Merkebe Getachew Demissie

COMBINING DATASETS FROM MULTIPLE SOURCES FOR URBAN AND TRANSPORTATION PLANNING:
EMPHASIS ON CELLULAR NETWORK DATA

Doctor of Philosophy Thesis in Transportation Systems, supervised by Professor Gonçalo Homem de Almeida Rodriguez Correia and Professor Carlos Lisboa Bento and submitted to the Department of Civil Engineering of the Faculty of Sciences and Technology of the University of Coimbra

June 2014

· U C ·

Universidade de Coimbra

• U C •

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Merkebe Getachew Demissie

COMBINING OF DATASETS FROM MULTIPLE SOURCES FOR URBAN AND TRANSPORTATION PLANNING: emphasis on cellular network data

PhD Thesis submitted to fulfil the requirements of the Doctoral program in Transportation Systems Department of Civil Engineering, Faculty of Sciences and Technology, University of Coimbra

Coimbra, June 2014

Supervisors

Professor Gonçalo Homem de Almeida Rodriguez Correia

Assistance professor
Civil Engineering Department
Faculty of Sciences and Technology of the University of Coimbra

Professor Carlos Lisboa Bento

Associate Professor with Aggregation
Informatics Engineering Department
Faculty of Sciences and Technology of the University of Coimbra

Financial support

This research work was financed by “Fundação para a Ciência e a Tecnologia” (FCT, Portugal) through the PhD grant with reference number SFRH / BD / 33749 / 2009.

Acknowledgment

I would like to thank all the people who contributed to make this thesis possible. First and foremost, I would like to thank Professor Gonçalo Correia and Professor Carlos Bento for being my supervisors and mentors. Professor Gonçalo, many thanks for your unreserved support and fruitful guidance in the course of the PhD work. You are truly gifted researcher and great educator. Professor Carlos, many thanks for your support and guidance while challenging me to move beyond my intellectual comfort zones. I also thank you for offering me the opportunity to work in the CityMotion project and for providing an excellent environment for my research in the Ambient Intelligence Lab of CISUC (AmIlab).

I thank Professor Pedro Bizarro, who supervised my work during the initial stage of this research. I would also like to thank Professor Antonio Pais Antunes for his genuine concern to my work and for making himself available to listen all my academic and non-academic problems and always come up with solutions. I would also like to thank Professor Marta C. González for hosting me at the HUMAN MOBILITY and NETWORKS lab (HUMNet), Department of Civil and Environmental Engineering, during my stay at the Massachusetts Institute of Technology (MIT).

I like to thank the following companies for letting us use their data:- MEO (Previously known as TMN), cellphone data; Geotaxi, taxi data; Companhia Carris de Ferro de Lisboa, bus data; Servidor de Apontadores Portugueses, Points of interest data; Municipality of Lisbon, and Estradas de Portugal, traffic count data; and Instituto Nacional de Estatistica, census data. I also thank Intergraph Corporation for the use of Geomedia Professional 6.1 (GIS software).

I thank all colleagues and friends with whom I worked and socialized during my PhD study. For this I thank my colleagues at the Spatial Planning and Transportation Engineering Group (Department of Civil Engineering, University of Coimbra) and at the Ambient Intelligence Lab (AmIlab) (Department of Informatics Engineering/Center for Informatics and Systems, University of Coimbra)

I would also like to thank my parents, my brothers and sisters and all close friends for their continued support during the long period to complete the thesis. Finally, the endless thanks go to the almighty God for all the blessings and for making everything possible.

Abstract

All around the World we experience the trends of the last decades on increased urbanization as more and more people shift their living to cities. However, many cities lack the resources to respond to the magnitude of the change in their urban areas, which forces people to compete for the use of land, roads, public transport, and other urban facilities.

As a result of the increasing number of people, cities face an increasing number of private vehicles and commuters which in turn cause various problems such as traffic congestion, parking difficulties, traffic accidents, loss of space for productive activities, public transport inadequacy and undesirable environmental impacts. In the past, public authorities followed approaches that nowadays are financially unsustainable, focused mainly on expanding the road network to alleviate the problem. However, many analysts argue that the solution for these problems is better addressed through intelligent planning and management of the existing urban and transportation systems.

Planning of the urban and transportation system traditionally relied on the knowledge of present and future problems that are associated to the urban growth such as how much travel will be generated, where these trips will take place, by which mode and on which routes. Creating such plans requires information regarding the movement of people and vehicles, knowledge of constituents of the urban system, and understanding the nature of activities at different places.

There are various traditional methods for gathering the raw data necessary for urban and transportation planning. Although these methods have the advantage of providing detailed information, their limited coverage and expensive costs of implementation often make them insufficient. More recently, the spread of massive sensoring, namely through the generalized use of cellphone, is producing massive amounts of data with spatio-temporal detail about our daily activities and traveling patterns, which could be important to the planning of urban and transportation systems given their pervasiveness, low cost, and real time nature.

In this thesis we explore the use of cellphone data for profiling the dynamics of urban activities and characterizing flows of people for planning of urban and transportation systems in cities. Three types of passive mobile positioning data were used: (1) Call Volume, which is the number of calls; (2) Erlang, which is used to measure the equivalent

cellphone traffic per hour; and (3) Handover, which is the process of transferring an ongoing call from one base station to another without interruption of service. Our observations are based on hourly aggregated cellphone data obtained from a dataset from a telecom company in Lisbon, Portugal.

Though passive mobile positioning data is extracted without incurring additional costs and operational risks for the network, it has two main limitations. Firstly, location acquired by this method is at the granularity of a cell sector, which gives uncertainty on the exact location of the collected variables; secondly, it is only acquired when a phone is engaged in a call or short message service. However, we argue that the aggregate cellphone data used in this study remains useful for our analysis, which is at a scale where the lack of a detailed level of precision is not essential. For validation of our results, we collaborated with other data providers in Lisbon to gather different ground truth datasets that could improve our understanding of urban dynamics such as census data, taxi movement, bus movement, traffic count, points of interest, and presence of people.

We proposed new approaches to reflect the potential of passive mobile positioning data for urban and transportation planning. Our approach comprises three stages: (1) exploratory data analysis aimed to discover the kind of relationship that emerges between cellular networks data and urban characteristics, activities, and dynamics at a city-scale; (2) use of cellphone data to detect activities associated to the urban areas in what respects to two aspects of activities: spatial patterns of urban activities, and intensities of urban activities along the hours of a day; and (3) extraction of cellular network data for development of models that predict hourly traffic status.

Our results confirm that passive mobile positioning data, taking the advantage of its pervasiveness and availability with reasonably less cost, can provide ways to analyse the dynamics of urban activities at a larger scale. In addition, our approach complements traditional urban data collection methods, which are usually made available less frequently to urban and transportation planners, and is especially useful for developing countries where other approaches are too expensive.

Keywords*: cellular network; Erlang; Handover; traffic estimation; transportation planning; urban planning; urban activity; urban dynamics.*

Resumo

Em todo o Mundo continua hoje a verificar-se a tendência das últimas décadas de crescente urbanização à medida que mais e mais pessoas mudam as suas vidas para as cidades. Apesar dessa mudança, muitas cidades não têm os recursos necessários para responder a estas alterações, o que força os seus habitantes a competir pela utilização de recursos escassos como sejam o solo, as estradas, os transportes públicos e outros serviços urbanos.

Como resultado deste processo de crescimento da população urbana, observa-se nas cidades o aumento do número de viagens pendulares e correspondente aumento do número de veículos particulares, o que tem como resultado vários problemas como o congestionamento, a escassez de estacionamento, os acidentes, o custo de oportunidade de utilização do espaço ocupado por infraestruturas, a redução do nível de serviço dos transportes coletivos e os impactos ambientais. No passado, as agências governamentais seguiram uma política com custos muito elevados, focada na expansão da rede de estradas para aliviar a pressão sobre o sistema de tráfego. Contudo, muitos analistas discutem esta abordagem argumentando que para mitigar estes problemas será preferível planear e gerir de forma mais inteligente o sistema de mobilidade.

Planear as cidades e a sua mobilidade tem tradicionalmente recaído sobre a importância de conhecer os problemas presentes e futuros que estão associados ao crescimento urbano como o número de viagens que são geradas, as suas origens e destinos, modo e caminhos escolhidos. Para um planeamento eficiente é necessária informação acerca dos movimentos das pessoas e dos veículos, conhecer bem as redes existentes, e compreender a natureza das diferentes atividades que são desempenhadas em cada parte da cidade.

Há vários métodos tradicionais para recolher os dados necessários ao planeamento urbano e de transportes. Apesar destes métodos terem a vantagem de dar informação muito detalhada, as suas limitações de cobertura e altos custos de implementação e manutenção, por vezes, tornam-nos incomportáveis. Mais recentemente, o aumento da utilização massiva de sensorização, nomeadamente a utilização generalizada de telemóveis está a produzir grandes quantidades de informação com detalhe espácio-temporal acerca das nossas atividades e padrões de deslocação, que poderão ser

importantes para o planeamento das cidades e da sua mobilidade dada a sua penetração no território, o seu baixo custo e disponibilidade em tempo real.

Nesta tese explorámos a utilização dos telemóveis para traçar o perfil das dinâmicas urbanas e caraterizar os fluxos de pessoas com o objetivo de planear as cidades e os seus sistemas de transportes. Três tipos de informação passiva foram utilizados: (1) volume de chamadas, que é o total de chamadas num intervalo de tempo; (2) Erlang, que é o tempo total de chamadas durante um intervalo de tempo; e (3) a entrega de chamadas num intervalo de tempo, que é o processo de transferir chamadas ativas de uma estação base (torre) para outra estação base. As nossas observações são agregadas ao intervalo de tempo de uma hora, obtidas numa base de dados de comunicações na cidade de Lisboa, Portugal.

Apesar da informação passiva móvel ser extraída sem incorrer em custos e riscos de operação da rede, esta tem duas limitações. A primeira está associada a que este método de localização tem lugar à escala do setor de cada célula (estação base), o que conduz à incerteza acerca do local da chamada; a segunda é que a informação diz respeito apenas a chamadas que foram realizadas ou a utilizações do serviço de mensagens. Apesar destas limitações considera-se que os dados dos telemóveis utilizados nesta tese permanecem relevantes para as analises efetuadas que são realizadas a uma escala em que a falta de precisão não deverá ser crítica. Para validar os nossos resultados utilizaram-se várias fontes de informação em Lisboa que permitiram caracterizar a realidade atualmente existente e melhorar a nossa compreensão das respectivas dinâmicas urbanas. As fontes consideradas para validação foram: dados dos sensos, movimento dos táxis, movimento dos autocarros, contagens de tráfego, pontos de interesse e presença de pessoas a cada hora em cada área da cidade.

Neste trabalho foram propostas novas abordagens que refletem o potencial dos dados passivos dos telemóveis para o planeamento das cidades e da sua mobilidade. As abordagens seguidas focaram-se em três pontos centrais: (1) exploração dos dados no sentido de descobrir o tipo de relações que se podem encontrar entre os dados dos telemóveis e as características urbanas, atividades e dinâmicas à escala da cidade; (2) análise dos dados dos telemóveis para detetar atividades associadas à área urbana no que respeita a dois aspetos: padrão espacial dessas atividades e intensidade dessas atividades ao longo do dia; (3) utilização dos dados dos telemóveis para desenvolver modelos de previsão do estado do tráfego na rede adjacente às torres.

Os resultados confirmam que os dados passivos da utilização dos telemóveis, considerando ainda a sua disponibilidade a baixo custo, podem constituir uma boa forma de analisar as dinâmicas das atividades urbanas a um nível abrangente. Além disso, a abordagem que foi utilizada complementa os métodos tradicionais de recolha de dados, que estão disponíveis com menos frequência para os planeadores da cidade e do sistema de transportes, especialmente em países em vias de desenvolvimento em que outras abordagens são demasiado dispendiosas.

Palavras Chave: *rede móvel; Erland; transferência de chamada; estimação de tráfego; planeamento de transportes; planeamento urbano; atividades urbanas; dinâmicas urbanas.*

Table of Contents

Figures

Tables

Chapter 1 Introduction

1.1. Problem statement

Urban areas are locations which accommodate persons living, working, shopping and carrying out different daily activities, and also support complex spatial structures that present patterns of constantly changing colours and shapes. In making all these activities and movements within the limited spaces of the urban areas there is natural competition for the use of shared space such as roads, and public transportation facilities. In order to manage those flows of people and their different interests in using the urban space there is the need to accurately characterize those movements. However due to their great variations with time there is considerable difficulty in obtaining this information in real time through traditional methods thus making it difficult to plan and manage transportation infra-structure.

The steadily increasing flow of people and vehicles in urban areas have caused several urban problems such as: traffic congestion, parking difficulties, pollution, loss of space for productive activities, traffic accidents and public transport inadequacy, which impose costs on the economy and generate multiple impacts on the urban environment. The majority of these problems are caused by an inept urban transportation system, especially when the system fails to satisfy the mobility requirements that are generated by the various activities in a city (Rodrigue et al., 2006). To address these problems efficiently and on time, urban and transportation planners should develop means to profile urban activities in a dynamic, almost real time way. Profiling of urban activities requires reliable and detailed information regarding the movement of people and vehicles; understand the nature of activities at different places; and knowledge of constituent of the urban system.

Most developed cities have comprehensive inventories that relate the existing and projected features for the urban and transportation planning: census and other demographic data, existing and planned land uses, street and transit facilities, movement of people and vehicles, spatial location of activities, etc. However, most inventories follow time consuming techniques to acquire data, and large scale inventories are laborious and expensive. In addition, there is the need to keep these databases up-to-date to provide good factual bases for urban and transportation planning and forecasting. On the other hand,

cities in developing countries have great difficulties in getting a grip on existing and future activities and dynamics of their urban systems. This is partly because these cities face huge changes in population and economic growth, changes on existing facilities and so on that change travel requirements every year. The other reason is that most of the developing cities do not have enough budget to collect detailed information for the urban and transportation planning.

Traffic management sectors use several techniques to gather raw traffic information. The two major data collection categories are point-detection and vehicle-based detection systems (OECD, 2007). These methods have the advantage of providing detailed information, but very often they are not sufficient because of their limited coverage and expensive costs of implementation and maintenance. Planners also use customized inventories to obtain different datasets that are useful for the planning of the urban and transportation systems. One particular example is the travel surveys which are used to estimate the mobility patterns of people. However, travel surveys suffer from relatively small samples due to their cost and are usually conducted at infrequent intervals, thus, many metropolis learn the presence of new trends only after the release of new census results and only in the case when mobility questions are part of those comprehensive surveys (Becker et al., 2011b).

The coverage of urban and transportation data collection has significant importance on the success of the planning, design, monitoring and management of activities in urban areas. One recent and innovative way that has been tried to obtain urban and transportation information is through analyzing data from the cellular network usage.

The use of cellular networks data for urban and transportation studies has been a topic of research for quite some time now (Caceres et al., 2008; Gundlegård and Karlsson, 2009); however, it is far from being mature. The fact that many people are users of cellphones makes the cellular network a natural alternative or/and complementary source of information to more traditional method. The potential of cellphone data is quite intuitive; Cellular networks are ubiquitous and cellphones can be found with a constant proximity to their owners that would allow transport planners and operators to opportunistically detect the locations of large populations.

Cellular networks data can be extracted in a cost efficient way by using existing signaling data without the need to invest on a sensor infrastructure. An example of this

application for comparison of cost estimation between traditional loop based surveillance and cellular networks based systems is the project CAPITAL (Cellular Applied to ITS Tracking and Location). As a case example the corridor between Baltimore-Washington D.C. which is 38.6 km in length was chosen. A 0.8 km spacing of loop detectors and a cellular network system with 23 towers was used to cover the corridor. Cost data for traditional loop based surveillance was obtained from average values of the cost estimates by Federal Highway Administration and Maryland State Highway Administration. The result showed that the cellular networks based system (cost of $2,736,300) is only a little less expensive than the loop based system (cost of $2,831,500) based on the average total cost estimate. However, the cost estimate for the cellular network system was done for the technology that existed before 1995. Since then there were some fundamental changes: a switch from vehicle based to portable cellular phones; and the typical transmission from the towers to a cellphone is changed to a bidirectional antenna that used to be an Omni-directional antenna. The cost for the cellular based system was exaggerated because of the 1995 cellular environment required to place a "direction finding system" that costs more than 85% of the average total cost estimate (University of Maryland Transportation Studies Center, 1997).

Following the steps of previous research and trying to go deeper in using this type of data in this thesis we aim at defining methods to use three types of passive mobile positioning data for profiling the dynamics of urban activities and characterizing flows of people with the aim to help planning urban and transportation systems in cities. The three types of data are: Call Volume, which is the number of calls; Erlang, which is defined as one person-hour of phone usage or two people talking for half hour each and so on; and Handover, which is the process of transferring an ongoing call from one base station to another without interruption of service. The significance of the results of the models that we develop is always examined through comparison with ground truth information. This includes establishing a comprehensive set of indicators that define the activity and mobility features at different locations in a city. Mobility and activity ground truth is represented by a number of indicators related to people´s presence, residential buildings and Points of Interest (POIs), traffic volumes, bus movement, and taxi movement.

1.2. Urban and transportation information from cellular networks

The fact that cellular networks' data is increasingly becoming available to researchers led them to carry out works that have produced important lessons for urban and transportation studies. In recent years, several experiments were performed to understand the implication of cellular networks data for monitoring of the urban activities and their dynamics. A first high level differentiation regarding the collection of mobility and activity related cellular networks data is associated with active and passive monitoring techniques (Gundlegård and Karlsson, 2009).

In the case of the active monitoring techniques, cellphones are located periodically. A typical technique applied to identify the location information of a cellphone is the paging procedure, which is usually performed to forward incoming calls/connections to the respective cell. This monitoring technique produces huge traffic in the network and causes increased network costs. An additional disadvantage of this technique is that it drains the cellphone battery during communication between the cellphone and the network. The position of cellphones through active monitoring can be extracted with good precision ranging from cell dimension (Cell ID-based positioning) down to a few meters (A-GPS) (Valerio, 2009).

Another category of active monitoring technique is the application-based active monitoring, where a cellphone runs dedicated software that reports its whereabouts to a server outside the cellular network. A typical example of application-based active monitoring technique is a car with GPS receiver equipped with a GPRS transceiver, which reports the position to a server via cellular network (Valerio, 2009).

In 2008, a trial experiment has been launched by the Mobile Century project, which used the application-based active monitoring technique. The demonstration was started by deploying 100 vehicles with GPS equipped Nokia cellphones to gather information such as continuous location and speed profile of vehicles. In theory this information has a higher potential to explain the traffic condition (Herrera et al., 2010). An extension of the Mobile Century project, Mobile Millennium, also planned to collect location and speed profile of vehicles from drivers who are willing to share data and in return receive value added traffic information as a reward.

The study by Hongsakham et al. (2008) gathered Cell Dwell Time (CDT) (the length of time that a mobile device remains registered to a base station until it switches to another

base station) data through Nokia cellphone loaded with cellular probe software to infer traffic congestion on a heavily utilized expressway in Bangkok. Puntumapon and Pattara-atikom (2008) used software installed on volunteers' cellphones for collecting CDT to develop a Naive Bayes model that differentiates pedestrian and sky train passengers by analyzing cellphone user permanence in a cell. Results showed classification accuracy of up to 93.1%.

A study by Cayford and Johnson (2003) investigated the impact of several system parameters, such as sampling frequency, accuracy of the locations, and number of locations available in a given area. A good review of projects that are based on application-based active monitoring techniques and various field-tests can be found in Yim (2003).

In the case of passive monitoring techniques, cellphone information is collected from the network without causing additional traffic. The disadvantage related to passive monitoring technique is that location data would be insufficient if a high number of users are not active (calling or connected) (Gundlegård and Karlsson, 2009).

One of the first big projects aimed at using mobile phone as traffic probes was the CAPITAL project, mentioned before, which was started in 1994. In this project data was collected from eight base stations with the objective of obtaining estimates of speed and travel times. Unfortunately, with the location accuracy of about one hundred meters there was no reliable speed estimate and incident detection (University of Maryland Transportation Studies Center, 1997).

The study by Ratti et al. (2006) presented preliminary findings of the "Mobile Landscapes" project: an application in the metropolitan area of Milan, Italy, based on the geographical mapping of passive mobile positioning by using Erlang values. Results enabled a graphic representation of the intensity of urban activities and their evolution through space and time.

The study by Ratti et al. (2005) demonstrated results of the "Mobile landscapes" project for the Case of Graz city, Austria. Three types of graphically appealing maps of the urban area of Graz were developed and shown in real-time. Cellphone traffic intensities, handover values, and traces of registered users as they move through the city were used.

Calabrese et al. (2011) also developed a real-time urban monitoring system for the city of Rome. The system uses the Localizing and Handling Network Event Systems (LocHNESs) platform developed by Telecom Italia for the real-time evaluation of urban

dynamics through the use of Erlang, handover, and cellphone trajectories of registered users. This study has also combined the instantaneous positioning of buses and taxis to provide information about urban mobility.

1.2.1 Lessons learned

Speed and travel time estimation

Since the start of cellular networks data use for the urban and transportation studies road traffic parameters estimation and prediction are the most common topic of research. The most common cellular networks' data used for travel time and speed estimation is the data on the Handover process (double handovers). A double Handover data has time information on the entrance and exit of a given cell. This data alone could be used for travel time estimation. The time information coupled with the distance between the initial and the final handover events will help to compute travel speed estimate for the road section running through the cell between two handover points. There are some limitations associated to the travel time and speed estimations: only cellphones that make sufficiently long calls to cross through the entire cell are considered in the sample, a situation that is not frequent; it is a hard task to identify the route followed by the cellphone probe when multiple road links pass through a cell; and it may not be always the case where cellphones moving on certain road links obtain the handover events exactly at the same location (Caceres et al., 2008).

Gundlegard and Karlsson (2009) investigated the location accuracy of handover points in both the GSM and UMTS systems. The experiment was carried out on a 900 meters long road segment in a "sparse" urban environment. The handover data was collected from a GSM terminal and a UMTS terminal simultaneously, with ongoing cellphone calls on a vehicle driven fifteen times back and forth. The first accuracy measure used was consistency, which was defined as the percentage of times a specific handover is completed in the same handover zone. In the UMTS case there was consistency 92.5% of the time, and the GSM handovers were much more scattered with consistency of 43.8%. The second accuracy measure was mean location error, which was computed as the average values of location deviations from the average handover point. The result showed that the mean location error for the GSM and UMTS were, respectively, below 40 meters and 20 meters.

Evolution from 2nd to 3rd generation

Currently, there are many countries operating with a combined UMTS and GSM terminals. Karlsson and Gundlegard (2006) investigated suitability of the UMTS terminals as a source of road traffic information when compared to the GSM terminals. In UMTS the cell breathing, which is a mechanism to change the size of the cells dynamically, may be utilized more often than in GSM and this may potentially induce error on travel time and speed estimations. In cell breathing overloaded cells are allowed to offload some traffic to neighboring cells. This can be done by reducing the cell size of overloaded cells while neighboring cells increase their service area to receive more traffic. Travel time and speed estimation require double handover points and the time between these two points. Cell breathing can cause these points to change over time. However, in UMTS the use of smaller cells and soft handover may have the potential to increase the information quality when road traffic information data is extracted compared to using the GSM system (Karlsson and Gundlegard, 2006).

Privacy issues

Cellular data can be aggregated at different spatial and temporal scales. For example datasets aggregated at the granularity of cell tower include: Call Volume (the number of calls), Erlang (total communication time), number of SMS, number of handovers, and number of location updates etc. Unlike aggregate cellular networks data, individual cellular networks data is rarely available. This is because individual data is not easily manageable and the tracking of individual cellphone data raises significant privacy concerns. Previous studies addressed the privacy concern in different ways: some studies obtained cellular networks data aggregated both spatially and temporally; and in their analyses and discussions they did not allow the identification of persons on geographical or temporal grounds (Reades et al., 2007; Reades et al., 2009). Some other studies apply traces of location and time of individual calls. However, these studies replaced the real identity of the phone users through randomly assigned IDs to deal with privacy issues (Ahas et al., 2010).

Integration of active and passive techniques

In the past, the majority of cellular network-based studies have been performed based on the active monitoring technique. In recent years, the focus is shifting towards the passive monitoring technique. However, there are some occasions where passive technique can be integrated with active monitoring technique (Valerio, 2009): (1) Low number of active users: passive monitoring requires sufficiently high number of active users (calling or connected). When the number of active users reduces, which means that the network is underutilized, active monitoring technique can be activated without fear of affecting the network performance, and (2) Event uncertainty: when passive monitoring technique senses some road anomalies which cannot be identified clearly, the active monitoring technique can be activated temporarily to assist detection.

1.3. Our approach

In the past, various research papers have been published presenting the results of several projects and independent studies that have been carried out on using information from cellular networks for urban and transportation planning. These studies investigated a wide range of issues such as, types of cellular networks data collection techniques (Valerio, 2009); qualitative and quantitative representation of the cellular networks data (Calabrese et al., 2011; Ratti et al., 2005; Ratti et al., 2006; Reades et al., 2007); travel time and speed estimation (Alger et al., 2005; Bar-Gera,2007; Caceres et al., 2007; Herrera et al., 2010; Liu et al., 2008); correlation between cellphone traffic and vehicular traffic (Becker et al., 2011a; Caceres et al., 2007; Thiessenhusen et al., 2003; Vaccari et al., 2009); origin-destination estimation (Calabrese et al., 2011a; Iqbal et al., 2014; Pan et al., 2006; White and Wells, 2002); congestion detection (Hongsakham et al.,2008; Thajchayapong et al., 2006); incident detection (University of Maryland Transportation Studies Center, 1997); route classification (Becker et al., 2011a); inferring land use patterns (Becker et al., 2011b; Soto and Frías-Martínez, 2011; Toole et al., 2012); and inferring frequently visited locations (Ahas et al., 2010; Csáji et al., 2013; González et al., 2008; Isaacman et al., 2011).

The aforementioned studies showed that cellular network based systems have great potential for the urban and transportation studies although more research is required in order to use it as efficiently as possible. The technology has shown very promising

advantages especially in detecting large samples of population (Becker et al., 2011b). In addition, unlike fixed sensor systems, cellular networks are ubiquitous in today's cities because they imply wide area coverage (Ratti et al., 2005). An additional advantage of cellular network based systems is their capability to produce results in a faster way in contrast with transitional methods. This happens in the case of traditional OD estimation through surveys that would take a long time from the initial data collection until the computation of the final result and they only characterize one point in time (Calabrese et al., 2011a).

In spite of the important efforts in applying the cellular network data for the urban and transportation studies, some challenges are still to be addressed by the research community. Hence there is a need to further explore this type of data and experiment methods that are able to extract meaningful conclusions from its use. The three main goals of the thesis are:

A. To build relationships between the cellular network usage and the urban and transportation usage.

Cellular networks are not primarily dedicated to the urban and transportation data collection purpose. The major advantage of applying cellular network data is to use them as complimentary or/and alternative data sources for the urban and transportation studies. However, previous studies miss to prove the following fundamental relationships that are important for the development of urban and transportation information from cellular networks at a city-scale: (1) the presence of a spatial relationship between road network and cellular network infrastructures; and (2) the presence of time relationship between different urban activities (vehicle, people) and cellular activities (previous studies addressed this to only a certain degree). For example, is there significant cellular traffic when there is higher vehicular traffic? One of the goals of this thesis is to investigate the presence of such relationships.

B. **To investigate the potential of the cellular networks usage in detecting the Spatio-temporal distribution of the urban activities**.

The complexity of urban activities greatly depends on people´s motivations and interests to travel, socio-demographic characteristics, economic, cultural and technological factors,

urban structures and forms, etc. Thus, one of the goals of this thesis is to investigate whether the cellphone data can be used to infer the pattern and intensity of urban activities and compare it with ground truth.

C. To develop models that can use cellular networks data to predict road traffic status.

The interaction within the traffic system can be quantified by the macroscopic traffic stream parameters that describe aggregate traffic flow characteristics and relationships: speed, density, and volume. Regardless of the important efforts in applying cellular networks data for the development of speed, density, and travel time estimation, traffic volume estimation has been neglected in prior works. Understanding the variation of traffic by hour of the day is one of the most important factors for traffic monitoring. For example, peak hour traffic in the morning and afternoon largely determine the requirements for facilities. Another goal of this thesis is to employ mobility related cellular networks data to predict hourly road traffic. To verify this, hourly Handover count is used as a proxy to predict the hourly road traffic level in the arterial road links.

1.4. Research questions and contributions

This thesis gives contributions to the following research questions of relevance to urban transport planners, transport geographers, and urban planners: Is cellular network data adequate and available in the form required for urban and transportation analysis? Is it possible to use cellphone data to detect intensity of urban activities in a city? Are we able to explain the varying urban activity patterns along the day in different parts of a city? How is it possible to highlight critical spots in the urban road network without incurring great costs?

This thesis contributes to an understanding of where the cellular networks data can be applied in cities at different growing stages. In a developed city with relatively stable spatial structure and slow projected population growth, the existing land use and transportation network may be expected to change slowly. However, there are situations that cause changes in the urban use over time. Therefore, planners should perform frequent estimates of the urban activity and its dynamics and measure the likely consequences of the changes upon transportation and urban uses. The analyses in this thesis provide bases to

carry out such estimation and an interesting approach should also be considering these approaches as means for updating traditional urban and transportation data, which are usually made available less frequently to planners and policy makers.

Most of the cities in the developing world have high projected population growth and uncertain development dynamics, such as new road development, provision of public transportation, and other public facilities. The task of urban and transportation planning would be more challenging as the travel patterns will change regularly as a result of the population trying to adjust to new types of land uses (Berke et al., 2006). In addition to that there is poor availability of data for urban and transportation planning, where most of the cities do not carry out traffic counting in a regular basis, and there are no comprehensive travel surveys. However, most of the developing cities are gaining good mobile penetration. Thus, these cities could take advantage of the approaches from this thesis to be used as alternative low-cost estimators of traffic and census data.

1.5. Organization of the thesis

This thesis is organized into six chapters. All chapters, except the introduction (Chapter 1), data description (Chapter 2), and conclusions (Chapter 6), were written in the format of papers in international peer-reviewed journals and they have not been altered in any meaningful way during the preparation of the PhD document. The chapters which are based on scientific articles (Chapter 3, Chapter 4, and Chapter 5) are each self-contained in the sense that each can be read and understood independently. Undoubtedly, the nature of this format involves the repetition of some background information throughout the thesis, but this is overshadowed by the advantage to the reader of having an approachable document clearly defined into chapters that relate to the specific subject. However, the chapters are interrelated and do form a consistent PhD formal document.

All the chapters address the application of cellular networks data for the profiling of the urban activities and their dynamics. Moreover, the results in the early chapters are consecutively used as a stepping stone for the ones that follow. The analyses in the early chapters were exploratory studies aimed to discover the kind of relationships between the cellular networks data and the urban characteristics, activities, and dynamics at a city-scale. The discovered relationships were used to narrow the scale of the analysis down to

detailed urban areas and road links to reflect the potential of the cellular networks data to a specific application in the urban and transportation planning.

Chapter 2 introduces the datasets obtained from different sources and used in this thesis. In it we identify the different attributes which are essential for our analysis. Since we have data from different sources that were developed and maintained independently to serve specific needs, the data in each source were represented differently and this results in a large degree of heterogeneity. Therefore, this chapter explains the type of data integration that was performed. This chapter also focuses on data cleaning, which is the process of structuring the dataset to make further manipulation, visualization and modeling easy. This process consists of a range of activities, such as outlier checking, date parsing, and missing value imputation.

Chapter 3 is dedicated to exploratory data analysis. The goal is to find relationships we did not know between the cellular network data and the different urban characteristics, activities, and mobility features. Initially, a pre-processing of the cellular network data was performed. Then, experiments were carried out to understand the city dynamics through Geographic Information System (GIS) visualization and statistical analysis. The GIS visualizations provided a qualitative explanation of how the movement of calls is useful in highlighting the flow of people in urban infrastructures. Statistical analysis was employed to discover the presence of significant relationships of the aforementioned data features. The analysis in this chapter allowed us to discover new connections and also works as a base to define future analyses.

Chapter 4 addresses the work done on the profiling of activities in urban areas. The goal is to apply cellular network data to detect the pattern and intensity of the urban activities. Profiling of urban activities has largely relied on the knowledge of land use patterns, and the transportation networks. Land use and transportation infrastructures once they are on the field remain with the same form for a long time, the activities, however, often change for the same land use and accessibility due to some trend or new movement in the urban area. Thus, planning practices that build structural relationships between trip rates and stable elements of urban spatial structure such as land use categories and aggregate accessibility indicators miss out the dynamics associated to these elements such as, having the same land use generating different types of activities in what respects to their trip generation and modal split, and the movements of people that can change in a short time

interval as a result of new events happening in a neighborhood. Previously, planners used traditional surveys to detect changes on the pattern of activities in urban areas. However, static information mostly comes from traditional survey methods that are expensive and time consuming giving planners only a picture of what has happened and needing the active involvement of the traveler to respond to the survey. In Chapter 4 the recent possibility of understanding activities in an urban environment using three types of cellphone data (Call Volume, Handover, and Erlang) is explored. Fuzzy c-mean clustering algorithm was applied to the cellphone data to create clusters of locations with similar features in what respects to two aspects of activities: daily patterns and intensity. In order to validate those clusters as actual predictors of human activity, the results were compared with clusters formed using ground truth variables: presence of people, buildings, POIs, bus and taxi movements.

Chapter 5 addresses predictive data analysis performed in the thesis. The goal is to use the handover count data to build a model that predicts road traffic status. Traffic management agencies carry out different types of traffic counts. Permanent Traffic Counter (PTC) is the preferred counter that provides traffic statistics throughout the year. Due to its expensive installation and maintenance costs, PTC has limited road network coverage; therefore, agencies choose to utilize sample traffic counts from Seasonal Traffic Counter (STC). However, road network covered with STC lack continuous observation to update changes in traffic in a regular basis, which causes a serious impediment to the effective use of active traffic management schemes. In the absence of a traffic count from a specific site, prediction is usually made through the average traffic obtained from other places or historical traffic data of a given location. However, this kind of approach forces traffic management authorities to rely on an incomplete picture of the traffic stream in a city. This chapter explores a complementary method to gauge the status of road traffic conditions through the use of cellular network handover counts. Two models, multinomial logit and artificial neural network, are used to relate traffic and handover counts.

Finally, a summary of the research work done in the thesis and its conclusions are summarized in Chapter 6 along with the discussion of future areas of research.

1.6. Publications

As mentioned in Section 1.5, part of this thesis is organized on the basis of peer-reviewed papers published or under revision in journals. The work in Chapter 3, "Exploring cellular networks handover information for urban mobility analysis", is published in the Journal of Transport Geography:

- Demissie, M.G., Correia, G.H., Bento, C., 2013. Exploring cellular network handover information for urban mobility analysis. Journal of Transport Geography, 31(2013), pp. 164-170).

The research work in Chapter 4, "Analysis of the pattern and intensity of urban activities through aggregate cellphone usage", is under review in the Transportmetrica A: Transport science. The research work in Chapter 5, "Intelligent road traffic status detection system through cellular networks handover information: Exploratory analysis", is published in the Transportation Research Part C: Emerging Technologies:

- Demissie, M.G., Correia, G.H., Bento, C., 2013. Intelligent road traffic status detection system through cellular networks handover information: An exploratory study. Transportation Research Part C: Emerging Technologies, 32(2013), pp. 76-78.

Along with the publications in the international journals the work in this thesis has also been presented and discussed on the following conferences:

- 17th EURO working group on transportation meeting, Seville, Spain, July 2 to 4, 2014.
- 4th MIT Portugal program conference, Coimbra, Portugal, June 27, 2014.
- 11th Annual Transports Study Group Conference, Covilha, Portugal, January 6 to 7, 2014.
- 13th World Conference on Transportation Research, Rio de Janeiro, Brazil, July 15 to18, 2013.
- CITTA 6th Annual conference on planning research, Coimbra, Portugal, May 17, 2013.
- 3rd MIT Portugal Program Conference, Guimarães, Portugal, May 28 to 29, 2012.
- 9th Annual Transports Study Group Conference, Tomar, Portugal, January 5 to 6, 2012.

Chapter 2 Data description

2.1. Introduction

In a rapidly changing environment of urban areas, there is a need to understand the status of urban activities and their dynamics to manage the urban and transportation systems. Understanding urban activities and their dynamics involves acquisition of various types of urban and transportation related datasets. Transport planners collect data for the evaluation, assessment, and design of transportation facilities, such as roadways, pedestrian and bike lanes, and public transportation routes. On the other hand, urban planners require a wide variety of information from different sources regarding the use of land, urban environment, and transportation networks in order to manage the smooth development of urban areas. The coverage of urban and transportation data collection has significant importance on the success of planning, design, monitoring and management of activities in urban areas.

In order to facilitate the assessment of present and future status of the urban activities, it is necessary to access accurate information and continuous monitoring of the urban and transportation systems. As a result, several attempts are now being made to adopt suitable data collection methods which can be operationally convenient and cheaper than traditional ones. This includes the use of both manual and automatic data collection methods along with suitable tools to analyze collected data.

The municipality of Lisbon is used as a case study area to illustrate the analyses carried out throughout this thesis. Based on the 2011 census, the total population in Portugal is 10,555,853, with dwelling size of 5,879,845. Lisbon is the capital of Portugal and the center of the Lisbon Metropolitan Area (LMA). The LMA has a population of 2.3 million and has 18 municipalities with a total area of 2957.4 km^2, where about 24.3% of the population resides in the municipality of Lisbon (INE, 2013). Figure 2.1 shows the municipality of Lisbon divided into 53 administrative parishes (Freguesias) (marked by bold colored borders) and its neighboring municipalities also divided into administrative parishes (marked by light colored borders).

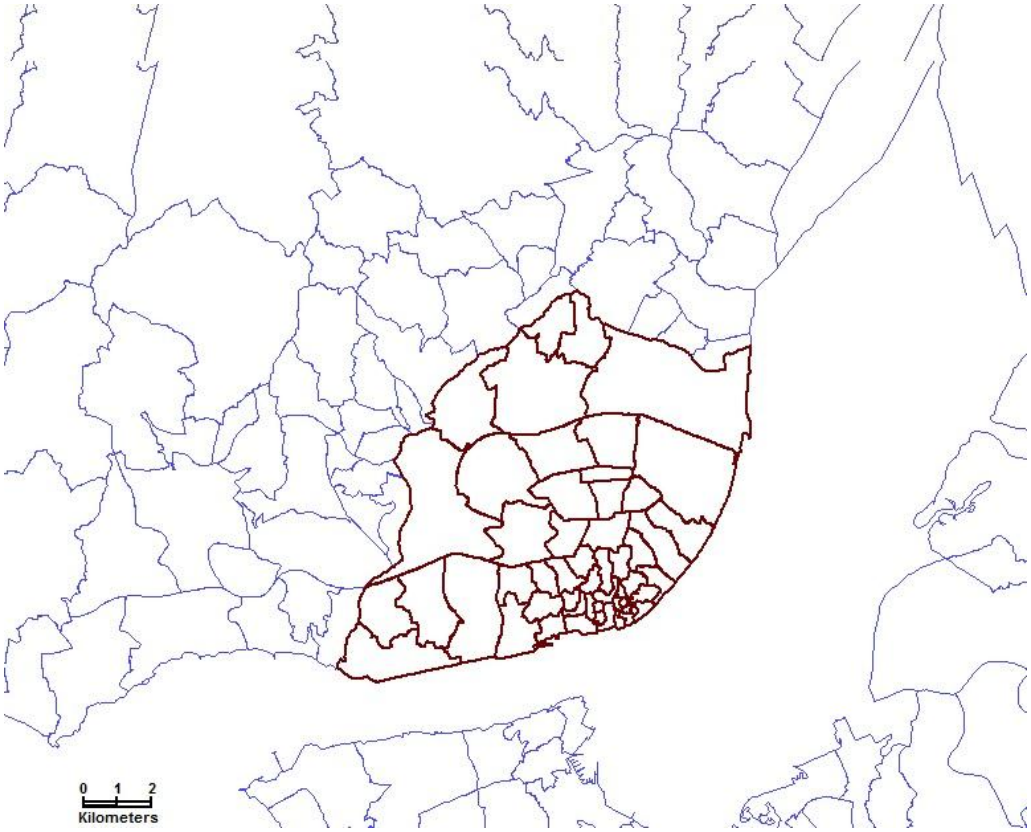


Figure 2.1 Lisbon and its neighboring municipalities

This chapter has been prepared to provide basic information regarding the urban and transportation data with the following three main objectives: (1) introduce the most common urban and transportation data collection methods; (2) introduce the urban and transportation datasets of the municipality of Lisbon, which were acquired from different sources and used in this thesis; and (3) explain data preparation procedures applied to the data obtained from different sources, which is the process of cleaning and structuring the datasets to make further manipulation, visualization and modeling.

2.2. Urban and transportation data collection methods

There are various methods for gathering the raw data necessary for urban and transportation planning. The methods can be split into two categories: human observation and remote sensing. Most of the data collection techniques have been deployed for many years and the most important ones are briefly described hereafter:

2.2.1. Point detection methods

Point detection methods involve engaging measuring equipment at a specific location and record the traffic data to estimate traffic performance over a segment of a roadway. The main problem of point detectors is that they record data from a single location and that location may not be an accurate representation of the remaining roadway segment to which those data are associated. This method is usually used to report data on vehicle volume, lane occupancy, vehicle speed and classification (with dual loop). The most common point detection methods are:

Inductance loops: this is an expensive, but well-known and reliable technology. The main limitation is that inductance loops require lane closure during installation and maintenance.

Video image detection: this technology is introduced partly to deal with the limitations of loop technology. In most cases cameras are mounted above the ground. Therefore, traffic lanes do not need to be closed during installation, repair or to adjust the data collection devices. The main limitation of this technique is that it works poorly in low-visibility weather conditions.

Microwave radar technology: this technology is developed, in part, in response to the limitations of loop and video technologies. It is not affected by bad weather or low light conditions. However, it reports slightly less accurate traffic volume counts than the once obtained with loop and video detection.

2.2.2. Vehicle based detection methods

Most vehicle based detection techniques are recent methods for traffic data collection. The Vehicles are considered as moving sensors that detect traffic information directly from the traffic stream. It is a good source of traffic information to obtain travel time and speeds.

Probe vehicles: Probe vehicles, which are not deployed into the traffic stream for the primary purpose of data collection, are referred to as passive probe vehicles. On the other

hand, instrumented vehicles that are driven at regular intervals down specific roadways for the purpose of gathering traffic data are referred to as active probe vehicles. For the case of active probe vehicles, datasets can be collected automatically or manually recorded.

Beacon-based probe vehicle: In this technique, a device (beacon), which is mounted along a roadway, probes electronic vehicle tags as vehicles pass that reader location. Then, it requires matching the time and location data associated with each vehicle that passes from one beacon location to the next to compute travel time of a given vehicle between two consecutive beacons.

Cellphone tracking: this technique is based on tracking of cellphone movements. By restricting the analysis to those cellphones along the roadways, cellphone traces could provide means to measure traffic conditions on the roadway. One of the major advantages of this technique is that the number of vehicles with cellphone is quite high and it will potentially be used to monitor the entire roadway system. The limitation associated to this method is that tracking of individual cellphone data enables transport analysts to identify the whereabouts of a person without the person´s permission. Data providers deal with this matter through anonymizing the data before they pass it to third party or ask the user to give permission such that positioning data can be used.

Satellite tracking: this technology (like GPS devices) provides information like current location, heading, and speed with a reasonable degree of accuracy. Currently, there are many services that use satellite-tracking devices as a primary component of location information. The difficulty with this system is that the information obtained is stored on-board in the vehicle itself and it is necessary to provide some communication means to/from vehicle in order to obtain the relevant data.

2.2.3. Manual counts

Manual counting is the most common traditional data collection method. Manual counts are usually used when the effort and expense of using automated equipment is not justified and to gather traffic data that cannot be efficiently obtained through automated counts, such as vehicle occupancy rate, pedestrians, and vehicle classifications. Most applications

of manual counting require small samples of data at any given location and are used for periods of less than a day. The equipment most frequently used for manual traffic counting includes tally sheets, mechanical count boards and electronic count board systems.

2.2.4. Urban and transportation inventories

Planners perform several inventories that relate the existing and projected features of the urban and transportation systems such as, census and other demographic data, existing and planned land uses, street and transit facilities, movement of people and vehicles, spatial location of activities, etc. A more common inventory for transportation planning is the travel survey, which is designed to obtain a comprehensive understanding of the travel behavior of people. This data provides important information regarding trip making and how travel patterns change over time. This information is required to plan for the future travel needs of both established and developing areas. The approach for travel surveys depends on the scope and quality of information required: nationwide travel surveys can be done to track long-term developments of trends in a country or it can be done to collect data about actual trips being made in a specific area. Travel survey data can be collected via different means: face to face interviews, telephone interviews, and vehicle registration number. However, travel surveys suffer from relatively small samples and require a huge investment of time and money. Coordinating travel surveys in regular bases is laborious and expensive.

2.3. Urban and transportation datasets

With the capabilities of advanced urban and transportation data collection methods, technically, we are now at a point where we can understand the majority of the urban movements. In this section, we discuss the urban and transportation datasets used in the analyses of the thesis. We obtained different datasets that show the existing mobility, activity and business data at Lisbon, our case study city.

2.3.1. Census data

The Instituto Nacional de Estatistica (INE) provides census that mainly address demographic, economic, social, and housing information. The data was based on Portuguese census collected during 2011. The counting was made on people whether they are present or absent in a given residence at the time of observation, but live at a given residence for a continuous period of at least 12 months before or after the census. Thus, census data provide information regarding the spatial distribution of people when they are at home (at night). The highest data resolution that can be found is at the sub-section level. The size depends on the number of people living in an area, where the sub-section encompasses on average 300 dwellings. Below is the hierarchy of census data is: Country→District→Municipality→Freguesia→ Section→sub-section.

Different types of statistics are computed to observe the distribution of residents, buildings, etc. Figure 2.2 shows the population density (number of residents per hectare) in Lisbon, which is a convenient way of showing how the population is distributed within the city. The population density is based on people who reside in 3623 sub-sections. It is computed as the number of residents divided by the area of each sub-section. Some build up areas, even though they are consumed by urban activities, Figure 2.2 (A), they do not have population density because no one is residing in those areas. Hence it is important to note that the density maps represent the intensity of the residents around midnight, not during the day. A specific characteristic of this map is that it indicates most of the starting points of people´s commuting daily trips.

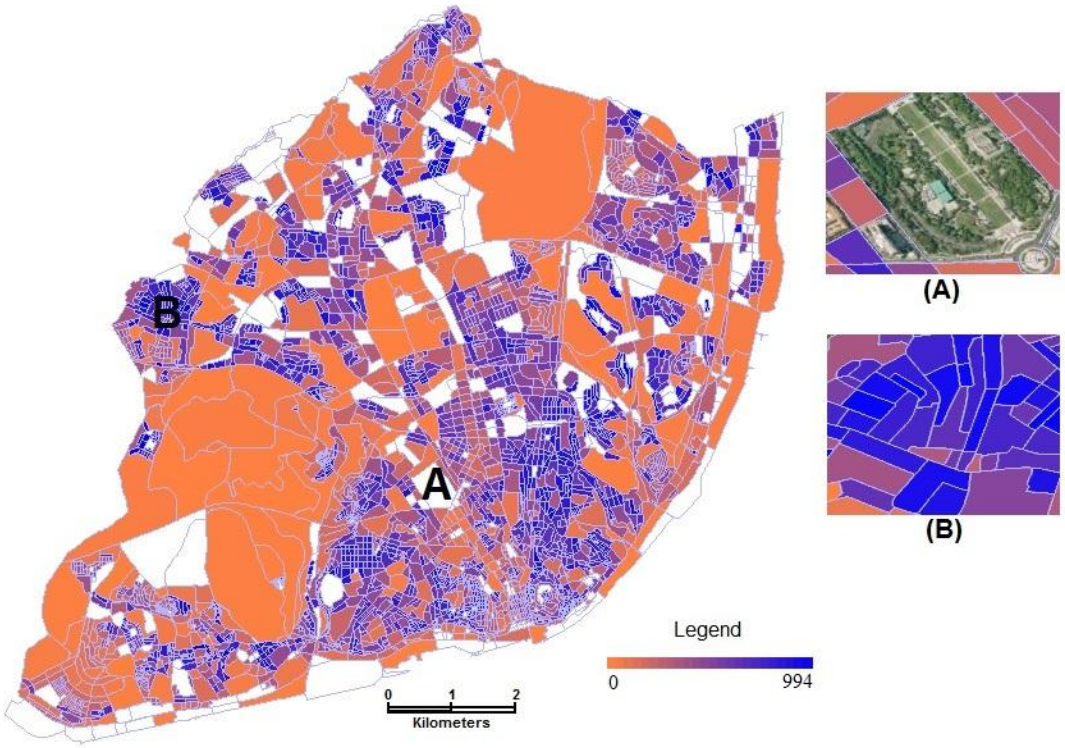


Figure 2.2 Population density of Lisbon in 2011

Figure 2.3 shows the density of exclusively residential buildings (number of exclusively residential buildings per hectare) at the granularity of sub-section. This is computed as the number of exclusively residential buildings divided by the area of each sub-section. The blank sub-sections are showing the absence of exclusively residential buildings in the area. In addition, Figure 2.3 shows a higher concentration of exclusively residential areas in the outskirts of the city.

