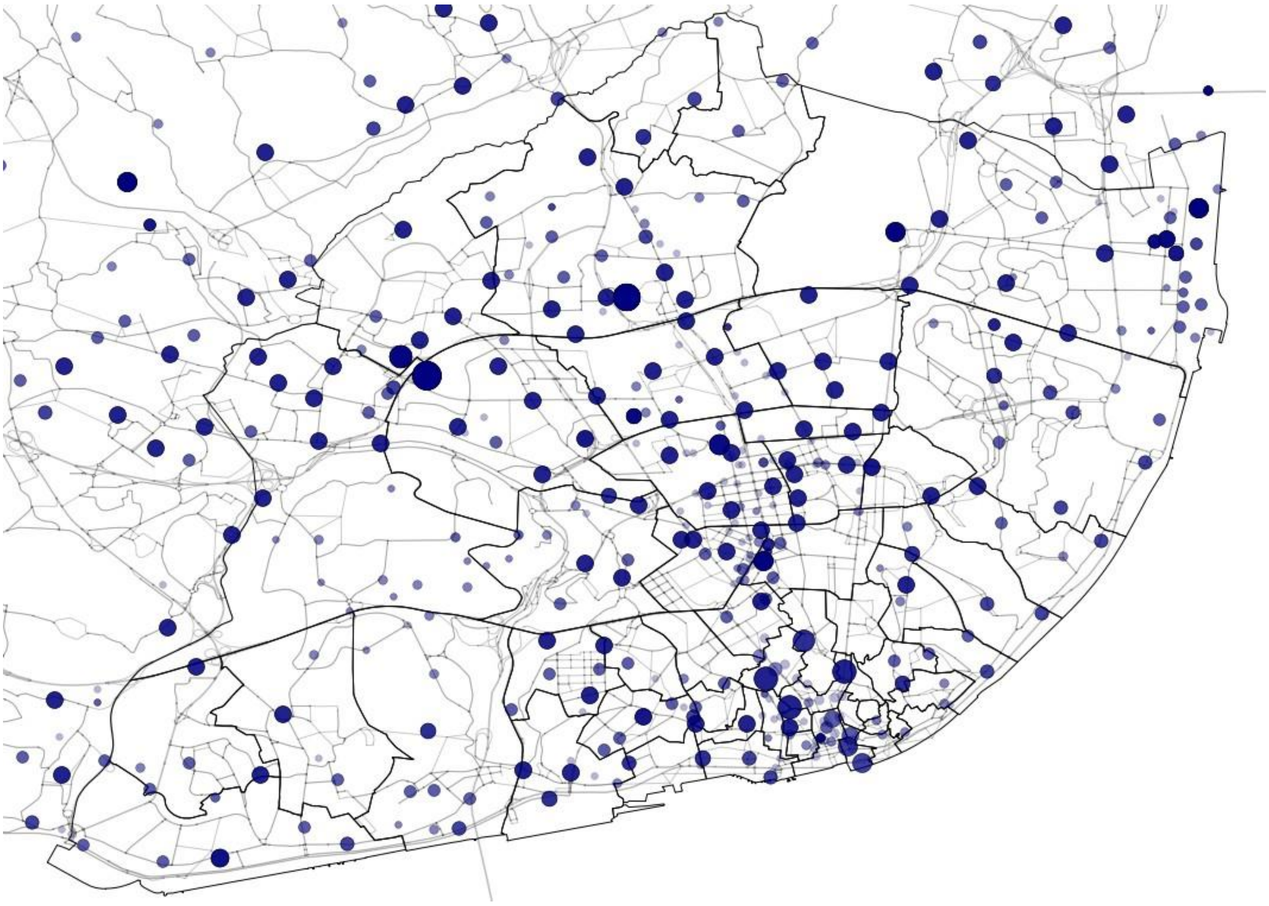


Marco António Morais Veloso

ANALYSIS OF TAXI DATA FOR UNDERSTANDING URBAN DYNAMICS

Doctor of Philosophy Thesis in Sciences and Information Technologies, supervised by Professor Carlos Lisboa Bento and Professor Santi Phithakkitnukoon, and submitted to the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra

September 2016

UNIVERSIDADE DE COIMBRA

· U C ·

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Marco António Morais Veloso

ANALYSIS OF TAXI DATA FOR UNDERSTANDING URBAN DYNAMICS

PhD Thesis submitted to fulfill the requirements of the Doctoral program in Sciences and Information Technologies

Department of Informatics Engineering, Faculty of Sciences and Technology, University of Coimbra

Coimbra, September 2016

Supervisors

Professor Carlos Lisboa Bento

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

Professor Santi Phithakkitnukoon

Associate Professor
Department of Computer Engineering
Chiang Mai University, Thailand

Acknowledgments

This study is the result of a long and difficult journey spanning several years, and was not possible without the valuable contribution of various people and institutions, which I would like to thank.

Firstly, I would like to express my gratitude to my supervisors: Prof. Carlos Bento and Prof. Santi Phithakkitnukoon. Their guidance, expertise, and motivation were fundamental throughout the entire project. They have spare considerable amounts of their personal time to discuss and review my work. They provided solutions and guidance when they were most needed, from setting the initial goals, all the way to the final validation. And above all, they have believed in me and my work, and have been at my side throughout this entire academic path.

No contribution or finding produced from this study was possible without the richness of the data set and the knowledge it possesses. Therefore, data providers are a key player in this study – as they are on every research. Besides collecting and providing data, data providers were available to share their knowledge and assist during the initial interpretation of the data sets. I would like to thank all the data providers that made this study possible: Geotaxi, TMN (currently rebranded as MEO), APA (*Agência Portuguesa do Ambiente*), CCDR-LVT (*Comissão de Coordenação e Desenvolvimento Regional de Lisboa e Vale do Tejo*), INE (Statistics Portugal), Lisbon Municipality, Sapo Maps, and Weather Undergroud.

The Doctoral Program was conducted concurrently with a full-time job. My sponsoring institution's comprehension and flexibility was essential, allowing me the space to conciliate my research with my professional obligations. Thus, I would like to thank the management and my colleagues at the College of Management and Technology of Oliveira do Hospital and the Polytechnic Institute of Coimbra.

In research, the support of a group to discuss and share our approaches and results is essential. They help us validate our work, alert us to any obstacles unforeseen by us, and sometimes identify new paths or courses of action. Also relevant is the companionship that emerges from this community, composed by empathetic fellows sharing similar experiences. A double thank you to all my lab colleagues at the Ambient Intelligence Laboratory (AmILab) of Centre for Informatics and Systems of the University of Coimbra (CISUC) for all the experiences we have shared and for their time.

Friendship is a major pillar in one’s life. Those who know us best are able to understand us without a word being said and help us without any help being asked. We share life experiences, good and bad, helping each other in difficult times. Thank you to all my friends, to whom I owe a debt of gratitude. But to a few of them, who rose above all others, I give special thanks: Nuno Gil, Ricardo, and Carla. They have been an important support system during this academic step. Our frequent meetings and encounters were both cathartic and a source of motivation. A special thanks to you three.

Finally and most importantly, family. Just like true friends, they are fundamental in one’s life - an everlasting source of motivation. They are with us before we have started a new path, and they will stick with us long after it is concluded, regardless of the results. Most of all, they support us during difficult times and share our joys during good times. A very special thanks to my mother, father, brother Luis, sister Ana, and my life partner Sofia.

Thank you everyone. This study is just as much the result of my work as it is of yours.

Abstract

The growth of urban areas poses both challenges and opportunities. Challenges due to the increase in demand for resources and services needed. However, it also allows the opportunity for the development of new services and, collectively, urban areas can produce data to help better understand urban mobility.

The taxi can be perceived as a probe for traffic conditions. Additionally, its flexibility and ubiquity can be used to retrieve large data sets of information, essential for studying urban mobility. In this study we explore a data set of taxi-GPS traces, collected in Lisbon, Portugal, to understand to what extent can taxi data represent urban mobility. More specifically, in this study we aimed to answer three research questions: (A) *Is it possible to develop a model to estimate the taxi demand throughout the city?* (B) *Are urban data sources correlated among them? More specifically, is taxi activity correlated with mobile phone activity, two of the major urban data sources?* (C) *Can taxi data be used as a probe to infer the concentrations of exhaust gases in urban areas?* To aid the analysis, additional data sets were collected for the same spatiotemporal period, regarding mobile phone activity, information on atmospheric pollutants and meteorological conditions.

In order to develop a model to estimate taxi demand, an exploratory analysis was performed. The study was able to visualize the spatiotemporal variation, identifying the main pick-up and drop-off locations and busy hours, and observe that trip distance and duration follow Gamma and Exponential distributions. The study was also able to identify the link between pick-up and drop-off locations, observing strong links between public transportation hubs. Additionally, an analysis of taxi driver behavior during downtime was performed. The analysis of taxi-GPS from top drivers have shown specific strategies used to maximize their profit. Either by waiting for passengers in locations related with main public transportation hubs, during specific hours of the day, or by avoiding traveling great distances to the next pick-up location. The inference analysis explored the possibility of estimating the next pick-up area given the current location (last drop-off), day of the week, hour, weather conditions and area type (characterized by points of interest). The inference engine is based on a naïve Bayesian classifier, achieving 56.3% of accuracy of the training sample. Current location turned out to be the main contributor to the algorithm, contrary to weather conditions which is the variable with the least weight in the calculation.

The investigation of the relationship between taxi and mobile phone activity started by performing an exploratory analysis of the mobile phone call intensity. The study showed a fairly regular pattern, consistent throughout the day and during the entire time series. During data analysis, a significant correlation between the taxi volume and mobile phone call intensity was found, with a coefficient of determination of 0.8047. The strongest correlation was achieved over active hours of the day (8 AM-10 PM) and active days of the week (weekdays), in areas with medium and high taxi activity. Moreover, mobile phone call intensity had a significant correlation with taxi volume of the previous two hours. Furthermore, we found that this inter-predictability could be modeled with a linear function and varied across different times of the day.

To model and estimate the concentration of exhaust gases, taxi activity and meteorological conditions (temperature, wind, humidity, and weather conditions) were considered. The study revealed the daily and seasonal patterns of exhaust gases, how they are correlated with the weather conditions, and how nitrogen dioxide - a marker for atmospheric pollution - is strongly correlated with other exhaust gases. Using a multilayer perceptron, with 15 hidden layers and a sigmoid activation function, we were able to estimate the nitrogen dioxide concentrations, with a coefficient of correlation of 0.7869, showing a relationship between the exhaust gas concentration and other urban variables, especially on traffic stations. The multicollinearity analysis was applied to ensure non-correlated predictor variables and avoid overfitting of the model.

This study contributes to a better comprehension of the complex interactions between the diversity of urban data sources. Our findings, to some extent, unveil the relationships between different urban data sources, especially the role of taxi service as a predictor variable for other urban variables.

Resumo

O crescimento das áreas urbanas apresenta tanto desafios como oportunidades. Desafios devido à crescente exigência de recursos e serviços necessários. No entanto, também permite oportunidades para o desenvolvimento de novos serviços e, colectivamente, as áreas urbanas podem produzir dados para auxiliar a melhor compreender a mobilidade urbana.

Táxi pode ser compreendido como uma sonda ou sensor para as condições de tráfico. Adicionalmente, a sua flexibilidade e ubiquidade podem ser usados para recolher largas quantidades de dados, essenciais para o estudo da mobilidade urbana. Neste estudo exploramos um conjunto de dados composto por trajectórias GPS de táxis, recolhidos em Lisboa, Portugal, para compreender até que ponto os dados de táxi podem representar a mobilidade urbana. Mais especificamente, neste estudo pretendemos responder a três questões de investigação: (A) *É possível desenvolver um modelo para estimar a solicitação de táxis numa cidade*? (B) *Estarão as fontes de dados correlacionadas entre si? Mais especificamente, estará a actividade dos táxis correlacionada com a actividade da rede móvel, duas das principais fontes de dados urbanos*? (C) *Os dados de táxi podem ser usados como sensor para inferir as concentrações de gases tóxicos em áreas urbanas*? Para auxiliar a análise, bases de dados adicionais foram recolhidas para o mesmo espaço físico e período temporal, correspondendo à densidade de chamadas da rede móvel, informação sobre poluentes atmosféricos e condições meteorológicas.

Para permitir o desenvolvimento de um modelo de estimação da solicitação de táxis, foi realizada uma análise exploratória. O estudo foi capaz de visualizar a variação espacial e temporal, identificar as principais localizações para entrada e saída de passageiros, bem como as horas de maior afluência e observar que a distância e duração das viagens seguiam as distribuições Gamma e exponencial. O estudo também foi capaz de identificar a ligação entre as localizações de entrada e saída de passageiros, observando fortes ligações entre centros de transportes públicos. Adicionalmente, uma análise aos comportamentos dos taxistas durante o período de procura de novos passageiros foi realizada. A análise de trajectos GPS dos condutores mais eficientes demonstraram estratégias específicas para maximizar o ganho. Tanto ao aguardar passageiros in localizações relacionadas com os principais centros de transporte públicos em horas específicas do dia, como ao evitarem viajar longas distancias para a próxima localização de embarque de um passageiro. A análise inferencial explorou a possibilidade de estimar a próxima área de embarque de passageiros, a partir da localização actual (a localização da última saída de passageiros), o dia da semana, as condições climatéricas e o tipo de área (definido por pontos de interesse). O motor de inferência é baseado num classificador simples Bayesian, conseguindo obter 56,3% de acuidade a partir das amostras de treino. A Localização actual revelou ser a principal variável que contribui para o algoritmo, contrariamente às condições climatéricas, que se mostraram ser a variável com menos peso no cálculo.

A investigação da relação entre actividades de táxi e da rede móvel começou por realizar uma análise exploratória da densidade das chamadas na rede móvel. O estudo mostrou um padrão razoavelmente regular, consistente ao longo do dia e durante toda a série temporal. Durante a análise de dados, foi identificada uma correlação significante entre a actividade de táxis e a densidade das chamadas na rede móvel, com um coeficiente de determinação de 0,8047. A relação mais forte foi obtida durante horas de expediente (8h-22h), em dias de semana, em áreas de média e elevada actividade do serviço táxi. Além disso, a densidade de chamadas da rede móvel apresenta uma significante correlação com a actividade dos táxis das últimas duas horas. Acima disso, verificámos que essa previsibilidade entre ambas as variáveis pode ser modelada com uma função linear, e varia ao longo das horas do dia.

Para modelar e estimar as concentrações de gases tóxicos, foi considerado a actividade de táxis e as condições meteorológicas (temperatura, vendo humidade e estado do tempo). O estudo revelou os padrões diários e sazonais dos gases tóxicos, como estes estão correlacionados com o estado do tempo e como o dióxido de azoto – um marcador para a poluição atmosférica – está fortemente relacionado com os restantes gases tóxicos. Usando um perceptrão multi-camada, com 15 camadas escondidas e uma função de activação sigmóide, fomos capazes de estimar as concentrações de dióxido de azoto com um coeficiente de correlação de 0,7869, demonstrando a relação entre as concentrações de gases tóxicos com outras variáveis urbanas, especialmente em estações de monitorização de tráfico. A análise de multicolinearidade foi aplicada para garantir variáveis preditoras não correlacionadas entre si e evitar sobre-ajuste do modelo.

Este estudo contribui para uma melhor compreensão das interacções complexas entre as diversas fontes de dados urbanos. As nossas observações, até certo ponto, revelam as relações entre diferentes fontes de dados, especialmente o papel do serviço de táxi como variável preditora para outras variáveis urbanas.

Keywords

Artificial neural network, exhaust gases’ concentrations, intelligent transport systems, linear regression, mobile phone activity, naïve Bayesian classifier, predictive analysis, spatiotemporal analysis, taxi-GPS traces, time series analysis, urban mobility.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ANOVA	Analysis of Variance
ANN	Artificial Neural Network
CO	Carbon Monoxide
ED	Euclidean Distance
GPS	Global Positioning System
GSM	Global System for Mobile Communications
IG	Information Gain
ITS	Intelligent Transport Systems
LBSN	Location-Based Social Networks
MAE	Mean Absolute Error
NBC	Naïve Bayesian Classifier
NO	Nitrogen Monoxide
NO_2	Nitrogen Dioxide
NO_x	Nitrogen Oxides
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
POI	Points of Interest
r	Coefficient of Correlation
r^2	Coefficient of Determination
RMSD	Root Mean Squared Deviation
RMSE	Root Mean Squared Error
VANET	Vehicular Ad-Hoc Network
VNS	Vehicular Sensor Networks

Chapter 1
Introduction

1.1 Motivations and objectives

According to the United Nations in 2008, for the first time in history, half of the world's population was living in urban areas (United Nations, 2008). In 1950 there were only two metropolises with at least 10 million inhabitants. In 1975 only three metropolises broke that barrier. Today there are 21 megacities with more than 10 million inhabitants, and in 2025, the United Nations estimates that there will be 27 cities (United Nations, 2012). This is clear evidence of the fast growth of urbanization in terms of population and size.

The demand for better services (e.g. public transportation, energy, communications) and urban planning (e.g. infrastructures, environments, policies) increases with the rapid growth of urban areas. In order to maintain a constant flow of people and vehicles, we need to reduce the use of individual means of transport (e.g. car) and stimulate the use of public transportation (e.g. bus, metro, train). Traffic is one of the major sources of toxic compounds present in combustion gases that negatively impact the health of urban inhabitants (EEA, 2011), (Borrego, et al., 2000), (Zavala, et al., 2006), (Ndoke & Jimoh, 2005), (Becker, et al., 2000). There is a need to address this issue today while low-carbon transport systems are still being developed. However, we need to improve the public transportation system in order to meet citizens' needs.

A more efficient public transportation system can lead to a reduction in traffic congestions and consequent reduction of energy consumption and pollution. Nevertheless, to optimize the public transportation network it is essential to understand what drives the common citizen and what their needs are. We need a better understanding of city dynamics. Gathering data from the traditional public transportation (e.g. bus, train, metro) can provide us with a relevant database and information on general passengers' movement. However, it does not provide the exact origin and destination for each passenger, since these transportation modes rely on pre-designated stops and paths. The taxi can be a way to retrieve a large data set of information with higher precision when we focus on the origin and destination of each trip. Taxis can pick-up the passengers right where they are standing, and drop them off precisely at their desirable destination, without being bound to a pre-determined path. The process of data collecting is transparent and non-intrusive to the passenger. Additionally, taxis can be used as a probe for traffic conditions (Castro, et al., 2012), (Yuan, et al., 2011a), (Gühnemann, et al., 2004), (Liu, et al., 2009a).

At the same time, we are experiencing new developments in pervasive and ubiquitous computing technologies, such as a global system for mobile communications (GSM) and a global positioning system (GPS), which provide useful tools for sensing social and traffic activities in cities. Nowadays we are able to access a wider variety of devices, with a growing number of features and computational capabilities. This technological diversity provides us with the tools to sense urban spaces. It allows us to either take a collective snapshot of all urban activity or simply follow the pattern of a single vehicle or individual. Analyzing GPS-enabled vehicle traces and mobile phone activity thus provides, to some extent, an overview of how the city functions.

Taxis are currently equipped with GPS devices for better monitoring and dispatching. Their traces have been used to study various aspects of the traffic network as they provide fine-grained data that reflects the state of traffic flow in a city. These traces typically carry occupancy information on pick-up and drop-off location. The ubiquity of taxis has attracted considerable attention for a while, in order to extract information and develop prediction systems, which led to a significant amount of research work being performed around the exploration of taxi-GPS traces.

Facing the challenges of growing cities and by taking the opportunistic sensing approach, a main question is posed: *to what extent can GPS traces of taxis be used to infer the city's dynamics, namely the inhabitants' patterns*? Furthermore, *what is the role of taxis in the complex relations amid the diversity of urban data sources*? Although previous research on this topic led to important findings, there are still challenges yet to explore that we aim to analyze in this thesis. Our work deepens the spatiotemporal analysis and the study of predictability of taxi trips by using complementary data (e.g. Points Of Interest, weather conditions); explores the underlining relationship between taxi volume and mobile phone activity, two important urban data sources; and further extends the study of the relation among urban data sources by examining the relationship between taxi mobility patterns, weather conditions and the level of concentration of exhaust gases, to estimate the concentrations of gas in urban areas.

1.2 Research questions and contributions

Taxi data has been used by various researchers. Due to its ubiquity and ability to reach any corner of urban areas, without being bound to predefined schedules or specific paths, taxi-GPS traces have been widely used as a mean to retrieve a snapshot of the city and to develop better solutions and services in urban areas. Several works use taxi-GPS traces to uncover underlying patterns and can be organized according to their goals:

- estimating optimal driving path (Ziebart, et al., 2008), (Yuan, et al., 2010), and (Zheng, et al., 2010);
- predicting next taxi pick-up location (Chang, et al., 2010), (Ge, et al., 2010), (Liu, et al., 2010b), (Yuan, et al., 2011b), and (Moreira-Matias, et al., 2012a);
- improving dispatching services and detecting anomalies and fraud (Ge, et al., 2011), and (Ivan & Popa, 2015);
- modeling driving strategies to improve taxis' profit (Ge, et al., 2010), (Liu, et al., 2010a), and (Moreira-Matias, et al., 2014a);
- identifying flaws and possible improvements in urban planning (Zheng, et al., 2011b), and (Chen, et al., 2013a);
- developing models for urban mobility, social functions, and dynamics between the different areas in the city (Qi, et al., 2011), and (Castro, et al., 2013).

These publications have intensively explored taxi data sets, proposing various results and solutions. The need to improve driving paths and to predict the location of the next passenger has attracted much of the attention. However, there is room for improvement and challenges to be explored. Considering the work published up to now by the research community, and the challenges of growing cities, we posed three research questions, explored in this thesis:

- *Is it possible to develop a model to foresee taxi demand throughout the city*?

- *Are urban data sources correlated among them? More specifically, is taxi density correlated with mobile phone activity, two of the major urban data sources*?

- *Can taxi data be used as a probe to infer the concentrations of exhaust gases in urban areas*?

As a result, of these questions, three main contributions emerged:

A. Development of a model to estimate taxi demand.

In order to efficiently manage and dispatch taxis, it is relevant to understand and anticipate their demand. As previously discussed, this has been a widely explored topic. Nevertheless, our work differentiates from the remaining studies in two key aspects: (1) while other authors propose solutions to estimate taxi demand in specific locations, we propose an approach that takes into consideration a comprehensive urban area, modeling the city by a grid and estimating the likelihood of passengers on every cell of the grid; and (2) our inference approach is based on a Naïve Bayesian Classifier, considering as input variable a geographic characterization of each cell (in the form of Points Of Interest). Additionally, we explore the spatiotemporal distribution of taxi volume; analyze how taxis connect different areas of the city; and investigate taxi drivers' behavior between trips.

B. Investigation of the relationship between taxi and mobile phone activity.

Up until now, research on urban mobility and intelligent transportation systems focused on the study of a single variable or a set of related variables (e.g. taxi related variables, such as occupancy, average speed or amount of pick-ups and drop-offs). Our work expands that research by investigating the interplay between two different urban data sources. While mobile phone call data has been used to study social aspect of the city, taxi-GPS traces have been explored to understand the state of traffic flow in urban areas. Our approach correlates both urban data sources to understand the interplay between them. In order to do that, the data sets are transformed into time series. Using a regression analysis on time series along with a shifting temporal window, we are able to attain the best correlation with a significant coefficient of determination.

C. Exploration of models that use taxi volume information in order to estimate environmental variable.

Environmental awareness has risen with the growth of cities and with it, the need for sustainability with an emphasis on air pollution. To keep track of air quality, monitoring stations have been built and deployed to measure the concentration of toxic gases and particles in the atmosphere. However, these stations are expensive, demand maintenance, and are scarce throughout the city. Most urban areas do not have any kind of continuous measurements of concentrations of toxic gases. Solutions have been proposed to encompass most urban areas with sensors to control air quality. However, they demand a dedicated infra-structure, with the costs associated with developing and maintaining it. Our work explores the ubiquity of taxis, where vehicles are continually moving throughout the city. By combining the information from meteorological conditions and taxi activity, we propose a model to estimate the concentrations of nitrogen dioxide, from which we are able to infer the concentration of other three pollutant gases (nitrogen monoxide, nitrogen oxides and carbon monoxide, often termed as *exhaust gases*, since they are the byproduct of combustion). This initial approach eliminates the need for a dedicated infrastructure, thus, reducing the cost of developing, deploying and maintaining a physical framework, while improving spatial resolution. Moreover, it allows us to further explore the relation between urban data sources.

1.3 Approach

Once the research questions were defined, several steps were taken to verify the validity of our hypothesis. Classic knowledge extraction from databases was followed, as suggested and described by (Witten & Frank, 2005), (Linoff & Berry, 2011), (Santos & Azevedo, 2005), and (Gama, et al., 2012).

The data was collected and provided from different sources:

- Taxi data (described on Chapter 2), provided by Geotaxi[1];
- Mobile phone data (described on Chapter 3), provided by TMN (currently rebranded as MEO)[2];

[1] GeoTaxi. http://www.geotaxi.com

- Data regarding air quality and gases' concentrations (described on Chapter 4), provided by '*Comissão de Coordenação e Desenvolvimento Regional de Lisboa e Vale do Tejo*' (CCDR-LVT)[3], and the '*Agência Portuguesa do Ambiente*' (APA)[4];

- Meteorological data (described in Chapter 4), including weather conditions, was retrieved from Weather Underground[5];

- Points Of Interest (described in Chapter 2), provided by Sapo[6];

- Shape files, GIS data and Census information, provided by '*Instituto Nacional de Estatística*' (Statistics Portugal)[7].

