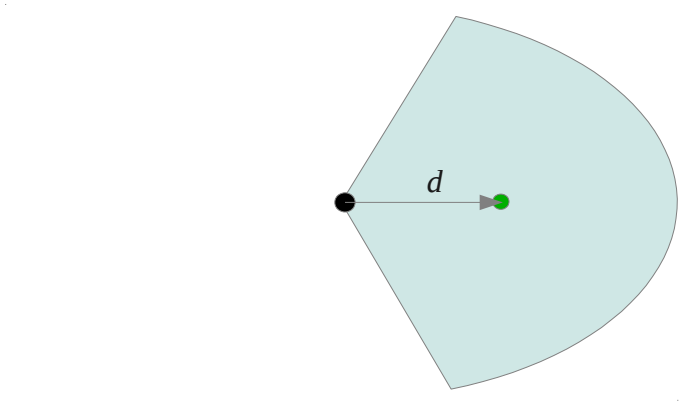

Figure 6.2: Finding the centre of a cell with a directional antenna.

The method used was to try different values of *d* and experimentally measure the average distance between the actual starting locations of the virtual handsets and the offset “centres” of the cells in which they resided. The results are shown in the dark blue line in figure 6.3. For comparison, the average location error was also broken down by 3G carrier (GSM carriers are not shown because nearly all of their antennae are omnidirectional). Note that the dark blue line representing the average location error is flatter than the carrier-specific lines because it contains large numbers of omnidirectional antennae, whose accuracy isn't affected by changes in the offset.

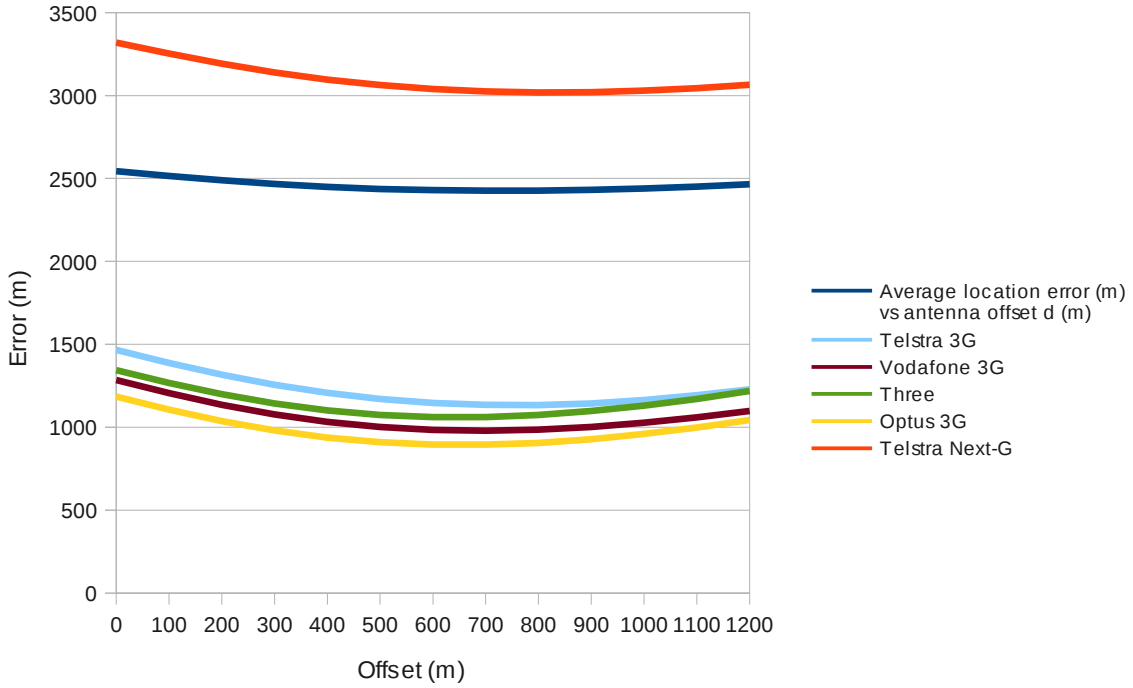

Figure 6.3: Average location error for different offset estimates of cell centres.

As can be seen on the graph in figure 6.3, when *d* is zero – i.e. handset positions are calculated from the antenna location – the average error is 2,545 metres. This decreases as *d* gets larger, until around 700 metres, when the error reduced to 2,426 metres. On a carrier-by-carrier basis, the average errors are minimized when *d* is between 600 and 900 metres, although 700 metres is not too far from the minimum in all cases.

As can be seen in table 6.1, this error is better than that obtained by using both polygon centroids (2,483 metres) and antenna locations (2,545 metres). Although it requires some effort to calculate the optimal offset – using a combination of census data, market penetration rates, and antenna locations – once the offset is known it is very easy to use it to estimate a handset's location given the cell antenna's latitude, longitude, and bearing.

Method for locating cell centre	**Average error**	**Median error**
Antenna location	2,545 metres	1,196 metres
Polygon centroid	3,372 metres	1,234 metres
Antenna location and polygon centroid	2,483 metres	1,101 metres
Average handset position	2,010 metres	894 metres
700 metre directional offset	2,426 metres	1,052 metres

Table 6.1: Average location error using different cell centre methods.

In summary, given the identity of an Australian handset's current cell, the handset can be located with an average accuracy of 2.48km using the polygon centroid technique, 2.54km using antenna locations, or 2.43km using a position 700 metres in front of a directional antennae. Note, however, that these average figures conceal a wide range of values, as shown in figures 6.4 and 6.5. These graphs show the distance error for Australia's handsets, ordered from best to worst.

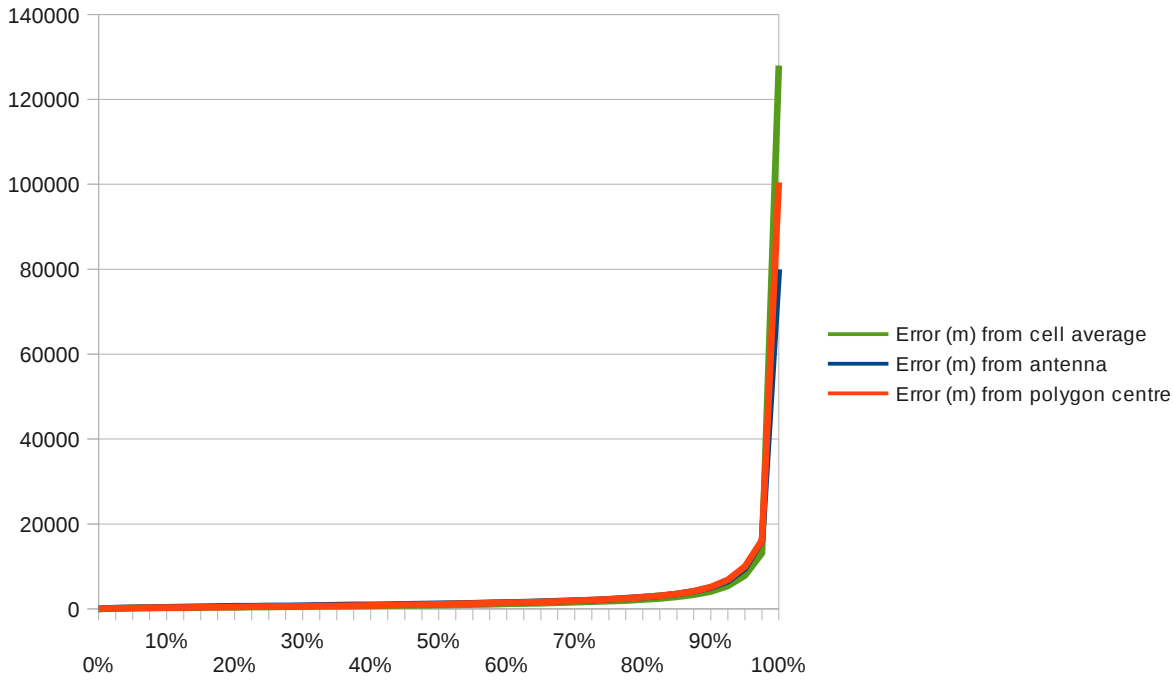

Figure 6.4: Location estimate errors for percentages of handsets.

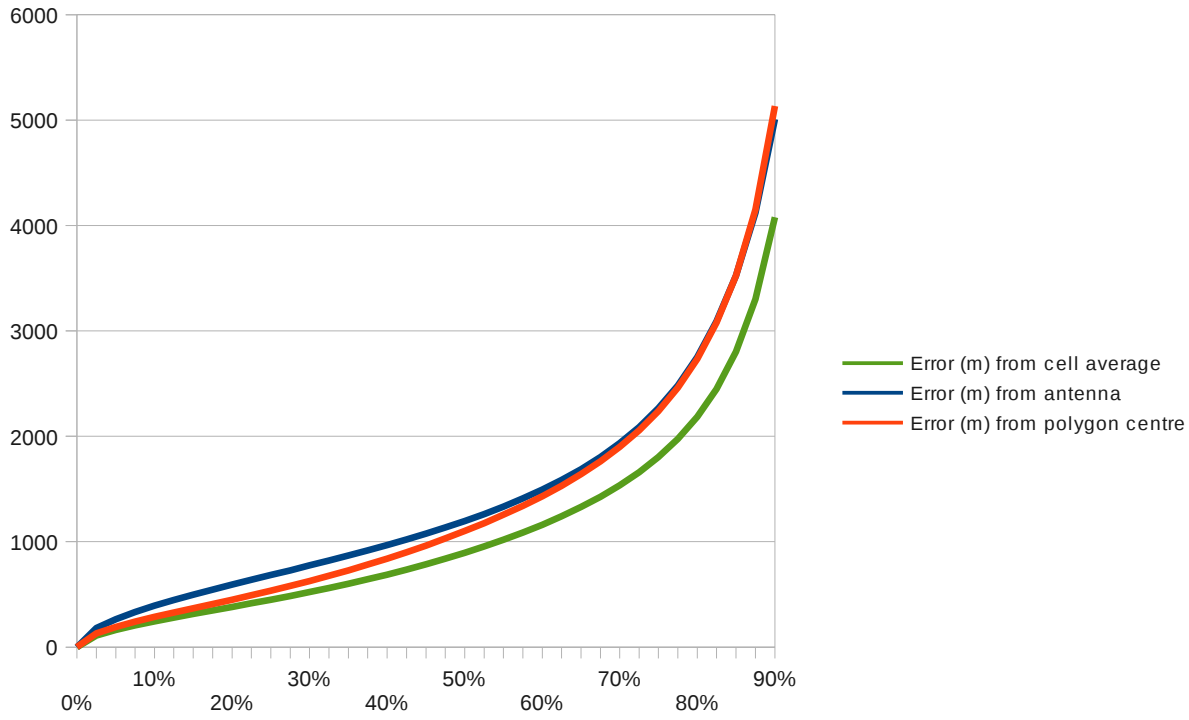

Figure 6.5: Location estimate errors for the best 90% of handsets.

Figure 6.4 shows the distribution for all handsets, while figure 6.5 shows the best 90 percent. As can be seen from the blue line in figure 6.5, the 50th percentile, or median, handset is located 1.2km from its cell's antenna, and 90 percent of handsets are located within about 5km. The small percentage that are located 50km or more away are due to handsets located in large 3G cells in sparely-populated rural areas, where antennae have a range of 80km.

For comparison, experimental measurements of cell accuracy have been calculated in the US and Italy (Trevisani & Vitaletti 2004), as shown in table 2.1 in chapter two. Note that those experiments were carried out on GSM networks with omnidirectional antennae, so antenna locations were used to estimate the handset position.

6.6 Spatio-temporal accuracy

There are two sources of error when estimating the position of a handset based on location updates. The first comes from using the coverage area of a cell as the estimate, as described above, and results in an average error of about 2.5km. The second is due to the frequency of location updates.

To reduce traffic on a mobile network, handsets only notify the network of their current cell

ID when they are switched on, change Location Area, make or receive a call, send or receive an SMS, access the internet, or periodically every hour or so. The duration between periodic updates is specified by the network, and is typically between 30 minutes and two hours. In the simulation it was set to one hour. No calls, SMSes, or internet accesses were simulated.

As a result, so long as a handset doesn't change its Location Area, it can move for up to an hour without generating a location update, or even longer if it loses coverage. If its most recent update is used to estimate its current position, the location error for a fast-moving handset could be significant. The fastest speed used in the simulation was 90km/h, so errors of this kind could potentially increase the error in location estimates by another 90km.

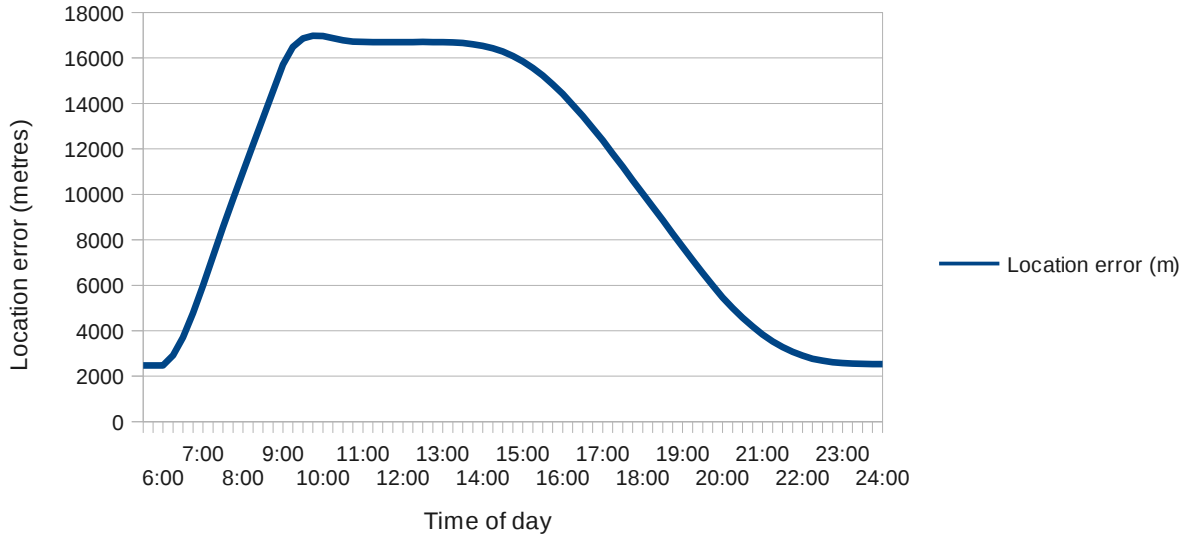

Figure 6.6: Average location error during the simulation.

The handset location errors during the simulation are shown in figure 6.6. Before 6am all handsets were stationary, so the average error was the 2.48km polygon centroid estimation error. As the day progressed, and handsets commenced their journeys between 6:00 and 8:50, the average error increased. Journey durations were between 15 and 90 minutes, so all handsets reached their initial destination by 10:20am.

The average error peaked at 16,985 metres at 9:50, then dropped slightly to level out at 16,700 metres. This slight drop was due to the arrival of hourly updates from handsets that had already reached their destination.

The average error of 16.7km around midday was higher than expected. Since all handsets at that time were stationary and had notified the network of their current cell, the average error was expected to be similar to the starting location error, around 2.48km.

However, the 2.48km average error was calculated from randomly-generated handsets that follow Australia's overnight population distribution. Since the mobile network itself follows Australia's population, the handsets tended to be fairly close to mobile phone towers. But when the handsets reached their simulated destinations – an average of 42km from their starting point in a random direction – their distribution no longer followed the mobile phone network so closely. In fact, over 20 percent of the handsets did not have any coverage, and those that did were an average of 8.6km from the centre of their cell.

For the more than 20 percent of handsets without coverage, their position was estimated using their last known cell. To achieve an overall error of 16.7km when nearly 80 percent of the handsets have an error of 8.6km, the handsets without coverage must, on average, be about 47km from their last known cell. Since handsets in the simulation could travel up to 135km, with an average of 42km, that sounds plausible.

The average error at the end of the simulation was 2,534 metres, slightly higher than the 2,483 metres at the start. The difference is due to the 0.28 percent of handsets that didn't have mobile coverage in their starting location. Because their cells were unknown at the start, they didn't contribute to the error calculation. But as they moved around during the simulation, some of them moved into the range of an antenna, and that cell was logged. Even though they left that cell when returning to their starting position, it was still recorded as their last known cell and thus used to calculate the overall error.

Because this small number of handsets had their final position estimated from cells where they no longer resided, the error in the estimate tended to be high. And since these estimates were used at the end of the simulation but not at the beginning, the error at the end was slightly higher.

6.7 Conclusions

This chapter used the mobile phone location data generated by the simulation in chapter five to determine the accuracy with which handsets can be located from cell IDs using different techniques. It found that the average error when estimating handset locations using the position of a cell's antenna was 2.55km, assuming handsets follow the population distribution described by Australia's census.

When the position estimates for cells with directional antennae were calculated using the cells' polygon boundary centroid this error was reduced to 2.48km, and when a position 700 metres

in front of the antennae was used, the average error was further reduced to 2.43km. Note that these averages incorporate a wide range of error values, including very large errors in rural areas, and that the median error is significantly lower, at around 1.2km.

The simulation of handset movements revealed average location errors of over 16km when handsets were at the maximum distance from their starting point. However, this was probably not a good estimate of the error that would occur if real mobile data were used, because the simulation moved handsets in random directions, often to locations with sparse or non-existent mobile coverage. In the real world, mobile networks tend to cover the areas where people live and work, so handsets are more likely to remain in areas with dense cell coverage and thus have lower errors in their position estimates.

The next chapter describes how the accuracy of the simulation was verified on a small scale using a programmable GPS-equipped mobile phone.

7 Chapter Seven: Verifying the simulation with real-world data

7.1 Introduction

This chapter describes how a programmable GPS-equipped mobile phone was used to collect cell information and GPS coordinates from around Melbourne. The aim was to verify the accuracy of the simulation data on a small scale using real-world data from an individual, and to test some of the simulation's underlying assumptions.

7.2 Background

Although the simulation was designed to be as accurate as possible, it had some limitations -

- It assumed that mobile phones always use the cell of their closest antenna. In reality, handsets have been observed using more distant antennae as much as 43 percent of the time (Trevisani & Vitaletti 2004).
- The simulated handsets moved on simplified, straight-line paths, unlike real movements, which would be more likely to follow roads, footpaths, and rail lines.
- The boundaries between Location Areas were estimated based on incomplete data. Since handsets generate a location update when they cross these boundaries, this would affect the rate at which the updates were generated in the simulation.

To address these limitations, real data was collected with a programmable mobile phone so it could be compared with the simulated results. The method of data collection was similar to that used by Trevisani & Vitaletti (2004) around Rome and New York, except that instead of recording data every two seconds it would record cell and GPS information at the times when a handset might send details of its current cell to the network.

However, unlike that study, carriers have not provided details of what cell ID has been assigned to each antenna, so data analysis becomes more complex and error-prone. Also, most of the cells sampled around Melbourne are likely to be using directional 3G antennae, whose coverage areas are more difficult to calculate than the omnidirectional GSM antennae around Rome and New York.

7.3 Data collection

Data was collected using a mobile phone application developed by the author called *Cell Logger*, running on an "HTC Tattoo" mobile phone as shown in figure 7.1. The Cell Logger application was written in the Java programming language, using the Eclipse development environment on a Fedora Linux PC.

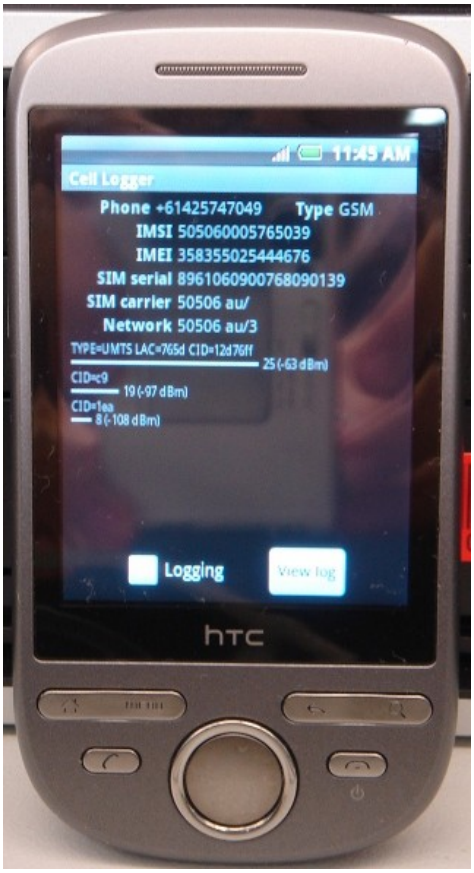

Figure 7.1: The *Cell Logger* application, running on an HTC Tattoo smart phone.

The HTC Tattoo was chosen because it was the cheapest GPS-equipped programmable handset available at the time, and it ran the Android operating system, which supports the development of complex applications. Software developed for Android phones can, among other things, poll the GPS receiver, access mobile phone cell information, and provide a graphical user interface to view and export data.

Cell Logger was designed to collect cell information from the handset as it moves around. Every 30 minutes, and whenever the handset changes to a new cell, Cell Logger records the

following information -

- The date and time, with millisecond resolution.
- The event that occurred (periodic update or cell change).
- The type of mobile phone network being used (3G or GSM).
- The mobile country code (MCC). This is 505 throughout Australia.
- The mobile network code (MNC), which identifies the cell's carrier.
- The cell's current location area identifier (LAI) and cell ID (CID).
- The signal strength of the current cell.
- The handset's GPS latitude and longitude to six decimal places (if available).
- The handset's GPS altitude to the nearest metre (if available).
- The accuracy of the GPS reading in metres.
- Details of all the other cells that are visible but not being used, including their LAI and CID (on GSM networks) or Primary Scrambling Code (on 3G networks), and their signal strength. In this chapter these are referred to as *neighbouring cells*.

7.3.1 Timing problems

Early testing of the software uncovered a few problems to do with timing. The original aim was to take a GPS reading and collect neighbouring cell information whenever the handset changed cells. However, obtaining a GPS fix can take over 30 seconds, assuming GPS coverage is even available, so the software would wait up to 35 seconds for a GPS fix before giving up. In the meantime, the handset had often changed cells again. Because this would associate GPS readings with incorrect cells, a lot of early data had to be discarded as invalid.

A similar problem occurred with the collection of neighbouring cell information. Normally the handset maintains a list of all visible neighbouring cells, and that list can be polled on demand. But testing revealed that this list is erased whenever a handset changes cell, and takes about two seconds to reacquire, so readings taken during a change of cell were always returning empty lists of neighbouring cells.