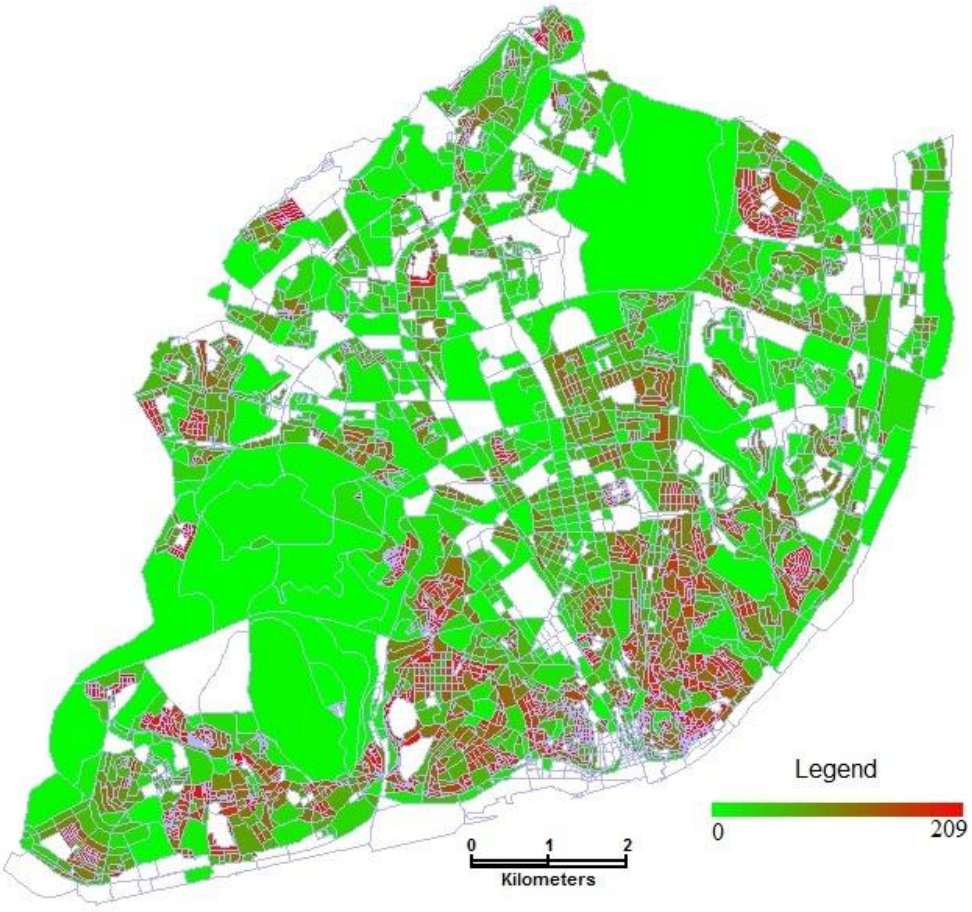


Figure 2.3 Density of exclusively residential buildings of Lisbon in 2011

2.3.2. Traffic counts

Traffic counts are performed to determine the number, movement, and classification of vehicles. The data were acquired from the municipality of Lisbon and from the website of Estradas de Portugal. Estradas de Portugal deploys a set of equipment and applications that collect and distribute real-time traffic data with the aim of reinforcing safety and management of the national road network. It installed more than 300 automatic traffic count recorders on the national road networks. There are different types of traffic data made available (e.g. traffic flow classified by vehicle category, average speed, vehicle weight) and it can be accessed for different time intervals (annually, monthly, daily, hourly, each 15min, 5min, and 1min). Hourly aggregated traffic count data was obtained from the Eixo Norte-Sul main itinerary road in the second week of April, 2010. The municipality of Lisbon monitors traffic volumes at several locations within its jurisdiction.

The traffic information is used to support planning, design, construction and operation of the municipality's road network. Hourly traffic volumes were obtained from 101 traffic counters equipped with inductive loops. The traffic volumes were obtained from a working day in the second week of April, 2010.

Traffic volume data for Lisbon was also obtained from a traffic assignment conducted using the software VISUM with a model of Lisbon's network and an OD matrix of the metropolitan area estimated through a home-based survey and traffic counting in 2008. More about this dataset is found at Correia and Viegas (2011).

Figure 2.4 shows the geocoded points of the traffic counters available in the city, which are represented by circles.

Figure 2.4 The geographical locations of traffic counters in the municipality of Lisbon

Figure 2.5 shows the average hourly traffic volume in Lisbon along the hours of a day. The traffic counts are taken from all the counters from Figure 2.4. The average hourly traffic volume shows the general patterns of traffic in the city. The first peak starts around

7 AM in the morning and stays until 10 AM followed by the valley between 11 AM to 14. The second peak occurred in the afternoon between 16 to 19 hr and a decrease in the average traffic from 19 hr on.

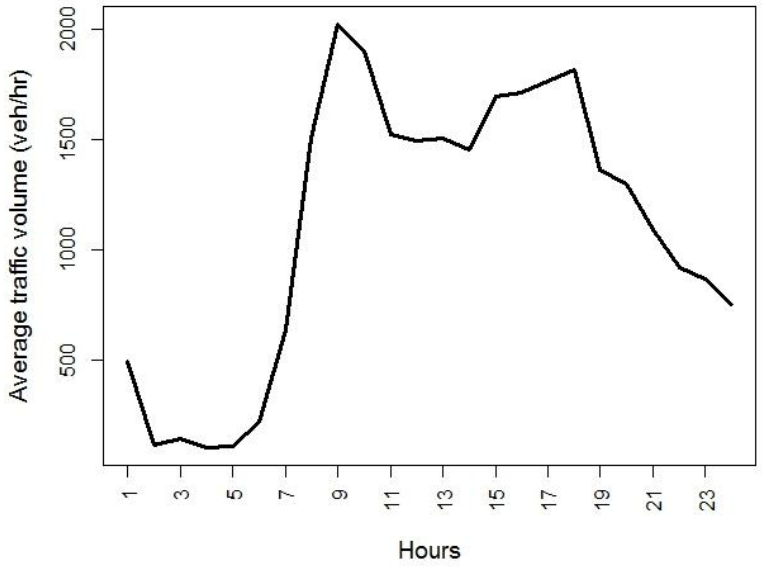


Figure 2.5 Average traffic volumes in Lisbon

2.3.3. Bus datasets

Bus related datasets were obtained from Companhia Carris de Ferro de Lisboa (CARRIS), which is a public transportation company in Lisbon that operates buses, trams, and funiculars. CARRIS has day and night time services. In 2010, the regular daytime passenger transport had 93 lines for the urban and suburban routes in Lisbon Metropolitan Area and it was assured by a fleet of 245 vehicles, such as bus, tram and elevators. This comprises about 78 bus lines served by 745 buses for a route length of 667km. Two datasets are acquired: (1) bus arrivals at each bus stop aggregated at 15 minutes interval, and (2) bus stop information.

Table 2-1 shows sample data about the bus stops regarding their name, ID, and geographical reference in terms of latitude and longitude.

Table 2-1 Sample data: Bus stop

Bus stop ID	Bus stop name	Latitude	Longitude
102	ESCOLA DONA LEONOR	38:45:7.506	-9:8:37.021
103	PRAÇA DE ALVALADE	38:45:15.188	-9:8:39.919
151	SANTO AMARO	38:42:6.48	-9:10:51.24
152	RUA DA JUNQUEIRA	38:41:59.28	-9:11:3.48
156	PALÁCIO DE BELÉM	38:41:50.28	-9:12:1.08
161	RUA PINTO FERREIRA	38:41:52.8	-9:11:25.8

Figure 2.6 shows a sample of the geocoded points for 2280 bus stops in Lisbon, which are represented by squares.

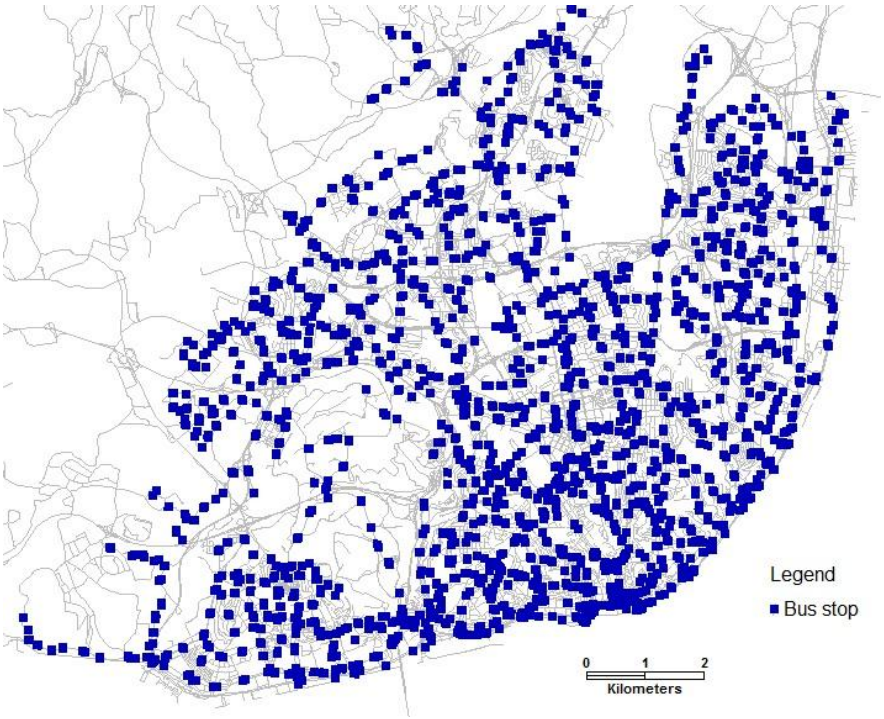


Figure 2.6 Location of Bus stops in Lisbon

Figure 2.7 shows the number of bus arrivals at each bus stop in the morning between 8 AM to 9 AM on April 12, 2010. Large circles (red colour) show a high number of bus arrivals and the small circles (green colour) show a low number of bus arrivals. In addition, Figure 2.7 shows that bus stops that are serving high movement are located adjacent to the main road network of the city.

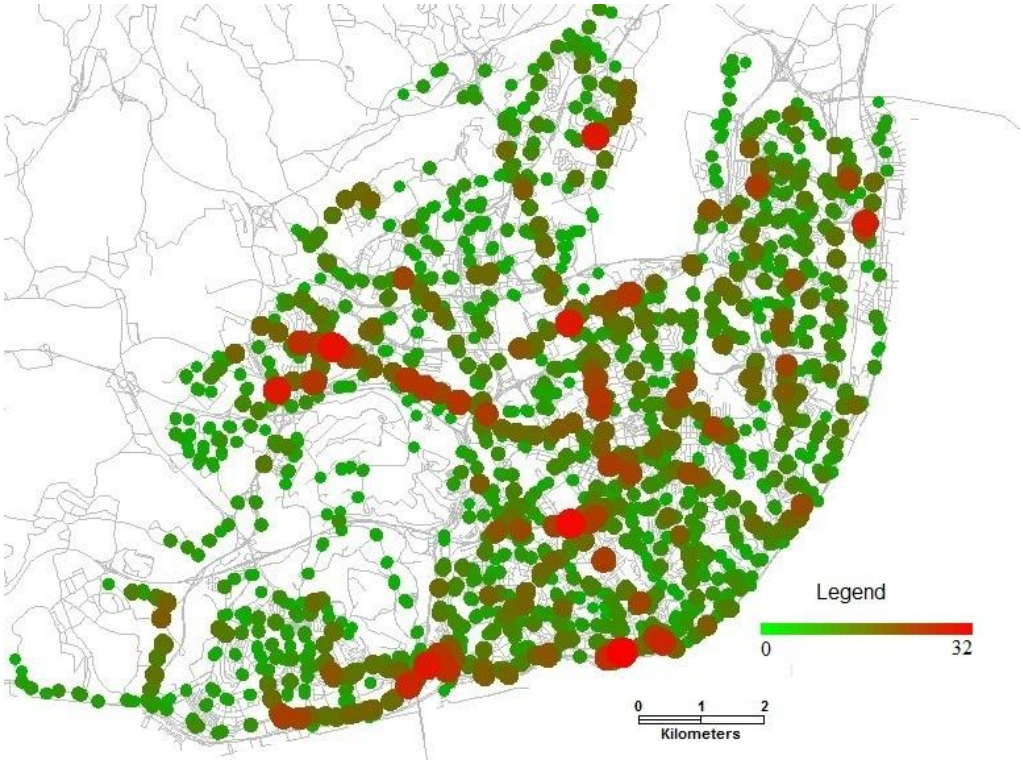


Figure 2.7 Bus movements between 8 AM and 9 AM in Lisbon April 12, 2010

2.3.4. Taxi data

Taxi data is acquired from GeoTaxi, a company that focuses on software development for fleet management, and holds around 20% share market in Portugal. The data were obtained from 230 taxis. It consists of the location (latitude, longitude), time, heading direction, and occupancy status of the taxi. The data were collected by in-vehicle equipment when pre-defined events occurred. There are different types of events: time intervals, distance interval, and sensor change. The meanings of distance interval and time interval are straightforward, and an example of sensor change is when a passenger is leaving or entering the vehicle. A sample of the taxi traces is shown in Table 2-2.

Table 2-2 Sample data: Taxi traces

Taxi ID	Date and Time	Taxi status	Latitude	Longitude
2575	02-11-2009 07:00:01	Free	37:7:33.6	-8:14:7.692
756	02-11-2009 07:00:01	Free	38:45:56.52	-9:18:1.512
83	02-11-2009 07:00:05	occupied	38:45:29.52	-9:5:47.76
2068	02-11-2009 07:00:03	occupied	38:42:27.36	-9:19:22.98
782	02-11-2009 07:00:05	Free	38:37:10.56	-9:6:43.992
118	02-11-2009 07:00:10	occupied	38:42:48.96	-9:8:20.004
754	02-11-2009 07:00:09	occupied	38:44:19.68	-9:7:32.808

Figure 2.8 shows the spatial distribution of taxi traces in Lisbon between 8 AM and 9 AM in the morning of November 2, 2009. Each taxi trace represents an observation obtained when a pre-defined event happened.

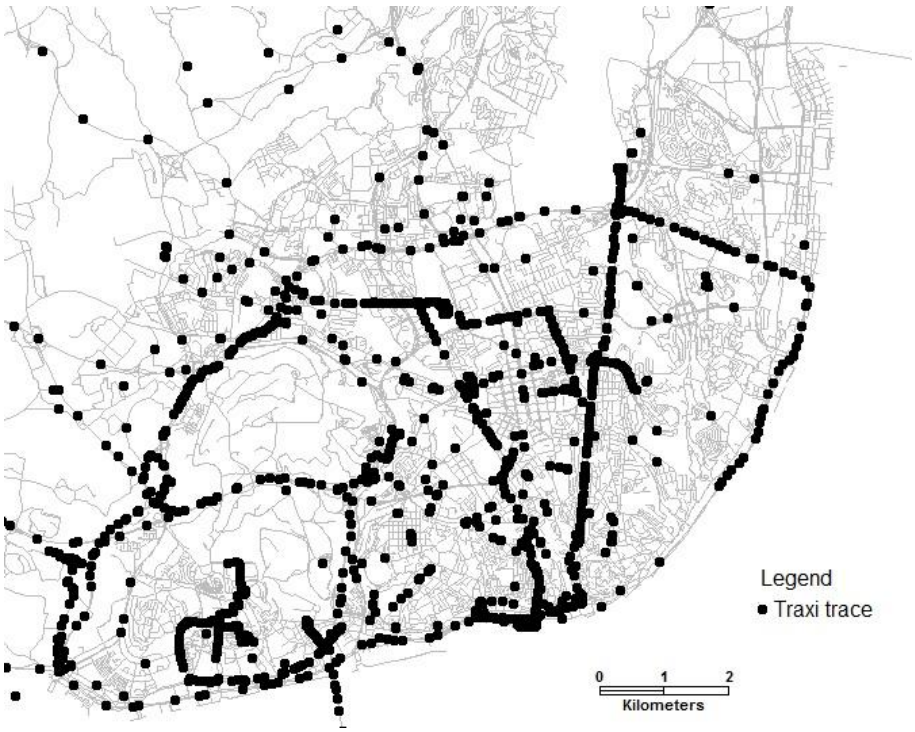


Figure 2.8 Spatial distribution of taxi traces between 8AM and 9AM in Lisbon November 2, 2009

The taxi volumes are computed based on the number of taxis that stopped in each area for the purpose of passenger pick-up and drop-off along the day. Thus, for a given location,

the taxi volume can be aggregated as taxi arrivals (traces during passenger drop-off), taxi departures (traces during passenger pick-up), and combination of taxi arrivals and departures (total taxi movements). Figure 2.9 shows the intensity of total taxi volume (taxi arrivals and taxi departures) in 800 meters by 800 meters grid-cell areas in Lisbon, where the intensity of taxi movement is represented by a color scale. The taxi traces were obtained on November 2, 2009.

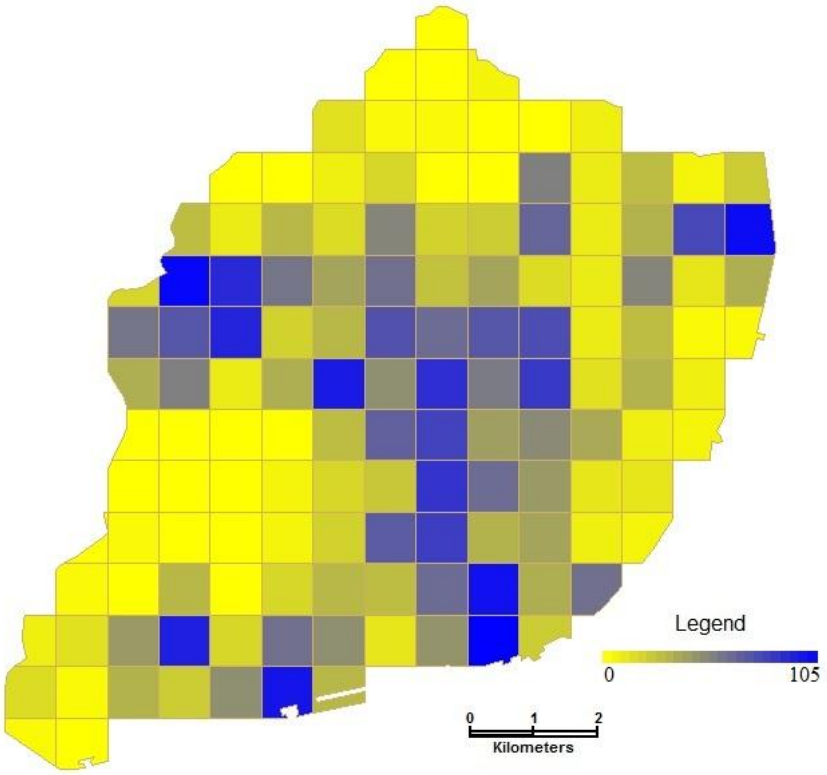


Figure 2.9 Taxi volumes in Lisbon (total daily taxi arrivals and departures)

2.3.5. Presence of people

Data about the presence of people was obtained in another study in the region (Martínez et al., 2009). This data was deduced from a mobility survey, on the origin and destination of trips and schedules. The data provides hourly aggregated number of people available in each area.

Figure 2.10 shows the geographical distribution of the people´s presence in Lisbon between 7 AM to 8 AM in the morning. The presence of people data is aggregated to an

800 meters by 800 meters grid-cell. For easy analysis and graphical representation, the data is normalized over space. This normalization gives the intensity of presence of people at each grid-cell relative to the total number of people in Lisbon at a certain hour. The Figure shows that between 7 AM to 8 AM, there is a higher intensity of people´s presence both in the city center and in the north-west part of the city which is predominantly residential area. The data on geographical distribution of presence of people for the remaining hours of the day is presented in Appendix A.1.

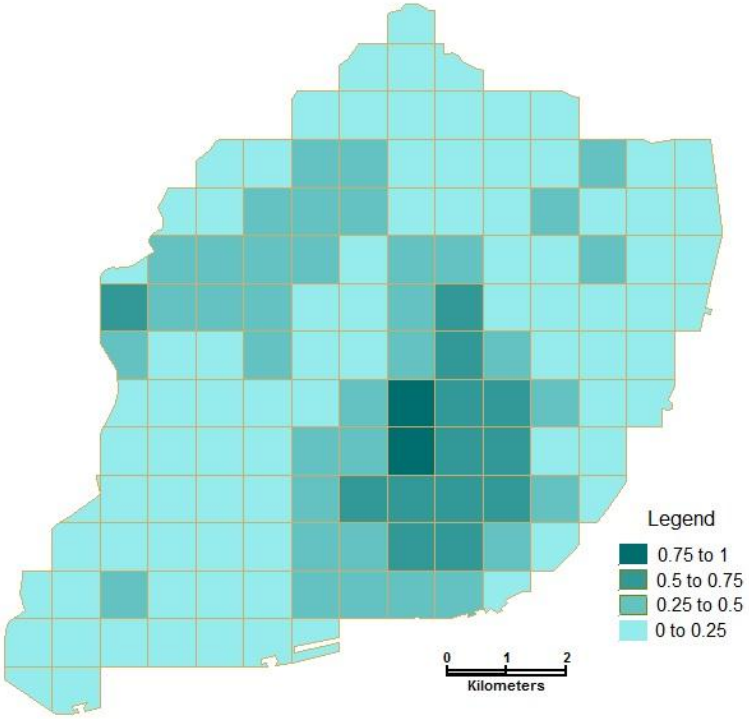


Figure 2.10 Distribution of people in Lisbon between 7 AM to 8 AM

2.3.6. Points of Interest

A Point of Interest (POI) is a feature on maps, navigation systems or route planners, which represents a particular activity on that place. The majority of POIs are relatively permanent structures, such as statues, buildings. Some other POIs have variable degree of permanence such as bars, restaurants, which may open at a day and close some months later, and others are temporal or periodic, such as the location of an annual festival. POIs data were acquired from Servidor de Apontadores Portugueses (SAPO). The dataset has 5471 points located within the Municipality of Lisbon. Table 2-3 shows a sample of the dataset, which

includes the name of the location, address, facility type and location of each POIs given by longitude and latitude.

Table 2-3 Sample data: Points of interest

Name	Address	Latitude	Longitude	Category
Hospital Egas Moniz	Rua da Junqueira 126	38:41:54.888	-9:11:16.951	Health Facility
Club	Largo de Santos	38:42:26.906	-9:9:15.707	Recreation
Fnac	Rua do Carmo no. 2	38:42:39.384	-9:8:22.222	Shopping
Estação de Metro da Avenida	Avenida da Liberdade	38:43:12.047	-9:8:42.572	Transport Facility
Millennium BCP	Rua Castilho 44B	38:43:24.42	-9:9:7.297	Service
Escola Alemã de Lisboa	Rua Filipe Duarte	38:45:28.548	-9:9:54.442	Education

Figure 2.11 shows the distribution of POIs in Lisbon, which includes different business categories (service, recreation, education, shopping, health, and transportation facilities). The figure displays that the majority of service and recreational areas are found in the downtown area, whereas education, health and shopping facilities are fairly distributed throughout the city.

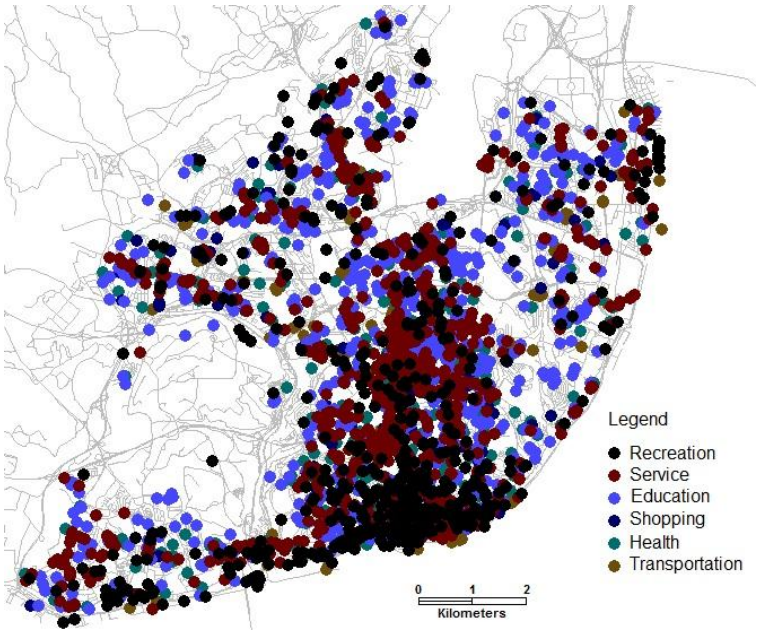


Figure 2.11 Distributions of POIs in Lisbon

2.3.7. Cellular network

A cellular network is a wireless network with service coverage divided into many smaller areas, known as cells, each of which served by a base station. Most cells are divided into three cell sectors, which overlap with cell sectors of the neighboring cells to deliver continuous network. Thus, when people move between cells, the signal is managed by a new base station without service interruption. The size and shape of each cell is determined based on the geographies of the surrounding area, such as buildings, mountains and trees, which can interfere with the signals.

The frequency spectrum allocated for cellular communication is scarce, thus the cellular network adopts a frequency reuse concept. This is one of the reasons why the coverage area is divided into many cells. To avoid interference and provide a guaranteed bandwidth within each cell, each cell uses a different set of frequencies from neighboring cells. However, the same frequency can be assigned to two cells that are well separated such that the radio co-channel interference between them is within a tolerable limit.

The system architectures for the Global System for Mobile communications (GSM) and the Universal Mobile Telecommunications System (UMTS) are relatively similar in their concept. Figure 2.12 illustrates the system architectures for both networks. A GSM Base Transceiver Station (BTS), also called base station, holds an antenna (or several antennas) to receive and send radio signals. Base stations transmit and receive radio signals to connect mobile devices to the network. The Base Station Controller (BSC) manages the received and transmitted radio signals of the connected base stations. In the case of UMTS, the base stations are called Node B and the device that controls these cells is named as Radio Network Controller (RNC). The interface between the BTS and BSC in GSM is designated A_{bis}, and in the case of UMTS, the interface is designated as I_{ub}. The interface between the GSM radio parts and the core network side is denoted by A and the corresponding interface for UMTS is denoted I_u.

Figure 2.12 also shows the Mobile Switching Center (MSC) and Serving GPRS Support Node (SGSN), connected to a simplified "network". Communications through the circuit switched network uses the MSC. The MSC is mostly associated with communication switching functions, such as call set up, release, and routing. It also carries out a host of other duties, such as routing SMS messages, conference calls, fax, etc. SGSN works to maintain a cellphone user´s connection to the internet and packet-based mobile

applications (Gundlegård and Karlsson, 2009). To accommodate a moving cellphone user, it is important for the MSC to determine the position of each cellphone to provide routing communications between them. To perform this, the MSC consults a large database named Home Location Register (HLR), which stores relevant information about each cellphone. Some of the information stored by the HLR includes if the user is active or not, and if active in which part of the network the user is located at the moment (Gundlegård and Karlsson, 2009).

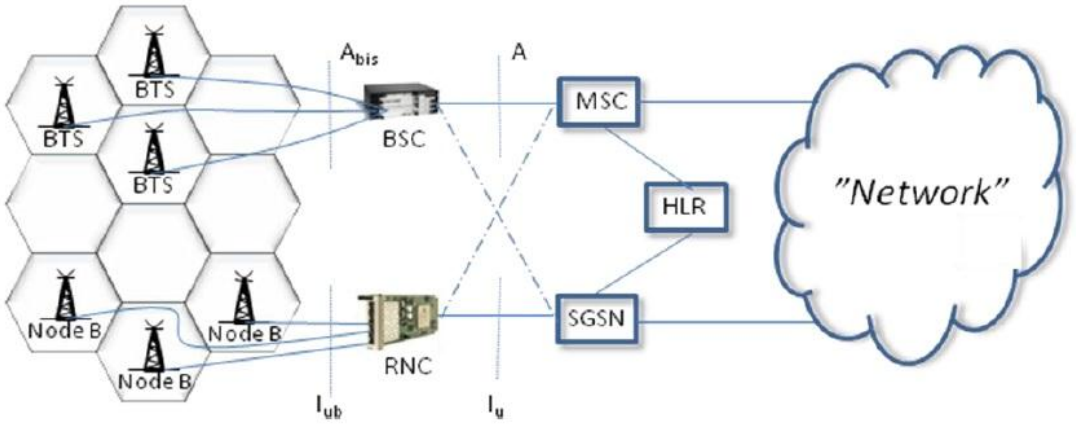


Figure 2.12 Cellular network architecture for the GSM (upper part) and UMTS (lower part) (Gundlegård and Karlsson, 2009)

2.3.7.1. Cellular positioning

Cellular positioning is a course of action to determine the spatial location of a cellphone. The main purpose of positioning is the measurement of different observables such as angles, time, time differences, or velocity. These observables reflect the spatial relation of a cellphone relative to a well-known point. Cellphone´s position is acquired through different methods that depend on the types of observables used. Examples of positioning methods include proximity sensing, lateration, and angulation (Küpper, 2007).

Cellular networks have always been capable of determining the generic locations of cellphones through the serving cell connection. However, in 1995 the United States federal communications commission launched an emergency initiative called enhanced 911 (e911), which was an important driving force for advanced positioning methods and their installation in cellular networks. This initiative imposed all the cellular network operators

to provide services to locate emergency 911 calls. Based on the initiative, mobile operators must be able to locate their users within 50 meters, 67% of the time, and 150 meters 95% of the time (Spinney, 2003). From July 2003, the European Union has also formulated a similar program to improve 112 emergency services. The program defines a set of minimum requirements for fixed and mobile network operators to locate people using 112 emergency lines in the best way possible (Ratti et al., 2006).

The imposed accuracy because of the e911 mandate was not achievable through cell-ID positioning. This created a chance for the invention of alternative cellphone positioning methods. These positioning methods can be grouped into two categories which are commonly named as network-centric positioning and device-centric positioning (Spinney, 2003).

2.3.7.1.1. Network-centric cellphone positioning

The network-centric cellphone positioning technique utilizes existing cellular network capabilities to determine cellphone positions. Enhanced network-centric cellphone positioning techniques include lateration and angulation based methods where the position of the cellphone is computed based on time, angle, and/or distance measurements. The following positioning methods fall into this category:

Cellular identification (Cell-ID) and Sectorized Cell-ID: These positioning methods are based on proximity sensing. In this method, location is expressed in terms of an area. The coordinates of the serving base station are then associated with the cellphone position. The location accuracy depends on the size and density of cells, where systems with smaller cells allow a higher precision (Küpper, 2007).

Cell-ID combined with Time advance: This method is used exclusively in GSM network. It is built on the concept of Cell-ID positioning and it includes additional dynamic variable, which is time. The method uses the known position of base stations in terms of accurate latitude/longitude coordinates. The cellular network operators have a database containing this data. The location of the base station is then coupled with distance estimated from the time and velocity needed for the radio signal to reach a cellphone. The limitation associated to this method is that it determines linear, single dimentional positions from the

base station, which only allows identifying a ring of potential positions of the cellphone with the serving base station in its center. Thus, the cellphone may be located anywhere along a signal radius which causes positional uncerainity (Spinney, 2003). However, in the case of sectorised antennas, the potential positions of the cellphone can be more specified to one-third or one-forth of a ring depends on the number of cell sectors (Küpper, 2007).

Time of arrival (Circular lateration): Time of arrival is another network based positioning technique. Unlike the Cell-ID and the time advance, which are single serving-cell position determination techniques, three base stations are employed to compute the cellphone position (lateration for three base stations is also known as trilateration). Each of the three distance estimated from the time needed for the radio signal to reach each base station is calculated. The intersection of the three distances is then used to interpolate the cellphone approximate position. This positioning technique addresses the azimuthal determination limitations of the time advance by employing multiple base stations to approximate cellphone positions. However, its positional accuracies depend on geographical distribution of base stations, signal strength, topography and weather conditions (Spinney, 2003). Another form of lateration is hyperbolic lateration (time difference of arrival). Unlike circular lateration that determines time; hyperbolic lateration is based on time differences. For detailed information on that, see (Küpper, 2007).

Angle of arrival: This is another method to estimate a cellphone position for the known coordinates of several bases stations. Unlike time of arrival, the observables here are the angles between the cellphone and a number of base stations. Angle of arrival is also known as angulation or direction of arrival. In order to obtain these angles, this positioning technique requires either the base stations or the cellphone to be equipped with smart antenna arrays. The smart antenna arrays on a single base station can specify the angle at which the cellphone is transmitting signal to one of the lobes in the antenna array. This allows determining the position of the cellphone through intersection of known signal angles from at least two base stations. This technique offers accuracies between 150 and 50 meters, but small angular errors could cause significant positioning inaccuracies for cellphones situated far from the base station (Ratti et al. 2006).

2.3.7.1.2. Device-centric cellphone positioning

Device-centric cellphone positioning technique differs from network-centric technique because the cellphone device performs its own position computations and it gives relatively higher accuracy. The limitation associated to this category is that additional hardware and software is required to each cellphone devices.

Server-Assisted GPS (A-GPS): A-GPS is device-centric cellphone positioning technique that uses both GPS and a terrestrial cellular network to obtain geographic position. This operation assists the functionality of the cellphone device by directing where the appropriate satellites are and allows the network to perform the majority of the calculation role that would otherwise be carried out by the cellphone device (ACA, 2004). A-GPS technique is capable of positioning cellphone devices within 3 meters accuracy in open-air environment and within 20 meters accuracy during dense urban environment (Spinney, 2003).

Enhanced-observed time difference: This is another device-centric positioning technique which is available in two configurations: circular and hyperbolic lateration. This requires the observation of the time of arrival or the time difference of arrival of pilot signals (radio signals) emitted from a number of base stations which are located adjacent to the cellphone. Moreover, for circular lateration, the cellphone has to synchronize with the base stations, while for hyperbolic lateration the base stations must be synchronized among each other. However, these are not originally fulfilled in GSM network, thus, GSM should integrate location measurement unit to measure time offsets and obtaining an a posteriori synchronization (Küpper, 2007). This method relies on good visibility of at least three base stations to compute the cellphone position. Thus, there will be difficulties in rural areas because of large distances between base stations (Ludden et al., 2002).

2.3.7.2. Location management

A cellular communication system has to track its users, because the exact location of the users must be known to the cellular network to direct calls to the relevant cell within a network. In a cellular network, keeping the track of users is the subject of location management, and it also tracks an active mobile device (cellphone), which is not in a call

(a mobile device is active as long as its power is on). There are two naïve options to perform this operation (Zhang and Stojmenovic, 2005): one is "never-update" scheme, where the mobile device does not tell its location when it moves around. This leaves the cellular network to page all cells in the service area to find out the cell in which the mobile device is currently located to forward an incoming call to the base station of that cell. However, it is an expensive procedure to page all cells in the service area. The second is "always-update" scheme. This involves a mobile device informing the network of its new location whenever it moves into a new cell. When an incoming call arrives, the network can just direct the call to the last cell reported by the mobile device. Obviously, there is no paging cost, but the successive location update of the mobile device when it moves form cell to cell would quickly overwhelm the network, and it can also be expensive.

Owing to these reasons, in current cellular networks, a combined approach of location update and paging is used, which is based on the concept of Location Area (LA). Cells within a network are grouped to form a mega-cell, which is called Location Area. Figure 2.13 (a) shows a coverage area of three location areas (LA1, LA2, and LA3) separated by bold wide lines. Thus, when a mobile device moves from Cell-B to Cell-C it updates its new location because the two cells are in different location areas. However, no location update is required if a mobile device moves from Cell-B to Cell-A, because the two cells are in the same LA. Thus, if an ongoing call is to be forwarded to a user, the network must only page cells in the LA to determine its precise cell.

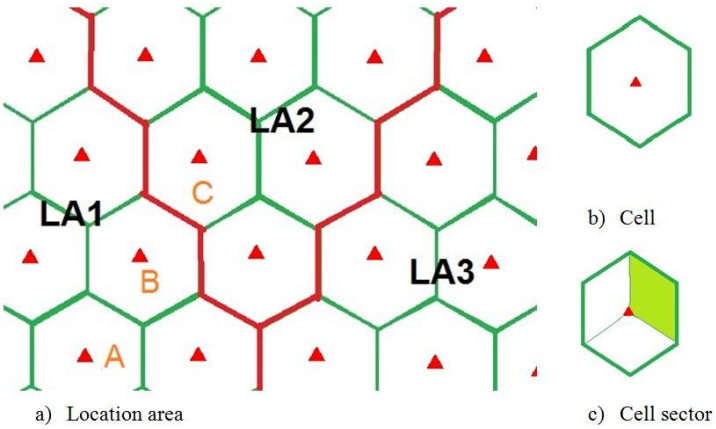


Figure 2.13 Typical cellular networks

2.3.7.3. Characteristics of cellular network data

Cellular networks produce massive amounts of data through the interaction of cellphone and serving telecommunication, which can tell us about the presence and movement of people in a given area. The resultant datasets are feed by three main events (Calabres, 2011b): (1) activity related events that include voice call, short message services (SMS), and internet usage; (2) periodic location update, which is generated on a periodic base and provides information on which cell tower the phone is connected to regardless of whether a new location area has been entered; and (3) mobility related events: this type of event is executed whenever subscriber enters a new location area or a new cell.

2.3.7.4. Portuguese cellphone operator

Portugal has a high number of cellphone users. According to statistics from ANACOM (2010), active mobile telephone cards per 100 Portuguese inhabitants grew to 159.9 by the end of year 2010, from 140.4 in the year 2008. This mobile penetration was obtained through the combined services provided by the three major mobile service operators: Vodafone, Optimus and TMN, each having their own network infrastructure.

The data available for this thesis was obtained from TMN operator. In December 31, 2010, TMN had 7.42 million subscribers in Portugal (this includes Mobile Virtual Network Operator (MVNO) supported by TMN´s network), which accounts for 45% of the total number of mobile subscribers. TMN uses Global System for Mobile communications (GSM) and Universal Mobile Telecommunications System (UMTS) technologies to provide its services. TMN's UMTS population coverage was approximately 93%, by 2010 and it was available over 4194 municipalities out of a total of 4252 in Portugal (ANACOM, 2010).

In the area of Lisbon and neighbor municipalities, TMN provides cellular network service by way of 487 base stations comprising 1669 cell sectors. The majority of the analysis throughout the thesis is based on the cellular network data that belongs to the Municipality of Lisbon, which accommodates 364 base stations comprising 1200 cell sectors. Table 2-4 shows sample data acquired by TMN regarding the cell sector information.

Table 2-4 Sample of data on cell sectors location

Cell sector ID	Cell sector name	Latitude	Longitude
10005	DAM1D	38:44:39,7	-9:13:1,44
10006	DAM2D	38:44:39,7	-9:13:1,44
10007	DAM3D	38:44:39,7	-9:13:1,44
1100	ALHC1D	38:55:54,51	-9:6:50,2
1101	ALHC2D	38:55:54,51	-9:6:50,2
11011	POAC1	38:50:25,72	-9:57:56
11012	POAC2	38:50:25,72	-9:57:56

Figure 2.14 shows the geographical distribution of base station positions in Lisbon. There is a higher concentration of base stations in the city center when compared to the outskirts of the city. However, each base station houses a varying number of cell sectors. Figure 2.15 shows the total number of cell sectors at each tower locations. The big circles show a higher number of cell sectors and smaller circles show a lower number of cell sectors available at each tower location.

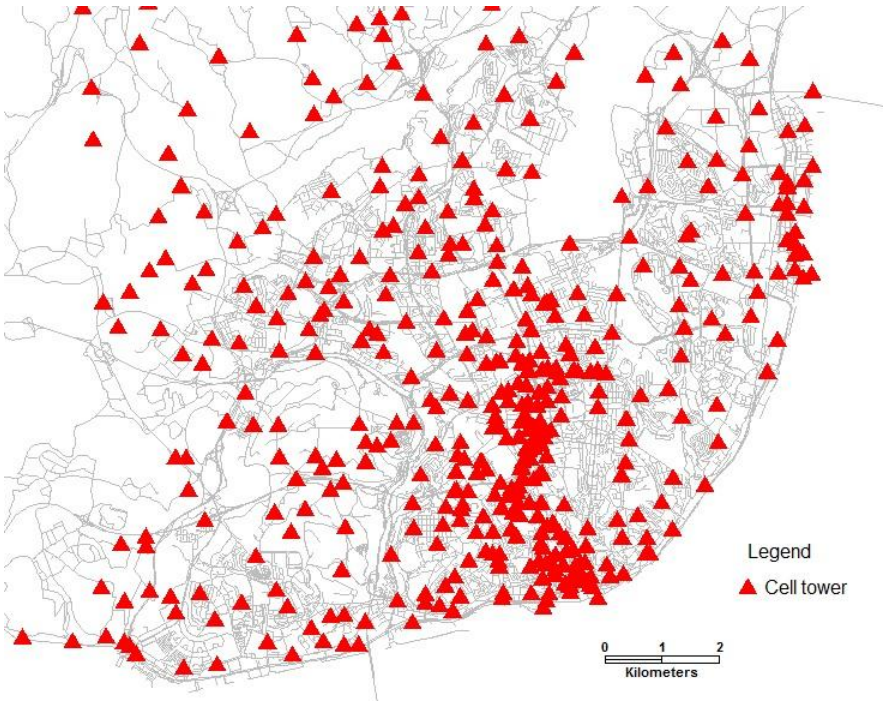


Figure 2.14 Location of TMN´s base stations in Lisbon

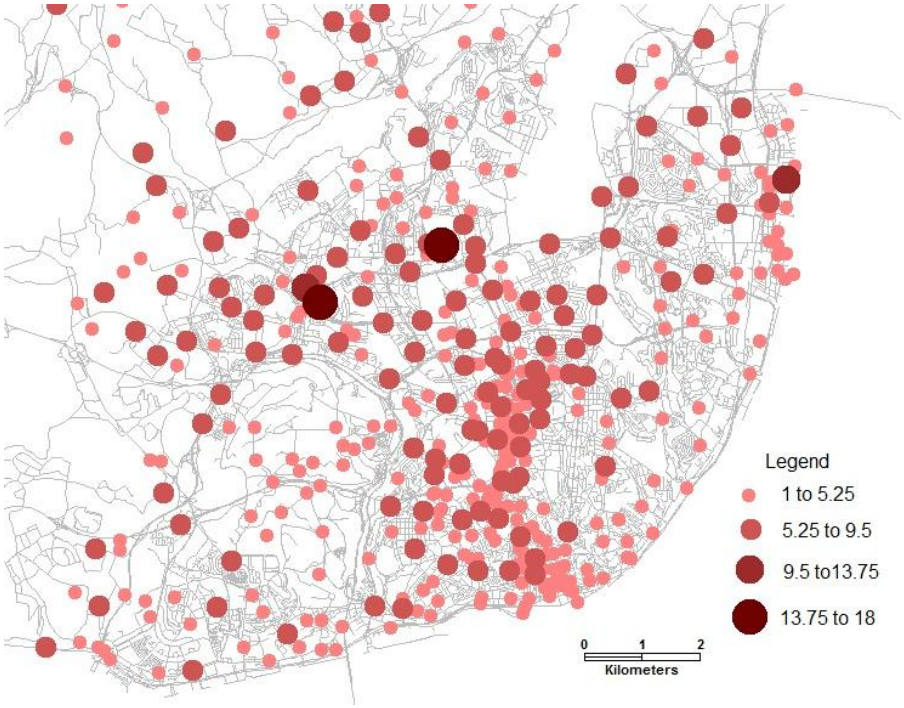


Figure 2.15 Cell sector densities

2.3.7.5. TMN´s cellular network data

Passive mobile positioning datasets are used to perform the analyses throughout this thesis. Passive mobile positioning data is automatically stored in the log files of the service providers. The usual source for passive mobile positioning is "a billing log", which is recorded for call commercial activities (Ahas et al., 2010). Other sources of data on call activity are also used for passive positioning, such as the Erlang value for a cell (Reades et al., 2007). TMN provides three passive mobile positioning datasets that are hourly aggregated at the granularity of a cell sector: Call Volumes, Erlang, and Handover values. The datasets were obtained on the second and third weeks of April, 2010. For easy analysis and graphical presentation, the cellphone datasets are normalized over space. This normalization gives the intensity of cellphone usage at each tower location relative to the total cellphone usage in the entire region of study at a certain hour. In the next paragraphs, we present a brief explanation on each dataset:

Call Volume: The Call Volume measures the phone activity in terms of the number of calls that happen within a given area and time window. Figure 2.16 shows the volume of

calls at each tower area in the morning between 7AM to 8AM on April 12, 2010. The higher Call Volumes are shown by bigger and darker circles, and lower call volumes are indicated by smaller and lighter circles. Intensity of Call Volume data for the remaining hours of the day is presented in Appendix A.2. Table 2-5 shows sample datasets of the Call Volume and Erlang. In each row, hourly aggregated value of Call Volume and Erlang values are given associated to their respective cell sector ID and cell sector name.

Table 2-5 Sample data: Call Volume and Erlang

Cell sector ID	Cell sector name	Date	Hour	Call Volume	Erlang
1110506612	RT_ALCANTAR_1-06612	04-06-2010	3 PM	887	13.94
1110506612	RT_ALCANTAR_1-06612	04-06-2010	4 PM	1067	16.20
1110506612	RT_ALCANTAR_1-06612	04-06-2010	5 PM	1401	15.94
1110506612	RT_ALCANTAR_1-06612	04-06-2010	6 PM	1482	18.55
1110506612	RT_ALCANTAR_1-06612	04-06-2010	7 PM	1151	13.44
1110506612	RT_ALCANTAR_1-06612	04-06-2010	8 PM	1140	14.62

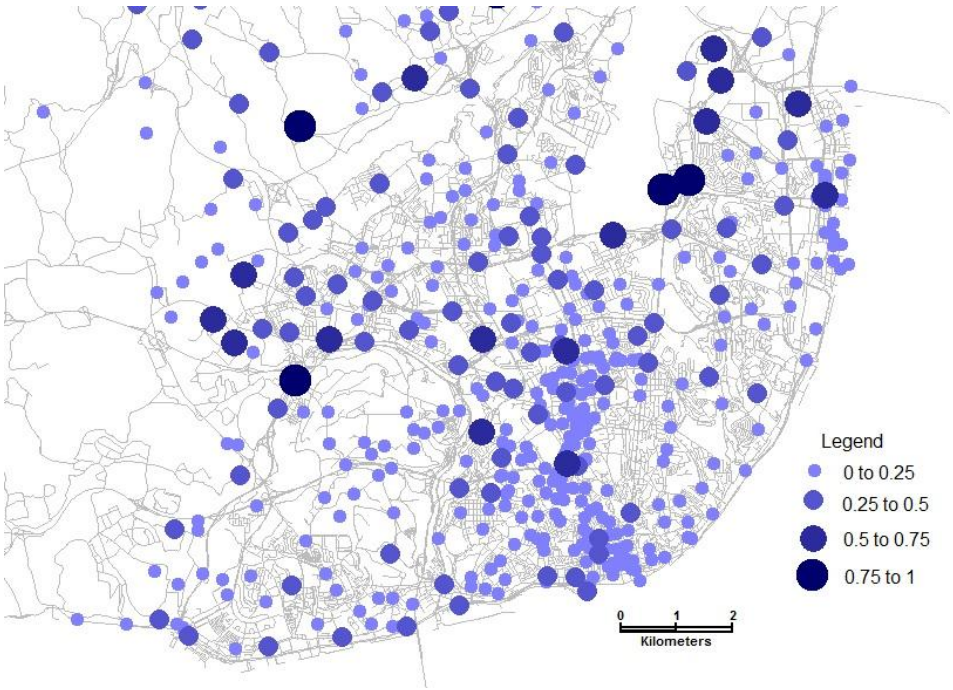


Figure 2.16 Volume of calls at each cellular tower locations in Lisbon between 7AM and 8 AM on April 12, 2010

Erlang: The Erlang measures the bandwidth usage, which is usually applied for network capacity planning on GSM networks and can be easily collected by the operator. Hourly aggregated Erlang values are obtained. An Erlang is defined as one person-hour of phone usage: alternatively, one Erlang can be obtained through two people talking for half hour each, sixty people talking for one minute each and so on. Consequently, Erlang data provides opportunities to demonstrate spatial as well as temporal dynamics as seen through network bandwidth consumption (Reades et al., 2007). Figure 2.17 shows the Erlang values at each tower in the morning between 7AM and 8AM on April 12, 2010. Large circles show a higher Erlang value and smaller circles show a lower Erlang value. Intensity of Erlang values for the remaining hours of the day are presented in Appendix A.3.