All data was collected considering the same time window. Meanwhile, data sets present distinct formats, and in some cases, different sampling rates. Additionally, we detected faulty, erroneous or missing data. Therefore, cleaning, alignment and transformation steps were needed. Furthermore, every sample was geo-referenced. In order to better handle the size of the metropolitan area, the city under study was modeled with an initial grid of 500x500m cells as suggested by (Huang, et al., 2010) and (Liu, et al., 2010a). Finally, data was stored in a relational database.

For each data set, an exploratory study was performed to understand the spatiotemporal distribution. This step allowed us to identify hotspots of activity or potential patterns to be further investigated during data analysis.

During the data analysis, different approaches were followed, according to the data sets and the goals of each research question. For the first research question, the goal was to estimate the most likely location to pick-up the next passenger, considering taxis' historical data and a set of independent variables. A probabilistic approach was considered in the form of a Naïve Bayesian Classifier due to its simplicity, how fast it is to build and train, and how adequate it is for problems where data is linearly separable, though assuming independence between (Zhang, 2004), and

[2] MEO. http://www.meo.pt

[3] CCDR-LVT. http://www.ccdr-lvt.pt/pt/

[4] Agência Portuguesa do Ambiente. http://www.qualar.org

[5] Weather Underground http://www.wunderground.com/

[6] Sapo Mapas. http://mapas.sapo.pt/

[7] Instituto Nacional de Estatística. http://www.ine.pt/

(Puntumapon & Pattara-atikom, 2008). Besides current location and time, characterization of the city's functions was also incorporated (Qi, et al., 2011), alongside with weather conditions (Yuan, et al., 2011a).

For the second research question, we aimed to identify a hypothetical correlation between two urban data sources (mobile phone activity and taxi volume). To explore both time series, a regression analysis was considered. To estimate the coefficient of the model, the method of least squares was applied (Pallant, 2005). Since data was normally distributed, a coefficient of correlation of Pearson (r) was used to analyze the linear association between the time series, and a coefficient of determination (r^2) was applied to attain the goodness-of-fit, representing the percentage of the response variable variation that is explained by a linear model (Kennedy, 2008). To verify the existence of a linear relation between the dependent variable and the predicted variable, the F test of ANOVA (Analysis Of Variance) was adopted, which is a measure of significance for the regression (Maroco, 2005).

Finally, the third research question seems to be more challenging, due to the nature of the data sets and the goal: estimate the concentration of exhaust gases, based on meteorological conditions and taxi activity. To investigate the relation between the variables, a multiple linear regression analysis, using the method of least squares to estimate the coefficients, was proposed (Pallant, 2005), (Donnelly, et al., 2015). As a measure of linear association between the variables, the coefficient of Pearson (r) was selected, while the coefficient of determination (r^2) was adopted to explain the percentage of variation of the dependent variable determined by the independent variables (Kennedy, 2008). The significance of the regression was tested using the F test of ANOVA, which verifies the existence of a linear relation between the dependent variable and the explanatory or predictor variables (Maroco, 2005). The Stepwise method was adopted to select the variables to be included in the model. Finally, the multicollinearity was verified using the VIF (Variance Inflation Factor) (Pallant, 2005).

However, due to the apparent complexity of the problem, combined with no prior information regarding how the variables could correlate, and uncertainty about whether linear relations were an adequate fit to the problem, a multilayer perceptron with backpropagation was additionally explored for comparison with the linear regression approach (Shi & Harrison, 1997), (Gardner & Dorling, 1999), (Kolehmainen, et al., 2001), (Perez & Reyes, 2001), (Kukkonen, et al., 2003), (Agirre-Basurko, et al., 2006), (Juhos, et al., 2008), and (Ahmed, et al., 2010). The amount of possible variables immediately poses a question: which variables could explain the variation of the target (or dependent variable), and to what degree? A factorial analysis Principal Component Analysis was considered (Pallant, 2005), as an alternative approach before applying the linear regression analysis, a method which presents no *a priori* requirements, using the Kaiser's rule for the eigenvalues (Larsen & Warne, 2010) and the scree plot approach to retain the most significant factors.

The models were compared with ground truth information to verify the significance of the results. Ground truth includes information about taxi volume, mobile phone activity, concentrations of toxic gases on the atmosphere, weather conditions, Points Of Interest, and census information. Additionally, technical reports were produced for each step of the process, to log the evolution of the work and compile experiments and results. Finally, main findings and achievements were submitted for peer-review at international conferences.

1.4 Organization of the thesis

The thesis is organized in five chapters, and describes the work around the analysis of taxi traces during the doctoral program. During that process, several publications were submitted to share the findings with the scientific community and to receive feedback and validation. The content of Chapter 2, Chapter 3 and Chapter 4 correspond to publications in peer-reviewed international conferences and journals. The content was not significantly changed, however, complementary information was added to better illustrate the analysis or results. These chapters are self-contained, allowing the user to read them without prior knowledge of the remaining chapters. In order to do so, some information is duplicated among these chapters. In each chapter the reader can find a description of the data set, state of the art, exploratory analysis an inferential analysis, and a chapter summary and conclusions. The chapters represent the sequence of steps taken during the study, and therefore, they are interrelated.

Chapter 1 introduces the motivation of the work. Describes the problem (fast growth of urban areas demanding improved transportation services) and opportunities (the advent of new technologies such as GPS and GSM devices) that led to the final goal (use opportunistic taxi data to infer the city's dynamics). It discusses the three main research questions which the work aims to deal with and the followed approach.

Chapter 2 focuses on the analysis of taxi data, uncovering how taxi-GPS traces can describe urban areas. It describes the main data set (taxi data, along with the city under study) and the state of the art on the analysis of taxi-GPS traces. The exploratory analysis encompasses a spatiotemporal analysis; a description how predominant locations are connected (termed *gravity map*); a description of how taxi-GPS traces are distributed (considering duration, distance and income); a study of driver strategies to find the next passenger and the behavior between services (termed *downtime*); and an analysis of how areas characterized by Points Of Interest (POI) affect the taxi service. The inferential analysis explores to what extend it is possible to infer the location for the next passenger. This analysis takes in consideration the effect of different set-ups, namely the temporal periods and cell size.

Chapter 3 explores the relation between taxi volume (as a representation of the traffic flow) and mobile phone activity (as a representation of the social aspect of the city), suggesting that distinct urban data sources can be correlated with each other. It describes a new data set (GSM data) and the state of the art on the analysis of mobile phone activity. The exploratory analysis describes the spatiotemporal distribution of mobile phone call intensity. The inferential analysis correlates taxi volume and mobile phone activity to extract the best fit using time series analysis.

Chapter 4 advances the study of taxi data in order to infer other variables relevant in urban areas. This chapter explores how taxi data and weather conditions can be used to estimate the concentration of exhaust gases. Similar to the previous chapters, it describes the data sets (exhaust gases and weather conditions) and the state of art on the study of exhaust gases and their pending relation with traffic conditions. The exploratory analysis describes the spatiotemporal distribution of different exhaust gases, how they correlate with each other and the effect of weather conditions on the dispersion of exhaust gases. The inferential analysis studies different algorithms and setups to estimate the concentration of exhaust gases from taxi data and weather conditions.

In Chapter 5 the reader will find the conclusions of this work. A summary of the work is presented, underlining the main contributions, alongside with the limitations. Finally it discusses future work.

1.5 List of publications

While working on the doctoral program, several publications were published in peer-review conferences and in journals, as leading author or co-author. The set of publications can the organized around three main topics: taxi traces analysis, bus ridership and data fusion for intelligent transportation systems. For each topic, the publications are arranged chronologically as conference papers, journals papers and technical reports.

The publications about taxi traces analysis form the core basis for the thesis. Chapter 2, Chapter 3 and Chapter 4 follow the content of these publications. The publications on bus ridership and data fusion are peripheral to the scope of the thesis. All conference and journal publications were subject to peer-review by international committees.

1.5.1 Taxi traces analysis

Chapter 2

The following publications correspond to Chapter 2 of the thesis.

Conference publications

Publications in international peer-reviewed Special Interest Groups (SIG) conferences.

- "***Taxi-Aware Map: Identifying and predicting vacant taxis in the city***". Santi Phithakkitnukoon, Marco Veloso, Carlos Bento, Assaf Biderman and Carlo Ratti. International Joint Conference on Ambient Intelligence, AmI 2010, Malaga, Spain, 2010. (Phithakkitnukoon, et al., 2010b).
- "***Exploratory Study of Urban Flow using Taxi Traces***". Marco Veloso, Santi Phithakkitnukoon and Carlos Bento, Patrick Olivier, Nuno Fonseca. First International Workshop on Pervasive Urban Applications (PURBA) in conjunction with the Ninth International Conference on Pervasive Computing, San Francisco, California, USA, 2011. (Veloso, et al., 2011a).
- "***Urban Mobility Study using Taxi Traces***". Marco Veloso, Santi Phithakkitnukoon and Carlos Bento. International Workshop on Trajectory Data Mining and Analysis (TDM) in conjunction with the 13th International

Conference on Ubiquitous Computing (UbiComp), ACM Digital Library and UbiComp Extended Proceedings, Beijing, China, 2011. (Veloso, et al., 2011c).

- "***Sensing Urban Mobility with Taxi Flow***". Marco Veloso, Santi Phithakkitnukoon and Carlos Bento. International Workshop on Location-Based Social Networks (LBSN) in conjunction with the 19th International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL), ACM Digital Library, Chicago, Illinois, USA, 2011. (Veloso, et al., 2011b).

Journal publications

Publications in international peer-reviewed journals.

- (Submitted on January 2016) "***Towards Recommendation System for Taxi Drivers***". Marco Veloso, Santi Phithakkitnukoon and Carlos Bento. Journal of Urban Technology, Taylor & Francis, 2016. Published work corresponding to Chapter 2 of the thesis.

Chapter 3

The following publication corresponds to Chapter 3 of the thesis.

Conference publications

- "***Exploring the Relationship between Mobile Phone Call Intensity and Taxi Volume in Urban Area***". Marco Veloso, Santi Phithakkitnukoon and Carlos Bento. 15th IEEE Intelligent Transportation Systems Conference (ITSC), Anchorage, Alaska, USA, 2012. (Veloso, et al., 2012).

Chapter 4

The following publications correspond to Chapter 4 of the thesis.

Conference publications

- "***Exploring Relationship Between Taxi Volume and Flue Gases' Concentrations***". Marco Veloso, Santi Phithakkitnukoon and Carlos Bento. Third International Workshop on Pervasive Urban Applications (PURBA) in conjunction with ACM International Joint Conference on Pervasive and Ubiquitous Computing, Zurich, Switzerland, 2013. (Veloso, et al., 2013).

- (Submitted on February 2016) “***Using Taxi and Meteorological Conditions as a Probe to Monitor Exhaust Gas***”. Marco Veloso, Santi Phithakkitnukoon and Carlos Bento. International Workshop on Intelligent Public Transports (IPT) in conjunction with the IEEE Intelligent Transportation Systems Conference (ITSC), 2016.

Journal publications

- (Submitted on February 2016) “***Monitoring Urban Exhaust Gas Concentration Using Taxi Location and Meteorological Information***” Marco Veloso, Santi Phithakkitnukoon and Carlos Bento. Journal of Urban Technology, Taylor & Francis, 2016. Published work corresponding to Chapter 4 of the thesis.

Technical Reports

To support the experimental work, several documents were produced, describing the procedures followed and results attained, termed *technical reports.* These documents were used in group meetings as a basis for discussion, and are the framework for the published papers.

- “***Exploratory study of taxi trajectories and taxi driver behavior***”. Marco Veloso. Ambient Intelligence Laboratory, Centre for Informatics and Systems of the University of Coimbra, May 2011. (Chapter 2)

- “***Study of average taxi speed on urban areas***”. Marco Veloso. Ambient Intelligence Laboratory, Centre for Informatics and Systems of the University of Coimbra, July 2011. (Chapter 2)

- “***Impact of area type characterized by POIs to the taxi service***”. Marco Veloso. Ambient Intelligence Laboratory, Centre for Informatics and Systems of the University of Coimbra, February 2012. (Chapter 2)

- “***Taxi Driver Assistant Framework***”. Marco Veloso. Ambient Intelligence Laboratory, Centre for Informatics and Systems of the University of Coimbra, October 2012. (Chapter 2)

- "***Taxi Driver Assistant – A Proposal for a Recommendation System***". Marco Veloso. Ambient Intelligence Laboratory, Centre for Informatics and Systems of the University of Coimbra, January 2013. (Chapter 2)

- "***Study of taxi demand and the use of GSM network***". Marco Veloso. Ambient Intelligence Laboratory, Centre for Informatics and Systems of the University of Coimbra, January 2013. (Chapter 3)

- "***Taxi as a Probe to Monitor Environmental Changes in Urban Areas***". Marco Veloso. Ambient Intelligence Laboratory, Centre for Informatics and Systems of the University of Coimbra, September 2014. (Chapter 4)

1.5.2 Bus ridership

Conference publications

Publications in international peer-reviewed Special Interest Groups (SIG) conferences:

- "***Predicting Bus Ridership***". Sourav Bhattacharya, Santi Phithakkitnukoon, Petteri Nurm, Arto Klam, Marco Veloso and Carlos Bento. Third International Workshop on Pervasive Urban Applications (PURBA) in conjunction with ACM International Joint Conference on Pervasive and Ubiquitous Computing, Zurich, Switzerland, 2013.

- "***Mining Temporal Patterns of Transport Behaviour for Predicting Future Transport Usage***". Stefan Foell, Gerd Kortuem, Reza Rawassizade, Santi Phithakkitnukoon, Marco Veloso and Carlos Bento. Third International Workshop on Pervasive Urban Applications (PURBA) in conjunction with ACM International Joint Conference on Pervasive and Ubiquitous Computing, Zurich, Switzerland, 2013.

- "***Catch Me If You Can: Predicting Mobility Patterns of Public Transport Users***". Stefan Foell, Santi Phithakkitnukoon, Gerd Kortuem, Marco Veloso and Carlos Bento. 17th IEEE Intelligent Transportation Systems Conference, 2014.

- "***A Tool for Exploratory Visualization of Bus Mobility and Ridership: A case study of Lisbon, Portugal***". Chalermpong Somdulyawat, Piyawat Pongjitpak, Santi Phithakkitnukoon, Marco Veloso, Carlos Bento. Fourth International

Workshop on Pervasive Urban Applications (PURBA) in conjunction with ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp ’15, Osaka, Japan, 2015.

Journal publications

Publications in international peer-reviewed journal:

- “***Predictability of Public Transport Usage: A Study of Bus Rides in Lisbon, Portugal***”. Stefan Foell, Gerd Kortuem, Reza Rawassizadeh, Santi Phithakkitnukoon, Marco Veloso and Carlos Bento. IEEE Transactions on Intelligent Transportation Systems, 2015.

- (Submitted on February 2016) “***Regularity of Public Transport Usage: A case study of bus rides in Lisbon, Portugal***”. Stefan Foell, Santi Phithakkitnukoon, Marco Veloso Gerd Kortuem and Carlos Bento. Journal of Urban Technology, Taylor & Francis, 2016.

1.5.3 Data fusion for intelligent transportation systems

Conference publications

- “***Data Fusion for Travel Demand Management: State of the Practice & Prospects***”. Christopher Zegras, Francisco Pereira, Andrew Amey, Marco Veloso, Liang Liu, Carlos Bento and Assaf Biderman. 4th International Symposium on Travel Demand Management, TDM 2008, Vienna, Austria, 2008.

- “***Multi-Sensor Data Fusion on Intelligent Transport Systems***”. Marco Veloso, Carlos Bento and Francisco Câmara Pereira. MIT Portugal, Transportation Systems, Working Paper Series (Paper# ITS-CM-09-02), 2009. (Veloso, et al., 2009).

- “***State of the Practice Overview of Transportation Data Fusion: Technical and Institutional Considerations***”. Andrew Amey, Liang Liu, Francisco Pereira, Christopher Zegras, Marco Veloso, Carlos Bento and Assaf Biderman. MIT Portugal, Transportation Systems, Working Paper Series (Paper# ITS-CM-09-01), 2009.

Chapter 2
How taxi patterns describe the city

2.1 Introduction

The evolution of society led to several changes in the organization of current demographics. Because of this fast growth, the urban areas are rapidly supporting the majority of the population (United Nations, 2008). Among other demands, there is the need to maintain a constant flow of people and vehicles. To optimize the public transportation it is essential to understand what drives the common citizen and what their needs are.

The taxi is a flexible way of transportation, since it is not bound to pre-defined paths or pick-ups and drop-offs locations. Taxi routes dynamically adapt to the flow and a city's need: it can pick-up the passengers right where they are standing, and drop them off precisely at the desirable destination. Therefore, taxis can provide more accurate information about the origins and destinations of passengers, in comparison to other traditional public transportation modes (e.g. bus, metro, train). Nevertheless, with the growth of urban areas, it becomes more difficult to move within the cities, and to efficiently search for new passengers.

New developments in ubiquitous computing technologies, alongside the wide variety of devices available, increase in processing and storage capabilities, and the integration of extended sensing capabilities, allow for new solutions. One of these solutions comprises collection of data from taxi movements enabling a study of their patterns. GPS-equipped taxis can be viewed as pervasive sensors, and the large-scale GPS traces produced, allow us to reveal facts about the social urban dynamics (Castro, et al., 2013).

We can envision a system that could help the taxi driver, by making recommendations of locations of where to find potential passengers, as pursuit by Yuan et al. (Yuan, et al., 2012a), Zheng et al. (Zheng, et al., 2010) and Yuan et al. (Yuan, et al., 2012a). These recommendations are supported by an inference engine and a database of past paths of the current driver and the taxi community. The inference engine should take in account a set of features, namely current location, hour of the day, day of the week, weather conditions, or proximity of points of interest (POI), in order to produce the best prediction for taxi drivers to find their next passenger, thus, optimizing resources (time and fuel), increasing sustainability (less pollution), and increasing service efficiency.

This chapter is focused on the research on taxi-GPS traces acquired in the city of Lisbon, Portugal, which will aid in better understanding urban mobility. The contribution of this work goes along the following aspects:

1. a spatiotemporal analysis of a data set of taxi-GPS traces, which identifies how taxis connect distinct regions of urban areas and uncovers taxi drivers strategies,
2. a proposal for a recommendation system, a type of inference engine,
3. a study of the predictability of taxi volume and its sensibility to various features.

For the first topic, we analyze taxi traces to identify relevant pick-up and drop-off locations referenced in time and space; study the relationship between those locations to produce a gravity map; characterize the search for new passengers (i.e. what happens between the latest drop-off and next pick-up) in order to improve taxi profit; and explore the value of Points Of Interest to analyze taxi flow. For the second contribution we evaluate a simple recommendation system based on a naïve Bayesian classifier, and explore the impact on the predictability by changing the configuration of variables, namely, temporal window, search space or taxi driver behavior. For the latter, we explore the possibility of estimating the likelihood of the next pick-up type of place given the previous drop-off hour of the day, day of the week, weather condition, and type of place.

2.2 State of the art

With the advent of pervasive technologies (e.g. GPS, GSM, Wi-Fi), several works have been performed to explore and improve urban mobility. Among them, mining taxi trajectories has recently attracted much attention. Taxi-GSP traces have been used in a number of studies to develop better solutions and services for urban areas such as:

- *Estimating optimal driving paths* (Krumm & Horvitz, 2006), (Yamamoto, et al., 2008), (Ziebart, et al., 2008), (Li, et al., 2009a), (Yuan, et al., 2010), (Zheng, et al., 2010), (Li, et al., 2011c), (Yuan, et al., 2011a), (Aslam, et al., 2012), (Hu, et al., 2012b), (Zhuang, et al., 2012), (Zhang & He, 2012), (Qian, et al., 2012), (Maciejewski & Nagel, 2013), (Yuan, et al., 2013);

- *Predicting next taxi pick-up locations* (Yang & Wong, 1998), (Yang, et al., 2000), (Wong, et al., 2001), (Wong, et al., 2008), (Yang, et al., 2002), (Yang, et al., 2005), (Chang, et al., 2008), (Lee, et al., 2008), (Chang, et al., 2010), (Chen, et al., 2010b), (Ge, et al., 2010), (Liu, et al., 2010b), (Phithakkitnukoon, et al., 2010b), (Li, et al., 2011a), (Li, et al., 2011b), (Powell, et al., 2011), (Takayama, et al., 2011), (Yuan, et al., 2011b), (Hu, et al., 2012a), (Moreira-Matias, et al., 2012a), (Yuan, et al., 2012a), (Zheng, et al., 2012b), (Qi, et al., 2013), (Tang, et al., 2013), (Gonzales, et al., 2014), (Moreira-Matias, et al., 2014b), (Qiu, et al., 2014), (Zhang, et al., 2014a), (Zhan, et al., 2014c), (Zheng, et al., 2014), (Wong, et al., 2015), (Yao, et al., 2015);

- *Improving dispatching services and detecting anomalies and frauds* (Liao, 2003), (Hao, 2004), (Lee, et al., 2004), (Li, 2006), (Tao, 2007), (Santani, et al., 2008a), (Santani, et al., 2008b), (Alshamsi, et al., 2009), (Cheng & Qu, 2009), (Xu & Huang, 2009), (Chen, et al., 2010), (Seow, et al., 2010), (Balan, et al., 2011), (Chen, et al., 2011), (Ge, et al., 2011), (Silva & Balassiano, 2011), (Lin, et al., 2012), (Orey, et al., 2012), (Sun, et al., 2012), (Wang, et al., 2012), (Wu, et al., 2012), (Chen, et al., 2013b), (Hou, et al., 2013), (Lee & Wu, 2013), (Ma, et al., 2013), (Santi, et al., 2013), (Wu & Lee, 2013), (Xiang, 2013), (Farkas & Dan, 2014), (Zhang, et al., 2014b), (Ivan & Popa, 2015), (Miao, et al., 2015);

- *Modeling driving strategies to improve taxi's profit* (Wong, et al., 2003), (Ge, et al., 2010), (Liu, et al., 2010b), (Yang, et al., 2010b) , (Moreira-Matias, et al., 2014a), (Yang, et al., 2015);

- *Identifying flaws and possible improvements in urban planning* (Gühnemann, et al., 2004), (Li, et al., 2007), (Li, et al., 2009b), (Wang, et al., 2009a), (Huang, et al., 2010), (Liu, et al., 2010c), (Bastani, et al., 2011), (Zhang, et al., 2011a), (Zheng, et al., 2011b), (Castro, et al., 2012), (Yue, et al., 2012), (Chen, et al., 2013a), (Grau, et al., 2013), (Martinez, et al., 2013), (Sun, et al., 2013), (Thompson & Bae, 2014), (Kartika, 2015), (Zhou, et al., 2015);

- *Developing models for urban mobility, social functions, and dynamics along different city's areas* (Schroedl, et al., 2004), (Matsushima & Kobayashi, 2007), (Zhang, et al., 2007), (Song, et al., 2008), (Liu, et al., 2009a), (Liu, et al., 2009b), (Lou, et al., 2009), (Yang, et al., 2009), (Yue, et al., 2009), (Bazzani, et al., 2010), (Chen, et al., 2010a), (Cooper, et al., 2010), (Girardin & Blat, 2010), (Austin & Zegras, 2011), (Liang, et al., 2011), (Qi, et al., 2011), (Salanova, et al., 2011), (Yue, et al., 2011), (Liu, et al., 2012a), (Liu, et al., 2012b), (Liu, et al., 2012c), (Liu, et al., 2012d), (Sagarra & Diaz-Guilera, 2012), (Yao & Cheng, 2012), (Yuan, et al., 2012b), (Zhang, et al., 2012a), (Castro, et al., 2013), (Pan, et al., 2013), (Salanova, 2013), (Zhan, et al., 2013), (Zhu, et al., 2013), (Amat, et al., 2014), (Farber, 2014), (Lee, et al., 2014), (Liu, et al., 2014), (Salanova, et al., 2014), (Thompson & Bae, 2014), (Wang, et al., 2014), (Zhang, 2014d), (Ding, et al., 2015), (Jianqin, et al., 2015), (Liang, et al., 2015), (Qian & Ukkusuri, 2015), (Shao, et al., 2015), (Tang, et al., 2015), (Wang, et al., 2015), (Moreira-Matias, et al., 2016).

Some of the most representative studies in each set are described in the following sections.

2.2.1 Estimating optimal driving paths

Significant work on taxi-GPS traces analysis focuses the problem of estimating optimal driving paths. The need to reduce time and fuel consumption is critical for taxis. Generally, proposed solutions rely on performing map-matching of trajectories and finding the shortest paths over large graphs as proposed by Ding et al. (Ding, et al., 2008) and Gonzalez et al. (Gonzalez, et al., 2008). The road network is usually perceived as a graph, with hotspots for pick-ups as nodes. Therefore, to minimize the cost of moving from one node to another, Dijkstra algorithm is often chosen, in order to find the best path.

Yuan et al. (Yuan, et al., 2010) presented the T-Drive system that identifies the optimal route for a given destination and departure time. The system uses a graph whose nodes are landmarks, defined as road segments frequently traversed by taxis. The authors proposed a Variance-Entropy-Based Clustering to adaptively split a day into different time segments, based on the travel time between landmarks. Improvements of the work were made by Yuan et al. (Yuan, et al., 2011a). The authors presented a cloud-based system to store historic information regarding traffic conditions, driver behavior and driving routes, alongside with information collected from internet sources, such as weather forecast or maps. Taxis were used as mobile and pervasive sensors to probe traffic conditions. The system is able to predict traffic conditions through Markov models, providing self-adaptive driving directions, considering the historic user behavior. The system is able to gradually learn from user behavior and traffic patterns.

Zheng et al. (Zheng, et al., 2010) described a three-layer architecture using the landmark graph to model knowledge of taxi drivers. The goal of the system is to provide the fastest route, given current location and departure time, relying on the taxi drivers' intelligence learnt from the historical taxi trajectories and an Interactive-Voting Based Map Matching Algorithm proposed by Yuan et al. (Yuan, et al., 2009). The system is supported by a database with taxi-GPS trajectories, generated by 33.000 distinct taxis. The authors claimed that the system outperformed concurrent approaches, such as speed constraint based and the real time traffic based methods, providing faster routes and saving up to 16% time of a trip. This is an implementation of the T-Drive system, presented by Yuan et al. (Yuan, et al., 2010).