The solution to both problems was to start constructing a location record the moment the handset changed cell, and at the same time to initiate a GPS fix and poll for neighbouring cell

information. As the GPS and neighbouring cell data became available, the location record would be filled out, and once it was complete or 35 seconds had elapsed it would be saved. But if the handset changed cell while waiting for that data then the location record would be saved as-is, potentially without a GPS location or with an empty neighbouring cell list.

7.3.2 GPS accuracy

When recording a GPS location, the logging software originally recorded the first value returned by the handset's GPS receiver. However, further testing using the Android *GPS Test* utility published by Chartcross Limited (Chartcross Ltd 2012) showed that the receiver returns better results as it acquires more satellites, and waiting a few extra seconds after the initial response can result in more accurate locations.

With this in mind, Cell Logger was modified to ignore all GPS readings with an accuracy worse than 50 metres. When a GPS location was required Cell Logger would continue to poll the receiver until the accuracy was better than 50 metres, the handset changed to a different cell, or it gave up after 35 seconds.

The original software ran for a period of six and a half weeks, between 18 July and 1 September 2010. After that, the modified software was used, with results as shown in table 7.1.

Period	Samples	Average accuracy	GPS samples	% with GPS
18 Jul – 1 Sep 2010	7247	294.3 metres	4730	65.27%
After 1 Sep 2010	8316	27.8 metres	5984	71.96%

Table 7.1: GPS readings recorded by Cell Logger.

Clearly, the modified software produced more accurate GPS readings. The average accuracy of the original software was 294.3 metres, and this was improved ten-fold in the modified software to 27.8 metres.

One risk with discarding initial inaccurate GPS readings was that more records might be saved without a GPS value, since the handset would be more likely to change cells or hit the 35 second limit while waiting for a better reading. However, this does not appear to be the case – the original software obtained GPS readings for 65% of its records, while the modified version got 72%.

Note that this does not necessarily imply that the modified software got a *higher* percentage of GPS readings. During the early days of testing, the Cell Logger application was often left running overnight while indoors, where GPS readings could not be obtained.

7.3.3 Carriers and locations

The handset was initially equipped with a Hutchison “3” SIM card, allowing it to log 3G antennae on the “3” network. The handset could also be switched to GSM-only mode, causing it to roam on the Telstra GSM network and log those antennae instead.

The handset was later equipped with a Vodafone SIM, and logged both 3G and GSM antennae on that network as well. However, the majority of the data was collected from the “3” network, as shown in table 7.2.

Carrier	Samples	GPS samples	Unique cells	Cells with GPS
Hutchison “3”	13067	9108	1016	943
Telstra GSM	1608	1019	173	152
Vodafone 3G	551	435	76	74
Vodafone GSM	318	135	44	38

Table 7.2: Number of samples collected from each carrier.

The handset was then carried around Melbourne with the Cell Logger software running. Over a period of several months more than fifteen thousand readings were taken, including details of over thirty-six thousand neighbouring cells (cells that were visible to the handset but not used as the primary cell). It should be noted that although the handset was taken out to Melbourne's outer suburbs on a few occasions, most readings were taken within three kilometres of RMIT's City Campus at 330 Swanston St, Melbourne.

An effort was also made to log every cell in the Location Area covering the Melbourne CBD on the “3” network, to gain an insight into the size and extent of Location Areas. The LA was found to contain at least 215 cells and cover an area about 5km across, centred on the south-eastern corner of the CBD.

7.4 Matching cells to antennae

In order to test the accuracy of handset positions derived from cell IDs, it was first necessary to match cell IDs to actual antennae. The RRL database (Australian Communications and Media Authority 2009b) provides spatial coordinates, transmitting frequencies, and carrier information for each antenna, but unfortunately no cell ID.

In section 5.6 an algorithm was described for matching the OpenCellID data (OpenCellID 2009) to antennae from the RRL database, but Cell Logger provides a richer set of data than OpenCellID (which only records latitude, longitude, carrier, LAI, and CID), so a more sophisticated matching algorithm could be used. The following assumptions were made.

First, Cell Logger logs data sequentially and indicates whether records are caused by a change of cell or a periodic timer. For example, say it records two sequential cell-change records, cell *A* at GPS coordinates (*x1*, *y1*) and cell *B* at (*x2*, y2*)*. This means that the handset was in cell *A* during the interval between those records. Now, assuming that cells cover areas with a convex boundary, it can be concluded that all points on the line connecting (*x1*,*y1*) and (*x2*,*y2*) must lie within cell A.

Second, Cell Logger records the accuracy of all its GPS readings, in metres. So instead of locating the handset at a particular point, it can instead be located within a probabilistic region. The GPS accuracy value returned by the Android operating system is defined only as "the accuracy of the fix in meters" (Google 2010c), so it is assumed that the true location lies somewhere in a circle around the GPS reading, with the radius of the circle equal to the accuracy value. It is also assumed that all points within the circle have an equal probability of being the true location.

Third, Cell Logger captures neighbouring cell information. These are cells which, although not used by the handset, are still visible to it. When attempting to assign a cell ID to an antenna, this additional information can be used to check that all occurrences of a cell ID fall within the antenna's coverage area. Note that for 3G networks, neighbouring cell information consists of Primary Scrambling Codes rather than cell IDs, and these first need to be converted to cell IDs as described below in section 7.5 below.

Fourth, the neighbouring cell information can be used to determine the locations that are *not* in a cell's coverage area. Because each record captures details of all the cells visible at a particular location, the absence of a cell indicates that the location isn't part of its coverage area.

Finally, Cell Logger captures signal strength information. This is an indication of the strength of an antenna's signal as seen by the handset, either as the current cell or as a neighbouring cell. The Android operating system provides the current cell's signal strength via the callback function **PhoneStateListener.onSignalStrengthChanged(int asu)** (Google 2010a), where the strength is measured in *asu* (defined below). Unfortunately, this function mostly returns one of the values 2, 6, 12, and 25, each corresponding to one of the four “bars” of strength displayed by the handset.

The signal strength of neighbouring cells is obtained by calls to a different function **NeighbouringCellInfo.getRssi()** whose return value is defined (Google 2010b) as

> received signal strength or UNKNOWN_RSSI if unknown. For GSM, it is in "asu" ranging from 0 to 31 (dBm = -113 + 2*asu). 0 means "-113 dBm or less" and 31 means "-51 dBm or greater". For UMTS, it is the Level index of CPICH RSCP defined in TS 25.125

Thus, for a GSM antenna, a recorded signal strength *s* corresponds to a power of -113 + 2*s* dBm. The unit *dBm* means *decibels above 1mW* (Lin & Chlamtac 2001), so for a signal strength *s* the power can also be expressed as

$$\text{Power} = 10^{\frac{-113+2s}{10}} \text{ mW}$$

Given that the power of a radio signal is inversely proportional to the square of its distance from the transmitter, the distance *d* to the transmitter should be proportional to the inverse square root of the signal power. This can be expressed with the following equation -

$$d \propto 10^{\frac{113-2s}{20}}$$

For UMTS (3G) antennae the calculations are slightly different. CPICH (*Common PIlot CHannel*) RSCP (*Received Signal Code Power*) is a signal strength value *s* in the range 0-91 (3GPP 2010), corresponding to a power of

$$\text{Power} = 10^{\frac{-116+s}{10}} \text{ mW}$$

The resulting distance relationship is thus

$$d \propto 10^{\frac{116-s}{20}}$$

Although the actual signal strength seen by a handset will be influenced by other factors, such as terrain, obstacles, and the orientation of the handset, in general it should, in theory, decline

with distance from the antenna. This could potentially be used to indicate whether a particular antenna has a given cell ID by plotting signal strength values for that cell ID against their distance from the antenna.

7.5 Primary Scrambling Codes

A complication was encountered when trying to use the neighbouring cell information from a 3G network. On a GSM network the Android operating system returns the 16-bit LAI and 16-bit CID of all neighbouring cells, which are the same as the 16-bit LAI and 16-bit CID values used for a current GSM cell. But on a 3G network, it was observed that neighbouring cells are identified by a 9-bit Primary Scrambling Code (PSC), which cannot be used to directly derive a 16-bit LAI and 32-bit 3G CID.

Android version 2.3, released in November 2010, added the function call **GsmCellLocation.getPsc()** which returns the PSC of the current cell. In theory this value could be correlated with the LAI and CID values returned by **getLac()** and **getCid()** to provide a definitive conversion between a PSC and an LAI/CID. Unfortunately, this function did not exist at the time of the study, and the HTC Tattoo was only running Android version 1.6, so it was necessary to develop an algorithm to match each PSC to an LAI/CID.

The matching algorithm was based on the following three assumptions -

1. When a cell is logged, the PSCs of its neighbouring cells do *not* correspond to that cell. If they are neighbours, they cannot refer to the current cell.
2. When a change of cells occurs, the PSCs of neighbouring cells with a strong signal strength are good candidates for matching the previous cell. In other words, the cell that the handset just left is likely to show up as a neighbouring cell with a strong signal.
3. A PSC is a 9-bit code, so there are only 512 possible values. Because there are more than 512 cells on each of the 3G networks in the study, the same PSC will be used by more than one cell. When such a conflict arises the physically closest cell with that PSC will be used, up to a limit of 5km. The 5km restriction was added after cells from over 100km away were being matched because they where the closest instance of a PSC.

Based on those assumptions, the PSC matching algorithm was as follows -

1. For each 3G network where data was available ("3" and Vodafone in this case), a two-dimensional matrix was constructed, with a row assigned to each different LAI/CID encountered, and 512 columns for all possible PSC values. The matrix was initially filled with zeroes.
2. Each 3G entry logged by Cell Logger contained the LAI/CID of the current cell and a list of neighbouring PSCs. In the matrix row corresponding to the LAI/CID, the columns corresponding to its neighbouring PSCs were set to -1 to indicate that they cannot correspond to that cell.
3. When an entry logged by Cell Logger was caused by a change of cell, the row of the previous cell's LAI/CID was identified. Within that row, the signal strengths of the current cell's neighbours were added to their PSC column if its value was not -1.
4. After processing all the records, for each LAI/CID the highest value in its row should be in the column of its PSC.

After running this algorithm, PSCs were assigned to 784 of the 1015 observed Hutchison "3" cells and 65 of the 76 Vodafone 3G cells. However, the accuracy of the assignment process could not be verified.

7.6 Cell matching algorithm

The algorithm used for matching a cell ID to an antenna was probabilistic and involved looking at all the locations where the cell ID had been observed and comparing them to the theoretical coverage areas of the antennae. The cell ID was then assigned to the antenna whose theoretical coverage area best matches those locations.

Cell ID observations are recorded with a GPS location and an accuracy value r measured in metres. Because of this measurement inaccuracy, observations do not occur at points but within circles of radius r with a probability density of

$$\frac{1}{\pi r^2} \text{ per square metre}$$

Consider the situation in figure 7.2 below. Each of the pale green circles represents an observation of cell X, either as a primary cell or as a neighbouring cell. The red circles represent observations where cell X was not visible, either as a primary or neighbouring cell, and the dashed lines indicate the theoretical coverage area of an antenna.

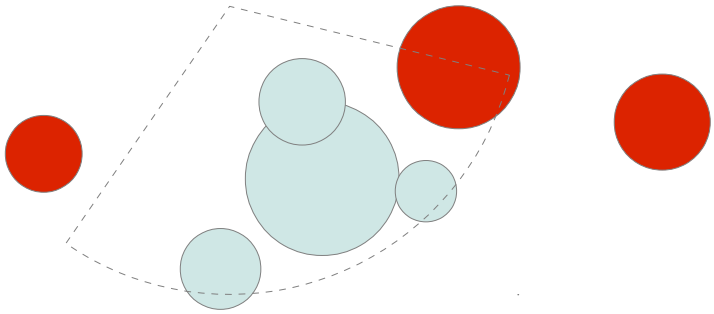

Figure 7.2: Example of cell observations.

Pale green = cell X; Red = no cell X.

The likelihood that cell *X* is covered by an antenna was defined as the probability that the antenna's coverage area contains all observations of *X* and *only* contains observations of *X*. The second requirement, that the antenna only contains observations of *X*, is to help rule out antennae with large theoretical coverage areas that intersect with the coverage areas of other antennae, and will thus contain observations of multiple cells.

This probability is equal to the fraction of all *X* observations that exist within the coverage area, multiplied by the fraction of observations in the coverage area that are of cell *X*, or

$$\frac{\sum_{cell(x)=X} area(x \cap C)}{\sum_{cell(x)=X} area(x)} \cdot \frac{\sum_{cell(x)=X} area(x \cap C)}{\sum_{x} area(x \cap C)}$$

where *C* is the coverage area of the antenna, *cell(x)* is the cell of observation *x*, and the area of an observation *x* is defined as being 1.

For example, in figure 7.2 there are roughly 3.6 *X* observations that overlap with the coverage area, out of a total of 4 *X* observations overall. So the probability that the antenna covers all *X* observations is 0.9. There is also half of a not-*X* observation in the coverage area, so the probability that the antenna only covers only *X* is 3.6 / (3.6 + 0.5), or roughly 0.88.

Multiplying these values together yields a score of 0.79 for cell *X* being assigned to the antenna whose coverage area is shown.

This score was calculated for each observed cell ID and each antenna whose range it was within. These cell ID/antenna pairs were then ranked by score. Going through these pairs in order of decreasing score, the cell ID was assigned to its paired antenna if neither the cell ID nor the antenna had previously been assigned.

Early testing using this algorithm yielded poor results, with many cells being assigned to distant antennae. For example, the assignment algorithm resulted in 37 Location Area changes occurring within the Melbourne CBD on the “3” network, which is unrealistic because the Melbourne CBD is covered by a single “3” LA. Also, the average observed distance from a “3” cell to its antenna in the CBD was 448 metres, which is higher than would be expected given the density of antennae in the area (see section 7.7 below).

To improve the results, the calculation was modified to give greater weighting to observations made close to the antenna, making it more likely that cells would be allocated to nearby antennae. When summing up the cell and non-cell observations in an antenna's coverage area, each observation was first divided by the ratio of the observation's distance from the antenna to the antenna's range, as shown in the following equation -

$$\frac{\sum_{cell(x)=X} \frac{area(x \cap C)}{ratio(x,C)}}{\sum_{cell(x)=X} \frac{area(x)}{ratio(x,C)}} \cdot \frac{\sum_{cell(x)=X} \frac{area(x \cap C)}{ratio(x,C)}}{\sum_{x} \frac{area(x \cap C)}{ratio(x,C)}}$$

where *ratio(x,C)* is the distance between observation *x* and the antenna covering area *C* divided by the range of the antenna.

As a result of this change, the number of “3” Location Area changes within the CBD was reduced to 23, and the average distance between a cell observation and its CBD antenna was reduced 358 metres. This is still worse than would be expected if cell IDs were being allocated correctly and the handset was always using the closest cell, but it is an improvement. More results are shown in tables 7.4 and 7.5 in section 7.9.

7.7 Predicted distance errors

When measuring the spatial accuracy of the cell data it was useful to compare the observed distance errors to those predicted by the actual distribution of mobile phone towers. This

helped to serve as a reality-check, especially because aspects of the cell data were generated heuristically and may have been incorrect.

Consider a region containing an evenly-distributed population and some mobile phone towers, as shown below in figure 7.3. Assuming that a mobile phone uses the cell of the nearest antenna, the predicted average error is the distance to the nearest tower, averaged across all points within the region.

Figure 7.3: A region with four mobile phone towers.

In theory this average distance can be calculated analytically. The technique described in Okabe *et al.* (2000) involves constructing Thiessen polygons around the towers, clipped by the bounding region. For all points within each Thiessen polygon the closest mobile tower will be the one the polygon was constructed around. It is then a case of calculating the average distance to the tower across all points within the polygon, then averaging these distances across all polygons, weighted by the area of each polygon. This technique was used by Okabe & Miki (1984) to show that book stores in Toshima, Japan are located closer to railway stations than would be predicted by a random distribution.

Although this analytical technique produces exact results, the calculation of the average distance to a point within a polygon requires decomposing the polygon into triangles and solving complex integral equations for each. Since none of the available GIS tools supported this calculation, it was decided to estimate the average distance computationally instead.

The technique used was to overlay a square grid of evenly spaced points over the region in question. For each grid point within the region, the distance to the nearest tower was calculated, and the average of these distances was used as the predicted error. The grid spacing was then continually decreased until the average distance value converged on a single value.

This technique was applied to the “3” network's towers within the Melbourne Central

Business District (CBD), a rectangular area bounded by Flinders, Spencer, Spring, and Latrobe streets. With dimensions of roughly 1870 by 950 metres and an area of 1.77 square kilometres, this region contained 33 towers, as shown in figure 7.4.

Figure 7.4: Location of "3" towers in the Melbourne CBD.

As can be seen in table 7.3, the average distance to the nearest “3” tower in the Melbourne CBD converged to about 107.5 metres as the grid spacing was reduced.

Grid spacing	Average distance to tower
200 metres	104.904 metres
100 metres	106.627 metres
50 metres	107.513 metres
20 metres	107.523 metres
10 metres	107.488 metres
5 metres	107.493 metres

Table 7.3: Average distance to the nearest "3" tower in the Melbourne CBD.

A similar calculation was carried out for Melbourne's north-eastern suburbs, where the density of towers is lower than in the CBD. The area used was a 10km by 10km square whose lower

left corner was located on the corner of Hoddle and Victoria Streets, Collingwood. Within this region there were 51 “3” towers, and the average distance to the nearest tower was found to be 677 metres.

7.8 Sources of experimental error

When using the Cell Logger data to measure the distance between an observation and the location of its cell's antenna, a number of sources of error were identified -

- GPS errors. Early GPS readings had an average accuracy of 294 metres, and although all readings since 1 September 2010 reduced this to 50 metres or less, the overall average was still 145 metres. To account for this, all distance measurements in the results are displayed with a +/- error. Also, some separate results will be calculated just using the newer, more accurate, data.
- Obsolete data. The cell IDs and GPS readings were collected in the second half of 2010, but were assigned to antennae taken from the July 2009 RRL database (Australian Communications and Media Authority 2009b). It is possible that antennae were added, moved, or taken away during the intervening year and a half, so that the database was no longer at accurate representation of antenna locations. This could result in cell IDs being assigned to antennae that have moved or no longer exist.
- Primary Scrambling Code (PSC) assignments may be incorrect. 3G cells were matched to a PSC using the algorithm described in section 7.6. However, only 849 of the 1091 3G cells were matched to a PSC, and only 18,496 of the 21,147 PSC observations could be matched to a cell. Any incorrectly-matched PSCs would affect the algorithm that matched cell IDs to antennae (see next point).
- Cell IDs may have been assigned to incorrect antennae. Cell IDs were assigned using the algorithm described in section 7.6, which relies heavily on probabilities, heuristics, and possibly obsolete data. 40 of the 152 observed Telstra cell IDs could not be assigned to an antenna, along with 5 of the 112 Vodafone cell IDs and 11 of the 943 “3” cell IDs. The consequence of cell IDs being assigned to the wrong antenna is the calculation of incorrect positional errors for that cell.
- Sample bias. Although not affecting accuracy, the choice of when to run the Cell Logger application may have influenced the type of observations that were collected. A large proportion of the samples were taken on the daily commute to and from RMIT

University, so the cells along that route are likely to be over-represented in the statistics.

Due to the likelihood that many cell IDs were assigned to an incorrect antenna, the results are unlikely to be numerically accurate. However, they should be sufficient to allow for comparisons between different types of observations (e.g. “cell change” vs “periodic update”) and observations from different regions.

7.9 Results

Over a seven month period between July 2010 and January 2011, 15,563 cell observations were recorded, along with details of 36,410 neighbouring cells. Of these cell observations, 10,714 contained valid GPS coordinates, with an average accuracy of 145 metres. Using improved software, 8316 samples were collected after 1 September 2010, of which 5984 had a valid GPS reading. The average accuracy of these samples was 27.8 metres.

7.9.1 Spatial accuracy of the observations

A summary of the spatial accuracy of the observations is shown in tables 7.4 and 7.5, where table 7.5 is calculated from the more accurate GPS samples collected after 1 September 2010. A few things to note about the tables -

- Only observations from the “3” network with a valid GPS reading are shown.
- The “City” column refers to antennae located within the Melbourne CBD.
- The “Suburban” column refers to antennae located between 3 and 100 kilometres from the centre of the Melbourne CBD (at the corner of Bourke and Elizabeth Streets).
- The “All samples” row refers to all samples from the “3” network.
- The “Periodic” row refers to the samples that were taken periodically every 30 minutes.
- “Cell change” samples were those taken when the handset changed cells.
- “LA change” samples were those taken when the handset changed Location Area.
- “Network signal” samples are those that would actually cause the handset to communicate with the network. These consist of Location Area changes and every second periodic sample. Periodic samples were collected every 30 minutes, but on real

networks the update rate is more likely to be 60 minutes, so only half the periodic samples were used.

- The “Predicted” row contains the estimated distance to the nearest antenna based on the layout of the mobile phone towers, as calculated in section 7.7. Note that this calculation uses a different definition of “suburban” to the other rows - the predicted suburban value here was calculated from a 10km square region north-east of Melbourne, while the other suburban samples refer to anything between 3km and 100km from the Melbourne CBD.
- The “Mean centre” row contains the average distance, not to the cell's assigned antenna, but to the mean location of all the cell's observations (for cells with two or more observations). This technique was covered in more detail in section 6.4 “Experimentally measured cell locations”.

	City		**Suburban**		**All**	
Sample type	**Samples**	**Antenna distance (m)**	**Samples**	**Antenna distance (m)**	**Samples**	**Antenna distance (m)**
All samples	1448	358 +/- 201	4084	1503 +/- 107	9087	1039 +/- 138
Periodic	25	401 +/- 239	115	1286 +/- 158	263	911 +/- 229
Cell change	1370	357 +/- 199	3793	1514 +/- 106	8374	1050 +/- 135
LA change	23	840 +/- 46	328	1634 +/- 67	734	1239 +/- 86
Network signal	37	684 +/- 101	385	1585 +/- 79	863	1193 +/- 104
Predicted		107.5		677		
Mean centre	1025	196 +/- 198	2699	401 +/- 90	8831	324 +/- 139

Table 7.4: Spatial accuracy of the observations on the “3” network.

	City		Suburban		All	
Sample type	**Samples**	**Antenna distance (m)**	**Samples**	**Antenna distance (m)**	**Samples**	**Antenna distance (m)**
All samples	705	341 +/- 34	2857	1505 +/- 27	5494	1109 +/- 28
Periodic	15	455 +/- 32	72	1235 +/- 30	142	966 +/- 31
Cell change	664	337 +/- 34	2652	1519 +/- 27	5083	1118 +/- 28
LA change	3	947 +/- 25	201	1625 +/- 25	379	1275 +/- 25
Network signal	12	582 +/- 33	234	1576 +/- 26	449	1230 +/- 26
Predicted		107.5		677		
Mean centre	468	143 +/- 36	2219	414 +/- 28	5230	332 +/- 28

Table 7.5: Spatial accuracy of the observations after 1 September 2010.