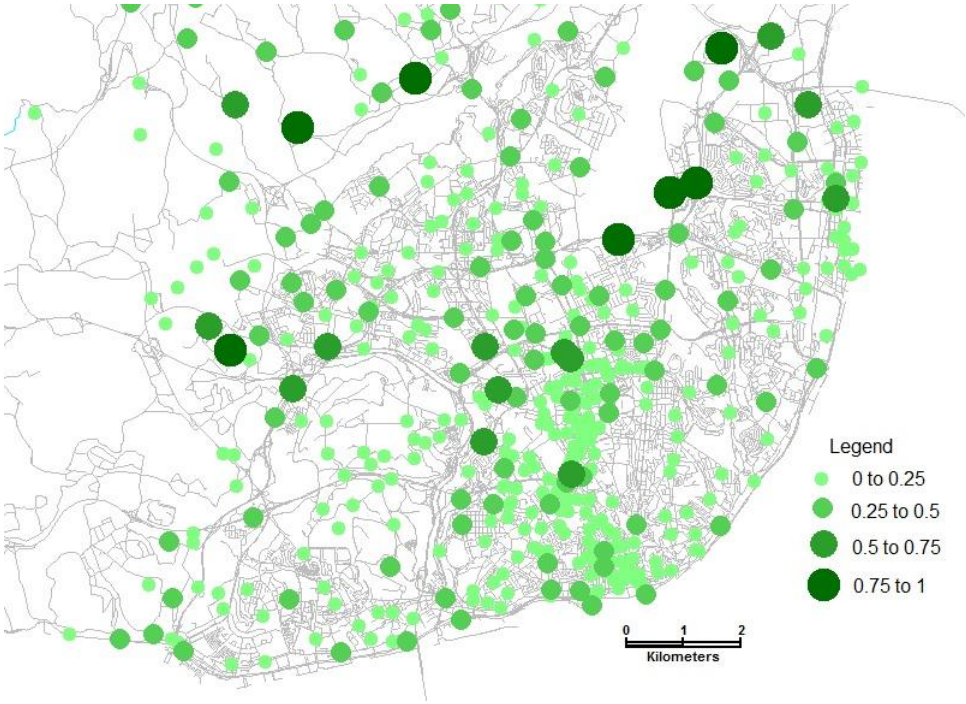


Figure 2.17 Erlang values at each cellular tower locations in Lisbon between 7AM and 8 AM on April 12, 2010

Handover: One of the major features of cellular network is its ability to provide data on the movement of a person when making a call. A mobile handset can move out of one cell to another one while in active communication (i.e., call or data session) or while it is in idle state. Handover (also called handoff) is the process of transferring an ongoing call or data session from one base station to another without loss or interruption of service. Table 2-6

shows a sample of the handover data obtained from TMN, where calls are transferred from the source cell sector to a neighbor cell sector.

Table 2-6 Sample handover data

Source cell sector ID	Neighbor cell sector ID	Date	Hour	Handover from Source to neighbor cell sector
1110506612	1110509355	04-06-2010	3PM	92
1110506612	1110509355	04-06-2010	4PM	133
1110506612	1110509355	04-06-2010	5PM	120
1110506612	1110509355	04-06-2010	6PM	134
1110506612	1110509355	04-06-2010	7PM	109
1110506612	1110509355	04-06-2010	8PM	109

Using a GIS platform, the handover events associated to the entire cell sectors sited at the same location were summed up and grouped into incoming and outgoing handovers. All calls ended but not originated at cell sectors in a given base station (cell tower) were summed up and named as incoming handover. All calls originated but not ended at cell sectors in a given cell tower were summed up and named as outgoing handovers. The occurrence of a handover event between cells within the same tower is more frequent because of the shorter distance required to cross the cell boundaries. This kind of handover event is filtered with the objective of reducing the noise from pedestrians' calls in two scenarios: during an exploratory analysis that seeks for relationships between traffic and handover counts (Chapter 3); and in a case of a predictive data analysis with the aim of predicting traffic levels through handover count (Chapter 5). The handover map (Figure 2.18) shows the flow of calls between cellular towers. The triangles sow the location of the cell towers, and the trajectories are showing the availability of a handover connection.

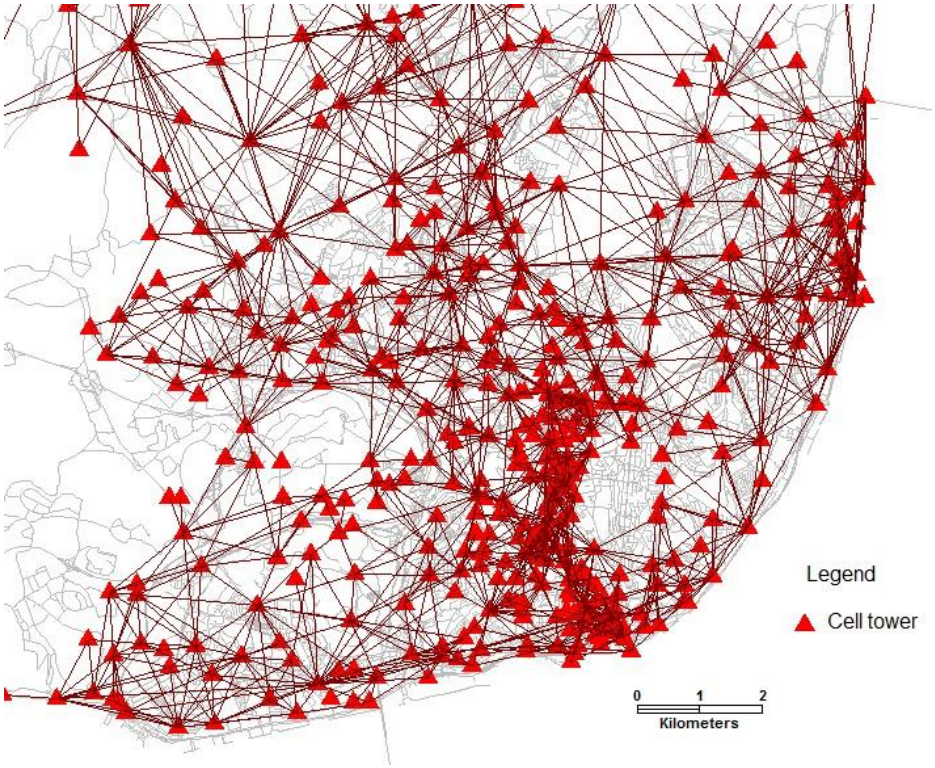


Figure 2.18 Handover map

Figure 2.19 shows the total number of outgoing and incoming handovers at each tower between 7AM and 8AM on April 12, 2010. The size of the circles is proportional to the intensity of the total handover. Large circles mean that there is a higher intensity of handover. Intensity of Handover data for the remaining hours of the day is presented in Appendix A.4.

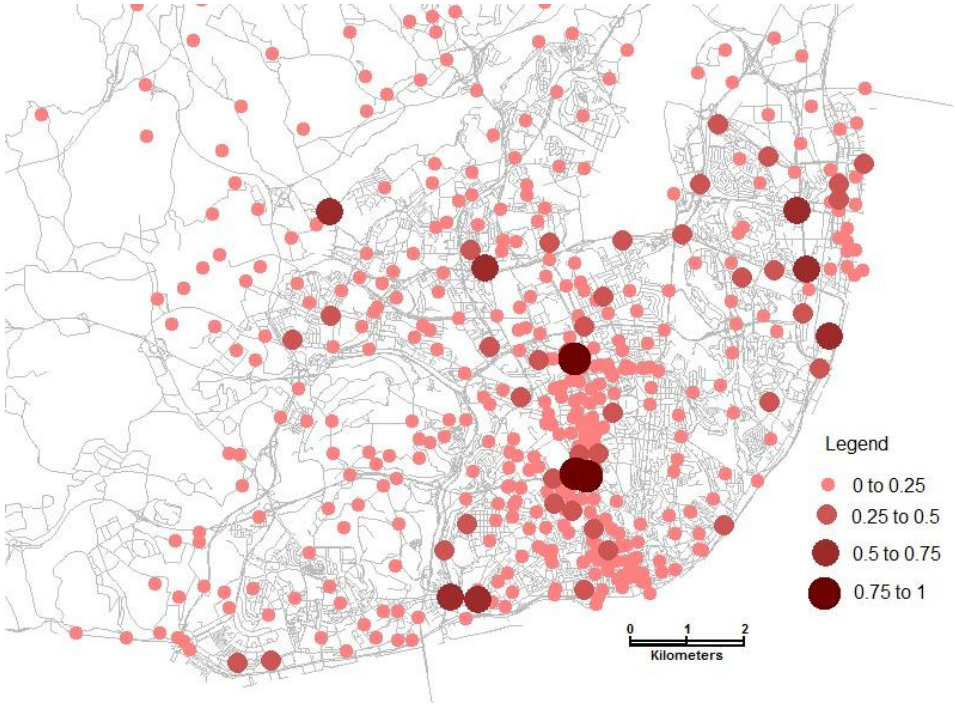


Figure 2.19 Total handover (incoming and outgoing) between 7AM and 8AM on April 12, 2010

Figure 2.20 shows normalized total cellphone usage in Lisbon on April 12, 2010. It illustrates total Call Volumes, Erlang and Handover values along the day. There is a high similarity in the general patterns of the three variables. There is a low Handover event between 11:00 to 16:00 hr during the day, showing call activities are denoting less movement. In the other hand, there is a relatively higher Erlang value at night with a decreasing trend in the call activities from around 19 hr on.

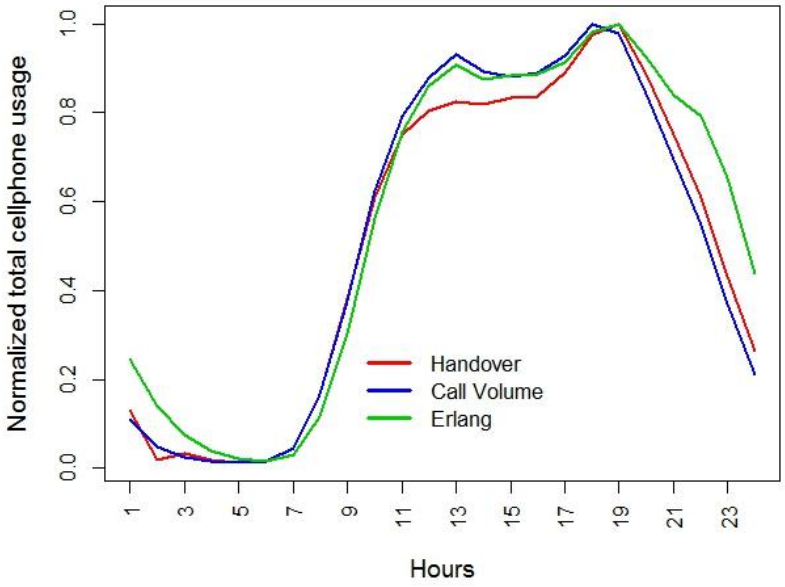


Figure 2.20 Normalized total cellphone usage in Lisbon on April 12, 2010

2.3.8. Lisbon road network

For our analysis we used the digitalized road network of Lisbon, which was supplied by the municipality of Lisbon. The digital road network has detailed information such as, direction, designed capacity, speed limit, name of the road, etc. For the purpose of our analysis; we divided the road links into two groups based on their designed capacities. Out of a total of 29,546 road links with different traffic capacities, road links accommodating 1500 vehicles/hour and above are categorized as main road links and the remaining road links are categorized as local road links. However, this approach resulted in some interruptions in the main network as it is possible to observe in Figure 2.21.

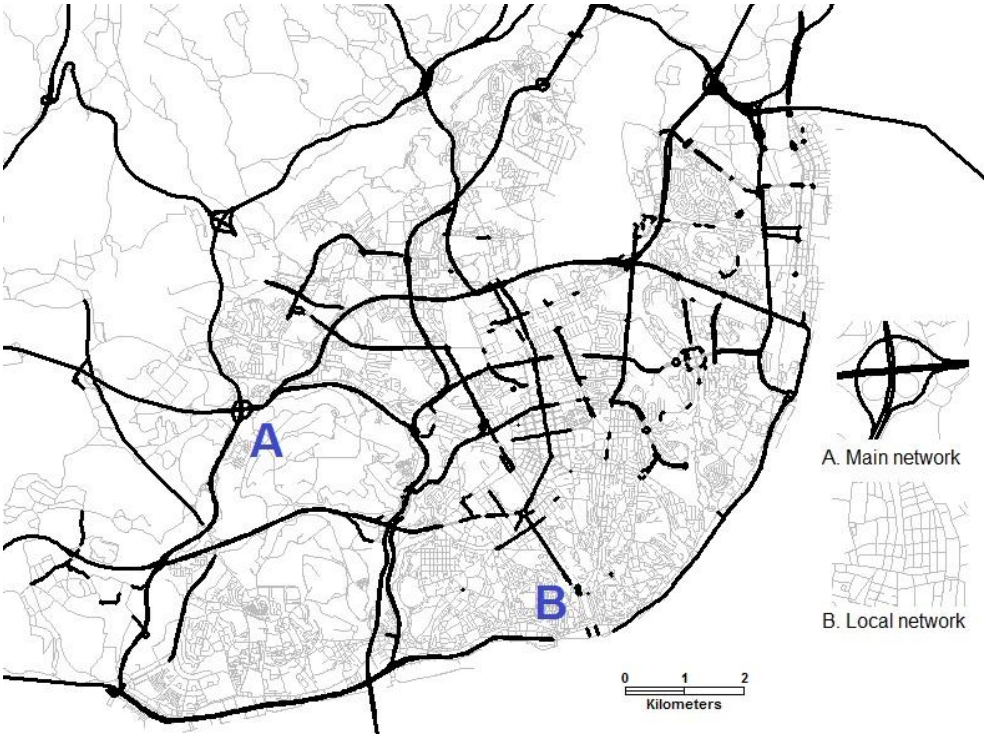


Figure 2.21 Digitalized Lisbon road network

2.4. Reconciliation of the spatial dimensions of the data

One of the main problems associated with studies that integrate different sets of data is the reconciliation of the spatial dimensions of the data. Since there are data from multiple sources that are developed and maintained independently to serve specific needs, the data in each source are represented differently and this results in a large degree of heterogeneity. In addition, different datasets have different spatial coverage and granularity as the different agencies have their own way of structuring the regions. For example, census data was provided at the sub-section level, and the location of cellphone data was provided in a coordinates system.

For data reconciliation, a layer of grid over Lisbon city is used to translate all the datasets to a uniform grid. This data integration would allow easy comparison and analysis. For example, a grid-cell size of 250 by 250 meters was necessary because of the much needed proximity during an exploratory data analysis to investigate the presence of relationships between urban activities and cell towers characterized by different cellphone usages (Chapter 3). In the other hand, an 800 by 800 meters grid-cell size was used to

study the intensity and patterns of activities in the urban areas (Chapter 4). This size was considered to be an adequate level of aggregation because it is coarse enough to accommodate enough cell towers per grid-cell and detailed enough in order not to mix areas of very different characteristics in the city.

With respect to the chosen grid-cell size, each of the dataset is allocated to a single grid-cell based on the most prevalent use within the area (in terms of fraction of area covered). Eqn. 2.1 gives an example of how census data is converted to a grid-cell based data.

$$CensusData_{grid-cell_j} = \sum_i \left(\frac{Area_{grid-cell_j} \cap Area_{SS_i}}{Area_{SS_i}} \times CensusData_{SS_i} \right) \qquad Eqn.\ 2.1$$

Where, sub-section (SS): is the smallest statistical unit in the hierarchy of census data.

$CensusData_{grid-cell_j}$ is the census data aggregated to a $grid - cell_j$.

$CensusData_{SS_i}$ is the amount of census data of SS_i (sub-section i).

$Area_{grid-cell_j} \cap Area_{SS_i}$ is the intersection of areas of $grid - cell_j$ and SS_i.

Chapter 3 Exploring cellular network handover information for urban mobility analysis

3.1. Introduction

3.1.1. Background

Urban areas have diverse activities and complex spatial structures that are supported by urban transport systems. Urban productivity is highly reliant on the competence of its transportation system to move passengers, and freights between multiple origins and destinations. The majority of urban transport problems occurred when the system failed to satisfy the mobility requirements that are generated by the various activities in the city. These problems include: heavily congested roads, parking difficulties, increased pollution, fragility of public transportation systems, and loss of space for productive activities (Rodrigue et al., 2006). To address these problems, transport and urban planners have to develop a means to understand people's mobility patterns. This requires reliable and detailed information regarding the flow of people in a city and understanding of activities at different places in a given city (Becker et al., 2011b). Analysis of activities at different places in an urban environment provides a perspective on human mobility and gives an opportunity to assess the spatial and temporal patterns and trends, which would help for better planning of city dynamics (Becker et al., 2011b).

Currently, planners learn about mobility patterns of people through different techniques, such as survey by interviewer or phone and vehicle counting. These methods have the advantage of providing detailed information about urban mobility patterns. Comprehensive commuting studies, however, require years to be completed and many metropolises learn the presence of new trends only after the release of new survey results (Becker et al., 2011b; Ratti et al., 2006). One way that has been tried to obtain this kind of information is through the use of cellular networks.

The pervasive use of telecommunication technology has changed our ways of exchanging information, interactions among individuals, movements, and use of urban space (Pulselli et al., 2008). These interactions generate mobility related events in the

cellular network, such as location area, route area, and cell updates (Valerio et al., 2009b). These events are useful to sense the movement of large populations of people more regularly, with reduced cost and in a large scale (Becker et al., 2011b). As a consequence, cellular networks mobility related events become an important new source of information for analysis of urban transportation and urban dynamics.

This study uses handover, which is the process of transferring an ongoing call from one cell to the other, to explore its usefulness for urban mobility analysis. Our analysis gives contributions to the following research questions of relevance to urban transport planners, transport geographers, and urban planners: How is it possible to highlight critical spots in the urban road network without incurring great costs? How do we understand the mobility patterns of people in a city? Is handover data adequate and available in the form required for urban analysis? How is the prospect of handover data to detect congestion?

Several authors have studied how cellular network data can be used for understanding people's movements and for urban planning purposes. The study by Asakura and Hato (2004) pointed out limitations associated with questionnaire type travel surveys such as person trip surveys and suggested tracking surveys that measure the precise space–time attributes of an object. Nobis and Lenz (2009) obtained a panel data from the years 2004 and 2007 to examine the complementarity between telecommunications and travel behavior at the level of individual persons. Calabrese et al. (2010) acquired nearly one million cellphone traces with the aim of associating these traces with social events. Results of the experiment showed there was a strong correlation in that: people who live close to an event were preferentially attracted by it; and events of the same type show similar spatial distribution of origins. Sohn et al. (2006) collected Global System for Mobile communications (GSM) traces walking and driving events from everyday lives of three people for a month. The result showed that 85% of the time it was possible to correctly recognize mobility modes among walking, driving, and stationary. González et al. (2008) obtained trajectory of 100,000 anonymized mobile phone users whose position was tracked along six months to form statistical models of how individuals move. Results showed that human trajectories exhibit a high degree of temporal and spatial regularity. Järv et al. (2012) used call detail records (CDRs) of mobile phones to investigate how and to what extent suburbanite commuters affect the evening rush hour traffic. The result shows the daily workplace-to-home trip contributes only 31% of the total evening rush hour traffic.

The study by Pulselli et al. (2008) showed how information from cellphone usage can be used to represent the intensity of urban activities and their evolution through space and time. The studies by Ratti et al. (2005) and Calabrese et al. (2011) also developed city-scale analysis that showed a real-time representation of city dynamics through Erlang values, handover and cellphone trajectories from registered users. The study by Reades et al. (2007) also investigated the correlation between Call Volume and urban activities at six distinct locations in Rome. Handover extracted from anonymized CDRs were also applied to identify which routes people take in a city (Becker et al., 2011a). The study by Ahas et al. (2010) developed a model that illustrates how passive mobile positioning data can provide information about regularly visited places. A comparison of the model results with the population registered data showed the developed model described the geography of the population relatively well.

3.1.2. Our approach

Previous studies have presented the use of cellular networks handover related data (double handover, cell dwell time, and CDRs) for traffic parameter estimation (Alger et al., 2005; Bar-Gera, 2007; Caceres et al., 2007; Herrera et al., 2010; Liu et al., 2008), OD estimation (Pan et al., 2006; White and Wells, 2002), analysis of urban dynamics (Becker et al., 2011b; Calabrese et al., 2010; Calabrese et al., 2011; Ratti et al., 2005), congestion detection (Hongsakham et al., 2008; Thajchayapong et al., 2006), and to understanding people's movements and for urban planning purposes (Ahas et al., 2010; Becker et al., 2011a; González et al., 2008; Järv et al., 2012). Even though many experts are convinced on the usefulness of cellular networks information for the analysis of urban mobility, some important limitations remain to be addressed. These limitations have different effects based on the type of urban problems under investigation:

a. **Limited accuracy:** Compare to other data collection technologies (e.g.: loop detectors), information from cellular based traffic data collection technologies can be provided with limited accuracy (such systems do not provide absolute traffic volume counts, do not differentiate between lanes, and suffer from low quantity of cellular data during night time) (Avni, 2007).

b. **Privacy issue:** Some techniques require individual cellphone signatures. This procedure threatens personal privacy, and cellphone operators should anonymize the data before being used for prediction.
c. **Limited information:** Cellular networks produce massive amounts of data as by-products of their interaction with clients. However, it is a hard task to infer the purpose of the trip, socio-demographic, economic, and psychological information from the cellular networks data that is needed to explain peoples travel behavior (Bolbol et al., 2010).

In spite of the important efforts in applying handover data for the different purposes, by far handover has been the less exploited source of information for the analysis of urban transport and urban dynamics in terms of understanding the mobility patterns of people and some challenges are still to be addressed by the research community.

In our analyses, even though the use of handover information for understanding people's movements is also the goal, a different approach is taken by using handover data to understand the flow of people in terms of the use of the main road infrastructure and the arrivals of travelers to their destination in the morning peak hour. In addition to the handover, we obtained data on traffic volumes and presence of people. We also used the digital road network of Lisbon as the infrastructure supporting flows from/to and between different parts of the city. Road network in urban area is the lowest level of linkage, which is the defining element of the urban spatial structure (Rodrigue et al., 2006).

The remainder of this Chapter is organized as follows: Section 3.2 describes the case study area and dataset collection procedures. Section 3.3 presents the methods that are used to conduct the analyses. In Section 3.4, we provide detailed description of the results. In Section 3.5, we summarize the main conclusions and future directions of our analyses.

3.2. Data description

A case study area of about 156 square kilometer was identified inside the Lisbon Metropolitan Area (LMA), which comprises of the Municipality of Lisbon and its surrounding. Lisbon is the capital of Portugal and the center of the LMA. The LMA has a population of 2.3 million and 18 Municipalities with a total area of 2957.4 square kilometers, where about 23.4% of the population resides in the Municipality of Lisbon (INE, 2013).

According to statistics from ANACOM (2010), active mobile telephone cards per 100 Portuguese inhabitants grew to 159.9 by the end of year 2010, from 140.4 in the year 2008. In this study handover data was obtained from TMN Company, which would be enough to represent the general patterns in the city.

The following discussion gives an understanding about the adequacy of the data to infer useful information by showing TMN´s service coverage. In December 31, 2010, TMN had 7.42 million subscribers in Portugal, which accounts for 45% of the total mobile subscribers. In addition to the increase in the number of subscribers, TMN´s voice traffic grew by 7.1% to 10.54 billion minutes in 2010, compared to 9.84 billion minutes in 2009 (ANACOM, 2010). TMN uses GSM and Universal Mobile Telecommunications System (UMTS) technologies to provide its services. TMN's UMTS population coverage was approximately 93%, and it was geographically available over 4194 municipalities out of a total of 4252 in Portugal (ANACOM, 2010).

The handover data was extracted from 487 cell towers that carry 1669 cell sectors pointed in various directions in Lisbon. The hourly handover counts were gathered in the second week of April, 2010, during working days. The handover data applied was the one generated through voice traffic with an average daily traffic of 2.5 million handovers from a total of 7.2 million calls.

Figure 3.1 shows the volume of calls that were made at each tower locations in the morning between 8AM to 9AM on April 12, 2010. The big circle shows a high volume of calls and the small circle shows a low volume of calls, where the Central Business District (CBD) of Lisbon is also shown in circle.

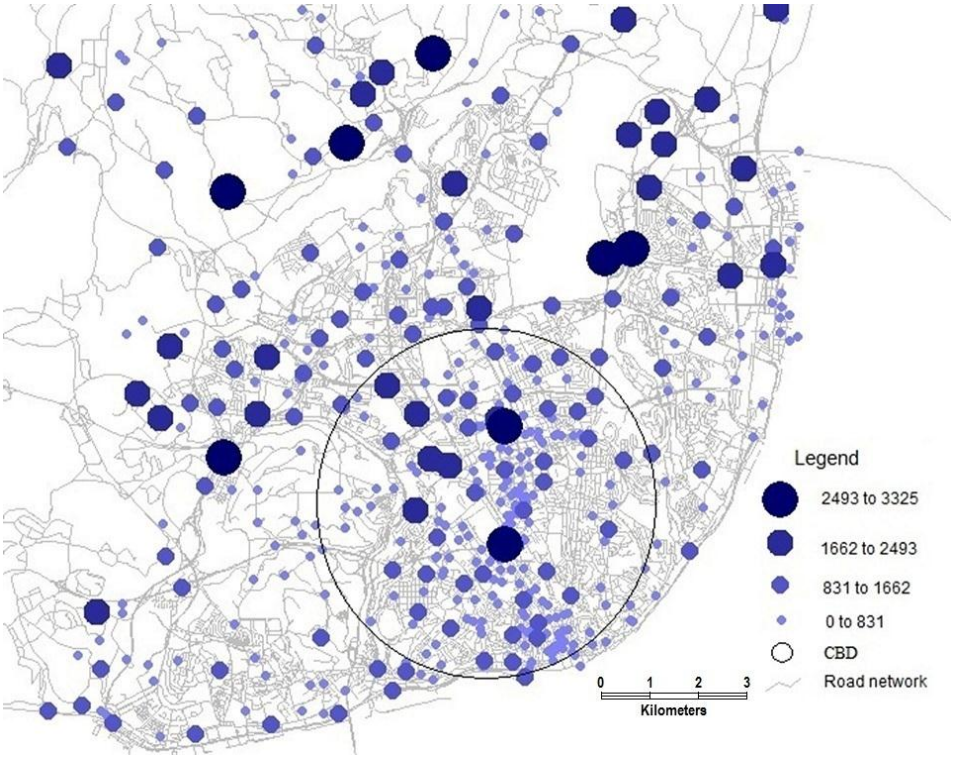


Figure 3.1 Volume of calls at each cellular tower locations in Lisbon between 8Am to 9AM on April 12, 2010

Traffic volume data for the city was obtained from a traffic assignment conducted using the software VISUM with a model of Lisbon's network and an OD matrix of the metropolitan area estimated through a home-based survey and traffic counting in 2008. The use of simulated traffic volume for the test is acceptable as the missing information in the simulation environment was acceptable (more about this dataset is found at (Correia and Viegas, 2011)). Data about the presence of people was obtained in another study in the region (Martínez et al., 2009). This data was deduced from a mobility survey, from the origin and destination of trips and schedules. The data provides the number of people available in real-time.

3.3. Methods

3.3.1. Visualization

Urban mobility analysis requires diverse expertise and advanced tools to identify issues like congested roads, increased pollution, and loss of open space. GIS software is selected from a pool of specialized tools that has been prominent for this kind of analysis. Several procedures were developed using the best functionalities of GIS software (Geomedia Professional) through its urban transportation analysis package. The primary goal of this analysis is to investigate the use of handover data to infer information about the movement of people in the vicinity of the antennas. To achieve this goal, geographic representations were produced through the GIS that allowed us to visualize various characteristics of call movements. Using GIS, handover events associated to the entire cell sectors sited at the same location were summed up and grouped into incoming and outgoing handovers. All calls ended but not originated at cell sectors in a given cell tower were summed up and named as incoming handover. All calls originated but not ended at cell sectors in a given cell tower were summed up and named as outgoing handovers.

3.3.2. Statistical analysis

Besides the visualization task, statistical analysis was performed that allows us to verify different relationships quantitatively. In our analysis we raised the following sets of questions:

- Is the location of cellular towers accommodating high number of moving calls significantly closer to the main road links when compared to the location of towers serving fewer movements?
- Is the amount of handover events significantly higher adjacent to a highest vehicular traffic link compared to the lowest?
- Is the presence of people significantly greater in the vicinity of towers with high incoming handovers compared to the towers with high number of outgoing handovers?

To answer these questions, a statistical method based on the comparison of two sample means was established. Two mutually exclusive hypotheses are required to examine the

difference between two means. Two population means, μ_1 and μ_2 are obtained from the sample of a given population designed in each case. The sampled data were analyzed by first computing the standard error (SE), degrees of freedom (DF), test statistic, and the P-value associated with the test statistic($Eqn.\ 3.1\ to\ Eqn.\ 3.3$).

The standard error of the sampling distribution is computed as:

$$SE = \sqrt{[(s_1^2/n_1) + (s_2^2/n_2)]} \qquad Eqn.\ 3.1$$

Where, s_1 is the standard deviation of sample 1, s_2 is the standard deviation of sample 2, $n_1$is the size of sample 1, and n_2 is the size of sample 2.

The number of degree of freedom is computed as:

$$DF = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \Big/ \left\{\left[\left(\frac{s_1^2}{n_1}\right)^2/(n_1 - 1)\right] + \left[\left(\frac{s_2^2}{n_2}\right)^2/(n_2 - 1)\right]\right\} \qquad Eqn.\ 3.2$$

Where, s_1 is the standard deviation of sample 1, s_2 is the standard deviation of sample 2, n_1 is the size of sample 1, and n_2 is the size of sample 2.

The test statistics, which is a t-score (t), is computed as:

$$t = [(\mu_1 - \mu_2) - d]/SE \qquad Eqn.\ 3.3$$

Where, μ_1 is the mean of sample 1, μ_2 is the mean of sample 2, d is the hypothesized difference between population means, and SE is the standard error.

3.4. Results and discussion

Figure 3.2 and Figure 3.3 are GIS representations that give a qualitative understanding of how the movement of calls can be used to perform urban analysis. The handover map (Figure 3.2) shows the flow of incoming and outgoing calls at the cellphone towers in Lisbon between 8AM to 9AM on April 12, 2010. Handovers intensity is shown using color differentiation. The connection is omitted if there were no handover events during that hour. The light color stands for low handover intensity and the heavy color for high handover intensity. The triangles are representing the location of the cell towers, and the direction of the handovers is shown by an arrow at the end of each trajectory. There is a higher concentration of cellular antennas in the city center having higher number of urban

activities when compared to the outskirts of the city. However, cellular activities between 8AM to 9AM in the morning caused more intense handover events in the east and north part of the city compared to the CBD.

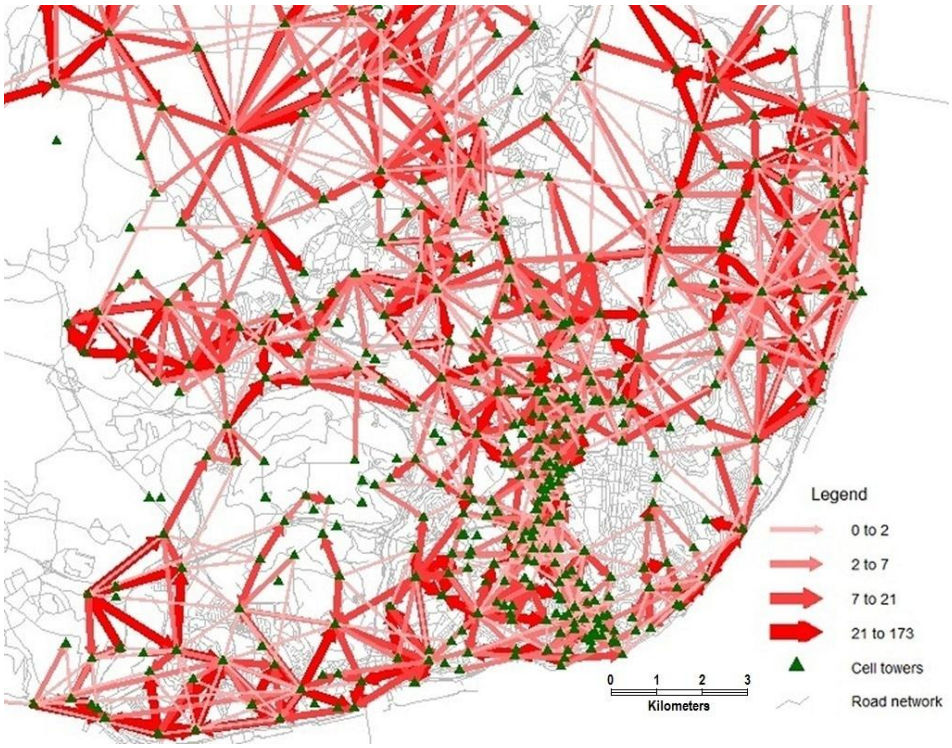


Figure 3.2 Handovers in Lisbon between 8AM to 9AM on April 12, 2010

Figure 3.3 shows the total number of outgoing, and incoming handovers at a given tower location. The size of the circles is proportional to the percentage of outgoing and incoming handovers to the total sum of handovers. This way, large circles mean that one type of handover direction dominates the other. The main road links of Lisbon is also represented at the background to provide the sense of how handovers can be related to mobility in the city.

The underlying assumption is that the pattern and amount of cellphone traffic movement is related with the intensity of urban movements and understanding this relationship will help in managing urban dynamics. In its most simplistic form, neighborhoods with cellular towers showing a high amount of handover events are likely to have high mobility. Neighborhoods with cellular towers showing a high and balanced number of handovers (equivalent number of incoming and outgoing handovers) are likely

to be intermediary mobility areas. Neighborhoods with cellular towers showing a high number of incoming handovers are likely to have a high unidirectional mobility pattern towards the neighborhood.

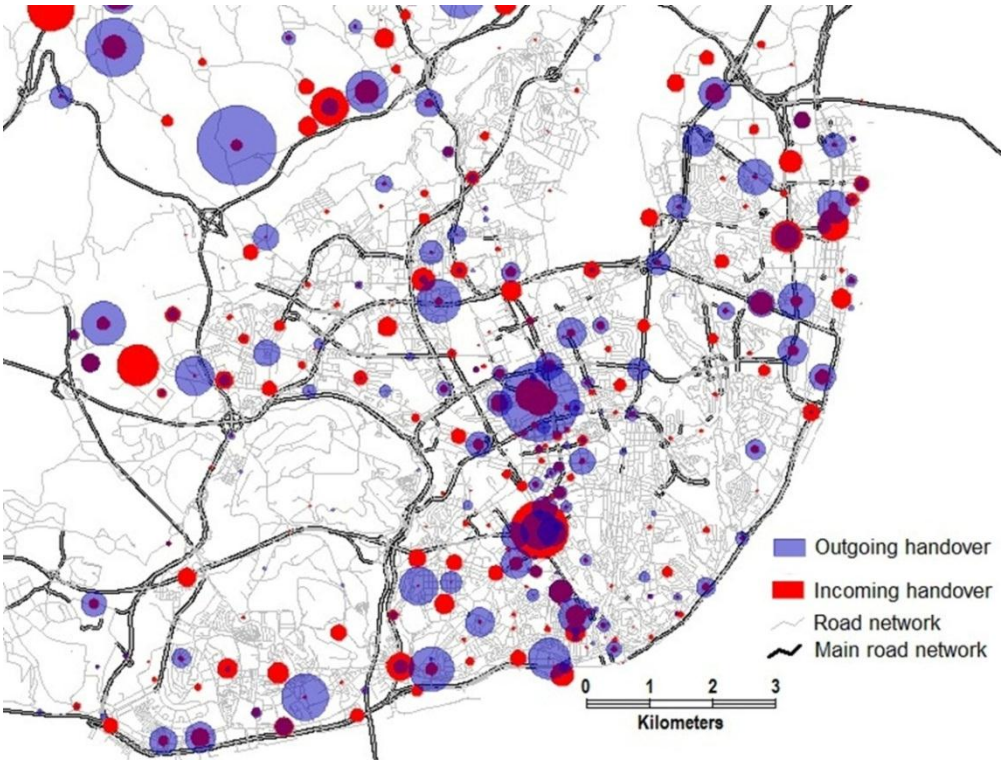


Figure 3.3 Incoming and outgoing handovers over Lisbon´s main road links between 8AM to 9AM on April 12, 2010

The visualization results give a qualitative view on the amount and patterns of call movements. Previous studies, using Call Volume (Pulselli et al., 2008), handover (Ratti et al., 2005), and Erlang (Calabrese et al., 2011), also achieved notable visualization results in the area of urban analysis. The visualizations that were obtained are graphically appealing; nonetheless, the relationship between intensity of cellular activities and the characteristics of a given district requires thorough validation (Ratti et al., 2006). In the following text, four different assumptions are formulated and according to the methodology defined in Section 3.3.; statistical analysis is employed to investigate the assumptions.

Assumption 1: It is claimed that cellular towers with a high number of handover events, both incoming and outgoing, should be adjacent to the main road links when compared to the rest of the cellular towers because the first should denote more movement.

The criteria to choose cell towers with equivalent (balanced) high number of incoming and outgoing handovers were the following: (1) the difference between the number of outgoing and incoming handovers has to be less than 30% of the total handover, and (2) the sum of the incoming and outgoing number of handovers has to be more than 144 per hour. The choice of 30% is arbitrary. Out of a total of 29,546 road links with different traffic capacities, road links with vehicular traffic beyond the 75th percentile were considered as main road links and the computed value was 1500 vehicles/hour. The same cut-off point of 75th percentile was applied to the total handover values and it was aligned to the value 144 handovers/hour.

Table 3-1 shows the results of the hypothesis test, where there are 4.4% chances that the null hypothesis is true given our sample. This shows that the alternative hypothesis is true at the observed and more specific confidence level of 95.6%. This confirms the assumption that the average distance of cellular towers accommodating high number of moving calls are closer to the main road links compared to the towers serving fewer movements.

Table 3-1 Hypothesis test: distance to the main road links (balanced towers)

Groups	Mean (m)	s	n	P-value
Towers with balanced handover	251	247	27	0.0440
All the other towers	341	384	403	

s: standard deviation; n: sample size

Assumption 2: It is claimed that cellular towers with high and equivalent number of incoming and outgoing handovers are located closer to the road links that carry the highest vehicular traffic within 250 meters radius when compared to the other towers.

We chose an optimistic distance of 250 meters assuming that the towers cover the roads within this distance. Other studies mentioned service coverage of cells: a distance of 400 to

500 meters from the antenna (Ratti et al., 2006), and a cell diameter of about 100 to 300 meters (Ratti et al., 2005).

Table 3-2 shows the results of the hypothesis test, where there are 15.95% chances that the null hypothesis is true given our sample. This shows that the alternative hypothesis is true at the observed and more specific confidence level of 84.05%. This confirms the assumption that there is a significantly high number of handovers adjacent to the highest vehicular traffic links compared to the lowest. Cellular towers located close to the main road links are accommodating higher movements and the relatively higher number of handovers in both direction, incoming and outgoing, is an indication of a two directional vehicular traffic passing through the main road links.

Table 3-2 Hypothesis test: highest traffic in a link inside 250 meters radius

Groups	Mean (Veh/hr)	s	n	P-value
Towers with balanced handover	1472	1608	27	0.1595
All the other towers	1152	1287	403	

s: standard deviation; n: sample size

Assumption 3: It is claimed that cellular towers with high number of incoming handover tend to be further away from the main road links compared to the other towers.

The criterion used to choose a cell tower with high incoming handover was set if the number of incoming handover is more than 80% of the total handover (incoming and outgoing). Table 3-3 shows the components of the hypothesis test, where there are 11.6% chances that the null hypothesis is true given our sample. This shows that the alternative hypothesis is true at the observed and more specific confidence level of 88.4%. This result suggests that trip ends related to the major trip destinations are not right beside the main road links which is consistent with what is observed in the city.

Table 3-3 Hypothesis test: distance to the main road links (incoming handover)

Groups	Mean (m)	s	n	P-value
Towers with high incoming handovers	238	202	141	0.1160
All the other towers	211	197	177	

s: standard deviation; n: sample size

Assumption 4: It is claimed that cellular towers with high incoming handovers are associated with the presence of a high number of people in its vicinity. Information about the presence of people is given in a 200 by 200 meters grid cell.

Table 3-4 shows the components of the hypothesis test, where there are 12.55% chances that the null hypothesis is true given our sample. This shows that the alternative hypothesis is true at the observed and more specific confidence level of 87.45%. The result proved the connection of a high number of incoming handovers and the presence of people around these towers in the morning. This result shows the potential of characterizing areas on the bases of mobility related features such as handover, unlike the previous studies that characterize neighborhoods on the bases of static information (Call Volume and Erlang) obtained from cellular networks (Ratti et al., 2006; Reades et al., 2007).

Table 3-4 Hypothesis test: people's presence adjacent to cell towers

Groups	Mean (no. of people)	s	n	P-value
Towers with high incoming handovers	1438	2340	141	0.1255
All the other towers	1168	1675	177	

s: standard deviation; n:sample size

The main findings in our analyses reveal that there exists a significant relationship between the handover and other urban activities. Therefore, our results contribute to an understanding of where handover data can be applied. The results can be used as a stepping stone for future studies that are aiming to develop predictive models in a city-scale that can be employed as low-cost estimators of traffic and population size, which are expensive to compute through traditional methods. This could be applied, namely in some developing countries as a cheaper way to understand mobility patterns at a city-scale. For example, in the year 2012, in twenty of the sub-Saharan African countries, active mobile telephone

cards per 100 inhabitants were ranging from 65% to 164% (Deloitte and GSMA, 2012). However, these countries have poor availability of data for urban and transportation planning, where most of the cities do not carry out traffic counting in a regular basis, and except South Africa, no other city in the sub-Saharan countries has household travel surveys (Williams, 2011).

For urban transport planners, transport geographers, and urban planners there are a whole range of issues arising out of the increase on demand for urban transportation. Congestion will remain as one of the challenges of urban areas. The other challenge will be availability of data for transportation studies that can be adequate in form and content for urban analysis (Rodrigue et al., 2006). Road traffic congestion poses a challenge for urban areas. In fact, urban areas are prone to congestion; however, road transport policies should seek to manage congestion on a cost-effective basis with the aim of reducing the impact on urban dwellers throughout the urban road network (OECD, 2007). Traffic congestion transcends local community borders, where local policy-makers may have limited ability to devise solutions because of the scarce financial resources they can access. However, transport policy and planning requires a broader perspective, one that thinks through different alternatives. Consideration of handover as an alternative or complementary data to sense the movement of large populations of people more regularly, with reduced cost and in large scale should be taken as a challenge that is relevant for local policy-making.

3.5. Summary

Over the past decades there has been explosion in the deployment of cellular services. Cellular networks produce a massive amount of data that can tell us about the presence and movement of people in a given area. This information can be used as a valuable source of information for the study of social dynamics and their interactions with the urban structural forms.

In this chapter, we applied cellular networks handover information. From the perspective of the quality of information, the data in this analysis have advantages over the data obtained through traditional surveys for estimating mobility patterns. Traditional studies of urban dynamics follow time consuming techniques to acquire data, and large scale studies can be laborious and expensive. On the other hand, handover data is collected nearly in real-time at very fine time resolution from a large portion of the population.

Two different analyses were carried out to uncover the mobility relevance of cellular networks handover information. In the first analysis, maps were produced to geographically visualize the handover information. In the second analysis, the hypothesis testing objectively proved four different assumptions. It was found that cellular towers characterized by high and balanced number of incoming and outgoing handovers are located in the vicinity of the main road links thus where there are the major flows of people moving in the city. There is a strong association between the presence of people in the city and the number of incoming handovers, supporting the idea that these towers are next to the main points of trip arrivals.

In our analyses, we explored the presence of significant associations between handover and traffic volume in the main road network. However, there are three main limitations associated to this analysis: Firstly, handover data is limited to mobile phones that are actively making calls, and the duration of the associated calls must be long enough to traverse the boundaries of two cells, thus it is not possible to make a direct correspondence of the handover and traffic counts. Secondly, even after giving priority to the cell towers close to the road link under analysis in order to consider the cells with information more relevant for characterization of a specific road link, it is a challenging task to sort out calls that were carried out while driving on those specific road links. This is because of the large geographic areas covered by individual antennas that could take multiple road links. Thirdly, the influence of cellphone use from pedestrian is severe in urban areas and may obfuscate the signature produced by calls from vehicles.

Information from cellular networks is ubiquitous and allows us to comprehend and visualize the flow of people from the entire urban system and its organization at a glance. The main results of our analyses reveal the existence of a significant relationship between the handover and other urban activities. Therefore, these findings can be used as a stepping stone for future studies that are aiming to develop predictive models in a city-scale that can be used as a low-cost estimator of traffic. In particular, the use of handover information for urban analysis promises to produce practical applications for urban management, route planning, traffic estimation, emergency detection and general traffic monitoring (Calabrese et al., 2011).

Future work will aim at developing traffic estimation and prediction models through the handover counts. This might require more detailed information on the handovers, such as

having a lower aggregation level than the hour and better information about the orientation of the cell tower sectors. Another interesting approach would be a hybrid application of handover data and results obtained from traditional surveys.

Chapter 4 Analysis of the pattern and intensity of urban activities through aggregate cellphone usage

4.1. Introduction

4.1.1. Background

Theories in the 1950´s and 1960´s regarding the spatial organization of urban activities portrait cities, usually at cross-sections in time, as resistant for a change. These theories highlight that new urban changes have small effect on the general day to day dynamics and form of a city. This is because historical inertia seems to dominate cities physical form at a macro level with cities making piecemeal adjustments to their structure in response to urban changes (Batty, 1996). The perception that spatial structures are long lasting has been reflected on urban studies that assume slow changes in the urban organization, where activities that cause such changes happen over months or years rather than shorter time cycles (Bertaud, 2004; Batty, 1996). For example, different transportation modeling techniques are developed from this standpoint in that routine trip-making behavior is the focus of study and not the variablility of activities that people perform in each area (Batty, 2002).

In spite of the aforementioned views, cities are clearly dynamic structures which present a complex pattern of constantly changing colours and shapes. Klapka et al. (2010) attempted to define the spatial organization of a city in terms of three factors: (1) spatial interaction, movement and distribution of population, (2) activities of different types that are associated to different land use, and (3) environments regarding distribution of economic activities and wealth, cultural and political conditions etc.

Sevtsuk and Ratti (2008) argue that there are sequences of urban activities that take place over shorter time periods which affect the spatial structure and forms of a city. Batty (2002) proposed the possibility of perceiving cities as clusters of “spatial events”, events that are characterized by duration, intensity, volatility, and location. Therefore, our understanding of cities would not remain at the stage whereby spatial structures are resilient and long lasting, but extends to urban activities performed by its residents.

Analysis of activities at different places in an urban environment provides a perspective on human mobility and gives an opportunity to assess the spatial and temporal patterns, which would help for better mobility and land use planning (Becker et al., 2011). There are situations where the same land use could mean different levels of activity over time, for instance, a station when a sport event occurs, a trendy pub area, a recently refurbished office building that attracts new companies.