Ziebart et al. (Ziebart, et al., 2008) presented a decision-modeling framework for probabilistic reasoning from observed context-sensitive actions. The model is able to make decisions regarding intersections, route, and destination prediction given partially traveled routes. Zhang & He (Zhang & He, 2012), also representing roads as segments, proposed the *pCruise*, a cruising system for taxicab drivers to find the optimal route to pick up a passenger. A graph (termed *Cruising Graph*), based on the road segments, is created with the location of nearby taxis in order to assign to each one the most probable and nearest passenger. The searching algorithm considers trip time and distance, to maximize the reduction of mileage.

Li et al. (Li, et al., 2009a) introduced the notion of road hierarchy, as an alternative do the classical Dijkstra algorithm. Roads are classified according to the frequency of use by taxi drivers. To produce a route, the hierarchical route planning algorithm will choose segments with high frequency of usage and computes the shortest path in duration. Using a similar algorithm, Yamamoto et al. (Yamamoto, et al., 2008) proposed a method to adjust the service dispatching assignments with each taxi driver's own behavior, which dynamically adapts according to the driver's current location and time. Routes are compiled to form a graph were a Dijkstra algorithm is applied to identify which customers are on the expected path of the driver.

Zhuang et al. (Zhuang, et al., 2012) proposed a weighted shortest path algorithm for route planning, based on past experiences of taxi drivers. The algorithm surveys a database to estimate the most chosen route between two locations, according to taxi drivers' behavior. Parameters such as turning frequency, number of signalized intersections, travel time and segment length are also taken into consideration to classify the shortest path. Aslam et al. (Aslam, et al., 2012) used taxi data to train a traffic congestion model in order to present a congestion-aware route planning system. This approach was able to reduce travel time by 15%.

2.2.2 Predicting next taxi pick-up hotspots locations

Identifying taxi passengers' hotspots is another subject attracting the attention of researchers. To avoid spending energy and time searching for passengers, as well as reducing the waiting time for customers, several authors proposed approaches to identify the most probable pick-up locations. Additionally, the identification of hotspots can also provide grounds for the development of services oriented to Location-Based Social Networks.

Yang & Wong (Yang & Wong, 1998) are among the first authors to analyze taxi drivers' behavior and develop a model to describe hotspots of vacant taxi. According to passengers Origin-Destination matrices, the model aims to provide taxi service equilibrium, using a set of non-linear functions to minimize the size of the taxi fleet, and distributing vehicles to expected areas of high demand. The authors also observed that average taxi utilization decreases with the number of taxis operating, and that the waiting time increases with higher taxi utilization. The work was later improved and extended by Wong et al. (Wong, et al., 2001), by incorporating traffic congestion into the model, and by introducing a new algorithm. This results in a two-level system that describes the movement of vacant and occupied taxis, as well as congestion on the road network. To meet customer demand and reduce waiting time, a Newtonian algorithm (a method for finding successively better approximations) with line search is applied. Further improvements were developed by Yang et al. (Yang, et al., 2002), (Yang, et al., 2005). The authors investigated the nature of demand–supply equilibrium in taxis. The new framework that they presented combines spatial road network, an origin-destination model according to customer demand and represents the competitive market by way of each individual taxi.

Lee et al. (Lee, et al., 2008) proposed a recommendation system for picking-up passengers. A clustering process using a k-means algorithm is used to define the granularity of the locations to be recommended. The authors observed that each location (cluster) has its own temporal patterns. Chang et al. (Chang, et al., 2008) (Chang, et al., 2010) proposed a four-step approach for mining historical data in order to predict taxi demand distributions based on time, weather, and taxi location. The model filters trajectories using contextual information (e.g. weather) and clusters GPS points into areas of high probability for predicting a passenger. Additionally, it defines a hotness score for each area according to the number of taxi requests divided by the size of the area. The authors show that different clustering methods have different performances on distinct data distributions.

Phithakkitnukoon et al. (Phithakkitnukoon, et al., 2010b) presented a model for predicting the number of vacant taxis for a given area of the city based on a naïve Bayesian classifier with an error-based learning algorithm (a weight linear function that emphasizes the recent errors from which the prediction is then adjusted). Additionally, a mechanism for detecting adequacy of historical data is explored, resulting in the conclusion that the latest 40 days of the data set are sufficient to maintain prediction accuracy. Liu et al. (Liu, et al., 2010b) classified taxi drivers according to their income. They observed that top drivers operate in a number of different zones while maintaining exceptional balance between taxi demand and traffic conditions. Regular drivers on the other hand operate in fixed zones with few variations. Additionally, the authors studied the influence of traffic conditions from a road segment to another road segment.

Ge et al. (Ge, et al., 2010) studied a way to extract energy-efficient transportation patterns from location traces, using taxi-GPS traces to develop a mobile recommendation system. Aiming for profit maximization and energy reduction, the authors proposed a Potential Travel Distance function to assess potential candidates and recommend routes. Li et al. (Li, et al., 2011a) studied two distinct strategies to find passengers: moving around (*hunting*) or waiting in specific locations. The authors observed that a combination of strategies is necessary, according to the time of the day: hunting for passengers in hotspots (especially transportation hubs) is preferable during commuting hours, while waiting is a better approach during hours with low taxi activity. To provide correct driving strategies to taxi drivers depending on time and location, the authors applied the L1-Norm regularization to linear Support Vector Machines (Bi, et al., 2003) to identify and select relevant features that determine taxi performances, i.e., according to the taxi driver location and time of the day, whether it is preferable to wait for a passenger, to go hunting, or to move to a distant location.

Xiaolong Li et al. (Li, et al., 2011b) aimed to predict pick-up hotspots for taxi drivers – areas with high demand. The authors proposed a method for predicting the amount of pickups at each hotspot by using a variant of the Auto Regressive Integrated Moving Average (ARIMA), a prediction method for time series analysis. By using passengers' mobility patterns and taxi drivers' picking-up/dropping-off behavior Yuan et al. (Yuan, et al., 2011b) proposed a system for passengers finding vacant taxis as well as for taxi drivers to find potential passengers, based on a probabilistic approach. Recommendations take into consideration time waiting for the passenger and the profit for the taxi driver. Later, the authors developed the T-Finder (Yuan, et al., 2012a), a recommendation system for both taxi drivers and passengers that takes into account the passengers' mobility patterns and taxi drivers' pick-up traces, using a database with taxi-GPS trajectories from 12.000 distinct vehicles during 110 days.

Considering the continuous increase of data provided by taxi logs and the rise of demand in the use of taxi service, Moreira-Matias et al. (Moreira-Matias, et al., 2012a) proposed a system to produce online short term predictions of passenger demand. To identify the best pick-up locations after a drop-off, the authors considered the number of vacant taxis in the area and passenger demand in the same area. The authors applied time series forecasting techniques such as time varying Poisson model and Autoregressive Integrated Moving Average. The system is able to accurately forecast in a 30 minutes time horizon, 76% of taxi demand. The authors further improved the system, using streaming data instead of traditional offline data set (Moreira-Matias, et al., 2013). Later, an online learning approach to predict profitability in taxi stops is proposed (Moreira-Matias, et al., 2014a), taking into consideration the type of services that are being requested by customers. This is achieved by computing an approximate revenue probability density function at each taxi stop, and by employing time series analysis techniques.

Zhang et al. (Zhang, et al., 2014a) analyzed historic taxi-GPS traces from experienced taxi drivers in order to characterize the efficient and less efficient taxi service approaches, according to passenger-searching strategies, passenger-delivery strategies, and service-region preference. A feature matrix is produced to describe the service approaches. A correlation between each service strategy and the corresponding revenue reveals the efficient and inefficient approaches in each time slot and location. The authors observed that hunting in specific locations is usually more efficient than waiting in order to find passengers. A few exceptions were identified, such as in airports. In suburban areas, the authors advise to return to hotspots of taxi demand, while during time slots with traffic congestion, taxi drivers should choose light-traffic routes in order to increase revenue.

2.2.3 Improving dispatching services

Associated with the identification of the most probable pick-up locations is the need to develop dispatching services that could deploy taxis more efficiently. These systems should avoid having different taxis compete for the same customers, leaving potential customers without transportation or waiting for a long period of time. With that goal, Lee et al. (Lee, et al., 2004) analyzed and proposed a new taxi dispatching system, which takes in consideration traffic conditions, in order to provide taxis in the shortest-time path possible, instead of the traditional nearest-coordinate method (in which the taxi assigned for each booking is the one with the shortest straight-line distance to the pick-up location). The proposed system is based on Dijkstra’s algorithm, to search for the shortest-time paths available for the taxis to reach the demand locations, considering real-time traffic conditions. Through a microscopic simulation model based on the Singapore Central Business District network, the authors were able to reduce passengers’ picking-up time by 50%. Balan et al. (Balan, et al., 2011) designed, implemented and deployed a real-time trip information system for taxis. The system aims to inform the passenger in real-time of the foreseeable duration, cost and path of the ride, based on a k-Nearest Neighbor clustering algorithm.

Chen et al. (Chen, et al., 2013b) proposed a different dispatching framework. By analyzing past logs, the system should rank road segments according to their attractiveness (amount of pick-ups and drop-offs). This should generate hubs where drivers will most likely be located. In a second phase, the authors developed two inter-hub routing algorithms: First-Come-First-Service (FCFS) and Destination-Closer

(DesCloser). Focusing on the quality and fidelity of customers, Ivan & Popa (Ivan & Popa, 2015) developed a dispatching system where the proximity of the taxis is not the fundamental parameter in the decision, but instead previous interactions between the passenger and the driver, though a feedback system. In order to achieve that, the authors introduced the Customer Relationship Management (CRM) component, which allows passengers to provide feedback on the service. After each taxi service, a passenger rates the taxi driver through a satisfaction survey. This information is used to decide which taxi driver will be assigned in future requests by the passenger.

Hou et al. (Hou, et al., 2013) proposed the Taxi Cruising Guidance, which aims to minimize the amount of vacant taxis by providing cruising route suggestions, instead of dispatching taxis on demand. The system models the road network on a graph with weights representing the likelihood of finding a new passenger and the state of traffic conditions. A Dijkstra's algorithm is implemented to identify possible routes, based on a set of heuristics to avoid competition among multiple taxis. The system is tested using real data and a microscopic traffic simulator.

2.2.4 Detecting anomalies and frauds

Detecting anomalous routes is also a concern in taxi services. Zhang et al. (Zhang, et al., 2011a), Ge et al. (Ge, et al., 2011) and Chen et al. (Chen, et al., 2011) studied this issue in order to uncover taxi driving fraud activities. Ge et al. developed a taxi driving fraud detection system, which is able to systematically investigate taxi driving fraud. The framework analyzes travel route and distance, comparing each individual path with standard routes from a database of taxi-GPS traces. A standard *routemark*, produced by statistical analysis of taxis logs, is used to identify possible evidence of deviations. In order to confirm evidences of fraud, the Dempster-Shafer theory is applied. As a result, the authors observed regular evidences of fraud in taxi driving activities. Zhang et al. proposed the *iBAT* (Isolation Based Anomalous Trajectory), a system to detect anomalous taxis' routes. The system relies on the road network being modeled by a grid of cells. Each trajectory is represented by a set of symbols (the cells of the grid), with statistical properties. Instead of using the traditional approach of distance or density measurements as used by Chen et al., the system analyses the amount of cells in common between the target trajectory and the historic log. Chen et al. applied a similar approach to detect anomalous routes, comparing the driver's route against historically standard routes. The authors developed the *iBOAT* system, which is based on the *iBAT*, able to detect anomalies in real-time and pointing out the road segments considered anomalous, by comparing Euclidean distances between paths.

Sharing the same taxi to reduce the amount of vehicles in circulation has been explored by Orey et al. (Orey, et al., 2012), Santi et al. (Santi, et al., 2013) and (Chen, et al., 2010). Orey et al. proposed a distributed and dynamic taxi-sharing algorithm to coordinate customers' requests. The system relies on distributed computing, based on passenger and taxi drivers' devices, without the need for a central computational system. It calculates the costs of a request (e.g. time waiting for the passenger, distance from the nearest taxi, destination of the passenger and nearby taxis) and decides if a vacant vehicle should be assigned or if a vehicle with a passenger passing by, should be detoured. Simulations showed benefits for both drivers and passengers. Santi et al. studied a similar problem in a bigger scale (Manhattan), applying classical methods from graph theory. The concept of taxi-sharing is also explored by Chen et al. The proposed system aims to identify fuel-saving paths according to request from passengers and current routes in progress, and reduce the amount of vehicles in circulation. The three-layer system models the city network into four matrices with size *n* (number of road intersections) representing transit time, time cost between two intersections, fastest path and fuel cost. The algorithm searches for the fastest path according to fuel consumption, using the Dijkstra's algorithm, and not the shortest path. A benefit function is introduced to maximize the taxi occupancy and minimize the fuel consumption.

2.2.5 Developing models for urban mobility and dynamics between the different areas of the city

The analysis of taxi-GPS traces is also used to study the city's design, in order to identity potential flaws and provide guidance for improvements in urban planning and discover specific dynamics between the different areas of the city. Considering that taxi patterns and origin-destination flows could represent the state of the traffic and the inhabitants' needs on commuting and transportation (Castro, et al., 2012), taxi data have been explored to improve other public transportation networks (e.g. bus and personal rapid transit), discover areas of the city that are disjointed or to identify social functions of the urban areas.

Zheng et al. (Zheng, et al., 2011b) identified flawed urban planning in region pairs with traffic problems and links among these regions, through their analysis of taxi traces. The urban areas are divided into segments according to main roads. By using taxi routes, the connections between different areas of the city are identified. The study of the density of traffic between pairs of segments can indicate flaws in the design of the city. To identify irregularities in the road network Huang et al. (Huang, et al., 2010) proposed the *MEtropolitan TAxis* (*META*) system. The framework explores taxi traces and analyzes three parameters (turn probability, road section speed and travel pattern) to evaluate the performance of road segments.

Wang et al. (Wang, et al., 2009a) used passenger pick-up and drop-off points to analyze the location and travel patterns to and from hotspots - as an indication of social interactions - aiming to improve location-based services. In order to achieve that, the authors used the Kernel Density Estimation and Agglomerative Hierarchical Clustering methods, applied to Origin-Destination matrixes from taxi-GPS traces. Liu et al. (Liu, et al., 2010c) had the same goal, but used a different approach. Instead of the traditional density-based clustering algorithm, the authors proposed a mobility-based clustering, where vehicles are used as a mobile sensor to perceive the crowdedness of nearby areas, and define hotspots according to vehicles speed. By mining historic taxi data, the authors are able to quantify the vehicle crowdedness of an area and investigate the evolution of hotspots. Along the same line, Zhang et al. (Zhang, et al., 2012a) aimed to identify social interactions from the analysis of taxi Origin-Destination (OD) pairs. The analysis reveals that the frequency of the most visited OD pairs follows a Zipf law (Powers, 1998).

Chen et al. (Chen, et al., 2013a) proposed the analysis of taxi-GPS traces to improve bus routes. The authors argued that by exploring the hotspots for pick-ups and drop-offs of taxi service at night, it is possible to identify the demand for transportation, and thus redesign the bus network. The authors proposed a probability based spreading algorithm and a set of heuristics to automatically build and prune a network graph for bus service, based on taxi traces. Thompson & Bae (Thompson & Bae, 2014) follow the same approach to propose a system that could redesign the personal rapid transit in Korea based on the analysis of taxi patterns.

Castro et al. (Castro, et al., 2012) used taxis as pervasive sensors for traffic conditions and proposed a method to construct a model of traffic density based on large scale taxi traces. The authors demonstrated that it is possible to predict traffic conditions alongside with the capacity of road segments. Much like Gühnemann et al. (Gühnemann, et al., 2004), the authors devised an approach to use traffic conditions in order to estimate emissions. Gühnemann et al. besides monitoring real-time traffic conditions and fleet management with a fleet of taxis equipped with GPS, proposed a framework to map between travel times, traffic flows and velocity profiles in order do estimate emissions, an improvement of the Handbook of Emission Factors (INFRAS, 1999).

Qi et al. (Qi, et al., 2011) investigated the relationship between regional pick-up and drop-off characteristics of taxis and the social function of city regions. They developed a simple classification method to recognize regions' social areas that can be divided into scenic spots, entertainment districts, and train/coach stations. The work is further improved by Pan et al. (Pan, et al., 2013). The authors introduced an improved clustering algorithm (iterative DBSCAN) to identify regions according to the characteristics of the data. Finally, Castro et al. work (Castro, et al., 2013), performed an extensive survey on the use of taxi-GPS traces to study a city's dynamics, classifying the existing research into three types of dynamics: social, traffic and operational.

To model the distribution of taxi trajectories, Liang et al. (Liang, et al., 2011) applied the Akaike Information Criterion (Akaike, 1974) to a data set of taxi-GPS traces. The authors observed that contrary to most models in human mobility, taxi displacement and elapsed time tend to follow an exponential distribution rather than a power-law, as identified by (Sagarra & Diaz-Guilera, 2012). By studying the GPS-traces of 35.000 vehicles in Florence, Italy, Bazzani et al. (Bazzani, et al., 2010) were able to identify three statistical laws for path lengths, activity downtime and degree of traffic behavior in urban areas, corroborating the finds of Liang et al. of an exponential distribution law.

The ubiquity and flexibility of taxis also suggests that they can be used to help city mapping (improving or creating new maps), since they cover most of the arterial roads. However, taxi devices provide low quality/coarse-grained GPS traces (with a low-sampling-rate of at around once every minute), which affects traditional map-matching algorithms such as incremental algorithm (local method) and Average-Fréchet-Distance (global method), which are more suitable for high-sampling-rate traces. Several approaches have been proposed by Schroedl et al. (Schroedl, et al., 2004), Lou et al. (Lou, et al., 2009), and Liu et al. (Liu, et al., 2012a) to overcome these constraints and use information from taxi-GPS traces.

Schroedl et al. (Schroedl, et al., 2004) aimed to improve digital maps by mining taxi-GPS traces and using a clustering algorithm. The proposed algorithm analyzes individual trajectories and divides them into road segments and intersections. From each segment, a center line is derived, and the center lines from distinct trajectories are clustered to determine the lane position. Intersections are obtained from the transitions between road segments, where traces may diverge and follow more or less constrained trajectories. The use of a large database enables the system to overcome noisy data.

Lou et al. (Lou, et al., 2009) proposed a map-matching algorithm called ST-Matching for low sampling-rate GPS trajectories, which considers the spatial geometric structures of the road network and the temporal and speed constraints of the trajectories. The algorithm constructs a candidate graph from which the best matching path sequence is identified, using Dijkstra's algorithm to compute the shortest path. The authors demonstrated that ST-Matching algorithm outperforms Average-Fréchet-Distance algorithm when dealing with low-samping-rate traces, with a tradeoff of reasonable increase of time complexity.

Liu et al. (Liu, et al., 2012a) proposed an algorithm to infer road maps from large scale coarse-grained vehicular GPS traces. The algorithm consists of three main steps: pruning low-quality samples, clustering relevant samples for the same road segments and applying shape-aware B-spline fitting technique, which treats the curves as smooth piecewise-polynomials. The resulting road network was compared against OpenStreetMap[8], proving to be more accurate and able to cover 93% of the arterial roads present in the online mapping platform.

2.2.6 Our approach

The majority of the described works focus on two main topics: identifying hotspots for picking-up potential customers and identifying the most efficient routes. From these two fields of study, other approaches arise such as analyzing efficient dispatching systems, identifying taxi frauds, improving profitability, or studying urban planning design.

[8] Open Street Map. http://www.openstreetmap.org/.

Our work addresses some of these topics. Although our main topic is the identification of hotspots for the next pick-up, we also explore how the city is inter-connected by taxis and analyze taxi drivers' strategies. Our approach is distinct and contributes in the following aspects:

(1) performs an exploratory analysis which describes taxis as bridges between transportation hubs in the city and identifies taxi drivers' strategies in order to improve their profit during the search for potential customers;

(2) performs an inferential analysis aiming to estimate probable pick-ups locations based on a Naïve Bayes Classifier, considering not only spatiotemporal variables, but also weather conditions and points of interest that describe functions of different areas of the city. Our approach, different from other approaches, considers the entire urban area - modeled by a grid - as potential pick-up locations (instead of just assuming a set of possible points).

2.3 Methodology

Our approach is based on the classic process of knowledge extraction from databases as described by Witten & Frank (Witten & Frank, 2005) and Santos & Azevedo (Santos & Azevedo, 2005). It comprises the following steps:

- Data collection;
- Data cleaning and transformation;
- Exploratory analysis;
- Inference analysis;
- Validation.

The main data set was collected and provided by Geotaxi. It comprises a database of taxi-GPS traces, described in the following section. Additional data was collected, namely weather conditions (retrieved from Weather Underground), Points of Interest (provided by Sapo) and shape files, GIS data and Census information (provided by '*Instituto Nacional de Estatística*').

A cleaning and transformation procedure is performed, in order to remove noisy and missing data, as well as to format the data set into a more suitable scheme. This step also allows for a reduction in the size of the data set. To better handle the size of the metropolitan area, the city under study is modeled on an initial grid of 500x500m cells, as suggested by Huang et al. (Huang, et al., 2010) and Liu et al. (Liu, et al., 2010a). Finally, data is stored in a relational database were every sample is geo-referenced.

An exploratory study is performed to understand the spatiotemporal distribution, aiming to identify hotspots of activity or potential patterns to be further investigated during inferential analysis. Data exploration is designed to examine the spatiotemporal distribution; to identify hotspots of activity and how they are connected; to explore the taxi strategies in order to find the next pick-up; and to probe how different areas of the city with distinct functions (e.g. residential, commercial, recreation, education) affect the taxi service.

The main goal for the inference analysis performed on taxi-GPS traces is to study to what extent it is able to estimate the most likely locations to pick-up new passengers, based on the current time, location and historic data. Moreover, Yuan et al. (Yuan, et al., 2011a) and Chang et al. (Chang, et al., 2010) hypothesized that weather conditions could affect the behavior of inhabitants, and therefore, taxi patterns, while Qi et al. (Qi, et al., 2011) and Pan et al. (Pan, et al., 2013) suggested that social functions of distinct regions of a city influence taxi flows. Considering these set of variables, a probabilistic approach is pursued in the form of a Naïve Bayesian Classifier, due to its simplicity, how quick it is to build and train, and how adequate it is for problems where data is linearly separable, as proposed by Zhang (Zhang, 2004) A Bayesian approach was also explored by Fusco (Fusco, 2003) to model urban mobility. Different setups are tested, exploring the effect of daily and weekly periods on the performance of the model, as well as the influence of the size of the grid.

To validate the model a ground truth is used, consisting of historic data from taxis, weather conditions and POI data set. Considering the size of the data set (170,000+ samples) and the existence of a temporal attribute, the samples are organized into training and testing subsets, following a holdout configuration (the oldest 2/3 forming the training set and the latest 1/3 forming the testing set). Entropy, from information theory, is used to access the randomness or uncertainty of variables. Main findings and achievements were submitted to peer-review international conferences, to validate our work with the scientific community (Phithakkitnukoon, et al., 2010b), (Veloso, et al., 2011a), (Veloso, et al., 2011b), (Veloso, et al., 2011c).

2.4 Data description

This section analyzes and describes taxi data in Lisbon, Portugal, alongside with points of interest and weather conditions. The data was collected from September to December 2009.

2.4.1 Target city

The area of study corresponds to the municipality of Lisbon, which, as of 2012, consisted of 53 parishes, an area of around 110 km2, and a population of 800,000 habitants, represented in Figure 2-1, where (*A)* is the Downtown; (*B)* Airport; (*C) Oriente* Train Station; *(D) Santa Apolónia* Train Station; (*E*) Ferry dock; (*F) Marquês do Pombal* (City Center); (*G)* Univ. Campus; (*H)* Commercial Area; and (*I)* Residential areas. In 2013, those parishes were fused due to an administrative process forming 24 new parishes.

The city’s downtown is characterized by a higher population density (red) including touristic, historic and commercial areas, and the interface for several public transportation services. Encircling the city center, there are residential areas surrounding business areas with lower population density (yellow).