Despite the larger GPS errors in table 7.4 and the smaller sample sizes in table 7.5, the measured antenna distances are broadly in agreement. From these results some tentative conclusions can be drawn -

- Observations taken in the city are more accurate than those in the suburbs.
- Because cell and LA changes occur at cell boundaries, some distance from the antenna, they would be expected to be less accurate than periodic samples, which occur at random locations within the boundaries. While this appears to be the case for the suburban samples (1286m periodic vs 1514m cell change), the opposite is true for the city results (401m periodic vs 357m cell change).
- The average distance between cell observations and their assigned antenna (1039m) is within a factor of four of the average distance to their mean observed location (324m), indicating that the assignment of cell IDs to antennae was reasonably accurate. However, there is no obvious way to quantify that accuracy.
- Using the more accurate data in table 7.5, in the city the average distance from the mean centre (143m) was about 40 percent greater than the predicted distance (107.5m), while in the suburbs it was significantly less (414m vs 677m). The higher than expected city distances may be due to the tall buildings in the CBD, which block and reflect radio signals, resulting in cell coverage areas that are more different from their predicted shape than suburban cells. The suburban mean centre value may also have been affected by sample bias, since many of the readings were taken while

stationary at a single venue, resulting in a low variation from the mean.

Assuming that assigning cells to the wrong antenna increases the average distance between observation of the cell and the location of its antenna, then performing correct cell assignments should reduce, or at least not increase, the average antenna distances in tables 7.4 and 7.5. Thus it can be assumed that the values in the tables serve as an upper bound on the true distances, even if their accuracy is unknown.

It is also likely that the distances to a cell's mean centre, which can be calculated without knowing the location of a cell's antenna, form an approximate lower bound on the true distance error. For the true distance error to be lower, the antenna would have to be located at a point even closer (on average) to the observations than their mean position, which is possible but unlikely.

7.9.2 Correlation between signal strength and distance

As mentioned in section 7.4, signal strength should be related to the inverse square of the distance to the antenna. In other words, there should be a strong positive correlation between an observation's distance from an antenna (in metres) and the inverse square root of the observation's signal strength (in milliwatts).

To test this prediction, a correlation coefficient was calculated for each antenna which had two or more observations containing GPS coordinates. More precisely, the Pearson product-moment correlation coefficient (Wikipedia 2010) between the distance from the antenna and the inverse square root of the signal strength was calculated for 904 antennae.

A Pearson correlation coefficient is a value between -1 and 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and zero indicates no correlation between the two variables. The distribution of the 904 correlation coefficients is shown in figure 7.5.

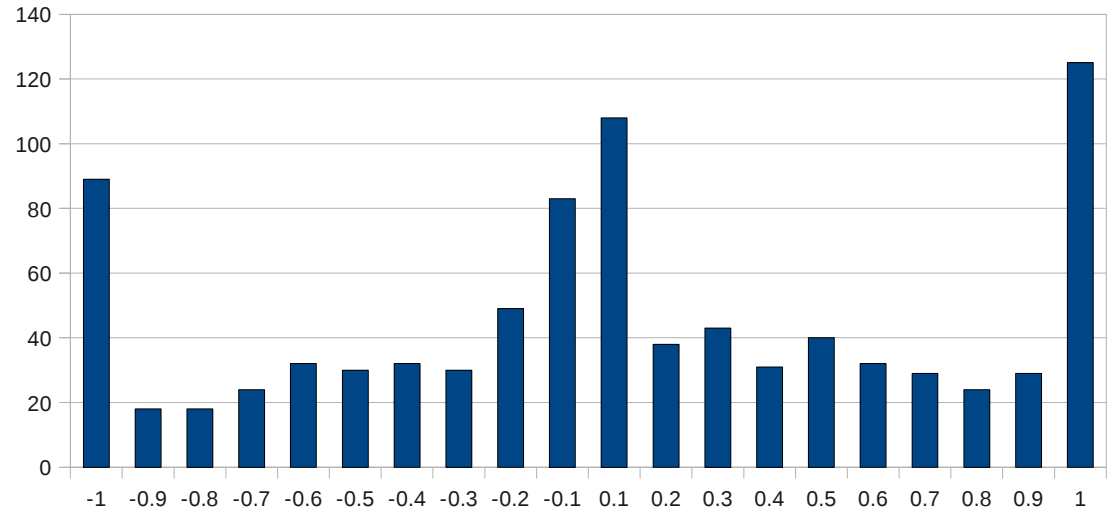

Figure 7.5: Distribution of correlation coefficients between distance and signal strength.

The graph clearly does not show a strong positive correlation for all antennae. In, fact the average correlation coefficient across all antennae was only 0.063, indicating that there is a very weak if any correlation between signal strength and distance in the set of observations.

There are a few possible reasons for this apparent violation of the inverse-square law -

- Cell IDs may not have been assigned to antennae correctly, so the distances were measured from incorrect locations. Similarly, many of the samples used in the calculations came from neighbouring cell observations, which rely on the PSC to CID conversion described in 7.5. This conversion may also have resulted in incorrect antennae being used.
- Signal strengths for current cells are usually quantized by the Android operating system to the ASU values 2, 6, 12, and 25, resulting in inaccurate signal strengths.
- The correlation coefficients of -1 and 1 all occur for antennae which have only two observations, as do many of the zero coefficients, so those results have large margins of error.
- Most observations were taken when the handset changed cell, which typically occurs where signals are weak. Without a corresponding set of strong signals observed at close range it may not be possible to establish a strong correlation. Weak signals might also be caused by terrain and obstacles, which could occur at random distances from the antenna.

7.10 Conclusions

The aim of this chapter was to collect cell data and GPS locations in the real world and use it to validate the assumptions behind the simulation in chapter six. Unfortunately, due to uncertainties in the assignment of cell IDs to antennae it was not possible to verify or refute many of the assumptions conclusively.

But assuming that the average antenna distances obtained in this study form an upper bound on the error of cell-based location estimates, and that the distances from mean cell observations form a lower bound, then using mobile phone signalling data around Melbourne and its suburbs would achieve an accuracy of between 324 and 1039 metres. The upper bound is roughly equal to the median accuracy obtained by the simulation in chapter six (1.2km), although the simulation was carried out across regions with a lower cell density than Melbourne and its inner suburbs.

While the simulation assumed there were clearly defined boundaries between cells, delimited by lines equidistant from the antennae, in the collected data cell boundaries were much more variable, depending on terrain, obstacles, and the orientation of the handset. It was common to see the test handset switch between two or three cells while moving around a small property, simply because of changes in obstacles and orientation, or to change cells when swapping the phone from one hand to another.

As a consequence, the simulation probably underestimated the error in calculating handset locations. If handsets regularly remain in a cell despite being closer to another cell's antenna, then not all cell observations will occur with an antenna's theoretical coverage area. This will increase the average distance between a cell observation and its antenna beyond that measured in the simulation, but the data collected in the study was too inaccurate to say by how much.

An attempt was made to test the correlation between signal strength and distance from the antenna, but there were large sources of experimental error and the results were effectively random. The lack of correlation could imply that cells were assigned to incorrect antennae or that signal strength doesn't correlate with the inverse square root of distance due to obstacles and reflection. Or it could be due to the inability of Android handsets to distinguish between signal strengths, or the fact that most readings took place during cell changes, when signal strengths are at their weakest. Either way, it was not possible to draw any useful conclusions about how signal strength varies with distance.

Due to the large amount of experimental error in the study – especially relating to the

assignment of cell IDs to antennae – the ability to draw many quantitative conclusions about the spatial accuracy of cell data around Melbourne is limited. Further research is needed, but using an up-to-date database of antenna locations with definitive cell IDs assigned to each, and a version of Cell Logger that can take accurate GPS readings and match Primary Scrambling Codes to cell IDs.

The next chapter looks at a case where actual mobile phone billing data was released to the public, allowing it to be analysed in detail.

8 Chapter Eight: A case study using publicly-available billing data

8.1 Introduction

This chapter analyses six months of one individual's Deutsche Telekom mobile phone billing records. Without access to data from Australian mobile phone carriers, this was the most comprehensive set of real-world billing data that could be found.

As well as describing the contents of the records, it suggests a technique for improving the spatial accuracy of the data, and explores the relationship between the frequency of billing records and the accuracy with which a handset can be located.

8.2 Background

As discussed in section 3.4.8, whenever a mobile phone communicates with a network it reveals its approximate location through the ID of the fixed antenna that received its signal. Given the location of the antenna and the direction it points in, its coverage area, or *cell*, can be estimated, providing an indication of where the handset is located. Additionally, when the communication consists of a *billable event* – such as a phone call, an SMS, or internet access – details of the event, including the ID of the cell where the transmission was received, are stored in a billing record.

Storing cell IDs in billing records gives carriers the ability to charge distance-based tariffs, since it is simple to calculate the distance between two antennae. However, the records also allow the movements of a handset to be tracked whenever it is used, accurate to the nearest cell. The accuracy of cells ranges from a few hundred metres in a city centre to a kilometre or two in suburbs, and up to 100km or more in rural areas (Trevisani & Vitaletti 2004).

Due to privacy laws and commercial sensitivity, mobile phone billing records are not generally available for analysis. However, in 2010 Malte Spitz, a German politician from the Green Party, sued Deutsche Telekom to release all billing data relating to his account (Biermann 2011). After redacting some fields to protect the privacy of others, he chose to make this data publicly available on the internet (in a Google Docs spreadsheet linked from Biermann's Zeit Online story (Biermann 2011)).

Although a number of other researchers have had private access to mobile billing data (Ahas

et al. 2007, Calabrese *et al.* 2010, González *et al.* 2008, Yuan & Raubal 2010), this appears to be the first dataset to be made publicly available. It shows exactly what information is available in billing records, something that was not described in the papers that had access to private data, perhaps due to confidentiality agreements.

In this chapter the contents of the records are described in detail, and the information they contain is summarized. A new technique is then applied to improve the spatial accuracy of the data, using estimates of cell centroids rather than antenna locations. Finally, by deliberately discarding records, the effect on accuracy of a reduced sample rate of billing data is investigated.

8.3 Data analysis

The spreadsheet of Spitz's billing data covers a six month period from 31 August 2009 to 27 February 2010, containing 35,830 records, or just under 200 per day.

As an indication of the data provided, the first ten lines of the file are as follows (Biermann 2011) ...

```
Beginn,Ende,Dienst,ein/ausgehend,Laenge,Breite,Richtung,Cell-Id_A,Cell-Id_B,Hint
ergrundinfo: http://www.zeit.de/vorratsdaten
8/31/09 7:57,8/31/09 8:09,GPRS,ausgehend,13.39611111,52.52944444,30,45830,XXXXXX XXXX
8/31/09 8:09,8/31/09 8:09,GPRS,ausgehend,13.38361111,52.53,240,59015,XXXXXXXXXX
8/31/09 8:09,8/31/09 8:15,GPRS,ausgehend,13.37472222,52.53027778,120,1845,XXXXXX XXXX
8/31/09 8:15,8/31/09 8:39,GPRS,ausgehend,13.37472222,52.53027778,120,1845,XXXXXX XXXX
8/31/09 8:20,,,ausgehend,,,,,XXXXXXXXXX
8/31/09 8:20,,SMS,ausgehend,13.38361111,52.53,240,9215,XXXXXXXXXX
8/31/09 8:39,8/31/09 9:09,GPRS,ausgehend,13.37472222,52.53027778,120,1845,XXXXXX XXXX
8/31/09 9:09,8/31/09 9:39,GPRS,ausgehend,13.37472222,52.53027778,120,1845,XXXXXX XXXX
8/31/09 9:12,8/31/09
9:12,Telefonie,ausgehend,13.37472222,52.53027778,120,1845,XXXXXXXXXX
```

The fields are fairly self-explanatory (although a working knowledge of German helps). All the records contain a beginning date and time, and, when the service is not instantaneous (e.g. SMS), an end date and time as well. The third column describes the type of service (e.g. GPRS (General Packet Radio Service) internet access, SMS, or voice telephony), and the fourth indicates whether it was inbound or outbound.

Columns five and six contain the longitude and latitude, respectively, of the antenna covering

the cell used by the phone. If the antenna is directional, column seven contains its direction, as a bearing in degrees clockwise from north, otherwise the value is **null**.

Column eight contains the cell's ID, and column nine contains the redacted cell ID of the other party to the service. Cell IDs range from 1 to 65523, suggesting that they are stored as 16-bit values. Note that the cell IDs are not all unique, since there are 2,830 different cells but only 2,797 different IDs. For example, in the records below cell ID 2334 refers to two different cells because they have different latitudes and longitudes.

```
9/9/09 14:56,9/9/09 15:04,GPRS,ausgehend,9.650833333,50.40638889,0,2334,XXXXXXXX XX

9/18/09 19:36,9/18/09 19:36,Telefonie,ausgehend,13.48472222,52.50194444,30,2334,
XXXXXXXXXX
```

As discussed in section 4.2, in a mobile network cells are grouped into Location Areas, each with its own Location Area Identifier (LAI) (Lin & Chlamtac 2001). It appears that the two records above refer to cells in different Location Areas, but since there are no LAI columns it isn't possible to uniquely identify cells by their ID. However, since cells can be distinguished by their latitude, longitude, and direction, this doesn't affect analysis.

The majority of the billing records are related to GPRS internet access, with SMS records making up another quarter, as shown in figure 8.1.

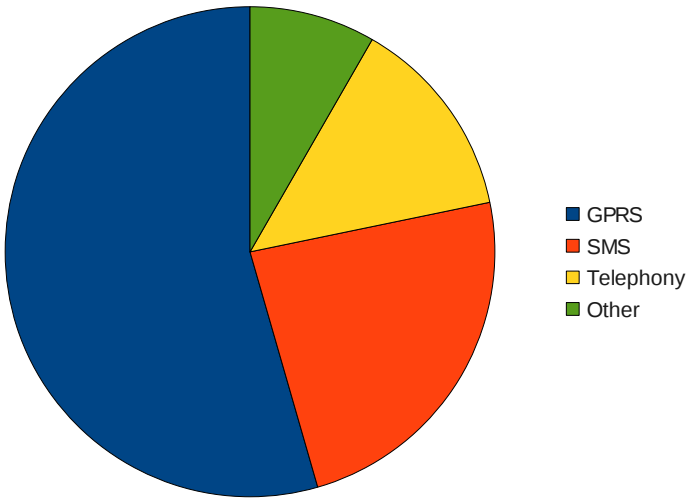

Figure 8.1: Breakdown of service types in the billing data.

Of the 35,830 records, 30,374 (84.8 percent) contain a latitude and longitude, averaging 168 per day. Figure 8.2 shows the numbers of these records, broken down by service type. According to the Zeit Online story (Biermann 2011) many of the SMS and voice call records

that don't provide a location occurred when the handset was overseas, roaming on a different network.

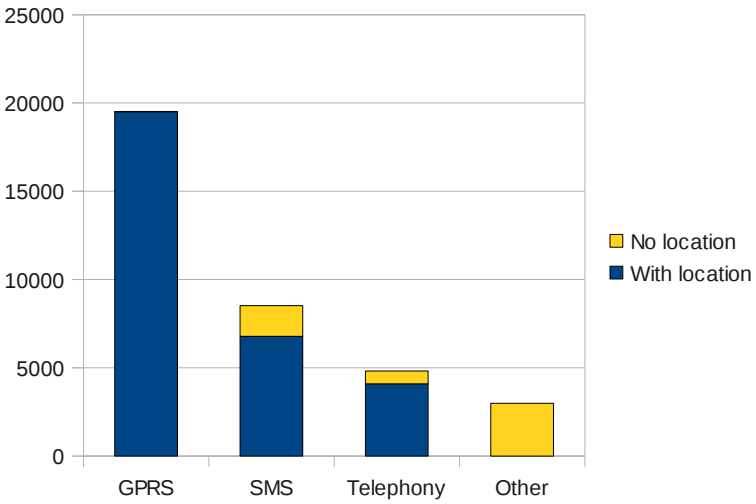

Figure 8.2: Breakdown of records containing location data.

8.4 Location data

The latitude and longitude are expressed to eight decimal places, but all appear to be decimal conversions of degrees-minutes-seconds values. When converted to that form, the seconds component is always within 0.00002 of an integer value, so the locations are presumably accurate only to the nearest second. This implies a north-south accuracy of 31 metres, and, at a latitude of 52ºN, an east-west accuracy of 19 metres.

Since many records provide both a start and end time, it was initially hypothesized that the handset remains in the same cell for the duration. However, there are numerous records where this cannot be the case, for example ...

```
2/27/10 17:30,2/27/10 17:35,GPRS,,13.3725,52.5575,null,47124,XXXXXXXXXX
2/27/10 17:33,,SMS,,13.41861111,52.49916667,null,52924,XXXXXXXXXX
```

The first record indicates that GPRS is being used from 17:30 to 17:35 in cell 47124. However, the second record shows an SMS being sent or received at 17:33 from cell 52924, indicating that the handset was *not* in cell 47124 at 17:33. Since there are 1,782 such conflicts in the data, the hypothesis appears to be incorrect. As a result, it is assumed that the location is only valid at the start time of a record.

8.5 Location accuracy

It should be noted that the locations in the dataset are actually those of the *antennae* covering the cells, which are not necessarily the best values for estimating handset locations. Ideally the values would be the centroid of each cell's coverage area, or, even better, the average position of all handsets ever to communicate with that cell.

As discussed previously in section 4.2, there are two types of antenna, directional and omnidirectional, with coverage areas shown in figure 8.3. Directional antennae tend to cover an arc between 120 and 180 degrees, and are usually mounted three to a tower. omnidirectional antennae, on the other hand, transmit in all directions and are usually alone on their tower.

Cells tend to be laid out in a hexagonal pattern, so the effective coverage area of omnidirectional antennae is usually hexagonal, while for directional antennae it is often diamond-shaped. In the dataset provided, most of the cells had directional antennae – of the 30,374 cells, only 1,111 (3.7 percent) were omnidirectional.

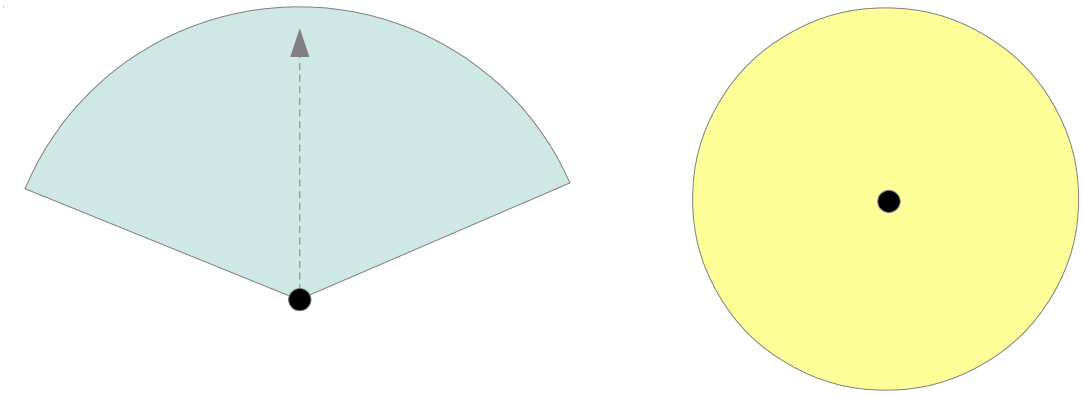

Figure 8.3: Coverage areas of directional and omnidirectional antennae.

8.6 Centroid estimation

When it comes to estimating the centroid of a cell, the location of an omnidirectional antenna is a reasonable approximation. It broadcasts with equal power in all directions, and so long as its neighbouring antennae are arranged roughly symmetrically it is likely to be located near

the centre of its coverage area.

However, as discussed in section 6.5, this is not the case for directional antennae, which are located at a corner of their coverage area. This becomes clear in cases where a handset moves between two cells covered by antennae on the same tower, as shown below.

```
8/31/09 13:35,,SMS,ausgehend,12.67444444,51.86833333,60,19583,XXXXXXXXXX
8/31/09 13:37,8/31/09 13:38,GPRS,ausgehend,12.67444444,51.86833333,290,59260,XXXXXXXXXX
```

Although the handset has clearly moved to a new cell – from cell 19583 with an antenna direction of 60º to cell 59260 with a direction of 290º – the location in both records is the same.

Better results would be obtained if the centroid of a directional antenna's cell could be estimated. But in order to find the centroid of a cell with a directional antenna it is first necessary to determine its cell boundary. In theory this can be estimated if the positions of neighbouring antennae are known, and assuming roughly equal transmitting power and homogeneous terrain. The Deutsche Telekom data provides the locations of 2,830 antennae on 1,642 different towers, so there may be enough information to do those calculations in some regions.

8.7 A simpler method

However, a simpler method, described in section 6.5, is to assume that an antenna transmits symmetrically on either side of the direction it points in, so the cell centroid will lie somewhere along its directional vector. Then it simply becomes a case of estimating how far along that vector the centroid lies.

One approach to calculating that vector distance is to assume it's the same for all cells, and to find the value that minimizes the average positional error when locating handsets. However, unlike the technique used in section 6.5, which used simulated data, the *actual* position of the handset is not known, so a proxy for positional error is needed. In this case the proxy used was overall distance travelled.

If we assume that a handset in the real world generally travels in straight lines, taking the shortest path between destinations, then cell centroid estimates that minimize total distance travelled should be the most accurate, since they generate the smallest amount of "zigzagging".

In figure 8.4, the estimated total distance travelled is plotted against cell centroid distances

from the directional antennae. As can be seen, that total distance is minimized when the centroid is roughly 320 metres in front of the antenna – a total distance of 35,687km, or nearly 200km per day. This leads to the conclusion that, on the Deutsche Telekom network at least, a more accurate estimate of a handset's location is 320 metres in front of a directional antenna, and not the antenna itself.

A similar calculation for Australia, using simulated handset positions based on census data, was carried out in section 6.5. It found that average location error is minimized when a position about 700 metres in front of the antenna is used. The difference in values may be due to Germany's higher population density, which would naturally lead to a higher antenna density and hence smaller cells.