Traditionally, planners perform surveys to estimate the mobility patterns and the whereabouts of people. This method has the advantage of providing detailed information. However, travel surveys suffer from relatively small samples due to their cost and are usually conducted at infrequent intervals, thus, many metropolis learn the presence of new trends only after the release of new census results when mobility questions are part of those comprehensive surveys (Becker et al., 2011; Ratti et al., 2006). In addition to that, it is a hard task to learn the distribution of residents´ activities and urban dynamics through traditional surveys. One recent and innovative way that has been tried to obtain this kind of information is through analyzing data from the cellular network usage.

The pervasive use of telecommunications technology has changed our ways of exchanging information, interaction between individuals, mobility, and use of urban space (Pulselli et al., 2008). These interactions generate huge amounts of cellular network data that hold a large potential for providing important information on places and activities. Lu and Liu (2012) noted that comparing to GPS; mobile positioning has the potential to support a wider range of studies that require large samples, because of its capability to access most locations from a large portion of the population. However, passive mobile positioning data have two significant limitations as a source of location information: Firstly, it is sparse in time as it is only acquired when a phone is engaged in a call or short message service. Secondly, it is coarse in space because the location record is made at the granularity of cell tower. The spatial accuracy depends on the density of those cell towers. The location error is estimated to be within a range of 300 to 600m in cities and could be of several kilometers in suburban areas (Ahas et al., 2007).

In recent years, some authors explored the use of cellular network data for urban mobility analysis. Most notably, Demissie et al. (2013) revealed the existence of a relationship between calls handover intensity between towers and different urban characteristics: proximity to main road network, presence of people, and traffic in the main

road network. Ahas et al. (2010) stated the use of traditional census and population register data for long-term planning processes and recommended alternative means, such as cellular network data for everyday mobility analysis. Asakura and Hato (2004) mentioned important limitations associated with questionnaire type travel surveys, such as person trip surveys and suggested tracking surveys that measure the precise space-time attributes of an object. Calabrese et al. (2010) investigated the association of cellphone traces with social events and identified a strong correlation in that: people who live close to an event were preferentially attracted by it; and events of the same type show similar spatial distribution of origins.

The studies by Ratti et al. (2005) and Calabrese et al. (2011) performed city-scale analysis to represent real-time city dynamics through Erlang and handover values, and cellphone trajectories of registered users. The study by Pulselli et al. (2008) showed how information from cellphone usage can be used to represent intensity of urban activities and their evolution through space and time. Zuo and Zhang (2012) employed Erlang values to detect and analyze different hotspots in the urban system. Calabrese et al. (2011a) used mobile phone location data to estimate dynamic OD matrices. Demissie et al. (2013a) developed a model that uses handover data to distinguish the intensity of traffic in urban environments and provides estimates of traffic levels in specific road segments. Calabrese et al. (2010a) applied a clustering analysis on a Wi-Fi network data to segment locations based on their associated digital signal. A comparison of this result with a reference data that displays classifications of access points by usage type showed a good match.

There have been studies that explored passive mobile positioning data for the purpose of identifying places that are meaningful to mobile phone users (Ahas et al., 2010; Csáji et al., 2013; Isaacman et al., 2011). Meaningful places were defined as frequently visited places that represent personal anchor points, such as home, work, and other important locations (Ahas et al., 2010). Extraction of meaningful places requires traces of location and time for individual calls. However, the aforementioned studies replaced the real identity of the phone users through randomly assigned IDs to deal with privacy issues. In the process of deducing meaningful places, most of the studies were confident in determining home and work-time anchor points (Ahas et al., 2010; Csáji et al., 2013; Isaacman et al., 2011).

Besides anchor point modeling, some authors have been exploiting the anonymous passive mobile positioning data to deduce the predominant land use in urban systems.

Toole et al. (2012) applied aggregated Call Detail Records (CDRs) to understand how the population of different areas of a city changes with time and investigated if zones of the same kind share common usage. Reades et al. (2009) attempted to associate cellular networks Erlang data to Points of interest (POI) derived from the Italian "yellow page". Becker et al. (2011) applied aggregated voice and SMS cellphone activities to capture different usage patterns of city dwellers. A comparison of this result with a reference data that displays classifications of access points by usage type showed a good match. A study by Soto and Frías-Martínez (2011) achieved good results by comparing automatically identified land uses from CDRs with expert knowledge of a city land uses.

4.1.2. Our approach

The generalized and ubiquitous adoption of cellphones opened new possibilities on sensing the urban space. Cellular networks provide new resources for spatiotemporal datamining and geographic knowledge discovery (Yuan et al., 2012). The majority of existing urban studies have focused on identifying a geographical domain in which a specified set of activities took place. Some authors investigated the number of places where a person spends a significant amount of time and/or visits frequently via surveying individual cellphone usage (Ahas et al., 2010; Csáji et al., 2013; González et al., 2008; Isaacman et al., 2011). Even though results are fairly in the same range, these studies showed that different outcomes could be obtained at different geographical areas: González et al. (2008) studied the trajectory of phone users tracked for six months and showed that human trajectories exhibit a high degree of temporal and spatial regularity and people spend their time at few locations. Ahas et al. (2010) argued that with the rising mobility of individuals, the dominance of home and work anchors has reduced, and people also spend significant amount of time in other locations. Study by Isaacman et al. (2011) used two months CDRs data and found that New Yorkers have a higher percentage of people with 1 to 4 important places, whereas Angelenos have a higher percentage of people with 5 to 8 important places. Csáji et al. (2013) applied a one week individuals cellphone usage and found that the average number of frequently visited locations per user in Portugal is approximately 2.14 and 95% of the users have less than 4 frequently visited locations. It has to be noted that these studies followed common procedure in determining which cell towers are relevant for characterizing frequently visited locations, which is extracting

consistent cells with highest number of days with calls. However, they used different number of days to track trajectories of cellphone users, and applied different threshold values to exclude irrelevant traces.

Some authors attempted to connect aggregate cellphone usage to a geography of human activity derived from actual land use that supplements zoning regulations (Toole et al., 2012), Points of interest (POI) obtained from the Italian "yellow page" (Reades et al., 2009) , expert knowledge of a city land uses (Soto and Frías-Martínez, 2011), and census data (Becker et al., 2011). However, Toole et al. (2012) argued that low classification accuracy was achieved because some zones might feature different usage from intended use that was dictated by zoning regulations implemented and enforced by local governments. Reades et al. (2009) concluded that the presence of a high degree of overlap between different activities in Rome hinders the possibility of inferring predominant land uses in urban environment.

Although the results from previous urban studies show great potential for using cellular network information to infer geographical areas in which a specified set of activities takes place, some of the following limitations remain to be addressed:

Lack of validation: Supported on cellular networks data, several authors inferred areas of the city exhibiting different characteristics; however, most of the studies did not validate their results.

Insufficient ground truth: Some authors attempted to validate their results, but the ground truth was either insufficient or unfitting for the following reasons: (1) information obtained from local governments regarding urban land uses usually tell how land use is planned, not the actual use, (2) most studies use population density alone based on census data as ground truth. However, census data shows spatial distribution of people when they are at home (at night). Census does not provide any information regarding where people are or patterns of their daily trips during the day.

Inaccurate representation of the cellular network: Alternative techniques that approximate the real coverage areas of cellphone towers would undermine the outcomes of

studies in this field. Better knowledge regarding the orientation of the cell tower sections is necessary.

Regardless of the important efforts in applying cellphone data to interpret areas of cities exhibiting different characteristics, by far aggregate cellphone usage has been the less exploited source of information for the purpose of qualitatively measuring urban activities and some challenges are still to be addressed by the research community. One important issue is how to define parameters used to approximate and measure the intensities and patterns of activity at different locations. In this Chapter we address the use of passive mobile positioning data: Call Volume (the number of calls), Erlang, and Handover values to detect the intensities and patterns of urban activities along the hours of a day. A fuzzy c-mean clustering algorithm is used to create clusters of locations with similar characteristics. We also examined the significance of the results through comparison with ground truth information. This includes establishing a comprehensive set of indicators that would define the pattern and level of activities at different locations. We addressed the characteristics of activity places with a number of indicators related to People´s presence, residential buildings and POIs, bus movement, and taxi movement.

The remaining of this Chapter is organized as follows: Section 4.2 describes the case study area and dataset acquisition procedures. Section 4.3 presents the developed methods used to explain urban activities through the cellphone data. In Section 4.4 we provide detailed description of the results. In section 4.5 we present a discussion on the results of our analysis as well as its main limitations, and in Section 4.6, we present a summary of the chapter addressing the main conclusions and future research directions.

4.2. Data description

The municipality of Lisbon was used as a case study to illustrate the analysis carried out in this Chapter. Lisbon is the capital of Portugal and the center of the Lisbon Metropolitan Area (LMA). The LMA has 2.3 million inhabitants and has 18 municipalities with a total area of 2957.4 square kilometer, where about 24.3% of the population resides in the municipality of Lisbon (INE, 2013). The municipality of Lisbon went through numerous morphological periods to achieve its current urban form. The consecutive periods of urban planning resulted in new urban forms and also influenced the expansion of the city, which is supported by the coexistence of structural urban elements from the past and integrates

new spatial uses and organizational proposals. Unlike some North American cities, Lisbon does not have high spatial segregation of activities; instead there is a high degree of overlap between activities within the city. The downtown of the city is composed of touristic, historical, and commercial areas with significant population density. The outskirts of the city are mainly occupied by residential areas and major infrastructures, such as airport and industrial facilities (Oliveira and Pinho, 2010).

By the end of the year 2010, the number of active mobile telephone cards per 100 Portuguese inhabitants was 159.9 (ANACOM, 2010). This mobile penetration was obtained through the combined services provided by three major mobile operators: Vodafone, Optimus, and TMN. For our analysis cellphone data was obtained from TMN Company. In December 31, 2010, TMN had 7.42 million subscribers in Portugal, which accounts for 45% of the total number of mobile subscribers. TMN uses Global System for Mobile communications (GSM) and Universal Mobile Telecommunications System (UMTS) technologies to provide its services. TMN's UMTS population coverage was approximately 93%, by 2010 and it was available over 4194 municipalities out of a total of 4252 in Portugal (ANACOM, 2010).

We obtained passive mobile positioning data. Passive mobile positioning data is automatically stored in the log files of service providers. The usual method for passive mobile positioning is "a billing log", which is recorded for call activities like call, SMS, etc (Ahas et al., 2010). Other sources of call activity are also used for passive positioning, such as Erlang value of a cell (Reades et al., 2007).

A cellular mobile network is built around base stations, which are responsible for transmitting signal from one cell to another one. When a call is performed, the signal is transmitted across a network of interconnected geographic areas called cells. A cell represents the area served by base station, which is often called mast, tower or cell-site. We analyzed cellular traffic handled by 361 base stations comprising 1194 cell sectors. We obtained hourly aggregated Call Volumes, Erlang, and Handover values at the granularity of a cell sector. The datasets were obtained on April 12, 2010, which was a working day. It follows a brief explanation on each dataset:

Call volume: The Call Volume measures the phone activity in terms of the number of calls that happened within a given area and time window. Figure 4.1 shows the volume of calls

at each tower area in the afternoon between 5 PM to 6 PM on April 12, 2010. Cell towers with a high number of Call Volume are represented with bigger and darker circles. The Central Business District (CBD) of Lisbon is also shown in circle.

Erlang: The Erlang measures the bandwidth usage, which is usually applied for network capacity planning on GSM networks and can be easily collected by the operator. We acquired hourly aggregated and anonymous Erlang data. An Erlang is defined as one person-hour of phone usage: alternatively, one Erlang can be obtained through two people talking for a half hour each, sixty people talking for one minute each and so on. Consequently, Erlang data provides opportunities to demonstrate spatial as well as temporal dynamics as seen through network bandwidth consumption (Reades et al., 2007). Figure 4.2 shows the Erlang values at each tower in the afternoon 5pm to 6pm on April 12, 2010. Large circles show a high Erlang value and the small circles show a low Erlang value.

Handover: One of the major features of cellular mobile network is its ability to provide data on the movement of a person when making a call. A mobile handset can move out of one cell to another one while in active communication (i.e., call or data session) or while it is in idle state. The handover (also called handoff), which is the process of transferring an ongoing call or data session from one base station to another without loss or interruption of service, is the most usual mobility management procedure (Bratanov, 1999). We acquired hourly aggregated and anonymous handover data. For our analysis we used total handover (outgoing plus incoming) of a given cell-site. Figure 4.3 shows the total handover (both incoming and outgoing handovers) associated to each cell tower location in Lisbon in the afternoon between 5 PM to 6 PM on April 12, 2010. The small circles are for low handover intensity and the large circles are for high handover intensity.

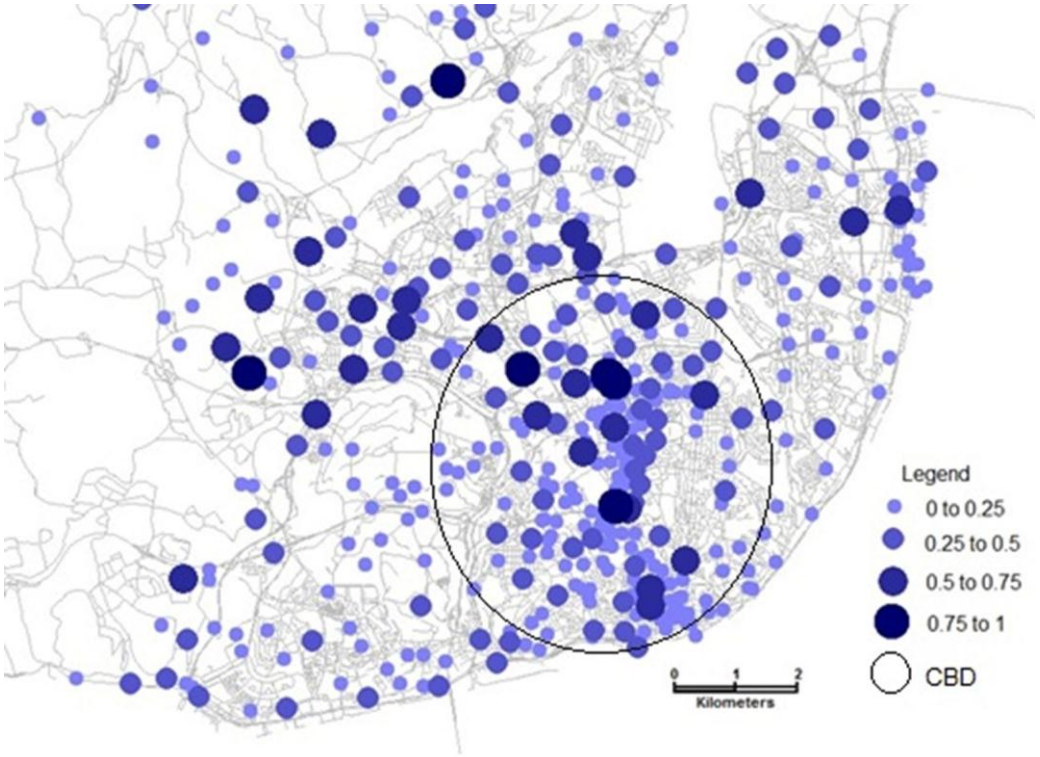


Figure 4.1 Lisbon cellphone traffic (Call Volume) between 5 PM to 6 PM on April 12, 2010

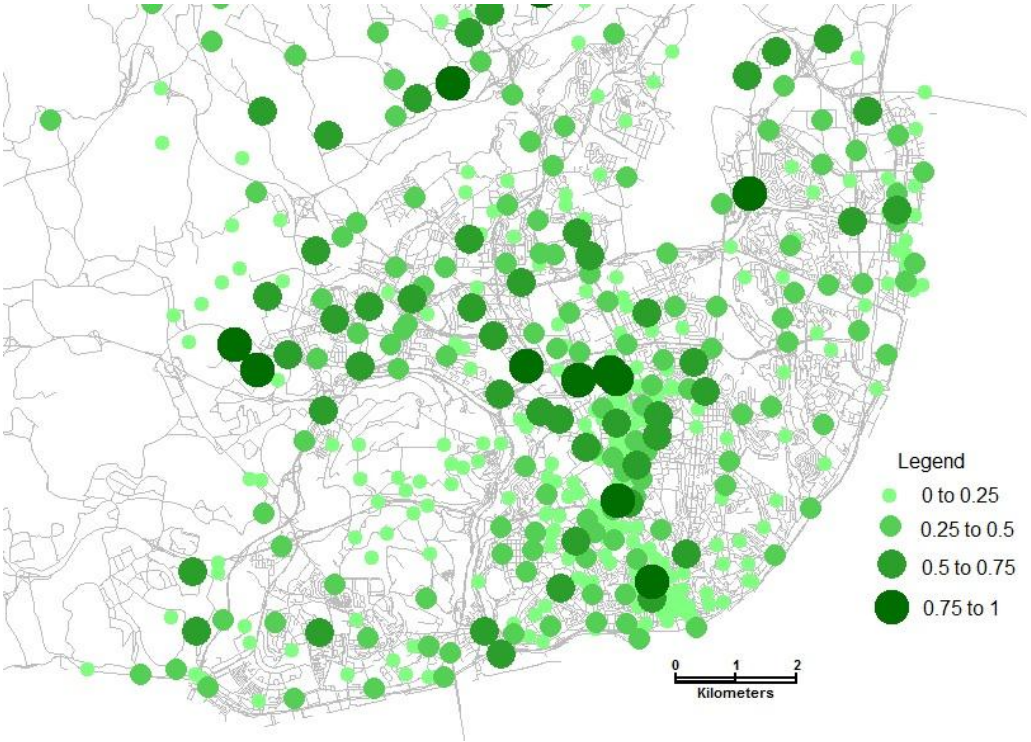


Figure 4.2 Lisbon cellphone traffic (Erlang) between 5 PM to 6 PM on April 12, 2010

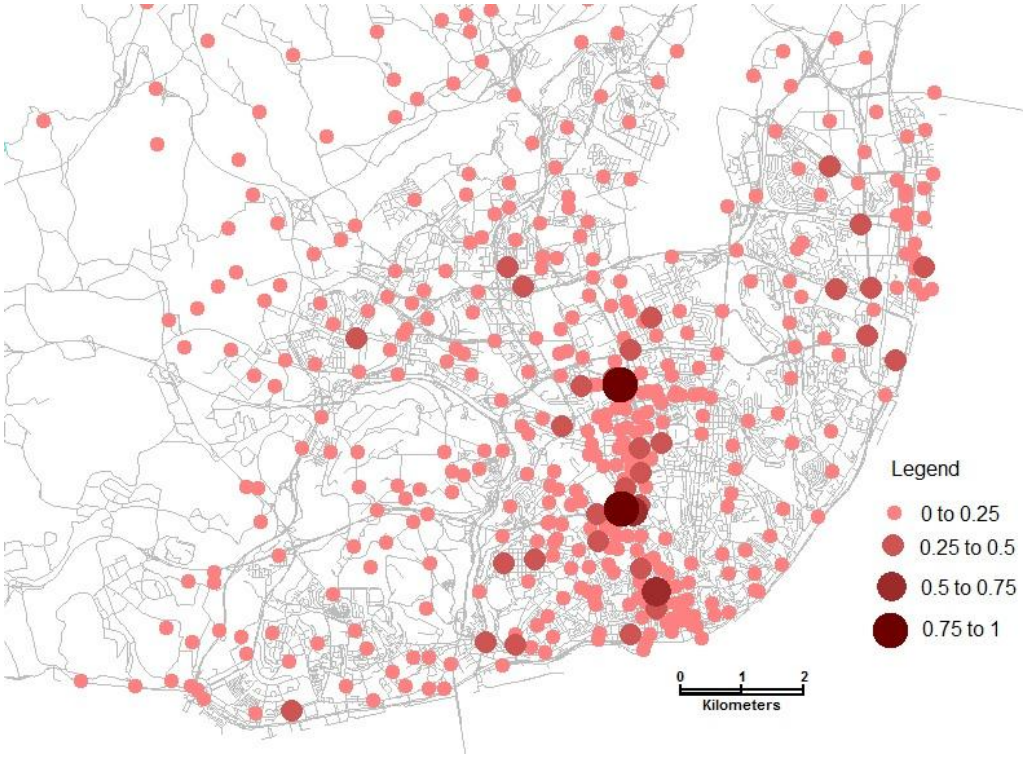


Figure 4.3 Lisbon cellphone traffic (Handover) between 5 PM to 6 PM on April 12, 2010

Residential buildings: we obtained the number of residential buildings from the Portuguese census data collected during the year 2011 with a granularity of the smallest statistical unit, sub-section (SS) (the SS encompasses on average 300 dwellings).

Points of interest: The POIs data were acquired from Servidor de Apontadores Portugueses (SAPO). The dataset includes different business categories (service, recreation, education, shopping, health, and transportation facilities), and location of each POI is given by longitude and latitude.

Bus movement: we obtained hourly aggregated bus arrivals at each bus stop in Lisbon on April 12, 2010.

Taxi movement: We also obtained hourly aggregated taxi movement. We summed up the number of taxis that stopped in each grid-cell for the purpose of passenger pick-up and drop-off along a day. We do not have taxi movement data from the same day the cellphone

data were obtained, April 12, 2010 (Monday). As a replacement, we used average taxi movement data from the same day of the week in the month of November 2009.

Presence of people: data about the presence of people was obtained in a previous study in the region (Martínez et al., 2009). This data was deduced from a mobility survey, from the origin and destination of trips and schedules. These data provide hourly aggregated number of people.

4.3. Merthods

In order to automatically infer population patterns along time and space, and to detect the level of activities we present the following techniques that use information obtained from a cellular network.

4.3.1. Reconciliation of the spatial dimensions of the data

One of the problems associated with studies that integrate different sets of data is the reconciliation of the spatial dimensions of the data. Different datasets have different spatial coverage and granularity as the different agencies have their own way of structuring the regions. For example, census data was provided at the SS level, and the location of cellphone data was provided in a coordinate system. For conciliation of this we used a layer of grid net over Lisbon city with a grid-cell size of 800 by 800 meters to transform all the datasets to a uniform grid. The choice of an 800 by 800 meters grid-cell size was considered to be an adequate level of aggregation because it is coarse enough to accommodate enough cell towers per grid-cell and detailed enough in order not to mix areas of very different characteristics in the city.

4.3.2. Normalization over time and space

The magnitude of the differences between cellphone usages at different grid-cells makes it hard for detailed analysis. Thus, we normalize the cellphone data over time and space for easy clustering and graphical analysis based on relative cellphone values.

Normalization over time: For a given grid-cell, the normalization shows the intensity of cellphone usage in each hour relative to its daily total cellphone usage. This way we can compare the varation in urban activity patterns between grid-cells. The normalization is given by $Eqn.\,4.1$.

$$A_i = \sum_{j=1}^{m} x_{ij} \ , \quad B_{ij} = \frac{x_{ij}}{A_i} \qquad Eqn.\,4.1$$

Where, $i = 1, \dots n$, n=118 grid-cells; $j = 1, \dots, m$, m=24 hours; A_i is the sum of daily cellphone usage of a given gird-cell i; x_{ij} is the cellphone usage of grid-cell i at hour j of the day; B_{ij} is the proportion of cellphone usage at grid-cell i with respect to its total daily cellphone usage at hour j of the day.

Normalization over space: this normalization shows the intensity of cellphone usage at each grid-cell relative to the total cellphone usage of the entire area under study at each hour of the day. The normalization is given by $Eqn.\,4.2$.

$$D_j = \sum_{i=1}^{n} x_{ij} \ , \quad E_{ij} = \frac{x_{ij}}{D_j} \qquad Eqn.\,4.2$$

Where, $i = 1, \dots n$, n = 118 grid-cells; $j = 1, \dots, m$, m=24 hours; D_j is the sum of the cellphone traffic from all the grid-cells at hour j of the day; x_{ij} is the cellphone traffic of grid-cell i at hour j of the day; E_{ij} is the proportion of cellphone usage at each grid-cell with respect to the total cellphone usage from all the grid-cells at hour j of the day.

4.3.3. Fuzzy clustering

The geographical area of the city is represented by a layer of 800 by 800 meters grid, and each grid-cell is characterized through the corresponding cellphone traffic. A clustering method will be used to create groups of grid-cells with similar activities of urban uses. The goal is that the objects within a cluster be similar to one another and different from the objects in other clusters. Clustering methods can be distinguished in different ways; one

possible classification of clustering methods can be whether the set of clusters are fuzzy or crisp. In our analysis, different clustering methods are possible candidates to create partitions based on the affinity between objects in the dataset. We chose a Fuzzy C-Mean (FCM) clustering approach, because it allows objects to belong to all clusters simultaneously with different degrees of membership. In many situations, fuzzy clustering is more natural than other type of clustering methods which segment the given observation as exclusive clusters (Sato-Ilic, 2006).

We used the FCM algorithm to partition objects $\mathrm{X} = \{x_1, x_2, \dots, x_n\}$ into a collection of C fuzzy clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of C cluster centers V, such that $V = (v_k), k = 1, \dots, C$ denotes the values of the centroid of a cluster C. The state of fuzzy clustering is represented by a partition matrix $U = (u_{ik}), i = 1, \dots, n, k = 1, \dots, C$. In general, u_{ik} satisfies the conditions in $Eqn.\,4.3$.

$$u_{ik} \in [0,1], \forall i, k; \ \sum_{k=1}^{C} u_{ik} = 1, \forall i \qquad Eqn.\,4.3$$

Where, u_{ik} is the degree of belongingness of an object i to a cluster k.

The algorithm aims to minimize the weighted within-class sum of squares shown in $Eqn.\,4.4$.

$$J(U, v_1, \dots, v_C) = \sum_{i=1}^{n} \sum_{k=1}^{C} (u_{ik})^m \, |x_i - v_k|^2 \qquad Eqn.\,4.4$$

Where, $v_k = (v_{ka}), k = 1, \dots, C, a = 1, \dots, p$ denotes the values of the centroid of a cluster C with respect to variable p , $x_i = (x_{ia}), i = 1, \dots, n, a = 1, \dots, p$ is the i^{th} object with respect to p variables, and $|x_i - v_k|^2$ is the squared Euclidean distance between x_i and v_k. The exponent m determines the degree of fuzziness of the clustering. For $m \to 1$, the method becomes equivalent to k-means clustering whereas for $m \to \infty$ all the data objects have identical membership to each cluster.

4.4. Analysis of the pattern and intensity of urban activity

We now analyze the results of the clustering analysis. First, we present our results concerning the patterns of activities, which show the patterns of activities of different locations of the city calculated through their cellphone usage. We then analyze the intensity of activities of the same locations. In both cases, we compare the results to independent statistics.

4.4.1. Analysis of the pattern of urban activity

One of the objectives of our study is to detect patterns of activities at different locations of the Municipality of Lisbon. For a given grid-cell, we define the patterns of activities as the evolution of its calling activity along the hours of the day. Thus, we normalized the cellphone usage over time. We employ cellphone activity to segment city areas into categories that can help urban planners and transport geographers answer various questions: How is a particular area of the city used? How is the population distribution in a given area over time? We hypothesize that this type of questions can be addressed by clustering urban areas based on their cellphone activities. Answering these questions through the use of cellphone data would be important to identify emerging short to medium run trends, and to make effective and efficient planning decisions on the city, especially in urban areas where traditional data are not available.

We applied an unsupervised fuzzy clustering algorithm that has no prior information regarding the activity profile of the grid-cells. We define the activity patterns of the grid-cells in terms of three different cellphone data: Call Volume, Erlang, and Handover. The input dataset consists of 118 grid-cells × 24 hours cellphone values. Each of the 24 values for each grid-cell represents a normalized cellphone usage for a specific hour of the day. In this case we used normalization over time defined in Eqn. 4.1.

In addition to the cellphone activity signatures, FCM requires as input the number of clusters, and the fuzziness coefficient (m) which influences the fuzziness of the resulting partition. Usually, the value of the fuzzier is set equal to two (Schwammle and Jensen, 2010). This may be considered as imposing an a priori fuzziness in the dataset. In our study, we applied the clustering procedure to our datasets for a range of fuzziness coefficients $m = \{1.5, 1.6, \dots, 2.5\}$. Two important conclusions can be drawn from the results: (1) the number of objects in each cluster stays relatively unchanged for all

coefficients; (2) there is a tendency to zero of both the value of the objective function and the distance between the cluster centroids when increasing the fuzzier. Therefore, a threshold for the fuzzier value m is reached as soon as we have: (1) no change in the number of objects in each cluster after the increment on the fuzzier value, (2) the smallest possible value for the objective function in Eqn. 4.4, and (3) a reduction in the percentage change of the average sum of squared error of the distance between the cluster centroids. Through these procedures we found fuzziness coefficient value of 2.4 for the ground truth data, 1.5 for the Handover data, 1.7 for the Call Volume data, and 1.6 for the Erlang data.

In order to find the best number of clusters we applied a statistical method based on the comparison of two sample means. In the first step, we applied the FCM clustering algorithm to generate clusters of size $C \geq 2$. In the second step, we applied a test to prove the presence of significant differences between mean values of variables in two separate clusters. For example: for $C = 3(C_1, C_2, C_3)$, we performed a test for difference between mean values of variables in $C_1\ C_2, C_1\ C_3$, and $C_2 C_3$. We run the test on the variables for the hours of the day where we think the variation in those hours would matter the most for the planning of a city and its transportation system: day time (7 AM to 5 AM) and night time (6 PM to midnight). The threshold for the number of clusters is reached as soon as the test does not provide significant differences between means of more than four variables of two separate clusters (in this case, each variable represents hourly cellphone data normalized over time). We found that $C = 2$ gives distinct activity patterns. We were interested in identifying varying activity patterns, but we noticed that when $C > 2$ is used, the resulting activity patterns are similar in some pairs of clusters in most hours of the day.

We validated the clusters obtained through cellphone data by comparing them to ground truth information, which is presence of people data. The presence of people data was deduced from OD data obtained in a mobility survey and it provides the number of people present at each grid-cell aggregated every hour. For a given grid-cell, we normalized the data over time and prepared a total of 24 values. We then carried out a second clustering analysis of the same grid-cells based on this variable.

Figure 4.4 illustrates average activity pattern in each cluster, which is obtained from cellphone data and the presence of people (ground truth). Observing Figure 4.4a, the first cluster has a lower first peak in the late morning and the maximum usage occurs during the evening, which suggests that the area is predominantly residential. The second cluster has

its first peak in the late morning followed by a very small dip at noon, and regains the second peak in the afternoon, and a decrease in the activity from around 6 PM on. This cluster represents a pattern related to an area with offices or/and commercial activity and we call it predominantly nonresidential.

It is obvious that more than two patterns do exist. However, most of the patterns are overshadowed by the two major patterns. For example, within the domain of predominantly nonresidential areas, there are commerce, touristic and office areas, which certainly share higher levels of routine or habitual activities. A previous study by Csáji et al. (2013) in the same region also obtained similar results. The study used anonymized individual cellphone traces to extract frequently contacted cell towers. Then, k-means clustering was performed to group aggregated calls from these towers and resulted in three different groups: two distinctive patterns of usage, home and office, and "others". However, the extracted patterns were not assigned to a specific urban area; instead, each tower location was related to home, office, and "others". In addition, the results in this study were not validated with clusters obtained from external datasets.

Comparing the clusters of the presence of people (Figure 4.4a) and the clusters with the cellphone data (Figure 4.4b to Figure 4.4d); there is a good similarity in the general patterns. However, there are also times of the day when the two sets do not match. For example, for the case of predominantly residential areas, the ground truth demonstrates an abrupt increase at 7am, but cellphone patterns do not exhibit this tendency. Another difference is that the proportion of ground truth data is unchanged in the morning period which is not observed in the cellphone data. These differences can be originated from the fact that cellphone use strongly depends on time. For example, there are some times of the day when users are more likely to make calls, but they do not call in others, such as late night or early morning. Nevertheless, we are not searching for a match between the two datasets; rather we hypothesize that the cellphone usage data may be able to distinguish between the two types of areas that are identified in the people's presence data.

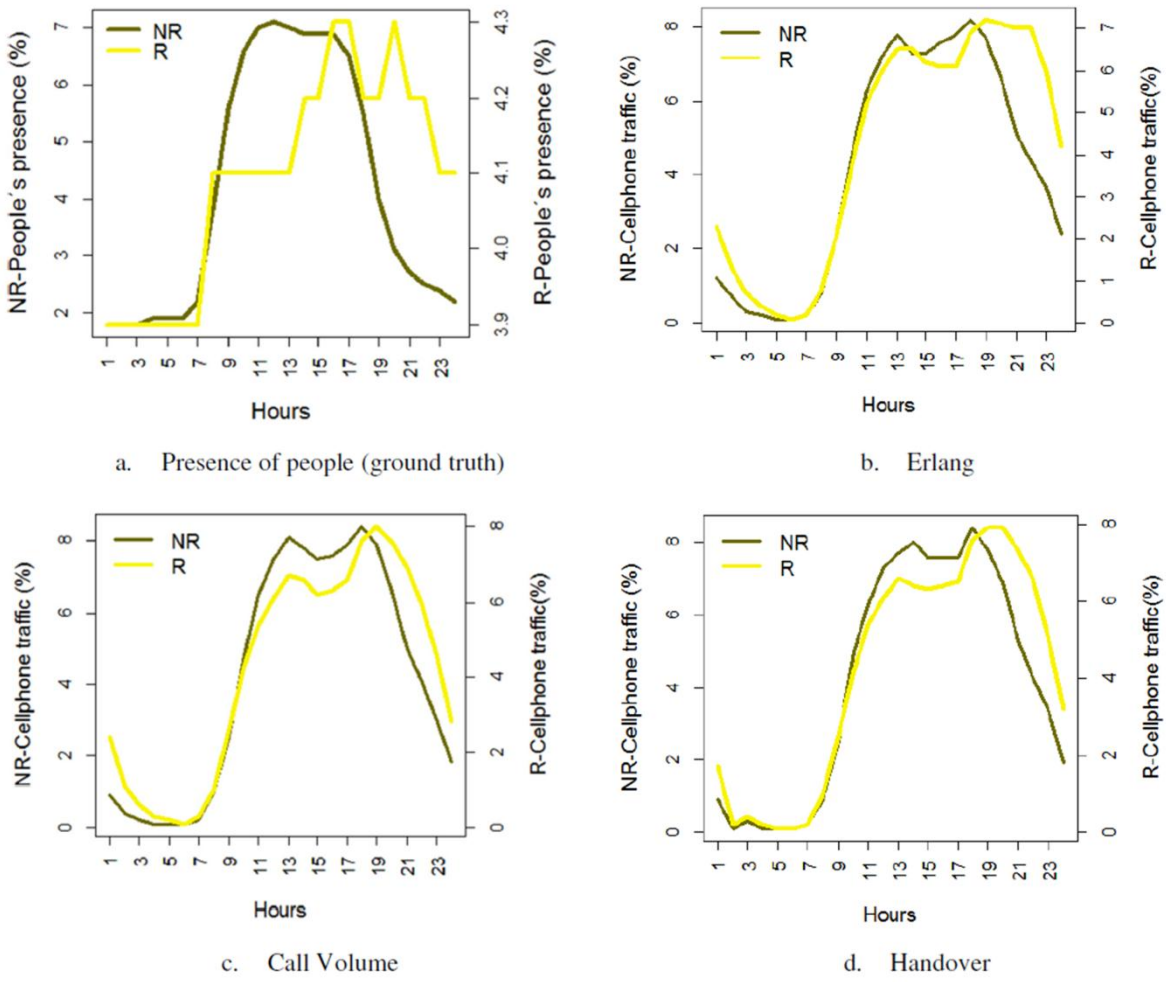


Figure 4.4 Patterns of average activity at the predominantly residential (R), and nonresidential (NR) areas

As we referred before in fuzzy c-mean clustering, each grid-cell belongs to each cluster, to the degree given by the membership value. Therefore, to assign a grid-cell to a specific cluster, we need to define a threshold. In the case of two clusters, the threshold is set to 0.5. Thus, in order to form two hard clusters, a grid-cell will be assigned to a given cluster if it has a higher degree of belongingness to that particular cluster (> 0.5).

Table 4-1 presents a comparison between activity patterns of grid-cells obtained through cellphone usage and their corresponding ground truth, which is presence of people information. Through ground truth, there are 66 grid-cells classified as predominantly residential and the other 52 grid-cells as predominantly nonresidential. The use of Call Volume provides a 62% and 72% match with the ground truth in the case of predominantly residential and nonresidential areas respectively. However, there is also 38% of predominantly residential area classified as predominantly nonresidential area. Out of the

three different types of cellphone data used to detect the activity patterns, Erlang shows a better agreement with the ground truth giving an overall correct classification accuracy of 69% of the grid-cells (the overall correct classification is used to measure accuracy, which is the ratio of correct predictions to the total number of predictions). On the other hand, compared with the other variables, the use of handover provides the highest match with the ground truth in terms of classifying the predominantly residential areas.

Table 4-1 Accuracy of FCM clustering algorithm: Patterns of cellphone activities

Observed		Area types through ground truth	Predicted		Overall (%)
			Predominantly residential (%)	Predominantly nonresidential (%)	
Call Volume	Predominantly residential	66	62	38	66
	Predominantly nonresidential	52	28	72	
Erlang	Predominantly residential	66	65	35	69
	Predominantly nonresidential	52	26	74	
Handover	Predominantly residential	66	68	32	68
	Predominantly nonresidential	52	32	68	

Figure 4.5 shows the geographical distribution of predominantly residential and nonresidential areas as calculated from presence of people data (Figure 4.5a) and from the cellphone data (Figure 4.5b to Figure 4.5d). Comparing the first map with the other three, we see that the geographic distributions are similar, especially at the city center dominated by nonresidential entities indicating a large concentration of office and commercial activities. We did not compute the pattern and intensity of the blank grid-cells, they are either not built up areas such as unused land, forests, bodies of water, and land used by airports (thirteen grid-cells) or have no cellphone data (five grid-cells).

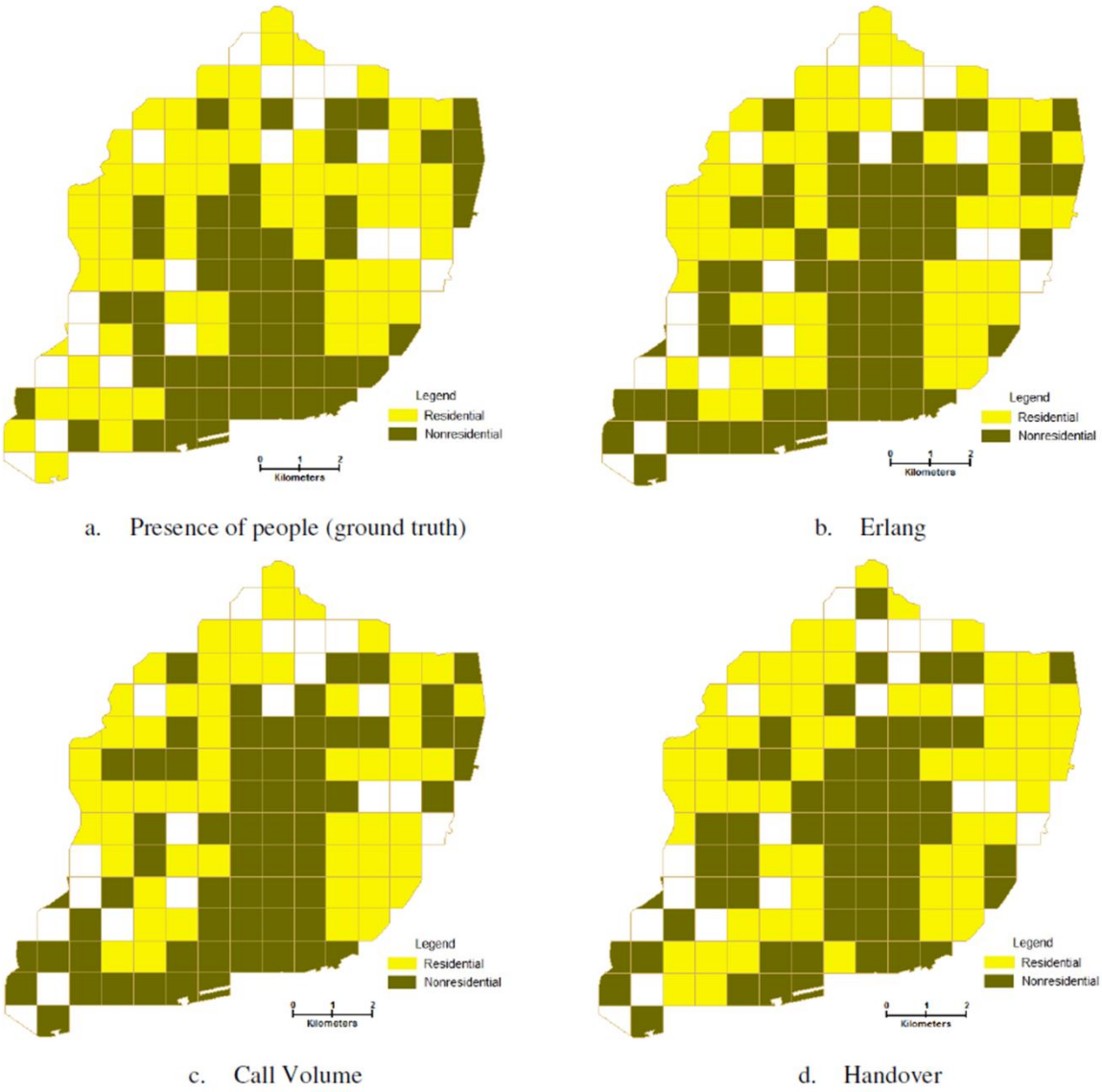


Figure 4.5 Geographical distribution of predominantly residential and nonresidential areas in Lisbon (hard clusters)

Though Figure 4.5 provides plenty of insights, assigning discrete type of land use may cloud the exact patterns of human activity. In Figure 4.6, we represent urban areas for both clusters simultaneously with different degrees of membership as calculated from presence of people data (Figure 4.6a) and from cellphone data (Figure 4.6b to Figure 4.6d).

Comparing the maps we may see that the geographic distributions are similar. Taking the comparison between Figure 4.6a and Figure 4.6b as an example, both maps reveal the predominantly nonresidential character of some areas in the city, such as locations C, and D. These places are the heart of Lisbon with a concentration of shopping streets in traditional neighborhoods of Baixa and the surroundings with its office and hotel services

that leave the city much deserted at night except from vehicular traffic. Locations A, and B are examples of predominantly residential neighborhoods. These urban areas belong to parishes, such as Benfica (Location A), which is a large residential area situated north of the city and Santa Maria dos Olivais, which is an upscale residential neighborhood (Location B). Additionally, mixed land uses are displayed by residential areas of Ajuda and district of Restelo which is also home to Lisbon´s famous monuments and museums (Location E), and Location F, which is a traditional business center of Lisbon that also has a low number of residential buildings.

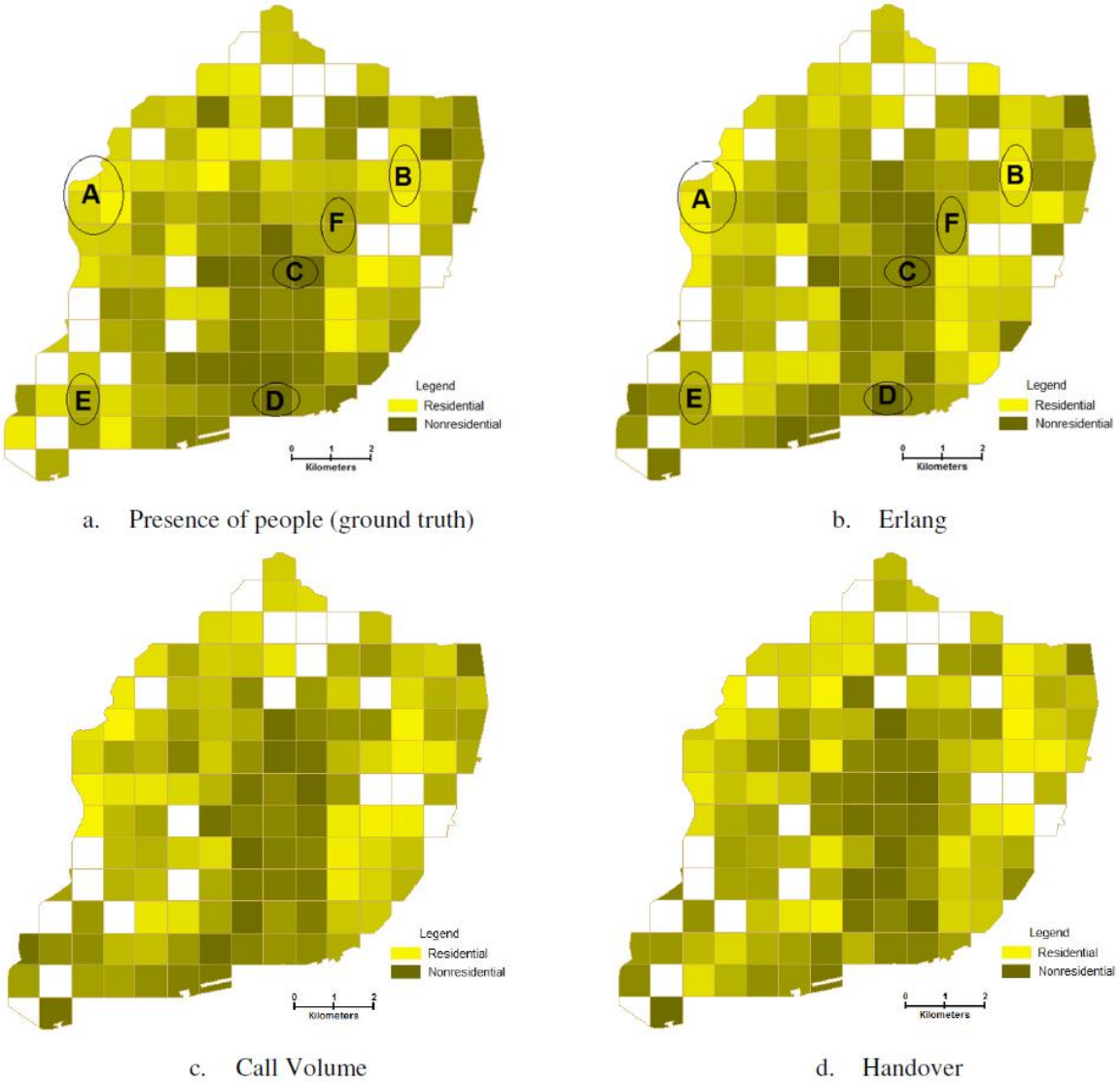


Figure 4.6 Geographical distribution of predominantly residential and nonresidential areas in Lisbon (soft clusters)

Knowledge of land use patterns is an important factor to understand the spatial distribution of urban activities as it is the preeminent index available to know the activities which people undertake in each part of a city. Observing Figure 4.5 and Figure 4.6, the CBD is dominated by nonresidential areas and the outskirts are mainly residential. This strengthens the fact that different activities have their own peculiar locational requirements, and this encourages people to travel in order to get from one type of activity area to another (Berke et al., 2006). However, knowledge of the uses only provides part of the picture. Within a given land use type, there are variations in the intensity of activities. To make our analysis more useful for transportation and urban planning, in the next section we explore the measurement of the intensity of activities, which influences the likelihood of trip making to a given place.

4.4.2. Analysis of the intensity of urban activity

In Section 4.4.1 we performed analysis of the patterns of activities in the city. However, this analysis is not suited to be used to detect the intensity of activities of a given urban land use. For a given grid-cell, we define the intensity of cell-phone use as its share of the total calling activity of the city for each hour of the day. Hence, unlike the analysis in Section 4.4.1, in this case we normalized the cellphone usage over space and we intend to relate that with the intensity of activity using ground truth computed with the same normalization procedure defined in Eqn. 4.2.

We compute the intensity of cellphone use for each grid-cell for the previous three different cellphone data: Call Volume, Erlang and Handover. A combination of Handover and Erlang variables is also considered. To combine the two variables: first, we computed the share of both Handover and Erlang values at each grid-cell relative to their respective total Handover and Erlang values of the entire grid-cells at a certain hour; then, we took the average of these values at each hour of the day. Thus, the input dataset consists of 118 grid-cells $\times$ 24 hours cellphone values. Each of the 24 values represents hourly cellphone usage, which is normalized over space.