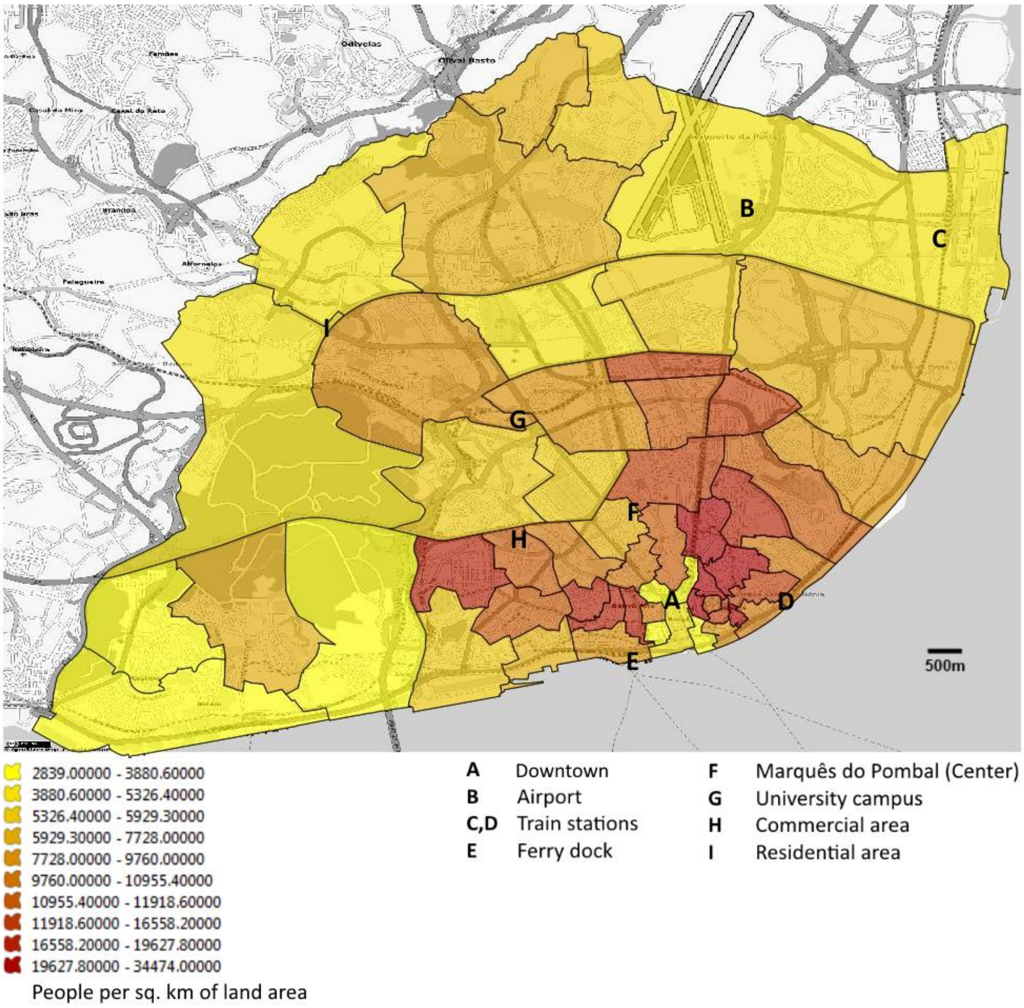


Figure 2-1– Lisbon municipality and population density.

Major infrastructures (e.g., airport and industrial facilities) are located in the city's outskirts. The public transportation system consists of buses, metro, trains, and ferry. All transportation systems (trains, buses and taxis) have station hubs within the city center, enabling a multimodal transportation system. Figure 2-2 represents the road network with average speed of circulation. Red segments correspond to higher average speed (between 100 and 120 km/h) and represent expressways.

The city is modeled by a 500mx500m-grid cells, splitting the urban space into disjoint areas (396 cells), facilitating the visualization process and further processing (e.g. eliminating empty areas), as suggested by Huang et al. (Huang, et al., 2010), Liu et al. (Liu, et al., 2010a) and Zhang & He (Zhang & He, 2012). Other authors chose to use digital maps and map-matching techniques (Krumm & Horvitz, 2006), (Yuan, et al., 2010), (Schroedl, et al., 2004), or split the city via road hierarchy (Zheng, et al., 2009), (Liu, et al., 2012a), (Gonzalez, et al., 2007), which are more complex processes and require further information about individual road segments and the road network. The cell size selection of 500m side by side is based on the average maximum walking distance travelers are willing to cover to reach a destination, without using any other mean of transportation (Dunning & Ford, 2003), (Daniels & Mulley, 2011), (Thompson & Bae, 2014), and a measure used by transportation authorities to design public transportation hubs (Public Transport Authority, 2003). However, the size of the cells will be subject to study later on in this work.

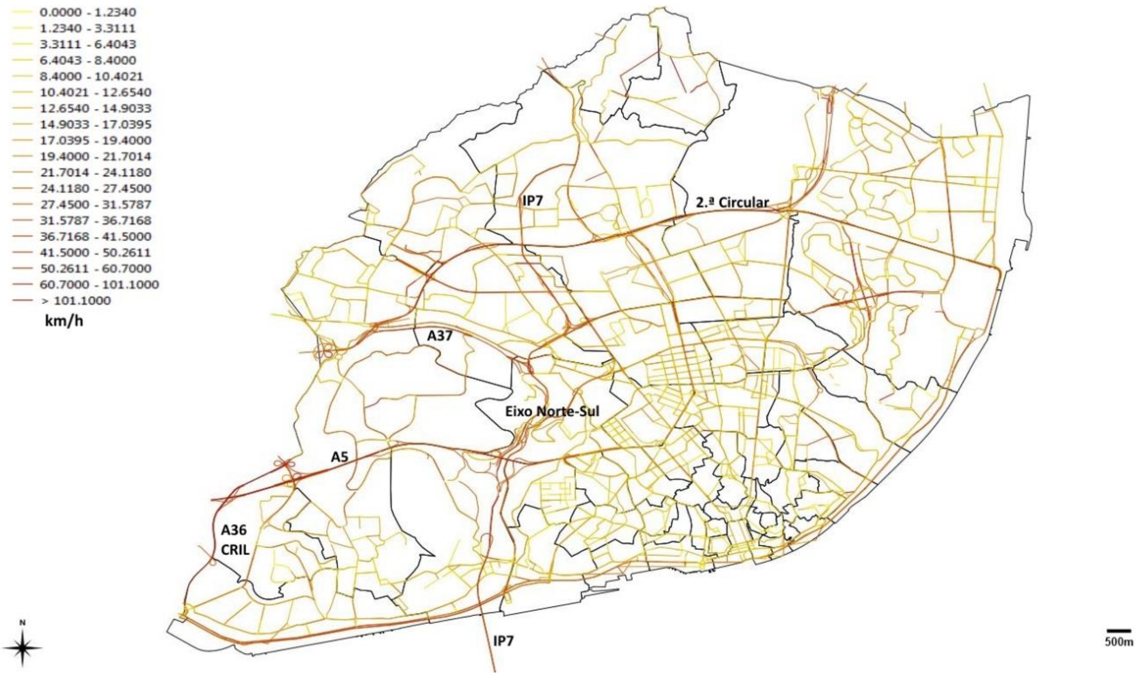


Figure 2-2 – Lisbon's road map and average speed.

Considering the Earth's curvature, the measures for the grid were computed using the *Haversine distance*, which computes the distance (*d*) between two points in a sphere (Brummelen, 2012), defined as:

$$d = 2.R.arcsin\left(\sqrt{sin\left(\frac{\emptyset_2 - \emptyset_1}{2}\right)^2 + cos(\emptyset_1).cos(\emptyset_2).sin\left(\frac{\lambda_2 - \lambda_1}{2}\right)^2}\right) \quad (1)$$

where $(\emptyset_1, \lambda_1)$ are the latitude and longitude of the first point, $(\emptyset_2, \lambda_2)$ are the latitude and longitude of the second point, and *R* the radius of the sphere (in this case the mean radius of the Earth, which is 6,371km).

2.4.2 Taxi data

The taxi data set was provided by *GeoTaxi*[9], a company that focuses on software development for fleet management, and that in 2009 held about 20% of the taxi market share in Portugal. The data set was composed of around 1,600,000 taxi-GPS location points and collected by 230 taxis, for a period of four months (from September 1st to December 31st). Along with the GPS location (latitude, longitude) information, it includes speed, bearing, engine status, and occupancy status of the taxi, as described in Table 2-1. A sample on the raw data is presented in Table 2-2. Our study focused mostly on the information provided by the taxi location (latitude and longitude) and occupancy. The sampling rate varies from trajectory to trajectory and can be triggered by three events: time intervals (DOS), change in status of the vehicle (COS), or distance covered (POS). On average, the system produced a sampling rate of 30 seconds per sample.

The amount of inferred pick-ups and drop-offs accounted for 177,169 distinct trips and is termed *taxi volume* in this study. Besides the features extraction for each trip, a data enrichment step is performed, adding features extracted directly from the observed trips: distance traveled (from pick-up to drop-off), trip duration, average speed, trip income[10], downtime distance (distance traveled for the next pick-up), and downtime duration (time taken searching for the next pick-up). Additional features are added using other data sources: nearest point of interest, and weather condition.

The data cleaning process is applied to remove trips with a length less than 200m and greater than 30km (the realistically longest trips from one side of the city to the other could be around 22km), and, in terms of a time window, less than a minute and longer than three hours. Erroneous data entries, missing data entries, false flags were also detected, and respective samples were removed.

[9] GeoTaxi. http://www.geotaxi.com

[10] The income was calculated from data using the ANTRAL standard formulation http://www.antral.pt/simulador.asp. ANTRAL is a national association for transportation.

Atribute	Type of data	Format	Description	Value
Tipo Mensagem	String	xxx	Message type: DOS = time intervals; COS = sensor change; POS = by distance	DOS; COS; POS
Sinal GPS	Char	x	GPS Signal: A = precision coefficient high; V = precision coefficient low	A; V
Se	Integer	xx	Sequential value of messages	
Latitude	float	xxx.xxx	Latitude	
Longitude	float	-xxx.xxx	Longitude	
Velocidade	float	x.x	Velocity	
Direccao	float	xxx.x	Heading	[0, 359]
Data Envio Caixa	String (date hour)	dd-mm-yyyy hh:mm	Timestamp for data acquisition	
idcaixa	Integer	xxxx	Device ID (usually one device per vehicle)	
Data	String (date)	dd-mm-yy	Date of data processing	
Hora	String (hour)	hh:mm	Hour of data processing	
Ignicao	String	xxx	Ignition state (ON/OFF)	ON; OFF
Estado Servico	String	xxxxxxx	Occupancy (Ocupado = a passenger in the vehicle; livre = no passenger)	Livre; Ocupado
Taximetro	String	xxx	Taximeter status (ON/OFF)	ON; OFF
Alimentacao	String	xxx	Energy status (ON/OFF)	ON; OFF
Sensor 4	String	xxx	Not used	ON; OFF
Sensor 5	String	xxx	Not used	ON; OFF
Sensor 6	String	xxx	Not used	ON; OFF
Sensor 7	String	xxx	Not used	ON; OFF

Table 2-1 - Attribute description from taxi-GPS traces’s raw log.

```
"POS";"A";30;38.7116;-9.42355;0;0;"06-12-09 01:34:37";2091;"2009-12-
06";"03:11:21";"ON";"Ocupado";"OFF";"ON";"OFF";"OFF";"OFF";"OFF"
"POS";"A";88;37.0914;-8.22116;0;0;"06-12-09 01:34:37";2506;"2009-12-
06";"03:11:22";"OFF";"Ocupado";"OFF";"ON";"OFF";"OFF";"OFF";"OFF"
"POS";"A";86;38.7272;-9.12833;14.91;182.42;"06-12-09 00:27:27";534;"2009-12-
06";"03:11:22";"OFF";"Ocupado";"OFF";"ON";"OFF";"OFF";"OFF";"OFF"
"POS";"A";49;38.7681;-9.16966;0.1;0;"06-12-09 01:34:38";540;"2009-12-
06";"03:11:23";"OFF";"Ocupado";"OFF";"OFF";"OFF";"OFF";"OFF";"OFF"
"POS";"A";26;37.0912;-8.24094;0;0;"06-12-09 01:34:39";2540;"2009-12-
06";"03:11:23";"ON";"Ocupado";"ON";"ON";"OFF";"OFF";"OFF";"OFF"
"POS";"A";47;38.7039;-9.40856;0.1;0;"06-12-09 01:34:39";2140;"2009-12-
06";"03:11:23";"ON";"Ocupado";"OFF";"ON";"OFF";"OFF";"OFF";"OFF"
"DOS";"A";65;38.783;-9.18566;0.1;118.44;"06-12-09 03:00:22";54;"2009-12-
06";"03:11:23";"OFF";"Ocupado";"OFF";"ON";"OFF";"OFF";"OFF";"OFF"
"DOS";"A";87;38.7272;-9.12833;14.91;182.42;"06-12-09 00:27:27";534;"2009-12-
06";"03:11:23";"OFF";"Ocupado";"OFF";"ON";"OFF";"OFF";"OFF";"OFF"
"POS";"A";0;38.7134;-9.29041;72.8;84;"06-12-09 01:34:38";2071;"2009-12-
06";"03:11:23";"ON";"Ocupado";"OFF";"ON";"OFF";"OFF";"OFF";"OFF"
"POS";"A";31;38.7087;-9.42206;12.6;184.9;"06-12-09 01:38:07";2091;"2009-12-
06";"03:11:23";"ON";"Ocupado";"OFF";"ON";"OFF";"OFF";"OFF";"OFF"
"POS";"A";29;38.7999;-9.43983;32.9;258.5;"06-12-09 01:35:10";721;"2009-12-
06";"03:11:24";"ON";"Livre";"OFF";"ON";"ON";"OFF";"OFF";"OFF"
```

Table 2-2 - Sample of taxi-GPS traces’s raw log (10 records).

The individual pick-up (red) and drop-off (green) locations of taxi service are depicted in Figure 2-3. Taxi service encompasses all the city, and some clusters appear to form (e.g. airport, train stations and downtown).

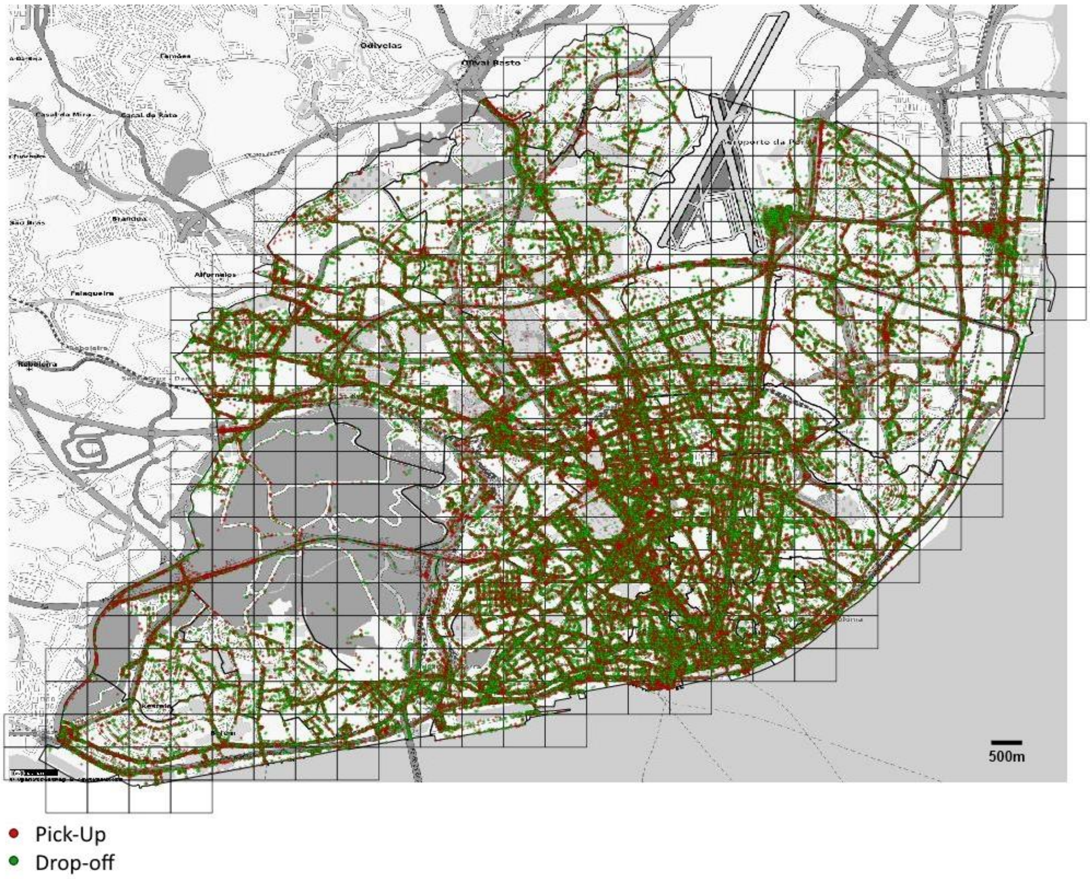


Figure 2-3 - Spatial distribution of pick-ups (red) and drop-offs (green).

To better visualize the data, samples are distributed to the corresponding cell of the grid that models the city. The density of taxi volume is represented by a color scale (red corresponds to cells with a higher number of pick-ups and drop-offs). The overall taxi volume's spatial distribution in Lisbon is shown in Figure 2-4 (on 500mx500m-grid cells). Some major locations are identified, such as downtown (*A*), airport (*B*), train stations (*C, D*) and ferry dock (*E*). Most of the taxi hotspots, for pick-up and drop-off, match taxi stations (specific stopping areas for taxi service), located near other transportation modes. Therefore, different public transportation modalities (airport, train, ferry, bus) are well connected through taxi services. Although these locations are predominant locations for pick-ups and drop-offs of taxi service throughout the day, other locations can become more active during different hours of the day, as presented in Figure 2-6.

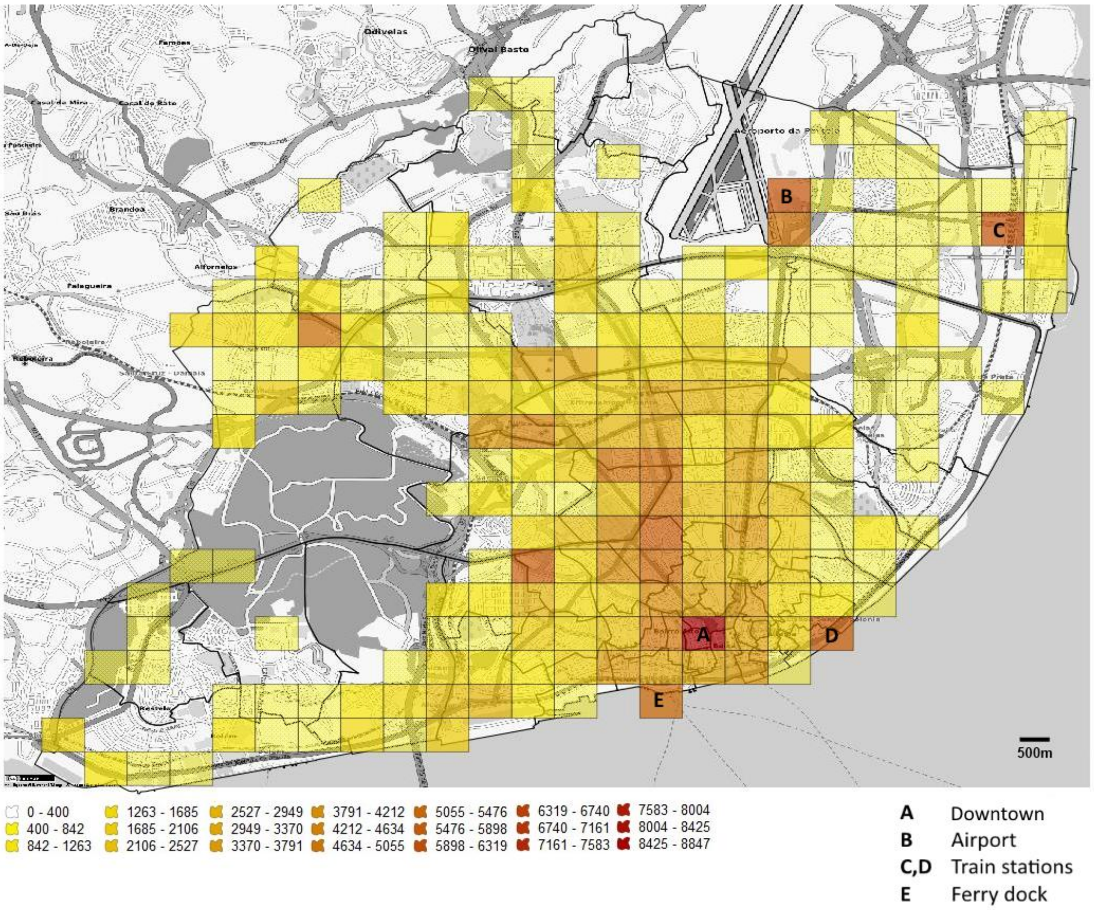


Figure 2-4 - Spatial distribution of taxi volume (number of pick-ups and drop-offs).

Based on the density of pick-ups and drop-offs, we can classify each cell into high taxi activity (top 30% of cells with higher taxi volume), medium taxi activity (middle 40% of the cells), and low taxi activity (bottom 30% of the taxi volume cells). From the high activity cells set, a subgroup can be highlighted: very high taxi activity, corresponding to the top 2% of the cells with highest taxi volume. The result is shown in Figure 2-5, where taxi activity is represented in a color scale, from red (very high taxi activity) to yellow (low taxi activity). In concordance with the scenario observed in Figure 2-4, the six cells with very high activity are located near the airport, train stations, ferry dock, downtown and a commercial area. Therefore, five (out of six) cells with very high taxi activity are located near transportation hubs.

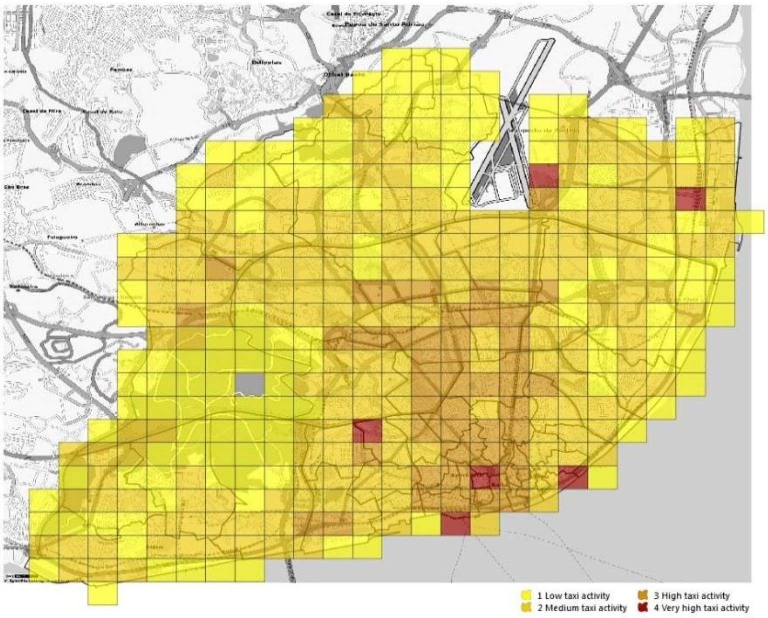


Figure 2-5 – Classification of cells according to taxi activity.

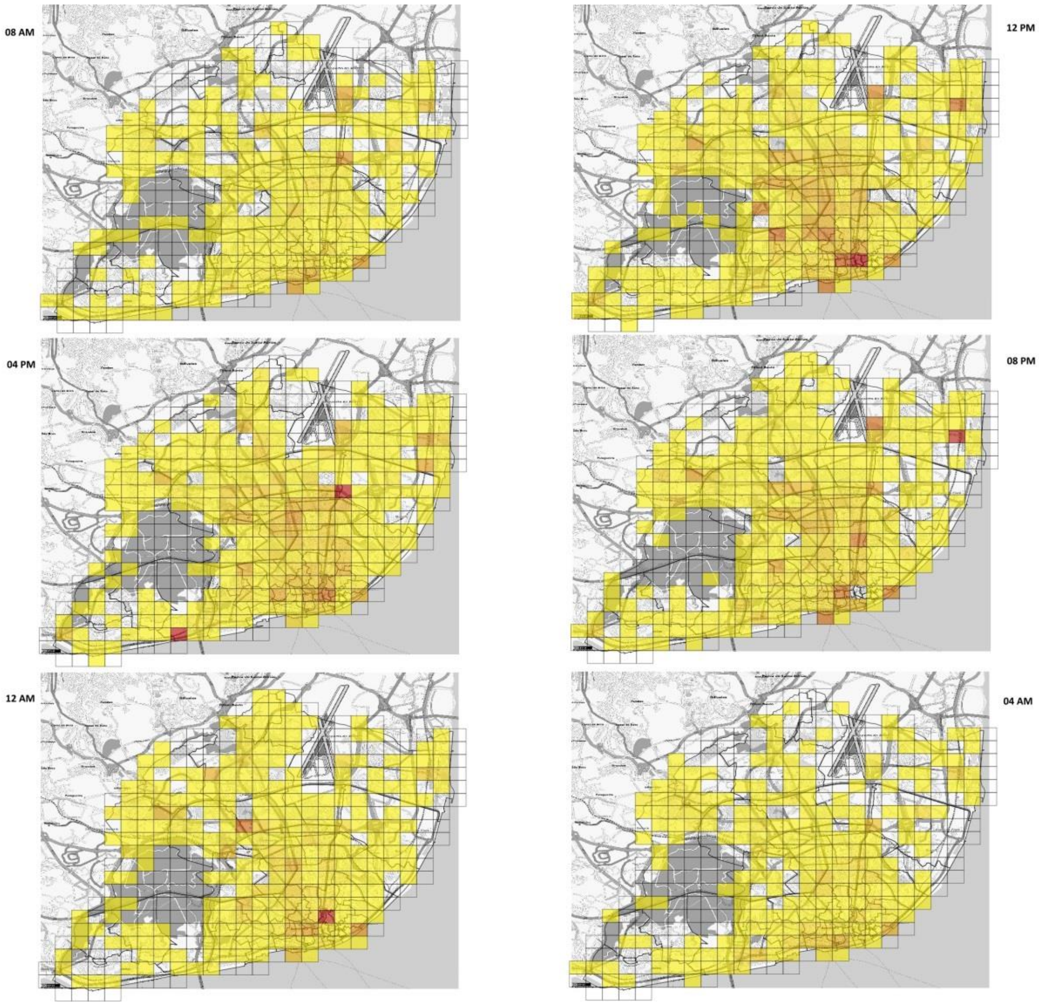


Figure 2-6 - Spatial variation of taxi pick-ups in different time slots (red corresponds to cells with a higher number of pick-ups).