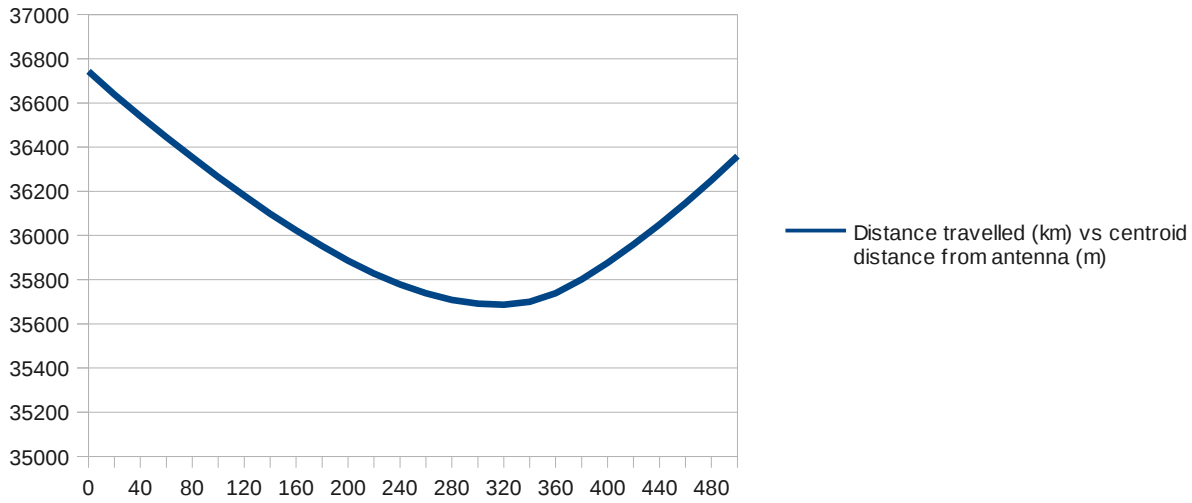

Figure 8.4: Graph of total distance travelled vs cell centroid distance from antenna.

8.8 Sample rate

As mentioned in section 8.3, the data released by Malte Spitz contains just under 200 records per day, of which 168 provide location data. However, previous studies of mobile phone billing data have revealed significantly lower sample rates than this. Due to his job as a full-time politician, Spitz may be an outlier in terms of mobile phone use.

Data from almost a million subscribers in the Boston area over the period 30 July to 12 September 2009 (45 days) consisted of 130 million records (Calabrese *et al.* 2010). This implies a sampling rate of only 2.9 records per subscriber per day, barely one sixtieth as many

as Spitz. However, it should be noted that this data was “aggregated and anonymous” so it is not clear if it contains all the original data.

But assuming that most mobile phone users generate fewer than 168 records per day, it is useful to know how much positional accuracy is lost due to lower sample rates. This can be estimated by comparing the original “high resolution” dataset with artificially-degraded “low resolution” versions taken at lower sample rates.

The high resolution dataset is used to create a “reference” set of locations, using the positions of omnidirectional antennae and 320 metres in front of directional antennae. During the period between samples the handset is assumed to be at the last known location.

A low resolution dataset is created the same way, except that samples are discarded based on the new sample rate. For example, to simulate a rate of eight samples per day, only every 21st record from the original dataset would be used. The accuracy of this dataset can then be evaluated by calculating distance between its estimated handset position and that of the high resolution data, averaged across the six-month sample period.

The results are shown in figure 8.5. Note that the X axis uses the mean time between samples rather than samples per day. The original high resolution data has a mean sample period of 8.58 minutes, while, for comparison, the Boston data from 2009 had a mean period of 500 minutes.

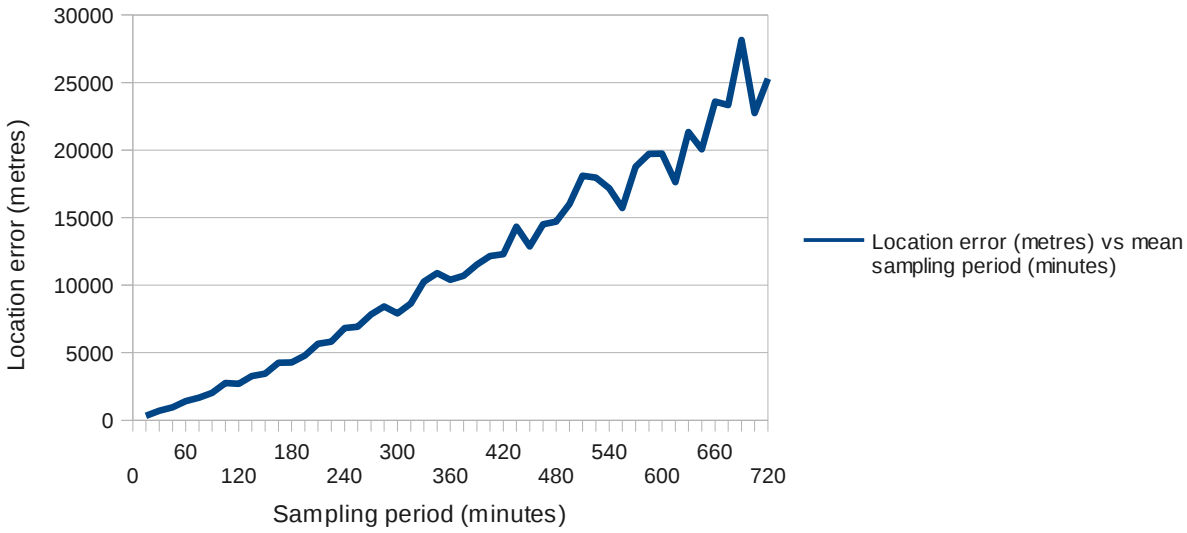

Figure 8.5: Average location error vs mean sampling period.

It should be noted that in the data provided by Malte Spitz the average distance travelled per

day was nearly 200km, which may also be an outlier among the general population. As a result, the average location error should be scaled down in line with the average distance travelled by people in the region of interest.

But regardless of the scale, the data shows a strong linear correlation between the mean sampling period and the average distance error, with a correlation coefficient of 0.9872. In fact the correlation is so strong it should be possible to estimate the average error of the original dataset by comparing it to an “ideal” dataset with a sampling period of zero.

The line of best fit through the points has a gradient of 35.6 metres average location error per minute of mean sampling period. Given an original sampling period of 8.58 minutes, that implies an error of 305 metres due to the sampling rate. Note that this error is in addition to errors caused by the use of cell centroids to estimate the handset position.

8.9 Conclusions

Although it is unwise to generalize based on the billing data of a single individual – an individual whose movements and phone usage may not be typical – the release of Malte Spitz's records has provided valuable insights into what information is collected by carriers, and how it can be used. The data showed that, assuming people take the shortest path between locations, using a point 320 metres in front of a directional antenna is a better estimate of a handset's location than the antenna itself (at least on the Deutsche Telekom network). Using this knowledge, applications making use of billing data to track movements can produce more accurate results.

The data also showed that the frequency of billing records directly affects the accuracy with which a handset can be located at any point in time. Knowing this, applications using billing data can more accurately estimate the errors involved and take appropriate action.

Finally, because the data is publicly available, other researchers can perform their own analysis and generate their own results. This research can then be reproduced or invalidated, something that is not possible with privately-held datasets. Moving forward, it would be desirable to have more billing data to analyse, from a diverse range of individuals, to see if the results apply generally.

The next chapter looks at various applications of location data in detail, and evaluates whether passively-scanned mobile phone data is suitable.

9 Chapter Nine: Applications

9.1 Introduction

The aim of this chapter is to evaluate the suitability of mobile phone data for a number of different applications. The requirements of each application will be analysed in detail and a solution involving mobile phone data will be proposed and evaluated.

As discussed in previous chapters, the spatial accuracy of a cell ID varies with the density of cells in the area, and will typically range from a few hundred metres in urban centres to tens of kilometres in rural areas. If a phone is *actively* queried, its distance from the antenna can usually be calculated as well, with an accuracy of 550 metres on GSM networks and 39 metres on 3G networks.

There is also a lot of variability in the rate at which a handset's cell ID is sampled. When relying on billing data, rates as low as 2.9 records per day (Calabrese *et al.* 2010) and as high as 200 per day (Biermann 2011) have been seen. The use of signalling data, which includes periodic location updates every hour or so, has the potential to increase this by 24 records per day. The use of active querying, on the other hand, can poll a handset's location as often as needed, subject to network capacity constraints.

9.2 Scenarios

The following scenarios were chosen because they serve a useful purpose and can potentially be addressed using passively-scanned mobile phone data -

1. Send alerts to people in the path of a something dangerous, like a bushfire or tsunami.
2. Predict the utilization of a new public transport route.
3. Track the movements of fugitives and missing persons.
4. Measure internal migration within Australia.
5. Identify abnormal population concentrations in real-time so that police and crowd-control resources can be deployed.
6. Measure the population of a region throughout the day/year.

9.3 Bushfire and tsunami alerts

Tsunamis and bushfires are both examples of disasters that cover large areas, but whose movements can usually be predicted in advance to some extent. Earthquakes capable of generating a tsunami can be detected within five minutes (Rudloff *et al.* 2009) and sophisticated modelling software can predict where and when they will strike (Titov *et al.* 2005). Similarly, there are models that can predict the spread of bushfires, although not as accurately (Gould *et al.* 2007).

As a result, it may be possible to give advance warning to people located in the path of the disaster if their position is known and there is some way of contacting them. Such a system should, ideally, meet the following requirements -

1. Everyone currently in the danger zone should receive the alert immediately.
2. If the danger is on-going, everyone who enters the danger zone after the initial alert should also receive an alert, unless they have already received one.
3. Handsets outside the danger zone should not receive an alert. Unnecessary alerts cost money, tie up network capacity, and may cause confusion.

9.3.1 Existing solutions

On 2nd March 2009, on a day of high fire danger, the author received the following SMS from the Victoria Police -

> Extreme weather in Vic expected Mon night & Tues. High wind & fire risk.
> Listen to ABC Local Radio for emergency updates. Do not reply to this msg.

This message was sent to every mobile phone with a Victorian billing address. While this may have been the most targeted form of communication available at the time, it wasn't ideal. Visitors from interstate and overseas, as well as residents who had recently moved and not updated their billing address, would not have received the message. But Victorians travelling interstate and, possibly, overseas, would have.

The message was also not targeted regionally. People living in the inner suburbs of Melbourne are not at risk from bushfires due to a lack of vegetation in the area, so there was no need for them to receive the message. In fact, most of Victoria's population was not at risk, so the Victoria Police would have sent millions of unnecessary messages, possibly at great expense (although it is not clear who paid for them).

There are, however, common examples of regionally-targeted SMS messages. Often when a mobile phone with international roaming enters a new country it receives a welcome message. For example, the author's phone, with a Vodafone Australia account, received the following SMS upon entering Norway -

> Welcome to Norway. To call home dial +61(area code – omit the 0)(phone no.) and for landline dial +61(mobile no. - omit the 0)

This indicates that there is a mechanism for detecting handsets arriving from a different country and which is capable of triggering a customized response. Similar messages have also been received when entering Switzerland, Germany, Qatar, and Thailand.

Another example occurred on the 2nd June 2010, on a train from Oslo to Bergen, in Norway, when the author received the following SMS from phone number 31000 -

> EXERCISE Aurland Municipality practice to alert the population today. You receive this message because you are in the practice area. EXERCISE

This message was received on a handset with a roaming Vodafone Australia account. A second handset with a pre-paid Telenor account did not receive it, and apparently neither did nearby passengers on the train, although the sample was small. Neither handset had been used to make a call or send an SMS that day, so they would not have generated any billing records using nearby cells. It is not known whether the Vodafone handset was roaming on Telenor, or some other Norwegian network.

Although the author does not recall exactly where the message was received, the Bergen Line passes through Aurland Municipality when it stops at the town of Myrdal, and according to the timetable (Norges Statsbaner 2011) it spends about half an hour within the municipality, so the message may have been received then. Despite two subsequent journeys through Myrdal over the following three days, no further messages were received.

The author has been unable to find out more information about the SMS. A Norwegian colleague has never received such a message, despite living in the country for the past decade. And it is worth noting that the SMS was written in English, not the local language, Norwegian.

There are two hypotheses as to how the message was sent -

- During the exercise period, the Mobile Switching Centre (MSC) controlling the Location Area (LA) around Aurland Municipality may have been configured to send

the alert to any international handsets entering the LA, just like a roaming handset welcome message. However, for this to be useful, the LA would have to roughly conform to the boundary of Aurland Municipality, otherwise the message would be sent to a lot of unnecessary handsets. Also, this technique on its own would not result in messages being sent to handsets that were already present in the LA at the start of the exercise.

- At the specified time, the SMS may have been sent to every handset known to be in the cells covering the region, with international handsets receiving an English language version. Handsets would only be known to be in the region if they had made a call or done a periodic update while in the targeted cells, which might explain why only one handset received the message.

Neither hypothesis is ideal. The first one is easier to implement, but the second would be more useful in a real emergency, which makes it the more likely candidate.

9.3.2 Network load during emergencies

One constraint facing any mobile phone-based warning system is network overloading. During an emergency many people may be calling for help, calling friends to see if they're OK, or even taking pictures and uploading them to the internet. For example, in the immediate aftermath of the London bombings on 7 July 2005, all four mobile networks were at capacity and callers were experiencing significant delays (Best 2005).

As a result of this network load, solutions involving the active querying of handsets or the sending of SMSes may not be feasible. Other forms of alert, such as pre-recorded voice calls and video calls to video-capable 3G phones, are almost certainly precluded as they would tie up network resources for the entire duration of the call. However, one approach that shows promise for sending alerts under heavy load is the use of Cell Broadcast SMS, which is described in more detail below.

If the active querying of handsets is not possible, the next best option may be to use the passively-scanned last known position of each handset. Although these positions don't include distance-from-tower information and may be up to two hours out of date – or days out of date for billing data – they don't generate extra traffic on the network and may be the only location information available.

9.3.3 Cell Broadcast SMS

Another method for sending SMS alerts in an emergency is with Cell Broadcast SMS (SMS-CB) (Lin & Chlamtac 2001), described in 3GPP standard TS 44.012 (3GPP 2007). Unlike regular SMS (known as point-to-point SMS, or SMS-PP), SMS-CB is broadcast one-way to all handsets in a cell.

Every cell in a mobile network has broadcast channels, which are used to provide information such as the carrier, the suburb, and the periodic update period. One of these channels can also transmit short messages, which will be received by every handset in the cell.

Cell Broadcast messages are divided into *pages*, each 82 bytes in length. Using the default GSM 7-bit character set, that is sufficient for 93 characters (compared with 140 bytes and 160 characters for SMS-PP). Up to 15 of these pages can be concatenated to form a single message, and a message can be broadcast every 1.833 seconds (Wikipedia 2012). It is common to repeatedly broadcast the same message in case some handsets missed the first one due to a weak signal or interference, and so that handsets entering the cell will receive also it.

However, because SMS-CB is a broadcast, rather than two-way, technology, handsets receiving the message do not send back any response or acknowledgement. As a result there is no way of knowing how many handsets received the message, which might be useful to know when evacuating an area.

On the other hand, because SMS-CB uses the cell's broadcast channel, it doesn't affect the cell's capacity to make calls or deliver point-to-point SMSes. And its ability to deliver a single message to thousands of handsets simultaneously is a very efficient use of the available radio spectrum.

Another potential issue with the use of Cell Broadcast SMS relates to terrorism, in particular SMS-triggered bombs. During the London bombings in 2005 the police in London considered shutting down the mobile phone network in the area to prevent mobile phones from being used to trigger additional bombs (BBC News 2005). If SMS-triggered bombs were in place, depending on how they were configured, then a Cell Broadcast SMS might have accidentally triggered them, causing additional casualties. On the other hand, Cell Broadcast SMS might be a useful tool for detonating mobile phone-activated bombs once an area has been evacuated.

The Daily Telegraph (Osborn 2011) reported that a suicide bomber in Moscow was killed

when her suicide belt was accidentally detonated by a “Happy New Year” SMS from her mobile phone carrier on New Year's Eve 2010. According to the article, many suicide bombers wear SMS-triggered bombs that are remotely detonated by a handler, timed to inflict maximum casualties, and the SMS from the carrier triggered the bomb while she was still at a safe house. If this story is true, then Cell Broadcast SMS may be useful tool for preventing suicide bombings at large events.

9.3.4 Evaluation

If the goal is to simply send an alert to everyone in a danger zone, then Cell Broadcast SMS is the best solution. It meets the original requirements – that everyone currently in the zone receives the SMS; everyone entering the zone afterwards also receives the SMS; and people outside the zone don't receive the SMS (except those in cells overlapping with the zone).

Equally important, SMS-CB works even when a network is under heavy load, which is common during a disaster. Under such conditions a regular SMS may be delayed indefinitely.

One minor drawback with SMS-CB is that it provides no information about the number of handsets receiving the message, or their identity. If information is needed about the handsets in the danger zone, heavy load on the network may preclude the active querying of handsets, leaving passively-scanned signalling data – or even billing data – as the only option. Such data is likely to be hours, or even days, out of date, but may still be accurate for many people depending on their mobility.

9.4 Predicting public transport utilization

Adding a new public transport route can involve significant up-front costs, especially when fixed infrastructure such as tracks and rolling stock are required. One of the key factors determining whether such a route will be cost-effective is the demand forecast – in other words, a prediction of how many passengers will use it.

A powerful tool in transport planning is the origin-destination (OD) survey. First carried out on a large scale in Detroit in the 1940's (Papacostas & Prevedouros 2005), these surveys record demographic information and details of journeys undertaken by participants in the recent past. The local region is typically divided into zones, and participants are asked to record the origin and destination zones of all their journeys, often with supplemental information such as the time of day, duration, method of transport, and cost of the journeys.

The survey data is then compiled into an origin-destination matrix, where the origin zones are listed down the left hand side, destination zones are listed across the top, and the matrix is filled with a count of journeys between those zones, as shown in table 9.1. These matrices can then be used to predict the utilization of new transport routes between particular zones.

	Destination			
Origin	**Zone**	**A**	**B**	**C**
	A	48	72	39
	B	72	33	86
	C	39	86	24

Table 9.1: An example of an origin-destination matrix.

However, conducting surveys and compiling the matrices is an expensive and time-consuming activity. The most useful data comes from household surveys, but these require the participation of, depending on the region's population, 1-20% of the population (Ortuzar & Willumsen 1994) or 2-4% in major cities (Papacostas & Prevedouros 2005). Other problems with household origin-destination surveys were identified by Ortuzar & Willumsen (1994) -

- They tend to record average, rather than actual, travel behaviour.
- Not all movement is recorded.
- Many details, such as travel time, are poorly estimated.

Some attempts have been made to generate OD matrices using mobile phone data; for example, a pilot study using billing data was carried out in Kent, UK (White & Wells 2002). It was able to produce a valid matrix, but since it could only record cell information when a call was made/received or an SMS was sent, the matrix was very sparsely populated compared to a reference matrix that was collated ten years earlier from roadside surveys. The authors concluded that producing a useful matrix would require a much larger sample, either using mobile signalling data, or by combining billing data from multiple days (which would only be accurate if travellers follow similar paths on those days).

Origin-destination information was also generated by Caceres *et al.* (2007) using simulated signalling data. Representing a stretch of road between the Spanish cities of Huelva and Seville, it simulated the signalling data that would be generated by handsets in cars travelling

along this route.

However, there were a number of problems with this simulation. First, it assumed that the origin and destination zones corresponded to Location Areas. Given the small number and large size of Location Areas (LAs) in Australian cities, these zones are unlikely to be useful to transport planners. The size and shapes of LAs also vary from carrier to carrier, and even between GSM and 3G networks for the same carrier, so there would be no common set of zones across a city's population.

Second, the simulation only used change-of-LA signalling events. This is adequate for recording journeys that cross LA boundaries, but not journeys that occur *within* Location Areas.

However, it might be possible to generate OD matrices using signalling data, especially the periodic update signals sent by mobile phones every hour or two. This would, however, have the following limitations -

- The method of transport would be unknown. It might be possible to make an educated guess based on the route taken, but with the spatial accuracy of cell information being a kilometre or so, this is unlikely to be reliable.
- Estimates of journey durations under a few hours are unlikely to be accurate. For example, if periodic updates arrive every two hours, a change of cell ID between samples would indicate that a journey had started, but it could have started any time in the previous two hours. Calculating the end time of the journey would be similarly inaccurate.
- A heuristic would be needed to distinguish between a single journey with a pause (e.g. waiting at a railway station for a connecting bus) and two separate journeys, and it may not be possible to do that reliably.
- Conversely, a multi-stage journey (e.g. drop the kids off at school, then continue to work) might be identified as a single journey, but for planning purposes it should be recorded as two.
- Since the journey origins and destinations would only be accurate to the nearest cell, transport planners would have to use zones that are at least as large as those cells.

Note that it may be possible to do away with the concept of zones altogether, and simply represent a journey as a pair of spatial coordinates (using latitude and longitude). OD matrices

were originally constructed by hand and processed manually, and aggregating journeys into pairs of origin-destination zones helped keep the size of the task manageable. However, since these tasks are now generally carried out by computer, there is no processing advantage to aggregating the journeys. In fact, aggregating journeys in OD matrices forces traffic planners to discard time-of-day and other attributes from the data, possibly resulting in less accurate predictions. Further research is needed, however, to quantify the gain in accuracy, if any, of using this approach.

9.5 Tracking missing persons and fugitives

When a person is missing, for whatever reason, someone trying to find them will want to know their last known and, ideally, current location. Whether this information is available depends on whether the target is carrying a phone and whether they want to be found.

1. **Has phone, wants to be found.** In this case the target may simply be lost or running late. The easiest solution is to call and ask them where they are.
2. **Has phone, doesn't want to be found.** Some fugitives and runaways may fall into this category, where they are carrying a phone but refuse to answer it. With the cooperation of the phone's carrier it should be possible to actively query the handset, revealing its current cell and possibly distance from the tower (Telstra Corporation Limited 2006).
3. **No phone, wants to be found.** This may apply to kidnap victims and people whose phone has been lost, stolen, or whose battery has gone flat. If the phone is separated from the target but still switched on, its current cell and distance from tower can be actively queried, which may provide some clues as to the target's current location.

   If the phone is not switched on, a database of historical passively-scanned location data would provide details of the phone's location up to the point where it stopped working. Signalling data generates more frequent location updates than billing data, and will thus provide a more recent last known location. However, billing data is already available with existing network infrastructure, and the ability to obtain more recent locations may not justify the additional cost of recording signalling data.
4. **No phone, doesn't want to be found.** Fugitives and runaways who are serious about not being found will probably not carry a switched-on mobile phone (at least, not one registered under their own name). A database of passively-scanned records could

> determine their previous locations, just like targets without a phone that want to be found, and again this may provide some clues.

In summary, passively-scanned mobile phone location data is most useful in cases where the target is no longer carrying their phone, since it allows for the tracking of their movements up to the point where they parted with or disabled their phone. Where the target is still carrying their phone and it is switched on, active querying is a more effective technique for determining their location.