The ground truth for the study of activity intensity is composed of nine different variables. Within the POI group, we have five categories representing different facility types (service, recreation, education, health, and transport facility locations). For each grid-cell we summed up the number of points for each POI category. The other ground truth

variables include residential buildings, presence of people, bus movement, and taxi movement. The values of POIs and residential buildings do not change over time. In the case of the presence of people, taxi and bus movements, we used the total daily statistics for each grid-cell. Thus, each of the ground truth is represented by a single input variable, which is normalized over space. We used nine datasets, one for each ground truth (118 grid-cells × 9). The ultimate goal is that the ground truth composed of different urban aspects can be used to emulate the characteristics of urban areas.

We applied FCM clustering algorithm to cluster the cellphone data and the ground truth associated to each grid-cell. In order to find the optimal number of clusters we applied a statistical method based on the comparison of two sample means that we explained in Section 4.4.1. In the case of the ground truth, we found that cluster size of more than two does not give statistically significant differences between the mean values of the clusters. Thus, we based our analysis on two clusters. We also used the procedures in Section 4.4.1 to compute the fuzziness coefficient value of 2.4 for the ground truth data, 2.3 for the Handover data, 2.4 for the Call Volume data, 2.3 for the Erlang data, and 2.2 for the combined use of Handover and Erlang data.

Figure 4.7 illustrates the intensity of average cellphone activity in each of the two clusters: high and low activity areas. Unlike the cellphone data (118 grid-cells × 24 hours of cellphone data), the input ground truth data for clustering do not have hourly statistics, thus, we do not have a corresponding graph along with the cellphone usage. The results in Figure 4.7 are important indicators of the density of human activities taking place along the day at locations within Lisbon. Taking the Call Volume variable as an example the average percentage change in calling activity, between 7am to 11am is increased by 16.16% in the high activity areas and reduced by 19.36% in the low activity areas; at 1pm it is reduced by 0.91% in the high activity areas and increased by 1.40% in the low activity areas; between 2pm to 3pm , a 1.47% increase in the high activity areas and a 2.22% decrease in the low activity areas; and between 5pm to 9pm, a 11.39% reduction in the high activity areas and a 15.64% increase in the low activity area (Figure 4.7a).

On both the low and the high activity areas, the major change in the cellphone usage occurs between 7am to 11am and 5pm to 9pm, which in fact is well associated with the time periods for traffic rush hours. However, it is interesting to note that the changes in the calling activity are reverse at the high and low activity clusters clearly denoting two

different areas. The average percentage change in the calling activity along the day is suggesting the general inward and outward movement of people and vehicles from residential to nonresidential areas and back again.

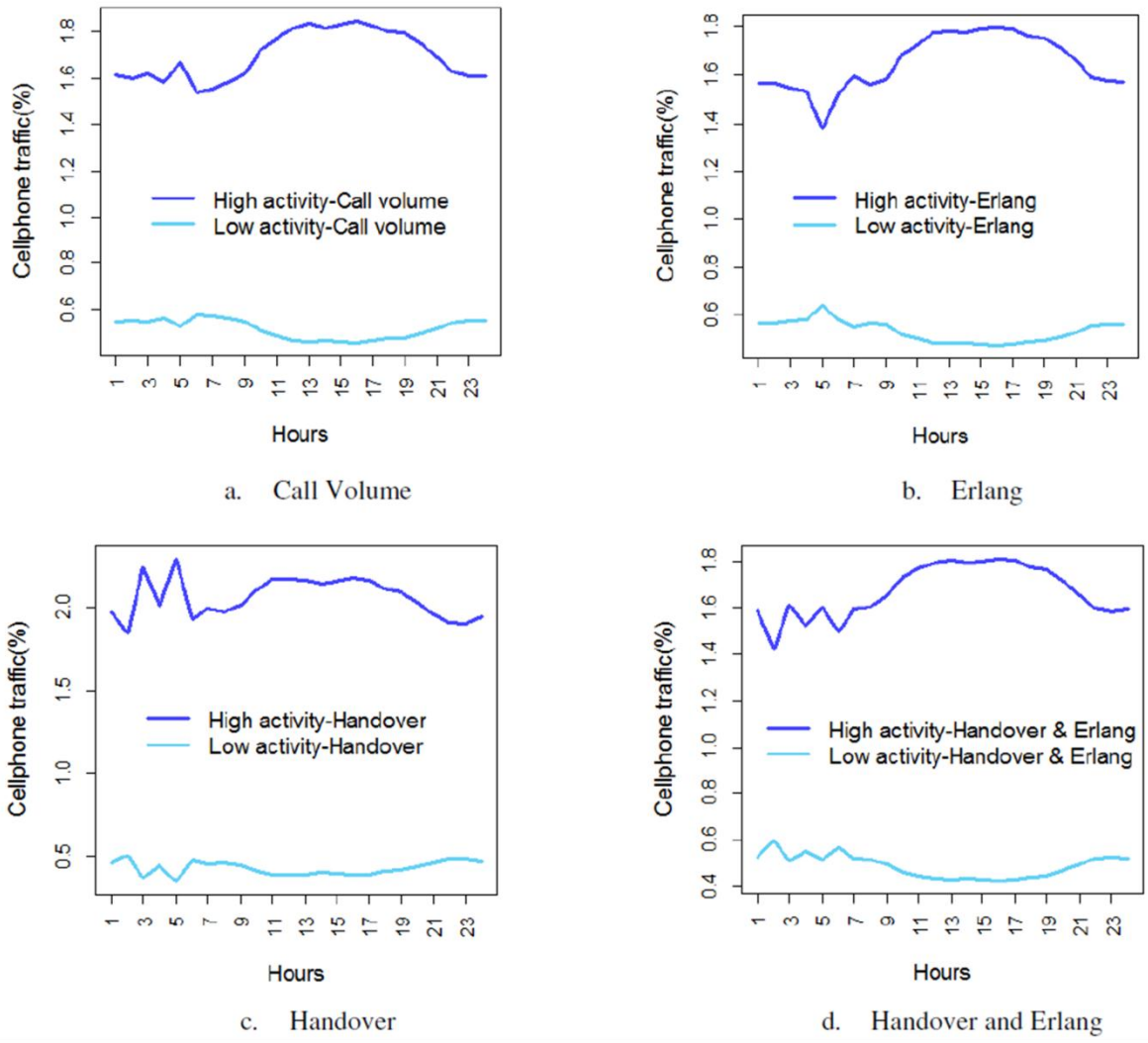


Figure 4.7 Intensity of average cellphone activity at the high and low activity areas obtained from cellphone data

Table 4-2 presents a comparison between activity levels of grid-cells obtained through cellphone usage and their corresponding ground truth information. Through clustering we identify two major types of grid-cells with high and low levels of activities. Through ground truth, there are 86 grid-cells classified as low activity area and the other 32 grid-cells as high activity area. As we had assumed, intensity of cellphone activity associated quite well with the ground truth. Out of the four different types of cellphone data used to

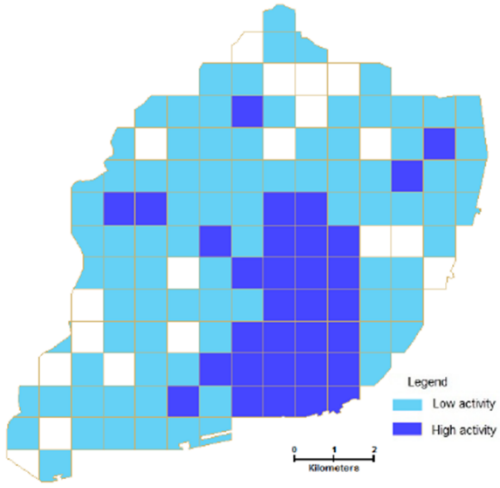
infer the activity levels, Handover, provides the highest overall correct classification accuracy of 80%. Observing the low and the high activity clusters separately, estimates through all the cellphone variables match better with the ground truth in the low activity cluster. We also tried different combinations of cellphone data to improve the results and we found that the mixed use of Handover and Erlang data provides better accuracy in terms of classifying the the high activity areas.

Table 4-2 Accuracy of FCM clustering algorithm: Intensity of cellphone activities

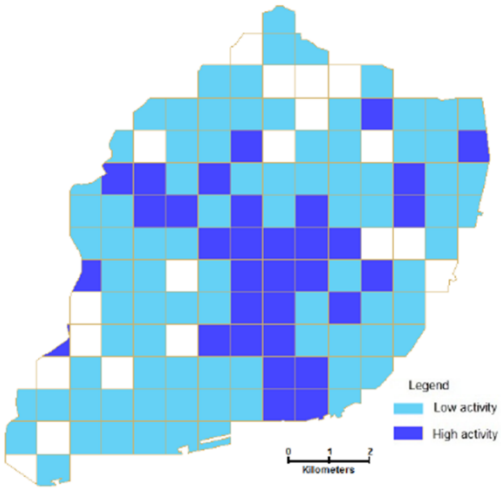
Observed		Activity levels through ground truth	Predicted: FCM		Overall (%)
			Low activity (%)	High activity (%)	
Call Volume	Low activity	86	83	17	75
	High activity	32	44	56	
Erlang	Low activity	86	81	19	74
	High activity	32	47	53	
Handover	Low activity	86	87	13	80
	High activity	32	41	59	
Handover and Erlang	Low activity	86	83	17	78
	High activity	32	34	66	

Figure 4.8 illustrates the geographical distribution of the high and low activity clusters in Lisbon. Figure 4.8 helps to visualize how clusters obtained through cellphone usage (Figure 4.8b to Figure 4.8e) are closely linked to clusters that are obtained through ground truth information (Figure 4.8a). It also provides validation, where clusters that represent different activity levels actually have different geographical locations. For example, most high activity grid-cells are situated at the city center, which is the venue for most offices and commercial activities. This provides important information for urban planners who have keen interest in gauging human activities over time and space to convey and improve public services in a city.

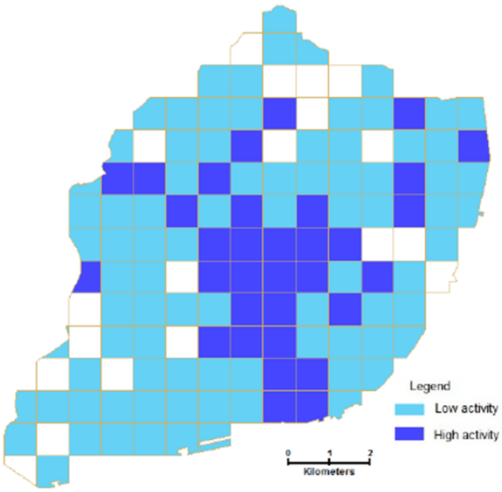
The maps in Figure 4.8 show a visual link between cellphone usage and city activities. The discrete nature of the hard clustering causes grid-cells to belong to a single cluster. However, some grid-cells might exhibit a mixed behavior of both high and low activity levels along a day. Thus, we used the soft clusters that would allow grid-cells to belong to both clusters simultaneously with different degrees of membership. Figure 4.9 shows the geographical distribution of grid-cells with their high and low activity membership corresponding to a probability of belonging to a specific cluster.



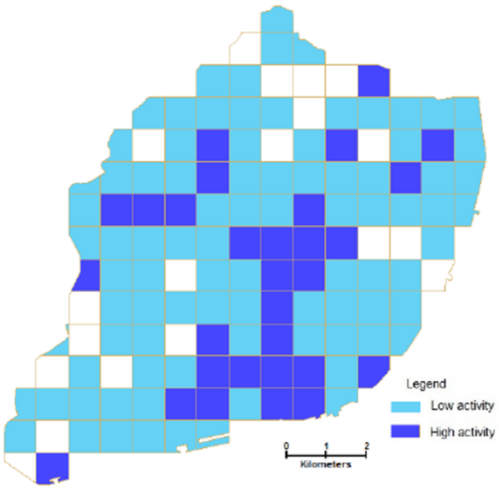
a. Ground truth



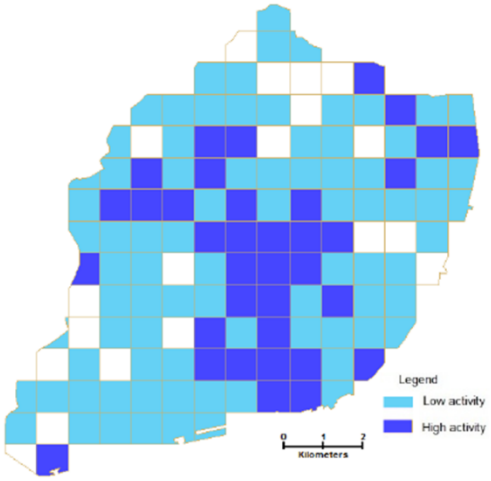
b. Call Volume



c. Erlang



d. Handover



e. Handover and Erlang

Figure 4.8 Geographical distribution of the high and low activity clusters in Lisbon (hard clusters)

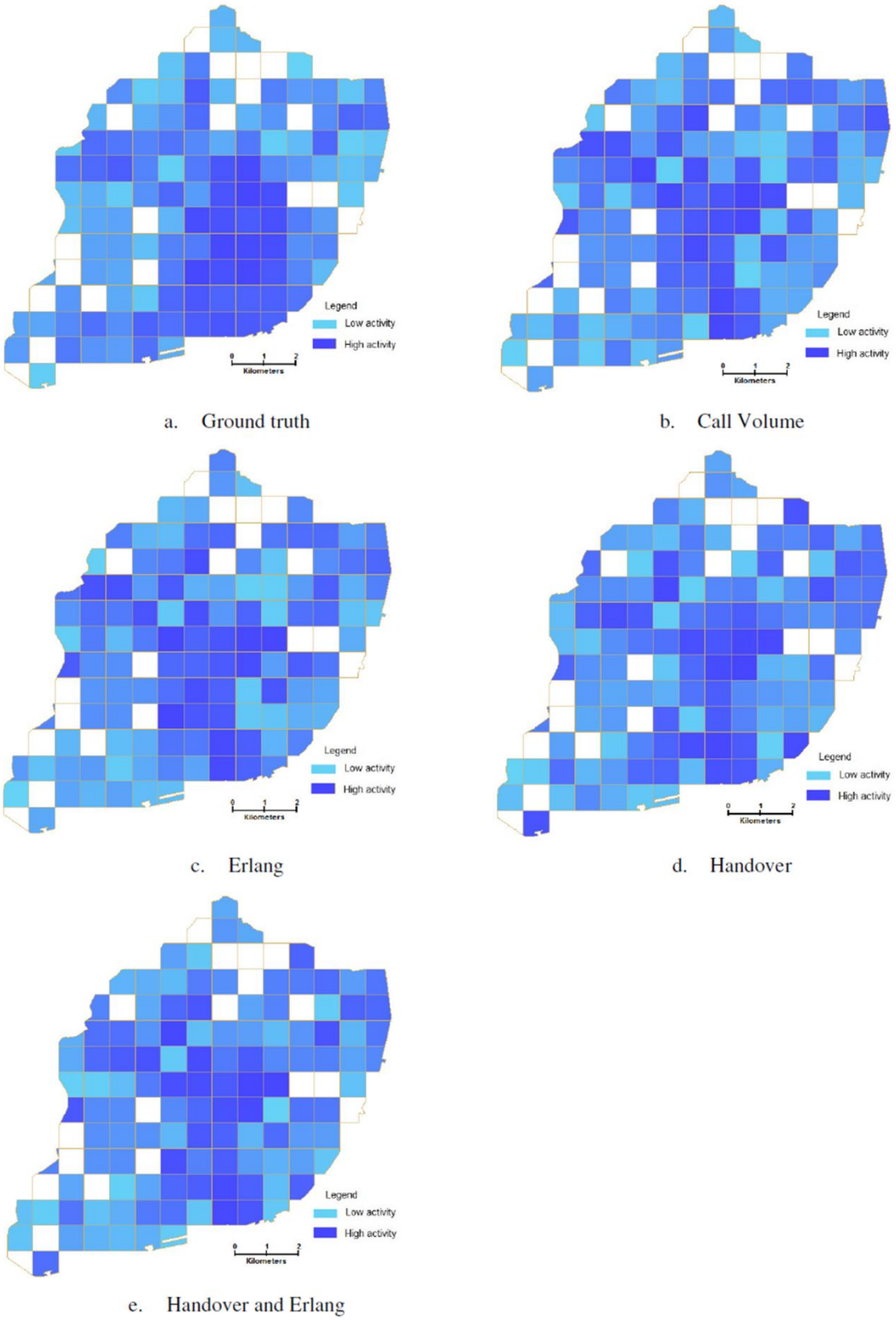
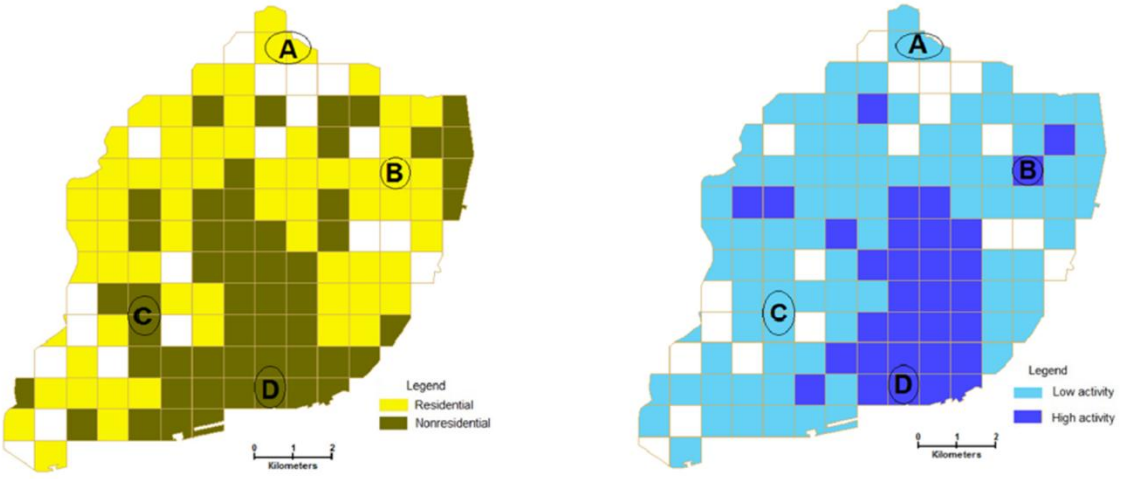


Figure 4.9 Geographical distribution of the high and low activity areas in Lisbon (soft clusters)

4.4.3. Combining the pattern and intensity of urban activity analyses

We showed that the pattern of cellphone usage can be used to specify the nature of urban land use (predominantly residential and nonresidential area), and the intensity of cellphone usage can be used to specify the intensity of urban activity (high and low activity area). Combining the two analyses, the pattern and intensity of urban activity yields interesting results.

Figure 4.10 shows the reality that within the same urban land use type, we can have different intensity of urban activities. Figure 4.10a shows land use categories and Figure 4.10b shows intensity of urban activity of the same locations, both calculated through ground truth. Locations A, and B are examples of areas with predominantly residential uses, but with different intensity of urban activity. Location A, which is in the northwestern side of Lisbon´s Portela Airport, is accommodating low population and residential building density and it is characterized as low activity area. On the other hand, Location B is an upscale residential neighborhood with a high activity area. An enlarged view of location B is shown in Figure 4.11b. Location C, and D are examples of areas with predominantly nonresidential uses, but with different intensity of urban activity. Location C is an urban area which is part of the biggest park in the city being characterized as a low activity area. An enlarged view of location C is shown in Figure 4.11a. Location D belongs to Lisbon´s traditional neighborhood accommodating shopping areas and restaurants, which attracts many tourists and it is characterized as a high activity area.



a. predominantly residential and nonresidential areas (ground truth)

b. high and low activity areas (ground truth)

Figure 4.10 Combining the pattern and intensity of urban activity analyses

a. low activity and nonresidential area

b. high activity and residential area

Figure 4.11 Examples of areas with different pattern and intensity of urban activity

4.5. Discussion, contributions, and limitations

Understanding urban activities and their dynamics is important for the urban and transportation planners. In this chapter, we showed how to use passive mobile positioning data to detect the pattern and intensity of urban activities. Specifically, this chapter makes the following contributions:

- In addition to Erlang and Call Volume, we tried the use of Handover data, which is mobility related events of the cellular network to characterize the pattern and intensity of urban activities. Previous studies also detect urban land uses on the bases of static information: through CDRs (Frías-Martínez, 2011; Toole et al., 2012); Erlang (Reades et al., 2009), and aggregate voice and SMS values (Becker et al., 2011).
- The main findings in our analyses reveal that pattern and intensity of activities in urban areas is related to cellphone usage. Previous studies also investigated the possibility of detecting urban activity patterns through cellphone data, which indicate how land is being used in different parts of the city (Frías-Martínez, 2011; Reades et al., 2009; Toole et al., 2012). However, knowledge of land use only provides part of the picture. Within a given land use type, there are variations in the intensity of activities. To make

our analysis more useful for urban and transport planners, we added the analysis on the intensity of urban activities.

- In an attempt to construct the characteristics of urban areas we integrated different types of datasets, which are existing mobility, activity, and business data in the city.
- Even though the use of cellphone data for urban and transportation planning has been validated by external datasets, the exercise has not been fully complete. Previously, some studies focused more on validating their results through comparing graphical representation of results from cellphone data against ground truth (Csáji et al., 2013; Becker et al., 2011b; Reades et al., 2009); some other studies performed validation through comparison of cellphone usage patterns on selected places of cities (Reades et al., 2007; Soto and Frías-Martínez, 2011). However, these approaches lacked derivation of quantitative information from ground truth for validation. In our validation process, we attempted to develop a city-wide system of measurements representing quantitative information and qualitative information (graphical representations) from both the cellphone data and ground truth.

Thus, our analyses provide plenty of opportunities to gain insights of how cellphone data can be applied for urban and transportation planning. The results could be used in a wide range of applications. For one thing results can be used to understand what is happening to an area with minimum cost as opposed to traditional methods, which are more accurate, but at the same time are usually more expensive and time consuming.

The models we developed could be employed, namely in regions such as developing sub-Saharan countries that are having high cellphone penetration as a cheaper and faster way to estimate the pattern and intensity of activities. Given the mobility needs and the development concerns of the sub-Saharan African cities, the challenge for transport policy makers would be to devise strategies that maximize the gains in economic efficiency in the transportation sectors. Because of the progressive urbanization and growing economies, most of these cities are planning to implement improved public transportation systems. One of the major problems associated to this kind of expansion would be estimating and forecasting public transport demand through conventional methods that require massive data collection work such as, expensive travel surveys, demographic, and land use characteristics (Mullen, 1975).

In our analysis we showed how cellphone activity can be used to estimate the patterns and intensities of activities. In the absence of comprehensive travel survey data, transport planners should take advantage of our results as an alternative or as a complementary method to the traditional transportation planning procedures. The predominantly residential and nonresidential land uses along with the intensity of use computed in our analysis provide the type and magnitude of human activities taking place at different locations within a city. Transport planners can use these results as a starting point to analyze the trip generation and distribution stages of the traditional four-step transportation forecasting model.

In a developed city, the existing land use pattern may be expected to hold through time. However, there are situations that cause changes in the urban use over time. Therefore, planners should perform frequent estimate of intensity and patterns of activities and measure the likely consequences of the changes upon transportation uses. The results in our study provide bases to carry out such estimation and an interesting approach should also be using our results as means for updating traditional urban data, which are usually made available less frequently to urban planners and policy makers.

We should emphasize two important limitations of the study. Firstly, despite TMN´s share market of 45% in Portugal, in principle; our analysis is only applicable to subscribers of TMN. In order to associate our findings to all inhabitants of Lisbon, we need to assure that our sample is representative of the current distribution of people in the city. However we do not have reasons to believe that our sample should be biased since there is no indication that there is a special preference of a certain type of population segments for one particular network.

Secondly, in urban environments, a mobile device can reach out multiple base stations at the same time, and selects a base station with the strongest signal and best signal quality for transmission (Küpper, 2007). Thus, it is not always the case where the call activity handled by a particular cell represents the actual people´s activity in its vicinity. Another bias could rise from the cellular network size and coverage representation. Alternative methods that imitate the real coverage of cellular towers (e.g. grid areas) would undermine the outcomes of the studies in this field.

4.6. Summary

Over the past decades there has been an explosion in the adoption of communication devices like cellphone, which generate data that can provide new opportunities for urban analysis. In this analysis we have explored the use of passive mobile positioning data, which is automatically stored in the log files of service providers, most specifically: Call Volume, Handover, and Erlang with the objective of detecting intensity and pattern of activities in an urban area. Currently, urban planners measure activity related variables via surveys. Though this method is time consuming and expensive, it has the advantage of providing detailed information, in contrast, cellphones reside with people and travel along with them as they move from place to place, thus offering the potential to sense the location of people and describe the city dynamics, their patterns, and evolution over time.

Unlike most urban studies that detect characteristics of places via surveying individual cellphone usage, we focused on analyzing the behavior of places through their aggregate cellphone usage rather than individuals. This information is anonymous thus easier to be obtainable. We raised research questions of relevance to urban transport planners, transport geographers, and urban planners: Is it possible to use cellphone data to detect intensity of urban activities in a city? And are we able to explain the varying activity patterns along the day in different parts of a city?

We tried to address these two interrelated questions by exploring cellphone usage over the city of Lisbon. As the available datasets have different spatial resolutions, we built a grid layer with cells of 800 by 800 meters over the city to convert all the datasets to a uniform grid. For validation purposes, we developed the characteristics of activity locations with indicators associated to spatial interactions, movement and distribution of population, which consists of presence of people, POIs, residential buildings, bus movement, and taxi movement.

We employed a FCM clustering algorithm that uses no prior information regarding the activity profile of the grid-cells. Two different experiments were carried out. In the first experiment, we performed analysis of the patterns of activities along the day. We applied a normalization of the cellphone usage over time, which shows the intensity of each hour's cellphone usage relative to its daily total usage. In practice, different grid-cells in a city would relate to different patterns of activities which indicate the presence of different population sizes at a given time in the grid-cell. Out of the three cellphone variables that

were used to infer the activity patterns, Erlang shows a better agreement with the ground truth giving an overall correct classification accuracy of 69%. In the second experiment, we performed analysis of the intensity of activities along the day. The interesting aspect of this analysis was observing how urban areas can be partitioned into high and low activity areas based on their share of the total cellphone usage in the city. This approach gives us a clue regarding where in a city we have high/low intensity of activity at a given hour. From the four cellphone variables that were used to infer the activity level, Handover provides the highest overall correct classification accuracy of 80%.

In spite of some of the limitations associated to our study, the analyses presented in this chapter show plenty of opportunities to understand the pattern and intensity of activities in urban areas. The results could be applied for urban and transportation planning. In particular, results from the analysis of predominantly residential and nonresidential areas can be consulted to approximate the connections which are made by trips between origins and destinations. These connections play significant role to estimating travel demands and streets and mass transportation routes.

We attempted to demonstrate the presence of a relevant link between cellphone usage and attributes that define spatial organization of a city such as, mobility, land use, and distribution of population indicators. The results confirm the existence of a relevant relationship between cellphone use and the ground truth that is worth exploring further.

Chapter 5 Intelligent road traffic status detection system through cellular networks handover information

5.1. Introduction

5.1.1. Background

The growing development of various Intelligent Transportation Systems (ITS) schemes needs comprehensive high quality information. However, obtaining heterogeneous road traffic information is still one of the key challenges of ITS. Traffic management sectors use several techniques to gather raw traffic information that basically fall into two major groups: point-detection and vehicle-based detection systems (OECD, 2007). A single set of traffic data collection system has functional limitations; that is, it does not offer the amount of data required to have a realistic and comprehensive view of the traffic stream in urban areas. Successful deployment of ITS schemes requires installation of large numbers of data collection points and deployment of a substantial amount of probes into the traffic stream to monitor the traffic status and detect bottlenecks in their early stage (Calabrese et al., 2011).

Arterial roads in an urban environment are used to conduct the analyses in this chapter, and traffic volume is the preferred parameter to characterize the state of traffic conditions. Traffic volume and flow rate are measures that quantify the amount of traffic traversing a point in a roadway system per unit of time. The duration of the traffic count defines the type of traffic volume: annual, daily, hourly, etc. Flow rate represents the equivalent hourly rate of vehicles traversing a roadway system during a time interval of less than 1 hour (TRB, 2000). Annual average daily traffic and average daily vehicle distance traveled are the two most used traffic statistics mainly for traffic planning purposes. However, traffic managing authorities should also collect hourly traffic volumes to look up to the operational characteristics of the road at different times of a day (FHWA, 2001).

The state-of-the-practice for data collection regarding traffic volume depends on either human observation or different forms of remote sensing (loop detectors, automatic video feed-based counts, etc.). The deployment of these conventional on-road sensors for traffic

data collection is indispensable, but because of their expensive installation and maintenance costs, they are only available in a limited part of the road network. In the absence of a traffic count from a specific site, prediction is usually made through the average traffic from other places or historical traffic data of a given location. However, this forces traffic management authorities to rely on an incomplete picture of the traffic stream in the city (OECD, 2007).

The growing pressure to improve traffic management services has brought alternative information sources to the road managers attention (Leduc, 2008). One way that has been tried to obtain this information is through the use of cellular networks. In comparison to on-road sensors, cellular networks provide mobility related events that can be obtained during conventional operation, such as Location Area (LA) update, Route Area (RA) update, and cell update (handover) (Valerio et al., 2009b). A cell is delimited by the area covered by a base station also called base transceiver station in the Global System for Mobile Communications (GSM) and node B in the Universal Mobile Telecommunications System (UMTS). Cells within a network are grouped into LA for the circuit switched GSM network, and RA for the packet switched networks (Valerio et al., 2009b). Handover is a cell based location update, which is the process of transferring an ongoing call or data session from one area to another without loss or interruption of service (Zeng and Agrawal, 2002). The LA update process can be initiated periodically or, while the cellphone equipment crosses the boundary of the LA.

Traffic estimation has been an important issue along the last decades. Greenshields conducted one of the first empirical studies on measurement of traffic volume, traffic density and speed through a snapshot of traffic by an aerial camera in the 1930s (Kühne, 2008). A recent study by Heydecker and Addison (2011) also investigated the relationship between speed, flow, and density that showed the direction of causality between speed and density differs on different circumstances. The development of ITS introduced new ways of obtaining road traffic data from alternative sources. The study by Varaiya et al. (2008) used wireless magnetic sensor networks to detect the presence and movement of vehicles in real time. Herrera et al. (2010) and Ahas et al. (2010) used GPS-equipped phones as probes to gather mobility related information within a cellular network.

In the recent years, studies regarding the analysis of data obtained from cellphone use to estimate traffic conditions have been carried out and resulted in a number of commercial

products (Cellint, 2007; INRIX, 2012). Valerio (2009) presented a brief summary of different projects and commercial products regarding cellular network use for traffic estimation. Studies by Ratti et al. (2006) and Reades et al. (2007) applied information from cellphone usage at a city-scale level to represent the intensity of urban activities and their evolution through space and time. Ratti et al. (2005) and Calabrese et al. (2011) also developed a city-scale study that shows a real-time representation of city dynamics through erlang, handover, and cellphone trajectories from registered users. The later study combines also the information from GPS equipped buses and taxis.

Alger et al. (2005) used information from double-handovers (handovers at the entrance and exit of a cell) for speed estimation. The data was collected from a 110 km long autobahn network, which consisted of 55 cells. The results of this study revealed a bimodal speed pattern with a double peak originated on Lorries at about 80-85km/h and cars at about 120km/h. Bar-Gera (2007) compared speed and travel time estimated from cellular network and loop detector data, which was collected from the Ayalon freeway in Tel-Aviv, Israel. Considering 20368 common time intervals (65% in the range of 8 to10 minute) from both sources during working days, the average absolute relative difference that was found is 10.7%, and the average absolute difference was 1.09 min.

Becker et al. (2011a) investigated the use of handover extracted from anonymized call detail records to estimate relative traffic volumes. The data was collected from 35 cell towers located in the center of Anytown, United States. The result showed a correlation coefficient of 0.77 between the handovers, and vehicle counts from loop detector. A result from a similar study by Vaccari et al. (2009) achieved an average correlation coefficient of 0.772, where the traffic flow data between the two boroughs of Brooklyn and Manhattan was considered for the analysis. The study by Thiessenhusen et al. (2003) also proved the coexistence of handover and traffic volume peaks during the morning and afternoon peak times.

Puntumapon and Pattara-atikom (2008) applied cell dwell time (the length of time that a mobile device remains registered to a base station until it switches to another base station) to develop a Naive Bayes model that differentiates pedestrian and sky train passengers by analyzing cellphone user permanence in a cell. Results showed classification accuracy up to 93.1%.

For traffic estimation methods that use handover information it is important that calls are being made while driving. However, this behavior differs in different countries. In 2005, observational studies on the drivers were carried out in the United States, Australia, and United Kingdom (UK) and the results showed that 1% to 4% of the drivers were using the phone while driving (Jeanne Breen Consulting, 2009). In the same study, the authors also referred an interview in the UK and in the Netherlands where the percentage of people who said that use their phones while driving was 36% and 50% respectively (Jeanne Breen Consulting, 2009). However, these studies overlooked the passenger phone usage and the increased usage of hands-free equipment's that is happening, allowing having more calls in a car.

5.1.2. Our approach

The most useful parameters to characterize the interactions in the traffic stream are speed, density and flow rate. The majority of past studies gave emphasis to the estimation of travel time, speed and traffic congestion through mobility related cellular network information. However, Bar-Gera (2007) had found 10% more noise on the estimated travel time through cellular networks information when compared against travel time estimated from loop detector data. Despite the massive amount of data that was used, Alger et al. (2005) and Caceres et al. (2008) found that the double-handovers data was rather too sparse in time and discovered a large variance in the estimated speed. An error of 20 to 30 km/h was observed when the speed estimated from the GSM network was compared against the speed estimated from probe vehicles and loop detector data (Thiessenhusen et al., 2003).

The study by (Becker et al., 2011a) claims the boundaries between cells are stable over different routes, speeds, direction and phone model conditions. However, other studies (Alger et al., 2005; Caceres et al., 2007; Smith et al., 2007) discovered spatial variability of cell boundaries, and it was considered to be one of the causes for low accuracy and error during travel time and speed estimation. Thajchayapong et al. (2006) and Hongsakham et al. (2008) applied cell dwell time information to infer congestion. However, this approach has the following problem: traffic stopped on a signalized intersection can be estimated as in a congested state even when there is no true congestion; due to the difference in the size

of the cell service area the cell dwell time from one cell may not represent the intensity of traffic in another cell region. Thus, even though many experts have argued the usefulness of the information from cellular networks for traffic estimation, some important issues remain to be addressed. These limitations have different effects based on the measured traffic parameters:

Sample size and cellphone position accuracy: Cellular networks produce large datasets, which can be generated in two ways: one is when the cellphone is on call, and the other is when the cellphone is on standby. Large amount of data can be obtained when the cellphone is on standby, but with less frequency and less location accuracy. Traffic volume estimation prioritizes availability of high sample rates, whereas, the preferred requirement for speed and travel time estimation is good positioning accuracy (Caceres et al., 2007; Valerio, 2009; Valerio et al., 2009a).

Legitimacy of cellphone data: The influence of cellphone use from pedestrian is severe in urban areas and may obfuscate the signature produced by calls from vehicles. Estimation that uses mobility related cellphone data should develop a way to isolate motorized and non-motorized clients.

Privacy issue: Some techniques require individual cellphone signatures in order to locate the cellphone at an individual level for speed and travel time estimation. This procedure threatens personal privacy, and cellphone operators should anonymize the data before being used for prediction.

Regardless of the important efforts in applying cellular networks information for the development of speed, density, and travel time estimation, by far cellular networks have been the less exploited source of information for the purpose of traffic volume estimation and some challenges are still to be addressed by the research community.

In this chapter, we explore the use of cellular networks handover information in order to complement the effort on traffic volume estimation. Unlike the previous experiments, which were limited to the analysis of the relationship between the handover and the traffic volume (Becker et al., 2011a; Caceres et al., 2007; Thiessenhusen et al., 2003; Vaccari et al., 2009), our experiments move forward the frontier of knowledge up to the level of

actually developing a model to estimate the road traffic status through handover information.

In addition to the advancements on the previous experiments, our analyses address the aforementioned limitations through different approaches. The handover data acquired in our analyses is hourly aggregated at a cell level that complies with the anonymity requirement on the used data as we base our approach on the number of handovers per base station and not on individualized handover information.

The occurrence of a handover event between cells within the same tower is more frequent because of the shorter distance required to cross the cell boundaries. Therefore, we filter these handover events with the objective of reducing the noise from the pedestrians' calls. We also give priority to the cell towers close to the road segment under analysis in order to consider the cells with information more relevant for characterization of a specific road.

The remainder of this chapter is organized as follows: Section 5.2 describes the case study area and dataset collection procedures. Section 5.3 presents the handover analysis and the developed models used to explain traffic through the handovers. In section 5.4 we provide detailed description of the results achieved by the models. In section 5.5, the chapter ends with the summary stating main conclusions and future research directions.

5.2. Data collection

We used Lisbon as our case study. Lisbon is the capital of Portugal and the center of the Lisbon Metropolitan Area (LMA). The LMA has a population of 2.3 million and has 18 municipalities with a total area of 2957.4 square kilometer, where about 24.3% of the population resides in the municipality of Lisbon (INE, 2013). Lisbon has a high number of cellphone users. According to statistics from ANACOM (2010), active mobile telephone cards per 100 Portuguese inhabitants grew to 159.9 by the end of year 2010, from 140.4 in the year 2008. This mobile penetration record was obtained through the combined services provided by the three major mobile service operators: Vodafone, Optimus and TMN, each having their own network infrastructure.

We selected 5 case study areas situated in various locations across the municipality of Lisbon. We selected these cases due to their diversity in terms of location, scale, and range

of mobility characteristics. In selecting the case studies, we considered whether the road had a high pedestrian movement or not. We also tried to select roads from the city center as well as from the outskirts of the city, and have selected roads with different average total daily traffic volumes. Figure 5.1 shows the location of the five case study areas: (A) Avenida Marechal Gomes da Costa, (B) Avenida da Igreja, (C) Avenida Joao XXI, (D) Avenida Almirante Reis, and (E) Avenida Engenheiro Duarte Pacheco and Rua Joaquim António de Aguiar. Geocoded points of the traffic counters available in the city are represented by circles and cellular tower locations by triangles.

Handover information was obtained from TMN Company. TMN has a share market of 45% in Portugal mobile service when compared to the other two companies: Vodafone and Optimus. In December 31, 2010, TMN had 7.42 million subscribers in Portugal. In addition to the increase in the number of subscribers, TMN´s voice traffic grew by 7.1% to 10.54 billion minutes in 2010, compared to 9.84 billion minutes in 2009 (ANACOM, 2010). TMN uses GSM and UMTS technologies to provide mobile communication services. At the end of 2010, TMN's UMTS population coverage was approximately 93%, and it was geographically available over 4194 municipalities out of a total of 4252 in Portugal (ANACOM, 2010).

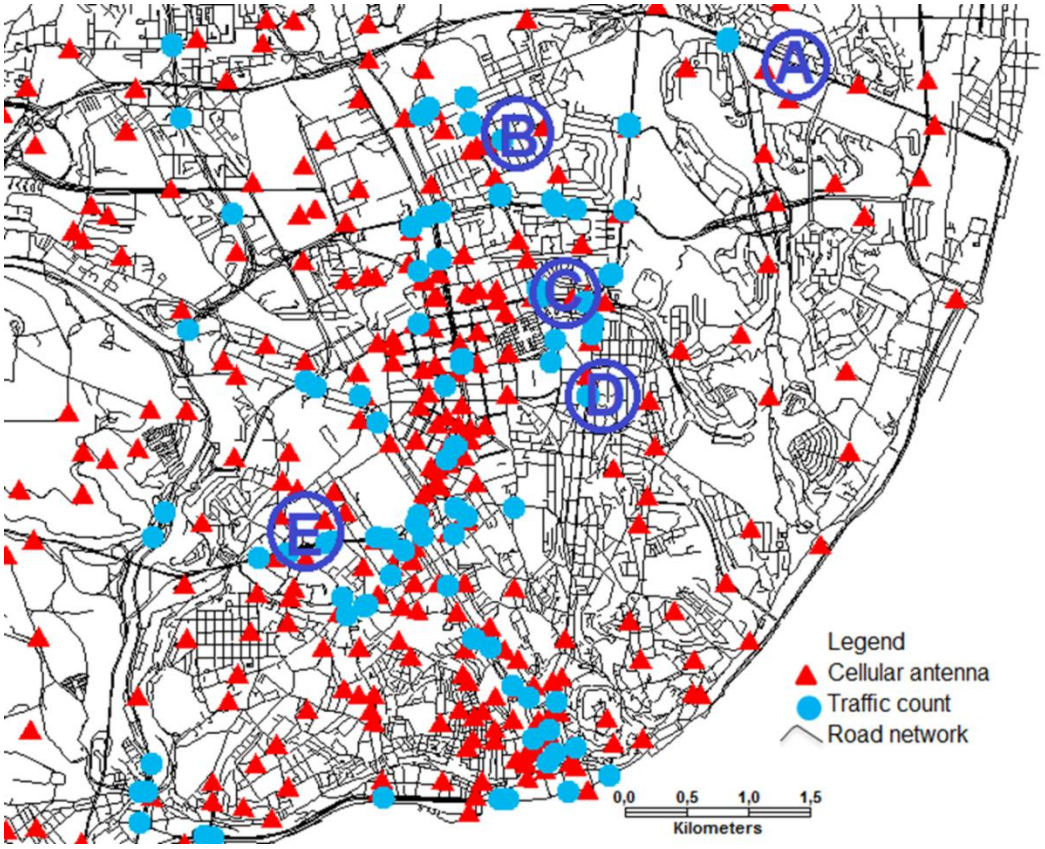


Figure 5.1 List of case study areas in the Municipality of Lisbon

We extracted handover counts from 39 cell towers in the five case study areas which have approximately 130 antennas pointed in the various directions (Figure 5.2). Our aim was to capture the cellular traffic in the vicinity of the roads and the number of cell towers chosen allowed us to cover the entire case study areas. The hourly handover counts were gathered from April 6, 2010.

Table 5-1 shows a sample of the handover data obtained from TMN, where calls are transferred from the source cell sector to a neighbor cell sector.

Table 5-1 Handover sample data

Source cell sector ID	Neighbor cell sector ID	Date	Hour	Handover from Source to neighbor cell sector
1110506612	1110509355	04-06-2010	3PM	92
1110506612	1110509355	04-06-2010	4PM	133
1110506612	1110509355	04-06-2010	5PM	120
1110506612	1110509355	04-06-2010	6PM	134
1110506612	1110509355	04-06-2010	7PM	109
1110506612	1110509355	04-06-2010	8PM	109

We also obtained the hourly traffic volumes from 101 traffic counters equipped with inductive loops that were made available to us by the municipality of Lisbon. The traffic volumes were obtained from a working day in same day as the Handover data, April 6, 2010. The average total daily traffic volume from the 12 traffic counters available in our case study areas is about 20500 vehicles. The arterial roads in Avenida da Igreja and Avenida João XXI have high pedestrian usage when compared to the remaining arterial roads chosen in the case study as they have considerable shopping activity.

Figure 5.2 shows a detailed representation of the cellular towers (triangles) and traffic counter locations (circles) on the five case study areas. The direction of the traffic movement is indicated with an arrow next to the counters considered in the study for model development and validation. In addition, Figure 5.2 shows the towers chosen to be related to the traffic counters in the study. Since we would have 12 pairs to show, for presentation simplicity we chose only those counters with the highest average daily traffic from each case study area for exhibition. Once the traffic counter is chosen, we give priority to the cell towers close to the road segment under analysis in order to consider the cells with information more relevant for characterization of a specific road. The other factor is availability of handover between the selected cells, which was carried out through a consultation of the handover database that we had available. The selected traffic counters in all the cases are labeled as T and the corresponding cell towers are labeled as C1 and C2, where the flow of calls is from C1 to C2.

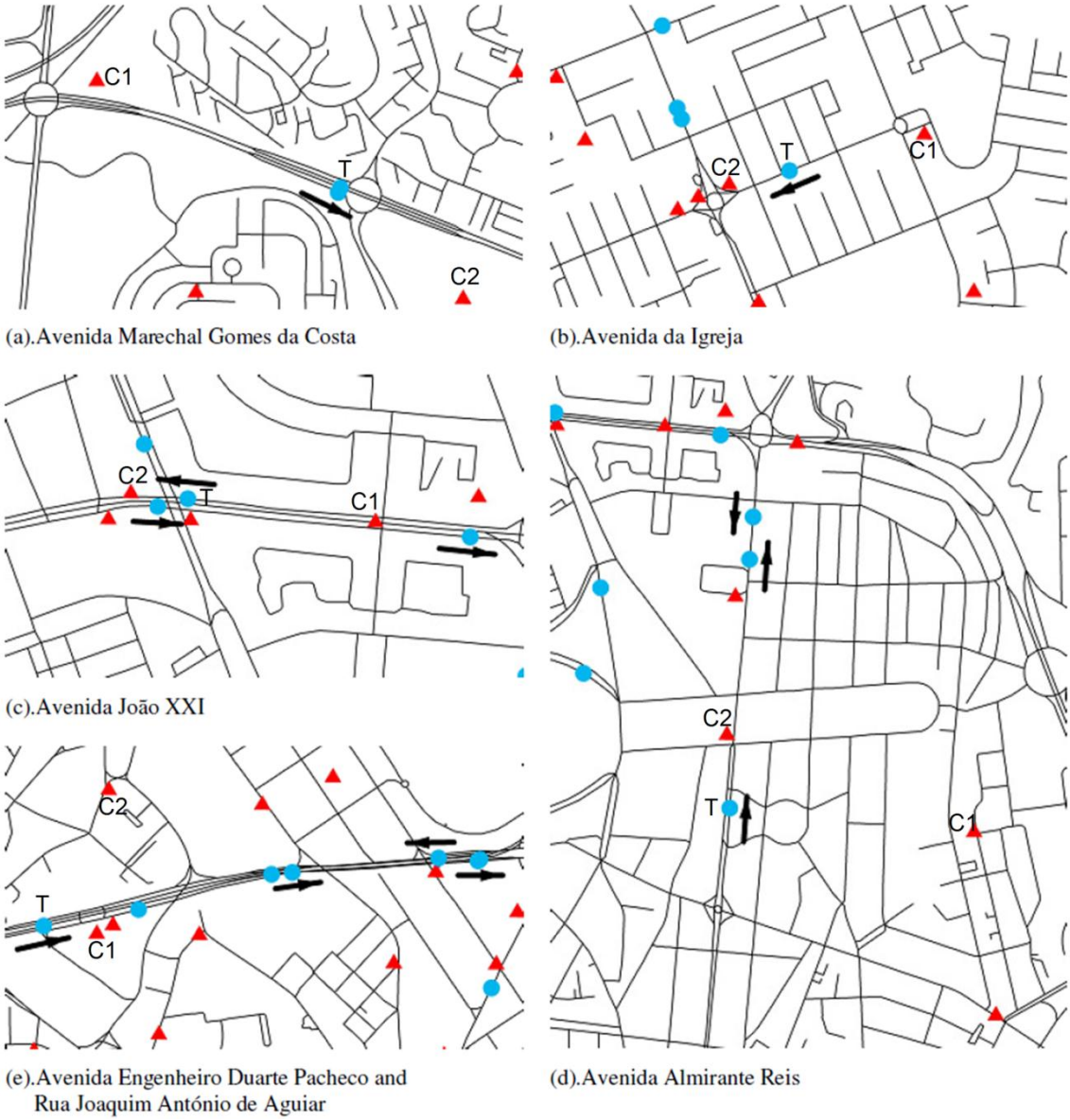


Figure 5.2 Detailed representation of cellular towers and traffic counter locations on the five case study areas

5.3. Predicting traffic through handover counts

5.3.1. Handover analysis

The handover process is triggered by the operational rules configured in the cellular networks. As a simplification the cells are represented as hexagonal areas (Figure 5.3).

However, the dimension of the cells depends on the geography of the area and cellular use within the cell (Zeng and Agrawal, 2002).

The handover process is represented through the schematic diagram in Figure 5.3 with an example of a trip maker who is on active call and driving in the eastbound direction between cell 1 and 4. As the vehicle crosses the roadway from cell 1 to cell 2, the call must be transferred between cells without interruption. Therefore, a handover event takes place (handover 1). This event occurs afterward at the location of handover 2 and handover 3 as far as the call is in progress. The idea behind the handover-based system is to use the handover lines at the cell borders as "virtual" traffic counters. These virtual traffic counters are assumed to capture the vehicle movement instead of using traditional on-road sensors represented by the circles.