Taxi volume also varies in time. Figure 2-7 presents temporal variation of the taxi services. As expected, the taxi service variation follows the business hours. It gradually increases from 7 AM, reaches the maximum between 11 AM and 1 PM, and slowly drops down in the afternoon. In the same way, there are more taxi services on working days than on weekends. On average, we observed a reduction of taxi volume of about 46.7% at night (from 10 PM to 7 AM) and 13.6% on weekends.

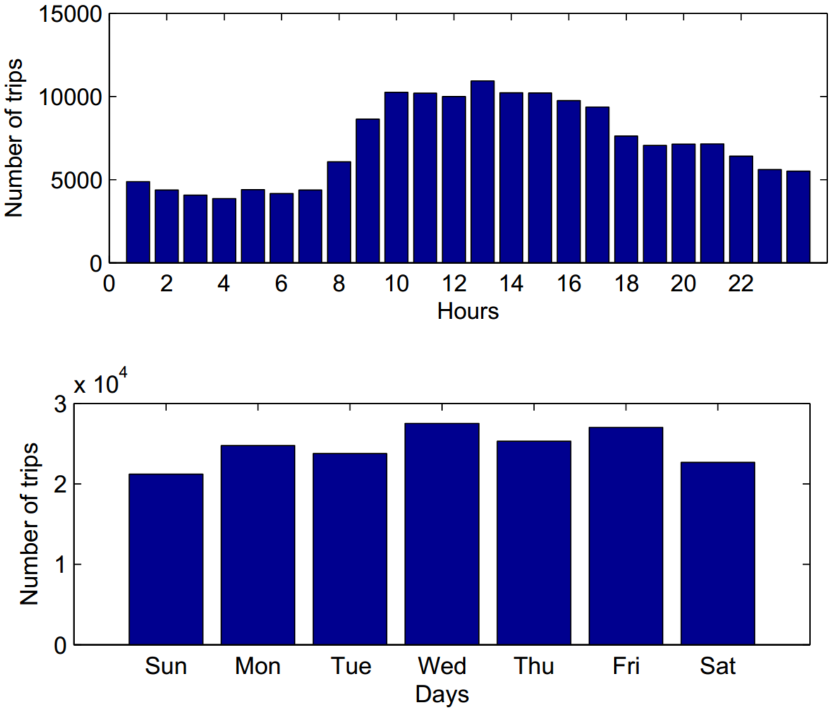


Figure 2-7 - Taxi volume variation according to hours of day (top) and days of week (bottom).

This overall pattern is consistent when cells with different taxi activity are analyzed individually. Figure 2-8 show the hourly variation of taxi activity on cells with very high, high, medium and low taxi activity, as described in Figure 2-5. Values of taxi activity are normalized in the interval [0,1] to enable the comparison. In all cells an increase of activity is visible from 7 AM on, with activity peaking between 11 AM (medium and low activity cells) and 1 PM (high and very high activity cells), followed by a decrease of activity in the afternoon. This reduction is more accentuated on very high activity cells than on low activity cells. Nevertheless, although individual cells follow the overall pattern, each one has its unique and specific minor deviations due to several factors (e.g. area type, proximity of other transportation means, road accessibility, type of commerce and services in the vicinities, events taking place, inhabitants' choice of transportation).

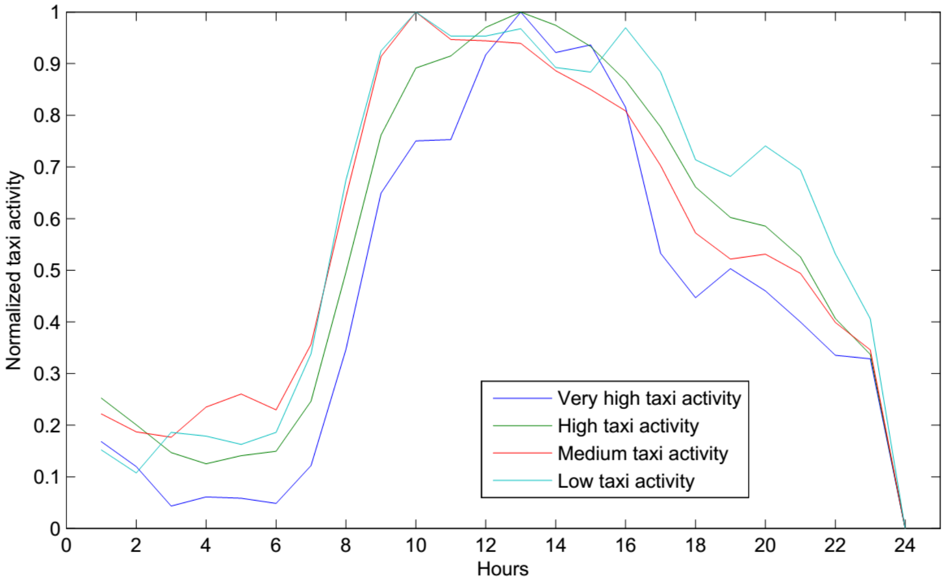


Figure 2-8 - Hourly variation of taxi activity on different cells (very high, high, medium and low taxi activity).

2.4.3 Points of interest

Sapo Mapas[11] provided a collection of 10,954 Points Of Interest (POIs) for Lisbon municipality, grouped into eight categories[12] (Services, Recreation, Education, Shopping, Police, Health facilities, Transportation and Accommodation), with a distribution represented in Figure 2-9. Those POIs are used to characterize the area type. Education facilities (e.g. kindergarten, high school, university, etc.), Recreation (e.g. bar, restaurant, etc.) and Services (e.g. bank, etc.) are the dominant POI categories (which account for over 70% of the total number of POIs).

[11] Sapo Mapas. http://mapas.sapo.pt/ .

[12] The classification was performed by the data provider.

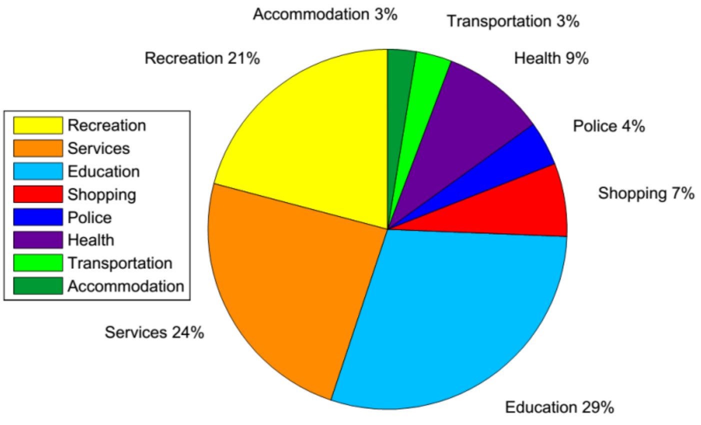


Figure 2-9 - POIs categories distribution.

Figure 2-10 represents the raw map of POIs and the underlying POI density distribution (red cells correspond to a higher concentration of POI, while yellow represents cells with low density of POI). As expected, the POIs are mainly present in areas with a higher population density or commercial facilities. The main cluster is located at downtown.

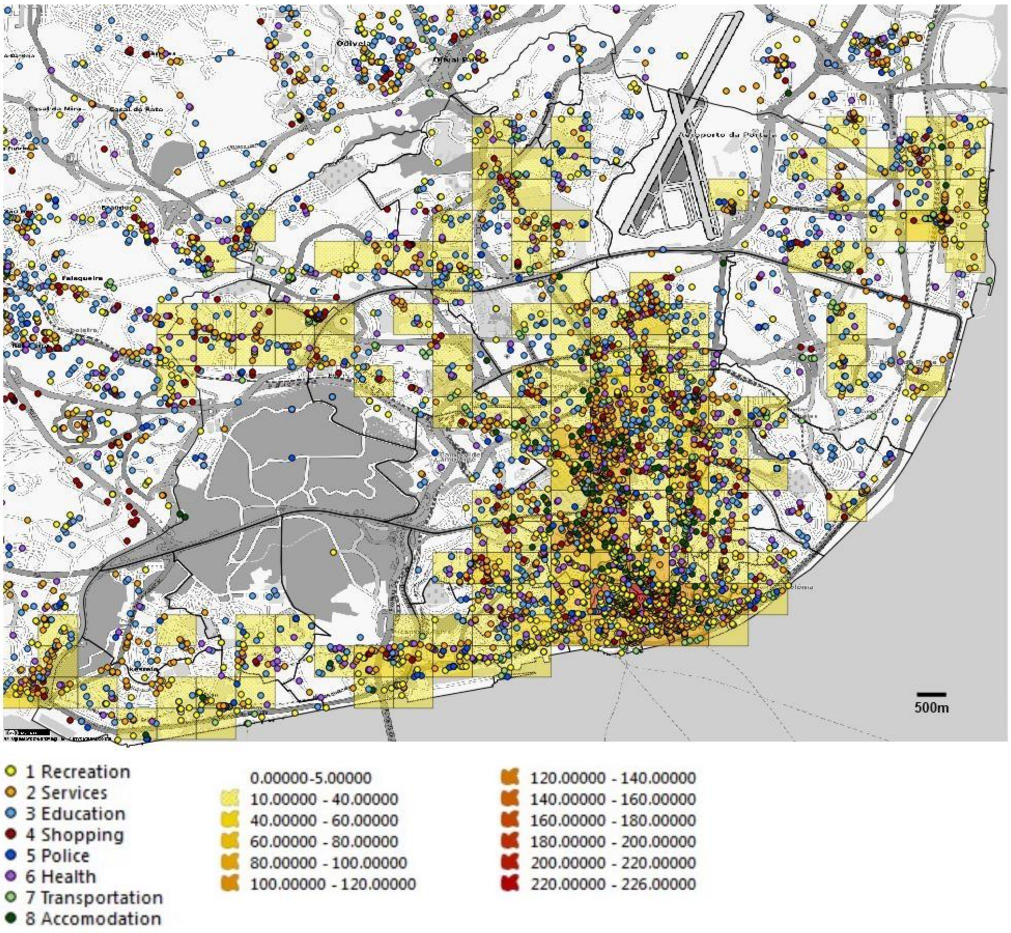


Figure 2-10 - POI's raw map and density distribution.

Figure 2-11 aggregates POIs in order to identify the predominant POI on each cell grid, according to the classification presented in Figure 2-9. Recreation by the riverside of the Tagus is the most predominant type of POI. The city center is characterized by services while education is predominant in the remaining areas.

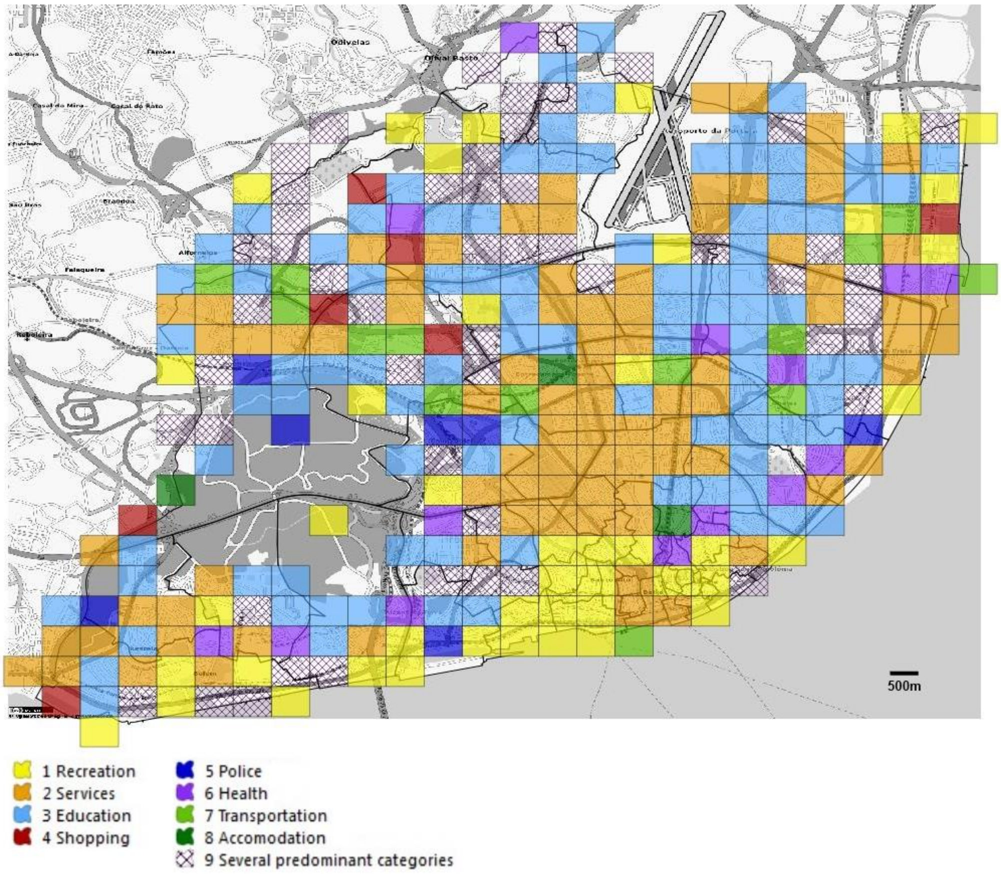


Figure 2-11 - Predominant POI category on each cell.

2.4.4 Weather conditions

Information on weather conditions was retrieved from the *Weather Underground*[13]; an online weather information service provider. The data set contains hourly measurements of weather conditions, arranged into three categories (Clear; Cloudy; Rainy), as shown in Figure 2-12. The online service provides a wide range of historic data, however only data from September to December 2009 were retrieved to match the same temporal window of the taxi data set. Clear weather days in

[13] Weather Underground http://www.wunderground.com/ ; http://www.wunderground.com/history/

September progress to an increased amount of cloudy days in October, eventually giving way to rainy days in November and December.

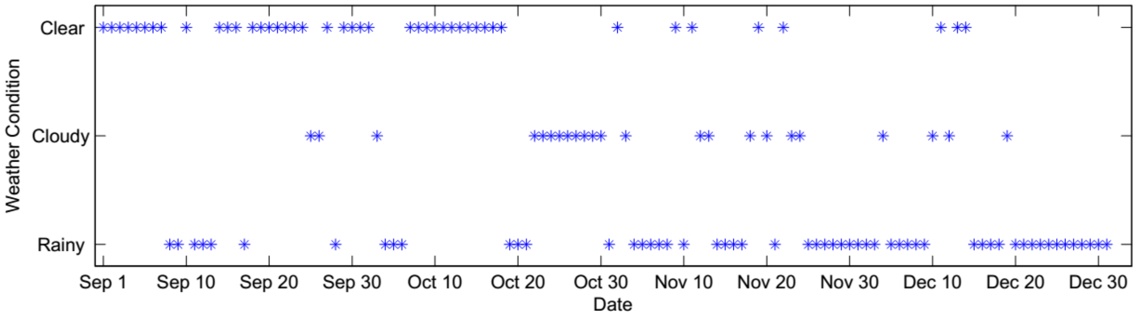


Figure 2-12 - Weather conditions for each day, from September 1 to December 31, 2009.

2.5 Data exploration

This section explores the spatiotemporal distribution of taxi-GPS traces in order to identify patterns and relationships to be further examined during the inference analysis. Taxi drivers' strategies in searching for new passengers are also studied.

2.5.1 Spatiotemporal analysis

Taxi demand varies in time and space, according to citizens' needs. Figure 2-7, of the previous section, presents the taxi service variation according to hours of the day and days of week. As expected, taxi service variation follows business hours. It gradually increases at 7 AM, reaches the maximum between 11 AM and 1 PM, and slowly drops in the afternoon. In the same way, there are more taxi services during working days than on weekends. In both cases the maximum is reached at the middle of the periods (11 AM to 1 PM and in Wednesday). This daily and weekly pattern is consistent throughout the four months of data analyzed (from September to December), with cyclic components, as shown on the time series of taxi activity in Figure 2-13, which represents the amount of taxi trips each hour. Visible punctual disruptions to this pattern take place during the last two weeks of December, on specific days, coinciding with Christmas Eve, Christmas day and New Year's Eve, characterized by a reduction of taxi service.

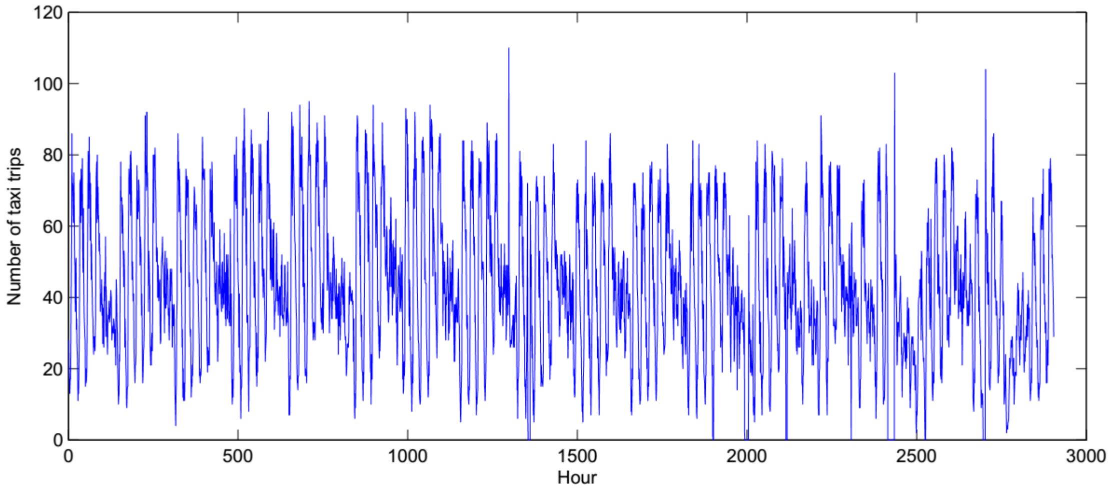


Figure 2-13 - Time series of taxi trips in Lisbon, from September 1 to December 31, 2009.

Figure 2-14 presents the taxi service distribution in Lisbon, according to the pick-ups (left) and drop-offs (right), where some major locations are identified, such as downtown (*A*), airport (*B*), train stations (*C*, *D*) and ferry dock (*E*). The predominant cells for taxi pick-ups are simultaneously the predominant cells for taxi drop-offs. As aforementioned, most of the taxi hotspots, for pick-up and drop-off, coincide with taxi stations, located near other transportation modes.

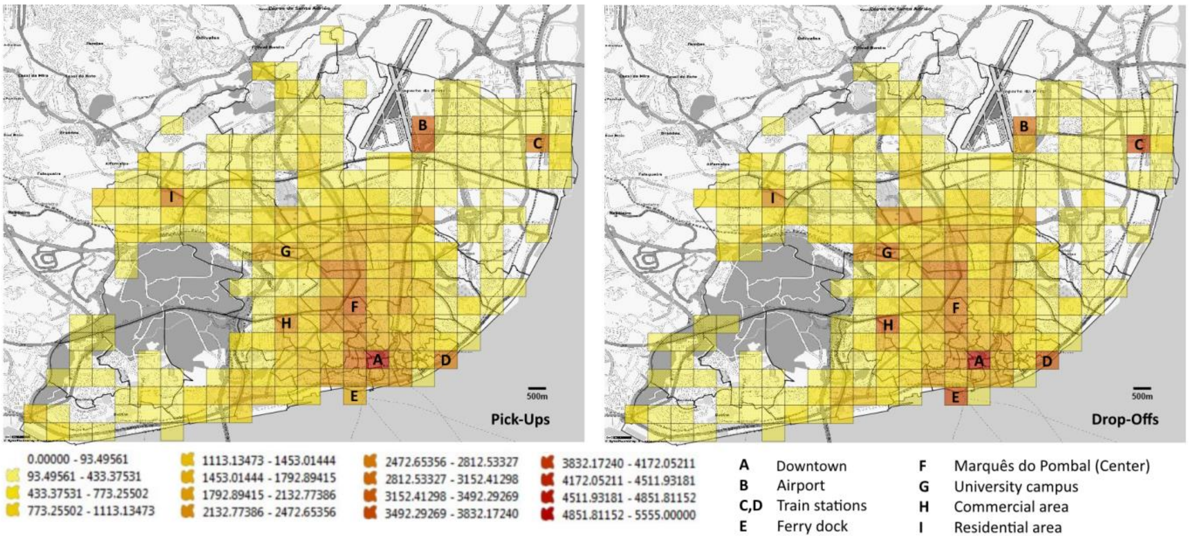


Figure 2-14 - Taxi pick-up (left) and drop-off locations (right).

Weather conditions could play an important role in deciding on how to travel (Yuan, et al., 2011a). As weather conditions worsen the average number of daily taxi trips slightly increases, in agreement with Farber's observations (Farber, 2014). Inhabitants prefer taxis and carpooling when weather conditions worsened, therefore avoiding walking or using public transportation hubs that could be located away from their starting and ending journey or if no shelter was provided, as observed by Stover & McCormack (Stover & McCormack, 2012). However, in our observations, this increase in usage is slight. From clear conditions to cloudy conditions the average increase of taxi volume is just 1.8%, and from cloudy conditions to rainy conditions, the increase is just 0.8%. Therefore, taxi volume does not change considerably given different weather conditions. This observation is in agreement with the time series depicted in Figure 2-13. Although the data set encompasses a warm summer month (September) and a cold winter month (December), the daily pattern is fairly similar (with the last two weeks of the data set being the exception, with a visible deviation from the traditional daily pattern). Although previous studies observed that taxi passengers are mainly frequent riders, city residents, and commuting between home and work locations (Schaller Consulting, 2006), our data set does not provide enough information to verify whether the customers using taxis during clear weather are the same ones using taxis during worse weather conditions.

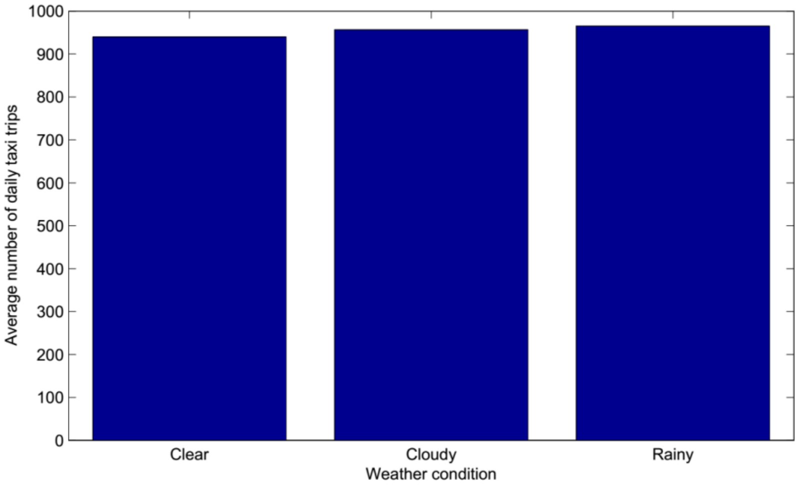


Figure 2-15 – Distribution of taxi trips given weather conditions.

2.5.2 Gravity map of taxi activity

In Figure 2-16 we can visualize how the pick-up and drop-off areas relate among them, where the thickness of the line represents the intensity between every two possible locations. Liang et al. (Liang, et al., 2011) define these connections as *taxi displacement* (line segments connecting the origin and destination). Strong relations can be observed in links *B-C*, *D-E*, *D-A*, *A-F*, and *F-B*. All those locations are characterized by some public transportation modality (airport, train, ferry, bus). *B* is the access to the airport, *C* and *D* are trains stations, *E* is a ferry dock, *A* and *F* are bus stops areas. It is important to stress that, although there is a subway in Lisbon, a direct subway line connection doesn't exist between the aforementioned locations[14].

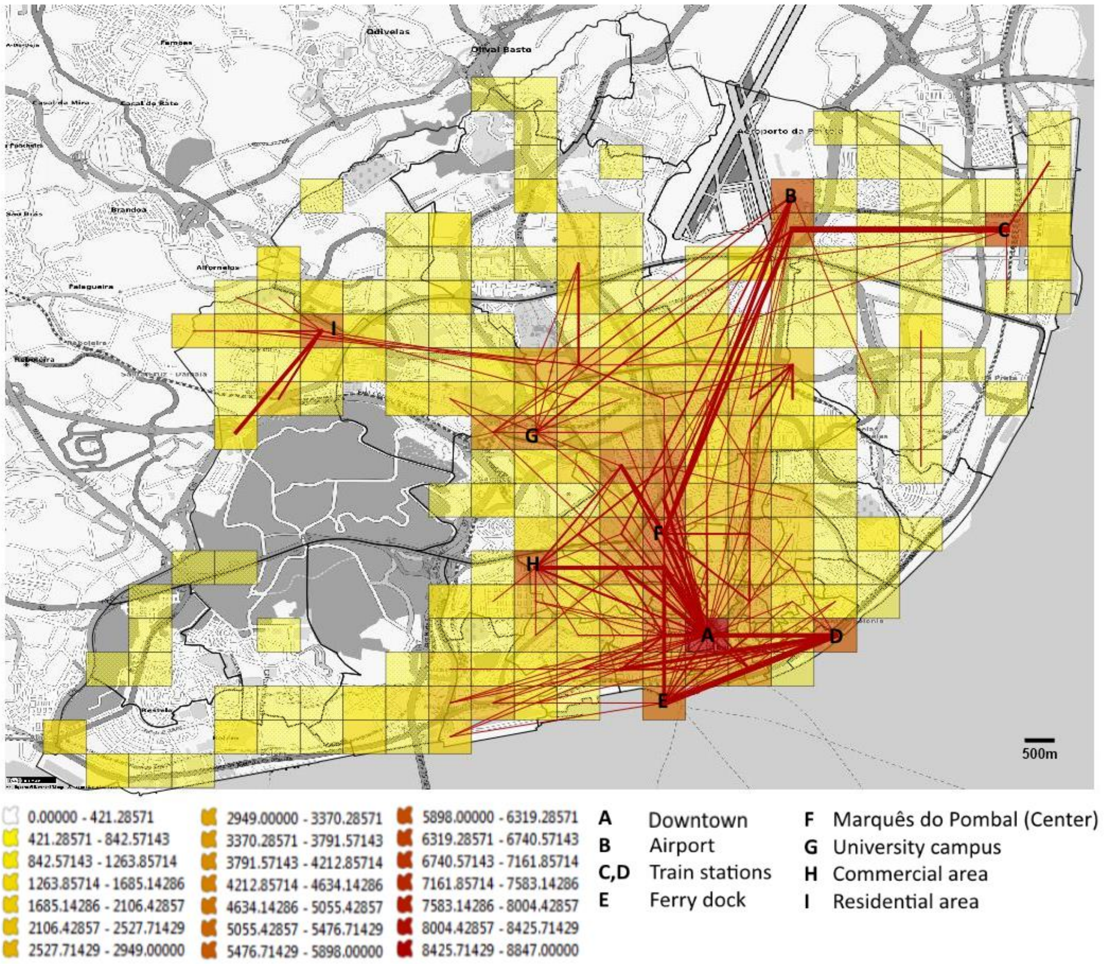


Figure 2-16 - How strongly connected locations are, according to taxi services.