9.6 Measuring internal migration within Australia

One of the roles of the Australian Bureau of Statistics (ABS) is to count Australia's population. Every five years they conduct a census that counts everyone in Australia and records their approximate location. During the intervening five years they use births, deaths, and immigration data to maintain an accurate running total of the overall population.

However, during this inter-census period the location information can become out-of-date. For example, if people move interstate, work temporarily at a mine site, or move to a share house while studying at university, that change in location may not be known until the next census. This is important, because the ABS is required to determine the population of each state in order to determine the number of seats allocated to each state in the House of Representatives (Australian Bureau of Statistics 2005).

> On 12 February 2002 the House of Representatives sat for the first time following the election of November 2001. In accordance with the Commonwealth Electoral Act 1918 (Cwlth), in the thirteenth month after that first sitting day (in this case, 19 February 2003), a determination was made as to the number of MPs there should be representing each state and territory. This exercise was based on the latest population statistics for the states and territories provided to the Electoral Commissioner by the Australian Statistician. At this stage, the first of two quotas used in the redistribution of electorates was calculated.

Note that the ABS has to provide the *population* of each state and territory, including foreign nationals, not the number of enrolled voters. And they also have to do this thirteen months after the first sitting day of parliament, which is unlikely to coincide with a census.

The ABS currently tracks interstate migration using Medicare change of address information

(Australian Bureau of Statistics 2009), but acknowledges that younger Medicare card holders are less likely to register a change of address, or do so long after the fact, and that foreign nationals do not use Medicare.

Mobile phone location data may provide another way to track interstate migration, but will have its own sources of error -

1. Distinguishing between migrants and short-term visitors.
2. Converting from a handset count to a population count.

To distinguish between an interstate migrant and a temporary visitor a simple rule of thumb may suffice. For example, after a period of 45 days where a handset only generates location data in a different state, it could be assumed that the handset's owner has migrated to that state. Whatever rule of thumb was chosen, it could be evaluated using the results of the next census.

Note that because location data is only needed over periods of days and weeks, locations extracted from billing data should be sufficient. Over those time scales most phones should make or receive at least one call or send or receive an SMS, so although signalling data would provide more accurate and timely data, the extra cost may not be justified.

At first glance, converting from a handset count to a population count should simply be a matter of dividing the handset count by the mobile phone penetration rate. For example, if 150,000 handsets have moved interstate and the penetration rate is 102%, then that should correspond to 147,058 people. But in practice it may not be that simple because of variations within the overall 102% penetration value.

In July 2008 Australia had a penetration rate of 102%, but only 72% of the population owned a phone (Australian Communications and Media Authority 2008). That means that 28% didn't own a phone, and it can be assumed that a similar proportion owned more than one. That would not be problem if the migrating population was representative of the population as a whole, but it would result in incorrect estimates if migrants were more likely to be working-age (and thus own at least one phone), or less likely to have a family with young children (who don't own a phone).

However, it might be possible to resolve the issue of multiple phone ownership if personal information can be accessed. When opening a mobile phone account in Australia, some form of identification, such as a driver's licence, must be shown, and if that information were

available then phones linked to the same identification could be removed from the count. This would, however, fail in the following cases -

- Where the second phone is a work phone linked to the employer, not the individual.
- Where extra phones are registered in a parent's name but used by their children.
- Where phones are registered using different forms of ID, for example a driver's licence and a passport, which may not be identified as belonging to the same individual.

Given these problems, and well as the additional data processing and privacy concerns of cross-referencing with personal ID, it may be best to simply divide the handset count by the overall penetration rate. This should provide a reasonable approximation that can be improved with future census data.

9.7 Identifying abnormal population concentrations

Large crowds can be dangerous – to bystanders, to property, and to themselves – so where possible authorities try to deploy crowd-control measures to keep them from getting out of hand. Most crowds, such as audiences at large events and participants in registered demonstrations, are organized in advance, giving authorities time to allocate appropriate resources. However, occasionally large crowds will form spontaneously, or are arranged in secret, and authorities need to detect them as quickly as possible in order to respond.

Detecting large crowds using mobile phone data has two problems that need to be addressed – counting the number of people in an area in real-time, and differentiating between “normal” crowds (e.g. around transport hubs during peak hour) and “abnormal” crowds that need a response.

It should be noted that it is not strictly necessary to track handsets in order to measure the number of people in an area. All that is needed is an anonymous count of the number of handsets in the cells at the time. For example, in the Real Time Rome study (Calabrese & Ratti 2006, Reades *et al.* 2007, Rojas *et al.* 2007) cell populations were estimated using *erlangs*, the average number of concurrent calls in each cell.

Clearly the volume of calls in a cell depends on the number of people present, but it also depends on factors such as the number of calls per person per hour and the average call length, which vary significantly between individuals and by time of day. If sudden increases in erlangs are more likely to be caused by an increase in per-capita call volumes than an

increase in population, then the authorities who are monitoring those values will receive numerous false alarms.

Although erlang data is useful, in practice it would be more accurate to use handset tracking data to measure populations, since each handset has a unique identifier than ensures that it is only counted once. Erlang data, on the other hand, is subject to significant variations that are unrelated to the number of handsets present.

When it comes to handset tracking data there are billing records and signalling data. Billing records are generated much less frequently than signalling data, and, given the need for real-time monitoring, are not as suitable for this application.

At a minimum, handsets generate signalling data once for each periodic update interval, somewhere between 30 minutes and two hours depending on the network configuration. If handsets carry out any other activity, such as making or receiving calls, or sending or receiving SMSes, then signalling data is generated more frequently. Whether this is frequent enough to provide sufficient warning to authorities is not known, and may depend on how quickly crowds form.

9.7.1 What is abnormal?

Distinguishing between normally-occurring crowds and abnormal ones is also important so that authorities don't waste resources on false alarms. A deployed system might display colour-coded current population densities on a map, and might normalize them against historical population densities at an equivalent time of day, on an equivalent day of they week, in order to highlight anomalies. However, that raises a few questions.

- What is an “equivalent” day to compare against? Same day of the week? Same time of the year?
- When normalizing the population density, is it done by subtracting the historical value or dividing by it? Subtraction shows the absolute “abnormal” crowd size – which is of interest to the authorities – but division is less likely to generate false positives due to small percentage variations in the daytime population of a large city.

The approach taken by Candia *et al.* (2008) was to accumulate population estimates (based on erlang data) at each cell at times modulo one week. For any particular cell, at any time of the week, the data could be used to calculate a mean and standard deviation of call volumes.

Anything exceeding two standard deviations was assumed to be an *anomalous* event, worthy of further investigation.

In that study erlang data was aggregated hourly, which may not be suitable for real-time applications. Smaller aggregation periods might be possible, but the trade-off would be more variation in call volumes, and hence a higher standard deviation and less sensitivity to anomalies.

The standard deviation approach would also work with signalling data, which could be used to estimate the number of handsets in each cell for each hour of the week. Because the estimates are based on the number of unique handsets recorded in a cell, rather than the number of handsets multiplied by per-capita call rates (which vary throughout the day), the relative standard deviation for a particular cell at a particular time of the week is likely to be significantly lower than for an estimate based on erlangs.

A lower standard deviation for the population density would make anomalous events easier to detect. Rather than displaying colour-coded "normalized" population densities on a map, densities could be colour-coded according to the number of standard deviations they are above or below the mean. Any values of two standard deviations or more above the mean could be assigned bright, obvious colours, that would stand out to anyone watching.

However, this approach may not work if a population varies significantly week by week, resulting in a high standard deviation. For example, a city such as Melbourne has a large number of commuters and numerous special events that cause the population to fluctuate. According to the Australian Bureau of Statistics (Australian Bureau of Statistics 2008), the resident population of postcode 3000 (the Melbourne CBD) in 2008 was around 14,500, but this is dwarfed by the estimated 380,000 workers who commute into the City of Melbourne (CBD and surrounding suburbs) on a typical weekday (City Research 2009).

With a normal variation of over 300,000 people between day and night, spotting an abnormal crowd of 10,000 or so would be nearly impossible. Variations of that order might occur regularly as a result of late-running trains, parades, and changes to university class schedules, and thus not be considered anomalous by the standard deviation measure.

There is also the possibility that crowds consist of people who are in the city anyway, for example office workers who take part in a demonstration during their lunch break. In this case the overall population of the city will remain constant, and unless the demonstration occurs in a clearly different set of mobile phone cells than regular lunchtime activities, it may not be

visible in the mobile phone data.

In summary, detecting abnormal population concentrations with signalling data is possible in areas which normally have small variations in population, though with a time lag of up to an hour. However, it is unlikely to work in places like cities with large numbers of commuters, whose population fluctuates significantly on a daily basis. At best it might provide a hint of a potential problem, which could be followed up with direct observation.

9.8 Measuring the population in a region throughout the day/year

A census is an accurate way to count a population, but, in Australia at least, it only records where people spend one weeknight every five years, which is usually their home address. It records no data about where people spend their weekdays, weekends, or summer holidays. This is useful information to have, since a lot of infrastructure (automatic teller machines, convenience stores, parking) needs to be located where people spend their days, not where they live, while other infrastructure (power, sewage) needs capacity to handle the peak population, not the average.

The requirements of this scenario are similar to those of measuring an abnormal population in section 9.7, in that it needs to know the number of people present in a cell at a point in time. Unlike the previous scenario it doesn't need the information to be available in real time, but it does need accurate absolute numbers, not variations from a mean.

A study carried out in Estonia (Ahas *et al.* 2007) to measure seasonal variations in tourist numbers used billing data from roaming mobile phones (i.e. phones from a foreign carrier). The data was able to show variations in tourist numbers by day-of-week and month-of-year, and had a correlation with independently-collected accommodation data of 0.97.

This indicates that billing data is closely *correlated* to population numbers, but it may not measure *absolute* numbers accurately. An unknown percentage of foreign visitors may not carry mobile phones or may not use them while in the country. However, a solution might be to compare census data with billing data from the day of the census. The census could provide absolute numbers that can be used to calculate a scaling factor for normalizing the numbers from the billing data.

The availability of signalling data would also overcome the problem of counting people who don't use their phones while in the region, but would still fail to count those not carrying one. As a result, the numbers should still be normalized against census data where possible. On the

other hand, signalling data is updated frequently enough to generate hour-by-hour population numbers, which is useful for applications such as counting the number of people who commute into a city every day.

In conclusion, billing data has been shown to accurately measure relative changes in populations at the temporal resolution of a day or more. Normalizing these results against census data may produce accurate absolute numbers. Signalling data has the potential to achieve similar numerical accuracy, but with a temporal resolution of an hour or so.

9.9 Conclusions

The use of mobile phone location data was evaluated for a number of real-world scenarios. The results were varied, as shown in table 9.2.

Scenario	**Results**
Send alerts to people in the path of a bushfire or tsunami	Cell Broadcast SMS: useful. Billing and signalling data: may be useful for counting handsets, but not for alerting due to network overloading.
Predict the utilization of a new public transport route	Billing data: may not be frequent enough to be useful unless movements are regular. Signalling data: useful.
Track the movements of fugitives and missing persons	Varies. Active queries are most useful when the target is still carrying their phone. Billing data is useful for reconstructing past movements, but signalling data would be more up-to-date.
Measure internal migration within Australia	Billing data: useful. Signalling data: useful.
Identify abnormal population concentrations in real-time so that police and crowd-control resources can be deployed	Billing data: too much variation and too out-of-date to be useful. Signalling data: only useful in areas with populations that don't vary much.
Measure the population in a region throughout the day/year	Billing data: useful for temporal resolutions of a day or more. Signalling data: useful, resolution of an hour.

Table 9.2: Summary of the usefulness of mobile phone location data.

Billing data was found to be most useful for applications requiring infrequent historical

information, such as measuring internal migration, reconstructing the movements of missing persons, and determining day-to-day populations in a region. It may also be useful for predicting the utilization of transport routes, depending on the frequency of the data and the day-to-day regularity of the movements being sampled.

Signalling data was found to be at least as useful as billing data, and may be applicable in some scenarios where billing data is not. In particular, it would be unambiguously useful for predicting transport utilization, and could determine regional populations hour-by-hour instead of day-by-day. However, whether this justifies the additional cost of collecting signalling data is unknown.

The next chapter looks at a number of techniques for visualizing data collected from a mobile phone network and describes a new technique for displaying population movements within a region.

10 Chapter Ten: Visualization of location data

10.1 Introduction

The aim of this chapter is to evaluate techniques for visualizing mobile phone location data, with a focus on techniques that show population movements. Some existing methods will be reviewed, along with their limitations when dealing with large, dispersed, populations. A new technique for visualizing population movements will then be described, along with the algorithms needed to generate its images.

10.2 Background

Due to the widespread use of mobile phones, data is becoming available about the location of people as they move around during the day. Whenever a mobile phone communicates with a network it reveals its current cell, and the details of many of these transmissions are recorded. When combined with time information and a unique handset identifier, this data can be used to track population movements.

In terms of population coverage, mobile phones are carried by most people in the developed world. For example, in June 2008 72 percent of the Australian population carried a mobile phone (Australian Communications and Media Authority 2008), with a total penetration rate of 102 percent (due to some people owning more than one). If people under 15 and over 50 are excluded, over 90 percent of the Australian population carried a mobile phone in 2008 and would have been generating location data.

With so much data available, methods are needed to visualize it in ways that are easy to interpret. While there has been much work done on visualizing population densities, for example using colours or 3D effects, little has been done to visualize the *movement* of populations.

10.3 Location data

Communications between a mobile phone and a base station fall into two categories - billable events and signalling events. Billable events are those that occur as a result of a billable activity, such as making/receiving a call, sending/receiving an SMS, or accessing the internet. Details of these events, including the cell ID, are stored as a matter of course by the carrier,

and are later used to calculate the customer's bill.

Signalling events are a superset of billing events and are used for network administration. They include registration when a phone is switched on, updates when it crosses a Location Area (a cluster of a hundred or more cells controlled by a single Mobile Switching Centre), and periodic updates every hour or so. Signalling events are not recorded by default, but companies such as AirSage (AirSage 2010) have installed equipment on many US networks that may be used to collect it.

In practice, billable events on their own provide a rich source of information, and they are, in theory, available from all carriers. Chapter eight described how in 2010 German Green Party politician Malte Spitz filed suit against Deutsche Telekom to release all the records they held relating to his account (Cohen 2011). The resulting data, consisting of 35,830 records over the six month period September 2009 to February 2010, was then made publicly available on the Zeit Online website (Biermann 2011).

Deutsche Telekom was recording around 200 records per day, of which 168 contained spatial information in the form of a cell ID, along with its latitude and longitude, and the bearing of the cell's antenna (if it was directional). Roughly 24 percent of the records were generated by SMS, 13 percent by voice calls, and 54 percent by GPRS internet access. The remaining records were not identified.

As a full-time politician Spitz may have used his mobile phone a lot more often than the average user, since other studies, e.g. Calabrese *et al.* (2010), report an average of 2.9 records per subscriber per day, over 60 times fewer than Spitz. However, mobile phone users are increasingly using their handsets to access the internet, and this will create more records since each access is recorded separately.

10.4 Visualization techniques

Given the enormous volumes of location data generated by mobile networks it is important to select a visualization technique that displays it in a useful form. Previous studies have used billing data to display population density (Yuan & Raubal 2010), sometimes broken down by country of origin (Ahas *et al.* 2007, Ahas *et al.* 2008) or activity (Phithakkitnukoon *et al.* 2010). Changes in population throughout the year have been shown on graphs (Ahas *et al.* 2007, Ahas *et al.* 2008), and animated population density maps have been used to show changes in distribution throughout the day (Airsage 2009a, Airsage 2009b). But no examples

have been found displaying the *movements* of populations based on billing data, showing either route or velocity information.

Section 2.11 described how the Hägerstrand space-time cube (Hägerstrand 1970) displays the movement of an individual with a line drawn in three-dimensional space, with the XY axes representing location and the Z axis representing time. This provides a concise depiction of an individual's movements, but Kwan (2000) finds that "Although the 'aquarium' is a valuable representational device, interpretation of patterns becomes difficult as the number of paths increases with the number of individuals examined". And when the number of paths increases to thousands, or even millions, a space-time cube is clearly not suited to the task.

In the case where individuals are moving between well-defined regions, such as countries, it is common to draw arrows between those locations, as described in section 2.12. This method, known as a *flow map* (Tobler 1987), is easy to interpret, and the use of variable-width arrows can be used to show volumes.

The problem with flow maps is that mobile phone users generally do not move en-mass between a small number of clearly defined regions, so when a population is spread throughout a city's suburbs it is unclear where the arrows would be drawn. There is also the difficulty of representing the many individuals whose journeys are round-trip, starting and ending at home.

When movements occur along well-defined routes, such as roads and railway lines, section 2.11 described how movements along each *segment* of the route can be clustered and displayed using arrows (Andrienko *et al.* 2007, Andrienko & Andrienko 2008), where thicker arrows indicate more users. This is another type of flow map. However, mobile phone data is not always accurate enough to determine a transport route, and in any case the technique as described requires significant manual input, making it impractical for large volumes of data.

Rather than display route information – which usually discards time-of-day and speed information – another approach is to display velocity information at a point in time. This discards journey information, but avoids the problem of having to decide what period of time defines a journey.

Although designed to represent wind velocity rather than population movements, one way to show velocity information is with a weather map, as shown in figure 10.1. This clearly displays location, direction, and speed, where the speed is represented by both the colour and the number lines sprouting from the side of each arrow.

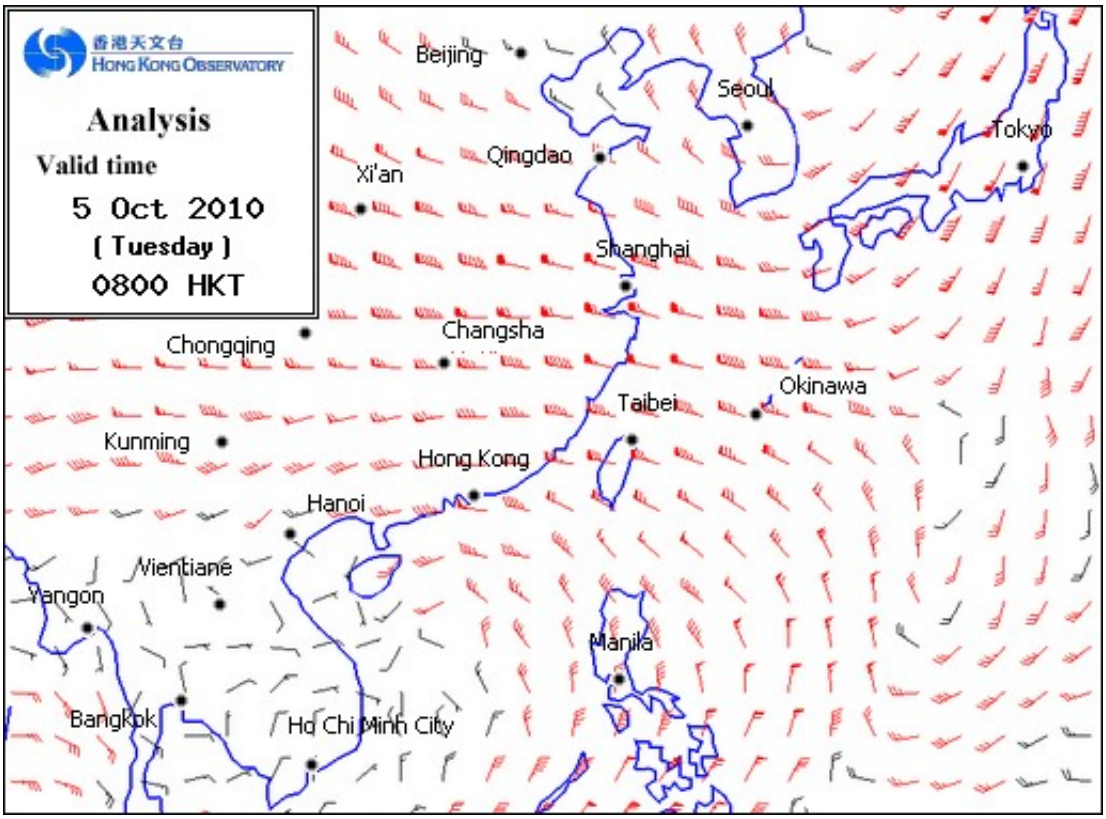

Figure 10.1: Wind speed and directions

http://www.weather.gov.hk/nwp/nwp200wd00e.htm

When used to display the movement of entities or volumes of materials, such a visualization is known as a *continuous* flow map (Tobler 1987). As described in section 2.12, arrows of varying widths or different colours can also be used to represent population volumes.

Unfortunately population movements differ from wind in one key respect – the wind only blows in one direction at each location, whereas people in the same location can be moving in many different directions. For example, at a typical cross-roads intersection traffic would be moving in four different directions – two directions on each of the intersecting roads – so a single arrow would not be sufficient.

One solution is to use a single arrow to display *net* flows at a location, as in figure 2.8 in section 2.12. But although this results in a simpler image, it discards a lot of potentially important information, especially if there are high volume flows moving in opposite directions along the same route.

Another option is to use multiple arrows at each point to display the highest volume flows, perhaps using the representation shown in figure 2.10. In the cross-roads example, four

arrows would be used. However, these additional arrows may add too much complexity to the image when laid out across a map.

But apart from those problems, continuous flow maps are an excellent way to visualize movement, and come close to meeting the requirements for displaying population movements. The new method that will be described here is similar, but incorporates features from other techniques as well.

10.5 Visualizing population movements

An intuitive way to display the instantaneous movement of an individual is with an arrow. The direction of the arrow shows the direction of movement, and the length of the arrow indicates speed. The centre of the arrow - about which it rotates - shows where the individual is located.

But although individual arrows are intuitive, millions of arrows drawn on a map, representing everyone with a mobile phone, would be nearly impossible for a viewer to interpret. Even if the image could be rendered accurately, it would be too complex to understand.

However, a simple way to reduce the complexity of the image would be to define a measure of arrow similarity and repeatedly cluster “similar” arrows into larger ones. The visual representation of a larger arrow is fairly straightforward – still an arrow, but wider than before. Its location would be the mean of the locations of the individual arrows in the cluster, and its velocity would be the mean of the velocities. Finally, its width would be related to the number of arrows merged together, although not necessarily in a linear relationship.

Obviously, when two different arrows are merged together, the resulting arrow is only approximating the actual movements of the handsets, since it just shows the mean, and as more non-identical arrows are merged the approximation becomes greater. Clearly with this technique there is a trade-off between accuracy and visual simplicity, but it is a trade-off that can be explicitly varied depending on how the visualization will be used.

10.6 Handset positions and velocities

Location data from a mobile phone network contains a series of records containing a time t, a handset ID h (usually anonymized), and a position p estimated from the coverage area of a cell.