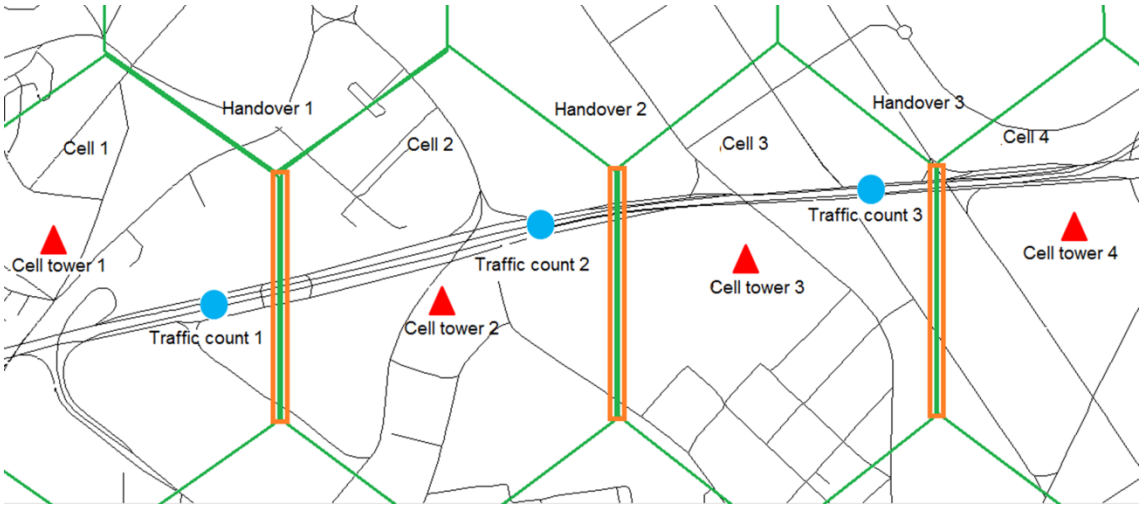


Figure 5.3 Schematic representation of handover-based system

We have followed this method to relate the traffic volumes from 12 traffic counters in the 5 case study areas and the associated handover counts (Figure 5.2). We started exploring this relationship with a correlation analysis between both sets of data to investigate the relationship between cellular and vehicular uses. We compared the hourly handover counts with the corresponding hourly traffic volumes from the inductive loops.

In Figure 5.4 the patterns of hourly traffic volumes and handover counts are plotted along the hours of the day. The left axis represents the traffic counts and the right axis the

handover counts. Since we would have 12 plots we chose only those counters with the highest average daily traffic from each case study area for presentation. The cell towers associated to the selected traffic counters are presented in Figure 5.2. An average coefficient of correlation of 0.76 was obtained from all the 12 counters. The result proved the existence of good correlation between the handover and the traffic volume which is in accordance to previous studies (Vaccari et al, 2009; Becker et al., 2011a). While the overall correlation values obtained in the previous as well as in our study seem impressive, this is not enough to say that handovers give site-specific traffic profile. We did a correlation analysis between traffic at one site and handover at another site. The result shows that there are times when the correlation between a handover from one site is better with a traffic count which is not close to it, thus we should not rely on the correlation result to prove the added value of handovers in providing site-specific traffic profile. The second observation is that different cases have different scales. Estimation of the absolute traffic volume through the use of the number of handovers is impossible if we wish to use the same model for different areas of the city. Using both scales in Figure 5.4 plot (a) there is a 100/2083=0.048 Handover/Vehicle relationship while in plot (b) there is a 200/414=0.49 Handover/Vehicle relationship, denoting a very different relationship between traffic counts and handover counts. In these two specific plots we should stress the fact that in (b) there are very high pedestrian movements when compared to (a), which has an influence on the number of calls.

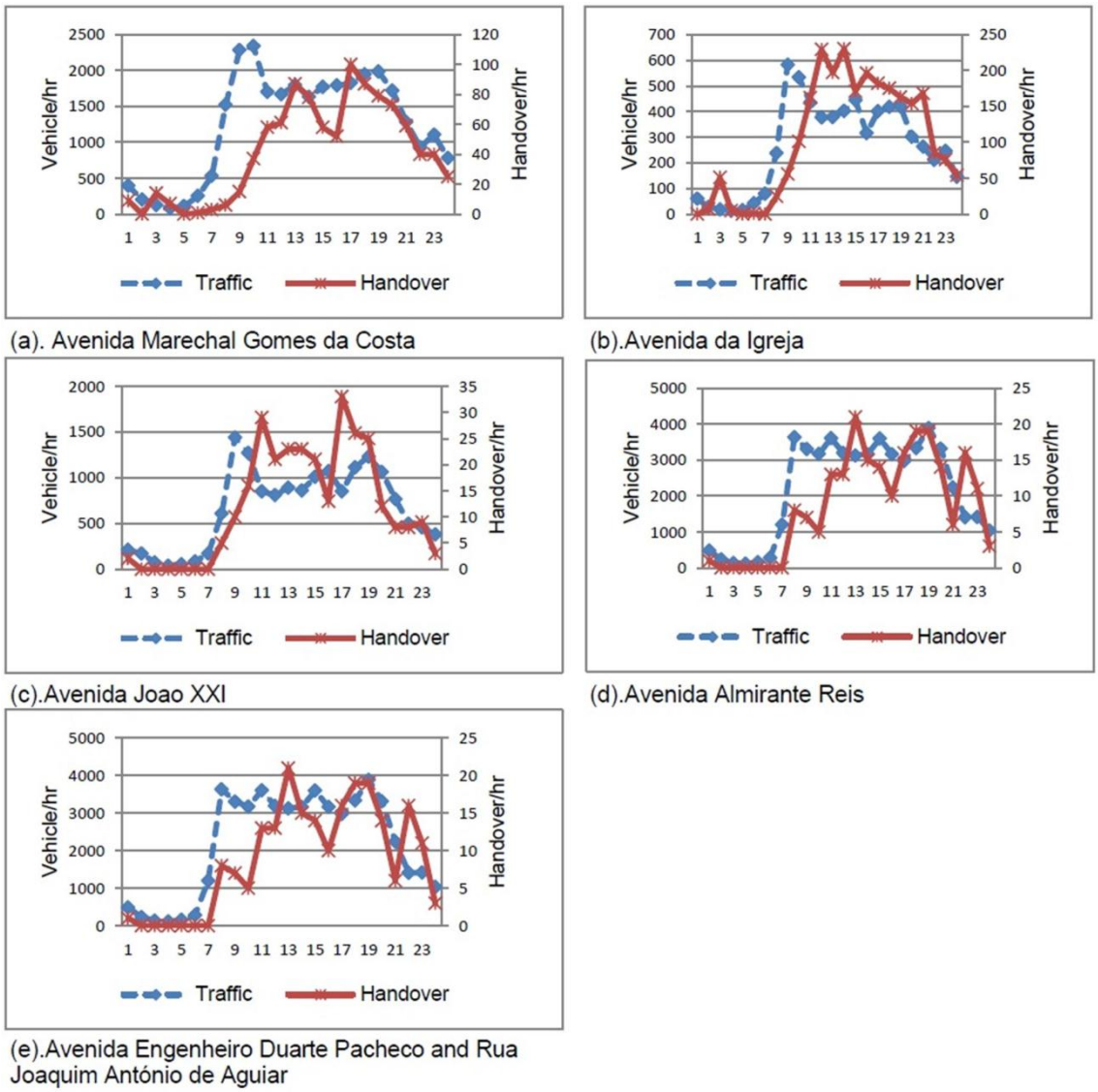


Figure 5.4 Traffic volumes and handover counts plotted over the hours of the day, where 1 implies midnight to 1AM, 2 implies 1AM-2AM, and so on

In order to tackle this problem we propose to estimate traffic levels instead of the absolute traffic volumes. The underlying assumption is that the pattern and amount of cellphone traffic movement is related with the intensity of traffic movements and understanding this relationship will help in managing the flow of urban traffic. The source and magnitude of handover events might vary depending on the kind of traffic activities. In its simplistic form, when there is a higher number of handover events in a cellular tower this should mean a traffic level increase and when this number decreases there should be a reduction on the traffic level passing near the tower. We aim to capture these differences in

the handovers and use them as signatures to classify traffic conditions. Many experts argue that the use of design capacity of each road as an appropriate choice to set the threshold values, one limitation of this approach would be the absence of a corresponding strategy to categorize the handover count as the notion of capacity of this variable does not relate to flow capacity. The other alternative is the use of some fraction of the highest observed traffic flow. While this approach is reasonable, it leads to biased outcomes especially when the highest observed traffic flow is far from the rest of the traffic counts of a day. For the purpose of our analysis that estimates the traffic levels on a given day at a given road link the thresholds were set at the 80^{th} and 50^{th} percentiles of the observed traffic. The observed hourly traffic of a day which is above the 80^{th} percentile is labeled as high traffic. The observed hourly traffic of a day between the 80^{th} and the 50^{th} percentile value is categorized as medium traffic level. The observed hourly traffic of a day below the 50^{th} percentile is classified as low traffic level. This notion of using percentiles as a threshold value is common for some traffic engineering applications. For example, posted speed limits in a highway are set primarily upon the 85^{th} percentile speed even though other factors like roadside development, road and shoulder surface characteristics, and pedestrian and bicycle activity are considered (Mark, 2010).

The choice of the 80^{th} and the 50^{th} percentile values in our study is arbitrary. The traffic level in this study does not necessarily show intensity of road congestion instead it is intended to show the proportion of traffic flow on a given day at a given road network based on the assigned cut-off points. This study stresses that while the 50^{th} and the 80^{th} percentiles give indicative values as to how much of the observed traffic is labeled as high, medium and low traffic of the day, there is no single acceptable threshold values. If this approach is operationalized in a real-time basis, local traffic management authority should decide how much of a daily observed traffic should be labeled as high, medium and low traffic that depends on local traffic conditions and the roadway types. For the handover levels we used the same rationale, dividing the variable in three categories using the same percentiles used for the traffic levels.

To relate the two categorical variables, two models were developed: a multinomial logit (MNL) and an artificial neural network (ANN). In the next two sections, we explain briefly the structure of these two models.

5.3.2. Multinomial logit

The Multinomial Logit (MNL) is a regression model that is appropriate for situations where the dependent variable is categorical. MNL is used to predict the probabilities of the different outcomes of a categorically distributed dependent variable, where the independent variables could be real valued, categorical or binary form (Stephenson et al., 2001). It has been widely used to study various problems in transportation, in particular the choice of mode of transportation in order to estimate the market share of new offers or changes in travel attributes such as travel time and cost (Ben-Akiva and Lerman, 1985).

For a given dependent variable that has j categories U_j measures the utility that characterizes the total contribution of all the explanatory variables to that category. U_j is composed of the known part V_j and unknown random part ε_j and it can be presented as $U_j = V_j + \varepsilon_j \, \forall j$ (Train, 2009). The probability of an alternative i among the possible j categories is computed as $P_i = \frac{exp^{V_i}}{\sum_j exp^{V_j}}$.

The known part of U_j, which is V_j, characterizes the level of importance of each variable in explaining the dependent variable and it is usually linear in parameters: $V_j = \beta' x_j$. Where, x_j is a vector of observed variables related to category j.

5.3.3. Artificial neural network

The Artificial Neural Network (ANN) is an artificial intelligence model that arose from an attempt to emulate the neurological pattern of human brain´s learning spot by combining many simple computing elements (neurons) into a highly interconnected system (Dougherty, 1995; Sarle, 1994). ANN takes both linear and nonlinear model approaches creating a better platform to understand the correlation between variables in a more valid basis (Sarle, 1994).

In its most general form, a neural network consists of several layers of neurons. The network type where neurons in each layer feed their output to the next layer is called feedforward (Nasr et al., 2003). Figure 5.5 shows a multilayered feedforward network with its nodes at each layer and the corresponding connection weights. In the first stage of the ANN model development, a considerable portion of the data is used for network training. During the training stage, the network tries to match the produced output with the values

provided by the training examples. In the second stage of model development, the network tries to perceive the model and this process is called perceptron process or hidden process (Kayri and Çokluk, 2010). In this stage, weights of the explanatory variables upon the dependent variables are produced. In a supervised learning method, the network tries iteratively to adjust weights of connections between neurons to produce the desired output. During this process, the error in the output is propagated back to the previous layers to adjust the weights, and the network uses the backpropagation method for propagating the error (Nasr et al., 2003). The final stage of the model development is the new model estimation where the ANN produces an output from an unknown input pattern.

The relationship between the neurons in a given layer to the neurons in the following layer can be explained mathematically. Let us assume the neuron in a given layer n_j is connected to the neurons in the previous layer designated as n_i. The connection between n_i and n_j is represented by weight w_{ij}. Therefore, the input and output of n_j is obtained through the use of Eqn. 5.1 and Eqn. 5.2, respectively.

$$input_j = \sum_{i=1}^{n} w_{ij} \times output_i \qquad Eqn. 5.1$$

The output from the n_j is calculated using a proper transfer function as

$$output_j = f(input_j) \qquad Eqn. 5.2$$

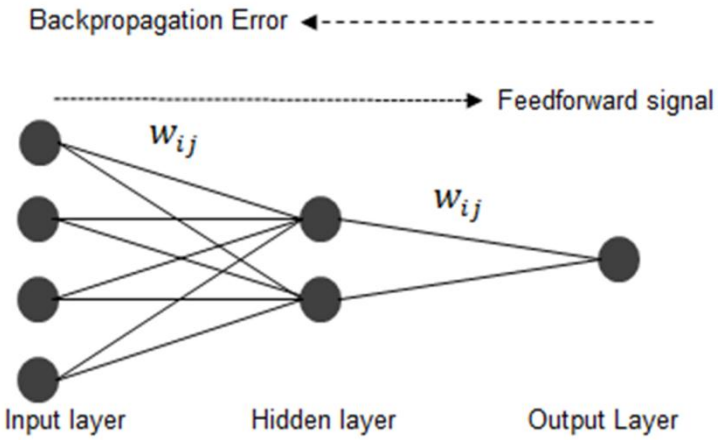


Figure 5.5 Schematic diagram of a typical feedforward neural network

The advantages of employing ANN over other statistical models were reflected in some previous classification problems of transportation studies. Rao et al. (1998) suggested the use of ANN over the MNL in modeling the travel behavior of individuals for the case of large number of alternatives and explanatory variables. Lingras and Adamo (1996) applied ANN and a multiple regression model for estimating average and peak hourly traffic volumes and ANN model achieved reasonable estimations when the data were not enough to develop multiple regression model. In spite of its advantages, ANN model has limitation regarding the coefficients corresponding to each explanatory variable that cannot be computed and presented as they are in the MNL model. Instead weights are produced in ANN model, but these weights are generally not used for interpreting the network results. This lack of interpretability at the level of individual explanatory variables is one of the prominent limitations of the ANN model.

5.4. Results and discussion

5.4.1. Multinomial Logit application

The MNL model was built through maximum likelihood estimation method using the SPSS software (IBM SPSS Statistics 20). Out of 288 hourly traffic levels collected from the five case study areas (12 counters x 24 hours), half of the data (144 traffic levels) were used to build the MNL model and the other half used for validation. Partitioning of the

calibration and validation datasets was made in such a way that the two sets had representatives from different traffic and pedestrian realities.

The dependent variables were built through the threshold values which were explained in section 5.3.1. The explanatory variables applied were the same for the three traffic categories. Knowing that the influential factor that matters in the distinction between the categories is the difference in their respective utilities (that characterize the state of the traffic condition through the corresponding explanatory variables) a reference alternative must be set, such that the remaining alternatives are compared to it. In our experiment we chose the low traffic level as a reference alternative.

Three explanatory variables were used to characterize the state of the three traffic levels. The categorized handover variable was used as a primary explanatory variable. The handover data was organized in the same way in which the observed traffic was organized i.e. using the aforementioned threshold values, 50^{th} and 80^{th} percentiles, as cut-off points to classify the data into HO_{low}, HO_{medium}, and HO_{high} .The low category was used as reference, therefore, it was not introduced in any of the utility functions, meaning that the difference between HO_{medium} and HO_{high} against the HO_{low} will be used to measure the difference in utility between high traffic and medium traffic in relation to low traffic. The remaining two explanatory variables were associated with pedestrian and the hours of a day. Ped, is a pedestrian density which was computed using Google street view for both sides of the street. Case study area A has 0.147 pedestrians/meter, case study area B has 0.239 pedestrians/meter, case study area C has 0.151 pedestrians/meter, case study area D has 0.135 pedestrians/meter, and case study area E has 0.0745 pedestrians/meter. We tried two different ways as to how to use the hours of a day as an explanatory variable in our model. The first alternative is the use of $\mathrm{Time_i}$ that represent dummy variables for each hour of the day (1 to 24). The second one is a dummy variable peak that represents a binary variable Peak/off-peak. Thus, it takes 1 if the time is in a peak hour period (8 to 10 in the morning, and 14, 15, 17, 18, and 19 in the afternoon) and 0 otherwise. The decision of peak/off-peak period was drawn from the average traffic flow for typical week days that were made available in the website of "Estradas de Portugal". Therefore, the inclusion of one of these two variables would capture the typical daily behavior. The utility functions that characterize each traffic level through the use of the explanatory variables ($HO_{high}, HO_{medium}, Peak,\ and\ Ped$) are given in $Eqn.\ 5.3\ to\ Eqn.\ 5.5$.

$$V(high\ traffic) = \alpha_h + \beta_{1h} \times HO_{high} + \beta_{2h} \times HO_{medium} + \beta_{3h} \times Peak + \beta_{4h} \times Ped \qquad Eqn.\ 5.3$$

$$V(medium\ traffic) = \alpha_m + \beta_{1m} \times HO_{high} + \beta_{2m} \times HO_{medium} + \beta_{3m} \times Peak + \beta_{4m} \times Ped \qquad Eqn.\ 5.4$$

$$V(low\ traffic) = 0\ (reference\ alternative) \qquad Eqn.\ 5.5$$

A coefficient β was estimated for each explanatory variable (HO_{high},HO_{medium}, Ped, $peak$, and $Time_i$). An independent coefficient α was estimated to an alternative specific constant (ASC) in the two first traffic categories (high and medium traffic) in order to represent the effect of other factors that are not captured through the available variables. The model was applied to the first 144 traffic levels. The coefficients of $Ped, Peak,\ HO_{high}$ and HO_{medium} variables were statistically significant at 5% significance level. The use of $Time_i$ dummy variables (the dummy variable for midnight to 1AM was kept as a reference variable) resulted in poor estimation and 22 of the dummy variables were not statistically significant at the usual level. The dummy variable for 7AM (which represents the time between 7AM to 8AM) was the only variable that was statistically significant at 5% significance level. Therefore, in the final MNL model we kept the peak variable instead of the $\mathrm{Time_i}$ variable as it gave better results. The MNL model estimated with the presence of the $peak$ variable is shown in Table 5-2.

Table 5-2 Multinomial logit model components

Choice sets	Variables	Coefficients(β)	P-value	Exp (β)	95% Confidence Interval for Exp(β)	
					Lower bound	Upper bound
High traffic	ASC_{high}	-6.188	< 0.001			
	HO_{high}	3.269	0.005	26.285	2.691	256.707
	HO_{medium}	3.836	< 0.001	46.360	5.893	364.733
	$Peak$	5.385	< 0.001	218.191	23.966	1986.497
	Ped	17.378	0.009	3.53×10^{7}	75.733	1.64×10^{13}
Medium traffic	ASC_{medium}	-3.093	< 0.001			
	HO_{high}	2.459	0.005	11.696	2.126	64.361
	HO_{medium}	2.380	0.002	10.801	2.451	47.600
	$Peak$	2.725	0.004	15.252	2.359	98.607
	Ped	17.663	0.002	4.68×10^{7}	547.306	4.01×10^{12}
Chi-square (significance)			152.970 (P-value: < 0.001)			
Pseudo R-Square			0.512			

We used the Chi-Square and pseudo R-square statistics to evaluate the overall model fit to the data. A good Pseudo R-Square value of 0.512 was obtained. This figure was mapped to a linear R square value of 0.90 based on the conversion graph developed by Domencich and McFadden (1975). A Chi-Square value of 152.97 was obtained confirming the possibility of estimating the traffic status through the use of the explanatory variables. The statistical significance of the chi-square test is well below the level of significance of 0.05. The null hypothesis that there is no difference between a reference model without explanatory variables and our model was rejected with great sureness.

Analysis of the estimated coefficients (β) was carried out to understand the relative importance of the explanatory variables and the low traffic level is treated as a reference category. Therefore, since the parameter estimates are relative to the reference category, the normal interpretation of the multinomial logit is that for a unit change in the explanatory variable, the logit outcomes (high and medium traffic levels) relative to the reference category is expected to change by its respective parameter estimate (which is in log-odds scale) given the variables in the model are held constant. The coefficients of the variables ASC_{high} and ASC_{medium} are both negative showing that there is a lack of explanation of the low traffic level. The presence of the variable peak clearly reveals its effect on the happening of high traffic level. This confirms the expectation to have a high traffic level during the usual rush hours.

The coefficients of HO_{high} and HO_{medium} have both positive values for the high traffic and medium traffic levels. However, the coefficient of $\mathrm{HO_{medium}}$ has a higher value for high traffic level and low value for the medium traffic level. This was not intuitive and posed an interesting question: why did not HO_{medium} have a higher effect on the medium traffic level? At this point our research does not point a reason for this result and we suggest further investigation in future work. The Exp(β) column presents the exponential of the coefficients, which converts the log-odds into odds ratios of the explanatory variables for easier interpretation. For example, the Exp(β) for $\mathrm{HO_{high}}$ in the high traffic level is 26.285. This shows for $\mathrm{HO_{high}}$ relative to $\mathrm{HO_{low}}$, the odds of inferring high traffic level relative to low traffic level would increase by a factor of 26.285. In other words, a high level of handover is more likely than a low level of handover to be associated with high traffic level compares to low traffic level.

The MNL model was validated using the remaining 144 traffic levels out of the total 288 hourly traffic levels collected from the arterial roads represented on Figure 5.2. Table 5-3 shows the classification results of the MNL model. Cells on the diagonal of the table represent the true positive classifications. An overall correct classification accuracy of 76.4% to the three traffic levels was obtained. However, the model performance on classifying the medium traffic level is the worst when compared to the other two traffic levels. There were 16.7% of the high traffic levels classified as low traffic.

Table 5-3 Classification accuracy: MNL model validation stage

Observed category	Observed traffic levels for model validation	Predicted Category			
		High traffic (%)	Medium traffic (%)	Low traffic (%)	Over all accuracy (%)
High traffic	30	73.3	10.0	16.7	
Medium traffic	42	40.5	52.4	7.1	76.4
Low traffic	72	0.0	8.3	91.7	

5.4.2. Artificial Neural Network application

We also used artificial neural networks for the estimation of the traffic levels. We considered a multilayer perceptron with a backpropagation algorithm. The training of the neural network was carried out by the same 144 traffic levels that were applied to build the MNL model.

We also tried use of the two different explanatory variables that were made out of the hours of a day in the ANN model (Time_i, and $Peak$). The results of the ANN model that used $Timei$ variable is discussed in this section and in section 5.4.4 we present the results of the ANN model that used the variable $peak$.

The input layer of the neural network architecture was made of 26 input nodes that comprise two nodes for handover variables (HO_{high} and $\text{HO}_{\text{medium}}$) and one node for the ped variable, and the remaining 23 nodes for $Time_i$ (the dummy variable for midnight to 1AM was kept as a reference variable). Decision on the number of hidden layers of the neural network requires understanding the complexity of the problem and the transfer function of the layer. We carried out a preliminary analysis to identify an appropriate number of hidden layers and the number of nodes in the hidden layers. A sensitivity analysis was done by altering different combinations of hidden layers along with the number of nodes that helps to manage the problem of over-fitting the ANN model. We found that one hidden layer with 7 nodes gives good classification accuracy both in the training and validation stages of the model development. We used the sigmoid function as a transfer function for the hidden layer along with conventional stopping criteria to prevent overtraining. In order to estimate different categories of traffic status, we applied a transfer function of multiple logistic which is called softmax that estimates the conditional probabilities of the three traffic levels.

Table 5-4 shows the classification results of using the ANN in the validation stage of the model development. Of all the traffic levels applied to validate the ANN model: 73.3% for the high traffic, 57.1% for the medium traffic and 93.1% for the low traffic are the true positive classifications. The overall correct classification accuracy in the validation stage of the model development reaches 78.1%. However, a high proportion of the medium traffic, 42.9%, was misclassified as high traffic level, where a similar pattern was also observed in the MNL model output. There are still 6.7% of the high traffic levels classified as low traffic and 2.8% of the low traffic classified as high traffic level.

Table 5-4 Classification accuracy: ANN model validation stage

Observed category	Observed traffic levels for model validation	Predicted Category			
		High traffic (%)	Medium traffic (%)	Low traffic (%)	Over all accuracy (%)
High traffic	30	73.3	20.0	6.7	
Medium traffic	42	42.9	57.1	0.0	78.1
Low traffic	72	2.8	4.2	93.1	

5.4.3. Prediction based on a City-wide time-of-day traffic profile

In the absence of a traffic count from a specific site, prediction is usually made through the average traffic from other places or historical traffic data of a given location and we call this information a City-wide time-of-day traffic profile. For our analysis the city-wide time-of-day traffic profile was built by averaging traffic volumes obtained from 95 traffic counters in Lisbon during April 6, 2010 (that excludes the traffic counts from the 6 counters used for validation). The average traffic of the day is then grouped with the same threshold values, 50^{th} and 80^{th} percentiles of the observed traffic, implemented by the MNL and ANN models. The traffic levels obtained represent the predictions through the City-wide time-of-day traffic profile. We did the validation of these traffic levels with the same dataset used in the MNL and ANN models. Table 5-5 shows the classification accuracy achieved by the City-wide time-of-day traffic profile. Use of the City-wide time-of-day profile delivered an overall correct classification accuracy of 70.8%, however, it only predicts 43.3% of the high traffic level correctly and the remaining 46.6% of the high traffic level is classified as medium traffic level and 10% as low traffic level.

Table 5-5 Classification accuracy: City-wide time-of-day traffic profile

Observed category	Observed traffic levels for model validation	Predicted Category			
		High traffic (%)	Medium traffic (%)	Low traffic (%)	Over all accuracy (%)
High traffic	30	43.3	46.7	10.0	
Medium traffic	42	38.1	54.8	7.1	70.8
Low traffic	72	1.4	6.9	91.7	

5.4.4. Comparison of predictions by the MNL, ANN, and city-wide time-of-day traffic profile

The overall correct classification obtained through the ANN and MNL models were, respectively, 7.3% and 5.6% better than the result delivered by the City-wide time-of-day traffic profile. The high traffic level predicted through the City-wide time-of-day traffic profile is inferior to the ANN and MNL models by 30%. This seems to indicate that relying on the City-wide time-of-day traffic profile in the absence of traffic data from a specific site would give less accurate prediction in the more important task of distinguishing the high traffic level.

A comparison of the ANN and MNL model was carried out by taking common explanatory variables and the same proportion of calibration and validation datasets. The explanatory variables were the ones which were used to build the MNL model and found to be statistically significant (HO_{high} ,HO_{medium}, $Peak$, and Ped). We obtained the same total correct classification accuracy of 76.4% to the three traffic levels. However, this comparison strategy constrained the ANN analysis by limiting the number of potential explanatory variables to a similar set of variables used in the MNL model. In the ANN, the use of $Time_i$ dummy variables instead of the peak dummy variable for the hours of a day delivered 1.7% more overall correct classification accuracy than the corresponding MNL model.

5.5. Summary

The growing development of ITS seeks to improve the efficiency of transportation systems through the use of emerging information technologies. Successful deployment of ITS schemes requires a pool of heterogeneous datasets to allow both road managers and road users to understand traffic conditions at any moment. The state-of-the-practice for road traffic data collection primarily depends on conventional on-road sensors. However, because of their high cost, on-road sensors are only available in a small portion of the roadway system. One way to solve this problem can come from using other sources of data. In this Chapter, we explored the use of cellular networks handover information for traffic status estimation in urban environment with the objective of assessing its possibilities and pointing directions for future research in this field.

We first investigated the correlation between cellphone handovers and traffic volumes in order to understand the relationship between cellular activity and vehicular traffic. We compared the hourly handover counts with the corresponding hourly traffic volume counts collected from 12 counter locations in Lisbon. We obtained an average coefficient of correlation of 0.76, however, two considerations must be made: (1) though the result confirms good relationship, correlation is not enough to claim the possibility of handover giving site-specific traffic pattern, (2) the relationship between traffic and handover happened in different scales, i.e., one handover cannot be related to the same number of vehicles in all cases.

Then a tentative method was developed aiming at predicting the traffic status in an urban environment through the handover counts. Unlike other studies which estimated specific traffic parameters of the traffic stream such as speed and density, our approach distinguishes the intensity of traffic and provides estimates on traffic levels in a specific road segment. For this purpose we applied the MNL and ANN models. The performances of our models were validated through the use of half of the dataset. The overall correct classification accuracies from the MNL and ANN models were, respectively, 76.4% and 78.1% that outperformed the overall accuracy of 70.8% obtained through a City-wide time-of-day traffic profile. This seem to indicate that in the absence of traffic data a traffic profile determined by handover based models could give better site-specific information compared to a City-wide time-of-day traffic profile.

We should emphasize some important limitations of the study. Our cellular network data was limited to mobile phones that are actively making a phone call, the caller must be a client of the TMN cellphone operator, and the duration of the associated call must be long enough to traverse the boundaries of two cells. Thus, it was not possible to make a direct correspondence of the cellphone activity (handover counts) and traffic. It is essential to our approach that calls are being made while driving. Previous studies came up with different statistics about the percentage of drivers who use the phone while driving, which is, as expected, low (Jeanne Breen Consulting, 2009). However, we may also point that the car passenger phone usage and the increased usage of hands-free equipment's were overlooked.

Another limitation is related to the handover counts. Even using rules to try to associate specific handover events to a particular road; it is a challenging task to sort out calls that

were carried out while driving on those specific roads. This bias must be further investigated by analyzing with greater detail the relationship between the cellphone location and the cell towers that connect to that equipment.

Our study was carried out on an urban road network, where the traffic flow is more likely to be influenced by entry and exit roads, traffic control devices, intersections, and presence of turning movements. The presence of such interruptions causes unevenness in the traffic flowing along the study segment. This is a challenge and a cause for estimation errors. Despite trying to avoid this bias by carefully selecting the counters and cell towers in the case study, this is undeniably present and in future work we should explore how the network complexity influences the prediction model.

Finally, further studies have to be carried out that consider more detailed information on the handovers: lower aggregation than the hour and better information about the orientation of the cell tower sections. One should also bear in mind that these models are always inferior to a system that uses a traffic count from loop detectors. However if an improved version is developed it could be a good option as a new development proposal in the absence of conventional on-road sensors or for a hybrid application combining the handovers with a limited number of on-road sensors. This could be useful in third world countries where the population is having access to mobile phones, but the infrastructure is not fully developed yet.

We believe our analyses show plenty of opportunities to gain valuable insights in the field of traffic estimation and implementation feasibility. If this approach is operationalized in a real-time basis, the result could be employed for a wide range of applications in the field of active traffic management. From the roadway manager point of view, the estimation outputs give indicative values about the intensity of the traffic and could be used as a baseline to trigger different road traffic management schemes.

Chapter 6 Conclusions

6.1. Introduction

In recent years, researchers are exploring ways to develop large scale urban sensing by employing the increasing capabilities found in the cellular networks system. It is possible to approximate the location of cellphone users whenever they make a call or send a short message service. In this thesis we explored the use of passive mobile positioning data for profiling the dynamics of urban activities and characterizing flows of people for planning of urban and transportation systems in cities. In the process, we proposed interesting and new approaches to address the urban and transportation problems. The data were obtained from the Portuguese cellphone operator, TMN, which accounts for 45% of the total number of cellphone subscribers in Portugal. In 2010, active mobile telephone cards per 100 Portuguese inhabitants were approximately 160 (ANACOM, 2010). Mainly, three datasets were used in this thesis: Handover, Call Volume and Erlang, which were extracted from 487 base stations comprising 1669 cell sectors geographically located within the Lisbon area and neighboring municipalities.

6.2. Main findings

In Chapter 3, we started our analysis by performing exploratory data analysis with the aim of extracting important cellular network data features and test underlying assumptions. We employed two different techniques, visualization and statistical analysis. As part of extracting the important features of cellular network data, we separated handover events into incoming and outgoing handovers for each tower location. We also set different criteria to differentiate cell towers of different characteristics based on the amount of handover events they were serving: (1) cell towers with balanced handovers: cellular towers with a high number of handover events, both incoming and outgoing; (2) cell towers with a high number of incoming handovers; and (3) cell towers with a high number of outgoing handovers.

In the first analysis, we used visualization techniques in order to uncover the mobility relevance of cellular network handover data. Interestingly, the handover maps were very

informative and gave a qualitative view on the amount and patterns of call movements. The visualizations provided a means to understand and visualize the flow of people from the entire urban system and its organization at a glance.

Besides the important results achieved through visualization, we also employed statistical analysis to investigate the presence of significant relationships between the handover data and different sets of urban characteristics (proximity to the main road network, presence of people, and traffic in the main road network). We have discovered the following four relationships: (1) it was found that cellular towers characterized by a high and balanced number of incoming and outgoing handovers are located in the vicinity of the main road links; (2) there is a strong relationship between cellular towers with a high and balanced number of incoming as well as outgoing handovers and road links that are serving the highest vehicular traffic within 250 meters radius; (3) it was found that cellular towers with a high number of incoming handovers tend to be further away from the main road links; and (4) there is a strong relationship between the presence of people in the city and the number of incoming handovers. The existence of a relationship between vehicular and pedestrian movements, and calling activities were addressed to a certain degree in the previous studies. However, the investigation regarding the relationship between the location of main road links and cellular towers is quite new.

Of more significance than identifying the movement of people and vehicles at a glance is the potential of cellular network data in detecting the characteristics of urban activities in terms of their intensity and pattern at different parts of a city along the hours of a day. The pattern and intensity of urban activities along the hours of a day reflects the distribution of people within an operationally sound cycle.

In Chapter 4, we raised research questions of relevance to transport planners, transport geographers, and urban planners: Is it possible to use cellphone data to detect the intensity of urban activities along the hours of a day? And are we able to explain the varying activity patterns along the hours of a day in different parts of a city? We employed three types of passive mobile positioning data: Handover, Call Volume, and Erlang to explore their potential in detecting the activity pattern and intensity of places. To validate our results, we used ground truth which was composed of indicators associated to distribution of people, buildings and movement of vehicles.

We applied fuzzy c-mean clustering to the cellphone data to create clusters of locations with similar features in what respects to two aspects of the urban activities: pattern and intensity along the hours of a day. The underlying assumption is that the different patterns and intensity of urban activities would relate to the presence of different size of calling activities at a given time in a given urban area. Thus, for a given urban area, the pattern of activity refers to the evolution of its calling activity along the hours of a day, and the intensity of activity refers to its share of the total calling activity of a city for each hour of a day.

We associated the pattern and intensity of calling activities to different land use categories. We used the pattern of urban activities to specify the nature of land use activity (predominantly residential and nonresidential area), and the intensity of urban activities to specify density of use (high and low activity area). The underlying assumption is that different urban areas can be distinguished based on the activities of people who use the land, which reflect the aggregate interactions between people and places over time.

Land use is a rich concept because of the uses variety, which goes far beyond the basic land use classifications and should include many other associated characteristics and components such as (1) land as functional space devoted to various uses (e.g., urban, rural, residential, commercial, industrial, etc.); (2) land as a setting for activities (e.g., working, studying, recreating, commuting, etc.); (3) land as part of an environmental system (e.g., wetland, forest, wildlife habitat, etc.); (4) land as real estate exchange commodity to be bought, developed, and sold; (5) land as publically planned, serviced, and regulated space; and (6) land as visual features for orientation and social symbolism (Berke et al., 2006). Our analysis has limitations in explicitly defining the land use in terms of land use type, location, amount, services, condition, design, and value, etc.

Land use systems are dynamic, where its use can expand and contract, persist and change, in response to different factors: population and economic growth; public and private decisions; and market and government actions (Berke et al., 2006). Traditionally, to plan for land use change, planners perform inventory of existing land as well as available land for future development; and monitor changes in these inventories. However, some analysts argue that planning the changes in land use requires a more dynamic land inventory and a continuous monitoring approach (Knaap, 2004). One of the advantages of

our analysis is that cellphone data can be obtained frequently to assess the changes in the pattern and intensity of urban land uses.

Observing the crossing between the two analyses, the pattern and intensity of urban activities, provided interesting results. These results show that the central business district of Lisbon is dominated predominantly by nonresidential areas accompanied by high intensity of activities and the outskirts of the city is mainly residential with low intensity of activities along the hours of the day. This strengthens the fact that different activities have their own peculiar location requirements. On both the low and the high activity areas, the major change in the cellphone usage occurs between 7am to 11am and 5pm to 9pm, which, in fact, is well associated with the time periods for traffic rush hours. However, it is interesting to note that the high and low activity areas exhibit opposite patterns of average percentage changes of calling activity. In the high activity area, the average percentage calling activity increases between 7am and 11am and reduces between 5pm to 9pm. This pattern is reversed in the low activity area. The average percentage change in the calling activity along the day is suggesting the general inward and outward movement of people and vehicles from residential to nonresidential areas and back again.

Knowledge of the spatial and temporal distribution of urban activities is an important factor to understand the urban land use patterns. Information regarding how the existing land is used and the continuous monitoring of the changes in the land use through time is necessary for legislators, planners, and local governmental officials to determine better land use policy, to project transportation and demand, and to implement effective plans for regional development (Anderson, 1976).

The other most common inventory for the transportation planning is the one performed to measure the usage of roadways. Annual average daily traffic and average daily vehicle distance traveled are the two most used traffic statistics mainly for traffic planning purposes. However, traffic managing authorities should also collect hourly traffic volumes to look up to the operational characteristics of the road at different hours of a day (FHWA, 2001). The variation in traffic by the hour of the day is an important factor for traffic management. For instance, the morning and afternoon peak hour traffic usage is used as a reference to estimate the requirements for facilities.

In chapter 5, we proposed a complementary method to explore the usefulness of cellular networks handover data to estimate vehicular traffic. To test this method, hourly handover

counts were obtained from 39 cellular towers in the vicinity of arterial roads that have 12 traffic counters. The traffic counters were selected from five case study areas with diverse characteristics.

A primary attempt to associate the traffic counts with the handover counts resulted in two shortcomings: (1) there were times when the correlation between a handover from one site was better with a traffic count which was not close to it, thus we should not rely on the correlation result to prove that handovers could provide site-specific traffic profile. The other shortcoming was that the number of vehicles to handovers ratio at different sites can change up to 10 folds, which showed estimation of the absolute traffic volume through handover count is impossible if we wish to use the same model for different areas of the city.

As an alternative to the absolute traffic volumes, we proposed to estimate traffic levels. The underlying assumption is that the changes in the intensity of vehicular traffic will correspond to the changes in the intensity of cellphone traffic and understanding this relationship will help in managing the urban traffic. We targeted to capture these changes in the handovers and use them as signatures to classify intensity of vehicular traffic conditions.

To relate traffic and handover counts, two models were developed: a multinomial logit (MNL) and an artificial neural network (ANN). Our approach distinguishes the intensity of traffic and provides estimates on traffic levels in a specific road segment. The dependent variables for the models were: high traffic level, medium traffic level, and low traffic level. The independent variables were composed of handover variables (high handover and medium handover); pedestrian density (which was obtained using google street view from both sides of the street); and time variables (peak and off-peak time variables drawn from statistics on Estradas de Portugal site).

The performances of the two models, MNL and ANN, were validated through the use of half of the datasets that were not used during model calibration. The decision of using half of the datasets for model calibration and the other half for model validation was made to accommodate enough representatives of different traffic and pedestrian realities at the two different stages of model development. The overall correct classification accuracies from the MNL and ANN models were, respectively, 76.4% and 78.1%. The two models also outperformed the prediction based on a city-wide time-of-day traffic profile. This seems to

suggest that the City-wide time-of-day traffic profile provides less accurate prediction compared to the handover based models in the absence of traffic count from a specific site.

6.3. Limitations

The analyses in this thesis showed that passive mobile positioning data could provide plenty of opportunities to support the urban and transportation planning and operation. In the meantime, there were certain limitations associated to applying these datasets for urban and transportation studies.

From the basic nature of passive mobile positioning data, there are two limitations. The data is sparse in time because it is limited to cellphones that are actively making a call or send short message service. It is also coarse in space because the location record is made at the granularity of cell sector or cell tower, which gives uncertainty on the exact location of the cellphone.

In urban environments a mobile device can reach out multiple base stations at the same time, and selects a base station with the strongest signal and best signal quality for transmission (Küpper, 2007). Thus, it is not always the case where the call activity handled by a particular cell represents the actual people´s activity in its vicinity. Another bias could rise from the cellular network size and coverage representation. Alternative methods that imitate the real coverage of cellular towers (e.g. grid areas) would undermine the outcomes of the studies in this field.

In Chapter 3, we have shown that the cellular networks handover information can be used as a proxy to provide information regarding the road traffic condition. In chapter 5, we proposed a method that predicts road traffic levels through the use of cellular network handover counts. These analyses were performed on an urban road network, where the traffic flow is more likely to be influenced by entry and exit roads, traffic control devices, and intersections. In addition, the presence of parallel roads/railways and major pedestrian avenues increases the complexity of the system. This undeniably causes estimation error because it poses a challenge to single out data associated to road users on the road of interest.

Some bias might also arise from the size of our data and its representativeness. However, the share market of the cellphone operator (TMN) which the dataset was obtained is 45%, and TMN accommodates all groups of customers and there is no special

preference of a certain type of population segments for this particular cellphone operator. Other factors that might obfuscate the results in our study would be calling plans, which might limit the sample size at certain hours of the day.

6.4. Contributions

In spite of some limitations associated to our analyses, the results of this thesis provided plenty of benefits relevant to the urban and transportation planning. We were able to demonstrate the presence of significant relationships between selected urban characteristics and cellular networks data; extracted the pattern and intensity of urban activities through aggregate cellphone usage; and showed the potential of cellular network data in providing site-specific traffic profile in terms of hourly traffic level estimates.

Compared to cellular networks data, it is quite clear that urban and transportation data obtained through traditional methods are superior in terms of providing detailed and accurate features of the urban and transportation use. However, it is not always the case that traditional methods give the necessary information at frequent interval with reasonable cost. On the other hand, our analyses contribute to an understanding of where cellular networks data could be used as complementary or/and alternative dataset, which could be collected nearly in real-time at fine time resolution from a large portion of a population. Our approach can be applied, for instance, in some developing countries, which have high cellphone penetration, but poor availability of data, as a cheaper way to understand urban activities and their dynamics.

Most developing cities are concerned with the urban and transportation problems, such as traffic congestion, parking difficulties, pollution, and shortage of public transportation. The authorities are constructing new roads and subways, providing public transportation facilities, and managing the existing facilities to meet the ever growing travel demand. However, needs for constructing and managing transportation facilities compete with other community development projects, such as schools, health facilities, parks etc. Thus, the budget requirements for other competitive projects will place limitations on the extent of the transportations planners' ability to fulfil the needs of transportation. In the light of all these considerations, transportation planners should weigh alternative courses of actions such as, taking advantage of cellular networks as alternative and/or complementary source of data as in their urban and transportation planning procedures.

6.5. Future works

This research on the exploration of cellular networks data for the urban and transportation planning was important, resulting in some models and the assessment of several results. However, there is most certainly room for further work as this is still a relatively new research field. Future studies should consider more detailed information on the cellular networks data: lower aggregation than the hour; better information about the orientation of the cell sectors; and the possibility of obtaining digitalized best serving maps that show cellular networks service coverage. A possible future research direction would be also to include cellphone data from other cellphone operators and it would be interesting to consider new case studies to be applied to the proposed approaches.

The analysis regarding the pattern and intensity of urban activities in Chapter 4 can be extended with larger sample sizes in terms of longer time series that might lead to higher classification accuracy. It would also be interesting to compare the cellphone usage collected from Lisbon with similar data from other places. Especially, cities with strong spatial segregation of urban activities are expected to exhibit strong signal differentiation.

With respect to the traffic estimation models in chapter 5, further research can be done on the combined use of handover points and conventional on-road sensors. For the purpose of traffic counting, a simulation approach can be applied to rigorously evaluate various scenarios, such as different combination of both systems at different traffic conditions and roadway types.

REFERENCES

ACA, 2004. Location, location, location: The future use of location information to enhance the handling of emergency mobile phone calls. The Australian Communications Authority (ACA). http://www.acma.gov.au/webwr/consumer_info/location.pdf (Accessed 15.12.2013).

Ahas, R., Silm, S., Järv, O., Saluveer, E., Tiru, M., 2010. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. Journal of Urban Technology, 17(1), 3-27.

Ahas, R., Aasa, A., Silm, S., Aunap, R., Kalle, H., & Mark, U., 2007. Mobile positioning in space-time behaviour studies: Social Positioning Method experiments in Estonia. Cartography and Geographic Information Science, 34(4), 259-273.

Alger, M., Wilson, E., Gould, T., Whittaker, R., Radulovic, N., 2005. Real-time traffic monitoring using mobile phone data. Vodafone Pilotentwicklung GmbH.

ANACOM, 2010. Autoridade nacional de comunicaçoes. http://www.anacom.pt/render.jsp?contentId=1027764. (Accessed 05.11.2011)

Anderson, J.R., Hardy, E.E., Roach, J.T., Witmer, R.E., 1976. A land use and land cover classification system for use with remote sensor data. United States government printing office, Washington. http://landcover.usgs.gov/pdf/anderson.pdf (Accessed 14.02.2014)

Asakura, Y., Hato, E., 2004. Tracking survey for individual travel behaviour using mobile communication instruments. Transportation Research Part C: Emerging Technologies, 12 (3-4), 273-291.

Avni, O., 2007. Performance and Limitations of Cellular Based Traffic Monitoring Systems. 6th European Congress and Exhibition on Intelligent Transport System and Services, Aalborg, Denmark.

Bar-Gera, H., 2007. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. Transportation Research Part C: Emerging Technologies, 15 (6), 380-391.

Batty, M., 1996. Urban change. Environment and Planning B: Planning and Design, 23 (5), 513-514.

Batty, M., 2002. Thinking about cities as spatial events. Environment and Planning B: Planning and Design, 29, 1-2.

Becker, R.A., Caceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., Volinsky, C., Ave, P., Park, F., 2011a. Route Classification Using Cellular Handoff Patterns. Proceedings of the 13th international conference on Ubiquitous computing. ACM, Beijing, China, September 17 -21.

Becker, R.A., Caceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., Volinsky, C., Ave, P., Park, F.A., 2011b. A Tale of One City: Using Cellular Network Data for Urban Planning. Pervasive computing, 10 (4), 10-18.

Ben-Akiva, M.E., Lerman, S.R., 1985. Discrete choice analysis: theory and application to travel demand. MIT Press, Cambridge, MA.

Berke, P.R., Godschalk, D. R., Kaiser, E.J., Rodriguez, D.A., 2006. Land use systems, In: 5th (ed.), Urban land use planning, University of Illinois Press, Urbana and Chicago. 2006. PP. 197-223.

Bertaud, A., 2004. The spatial organization of cities: Deliberate outcome or unforeseen consequence? http://alain-ertaud.com/images/AB_The_spatial_organization_of_cities_Version_3.pdf (Accessed 09.08.13).

Bolbol, A., Cheng, T., Paracha, A., 2010. GEOTRAVELDIARY: Towards Online Automatic Travel Behaviour Detection. WebMGS: 1st International Workshop on Pervasive Web Mapping, Geoprocessing and Services, Politecnico di Milano, Como, Italy, August 26-27.

Bratanov, P.I., 1999. User Mobility Modeling in Cellular Communications Networks. Ph.D thesis, Technischen Universität Wien Fakultät für Elektrotechnik.

Caceres, N., Wideberg, J.P., Benitez, F.G., 2007. Deriving origin–destination data from a mobile phone network. IET Intelligent Transport Systems, 1 (1), 15-26.

Caceres, N., Wideberg, J.P., Benitez, F.G., 2008. Review of traffic data estimations extracted from cellular networks. IET Intelligent Transport Systems, 2 (3), 179-192.