From this observation, we hypothesize that the taxi service is often used as a bridge between public transportation modalities. It is also important to point out that the locations *A*, *C* and *F* (some of the most frequent pick-up or drop-off locations) give access to services and commercial areas.

In Figure 2-17 we can observe the relation between pick-ups and drop-off locations, considering only the most frequent destination for each location. By filtering the remaining destinations, we can visualize the predominant relations between locations, and their strength. The links *B-C* (airport and / station); *A-D* (downtown / train station), *D-E* (train station / ferry dock) and *A-F* (downtown / *Marquês do Pombal*,

[14] At the time of data collection there wasn't a subway line connecting the airport (*B*) and the *Oriente* train station (*C*). However, a new line is now available between these two transportation hubs, beginning 2012.

city center), now become visible. Once again, a bridge between transportation modalities is observable. These findings are in agreement with Zheng et al. (Zheng, et al., 2011b) who also observed a similar scenario of strong connectivity between transportation hubs.

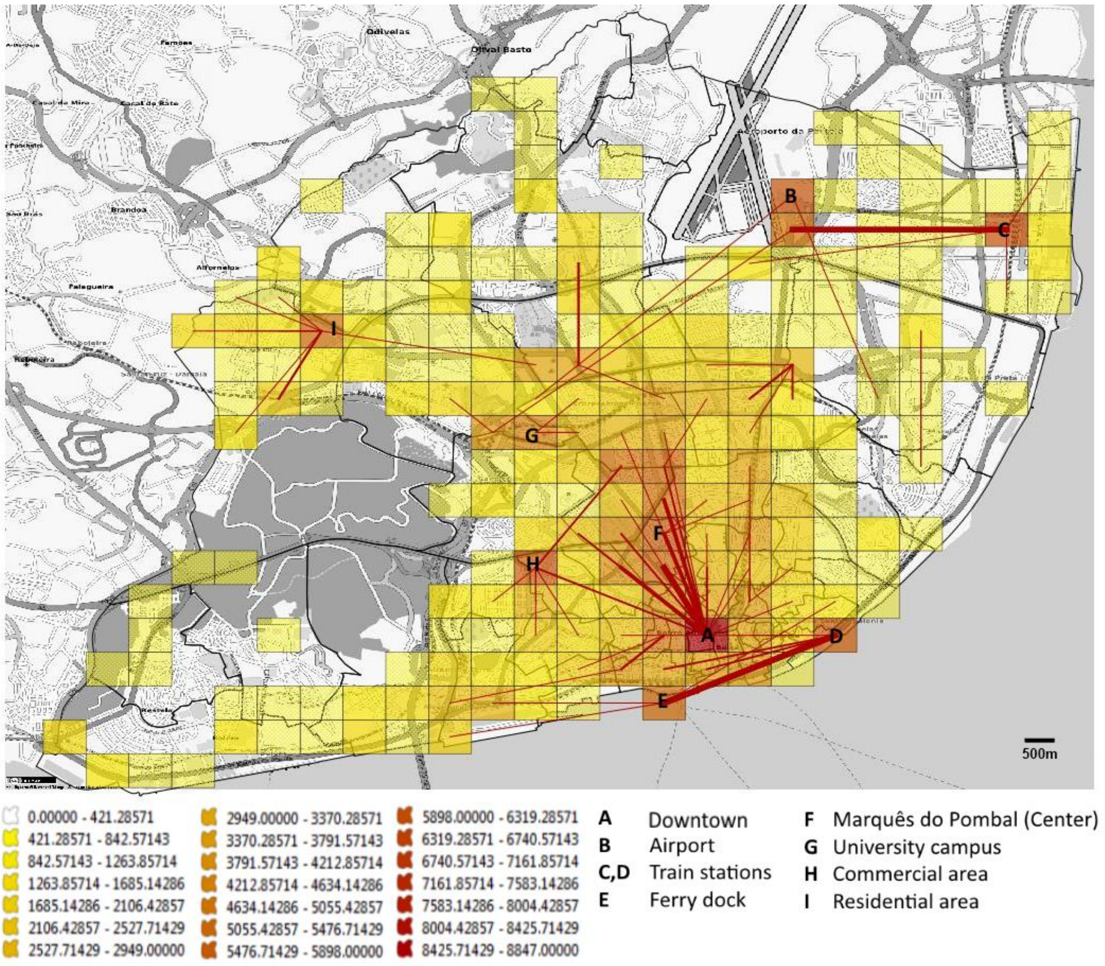


Figure 2-17 - Relation between pick-ups and drop-offs considering only the most frequent destination for each location.

2.5.3 Data distribution

To better understand the patterns from taxi services, we plot the taxi trips according to the distance, duration and income in Figure 2-18.

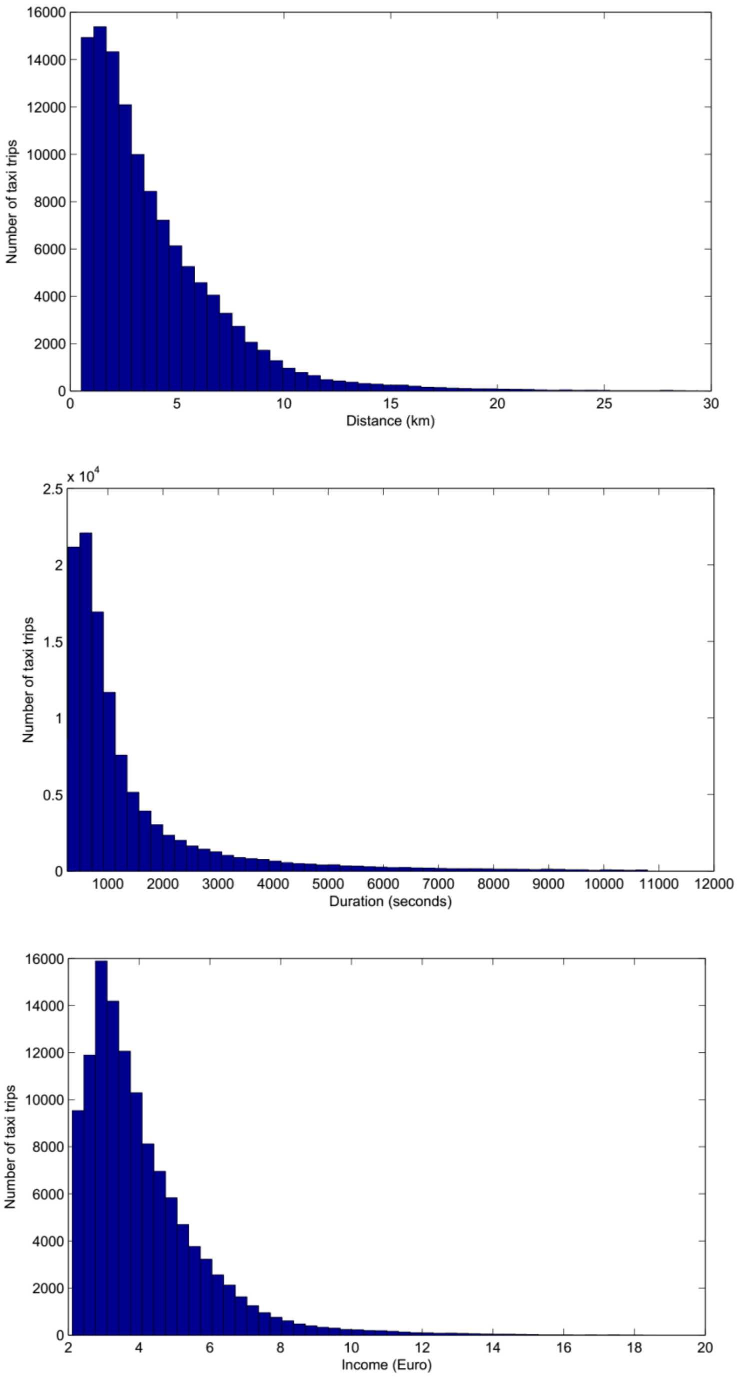


Figure 2-18 - Taxi volume distribution according to distance (top), duration (middle) and income (bottom).

After the data cleaning process, we examine the trip distance distribution and find that we can fit it with a gamma distribution with α = 2.7 and β = 1.2 as follows:

$$f_{\alpha,\beta}(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}} \tag{2}$$

where the parameters α (shape parameter) and β (scale parameter) satisfy α >0 and β >0, x >0, and $\Gamma(\alpha)$ represents the gamma function (Devore & Berk, 2012).

This observation does not agree with the results from different authors, where an exponential fit was observed using data collected in the Florence urban area, Italy (Bazzani, et al., 2010), or in Beijing, China (Liang, et al., 2011). However, Devore (Devore & Berk, 2012) demonstrated that exponential distribution is a special case of gamma distribution, when the shape parameter α is lower than 1. If the first interval of the histogram is removed (distance < 1000m), trip distance could be fitted with an exponential distribution *exp(λ)* (with λ = 0.26). In the same way, if the first interval of the trip duration is removed (duration < 5 minutes), then the trip duration can also be fitted with an exponential distribution. Trip income[15] follows a normal distribution (mean μ = 4.1166 and standard deviation σ = 1.7419) with a positive skew (2.1643). This observation is in agreement with the study of Liu et al. (Liu, et al., 2010a) using data collected in Shenzhen, South China, which also observed a normal distribution for trip income.

The difference in results compared with other authors can be due to following aspects: a) distinct data set (e.g. Liu et al. (Liu, et al., 2010a) worked with 3,000 distinct taxi drivers, whilst our data set contains only 217 distinct taxi drivers), b) to specific taxi drivers' behaviors (e.g. it was observed that a considerable amount of trips were from the airport to a nearby bus stop area, (roundtrip), being less than 1000m, a behavior that affected the overall distributions), and c) due possible noisy data.

2.5.4 Driver strategies

We further analyze taxi driver strategies with respect to income. A taxi driver may choose to pick up customers at a particular location (e.g. airport), drive around the city to find passengers at random places, or combine the two approaches. We find that only 8.37% of the taxis chose to stay at the same location for more than 50% of their waiting time. The airport seems to be one of the main pick-up and drop-off

[15] The income was calculated from data using the ANTRAL standard formulation http://www.antral.pt/simulador.asp. ANTRAL is a national association for transportation.

locations. Table 2-3 shows the statistics of the five best drivers according to the total income and number of trips, during the four months of study (September to December, 2009). We observe a low percentage of trips from the airport for these top drivers. Table 2-4 shows the five best drivers according to the income and number of trips from the airport. These results show that taxi drivers can improve their revenue by adopting a strategy of driving around the city instead of targeting one particular place like the airport, similar to what was observed by Liu et al. (Liu, et al., 2010b) and Li et al. (Li, et al., 2011a). Nevertheless, it is important to stress out that the reported numbers could not be entirely reliable or represent the real scenario with accuracy. Data was collected with devices being used for the first time, not fully tested, therefore prone to possible measure errors.

Driver ID	Total number of trips	Total income (€)	Number of active days	Average number of trips per day	Number of trips from airport	% of airport trips (from total number of trips)	Total income from airport (€)	% of airport income (from total income)
792	15,789	41,691.10 €	148	106	73	0.46 %	181.58 €	0.44 %
754	8,504	31,778.50 €	180	47	122	1.43 %	486.16 €	1.53 %
782	9,202	26,399.60 €	99	92	40	0.43 %	157.81 €	0.60 %
90	6,693	24,103.20 €	173	38	4	0.06 %	20.42 €	0.08 %
68	5,552	22,660.70 €	160	34	68	1.22 %	340.34 €	1.50 %

Table 2-3 - Statistics for top five taxi drivers, considering total income and number of trips, during four months.

Driver ID	Total number of trips	Total income (€)	Number of active days	Average number of trips per day	Number of trips from airport	% of airport trips (from total number of trips)	Total income from airport (€)	% of airport income (from total income)
37	3,003	6,585.83 €	140	21	2,343	78.02 %	5,066.81 €	76.94 %
134	4,377	18,511.60 €	151	28	813	18.57 %	3,873.69 €	20.93 %
538	1,947	7,711.46 €	114	17	817	41.96 %	3,094.13 €	40.12 %
193	2,829	10,835.30 €	159	17	348	12.30 %	1,676.02 €	15.47 %
2094	158	5,430.55 €	71	2	54	34.18 %	1,424.92 €	26.24 %

Table 2-4 - Statistics for top five taxi drivers, considering total income and number of trips from the airport, during four months.

We find that the majority of the taxi drivers appear to be using combined strategies – mainly driving around the city and in certain time periods staying at a fixed location. This phenomenon can be observed in the Figure 2-19, where there is an increase of trips from the airport between 6 AM and 8 AM and elsewhere during other time slots. On the other hand, in the same period there is a significant decrease of trips from the remaining locations. These observations are in line with Zhang et al. (Zhang, et al., 2014a).

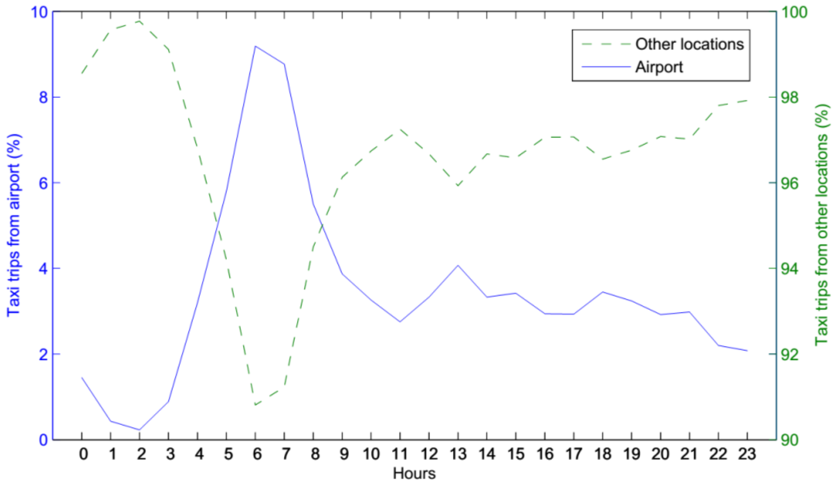


Figure 2-19 - Variation of number of trips aggregated by hour. Pick-ups from the airport (solid blue) and pick-ups from other locations in the city (dashed green).

To confirm our hypothesis, we examine mobile phone data collected from GSM networks. The GSM data consists of samples from December 2009 providing statistical measures of carried load (*erlang*) within one-hour period. In Figure 2-20 we can observe that immediately after the taxi rush hour there is an increase of the GSM network usage in the airport, while the mobile phone activities in other locations of the city begins to rise two hours later. This is an indication of possible taxi passengers at the airport.

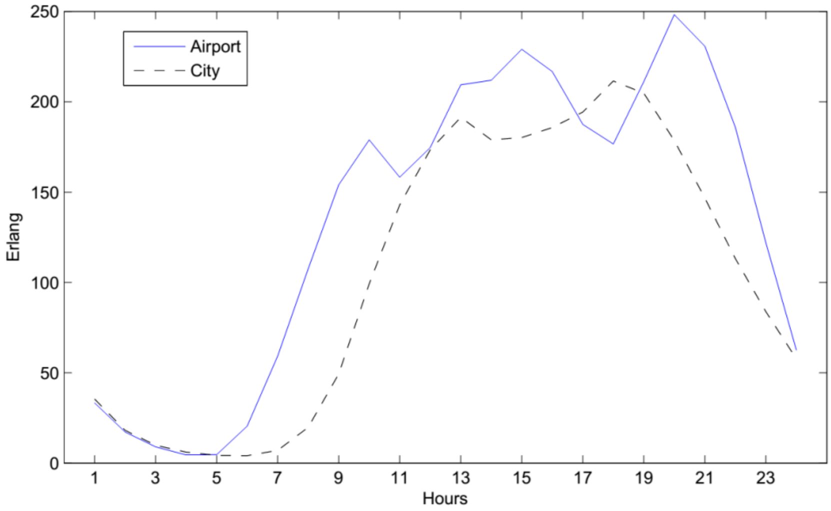


Figure 2-20 - GSM network usage aggregated by hour.

Another interesting observation in this preliminary study is that the main drop-off locations from the airport are (as expected) the downtown area and the main train station, but also the airport itself (Figure 2-21 right). The main pick-up locations with the airport being the final destination are also the *Oriente* train station (the main train station towards the north of the country); downtown; and also (surprisingly) the airport itself (Figure 2-21 left). By analyzing these airport trips individually, we observe that passengers take taxis to reach a nearby bus station and parking area.

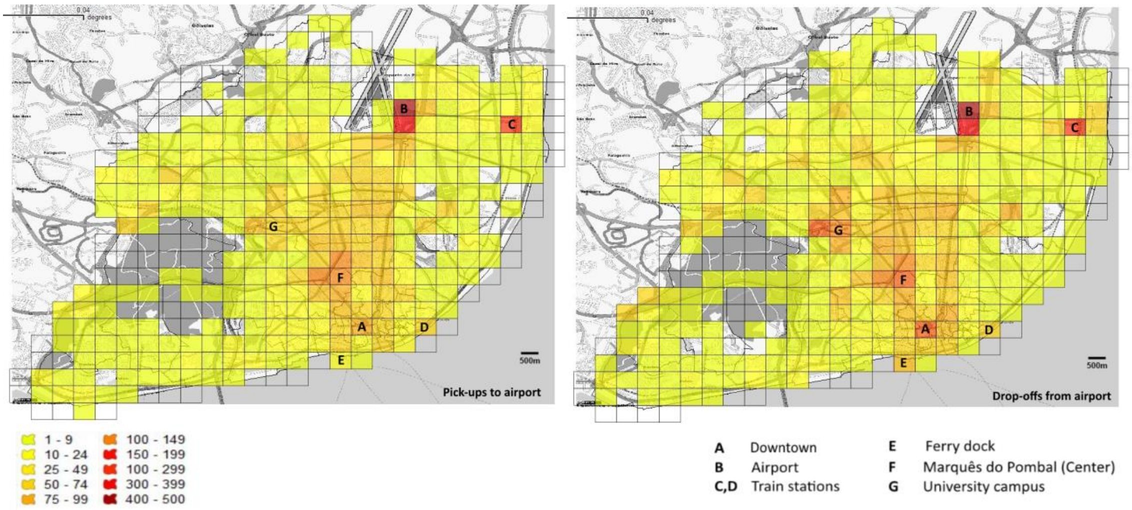


Figure 2-21 - Distribution of pick-up locations with destination to airport (left), and drop-off locations from the airport (right).

As aforementioned, overall revenue of taxi service seems to follow a normal distribution (mean μ = 4.1166 and standard deviation σ = 1.7419) with a positive skew (2.1643). We also observed that top 30% drivers perform 76% of the taxi trips undertaken by the entire taxi fleet. This is a clear indication of the top drivers' efficiency, who seem to be able to correctly choose hotspots to pick-up new passengers and decide on the shortest-time route to destination. Similar observations were made by Liu et al. (Liu, et al., 2010a), Yuan et al. (Yuan, et al., 2011b), (Yuan, et al., 2012a), and Zhuang et al. (Zhuang, et al., 2012), which have extracted knowledge from the behavior of top drivers to build recommendation systems.

2.5.5 Characterization of downtime: time spent searching for next pick-up

The previous analysis focused on the taxi service, in other words, the relation between the pick-up and the corresponding drop-off. It is also interesting to understand what happens in between services (i.e. downtime – time spent looking for next pick-up), since it can help improve taxi drivers' income.

Figure 2-22 presents the areas with high (red) and low (yellow) average distance traveled when taxis search for new pick-ups, and the relationship (link) between the previous drop-off locations and the following pick-up locations (line thickness represents strength).

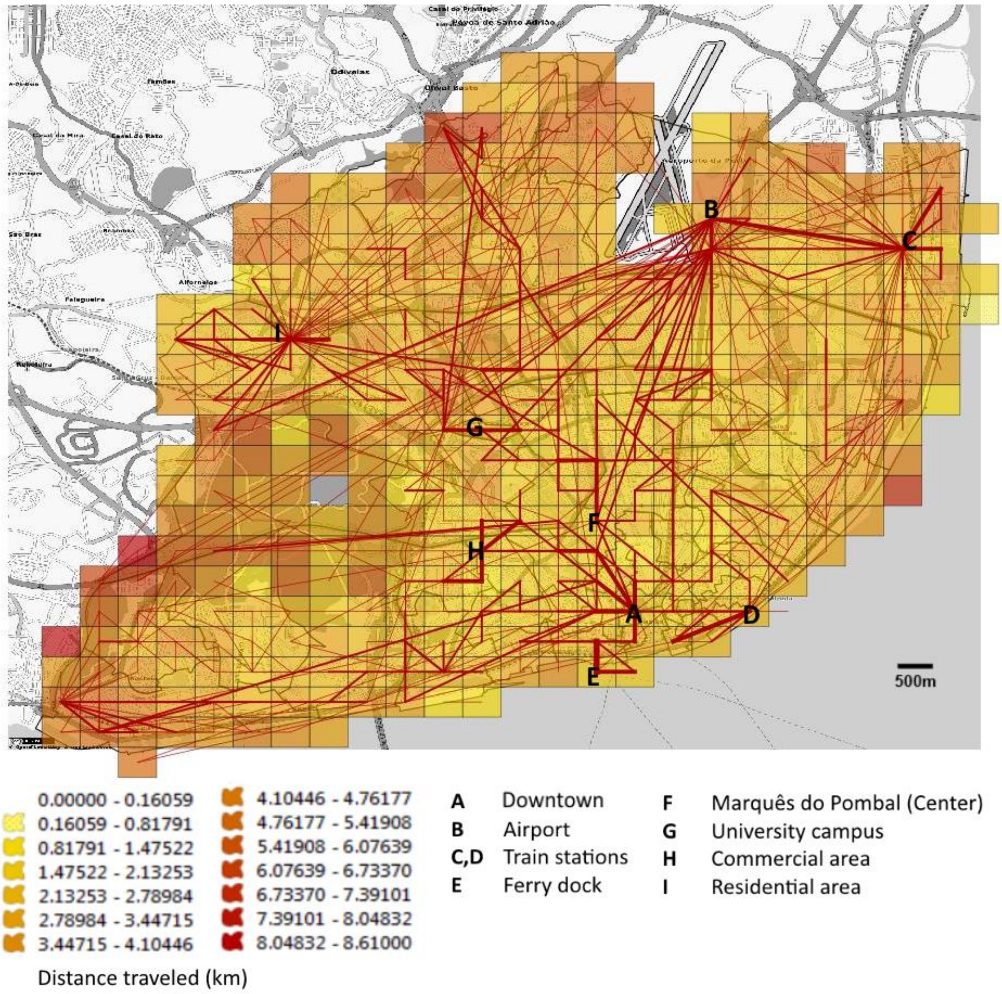


Figure 2-22 - Spatial distribution of average distance traveled during downtime and the relationship between previous drop-off and next pick-up.

The outer city areas (characterized by a higher number of residential buildings) show higher average distances traveled between services (cells in red), whereas in the inner city areas the distances traveled are relatively small (cells in yellow). In other words, a taxi driver after a drop-off in the city suburbs, on average has to make a longer trip to the next pick-up than a taxi driver dropping-off a passenger in the city center.

In the same way, strong relationships between adjacent locations are observed in the city's inner areas (the next pick-up takes place in the vicinity of the last drop-off), while in the outer city areas a strong link is observed between distant locations (the next pick-up takes place in a distant area). It appears to us that after a drop-off in the city's outer areas, a taxi driver typically heads to locations with higher probability of picking up new passengers (e.g. airport, city center, train station) even if it means traveling a farther distance to the next pick-up location.

In Figure 2-23 we can see a density grid, where next pick-up location takes place in the same location as the previous drop-off. Red means higher number of pick-ups on the same area of the previous drop-offs, while yellow represents areas with a low amount of the pick-ups on the same area of the previous drop-off. Downtown (*A*), *Marquês do Pombal* or city center (*F*), airport (*B*) and train stations (*C* and *D*) are the locations with higher probability for a taxi driver to pick-up a new passenger in the same area, after a drop-off.