The accuracy of the position tends to be a few hundred metres in a city, a few kilometres in

suburbs, and tens of kilometres - ranging up to 100km - in rural areas. For the purposes of visualization these errors are usually small in relation to the area being displayed, and are not taken into account.

Given this set of data, it is straightforward to determine the location and velocity of a handset at time t. From the handset's list of records (t_1, p_1), (t_2, p_2), ... the procedure is to find the records immediately prior to and after time t, i.e. (t_{prev}, p_{prev}) and (t_{next}, p_{next}). If there is no record before or after time t, assume the handset's velocity is zero, and its position p_t is the position of the record closest to time t.

Otherwise, the handset's velocity vector v_t is given by $\frac{(p_{next} - p_{prev})}{(t_{next} - t_{prev})}$ and its position is calculated by interpolating between the two points, so $p_t = p_{prev} + (t - t_{prev})\, v_t$

For the purposes of visualization, only handsets with a non-zero velocity are used. Visualizations showing both movement and population density would make use of this stationary handset information, but it is not necessary when just showing movement.

10.7 Defining a "good" cluster

As part of the new visualization technique, handsets with non-zero velocities are formed into clusters based on their position and direction attributes. The s*peed* component of their velocity, however, is not taken into account when clustering, only the direction. There are two reasons for this.

First, speed calculations are very inaccurate when mobile phone data is used, owing to the spatial error and infrequent nature of the location updates. And second, clustering handsets according to speed would create visual clutter. For example, two handsets at the same location, moving in the same direction but at different speeds, might not be clustered because of that speed difference. The result would be two arrows drawn on top of each other, potentially concealing information as well as being confusing. So speed is ignored as a clustering criteria.

In order to form clusters of handsets it is first necessary to define what makes a "good" cluster, so that handsets can be grouped in a way that represents them as accurately as possible. Also, by quantifying the "goodness" of a cluster, it becomes possible to explicitly trade off accuracy against simplicity.

To begin with, each handset h has a position and direction of movement, where its position is

defined as (h_x, h_y). The value h_x is the distance in metres east of an arbitrary point, and h_y is the distance north. In other words, an easting/northing pair. Its direction is expressed in degrees clockwise from north, a value in the range 0-360, denoted by h_d.

Now imagine that this handset information was displayed in three dimensional space, with each handset shown as a point located at (h_x, h_y, h_d). These points are analogous to a galaxy of stars in a space that wraps around in the Z axis. Intuitively, the densely populated regions of the space could be thought of as clusters, and the handsets in those clusters should be grouped together because they share similar positions and directions.

To achieve this, a clustering algorithm first needs a measure to evaluate clusters of handsets so that the most "accurate" ones, containing the most similar handsets, can be identified. The evaluation algorithm also needs to be efficient, preferably only referring to attributes of clusters that can be calculated without referring to the handsets contained within them (which could number in the hundreds of thousands).

10.8 Cluster evaluation algorithm

Since a cluster's position and direction are set to the means of the positions and velocities of its component handsets, it will not represent those handsets exactly unless they all have identical positions and directions. The degree to which the component handsets vary from those means is defined as the *error* of the cluster, or $Err(C)$. This error should be zero when clusters consist of handsets with identical positions and directions, and increase the more those positions and directions vary.

Using this error calculation, the aim of the clustering algorithm is to produce a small enough set of clusters to simplify the visual representation, while minimizing their total error, which is given by the equation $\sum_C Err(C)$

When it comes to measuring positional error of a cluster, the metric chosen was the sum of squares of the distances from the mean, given by

$$PosErr(C) = \sum_{h \in C} (h_x - \bar{h}_x)^2 + (h_y - \bar{h}_y)^2$$

This can also be expressed as

$$PosErr(C) = \sum_{h \in C} h_x^2 + \sum_{h \in C} h_y^2 - \frac{(\sum_{h \in C} h_x)^2 + (\sum_{h \in C} h_y)^2}{|C|}$$

where |*C*| is the number of handsets in the cluster. An error value of zero means that all the handsets in the cluster have the same position, while larger values indicate more variation.

As well as providing a useful measure of positional error, this metric has the advantage of being efficient to calculate. Each cluster only needs to maintain running totals of $\sum_{h \in C} h_x$, $\sum_{h \in C} h_y$, $\sum_{h \in C} h_x^2$, $\sum_{h \in C} h_y^2$, and the number of handsets, all of which can simply be added together when clusters are merged.

The directional error within a cluster is calculated by first summing the dot products of the individual handset unit velocity vectors with the cluster's overall unit velocity vector. Each dot product returns a value between -1 and 1, with 1 indicating the same direction, and -1 the opposite direction. Averaged across all the handsets in a cluster, a value of 1 indicates that all handsets are moving in the same direction, while lower values indicate more divergence. Since this value more accurately describes the *correlation* of the handset directions than their error, it will be termed *DirCorr*(*C*), defined as $DirCorr(C) = \frac{\sum_{h \in C} \hat{h}_v \cdot \hat{C}_v}{|C|}$

This metric can also be calculated using running totals. If *X*(*v*) and *Y*(*v*) are defined to be the x and y components of the vector *v*, then the dot product of two vectors *u* and *v* is given by

$$u \cdot v = X(u)\,X(v) + Y(u)\,Y(v)$$

which means that

$$DirCorr(C) = \frac{\sum_{h \in C} X(\hat{h}_v)\,X(\hat{C}_v) + Y(\hat{h}_v)\,Y(\hat{C}_v)}{|C|}$$

Since the cluster's velocity unit vector is given by $\hat{C}_v = \frac{\sum_{h \in C} h_v}{\left|\sum_{h \in C} h_v\right|}$, this expands to

$$DirCorr(C) = \frac{\sum_{h \in C} X(h_v) \sum_{h \in C} X(\hat{h}_v) + \sum_{h \in C} Y(h_v) \sum_{h \in C} Y(\hat{h}_v)}{|C| \sqrt{\left(\sum_{h \in C} X(h_v)\right)^2 + \left(\sum_{h \in C} Y(h_v)\right)^2}}$$

which can be calculated from running totals of $\sum_{h \in C} X(h_v)$, $\sum_{h \in C} Y(h_v)$, $\sum_{h \in C} X(\hat{h}_v)$,

$\sum_{h \in C} Y(\hat{h_v})$, and the number of handsets, which again can all be added together when clusters are merged.

In theory *DirCorr*(*C*) can be in the range -1 to 1, but in practice negative values are very unlikely. When they do occur, they indicate an unacceptable amount of variation in direction within the cluster, so that particular cluster of handsets will not be used.

The positional and directional metrics are then combined into an overall error value, as follows

$$Err(C) = PosErr(C) + k|C|(\frac{1}{DirCorr(C)} - 1), DirCorr(C) > 0$$

where *k* is a parameter used to vary the relative importance of direction error versus position error when forming clusters.

10.9 Forming clusters

The main difficulty when clustering arrows turns out to be computational. If each arrow has to be compared to all the other arrows to find similar ones, the complexity is usually of the order $O(N^2)$ or greater. In other words, for *N* handsets, the computational effort is proportional to N^2.

Given that a medium-sized city will contain a million or more mobile phones, and a mega-city such as Tokyo or Shanghai could contain ten million, an $O(N^2)$ algorithm is not feasible using readily-available computing resources, since clustering a million handsets would require at least half a trillion comparisons.

On top of this, visualization involves a lot of trial and error as parameters are varied, and waiting days or even weeks to produce an image is unacceptable. Also, because arrows visualize movement at a point in time, it can be useful to create an animation that shows how movement varies throughout the day. Clearly this is impractical if it takes days to render each frame.

In theory the use of an octree could be used to group together similar handsets in $O(N\log N)$. However, the additional overhead of constructing and deleting octree nodes, as well as the difficulty of adapting an octree to handle wrap-around direction values, meant that this approach was not used.

The solution eventually chosen was to pre-cluster the arrows based on whether their location

and direction fell into certain pre-defined ranges. These ranges can be thought of as three-dimensional cubes laid out in a grid pattern over the "galaxy" of handsets. Handsets falling within a particular cube were automatically pre-assigned to the same cluster.

These initial clusters can be created in $O(N)$, and, although merging them afterwards is still an $O(N^2)$ operation, by then N is orders of magnitude smaller than the original number of handsets.

For example, assume the visualization data is going to be overlaid on a map with a 640 x 480 pixel resolution, and a spatial resolution of 32 pixels is acceptable for an arrow. The grid in that case would have 20 cubes along the X axis and 15 along the Y axis. Direction might be divided into 30 degree intervals, resulting in 12 cubes along the Z (direction) axis.

The algorithm for assigning handsets to initial clusters is as follows.

1. Assume the grid is being used to visualize the area from x_{min} to x_{max} along the X axis, and from y_{min} to y_{max} along the Y axis, and that each cube in the grid is w units wide in both the X and Y directions.
2. For each handset h, if $h_x < x_{min}$ or $h_x > x_{max}$ or $h_y < y_{min}$ or $h_y > y_{max}$, then discard the handset.
3. The grid X coordinate of the handset is $floor(\frac{h_x - x_{min}}{w})$ where *floor* rounds the number down to the nearest integer. Similarly, the Y coordinate is $floor(\frac{h_y - y_{min}}{w})$
4. The grid Z coordinate of the handset is $floor(\frac{h_d}{30})$ where h_d is the direction of the handset's movement, in degrees clockwise from north.
5. If there is already a cluster at that (X,Y,Z) position on the grid, add the handset to it. Otherwise create a new cluster at that position containing the single handset.

Using the example above, the end result would be, at most, 20 x 15 x 12 = 3600 initial clusters, calculated in $O(N)$.

These initial clusters can then be merged with the following $O(N^2)$ algorithm -

1. Set the error weighting parameter k in the $Err(C)$ equation to $7.3w^2$. This value was chosen so that, assuming handsets are evenly distributed in all directions and were pre-clustered at 30 degree intervals in the Z axis, the error from merging clusters from

adjacent cubes will be roughly equal in the X, Y, and Z directions.

2. For all clusters C_A find the cluster C_B that would increase the total error by the least amount if they were combined. In other words, for each cluster C_A find the cluster C_B that minimizes $Err(C_A \cup C_B) - Err(C_A) - Err(C_B)$. These are recorded as a C_A/C_B pairs.
3. Go through the list of C_A/C_B pairs and find the one that increases error by the least.
4. Merge them together to form C_{AB}.
5. Remove the two pairs C_A/C_B and C_B/C_x from the list.
6. Find the best cluster C_y to pair with C_{AB} and add C_{AB}/C_y to the list of pairs.
7. For all occurrences of C_x/C_A and C_x/C_B in the list, recalculate the best cluster to pair with C_x.
8. If the total number of clusters is below a target threshold, or if the total error of the clusters exceeds a target threshold, finish. Otherwise go to step 3.

The clusters remaining at the end of this process can then be drawn as an overlay on a map.

10.10 Results

In the absence of real data from a mobile phone network, the simulated data from chapter five was used. Handsets were simulated in the Melbourne area with starting positions based on 2008 Australian census data (Australian Bureau of Statistics 2008). A random 10 percent of the population was used, and handsets were generated assuming a 72 percent ownership rate and according to the market shares of the Australian mobile phone carriers at the time. The simulated movements were simple out-and-back trajectories, in random directions and at random speeds.

These simulated handset records were stored in a file-based database which was accessed via the Python programming language, and the clustering algorithm was implemented as a Python script. The resulting arrows were then rendered using the Python Imaging Library and the *aggdraw* module (PythonWare 2011).

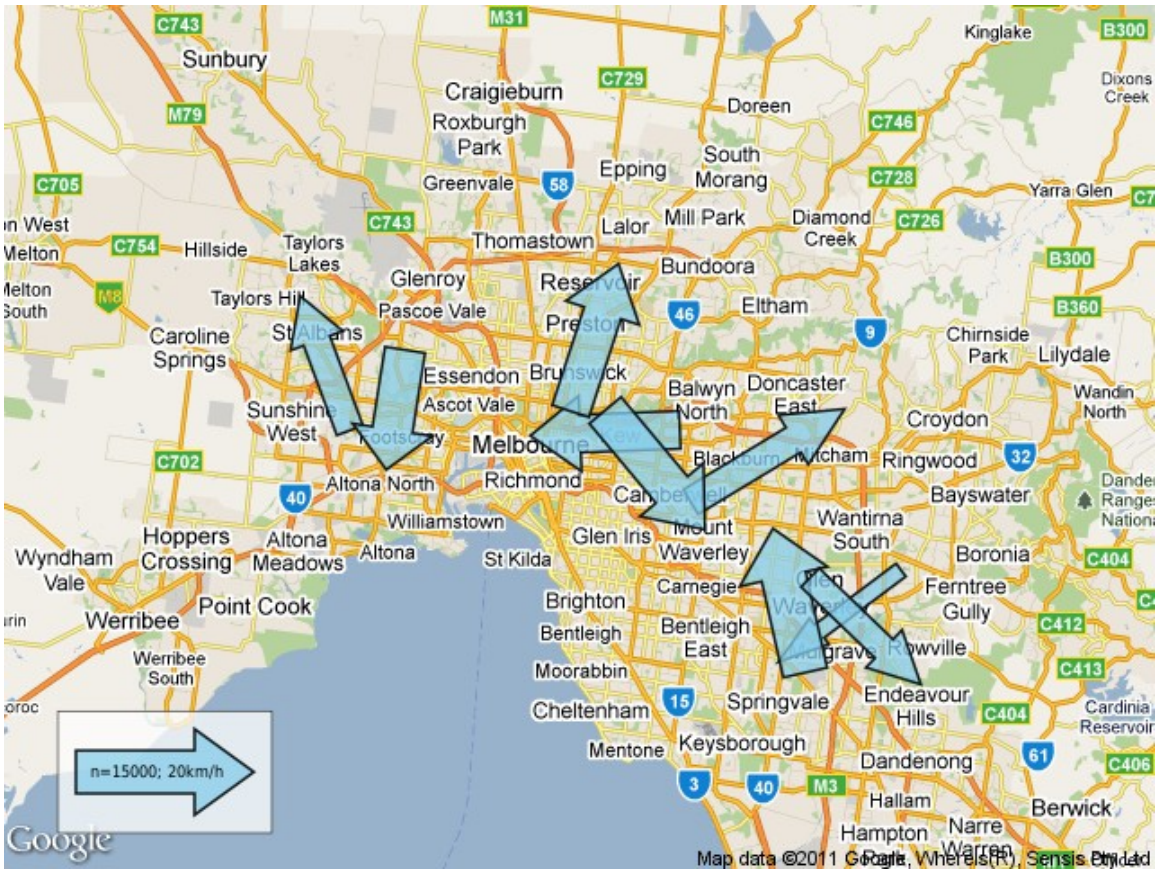

Figure 10.2: Visualization of population movements around Melbourne.

The output of the software is shown in figure 10.2. At the simulated time of 8am, a total of 106,821 non-stationary handset vectors were pre-clustered on a 16 x 12 x 12 grid, resulting in 1724 initial clusters. These were repeatedly merged until a target threshold of 9 clusters was reached (the number 9 was chosen through a process of trial-and-error for its visual appeal). The width of the arrows is linearly proportional to the number of handsets they represent.

The fact that the arrows are pointing in all directions accurately reflects the fact that the underlying simulated handsets were also moving in all directions. And the locations of the arrows largely follow the distribution of the population used in the simulation.

10.11 Animation

The low computational requirements of the algorithm and the fact that it displays a point in time means that it may be suitable for producing animations. Generating an image at one-minute intervals would result in a 60 second video at 24 frames-per-second showing how a population moves throughout the day.

The only uncertainty is whether sequential images would be sufficiently similar to allow for smooth animation. In other words, would two similar sets of handset positions and velocities, separated by only a minute in time, generate similar images with only slightly different arrow positions, or would the clustering algorithm produce entirely different arrows? And to minimize differences between frames, should the clustering algorithm aim for a constant number of clusters in each frame, use a constant maximum error, or apply some other metric?

However, it was decided not to try animation because the simulated movement data was not suitable for testing, and no real-world location data was available. The simulation involved handsets moving in straight-line trajectories in randomly-distributed directions, rather than the clusters of handsets moving along similar routes that might be seen in the real world. Given the way the clustering algorithm works, an animation of these trajectories was unlikely to produce recognizable movement patterns, even if animation could be made to work.

10.12 Conclusions and further research

Before mobile phone billing records became available it was not possible to track the movements of large populations in detail. Small numbers could be tracked using questionnaires or GPS units, larger numbers could be counted along road, rail, and air routes, but for the population as a whole there was only census data, and that only records residential locations every 5 - 10 years.

This historical lack of location data is reflected in the types of visualization techniques that were developed. They focused on displaying the movements of a few individuals in detail, movements along routes, and static information about large populations. However, the increasing availability of population-wide location data means that there is now a need for ways to visualize population-wide movements.

The use of clustered arrows largely achieves this goal because it displays locations, directions, speeds, and volumes in an intuitive manner. It also comes with a parameter that can explicitly trade off accuracy against simplicity, allowing images to be generated with a desired level of complexity. And it is computationally efficient to calculate, even when dealing with millions of individuals.

Although it was not tried, additional population characteristics could also be displayed by varying the colour or opacity of the arrows, perhaps showing age or sex information if that was available. Further, displaying the arrows on a 3D mesh whose “height” represents

population density could show the interaction between population density and movement.

Animation of these images may also be possible, although some research is needed into techniques for minimizing differences between sequential images. This is an area for further research, preferably using of real-world location data.

Further research is also needed to see if the clustering technique is actually suitable for real-world applications. Although the arrows provide an intuitive view of a population, this may not be sufficient for applications such as transport infrastructure planning or evaluating tourism campaigns, for example.

However, the clustered arrows technique is an effective way to display population movements within a region, and is flexible enough to be adapted to specific applications. As such, it should be a valuable addition to the toolkit of anyone working in visualization.

The next chapter evaluates the results obtained in this thesis from the perspective of the research objectives.

11 Chapter Eleven: Evaluation of results

11.1 Introduction

The aim of this chapter is to evaluate the results of the thesis. More specifically, the results will be evaluated from the perspective of the research objectives, defined in section 1.3 as

> ... to develop a method for monitoring and visualizing the spatial behaviour of populations in a large urban environment using passively-collected mobile phone location data.

More specifically, the following questions were to be addressed -

- What location information is available in a mobile phone network, and how can it be extracted?
- How accurate is the data, spatially and temporally?
- Can the data be used to support *real-time* applications such as measuring crowds, finding missing persons, and managing emergency evacuations?
- Can the data be used for *historical* analysis, to investigate movement patterns and, for example, optimize the design of transport infrastructure and the urban environment?
- What is the best way to present this information visually?

This chapter will evaluate the results for each question in turn.

11.2 What location information is available in a mobile phone network?

Chapter four showed that whenever a mobile phone communicates with its network it reveals its approximate location via the ID of the receiving base station antenna, known as a *cell ID*. As part of the communication the handset also includes its unique identifier.

This information can be used to generate records containing a handset identifier, a cell ID, and a time stamp. The cell ID can in turn be used to derive either a latitude/longitude location or an approximate boundary region. In other words, the data flowing through a mobile phone network can be used to determine the approximate location of unique handsets at various points in time.

Handsets communicate with the network when they -

- Are switched on.
- Re-acquire a network after being out of range.
- Make or receive a voice or video call.
- Send or receive an SMS.
- Access the internet.
- Change to a different Location Area (a group of cells usually associated with a single Mobile Switching Centre).
- Haven't communicated with the network for a period of time, usually an hour or so.

This traffic, except for the actual voice/video/data component of calls, is known as *signalling data*. This data is sent to a Visitor Location Register (VLR), a device that maintains a registry of the handsets currently in their Location Area. In theory this traffic could all be captured by reading the data going into and out of every VLR on the network.

It is not clear whether signalling data is actually collected anywhere in the world. AirSage Inc (AirSage 2010) claims to have partnerships with the major network equipment manufacturers and to be collecting real-time data from 127 cities across the United States, so they may have the capability. Also, raw data from the “A Interface” of Shanghai's mobile phone network has been collected since March 2005 (Qiu & Cheng 2007), although it is unclear whether this contains persistent handset identifiers, or just temporary ones. In any case, no examples have been found of an organization publicly admitting to using signalling data, nor of such data ever been produced in court cases or to find missing persons.

Another type of information that *is* collected, however, is billing data. This is the subset of signalling data that is used for customer billing. A billing record is created whenever a handset -

- Makes or receives a voice or video call.
- Sends or receives an SMS.
- Accesses the internet.

As discussed in chapter eight, an example of billing data was published by German member of parliament Malte Spitz (Biermann 2011), who obtained it by filing a lawsuit against

Deutsche Telekom. Among other things, each record contains a date and time, a cell ID, and the latitude, longitude, and bearing of the cell's antenna.

There are many examples of mobile phone carriers providing billing data, whether it was subpoenaed for court cases (Schmitz *et al.* 2000), released for research purposes (Yuan & Raubal 2010), or sold commercially (Withers 2009). Since carriers need to store cell IDs in billing records to facilitate distance-based tariffs, it is reasonable to assume that this data is collected by all carriers as a matter of course. If so, there are no serious technical impediments to accessing the data, only legal and commercial ones.

11.3 How accurate is the data?

As discussed in chapters six and eight, when it comes to mobile phone data accuracy there are two aspects to consider, spatial and spatio-temporal. Spatial accuracy is a measure of how closely a handset's estimated position matches its true position, while spatio-temporal accuracy is a measure of the handset's spatial accuracy throughout the day, which affects the accuracy of location estimates at an arbitrary point in time. Spatio-temporal accuracy depends on the frequency with which data is generated and the speed at which the handset is moving, as well as the spatial accuracy of the data.

11.3.1 Spatial accuracy

Chapter six described a number of techniques for estimating a handset's location from a cell ID, falling into two broad categories – *geometric* techniques and *experimental* techniques – and some which are a combination of the two.

11.3.2 Geometric techniques

Geometric techniques calculate the coverage area of a cell and use the centroid of that area as the location estimate. For simplicity, when cells use an omnidirectional antenna the centroid is generally assumed to be the antenna itself, since it broadcasts in all directions with equal power.

However, the centroid calculation is more complicated when a cell has a directional antenna. The standard approach is to estimate the boundary of the cell, based on the maximum range and the coverage arc of the antenna, along with the positions of neighbouring antennae. In sophisticated models the relative power of the neighbouring antennae is also taken into

account. The coverage area is then calculated assuming that handsets use the cell with the strongest signal. For simplicity it is usually also assumed that the terrain is homogeneous with respect to signal propagation.

Having somehow calculated the shape of the directional antenna's coverage area it is then a simple matter to calculate its centroid and use that as the location estimate for the cell.

Chapter six also estimated cell ID accuracies using simulated handsets whose location was generated from Australian census data. As would be expected in a country with such widely-varying population densities, the distribution of accuracies varied enormously, as shown in figure 6.4.