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., Ratti, C., 2011. Real-time urban monitoring using cellphones: A case study in Rome. IEEE Transactions on Intelligent Transportation Systems, 12 (1), 141-151.

Calabrese, F., Lorenzo, G., Liu, L., Ratti, C., 2011a. Estimating origin-destination flows using mobile phone location data. Pervasive Computing, IEEE, 10 (4), 36-44.

Calabrese, F., 2011b, Urban sensing using mobile phone network data, lecture notes distributed at the 13th international conference on Ubiquitous, Beijing, China on September 17-21. researcher.watson.ibm.com/researcher/files/ie-FCALABRE/Urban%20sensing%20using%20mobile%20phone%20network%20data.pdf (Accessed 18.12.2013).

Calabrese, F., Pereira, F.C., Lorenzo, G.D., Liu, L., Ratti, C., 2010. The Geography of taste: Analyzing Cell-Phone mobility and social events. Proceedings of the Eighth International Conference on Pervasive Computing. Springer-Verlag Berlin, Heidelberg 2010, Helsinki, Finland, 22-37.

Calabrese, F., Reades, J., Ratti, C. 2010a. Eigenplaces: Segmenting Space through Digital Signatures. IEEE Pervasive Computing, 9(1), 78-84.

Cayford, R., Johnson, T., 2003. Operational Parameters Affecting the Use of Anonymous CellPhone Tracking for Generating Traffic Information. Technical report, Transportation Research Board Annual Meeting, TRB 2003, Washington, D.C., USA, Jan. 2003.

Cellint, 2007. TrafficSense - Road Traffic Monitoring and Traffic Information Services. Available at: http://cellint.com/index.html (Accessed 02.12. 2011).

Correia, G.H., Viegas, J.M., 2011. Carpooling and carpool clubs: Clarifying concepts and assessing value enhancement possibilities. Transportation Research Part A: Policy and Practice, 45 (2), 81-90.

Csáji, B., Browet, A., Traag, V., Delvenne, J., Huens, E., Dooren, P., Smoreda, Z., Blondel, V., 2013. Exploring the mobility of mobile phone users. Physica A Statistical Mechanics and its Applications, 392 (6), 1459-1473.

Deloitte and GSMA, 2012. Sub-Saharan Africa Mobile Observatory. http://www.gsma.com/publicpolicy/wp-content/uploads/2012/03/SSA_FullReport_v6.1_clean.pdf (Accessed 27.04.2013).

Demissie, M.G., Correia, G.H., Bento, C., 2013. Exploring cellular network handover information for urban mobility analysis. Journal of Transport Geography, 31(2013), 164-170.

Demissie, M.G., Correia, G.H., Bento, C., 2013a. Intelligent road traffic status detection system through cellular networks handover information: An exploratory study. Transportation Research Part C: Emerging Technologies, 32(2013), 76-78.

Domencich, T.A., McFadden, D., 1975. Statistical estimation of choice probability function, in: J.Tinbergen, D.W.Jorgenson, J.Waelbroeck (Eds.), Urban Travel Demand: A Behavioral Analysis. North-Holland publishing company-Amsterdam: Noth-Holland Publishing company.LTD.-Oxford.

Dougherty, M., 1995. A review of neural networks applied to transport. Transportation Research Part C: Emerging Technologies, 3(4), 247-260.

FHWA, 2001. Traffic Monitoring Guide. Office of highway policy information, Washington, D.C.

González, M.C., Hidalgo, C.A., Barabási, A.-L., 2008. Understanding individual human mobility patterns. Nature, 453(7196), 779-782.

Gundlegård, D., Karlsson, J. M., 2009. Route classification in travel time estimation based on cellular network signaling. Proceedings of 12th International IEEE Conference on Intelligent Transport Systems (ITSC), St. Louis, USA, October 3-7.

Herrera, J.C., Work, D.B., Herring, R., Ban, X.J., Jacobson, Q., Bayen, A.M., 2010. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. Transportation Research Part C: Emerging Technologies, 18 (4), 568–583.

Heydecker, B.G., Addison, J.D., 2011. Analysis and modelling of traffic flow under variable speed limits. Transportation Research Part C: Emerging Technologies, 19(2), 207-217.

Hongsakham, W., Pattara-atikom, W., Peachavanish, R., 2008. Estimating road traffic congestion from cellular handoff information using cell-based neural networks and K-means clustering. 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology. IEEE, Krabi, Thailand, May 14-17.

INE, 2013. http://www.ine.pt/xportal/xmain?xpgid=ine_main&xpid=INE (Accessed 28.10.2013).

INRIX, 2012. Mobile Solutions. http://www.itisholdings.com/mobile.asp (Accessed 02.12.2011).

Iqbal, M. S., Choudhury, C. F., Wang, P., González, M. C., 2014. Development of origin–destination matrices using mobile phone call data. Transportation Research Part C: Emerging Technologies, 40 (2014), 63–74.

Isaacman, S., Becker, R., Caceres, R., Kobourov, S., Martonosi, M., Rowland, J., Varshavsky, A., 2011. Identifying important places in people's lives from cellular network data. 9th International Conference, Pervasive 2011, San Francisco, USA, June 12-15.

Jeanne Breen Consulting, 2009. Car telephone use and road safety. Final report. An overview prepared for the European Commission. North Yorkshire, UK.

Järv, O., Ahas, R., Saluveer,E., Derudder,B., Witlox, F., 2012. Mobile Phones in a Traffic Flow: A Geographical Perspective to Evening Rush Hour Traffic Analysis Using Call Detail Records. PLoS ONE 7(11): e49171.

Karlsson, J.M., Gundlegard, D., 2006. Generating Road Traffic information from cellular Networks-New Possibilities in UMTS. 6th International conference on ITS telecommunications proceedings, Chegdu, China, June 21-23.

Kayri, M., Çokluk, Ö., 2010. Using multinomial logisitc regression analysis in artificial neural network: An application. Ozean Journal of Applied Sciences, 3(2), 259-268.

Klapka, P., Frantál, B., Halás, M., Kunc, J., 2010. Spatial organization: development, structure and approximation of geographical systems. Moravian geographical reports, 18 (3), 53-66.

Knaap, G.J., 2004. Monitoring land & housing markets: An essential tool for smart growth. National Center for Smart Growth Research & Education, Maryland, United States.

Kühne, R.D., 2008. Foundations of Traffic Flow Theory I: Greenshields' Legacy-Highway Traffic, Symposium on the Fundamental Diagram: 75 Years (Greenshields 75 Symposium). Transportation Research Board, Woods Hole Massachusetts, United States.

Küpper, A., 2007. Location-based services, Fundamentals and operations. John Wiley & Sons, Ltd.

Leduc, G., 2008. Road traffic data: Collection methods and applications, Working Papers on Energy, Transport and Climate Change N.1. European Comission-Joint Research Centre- Institute for Prospective Technologica.

Lingras, P., Adamo, M., 1996. Average and peak traffic volumes: neural nets, regression, and factor approaches. Journal of Computing in Civil Engineering, 10(4), 300-306.

Liu, H.X., Danczyk, A., Brewer, R., Starr, R., 2008. Evaluation of cellphone traffic data in Minnesota. Transportation Research Record: Journal of the Transportation Research Board, 2086 (2008), 1-7.

Lu, Y., Liu, Y., 2012. Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies. Computers, Environment and Urban Systems, 36 (2), 105-108.

Ludden, B., Pickford, A., Medland, J., Johnson, H., Brandon, F., Axelsson, L. E., Viddal-Ervik, K., Dorgelo, B., Boroski, E., Malenstein, J., 2002. Coordination group on access to Location Information for Emergency Services (E112). Cgalies plenary.

Mark A. Marek, P.E., 2010. Roadway Design Manual. Texas Department of Transportation, US, Texas.

Martínez, L. M., Viegas, J.M., Silva, E.A, 2009. A traffic analysis zone definition: a new methodology and algorithm. Transportation 36(5), 581-599.

Mullen, P., 1975. Estimating the demand for urban bus travel. Transportation, 4(1975), 231-252.

Nasr, G.E., Badr, E.A., Joun, C., 2003. Backpropagation neural networks for modeling gasoline consumption. Energy Conversion and Management, 44 (6), 893-905.

Nobis, C., Lenz, B., 2009. Communication and mobility behavior - A trend and panel analysis of the correlation between mobile phone use and mobility. Journal of Transport Geography, 17 (2), 93-103.

OECD, 2007. Managing urban traffic congestion. European conference of ministers of transport. http://www.internationaltransportforum.org/Pub/pdf/07Congestion.pdf (Accessed 07.12.2011).

Oliveira, V., Pinho, P., 2010. City profile Lisbon. Cities, 27 (2010), 405–419.

Pan, C., Lu, J., Di, S., Ran, B., 2006. Cellular-Based Data-Extracting Method for Trip Distribution. Transportation Research Record: Journal of the Transportation Research Board, (2006-1945), 33-39.

Pulselli, R.M., Romano, P., Ratti, C., Tiezzi, E., 2008. Computing urban mobile land scapes through monitoring population density based on cell-phone chatting. International Journal of design and nature and ecodynamics, 3 (2), 121-134.

Puntumapon, K., Pattara-atikom, W., 2008. Classification of cellular phone mobility using Naive Bayes model. In: Vehicular Technology Conference. IEEE, Singapore.

Rao, P.V., Sikdar, P.K., Rao, K.V., Dhingra, S.L., 1998. Another insight into artificial neural networks through behavioural analysis of access mode choice. Computers, Environment and Urban Systems, 22 (5), 485-496.

Ratti, C., Pulselli, R.M., Pulselli, R.M., Williams, S., Frenchman, D., 2006. Mobile Landscapes: Using location data from cellphones for urban analysis. Environment and Planning B: Planning and Design, 33 (5), 727-748.

Ratti, C., Sevtsuk, A., Huang, S., Pailer, R., 2005. Mobile landscapes: Graz in real time. Proceedings of the 3rd Symposium on LBS & TeleCartography, Vienna, Austria, November 28-30.

Reades, J., Calabrese, F., Ratti, C., 2009. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. Environment and Planning B: Planning and Design, 36(5), 824-836.

Reades, J., Calabrese, F., Sevtsuk, A., Ratti, C., 2007. Cellular census: Explorations in urban data collection. IEEE Pervasive computing, 6 (3), 30-38.

Rodrigue, J-P., Comtois, C., Slack, B., 2006. The Geography of Transport Systems. Routledge Taylor and Francis Group, London and NewYork.

Sarle, W.S., 1994. Neural networks and statistical models, Proceedings of the nineteenth annual SAS users group international conference, Dallas, Texas.

Sato-Ilic, M., Jain, L.C., 2006. Introduction to Fuzzy Clustering. In: Kacprzyk, J. (Ed.), Innovations in Fuzzy Clustering: Theory and Applications. Springer-Verlag Berlin Heidelberg. PP. 1-8.

Schwammle, V., Jensen, O.N., 2010. A simple and fast method to determine the parameters for fuzzy c–means cluster validation. http://arxiv.org/pdf/1004.1307.pdf (Accessed 10.10.2013).

Sevtsuk, A., Ratti, C., 2008. Explorations into urban mobility patterns using aggregate mobile network data. http://senseable.mit.edu/papers/pdf/2008_Sevtsuk_Ratti_Journal%20of%20Urban%20Technologies.pdf (Accessed 24.10.13).

Smith, B.L., Fontaine, M.D., Beaton, J., Dejarnette, E., Hendricks, A., Tennant, L.L., 2007. Private - Sector provision of congestion data.

http://onlinepubs.trb.org/onlinepubs/trbnet/brd/NCHRP_70-01_Contractor_Final_Report.pdf (Accessed 06.12. 2011).

Sohn, T., Varshavsky, A., LaMarca, A., Chen, M.Y., Choudhury, T., Smith, I., Consolvo, S., Hightower, J., Griswold, W.G., Lara, E.d., 2006. Mobility detection using everyday GSM traces. Proceedings of the 8th international conference on Ubiquitous Computing. Springer-Verlag Berlin Heidelberg 2006, Orange County, CA, USA, 212-224.

Soto, V., Frías-Martínez, E., 2011. Robust land use characterization of urban landscapes using cellphone data. Workshop on pervasive urban applications in conjunction with 9th international conference on pervasive computing, San Francisco, CA, June 12-15.

Spinney J. E., 2003. Mobile positioning and LBS Applications. Geography, 88 (4) 256-265.

Stephenson, W.R., Cook, D., Dixon, P., Duckworth, W.M., Kaiser, M.S., Koehler, K., Meeker, W.Q., 2001. Binary response and logistic regression analysis, Advanced Statistical Methods for Research Workers.

Thajchayapong, S., Pattara-atikom, W., Chadil, N., Mitrpant, C., 2006. Enhanced detection of road traffic congestion areas using cell dwell times. 2006 IEEE intelligent transportation systems conference. IEEE, Toronto, Canada, September 17-20.

Thiessenhusen, K.-U., Schäfer, R.-P., Lang, T., 2003. Traffic Data from Cellphones: A Comparison with Loops and Probe Vehicle Data.

Toole, J., Ulm, M., González, M., Bauer, D. 2012. Inferring land use from mobile phone activity. The 18th ACM SIGKDD international workshop on urban computing, Beijing, China, August 12-16.

Train, K.E., 2009. Discrete Choice Methods with Simulation, Second ed. Cambridge University Press, New York, USA.

TRB, 2000. Highway capacity manual. Transportation Research Board, US, Washington, D.C.

University of Maryland Transportation Studies Center, 1997. Final evaluation report for the CAPITAL-ITS operational test and demonstration program. University of Maryland, College Park, US.

Vaccari, A., Liu, L., Biderman, A., Ratti, C., Pereira, F., Oliveirinha, J., Gerber, A., 2009. A holistic framework for the study of urban traces and the profiling of urban processes

and dynamics. 12th International IEEE Conference on Intelligent Transportation Systems. IEEE, St. Louis, MO, October 04-07.

Valerio, D., 2009. Road traffic information from cellular network signaling. Forschungszentrum Telekommunikation Wien, Vienna, Austria.

Valerio, D., Alconzo, A.D., Ricciato, F., Wiedermann, W., 2009a. Exploiting cellular networks for road traffic estimation : A survey and a research roadmap. IEEE 69th Vehicular technology conference, Barcelona, Spain, April 26-29.

Valerio, D., Witek, T., Ricciato, F., Pilz, R., Wiedermann, W., 2009b. Road traffic estimation from cellular network monitoring : A hands-on investigation. Personal, Indoor and Mobile Radio Communications, 2009 IEEE 20th International Symposium on, Tokyo, Japan, 3035- 3039.

Varaiya, P., Haoui, A., Kavaler, R., 2008. Wireless magnetic sensors for traffic surveillance. Transportation Research Part C: Emerging Technologies, 16(3), 294-306.

White, J., Wells, I., 2002. Extracting origin destination information from mobile phone data. 11th International Conference on Road Transportation and Control, London, 30-34.

Williams, B., 2011. Sustainable urban transport in africa: issues and challenges. Available at: http://ebookbrowse.com/sustainable-urban-transport-in-africa-brian-williams-pdf-d64725155 (Accessed 27.04.2013).

Yim, Y., 2003. The State of Cellular Probes. Technical report, California PATH Research Project, University of California, CA, USA, July 2003.

Yuan, Y., Raubal,M., Liu, Y., 2012. Correlating mobile phone usage and travel behavior – A case study of Harbin, China. Computers, Environment and Urban Systems, 36 (2), 118-130.

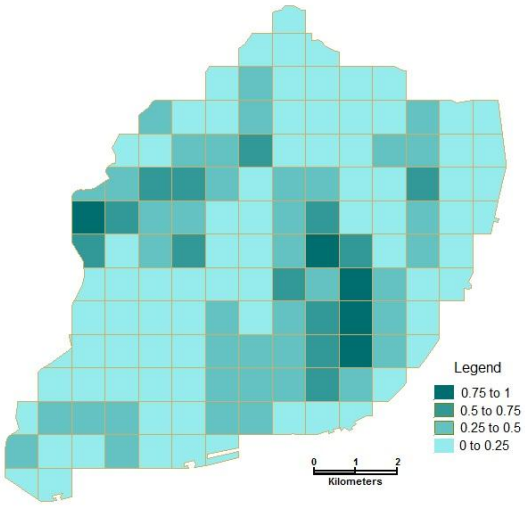
Zeng, Q.-A., Agrawal, D.P., 2002. Handoff in Wireless Mobile Networks, in: Stojmenovic, I. (Ed.), Handbook of wireless Networks and Mobile computing. John Wiley & Sons, Inc.1-26.

Zhang, J., Stojmenovic, I., 2005. Cellular Networks. http://www.site.uottawa.ca/~ivan/cellular.pdf (Accessed 18.11.2013).

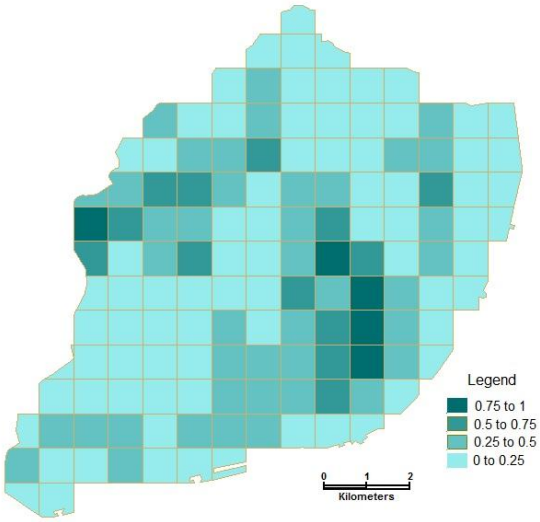
Zuo, X., Zhang, Y., 2012. Detection and analysis of urban area hotspots based on cellphone traffic. Journal of Computers, 7(7), 1753-1760.

Appendixes

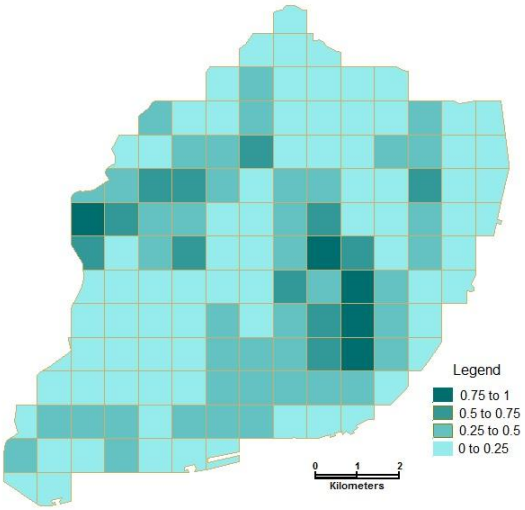
Appendixes A1: Presence of people in Lisbon along the hours of the day.



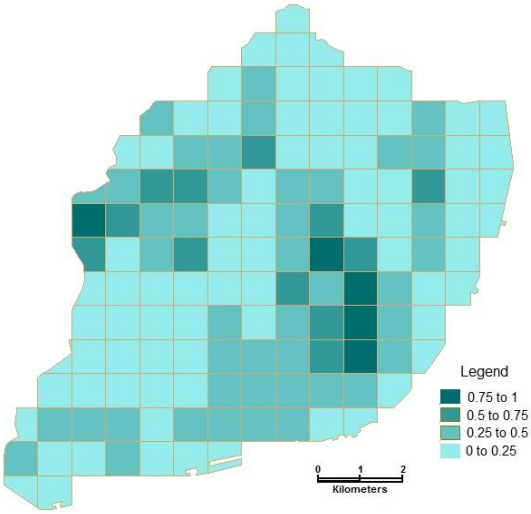
Presence of people: Midnight



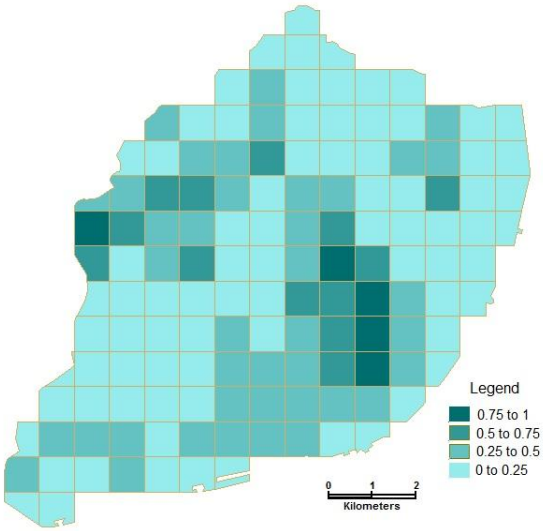
Presence of people: 1AM



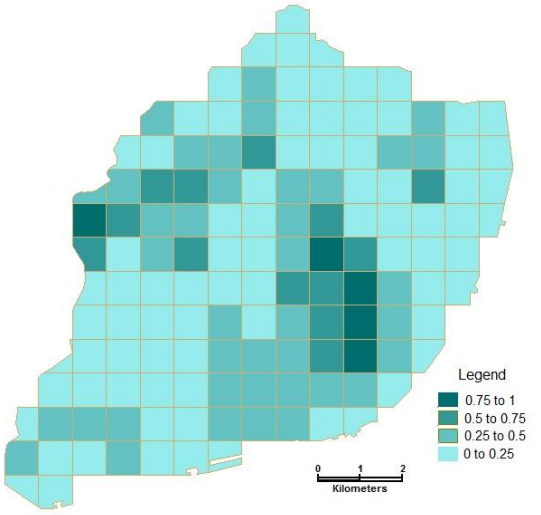
Presence of people: 2AM



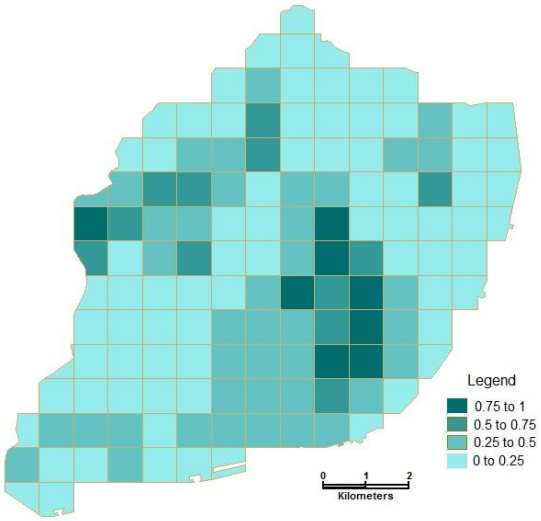
Presence of people: 3AM



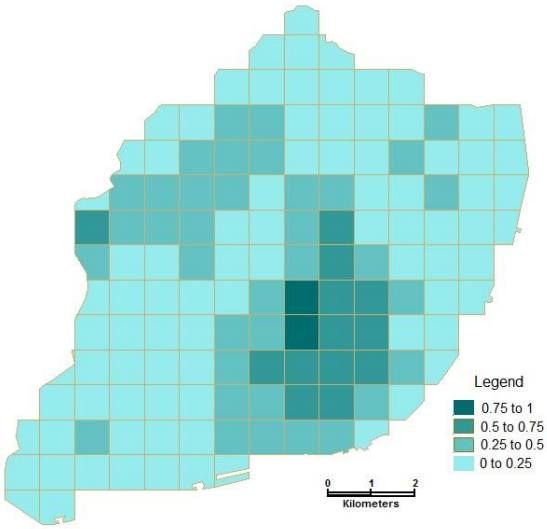
Presence of people: 4AM



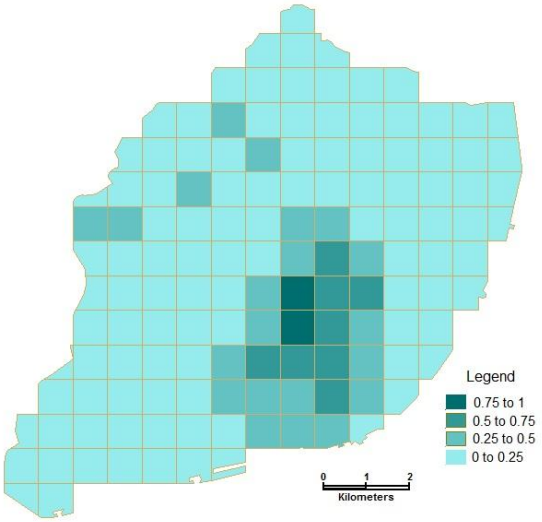
Presence of people: 5AM

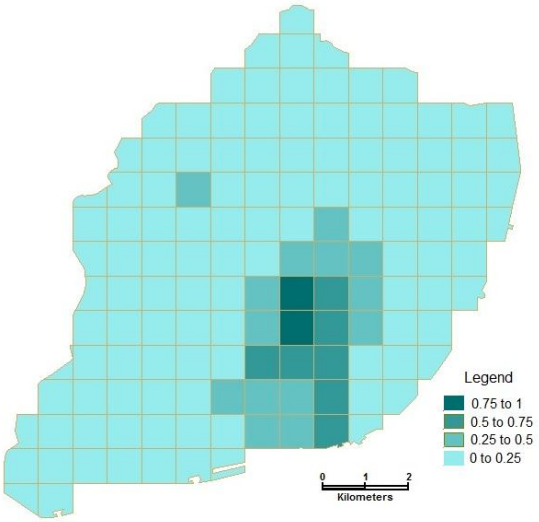


Presence of people: 6AM

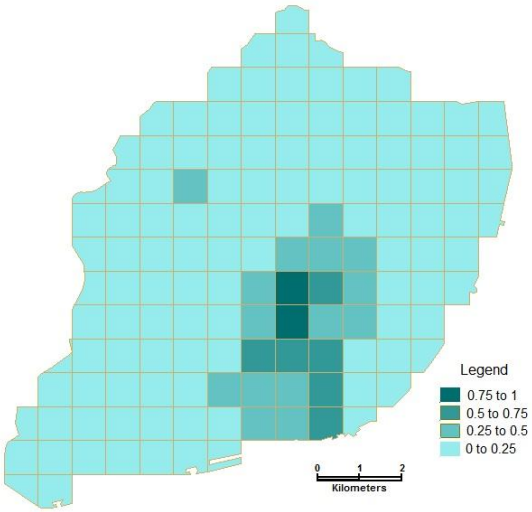


Presence of people: 7AM

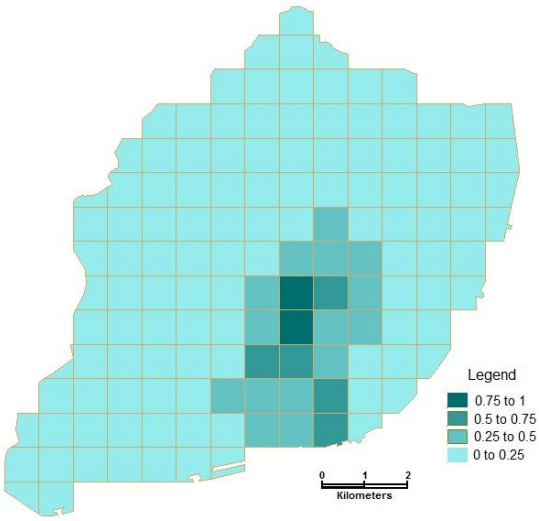




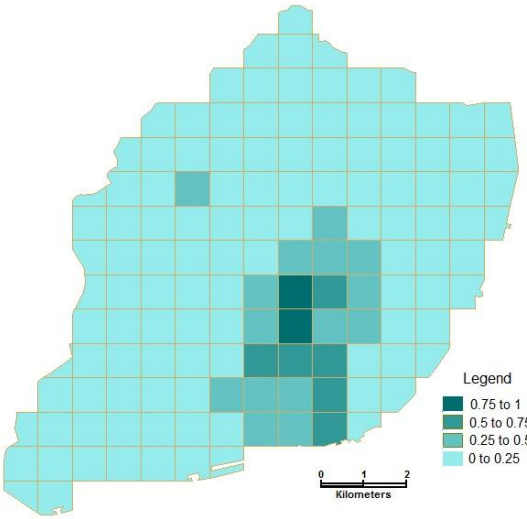
Presence of people: 8AM



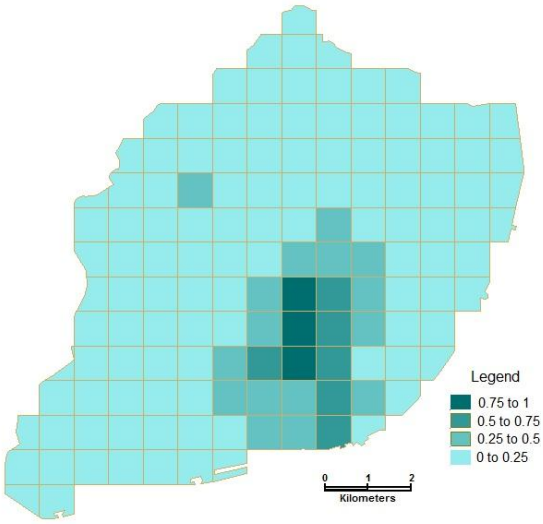
Presence of people: 9AM



Presence of people: 10AM

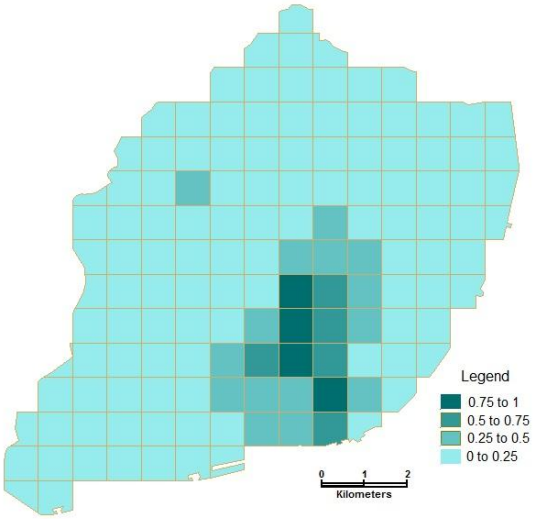


Presence of people: 11AM

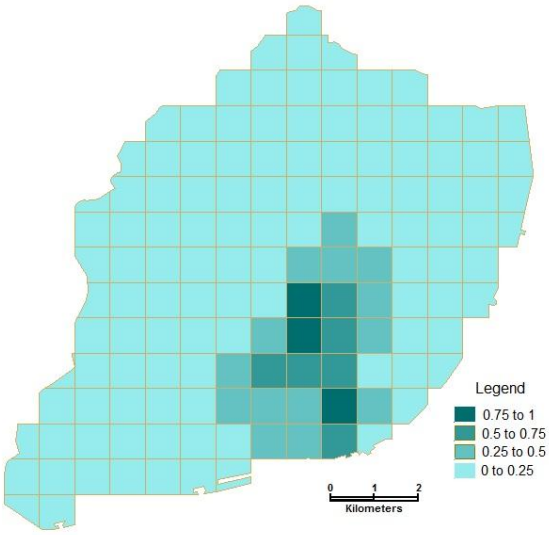


Presence of people: NOON

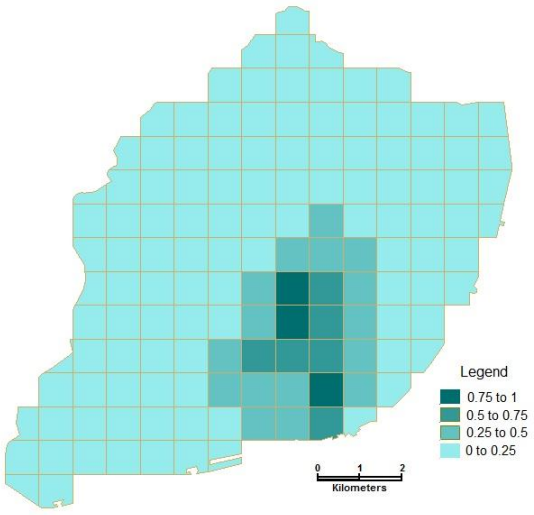
Presence of people: 1PM



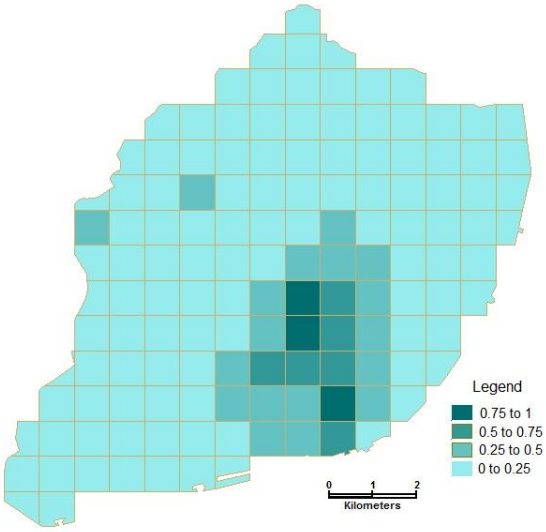
Presence of people: 2PM



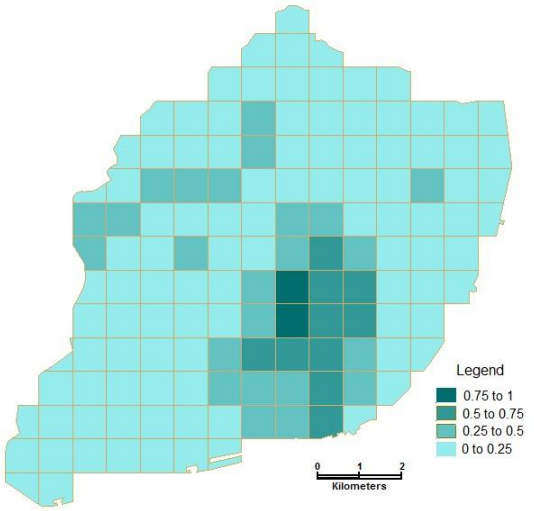
Presence of people: 3PM

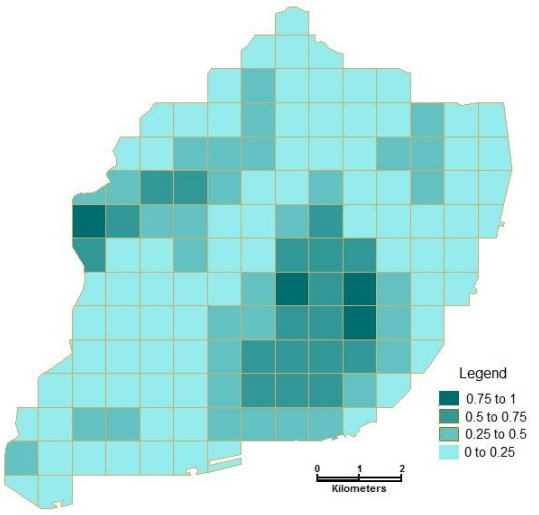


Presence of people: 4PM

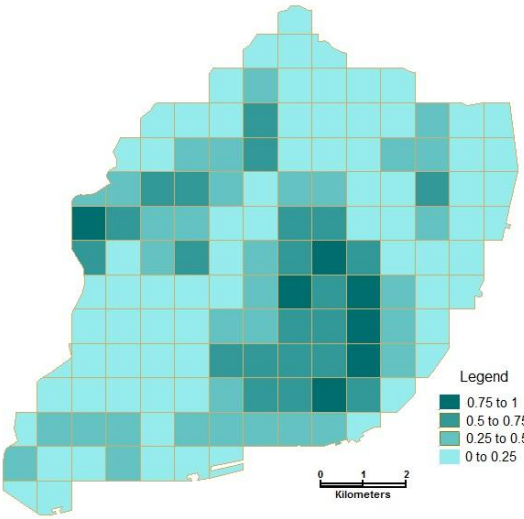


Presence of people: 5PM

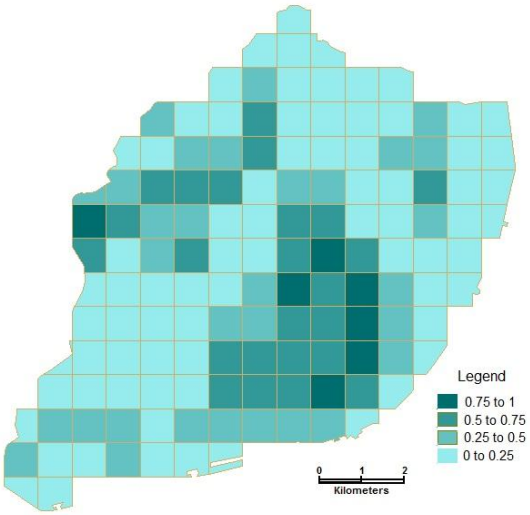




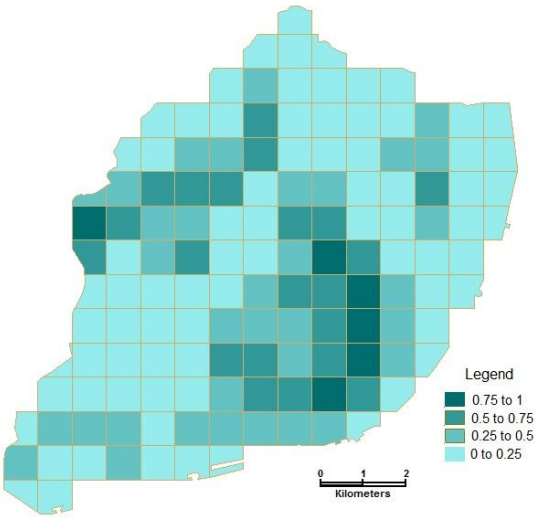
Presence of people: 6PM



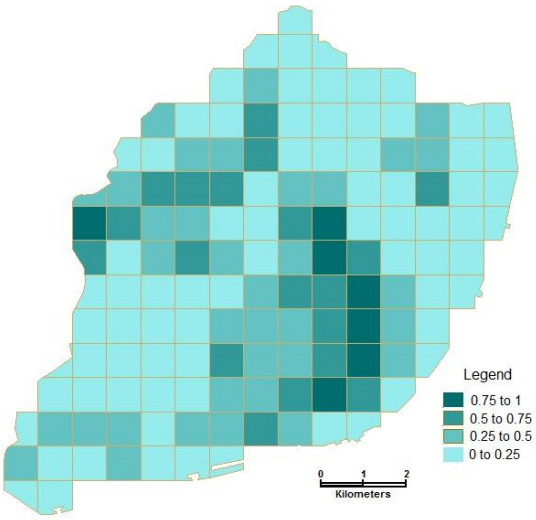
Presence of people: 7PM



Presence of people: 8PM



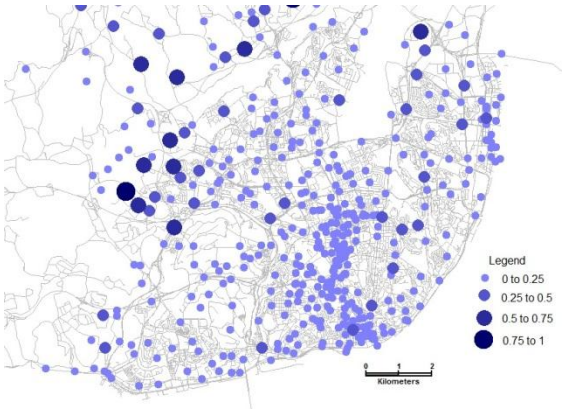
Presence of people: 9PM



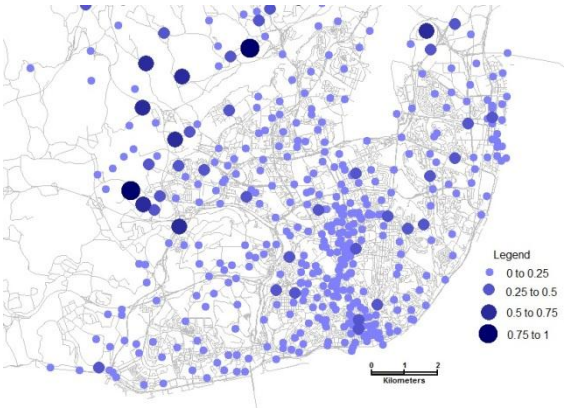
Presence of people: 10PM

Presence of people:11PM

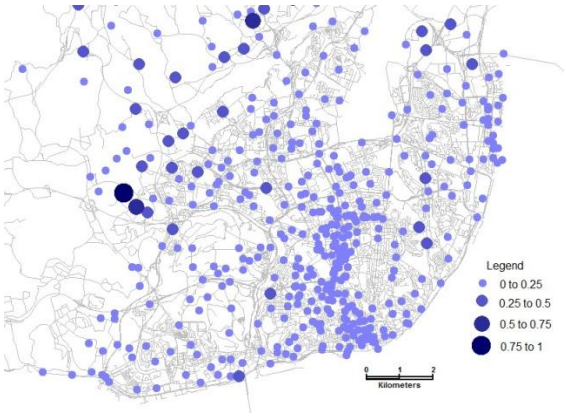
Appendixes A2: Intensity of Call Volumes in Lisbon along the hours of the day on April 12, 2010.