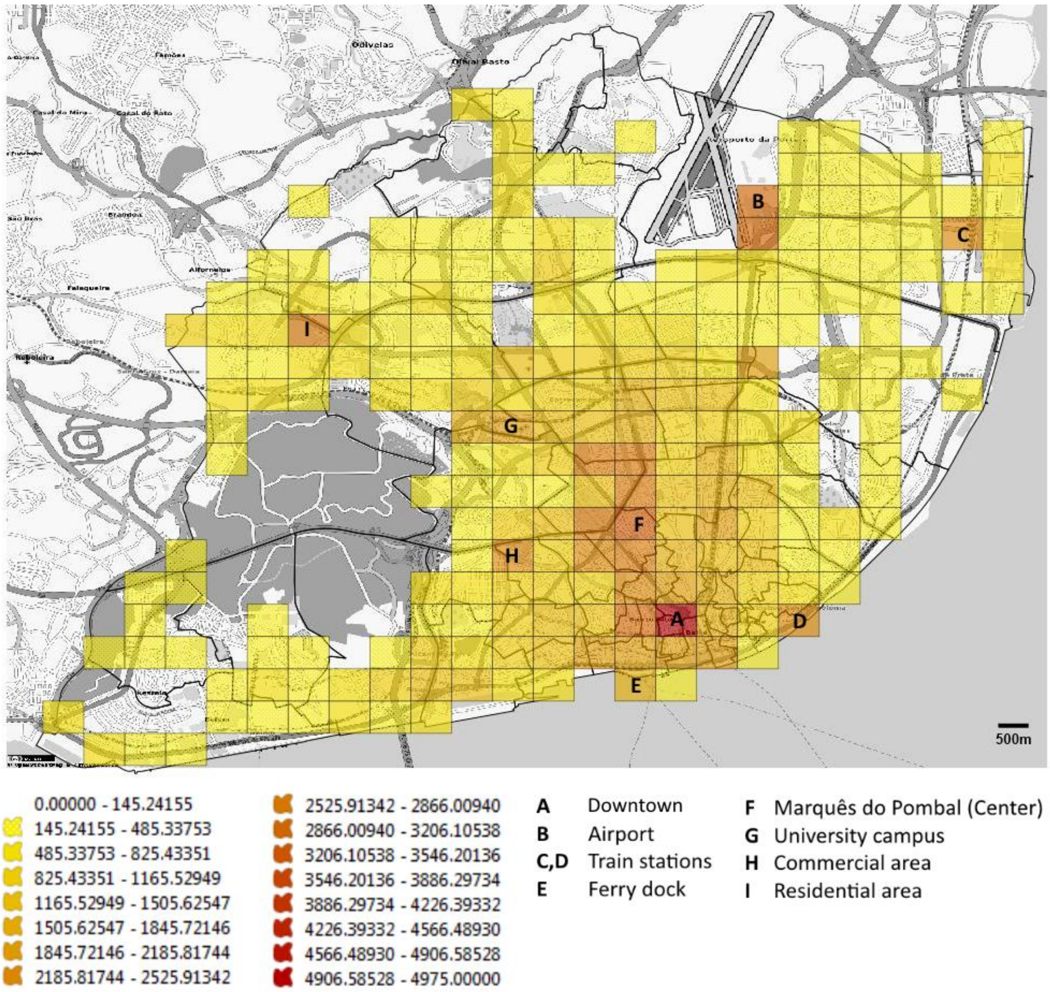


Figure 2-23 – Amount of pick-ups taking place in the same location as the previous drop-off.

From these preliminary results we can estipulate that taxi drivers may want to improve their income by targeting the above-mentioned locations (transportation hubs), or, at least, move to those locations after the latest drop-off, since it can improve the probability of picking-up a new customer in a reasonable amount of time and without the need to travel great distances.

Figure 2-24 (top) shows the hourly taxi service (percentage of taxi trips, in blue) and the percentage of taxis in service throughout the day (green), whereas Figure 2-24 (bottom) shows the average time spent (blue) and distance travelled during downtime (green).

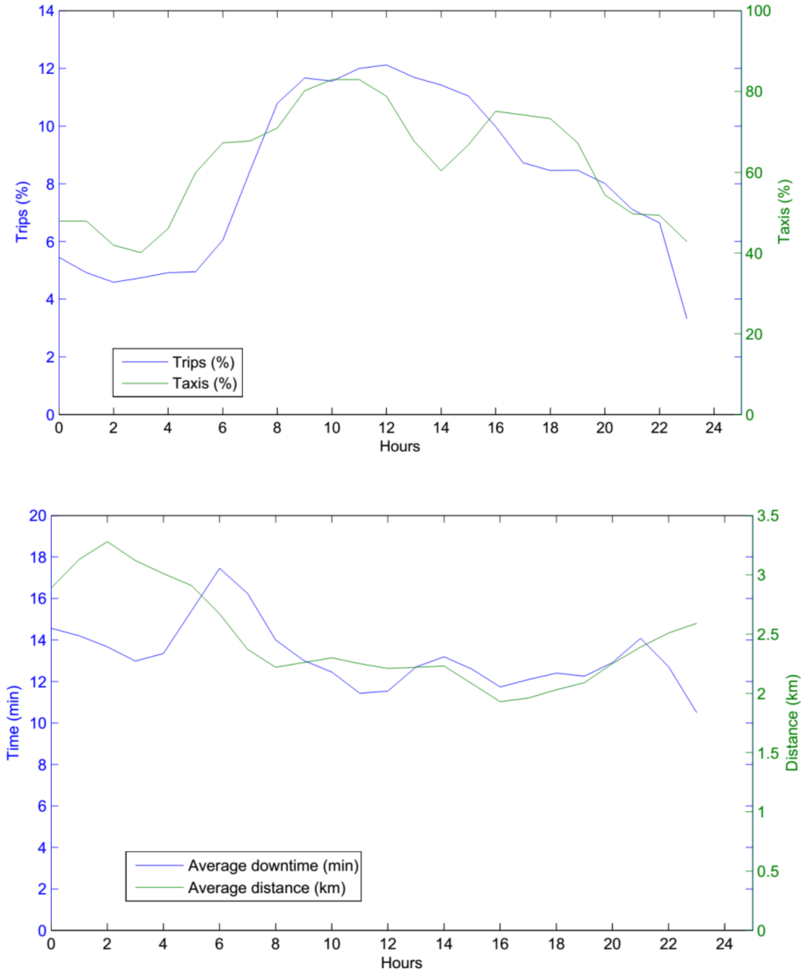


Figure 2-24 - Top: Amount of trips (blue) and number of taxis in service (green) throughout the day; Bottom: Average downtime (blue) and distance traveled (green).

Due to the low amount of taxis in service in the early hours in the morning (12 AM to 7 AM), the average downtime and distance traveled searching for new passengers are relatively high. The average downtime remains almost constant from

10 AM to 10 PM. There is a sudden drop in downtime at 10 PM but a rise of distance traveled. The lower number of taxis in service as well as potential passengers during this late hour presumably causes longer time spent searching for pick-ups.

Both distance traveled and downtime (searching for the next pick-up), appear to follow an exponential distributions as argued by Bazzani (Bazzani, et al., 2010) (Figure 2-25).

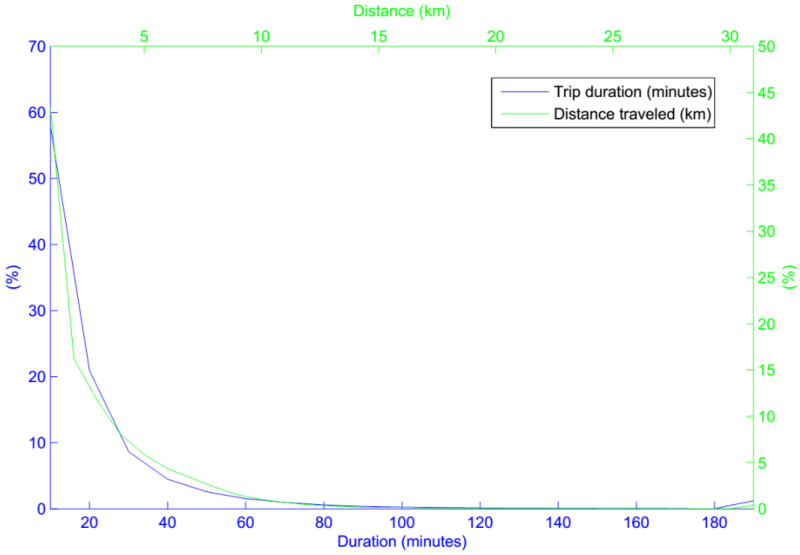


Figure 2-25 - Distribution of distance travelled (blue) and time spent (green) during downtime.

In Figure 2-26, we can observe the relationship between the distance traveled during downtime and the resulting service distance with corresponding average income. From this figure, we can conclude that a farther distance traveled during downtime does not guarantee a more profitable service. For instance, if taxi driver *A* travels less than 1km to the next pick-up, and taxi driver *B* travels around 8km to the next pick-up, they both end up providing similar taxi services with equal revenues (at about 4.50€), although taxi driver *B* drove a considerably longer distance to pick-up his/her customer.

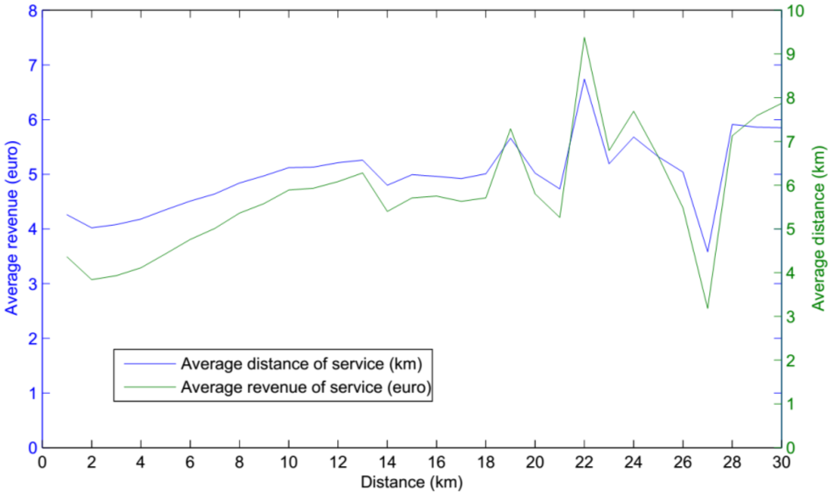


Figure 2-26 - Resulting service distance (blue) and income (red) as function of the distance traveled for the next pick-up.

We can argue that in order to improve profit, it is preferable for a taxi driver to wait for passengers in locations related with main public transportation hubs (airport, train stations, ferry dock or main bus stops), and to avoid traveling great distances to the next pick-up location, unless it is to return to the aforementioned locations. If the drop-off location coincides with a public transportation hub, it is preferable to wait for new passengers at that location. This is the behavior pattern observed by top drivers, as described in the previous sections. These findings are in line with Li et al. (Li, et al., 2011a). Although Zhang et al. (Zhang, et al., 2014a) observed that hunting is usually more efficient than waiting in order to find passengers, the authors also advise that there are exceptions, particularly in locations associated with main public transportation hubs (e.g. airport), where it would be preferable to wait during specific time slots. The authors also suggest that in the suburbs, the best strategy would be to consider returning to hotspots of taxi demand. Therefore, Zhang et al. corroborate our findings concerning this subject.

2.5.6 Impact of POIs on taxi service

By embedding spatial profiles, like Points Of Interest (POI), onto the map, we can further observe taxi dynamics according to the area characteristics. In section 2.4.3, Figure 2-11 shows the map with POIs that are grouped into eight different categories, and Figure 2-9 shows the distribution of these categories. One can observe that Education facilities (e.g. kindergarten, university, etc.), Recreation (bar, restaurant, etc.) and Services (e.g. bank, etc.) are the dominant POI categories (which account for over 70%).

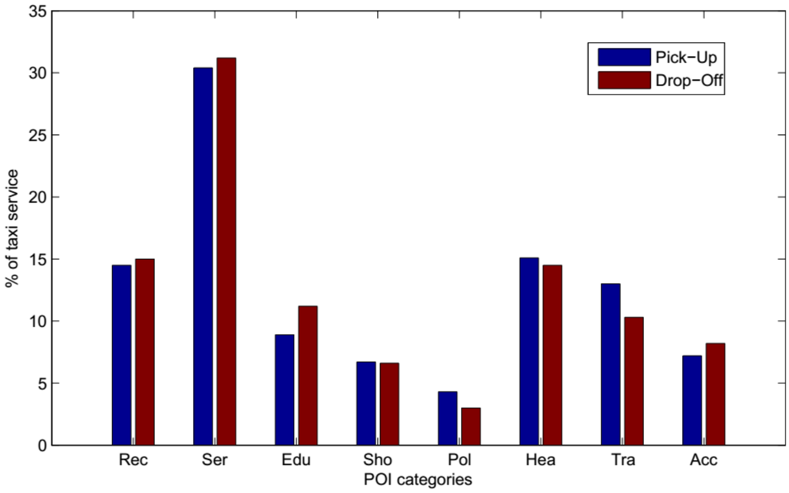


Figure 2-27 – POI categories distribution according to taxi service.

Figure 2-27 shows the POI categories distribution considering the amount of taxi pick-ups and taxi drop-offs. To each taxi pick-up/drop-off was assigned the nearest POI. Services are clearly the most predominant origin and destination for taxi services, with Recreation and Health following right behind. Weather conditions do not seem to affect this distribution. However, this distribution changes throughout the day. Taxi service is more active for the majority of POI categories during working hours (e.g. schools, museums, and shopping centers are mostly opened during working hours). Nevertheless, for Recreation areas (a category which includes bars and restaurants) the scenario is the opposite, where taxi service is more active between 7 PM and 5 AM. Figure 2-28 demonstrates this scenario. From 6 AM to noon, taxi pick-ups to Service locations increase, while they decrease to Recreation areas. From 7 PM onwards, while taxi pick-ups from Service areas decrease, taxi pick-ups from Recreation areas increase. The scenario is similar for taxi drop-offs. Moreover, taxi services for Recreation and Shopping locations increase on weekends, while taxi services decrease for Services, Health and Education locations during the same period.

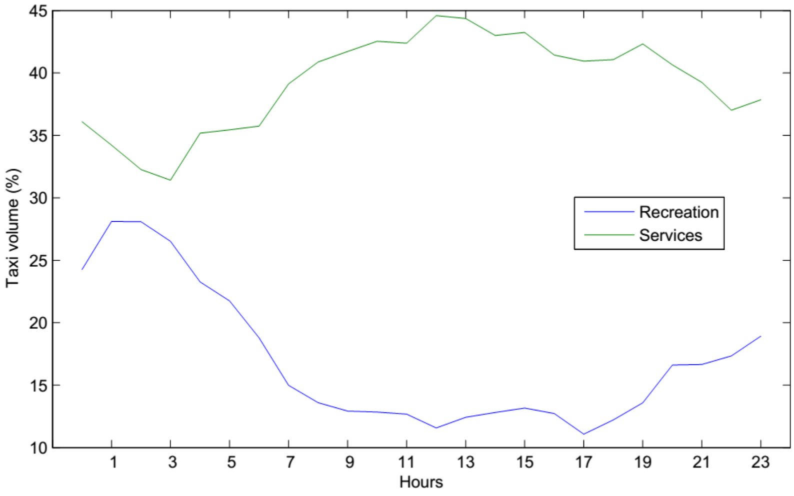


Figure 2-28 – Hourly variation of taxi service for recreation and service locations.

The POI distribution allows us to further explore the taxi origin-destination patterns. Figure 2-29 shows taxi origin-destination distribution according to the predominant POI in the area. Services and Recreation are the most frequent drop-off areas – independently of the pick-up location. Transportation is the most likely drop-off area if the pick-up is located in Shopping, Transportation, Health, or Education areas. Drop-offs in Education areas are mainly connected to pick-ups from Health areas. These observations are an indication that the area type can be a possible predictive attribute to consider when looking for taxi demand.

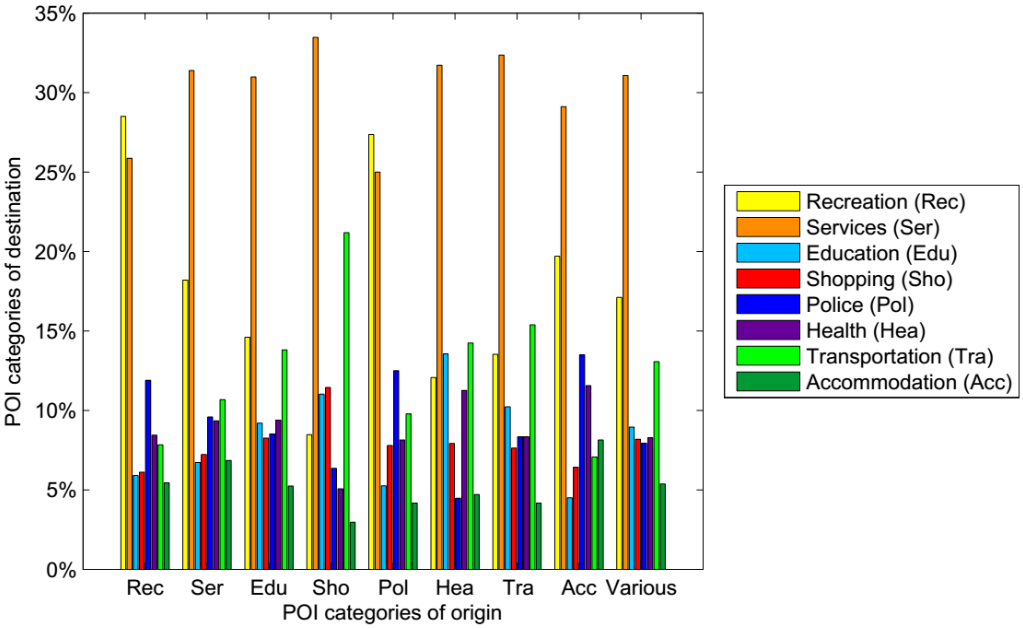


Figure 2-29 – Taxi origin-destination distribution according to the predominant POI in the area.

2.6 Data analysis and results: discovering the next pick-up location

Taxis are a flexible way of transportation, since they are not bound to pre-defined path or pick-ups and drop-offs locations. Therefore, taxis dynamically adapt to the flow and the needs of passengers. This flexibility can lead to a large search space and the prediction of taxi movements can be challenging. However, day of the week, time of the day, and weather conditions are promising features in predicting taxi volume and our exploratory study shows the possibility that some movement patterns (e.g. temporal and spatial density of pick-ups and drop-offs, the relation between pick-ups and drop-offs) can be predicted by these variables, as proposed by Chang et al. (Chang, et al., 2010) and Yuan et al. (Yuan, et al., 2011a).

In this section, we aim to explore to what extent the next taxi pick-up location can be predicted, given the current drop-off. This work aims to make possible the development of a recommendation system, which can help taxi drivers find the next pick-up location. Based on a Naïve Bayesian Classifier, the system should provide the likelihood of findings passengers along the urban space. Because the city is modeled by a flexible grid system, the user can interact with the system and obtain a personalized visualization. We have observed that area type, characterized by the predominant POI, can potentially be used along with other aforementioned features explored in previous work, following the findings of Qi et al. (Qi, et al., 2011) and Pan et al. (Pan, et al., 2013), which identified that different areas in the city have distinct functions and can affect taxi demand. Here we apply a simple probabilistic approach. The focus of the study will be the inference analysis and not the development of a full framework.

2.6.1 Overview

A system overview is shown in Figure 2-30 (left). In order to make a recommendation, the system extracts data from a database of taxi-GPS traces, and pre-processes it to select relevant features that match the current scenario, namely location, day of the week, hour, weather conditions or area type (characterized by the predominant POI). The selected data is processed by a classifier and the output presented to the user (the taxi driver).

The graphical interface is formed by three layers that provide enriched information and allows the user to filter or select specific views (Figure 2-30, right). The first layer represents the map of the urban area. The second layer provides the likelihood of each possible pick-up location according to the available data from the taxi community. The third layer provides similar information, but using only the historical data from the current driver. This configuration should be flexible enough to allow the user to select the desirable features that should be taken into account by the inference engine and provide controls so that the user can zoom in and zoom out to explore the map in detail, focusing on specific areas.

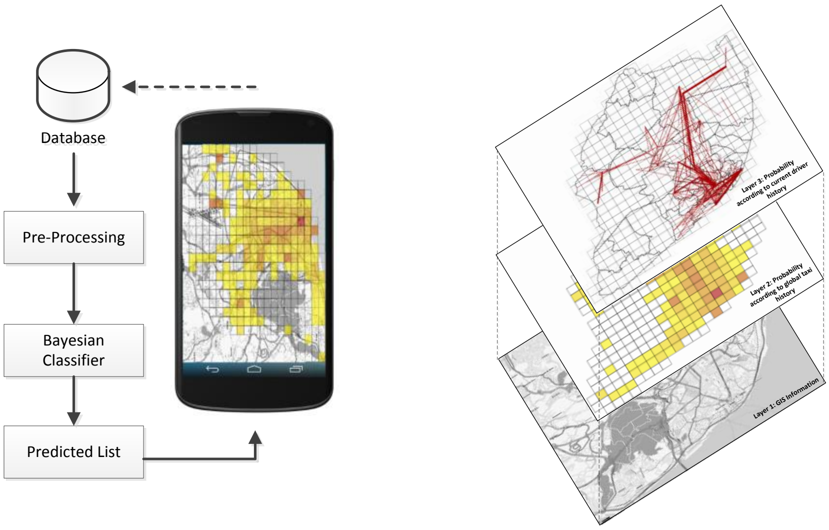

Database
Pre-Processing
Bayesian Classifier
Predicted List
Layer 3: Probability according to current driver history
Layer 2: Probability according to global taxi history
Layer 1: GIS Information

Figure 2-30 - Basic system overview (left) and system's visualization layers (right).

The data used for the inference engine can contain samples from the taxi community or solely past information from the current driver. In the same way, the system could provide a personalized recommendation, considering only the current location (samples which the previous drop-off matches the current location), or global information - taking into account all history available (samples which the previous drop-off took place at any location). Additionally, besides the current location, the system can also process information from neighbors' cells to improve the accuracy of the system.

The likelihood coefficient is presented within a grid and a scheme of colors (e.g. red, more likely to find a passenger; yellow, less likely). The size of the cells can be changed to meet user demand. However, as it is demonstrated in the next section, the use of smaller cells will reduce the performance of the inference engine. After each drop-off, the new taxi-GPS trace is pre-processed and transformed to be stored in the database.

The system is thought to be used individually and independently by each taxi driver. However, a centralized solution could be explored to take into consideration traffic conditions and the possible competition between different taxi drivers for the same potential pick-up. Knowing the current destination of each taxi driver would allow the system to avoid suggesting pick-ups already assigned.

2.6.2 Inference Engine

We apply a naïve Bayesian classifier, which is a simple probabilistic classifier based on Bayes' theorem with independence assumptions. Bayes rule of conditional probability (MacKay, 2003) is defined by:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{3}$$

where *P(A|B)* is the posterior probability, which is the probability *A* given the feature *B*, *P(B|A)* is the likelihood of *A* with respect to *B*, *P(A)* is called prior probability and *P(B)* the evidence factor.

The objective is to compute the likelihood of each possible pick-up area (*Y*) given:

- hour of the day (*T* = {1, 2, …, 24});
- day of the week (*D* = {Sunday, …, Saturday});
- weather condition (*W* = {Clear, Cloudy, Rainy});
- area type, defined by the predominant POI (*I* = {Services, Recreation, Education, Shopping, Police, Health, Transportation, Accommodation}); and
- current location (grid cell, *C* = {1, 2, …, 396}) of the last drop-off.

The conditional probability can be formulated as follows:

$$P(Y = y_i|T, D, W, I, C) = \frac{P(Y = y_i)P(T, D, W, I, C|Y = y_i)}{P(T, D, W, I, C)} \tag{4}$$

The prediction is based on the *maximum a posteriori probability* (MAP) decision rule:

$$\begin{aligned} y_{MAP} &= \arg\max_{y_i \in Y} P(Y = y_i | T, D, W, I, C) \\ &= \arg\max_{y_i \in Y} P(Y = y_i) P(T, D, W, I, C | Y = y_i) \\ &= \arg\max_{y_i \in Y} P(Y = y_i) \\ &\prod_i P(T|Y = y_i) P(D|Y = y_i)\, P(W|Y = y_i) P(I|Y = y_i) P(C|Y = y_i) \end{aligned} \tag{5}$$

2.6.3 Global accuracy and model improvements

Based on a holdout validation (oldest 2/3 of the samples for training the model and the most recent 1/3 of the samples for testing) the classifier is able to predict the next pick-up location with an accuracy of about 56.3%. A prediction is considered correct if the classifier suggests as the next pick-up location either the corresponding cell in the testing set or the bordering neighbor cells. Due to the proximity of neighbor cells, any prediction falling in the vicinity should also be considered correct, since it is reasonably close to the correct location (less than 500m). Moreover, due to the process of grid creation that models the city, and the *User Equivalent Range Error* (UERE) of GPS (El-Rabbany, 2006), samples with pick-ups in the same location could be split between two neighbor cells, forming a local cluster spread between cells.

To improve the model, cells with low taxi activity can be removed from the process (as described in Figure 2-5, section 2.4.2). Due to the low activity, taxi patterns can be random in those areas, which negatively affect the estimation. This scenario can be observed in Figure 2-31, where the percentage of prediction error is represented for each area through a color scale (yellow means lower prediction error, red means higher prediction error). Areas with low taxi activity have a higher prediction error (marked as red) when estimating the next pick-up location, i.e. it is more difficult to correctly estimate these cells as the next pick-up location. These cells have, on average, less than 10 pick-ups during the entire period under study (four months, from September to December), and this low amount of available samples is not sufficient to correctly train a model.

The model can be further improved by processing the data set using different layouts: weekdays or working hours. The highest accuracy (56.3%) is achieved with a combination of factors: weekdays (excluding weekends); working hours (excluding late-night hours, from 22 PM to 7 AM); and excluding cells with low taxi activity.