Using geometric techniques, the simulation produced a mean error of 2,545 metres and a median of 1,196 metres using the antenna locations as the estimate. When the location was estimated using the antenna location for omnidirectional cells and the coverage area centroid for directional cells, the result was a mean error of 2,483 metres and a median of 1,101 metres.

11.3.3 Experimental and hybrid techniques

The second broad method for estimating cell locations is using experimental techniques. These involve taking the actual positions of handsets in a cell, usually determined by GPS, and using their average recorded position as an estimate of the cell's location. This has the benefit of not needing to know the location of the network's antennae (which is often commercially sensitive information) and it accounts for variations in signal propagation due to terrain. On the other hand, it also requires samples from every cell on the network to be useful, and this may not be practical in a large country like Australia.

Using the average position of handsets located within each cell, the simulation in chapter six achieved a mean error of 2,010 metres and a median error of 894 metres. This, however, is a best-case scenario, since it is effectively using every handset in the country to record location samples, whereas in practice the samples would be collected by a small subset of the population.

It is also possible to use a hybrid of geometric and experimental techniques. Assuming that the centroid of a directional antenna's coverage area lies somewhere along its line of sight, experimental data can be used to calculate an approximate distance to the centroid along that line. Using a simulation with Australian data, this distance was found to be 700 metres. When

this value was used to estimate handset locations, the mean error was 2,426 metres, better than either of the geometric techniques on their own.

In section 8.7 a different technique was used to find the distance to the centroid, working on the assumption that people generally take the shortest path between destinations. Using experimental data, it came up with a centroid distance of 320 metres in front of a directional antenna on the Deutsche Telekom network. However, in the absence of GPS data to correlate with the cell IDs, the result could not be verified.

11.3.4 Spatio-temporal accuracy

The spatio-temporal accuracy of location data is a combination of its spatial accuracy and the frequency with which it is produced. It is essentially the error when trying to estimate the position of a handset at a particular point in time, and is of particular interest when trying to determine the position of a handset *right now*.

However, there was a limited amount of real-world data available for measuring spatio-temporal accuracy. Values were calculated from a simulation of Australia's population in section 6.6, but that used simplified movement trajectories, not real ones, which often took the handsets to remote destinations with large cells. The only real data was the Deutsche Telekom billing dataset for Malte Spitz (Biermann 2011), but his movements were probably not representative of the population as a whole.

Still, the available data does allow some trends to be observed. First of all, spatial accuracy depends on where people spend their days. In the simulation, the initial spatial accuracy was calculated from virtual handsets distributed according to where people live. These handsets then moved in straight lines at various speeds before coming to a stop for a few hours.

The average accuracy of the initial positions was 2.48km, but, as shown in figure 6.6, at the end of their straight line movement the average accuracy of the handsets increased to 16.7km. The reason for this big increase in error was that many of the handsets had moved to remote locations where the mobile phone cells are larger. Although this isn't a realistic outcome, it does highlight the fact that where people are located throughout the day directly affects their average spatial error.

For example, people who commute from suburbs to work in cities, where cells are a few hundred metres across, will have a lower average spatial error than people who spend their whole day in the suburbs, where cells are a few kilometres across. However, to quantify this

lower error, data is needed showing how people are distributed during work hours, and this was not available. In theory, if this data *was* available on an hour-by-hour basis during a typical week, it would allow the average spatial accuracy to be estimated based on time-of-day and day-of-week.

The second trend with spatio-temporal accuracy is that it varies almost linearly with the average time interval between location records. Although the simulation carried out in section 6.6 provided little useful data on spatio-temporal accuracy – errors due to sampling rate were drowned out by errors due to handsets moving to remote areas – the Deutsche Telekom billing data allowed a simple test to be carried out to measure the relationship.

Using the original records as a baseline, calculations were performed using smaller and smaller subsets of the data to see how spatial accuracy varied with reductions in the number of records. Because the handset location was estimated using its most recent billing record, fewer records meant locations were more out-of-date.

As shown in figure 8.5 in section 8.3, the spatio-temporal error of the location estimates increased linearly with the mean interval between records, at a rate of 35.6 metres per minute of interval. The correlation between the error and the interval was very strong, with a coefficient of 0.9872.

Although the billing data may not be typical in terms of the actual numbers – the handset was travelling 200km per day, much more than an average handset – the relationship to the sampling interval was not unexpected. Intuitively, so long as the mean sampling interval is greater than the mean time between cell changes, any increases in the sampling interval would be expected to worsen the spatial accuracy of the estimates. And assuming the handset is moving in roughly straight lines during the sampling intervals, the relationship is likely to be linear.

Related to this trend, the average speed of handset movement is also likely to affect spatio-temporal accuracy when positions are estimated using the last known cell. Simply put, the faster a handset is moving, the further it will be, on average, from its last known position. Unfortunately this cannot be verified from the available data, since the only real data relates to a single handset. Testing the hypothesis would require billing data from multiple handsets moving at different average speeds.

In summary, spatio-temporal accuracy depends on the size of the cells where handsets spend their day, the average time interval between location records, and, most likely, the average

speed of handset movement.

11.4 Is the data suitable for real-time applications?

Real-time applications are those that need to know where handsets (a proxy for people) are located at that very moment. The following real-time scenarios were evaluated -

- Send alerts to people in the path of a bushfire or tsunami.
- Track the movements of fugitives and missing persons.
- Identify abnormal population concentrations in real-time so that police and crowd-control resources can be deployed.

11.4.1 Emergency alerts

Section 9.3 described how the aim of emergency alerts is to get a warning message to everyone in a region, typically warning them of an impending disaster such as a bushfire or tsunami. To be successful, such an alert must reach every handset in the region, and do so as quickly as possible.

Alerts can take many forms, but in practice, given the need to notify a lot of people in a short amount of time, SMS messages are the most feasible. Not only do they consume fewer network resources than voice calls, they are also guaranteed to work on all handsets, unlike internet-based solutions such as e-mail.

In a scenario where emergency alerts have to get to everyone in the region, but there is no need to count or identify those people, the best solution was found to be cell broadcast SMS, or SMS-CB. This is a technology that broadcasts the same SMS message to every handset in a cell, and can efficiently reach all the handsets in large regions without tying up network capacity.

The drawback to SMS-CB is that it's strictly a one-way broadcast, with no response from the receiving handsets. This means there is no way to count the number of handsets that received the message, or to know their identity. Knowing the number of handsets would help with managing an evacuation, while knowing the identity of the handsets would allow authorities to keep track of who is in danger and who has been accounted for.

Although a system that relied on passively-scanned location data would be able to identify all the handsets in the region, the low update frequency of the data would miss handsets that had

recently arrived in the region, and incorrectly identify those that had recently left. It would also be slow to identify handsets entering the region after the initial alert. Although this is better than nothing, it does leave some people at risk.

A complete list of handsets in a region could be obtained by combining passively-scanned data with active scanning, where the location of all handsets that *might* be in the region is confirmed by an active query. This list could then be used to manage an evacuation.

Unfortunately, carrying out active queries of every handset would place a heavy load on the network, just at a time when communication is critical. This is in addition to the load caused by actually sending out the SMS alerts (unless SMS-CB is used). With this in mind, it may be better to simply rely on passively-scanned data and SMS-CB alerts, and leave network capacity free so people can call for help and check up on friends and family.

11.4.2 Tracking fugitives and missing persons

Section 9.5 discussed the use of mobile phone data to track fugitives and missing persons. It found that there are four possibilities when trying to track someone, depending on whether they want to be found and whether they are carrying a switched-on handset whose number is known. In the cases where they have a handset and it's switched on, passively-scanned data is somewhat useful for tracking their movements, but active queries are more timely and spatially accurate.

In the cases where they aren't carrying a switched-on handset, passively-collected data can show their movements up to the point where the handset was switched off or discarded. This may provide clues to their current location. However, strictly-speaking, this is an historical, rather than real-time, use of location data.

11.4.3 Identifying abnormal crowd concentrations

Two aspects to identifying abnormal crowd concentrations were identified in section 9.7 – identifying a crowd as it forms, and determining what is abnormal. In general, passively-scanned location data is poor at the former, but good at the latter.

The main problem with passively-scanned data for measuring crowds is its low update frequency, often less than one update per hour. As a result, crowds might only be detected a long time after forming, when all the handsets carried by the participants had finally communicated with the network. By then it may be too late to deploy crowd-control

resources.

On the other hand, the ability to mine historical location data allows for a statistical definition of what is normal and abnormal in terms of crowd sizes. An effective technique is to measure the number of handsets in each cell – or group of cells – at particular times of the day and days of the week using historical records. Over time it is then possible to calculate the mean and standard deviation of the number of handsets in those cells at different times of the week. An “abnormal” crowd is simply one that deviates from the mean by a sufficient number of standard deviations.

Of course, this doesn't address the problem of counting the size of the crowd in real-time, which simply may not be possible with passively-scanned data. There is also the problem of identifying crowds in areas such as city centres that have large populations that vary significantly throughout the day. The standard deviation in those cases may be so large that actual crowds won't be flagged.

11.5 Is the data suitable for historical applications?

Historical applications are those that need to know where people were located in the past. The following scenarios were evaluated -

- Predict the utilization of a new public transport route.
- Measure internal migration within Australia.
- Measure the population in a region throughout the day/year.

11.5.1 Predict transport utilization

Section 9.4 described how one of the most commonly-used transport planning tools is the origin-destination (OD) matrix. These matrices record the number of journeys between different “zones” on a typical day, and can be used to predict the number of travellers that will take a particular route.

A study using billing data to generate OD matrices (White & Wells 2002) found that the records were not frequent enough to produce a matrix as accurately as using roadside surveys. However, if people are consistent in their travel patterns on a day-by-day basis, then potentially OD matrices could be created by combining many days worth of billing data.

It should be noted that the increased use of smart-phones in recent years, many of which

access the internet every few minutes, would produce more frequent billing records. The data released by Malte Spitz (Biermann 2011) contained nearly 200 records per day, and this is more frequent than the rate of signalling data which, according to White & Wells (2002), would be sufficient to generate good matrices.

In general, billing data or, even better, signalling data, would provide useful origin-destination information, but with some limitations -

- The method of transport would not be known.
- It is very difficult to distinguish between multiple back-to-back journeys and longer single journeys.
- Journey times are likely to be inaccurate unless the sample rate is very high.
- The zones will not be exact due to different cell geometries on different networks.

On the other hand, origin-destination information generated from billing data has some advantages that can be exploited using computer processing. Rather than using a small number of manually-defined zones, each cell could be its own zone, providing more finely-grained route predictions. And using the date and time information from the billing data, which is not used by traditional OD matrices, would allow for route predictions based on time-of-day and day-of-week.

In summary, frequently-sampled billing data is suitable for generating origin-destination matrices, which can be used to predict transport utilization. However, the data has the potential to be *better* than OD-matrices at predicting transport utilization if it is exploited directly rather than first being converted to an OD matrix.

11.5.2 Measure internal migration

In section 9.7 it was discussed how during the five years between censuses, the Australian government does not directly count the number of people living in each state, even through that number is needed for allocating seats in parliament. The nation's *total* population can be calculated using the previous census and adjusted for births, deaths, and net immigration, but to calculate *state* populations the rate of internal migration is also needed. To some extent this is estimated using Medicare change-of-address information, but that has limitations.

Because of the time scales involved, billing data was found to be suitable for measuring internal migration. Even handsets that are used less than once per day will generate enough

billing data to indicate a move interstate, although rules will be needed to distinguish between a temporary visit and permanent migration.

The only serious problem identified was how to convert a count of handsets into a count of people. After considering the alternatives, it was decided that simply dividing the handset count by the nationwide mobile penetration rate would yield adequate results. Any errors resulting from this process could be detected at the next census, and future conversions could take those errors into account.

11.5.3 Measure regional changes in population during the day/year

As mentioned in the previous section, the weeknight location of Australia's population is recorded every five years in a nationwide census. However, section 9.8 describes how movements that occur in between census dates are not recorded. As a result, there is no official count of population changes that occur seasonally or within a day. Examples of such population changes are people heading to coastal towns during summer holidays and suburban residents who commute to and from a city every weekday.

Measuring seasonal changes turns out to be very similar to measuring internal migration. The time scales are large enough to use billing data, and the main problem is converting a count of handsets into a count of people. A study in Estonia (Ahas *et al.* 2007) using data from roaming handsets was able to show daily and weekly fluctuations in tourist numbers that correlated very strongly with independent accommodation data, but was not able to provide accurate absolute numbers. To calculate such numbers in Australia, the billing data from the day of a census could be compared to the actual census count to determine a scaling factor for calculating absolute numbers.

Measuring population changes within a day is similar to measuring seasonal changes, but more frequent location data is needed, ideally every hour or two. Although internet-accessing smart-phones may generate billing data at that rate, to cover the large non-smart-phone owning population signalling data may be needed.

11.6 What is the best way to present the data visually?

The raw location data produced by a mobile phone network consists of records containing a handset ID (possibly anonymized), a time stamp, and a cell ID. The cell ID corresponds to an antenna with a known latitude, longitude, and bearing, which can be used to derive the cell's

approximate centroid and error bounds. From the list of times and locations for each handset, approximate velocities can also be calculated by interpolating between the locations.

This enormous volume of information – probably amounting to billions of records per week in Australia alone – cannot realistically be understood as a table of numbers. A visual representation is needed, and given the spatial nature of the data this is best displayed on a map of some kind.

Three broad categories of information can be extracted from the mobile phone data -

1. Where are all the handsets located at a point in time?
2. What are the routes travelled by the handsets?
3. What are the velocities of the handsets at a point in time?

Different techniques can be used for each, as described below.

11.6.1 Displaying handset locations

Displaying the location of handsets at a point in time is essentially the same as displaying the geographical spread of a population, and the same well-established techniques described in section 2.8 can be used. Historically, population density has been represented using choropleth maps, 3D maps, or some combination of the two. These techniques have been found to provide an effective overview of population distributions, although they are not always suitable for detailed analysis.

For displaying changes in handset locations over time multiple techniques are described in sections 2.8 and 2.9. These include time sequences of static maps, animated time sequences, and cartesian graphs.

Time sequences of static maps have the benefit of being available on paper, but this also imposes space constraints that limit the amount of images that can be displayed. It can also be difficult to detect small changes between images when they are laid side-by-side.

Most of these problems can overcome by animating the time sequence, although this limits availability to electronic media. An additional benefit of animating a time sequence is that it can exploit the ability of the human eye to detect patterns in moving images, allowing insights that would not be possible from static time sequence images.

Finally, cartesian graphs were found to be effective when detailed analysis was required on a

small number of variables that change with time, such as the number of handsets within a handful of cells.

11.6.2 Displaying handset routes

Mobile phone location data that contains some kind of handset identifier, anonymized or not, can be used to determine the route taken by a handset, as well as the times at which it passed each point. An effective way to display this information is with the Hägerstrand space-time prism (Hägerstrand 1970) described in section 2.11, a pseudo-three-dimensional line representing the handset's journey, where the X and Y coordinates of the line represent its spatial coordinates, and the Z coordinate represents time.

While space-time prisms are an effective way to display the movements of an individual – especially if accessed through an interface that allows the image to be rotate in three dimensions – that have been found to be much less effective at showing the movements of large numbers of people. As more and more paths are displayed the image becomes too complex to interpret, and displaying the millions of paths from a typical city would result in a solid, impenetrable block of tangled lines.

Another technique for displaying route information involves combining common segments from many individuals, usually along major transport routes, and displaying these with arrows whose width varies with the number of individuals using the segment (Andrienko *et al.* 2007, Andrienko & Andrienko 2008).

Although visually intuitive, this technique has a number of drawbacks when using mobile phone data. For a start, it is usually not possible to identify whether a handset is using a particular transport route, especially when the frequency of its billing data is low. This makes it difficult to construct the visualization in the first place. In addition, the visualization discards important information, such as time-of-day and velocity.

But more important, it discards the overall structure of a handset's journey. By displaying only a collection of route segments it is no longer possible to show beginning and end points, and to estimate the distances travelled.

One way around this is to only group together entire journeys, rather than segments. This is commonly used to display migration routes, as discussed in section 2.12. However, this technique is only effective if there are large numbers of individuals travelling along common paths between well-defined endpoints. When there is a lack of commonality among the paths,

as is likely with population movements within a city, the large number of distinct routes results in a very complex image that is difficult to interpret.

In conclusion, no technique could be found that can reliably display route information for a large population. Space-time prisms are suitable for small numbers of individuals, and route clustering works when movement is along a small number of clearly-defined paths, but neither is suitable for visualizing population movements within a city. As such, this is a field where further research is needed.

11.6.3 Displaying handset velocities

Access to mobile phone location data provides previously-unavailable information about the movements of entire populations, in particular their velocity at a point in time. Although existing techniques, such as the wind map described in section 2.12, can show some velocity information, to take full advantage of the data a new visualization technique was needed.

In chapter ten, a method was described for clustering large numbers of individual velocity vectors into a smaller number of arrows. Each arrow, through its position, direction, length, and width, represents the average location, direction, velocity, and size, respectively, of a cluster of individuals. These arrows provide an intuitive view of a population's movements at a point in time.

In addition, the number of arrows generated can be varied to explicitly trade off visual complexity against representational accuracy, allowing visualizations to be adapted to specific needs. For example, an image containing a half dozen arrows may be sufficient for a brief overview of commuting patterns, but twenty or thirty arrows may be needed for detailed planning of transport infrastructure.

Due to their portrayal of velocities at a point in time, clustered arrows could potentially be used to generate animations. As a day progresses they could show the changes in the speed and direction of population movements as people go about their daily business.

However, it is not clear whether arrow clustering would produce smooth animations. For that to occur, small incremental changes in population movements need to result in small incremental changes in the arrows. It is possible that small changes in movements would result in a completely different set of arrows, leading to jerky, difficult-to-understand animations. This is an area for further research, but it will require access to real-world location data.

11.7 Conclusions

This chapter evaluated the results obtained so far from the perspective of the research objective, which was “to develop a method for monitoring the spatial behaviour of populations in a large urban environment using passively-collected mobile phone location data”.

The results were used to answer the five research questions, namely the availability of data, the accuracy of the data, its use in real-time applications, its use in historical applications, and visualization of the data.

The next chapter provides conclusions and suggests areas for further research.

12 Chapter Twelve: Conclusions and further research

12.1 Introduction

This chapter summarizes the results of the thesis and suggests areas that would benefit from further research.

12.2 Summary of findings

The research objective of this thesis was “to develop a method for monitoring and visualizing the spatial behaviour of populations in a large urban environment using passively-collected mobile phone location data”. The objective was broken down into five research questions, addressing the availability of data, the accuracy of the data, its use in real-time applications, its use in historical applications, and visualization of the data.

In terms of data availability, chapter four showed how mobile phone signalling data can theoretically be collected from a mobile phone network, although it may not be practical to do so due to the need for extra equipment throughout the network. Billing data, on the other hand, is already collected by mobile phone carriers. Chapter eight showed that it is readily available – subject to legal and commercial constraints, but not technical ones.

The spatial accuracy of the mobile phone cell data was found to vary significantly depending on the density of cells. Analysis of the simulated data in chapter six and the smartphone-collected data in chapter seven show that accuracies in Australia were found to be broadly similar to those found elsewhere in the world, with a median accuracy of just over a kilometre, although Australia has a long tail of very inaccurate results due to the very low density of cells in rural areas.

Spatio-temporal accuracy was found to depend on the average speed at which people move and the average time interval between samples. Real-world numbers were calculated in chapter eight from a German individual's billing data, but equivalent numbers for Australia could not be generated due to a lack of data.

Chapter nine described how passively-collected location data was found to have *some* use in real-time applications, but usually there were better ways to do the job, for example by using actively-queried data. The application with the most potential was finding missing persons, but only by making use of historical, rather than real-time, data. Essentially, passively-

collected location data is of limited use for real-time applications due to its infrequent sample rate.

On the other hand, passively-collected data was found to be very useful in historical applications, potentially providing information that is not available through any other means. In particular, it was found to be useful for planning transport infrastructure, and measuring populations short-term (e.g. how many people commute every day to a city), medium term (e.g. how many people spent their holidays in a coastal town), and long term (e.g. how many people have moved to a different state).

In terms of visualization, existing techniques were found to be suitable for displaying static handset (and thus population) densities, but not adequate for displaying movements. For example, no techniques were found that could reliably display route or flow information for large populations in a urban environment.

Although some techniques were found for displaying population velocities, they were not entirely suitable for visualizing mobile phone data. To address this, a new method was described in chapter ten which made use of clustered velocity vectors. It has the ability to show the locations, directions, speeds, and volumes of population movements at a point in time, and can explicitly trade off representational accuracy against visual simplicity.

In summary, the research objective was met. Mobile phone billing data was found to be an effective way, subject to moderate spatial and spatio-temporal errors, to monitor the spatial behaviour of populations in a large urban environment. In terms of visualization, existing techniques were identified that are suitable for some applications, and a new technique was developed to show population movements at a point in time.

12.3 Further research

The field of monitoring population movements with mobile phone data is relatively new and shows enormous potential, providing capabilities that are not possible, or not cost-effective, with other techniques. As well as being a recent area of study, what research has been done has often been hampered by a lack of access to data, be it for privacy or commercial reasons. As a result there are still many unexplored – or under-explored – areas that would benefit from further research.

12.3.1 Data availability

On the data collection front, most of the published research that had access to real-world data made use of billing records. But in theory the use of signalling data would provide more frequent updates. Further research is needed to determine how much more frequent, and at a more basic level, how feasible is it to collect signalling data across an entire network?

Although billing data has been made publicly available for an individual user in Germany, if data was available for *all* the subscribers on a network it would be possible to answer questions such as “how often do they generate a billing record?” and “how far do they travel each day?”. Answers to these questions would provide an estimate of the spatio-temporal accuracy of mobile phone billing data, which is needed when evaluating its suitability for a number of applications.

An attempt was made to validate the spatial accuracy of mobile phone data using some custom software developed by the author that ran on a GPS-equipped Android smart phones. While this produced some useful results, the distance measurements relied on the correct assignment of Location Area Codes, cell IDs, and Primary Scrambling Codes to cell antennae. Since the assignment algorithms relied on heuristics, it is safe to assume that many of the assignments were incorrect. It would therefore be valuable to redo the calculations using the *actual* LAC and CID/PSC of each antenna, which would yield valuable data about the accuracy of cell-based positioning in Melbourne.

12.3.2 Transport planning

A tool used in transport planning is the origin-destination matrix, representing the number of individuals travelling between regions. Although billing data can be used to generate origin-destination matrices, it is an open question whether it can be used directly for predicting route utilization. It may depend on the software and procedures used by transport planners, but the direct use of billing data could potentially generate more accurate results, since it includes time-of-day information and potentially more regions than an origin-destination matrix.