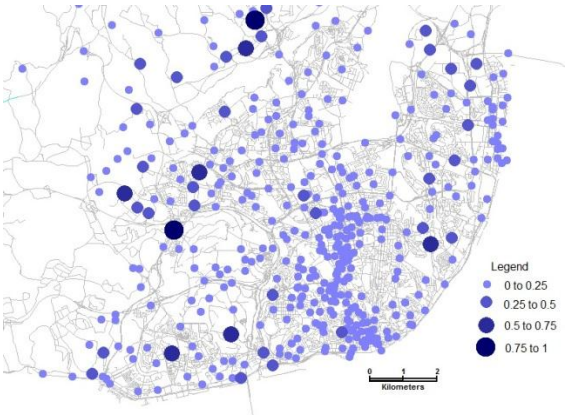
Call Volume: Midnight



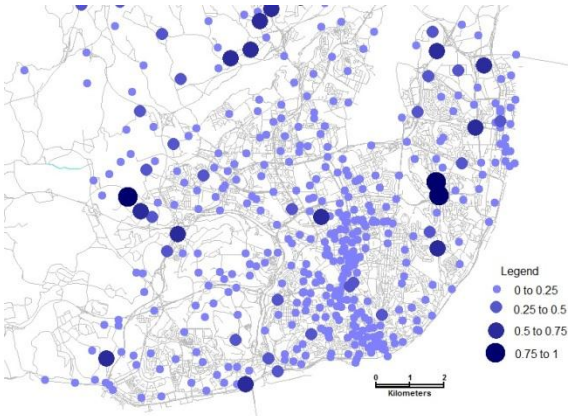
Call Volume: 1AM



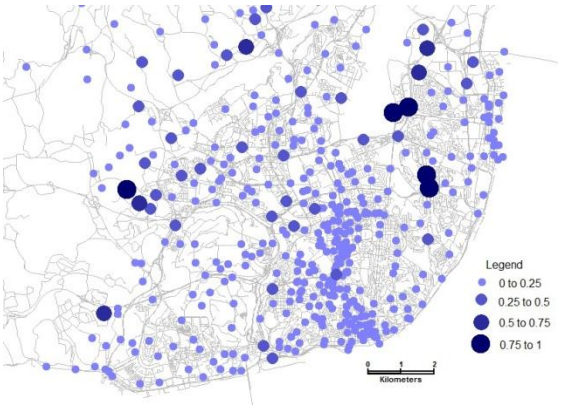
Call Volume: 2AM



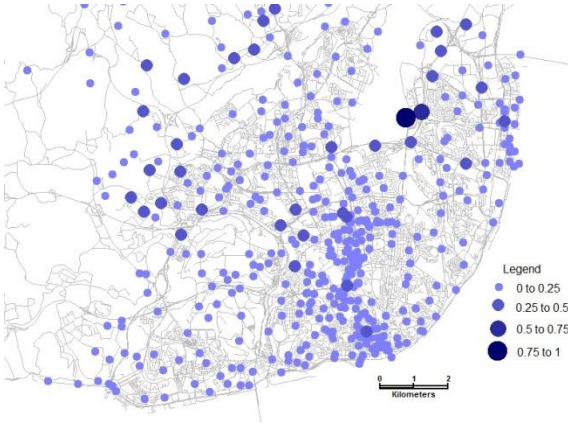
Call Volume: 3AM



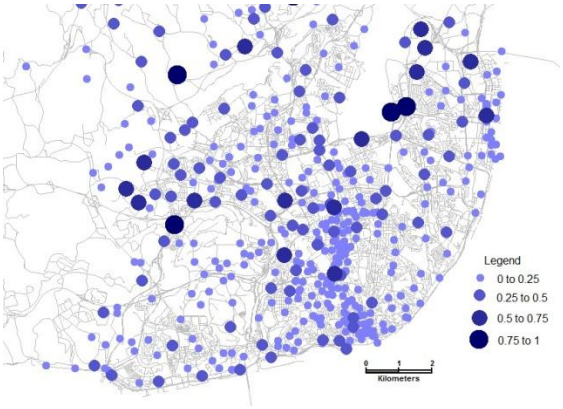
Call Volume: 4AM



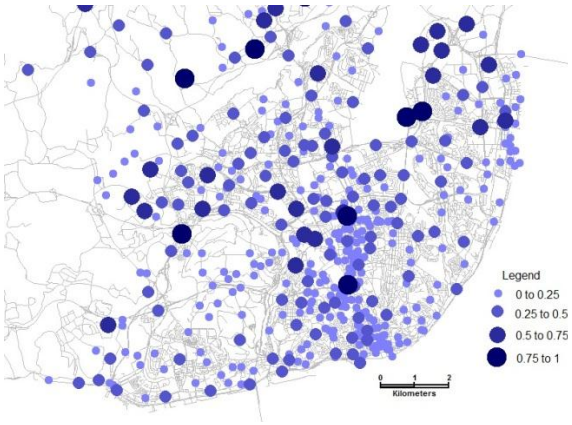
Call Volume: 5AM



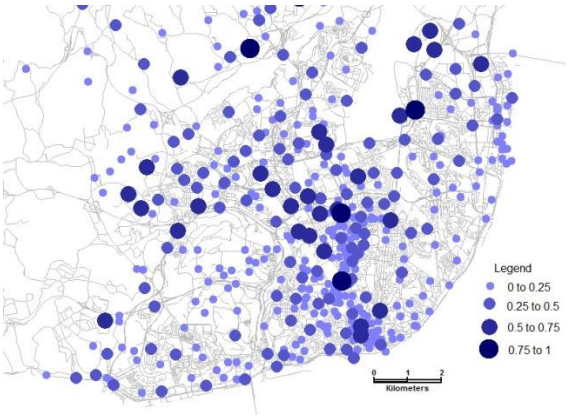
Call Volume: 6AM



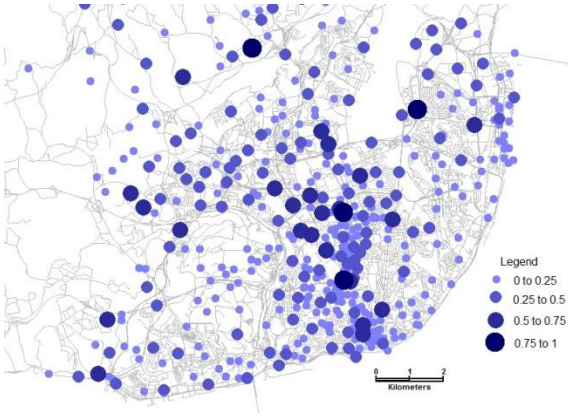
Call Volume: 7AM



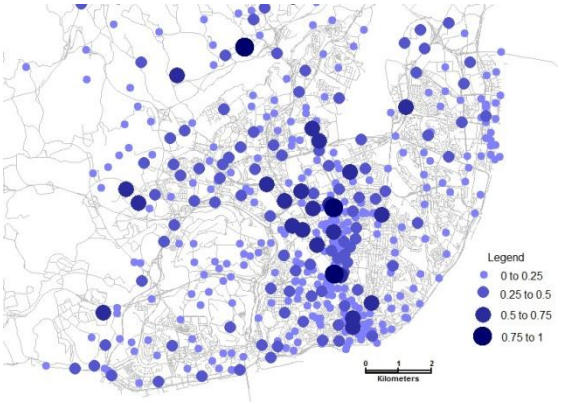
Call Volume: 8AM



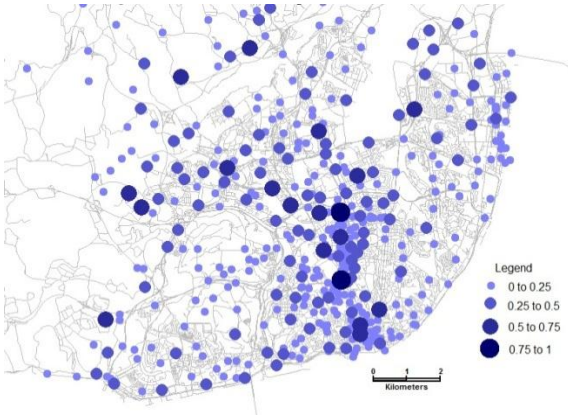
Call Volume: 9AM



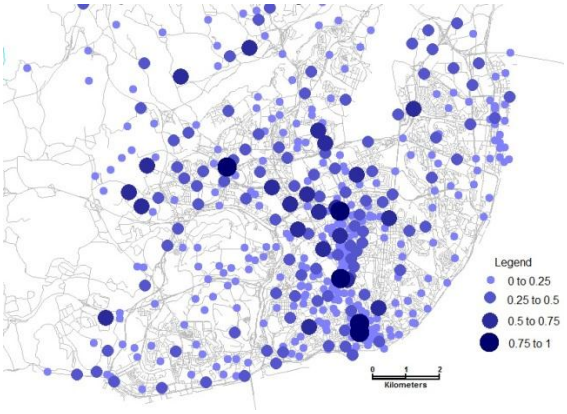
Call Volume: 10AM



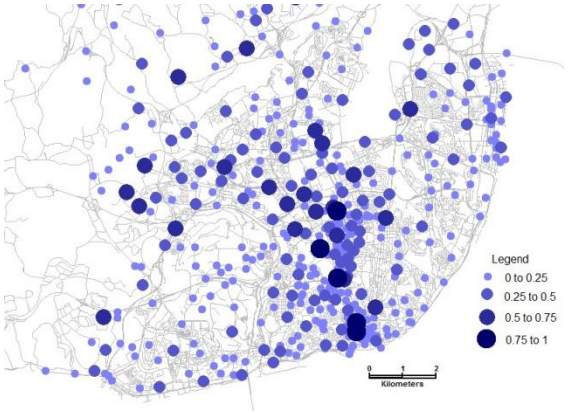
Call Volume: 11AM



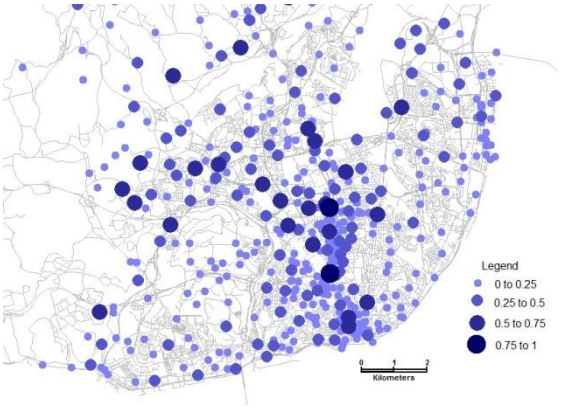
Call Volume: NOON



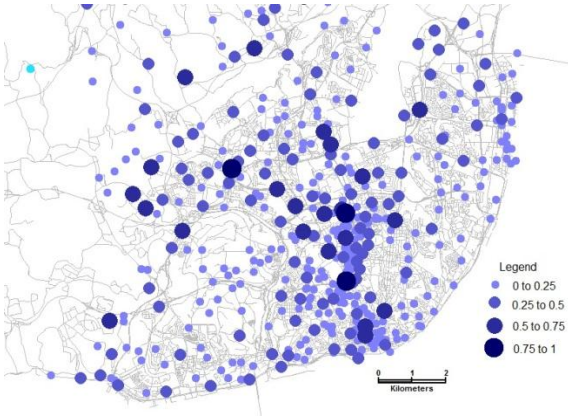
Call Volume: 1PM



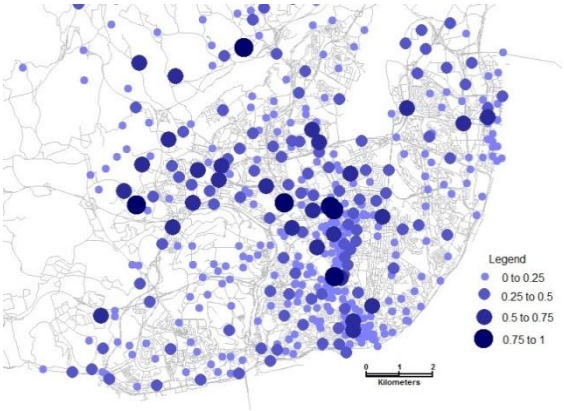
Call Volume: 2PM



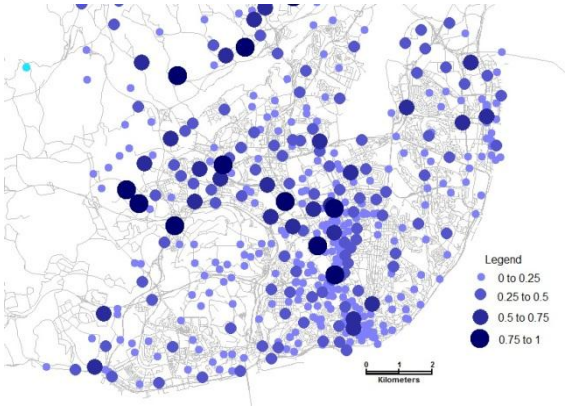
Call Volume: 3PM



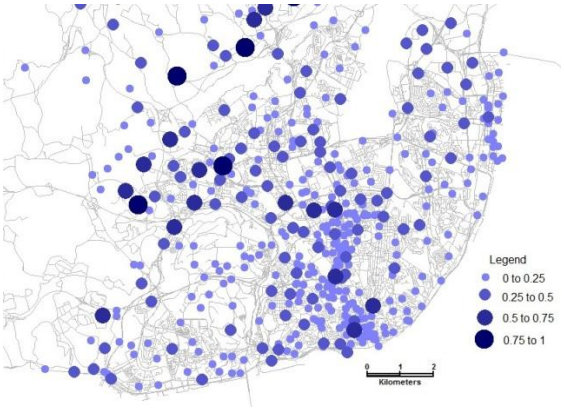
Call Volume: 4PM



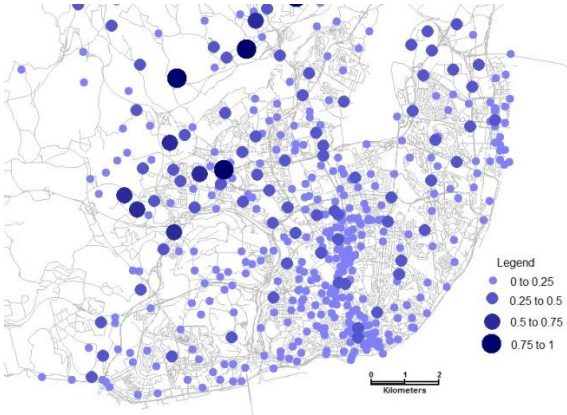
Call Volume: 5PM



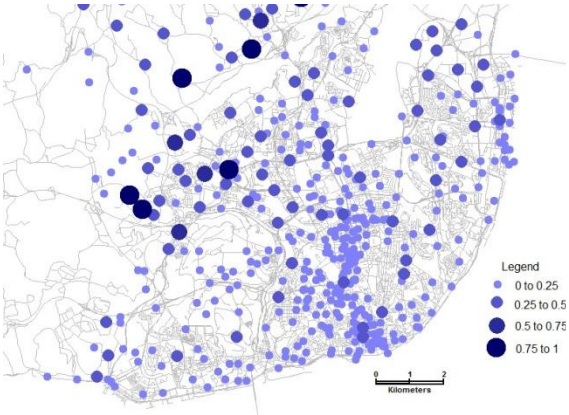
Call Volume: 6PM



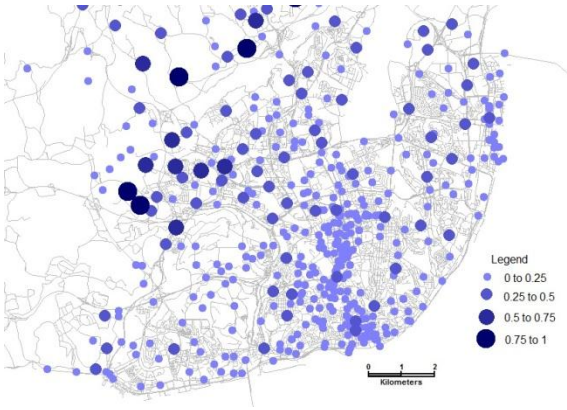
Call Volume: 7PM



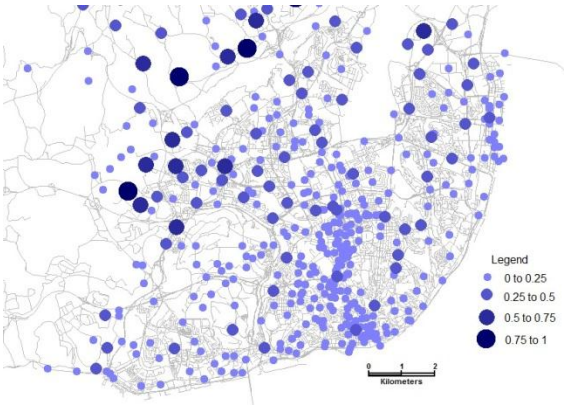
Call Volume: 8PM



Call Volume: 9PM

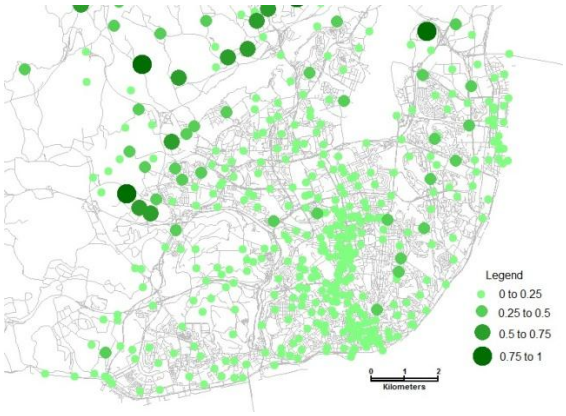


Call Volume: 10PM

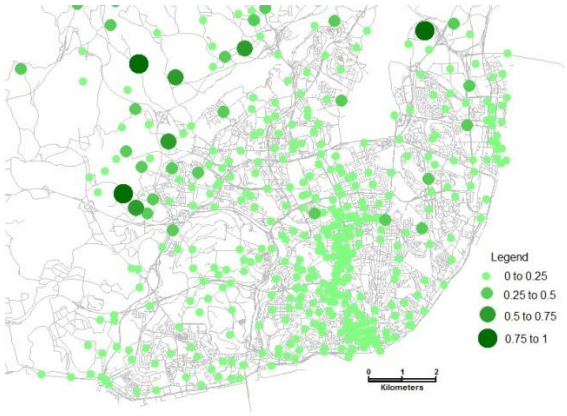


Call Volume: 11PM

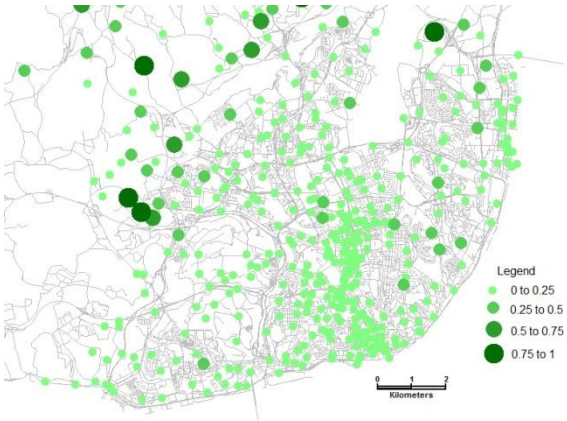
Appendixes A3: Intensity of Erlang values in Lisbon along the hours of the day on April 12, 2010.



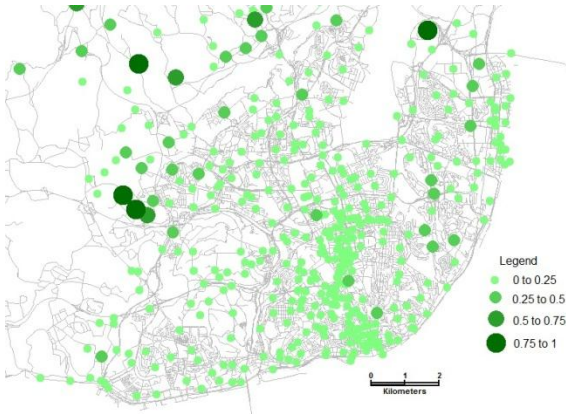
Erlang: Midnight



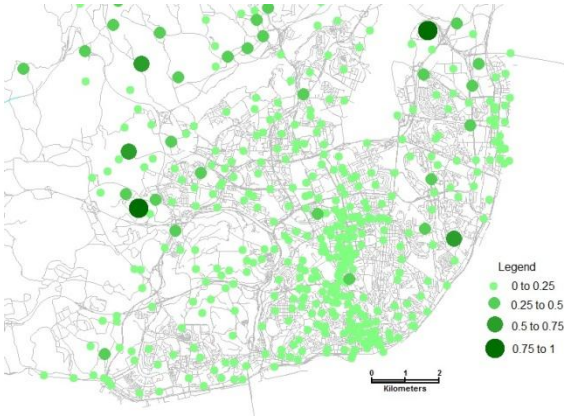
Erlang: 1AM



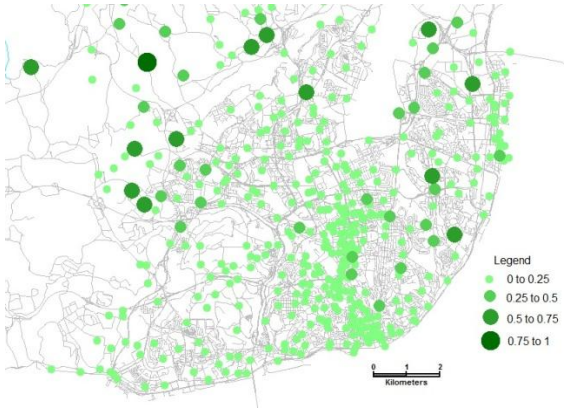
Erlang: 2AM



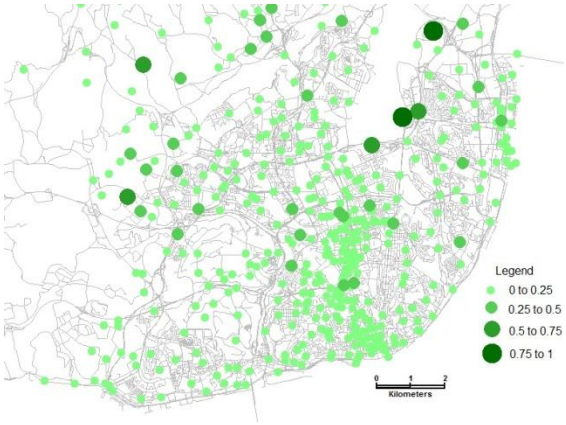
Erlang: 3AM



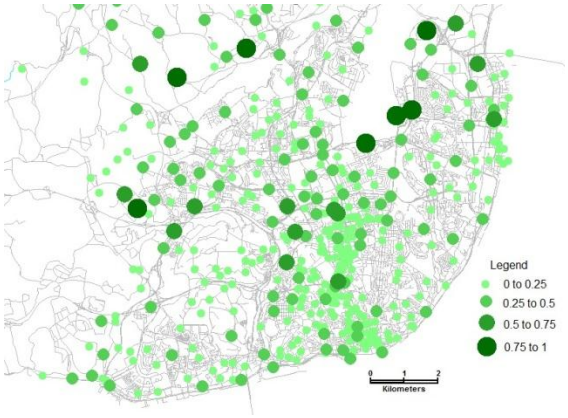
Erlang: 4AM



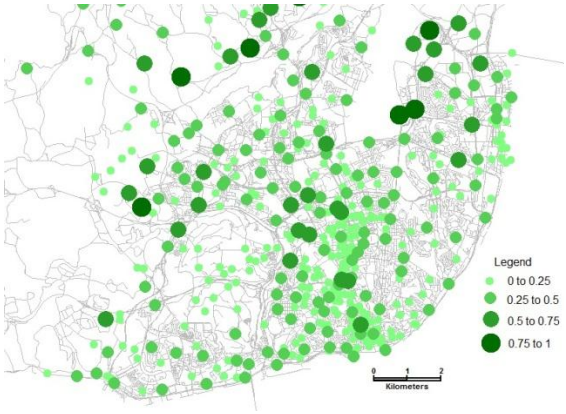
Erlang: 5AM



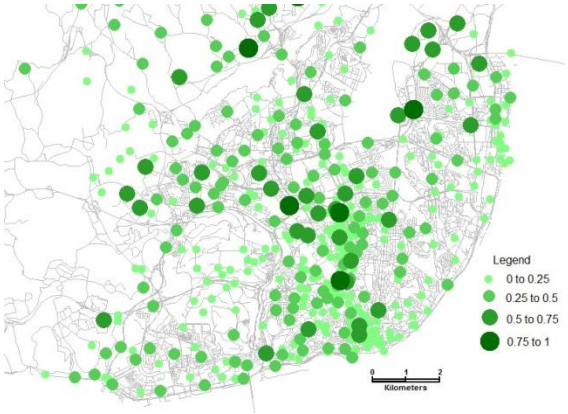
Erlang: 6AM



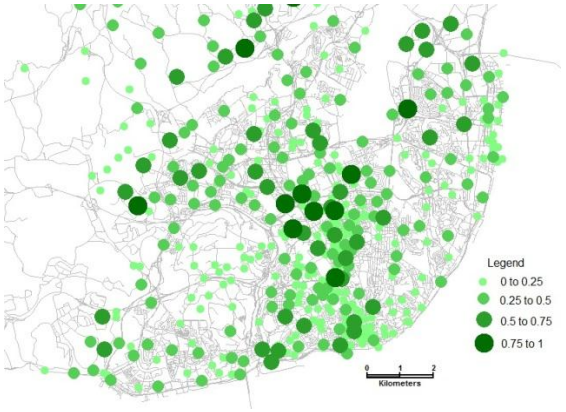
Erlang: 7AM



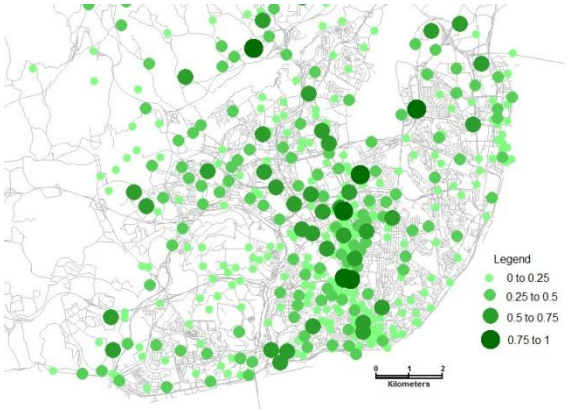
Erlang: 8AM



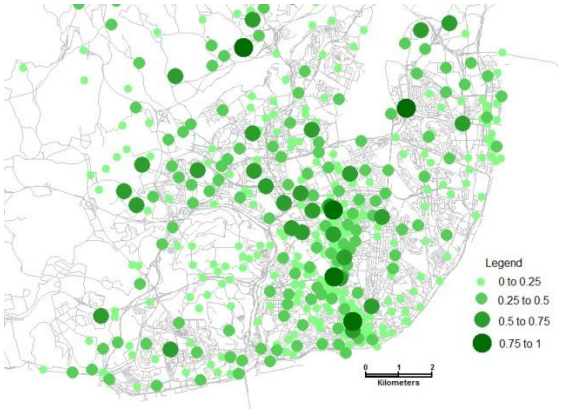
Erlang: 9AM



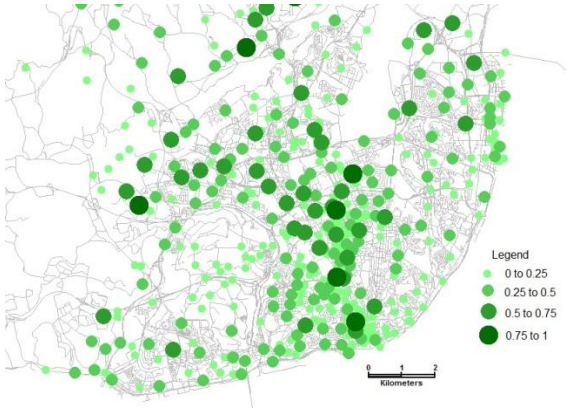
Erlang: 10AM



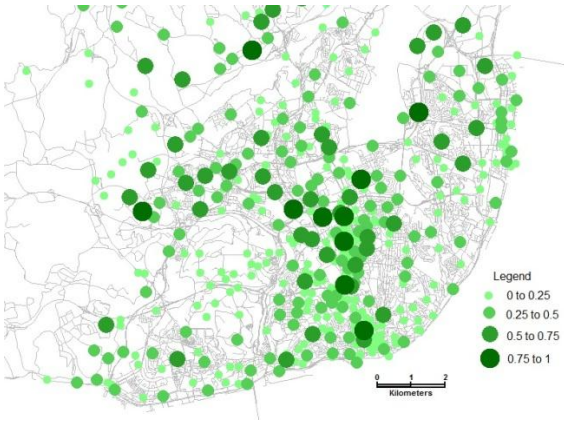
Erlang: 11AM



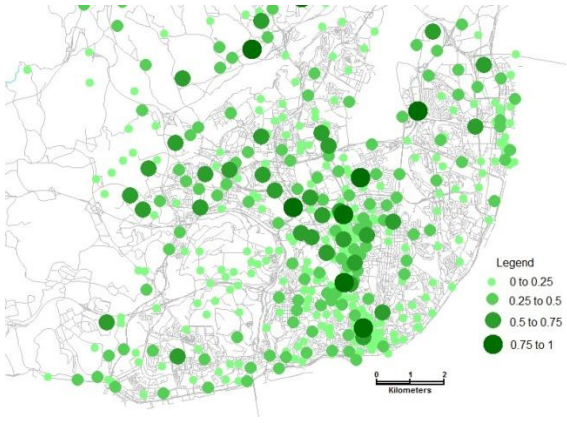
Erlang: NOON



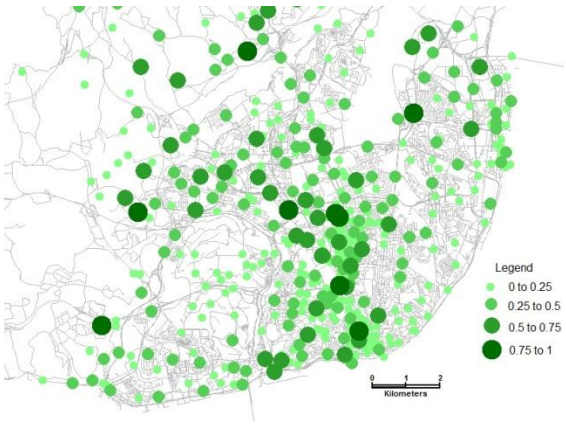
Erlang: 1PM



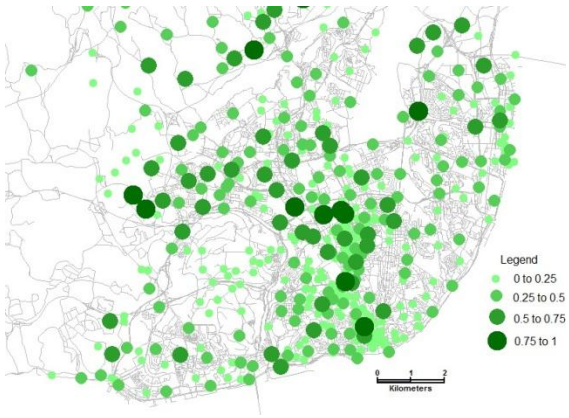
Erlang: 2PM



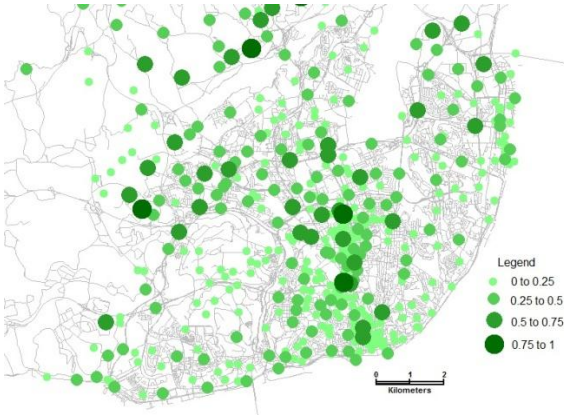
Erlang: 3PM



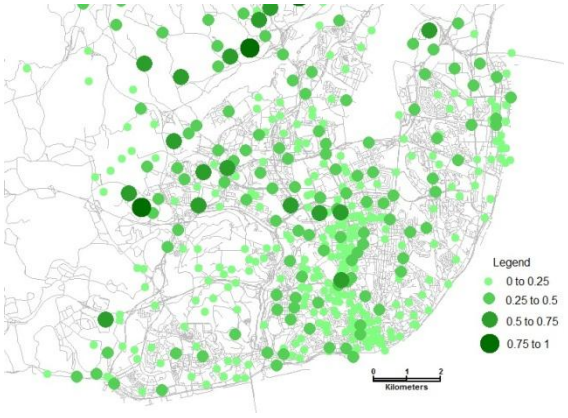
Erlang: 4PM



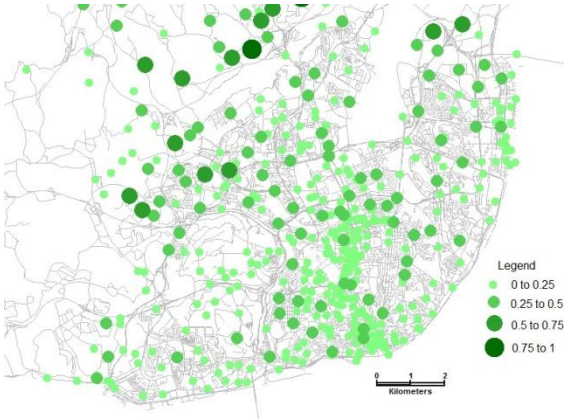
Erlang: 5PM



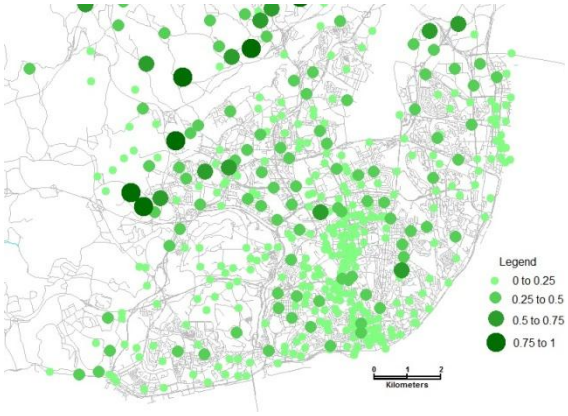
Erlang: 6PM



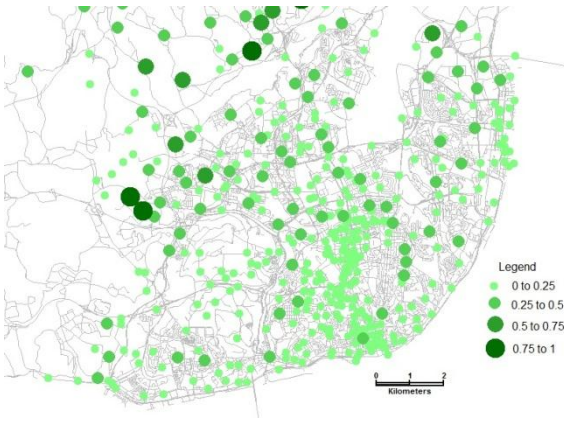
Erlang: 7PM



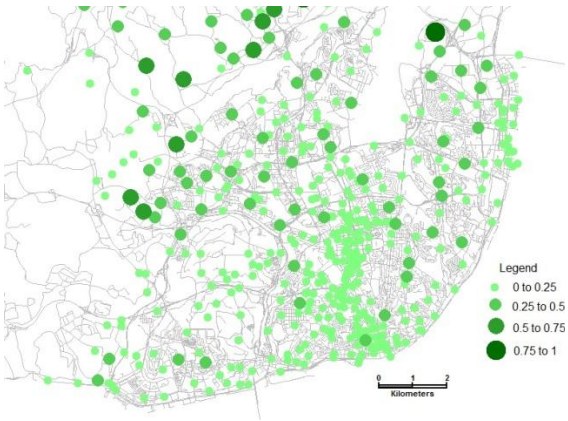
Erlang: 8PM



Erlang: 9PM

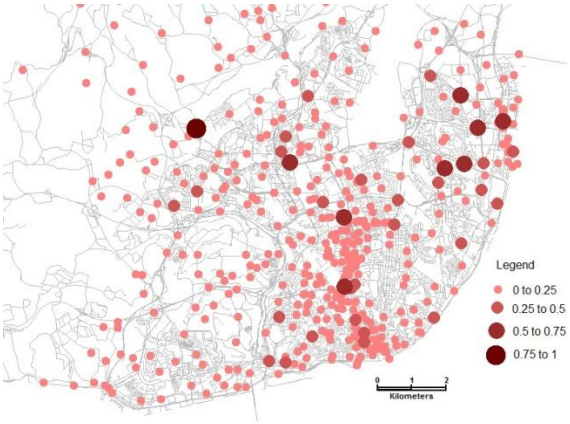


Erlang: 10PM

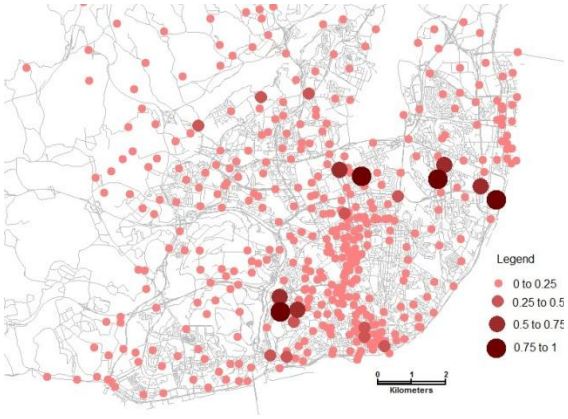


Erlang: 11PM

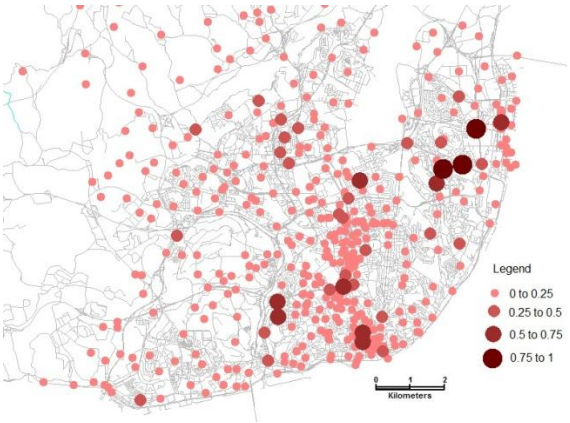
Appendixes A4: Intensity of Handover values in Lisbon along the hours of the day on April 12, 2010.



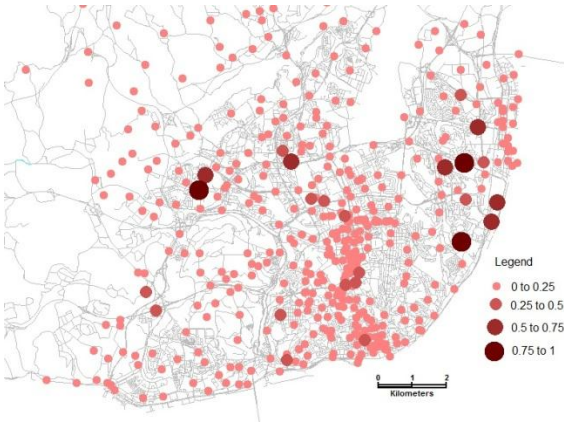
Handover: Midnight



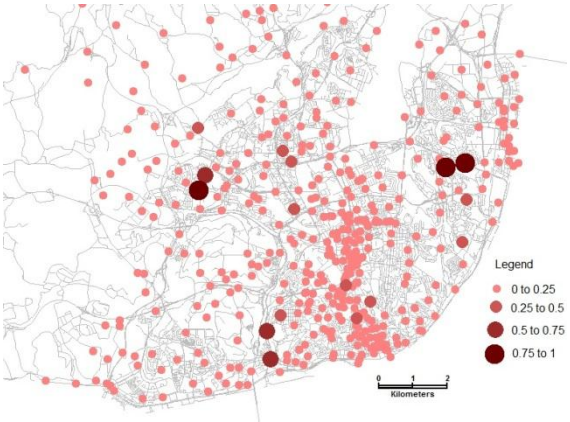
Handover: 1AM



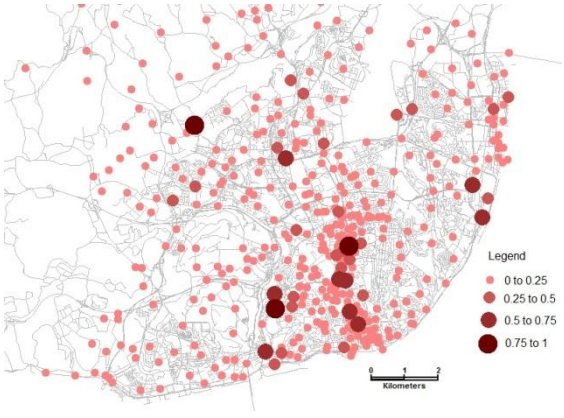
Handover: 2AM



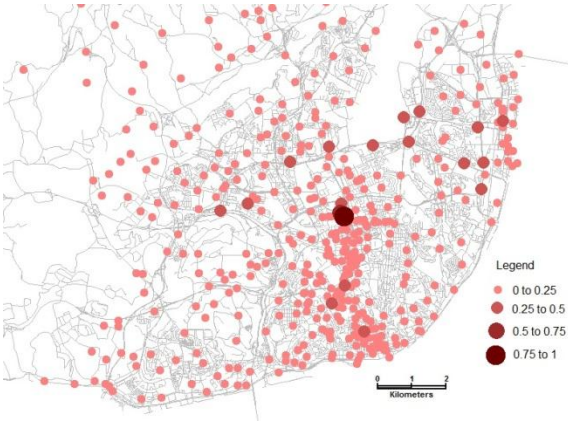
Handover: 3AM



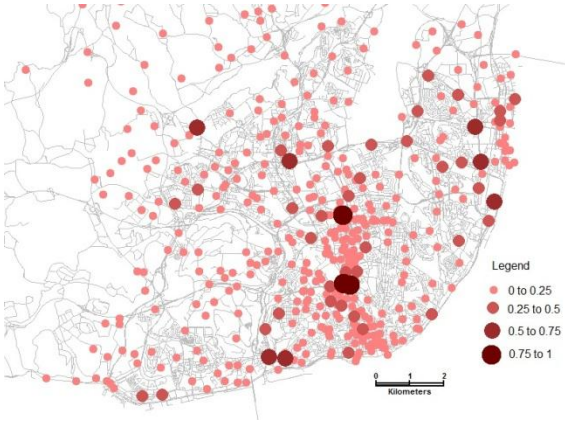
Handover: 4AM



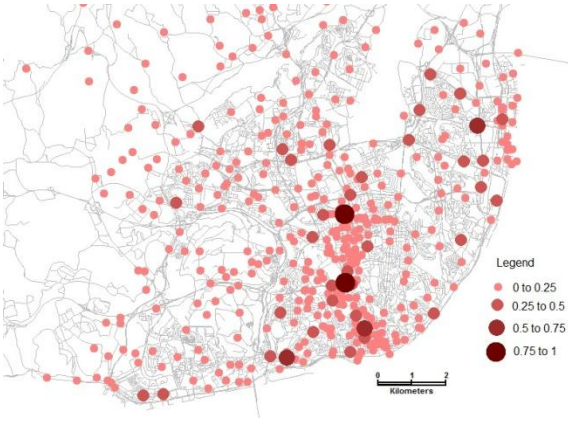
Handover: 5AM



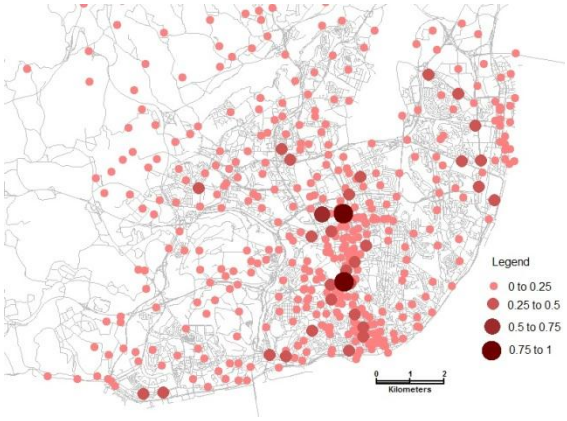
Handover: 6AM



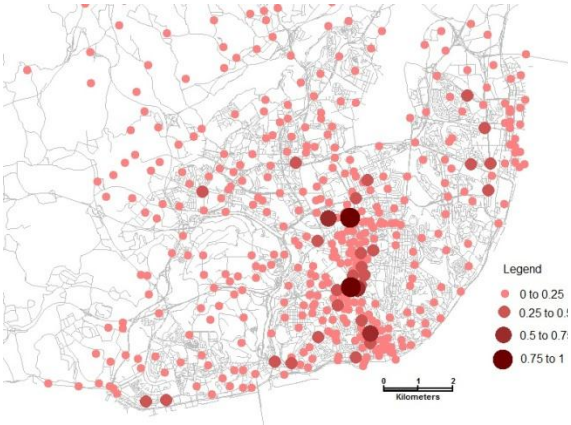
Handover: 7AM



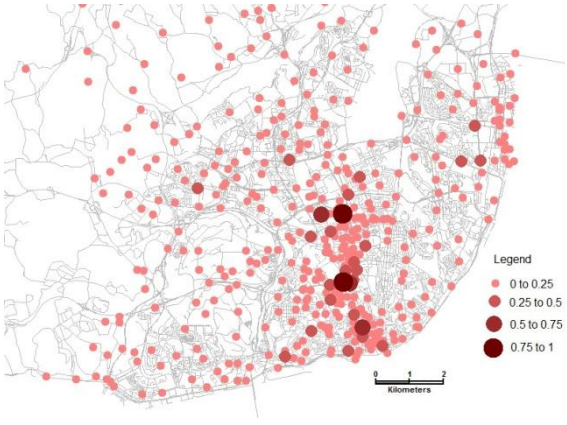
Handover: 8AM



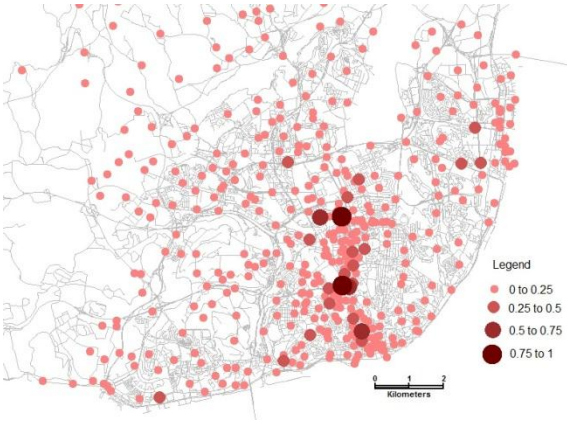
Handover: 9AM



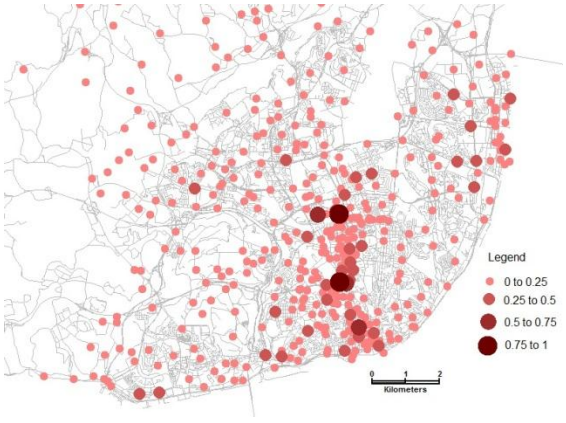
Handover: 10AM



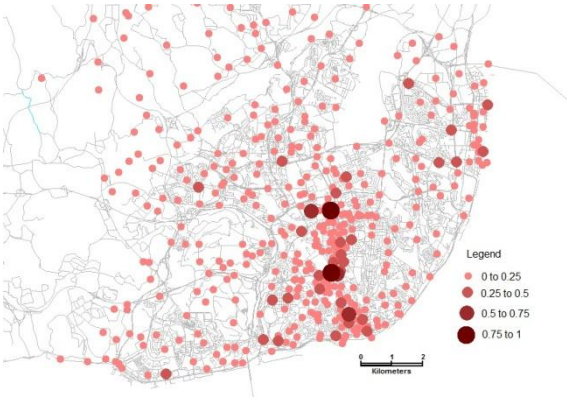
Handover: 11AM



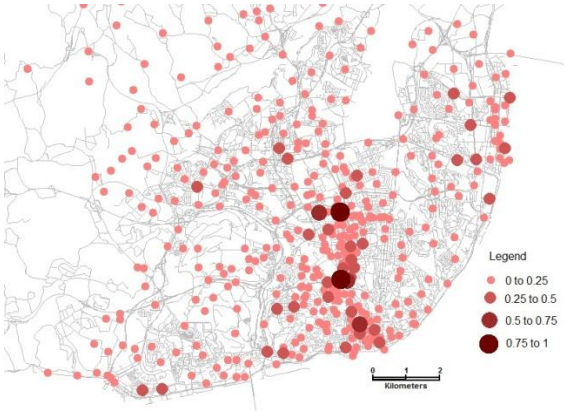
Handover: NOON



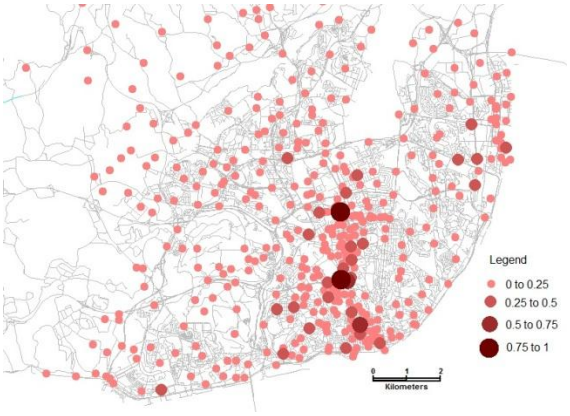
Handover: 1PM



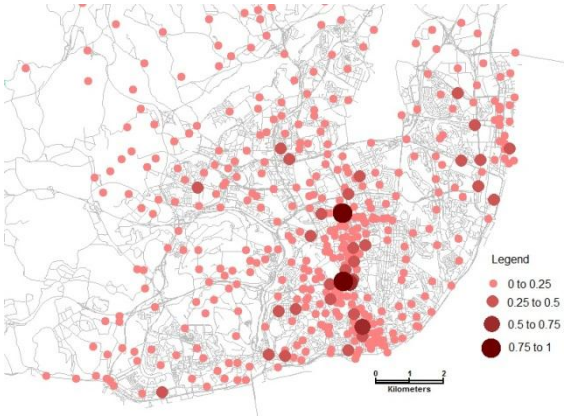
Handover: 2PM



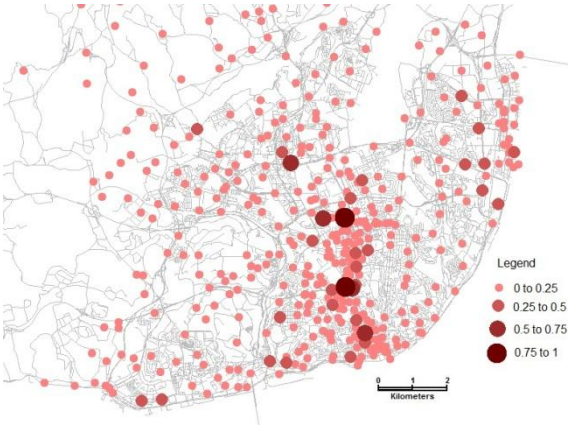
Handover: 3PM



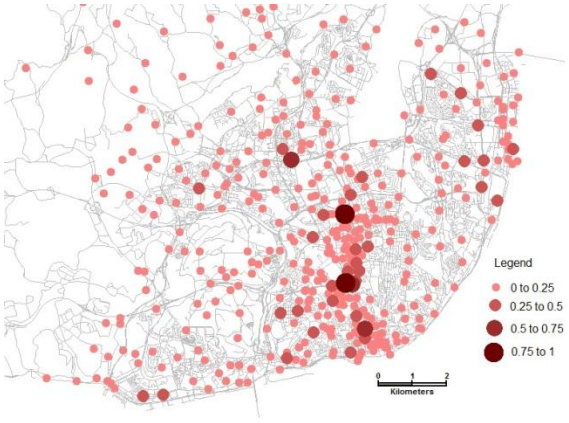
Handover: 4PM



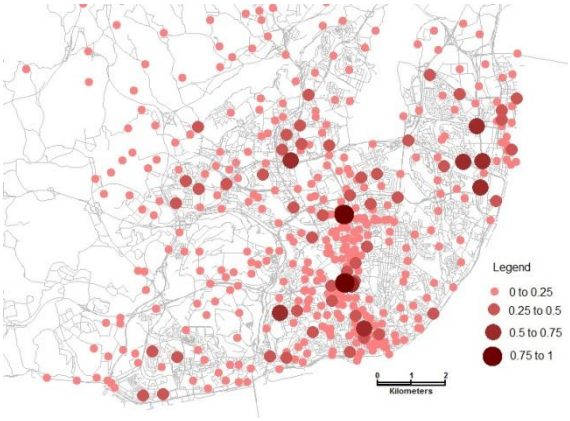
Handover: 5PM



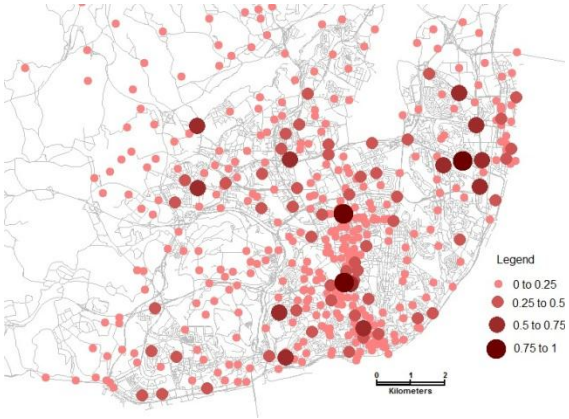
Handover: 6PM



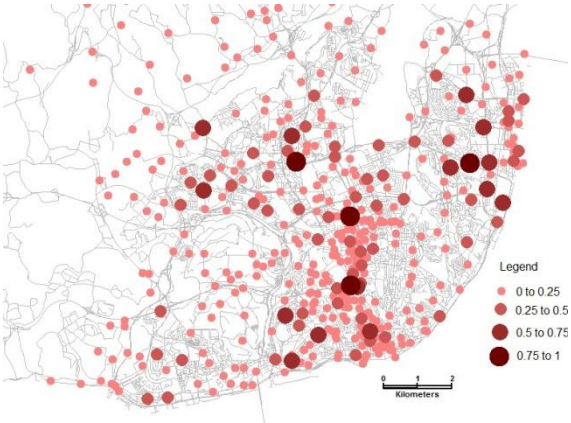
Handover: 7PM



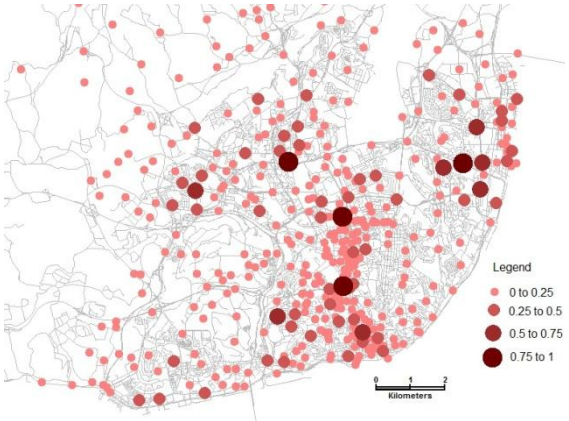
Handover: 8PM



Handover: 9PM



Handover: 10PM



Handover: 11PM