12.3.3 Visualization

In the area of visualization, a new technique was developed involving the use of clustered velocity vectors to show population movements. The technique could be extended to overlay the arrows on a 3D mesh (to show population density) and to use colour-coded arrows (to

show population attributes such as age and sex). It is an open question whether the technique is suitable for generating smooth animations, and whether the information it presents is actually useful for any real-world applications.

Another type of visualization that is worthy of further research is the multi-directional flow map. Continuous flow maps have traditionally displayed the *net* flows of items at a point in space, but the use of multi-directional diagrams could show speeds and volumes in (usually) eight directions, as well as the volume that isn't moving. This could show movement information across a region, and, unlike the clustered arrow approach, also show details of stationary items.

Finally, better ways are needed to automatically detect routes, trajectories, and flows from mobile phone data, and display them in a form that makes sense. The current detection methods require too much manual intervention and do not scale due to computing requirements, while the visualization techniques either discard too much information or are difficult to understand. Without better ways of presenting the information, much of the potential of mobile phone location data may go unfulfilled.

Bibliography

3GPP 2004, *3GPP TS 09.02 Digital Cellular Telecommunications System (Phase 2+), Mobile Application Part (MAP) Specification ETSI TS 100 974 V7.15.0 (2004-03)*, 3GPP.

3GPP 2007, *3GPP TS 44.012 Short Message Service Cell Broadcast (SMSCB) support on the mobile radio interface*, 3GPP.

3GPP 2010, *3GPP TS 25.133 V9.4.0 (2010-06)*, 3GPP.

Australian Bureau of Statistics 2005, *1301.0 - Year Book Australia, 2005 - Drawing House of Representatives electorate boundaries*, Australian Bureau of Statistics.

Australian Bureau of Statistics 2008, *3201.0 - Population by Age and Sex, Australian States and Territories, Jun 2008*, Australian Bureau of Statistics.

Australian Bureau of Statistics 2009, *3228.0.55.001 - Population Estimates: Concepts, Sources and Methods, 2009*, Australian Bureau of Statistics.

Australian Communications and Media Authority 2008, *ACMA Communications Report 2007-2008*, Australian Communications and Media Authority.

Australian Communications and Media Authority 2009a, *Convergence and Communications Report 1: Australian household consumers' take-up and use of voice communications services*, Australian Communications and Media Authority.

Australian Communications and Media Authority 2009b, *Record of Radiocommunications Licences (RRL), 1 July 2009 CD-ROM*, Australian Communications and Media Authority.

Ahas R, Aasa A, Mark Ü et al. 2007, 'Seasonal tourism spaces in Estonia: Case study with mobile positioning data', *Tourism Management*, vol. 28, no. 3, pp. 898-910.

Ahas R, Saluveer E, Tiru M et al. 2008, 'Mobile Positioning Based Tourism Monitoring System: Positium Barometer', *Information and Communication Technologies in Tourism 2008: Proceedings of the International Conference in Innsbruck, Austria, 2008*.

Airsage 2009a, *Obama Inauguration: The Power of Location Data*, viewed 27 April 2009, http://www.youtube.com/watch?v=n2YKlxmvLGs.

Airsage 2009b, *Wake Up San Diego*, viewed 1 June 2009, http://www.youtube.com/watch?v=NBrnY7A6EsE.

AirSage 2010, *AIRSAGE: Live, Real-Time Traffic Information*, viewed November 2010, http://www.airsage.com.

Andrienko G, Andrienko N & Wrobel S 2007, 'Visual analytics tools for analysis of movement data', *ACM SIGKDD Explorations*, vol. 9, no. 2, pp. 38-46.

Andrienko G & Andrienko N 2008, 'Spatio-temporal Aggregation for Visual Analysis of Movements', *IEEE Symposium on Visual Analytics Science and Technology, 2008*.

Andrienko G, Andrienko N, Kopanakis I et al. 2008, 'Visual Analytics Methods for Movement Data', *Mobility, Data Mining and Privacy*, Giannotti F & Pedreschi D (Eds.), pp. 375-410.

BBC News 2005, *Phone network shutdown over bombs*, viewed 20 January 2011, http://news.bbc.co.uk/2/hi/uk_news/england/london/4490372.stm.

Bengtsson L, Lu X, Garfield R et al. 2010, *Internal Population Displacement in Haiti - Preliminary analyses of movement patterns of Digicell mobile phones: 1 January to 11 March 2011*, Karolinska Institutet and Columbia University.

Biermann K 2011, *Betrayed by our own data*, viewed 26 March 2011, http://www.zeit.de/digital/datenschutz/2011-03/data-protection-malte-spitz.

Caceres N, Wideberg JP & Benitez FG 2007, 'Deriving origin destination data from a mobile phone network', *IET Intelligent Transport Systems*, vol. 1, no. 1, pp. 15-26.

Calabrese F & Ratti C 2006, 'Real Time Rome', *Networks and Communication Studies - Official Journal of the IGU's Geography of Information Society Commission*, vol. 20, no. 3&4, pp. 247-258.
Calabrese F, Pereira FC, Di Lorenzo G et al. 2010, 'The geography of taste: analyzing cell-phone mobility and social events', *International Conference on Pervasive Computing.*.
Candia J, Gonzalez MC, Wang P et al. 2008, 'Uncovering individual and collective human dynamics from mobile phone records', *Journal of Physics A: Mathematical and Theoretical*, vol. 41, no. 22, pp. 1-11.
Castilloa E, Menéndez JM & Jiménez P 2008, 'Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations', *Transportation Research Part B*, vol. 42, no. 5, pp. 455-481.
Chartcross Ltd 2012, *GPS Test*, viewed 8 June 2012, https://play.google.com/store/apps/details?id=com.chartcross.gpstest&hl=en.
Chen M, Sohn T, Chmelev D et al. 2006, 'Practical Metropolitan-scale Positioning for GSM Phones', *Proceedings of Ubicomp 2006*.
Best J 2005, *U.K. mobile service strained after explosions*, viewed 20 January 2011, http://news.cnet.com/U.K.-mobile-service-strained-after-explosions/2100-1039_3-5777715.html.
Cohen N 2011, *It's Tracking Your Every Move and You May Not Even Know*, viewed 26 March 2011, http://www.nytimes.com/2011/03/26/business/media/26privacy.html.
Dialogic 2010, *Adding Location-Based Services to Existing Architectures*, viewed 10 August 2010, http://www.dialogic.com/products/signalingip_ss7components/docs/9862_Add_Locationbased_Servs_an.pdf.
Dufkova K, Ficek M, Kencl L et al. 2008, 'Active GSM cell-id tracking: "Where Did You Disappear?"', *Proceedings of the first ACM international workshop on mobile entity localization and tracking in GPS-less environments, San Francisco, 19 September 2008*.
Gelernter D 1992, *Mirror Worlds: or the Day Software Puts the Universe in a Shoebox...How It Will Happen and What It Will Mean*, Oxford University Press.
GeoPKDD 2010, *Geographic Privacy-aware Knowledge Discovery and Delivery*, viewed 2 September 2011, http://www.geopkdd.eu.
Gianotti F & Pedreschi D 2008, *Mobility, Data Mining and Privacy*, Springer.
González MC, Hidalgo CA & Barabási A 2008, 'Understanding individual human mobility patterns', *Nature*, vol. 453, no. 7196, pp. 779-782.
Google 2010a, *PhoneStateListener*, viewed viewed 7 July 2010, http://developer.android.com/reference/android/telephony/PhoneStateListener.html#onSignalStrengthChanged(int).
Google 2010b, *NeighbouringCellInfo*, viewed 7 July 2010, http://developer.android.com/reference/android/telephony/NeighboringCellInfo.html#getRssi().
Google 2010c, *Location*, viewed 15 July 2010, http://developer.android.com/reference/android/location/Location.html#getAccuracy().
Gould J, McCaw L & Cheney P 2007, *Project Vesta - Fire in dry eucalypt forest*, CSIRO and Department of Environment and Conservation, WA.
Greene-Roesel R, Diogenes MC, Ragland DR et al. 2008, *Effectiveness of a Commercially Available Automated Pedestrian Counting Device in Urban Environments: Comparison with Manual Counts*, Traffic Safety Center, University of California, Berkeley.
Gregory I & Ell P 2007, *Historical GIS*, Cambridge Univeristy Press.
Hägerstrand T 1970, 'What about people in regional science?', *Papers in Regional Science*, vol. 24, no. 1, pp. 6-21.

Harris RL 1999, *Information Graphics*, Oxford University Press.
Harrower M & Fabrikant S 2008, 'The Role of Map Animation for Geographic Visualization', *Geographic Visualization*, Dodge M, McDerby Mary & Turner M (Eds.), pp. 49-65.
Hill MR 1984, 'Stalking the Urban Pedestrian: A Comparison of Questionnaire and Tracking Methodologies for Behavioral Mapping in Large-Scale Environments', *Environment and Behavior*, vol. 16, no. 5, pp. 539-550.
Holmes D 2000, 'The electronic superhighway: Melbourne's CityLink Project', *Urban Policy and Research*, vol. 18, no. 1, pp. 65-76.
Johnson C 2008, *Radio Access Networks for UMTS*, Wiley & Sons.
Klein LA 2001, *Sensor technologies and data requirements for ITS*, Artech House.
Kraak M 2008, 'Geovisualization and Time - New Opportunities for the Space-Time Cube', *Geographic Visualization*, Dodge M, McDerby M & Turner M (Eds.), pp. 293-306.
Kucharson MK 2006, 'GPS monitoring: A viable alternative to the incarceration of nonviolent criminals in the state of Ohio', *Cleveland State Law Review*, vol. 54, no. 4, pp. 637-670.
Kwan M 2000, 'Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set', *Transportation Research Part C: Emerging Technologies*, vol. 8, no. 1-6, pp. 185-203.
Kwan M 2003, *Space-Time Paths*, viewed 6 July 2011, http://www.geography.osu.edu/faculty/mkwan/Gallery/STPaths.htm.
Langran G 1992, *Time in Geographic Information Systems*, Taylor & Francis.
Lin Y & Chlamtac I 2001, *Wireless and Mobile Network Architectures*, Wiley & Sons.
McNamara J 2008, *GPS For Dummies, 2nd Edition*, Wiley.
City Research 2009, *Melbourne City User Estimates and Forecasts, 2004-2020*, City of Melbourne.
Melville S & Ruohonen J 2004, 'The development of a remote-download system for visitor counting', *Policies, Methods and Tools for Visitor Management. Proceedings of the Second Conference on Monitoring and Management of Visitor Flows in Recreational and Protected Areas. Rovaniemi, Finland, 2004*.
Miluzzo E, Oakley JMH, Lu H et al. 2008, 'Evaluating the iPhone as a Mobile Platform for People-Centric Sensing Applications', *Proceedings of the International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems (UrbanSense 2008)* .
Monheim R 1998, 'Methodological aspects of surveying the volume, structure, activities and perceptions of city centre visitors ', *GeoJournal*, vol. 45, no. 4, pp. 273-287.
Muehrcke PC, Muehrcke JO & Kimerling AJ 1998, *Map Use*, JP Publications.
Naor Z & Levy H 1998, 'Minimizing the wireless cost of tracking mobile users: An adaptive threshold scheme', *IEEE INFOCOM*, vol. 2, no. 1, pp. 720-727.
NAVTEQ 2008, *Groundbreaking Debut of Traffic Probe Data at ITS World Congress*, viewed 17 November 2008, http://corporate.navteq.com/webapps/NewsUserServlet?action=NewsDetail&newsId=680&lang=en.
NAVTEQ 2010, *NAVTEQ Privacy Policy*, viewed 11 November 2010, http://corporate.navteq.com/privacy.html.
Neumann A 2005, 'Thematic Navigation in Space and Time', *Proceedings of the 4th Annual Conference on Scalable Vector Graphics, The Netherlands 2005*.
Norges Statsbaner 2011, *Oslo S - Bergen timetable*, viewed 10 January 2011, http://www.nsb.no/getfile.php/www.nsb.no/nsb.no/Bilder/Rutetabeller/PDF-41-Oslo-Bergen-2011(1).pdf.
Okabe A & Miki F 1984, 'A conditional nearest-neighbor spatial-association measure for the

analysis of conditional locational interdependence', *Environment and Planning A*, vol. 16, no. 2, pp. 163-171.
Okabe A, Boots B, Sugihara K et al. 2000, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, Second Edition*, Wiley.
OpenCellID 2009, *OpenCellID project*, viewed 19 September 2009, http://www.opencellid.org.
Ortuzar JDD & Willumsen LG 1994, *Modelling Transport, Second Edition*, Wiley.
Osborn A 2011, *Text message blows up suicide bomber by accident*, viewed 26 January 2011, http://www.leaderpost.com/news/Text+message+blows+suicide+bomber+accident/4172966/story.html.
Papacostas CS & Prevedouros PD 2005, *Transportation Engineering and Planning*, Prentice Hall.
Phithakkitnukoon S, Horanont T, Di Lorenzo G et al. 2010, 'Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data', *International Conference on Pattern Recognition (ICPR 2010), Workshop on Human Behavior Understanding (HBU)*.
Pulselli RM, Pulselli FM, Ratti C et al. 2005, 'Dissipative Structures for Understanding Cities: Resource Flows and Mobility Patterns', *Proceedings of the 1st International Conference on Built Environment Complexity, Liverpool, UK, 11-14 September 2005*.
Pulselli RM, Ratti C & Tiezzi E 2006, 'City out of Chaos: Social Patterns and Organization in Urban Systems', *International Journal of Ecodynamics*, vol. 1, no. 2, pp. 125-134.
Pulselli RM, Romano P, Magaudda S et al. 2008, 'Monitoring human mobility in urban systems: a new technique based on cell-phone activity', *Proceedings of REAL CORP 008 - 13th International Conference on Urban Planning and Regional Development in the Information Society*.
PythonWare 2011, *Python Imaging Library*, viewed 14 April 2011, http://www.pythonware.com/products/pil/.
Qiu Z & Cheng P 2007, 'State of the Art and Practice: Cellular Probe Technology Applied in Advanced Traveler Information System', *86th Annual Meeting of the Transportation Research Board, Washington, January 2007*.
Raja K, Buchanan WJ & Munoz J 2004, 'We know where you are', *IET Communications Engineer*, vol. 2, no. 3, pp. 34-39.
Ratti C, Sevtsuk A, Huang S et al. 2005, 'Mobile Landscapes: Graz in Real Time', *Proceedings of the 3rd Symposium on LBS & TeleCartography, Vienna, Austria, 28-30 November 2005*.
Ratti C, Pulselli RM, Williams S et al. 2006, 'Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis', *Environment and Planning B*, vol. 33, no. 5, pp. 727-748.
Reades J, Calabrese F, Sevtsuk A et al. 2007, 'Cellular Census: Explorations in Urban Data Collection', *IEEE Pervasive Computing*, vol. 6, no. 3, pp. 30-38.
Rinzivillo S, Pedreschi D, Nanni M et al. 2008, 'Visually–driven analysis of movement data by progressive clustering', *Information Visualization*, vol. 7, no. 3, pp. 225-239.
Rojas F, Calabrese F, Dal Fiore F et al. 2007, 'Real Time Rome', *Urban_Trans_Formation. Proceedings of the Holcim Forum for Sustainable Construction, Shanghai, 18-21 April 2007*.
Rose G 2006, 'Mobile Phones as Traffic Probes: Practices, Prospects and Issues', *Transport Reviews*, vol. 26, no. 3, pp. 275-291.
Rudloff A, Lauterjung J, Münch U et al. 2009, 'The GITEWS Project (German-Indonesian Tsunami Early Warning System)', *Natural Hazards and Earth System Sciences*, vol. 9, no. 4, p. 1381–1382.
Saffo P 1997, 'Sensors: The next wave of innovation ', *Communications of the ACM*, vol. 40, no. 2, pp. 92-97.

Schmitz P, Cooper A, Davidson A et al. 2000, 'Breaking Alibis Through Cell Phone Mapping', *Crime mapping case studies: Successes in the field*, vol. 2, no. 1, pp. 56-72.
Sevtsuk A & Ratti C 2010, 'Does Urban Mobility Have a Daily Routine? Learning from the Aggregate Data of Mobile Networks', *Journal of Urban Technology*, vol. 17, no. 1, pp. 41-60.
Shepherd IDH 2008, 'Travails in the Thrid Dimesion: A Critical Evaluation of Three-dimensional Geographical Visualization', *Geographic Visualization*, Dodge M, McDerby M & Turner M (Eds.), pp. 199-222.
Simonite T 2009, 'Smart software promises to create a viable electric commuting car', *The New Scientist*, vol. 204, no. 2732, p. 23.
Skyhook Wireless 2009, *Skyhook Wireless, Inc. Privacy Policy*, viewed 11 November 2010, http://www.skyhookwireless.com/howitworks/privacypolicy.php.
Skyhook Wireless 2010, *Skyhook Launches SpotRank*, viewed 25 March 2010, http://www.skyhookwireless.com/developers/blog/2010/03/25/skyhook-launches-spotrank/.
Smith CW, Wilkinson C, Carlson K et al. 2002, *System and method for providing traffic information using operational data of a wireless network* , United States Patent 6842620.
Song C, Qu Z, Blumm N et al. 2010, 'Limits of Predictability in Human Mobility', *Science*, vol. 327, no. 5968, pp. 1018-1021.
Strategy Analytics 2010, *From Probes to Crowd to Community to Ads – Traffic Data Evolving Rapidly*, viewed 18 May 2010, http://blogs.strategyanalytics.com/auto/?p=105.
Telstra Corporation Limited 2006, *Quick-start Guide for Location Developers Version 3.1, Doc No: TAF0001-123607*, Telstra Corporation Limited.
Thomas JJ & Cook KA 2005, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, IEEE Computer Society.
Titov VV, González FI, Bernard EN et al. 2005, 'Real-Time Tsunami Forecasting: Challenges and Solutions', *Natural Hazards*, vol. 35, no. 1, pp. 35-41.
Tobler WR 1987, 'Experiments in migration mapping by computer', *Cartography and Geographical Information Science*, vol. 14, no. 2, pp. 155-163.
Trevisani E & Vitaletti A 2004, 'Cell-ID location technique, limits and benefits: an experimental study', *IEEE Workshop on Mobile Computing Systems and Applications, December 2004*.
Tufte ER 1990, *Envisioning Information*, Graphics Press.
Varshavsky A, Chen M, de Lara E et al. 2006, 'Are GSM phones THE solution for localization?', *7th IEEE Workshop on Mobile Computing Systems and Applications (HotMobile 2006)*.
Wang H, Calabrese F, Di Lorenzo G et al. 2010, 'Transportation Mode Inference from Anonymized and Aggregated Mobile Phone Call Detail Records', *Proceedings of 13th International IEEE Annual Conference on Intelligent Transportation Systems, 2010*.
White J & Wells I 2002, 'Extracting origin destination information from mobile phone data', *Eleventh International Conference on Road Transport Information and Control, London, 19-21 March 2002*.
Wigren T & Wennervirta J 2009, 'RTT Positioning in WCDMA', *Fifth International Conference on Wireless and Mobile Communications, 2009*.
Wikipedia 2009a, *UMTS frequency bands*, viewed October 2009, http://en.wikipedia.org/wiki/UMTS_frequency_bands.
Wikipedia 2009b, *GSM frequency bands*, viewed October 2009, http://en.wikipedia.org/wiki/GSM_frequency_bands.
Wikipedia 2010, *Pearson product-moment correlation coefficient*, viewed October 2010, http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient.
Wikipedia 2012, *Cell Broadcast*, viewed January 2012, http://en.wikipedia.org/wiki/Cell_Broadcast.

Withers S 2009, *Optus mobile phones tracked for traffic data*, viewed January 2010, http://www.itwire.com/your-it-news/mobility/26115-optus-mobile-phones-tracked-for-traffic-data.

Yuan Y & Raubal M 2010, 'Spatio-temporal knowledge discovery from georeferenced mobile phone data', *Proceedings of the Workshop on Movement Pattern Analysis 2010, Zurich, Switzerland, 14 September 2010.*

Zhao J, Rahbee A & Wilson NHM 2007, 'Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems', *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, no. 5, p. 376–387.

Appendix A: Discarded mobile antenna licensees

AAPT Ltd	AAPT Wireless Holdings Pty Ltd
AAPT Wireless Pty Ltd	ACA Canberra Operations Centre
ACA Central Office	Accredited Person
Administration of Norfolk Island	Airservices Australia Att Bruce Bilton
Apache Energy Limited	Australian Rail Track Corporation Ltd
Bartter Enterprises Pty Ltd	BGC (Australia) Pty Ltd
BHP Billiton Iron Ore Pty Ltd	BHP Billiton Nickel West Pty Ltd
BHP Billiton Olympic Dam Corporation Pty Ltd	BHP Billiton Petroleum Pty Ltd
BHP Bulwer Instrument/Electrical Engineer Reliability and Maintenance	BlueScope Steel (AIS) Pty Ltd
Boyne Smelters Limited	Brizman Pty Ltd
Broadcast Australia Pty Ltd	Burswood Resort (Management) Limited
Central Norseman Gold Corporation Ltd	Channel Seven Queensland Pty Limited
Citipower Pty	City of Wanneroo
Comalco Aluminium Bell Bay Limited	Comalco Limited
Comgroup Australia Pty Ltd	Commissioner of Police NSW Police Force
Comsource International Pty Ltd	Crown Melbourne Limited
CSPB Limited	Department of Defence
Department of the Attorney General (WA)	ElectraNet Pty Ltd
Electricity Networks Corporation	ESSO Australia Pty Ltd
GMG Solutions Pty Ltd	Hamersley Iron Pty Ltd IS&T Comms Team (Seven Mile-Dampier WA) Accounts Payable
Lanfranchi Nickel Mines Pty Ltd	Memo Communications Co Pty Ltd
Mobile Communication Systems Pty Ltd	Motorola Smartnet Pty Ltd
North Flinders Mines Ltd	Northern Rivers Television Pty Ltd
OneSteel Manufacturing Pty Ltd	Power and Water Corporation/Power Directorate
Prime Television Southern Pty Ltd	Qantas Information Technology
QR Network Pty Ltd (Att David Barbeler)	Radlink Communications
Regional Power Corporation	Robe River Iron Associates IS&T Comms Team

	(Seven Mile-Dampier WA) Accounts Payable
Rockhampton Regional Council	Santos Limited (Accounts Payable Supervisor)
Shell Co of Australia Ltd	Snowy Hydro Limited
Tasmania Police	TEC Desert Pty Ltd and AGL Power Generation (WA) Pty Ltd
Transend Networks Pty Ltd	TransGrid
Western Australian Police Service	Woodside Energy Ltd
Worsley Alumina Pty Ltd	Xantic BV