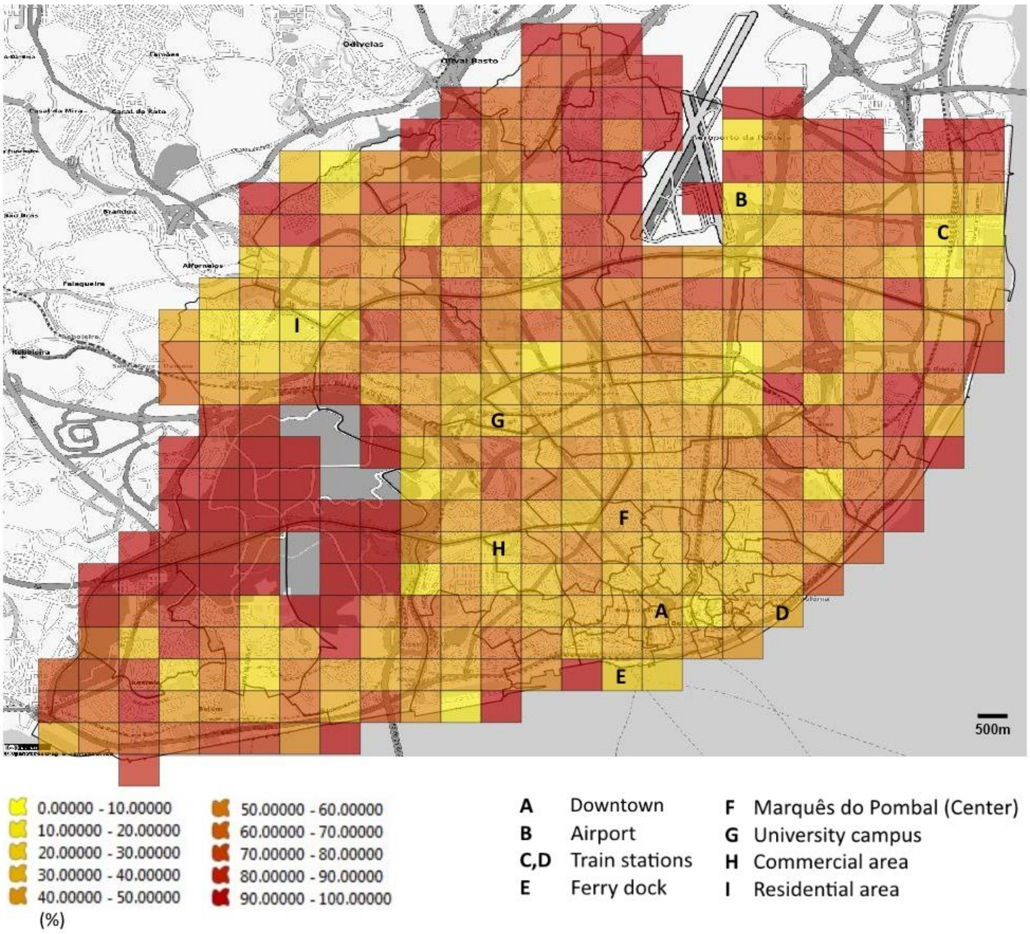


Figure 2-31 – Percentage of prediction error for next pick-up.

2.6.4 Inspecting the model accuracy

The highest accuracy of the model (56.3%) is achieved considering a set of conditions: on weekdays; working hours; in areas with reasonable taxi activity (cells with low taxi activity were disregarded). More importantly, the evaluation of the model considers that neighbor cells are also a positive prediction. This approach is due to insufficient data in some temporal (nighttime) and spatial slots (areas with low taxi activity). Nevertheless, we are interested to analyze the behavior of the model under a more rigorous evaluation, stripping down the aforementioned conditions.

In this second scenario, a prediction is considered correct only if the classifier suggests as the next pick-up location the exact cell of the testing set. Bordering neighbor cells are not considered in this scenario as a correct prediction. All cells and time slots are considered (even with low taxi activity). Therefore, we are increasing the search space and including cells with low data.

Based on a holdout validation (oldest 2/3 of the samples for training the model and the most recent 1/3 of the samples for testing), the classifier is able to predict the next pick-up location with an accuracy at about 31.1%, an expected lower performance since the search space was increased, and includes areas without sufficient data (low taxi activity cells). This is, in reality, a reasonable outcome once we consider that the search space is composed by 370[16] possible locations, where each cell, on average, has a probability of only 0.27% to randomly receive the next taxi pick-up. The *a priori* probability for the best cells (those with very high taxi activity) is 2.4%. Therefore, the ability of our approach to correctly predict the next location is 13 folds greater than randomly selecting one of the cells with very high activity.

This initial accuracy of 31.1% in the second scenario corresponds to the classifier without applying any special conditions. As presented before, the classifier can be improved by applying a set of restrictions to data. Filtering out cells with low taxi activity will improve the model to 36.5%. The restriction to consider only weekdays and working hours additionally slightly boost the accuracy to 38.6%. Moreover, considering the neighbor cells as correct predictions improves the model accuracy to 56.3%. This setup corresponds to the model observed in the first scenario.

The next sections will continue to explore the effect of different variables on the classifier, namely, the contribution of each variable, daily and weekly patterns, cell size, and taxi drivers' strategies. Specifically, the classifier will be explored without special conditions, considering all grid cells (with high and low taxi activity), and all daily and weekly periods. This setup corresponds to the second scenario described in this section.

[16] From the 396 cells that compose the grid which models the city, 26 cells have no taxi activity (because they are located in areas where traffic is not allowed, such as the *Centro Desportivo Nacional do Jamor* or the airport runways), or have very low taxi activity (less than 10 trips during all periods under study).

2.6.5 Impact of extracted features

To understand the effect of the features hour of the day, day of the week, weather condition and area type, the Information Gain (IG) of each variable is computed: 2.14543 for current area, 1.03384 for the predominant POI of current area, 0.17016 hour of the day, 0.03558 for day of the week and 0.00903 for weather conditions. The results show that the current location (the area of the previous drop-off) is the most important factor in determining the next pick-up location. Hour of the day and the predominant POI of the area are also relevant factors. However, weather conditions seem to be the least relevant factors in the process, confirming the observation in Figure 2-15 (section 2.5.1).

Information Gain (also Kullback–Leibler divergence of a conditional probability distribution), from the Information Theory, measures the differences in entropy (reduction of uncertainty) after splitting a data set *T* on attribute *A* (MacKay, 2003), and it can be defined by:

$$IG(T, A) = H(T) - H(T|A) \tag{6}$$

which can be expanded to:

$$IG(T, A) = H(T) - \sum_{t \in T} p(t)H(t) \tag{7}$$

where $H(T)$ is the entropy of the (training) set *T*, $p(t)$ the proportion of the number of elements in *t* to the number of elements in set *T*, and $H(t)$ the entropy of subset *t*. The entropy (represented by *H*) measures the randomness or uncertainty associated with a random variable, and has been studied and defined by Claude Shannon (Shannon, 1948) as follows:

$$H(X) = -\sum_{i} p(x_i) \log_2 p(x_i) \tag{8}$$

where $H(X)$ is an entropy of random variable *X*, *x_i* ∈ *X*, and *p(x_i) = Prob(X= x_i)*. When $H(X) = 0$ the set *X* is perfectly classified.

Since the target variable is nominal, we can also compute the Cohen's Kappa statistic, which is a measure of agreement, with value 0.3037. According to Fleiss (Fleiss, 1981), values between 0.21 and 0.40 are characterized as a fair agreement for the model. The Kappa statistic (*k*), defined by Jacob Cohen (Cohen, 1960), is expressed as:

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{9}$$

where $\Pr(a)$ is the observed proportional agreement between *X* and *Y* is defined as:

$$\Pr(a) = \frac{1}{n}\sum_{i=1}^{n} f_{ii} \tag{10}$$

where f_{ii} represents the frequency of the number of subjects with the i^{th} categorical response for variable *X* and the j^{th} categorical response for variable *Y*. $\Pr(e)$ is the expected agreement by chance, defined as:

$$\Pr(e) = \frac{1}{n^2}\sum_{i=1}^{n} f_{i+}\, f_{+i} \tag{11}$$

where f_{i+} is the total for the i^{th} row and f_{+i} is the total for the i^{th} column of a square contingency table that displays the frequency distribution of the variables.

2.6.6 Effect of daily and weekly periods

During the exploratory spatiotemporal analysis, we observed temporal patterns, where active hours of the day (8 AM - 8 PM) and active days of the week (weekdays) present a slightly higher predictability. Figure 2-32 represents this behavior through a predicted list (Phithakkitnukoon & Dantu, 2008) – the list of the most likely destinations, where the top of the list contains more likely destinations than the ones lower on the list.

This behavior can be explained by the existence of more activities (mostly repeated activities in temporal orders such as commuting to work, lunch time at similar place, school activities, and so on) on weekdays and active hours than on weekend. A similar observation is performed in the next chapter, where mobile phone call intensity has a strong correlation with taxi volume on weekdays and active hours.

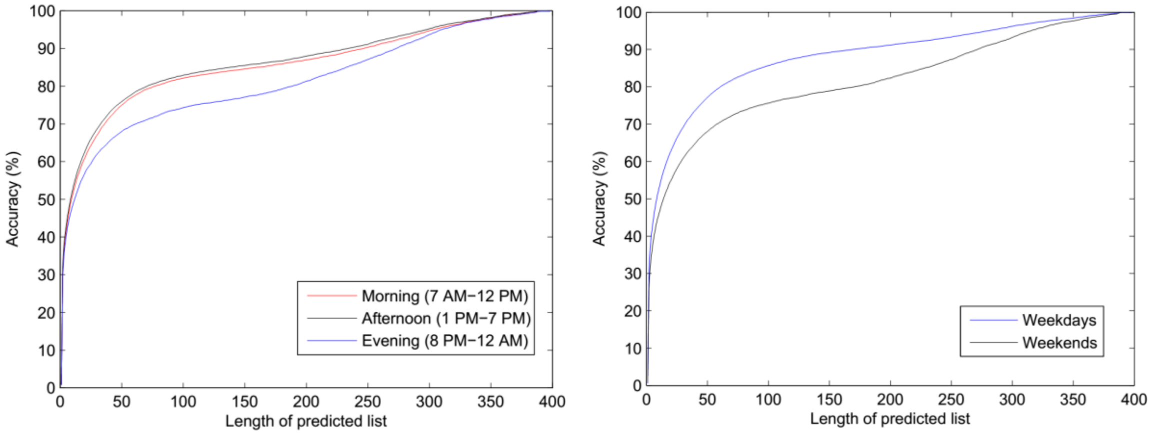


Figure 2-32 - Overall performance of the predicted list for different daily periods (left) and between weekdays and weekends (right).

2.6.7 Effect of taxi drivers' strategies

Taxi drivers have different strategies to identify the best location for the next pick-up or a fastest path to drop-off. Liu et al. (Liu, et al., 2010a) and Zheng et al. (Zheng, et al., 2010), (Zheng, et al., 2011b) explored this behavior to model taxi drivers' knowledge. Liu et al. (Liu, et al., 2010a) classified drivers as top drivers and ordinary drivers according to their income. In our work we also explore the difference between top performance drivers, average performance drivers and low performance drivers, considering their income and amount of trips. Although top drivers present specific strategies, they are also characterized by a high amount of trips achieved.

In order to improve income, top drivers search for passengers in specific locations related with transportations hubs, at specific hours of the day (e.g. searching for passengers near the airport between 7 AM and 9 AM, when the majority of international flights from the West arrive at Lisbon Airport), and avoid traveling long distances for their next pick-up. Therefore, they move from one location to another according to the hour of the day, and stay (or return to that location after a drop-off) for a period of time.

On the other hand, two main patterns are observed amongst low performance drivers: either staying (and returning after a drop-off) to the same location for long periods of time, or randomly picking-up passengers throughout the city. However, the second scenario could be generated by the taxi central dispatching service, redirecting taxis to specific locations, according to customers' requests via phone calls, and not via a specific hunting strategy.

Figure 2-33 shows the overall performance of the predicted list for top, average and low performance drivers. The former rapidly increases the accuracy of prediction, needing to grow to 37 possible destinations in order to predict the correct destination to a high accuracy (70%). Average drivers must grow to 177 and low performance drivers to about 250 to attain similar accuracy. Since average and low performance drivers do not appear to have specific driving strategies to improve their income, their behavior can be characterized by a certain randomness.

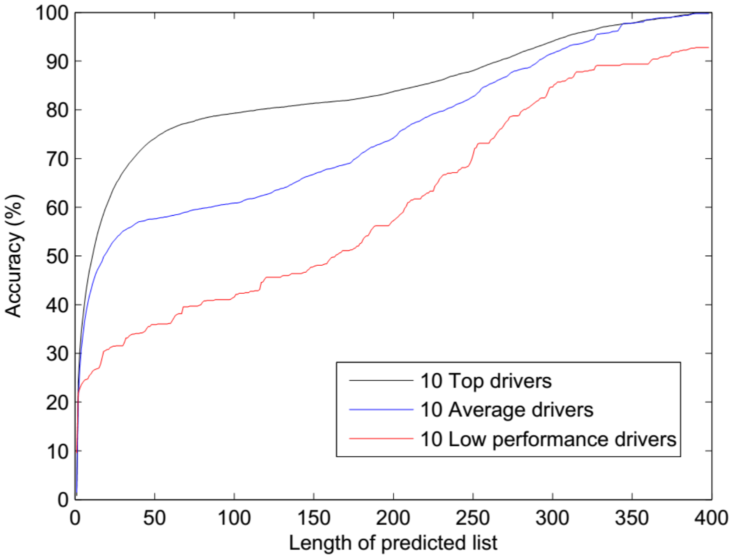


Figure 2-33 - Overall performance of the predicted list for different taxi drivers' types (top, average and low performance drivers).

Similar to Phithakkitnukoon & Dantu (Phithakkitnukoon & Dantu, 2008), we adopt the information entropy, defined by equation (8), to characterize the randomness of finding the next pick-up (*X*). Unsurprisingly, top drivers have a lower value of entropy (4.8842) than the average driver (4.9650), meaning they are less random and more predictable. Moreover, since the low performance drivers are characterized by a low amount of trips, the scarce amount of training data could also influence the performance of the classifier, as pointed out before.

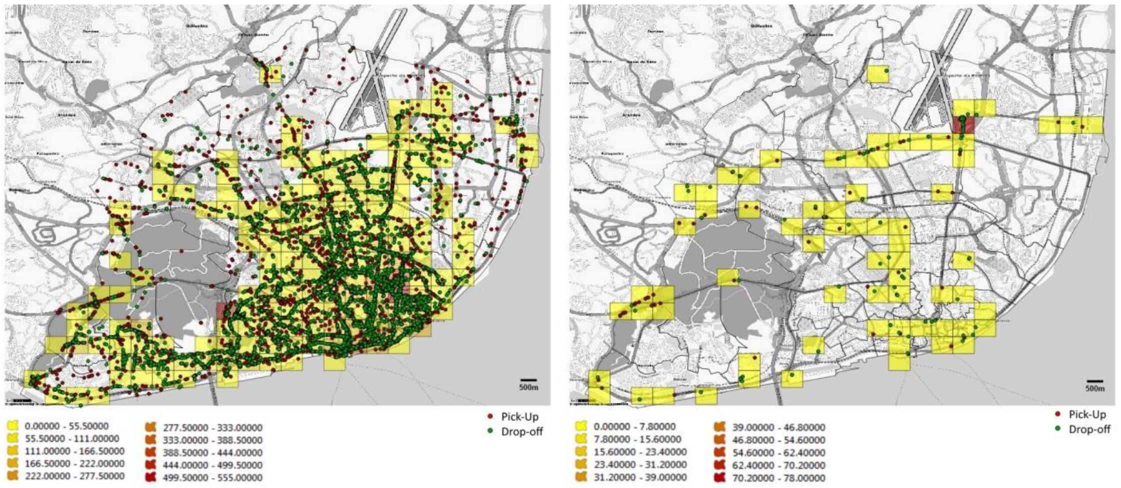


Figure 2-34 - Pick-ups and drop-offs distribution for a top driver (left) and a low activity driver (right)

Figure 2-34 shows the pick-ups (red) and drop-offs (green) for a top performance driver (left) and a low performance driver (right), overlapping a heat map indicating areas with higher activity for both drivers. Although the top driver has a high amount of pick-ups and drop-offs spread throughout the city, several clusters are visible (areas marked in red), which can help the prediction process. Generally, top drivers are more active than average and low performance drivers, and as pointed out, change locations (to specific hotspots) throughout the day, reducing their waiting period. This constant movement could have a negative impact on the predictability. However, if we associate hour of the day to the spatial movements, we can observe patterns (e.g. airport between 7 AM and 9 AM), which in turn benefits the inference engine.

The low activity driver has a single predominant location (airport), along with several isolated pick-ups and drop-offs. Usually, average and low performance drivers choose to stay at same pre-defined location for long periods of time, waiting for the next pick-up, hence the usual single predominant location. The consequence is longer waiting times.

Every sample from a data set contains on its own some degree of knowledge and should be preserved. However, these results are a possible indication that, to improve the recommendation system efficiency, samples should be filtered to retain data from the most efficient drivers and disregard data that carries the experience from low performance drivers. The use of all samples could introduce outliers, which in turn would negatively affect the outcome of the inference engines

2.6.8 Effect of taxi activity areas

The spatial analysis demonstrated that different locations have different intensities of taxi activity. Previous sections have shown that taxi service is often used as a bridge between transportation modalities, namely train stations, airports or ferry docks, an observation in line with the work of other authors, such as Zheng et al. (Zheng, et al., 2011b). Therefore, those locations have a higher taxi activity. As presented in Figure 2-5 (section 2.4.2) we divide the cells in three groups according to the taxi activity: high (or predominant cells), average, and low taxi activity cells. Figure 2-35 (left) shows the accuracy rate varying with the length of the predicted list for the three cells' types.

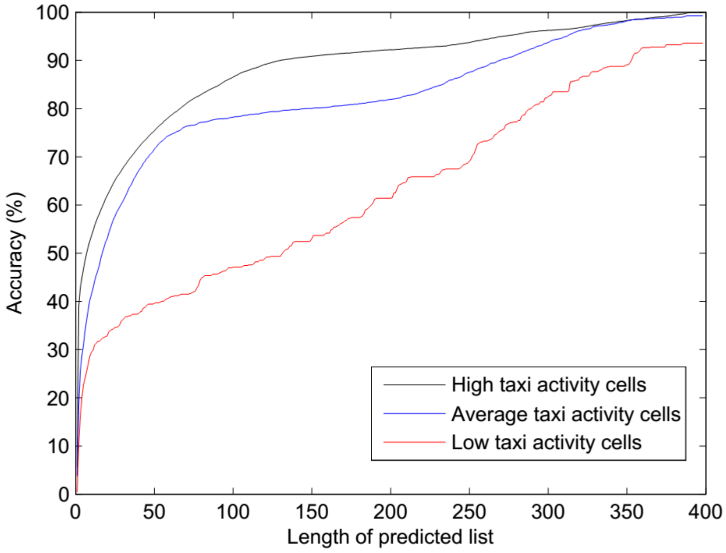


Figure 2-35 - Overall performance of the predicted list for high, normal, and low activity cells (left), and the contribution of neighbor cells (right).

The predictability of taxi activity on high taxi activity cells rapidly reaches a high accuracy (70%) with the first 37 possible destinations. Average taxi activity cells quickly attain the same value with the first 47 possible destinations, while low taxi activity cells need about 250 possible destinations from the prediction list. The corresponding entropy values are: 4.3851, 4.7395 and 4.8141. Additionally, as described before, we also observe a better accuracy predicting the next pick-up location when the data from the neighbor cells of the current location are also included in the classification.

Similar to the analysis of the effect of taxi drivers' strategies, the scarce amount of training data could also influence the performance of the classifier for the cells with low taxi activity. Therefore, considering the low predictability, higher value of entropy and scarce (or insufficient) data of cells with low taxi activity, in order to improve the recommendation system efficiency these set of cells should be disregarded.

2.6.9 Effect of cell size

One important feature for a recommendation system is the ability to allow the user to zoom in and out the interface in order to obtain more detailed information about a specific location. Moreover, since a grid to model the city is adopted in this study, the size of cells will affect the performance of the inference engine. Figure 2-36 represents the city modeled with different cell-sizes (1000mx100m, 500mx500m, 250mx250m, and 100mx100m) and corresponding taxi activity (red represents high taxi activity cells, yellow represents low taxi activity cells). Cells with the least amount of taxi activity were removed.

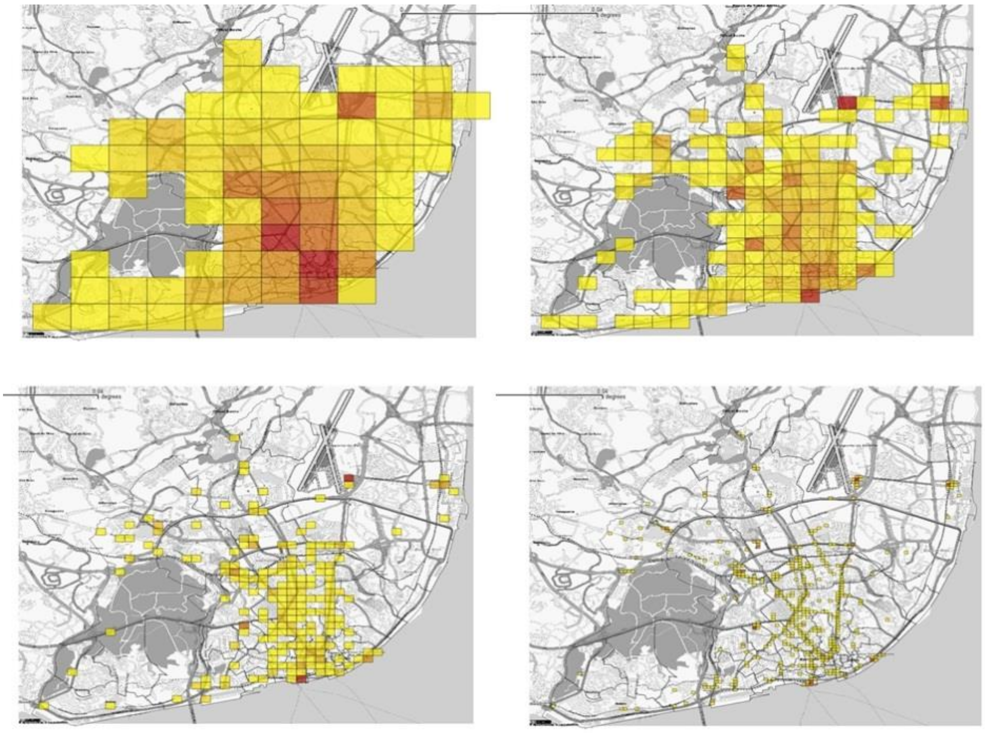


Figure 2-36 - Spatial distribution of taxi volume for different cell size: 1000mx1000m (top left), 500mx500m (top right), 250mx250m (bottom left) and 100mx100m (bottom right).

By reducing the size of each cell, there is an increase of the number of possible destinations in the search space and a substantial reduction of the amount of instances on each cell. As a consequence of the reduction of the cell size, the performance of the inference engine diminishes (Figure 2-37).

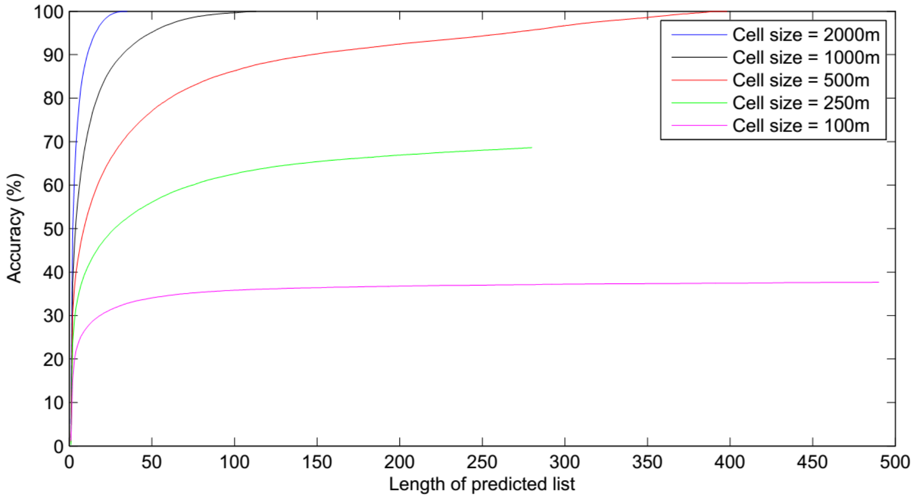


Figure 2-37 - Overall performance of the predicted list for different cell sizes.

As expected, by increasing the size of the cell, higher prediction accuracy is achieved, and with fewer destinations needed to achieve it, since there is a reduction of the search space for the inference algorithm. However, the size of the biggest cells (e.g. 2000x2000m) should not be considered practical for a recommendation system due to the sheer size where the taxi drivers would have to search for the potential customer. Smaller cells (e.g. 100x100m) provide a useful recommendation since they narrow the space for taxi drivers to find the next pick-up, but at the cost of very low accuracy. All things considered, cells with 500x500m size are the ones which appear to best balance between prediction accuracy and the usefulness recommendation for taxi drivers.

2.6.10 Adequacy of historic data

Since the data set can grow quickly, the increase in number of samples could improve the accuracy of the model, however, it can also increase the computing time and storage demands. Therefore, an interesting question is *how much data is adequate to characterize the taxi patterns*. While exploring the same data set in order to build a predictive model for the number of vacant taxis in a given area, Phithakkitnukoon et al. (Phithakkitnukoon, et al., 2010b) analyzed this question. The authors applied the information theory's mutual information (Cover & Thomas, 2006), which is a measure of the amount of information that one random variable contains about another random variable. The mutual information between two random variables $I(X;Z)$ is the reduction of the uncertainty in X due the knowledge of Z, where X is a random variable representing the entire data set and Z is a random variable representing some amount of the most recent data in X. The mutual information can be defined as follows:

$$I(X;Z) = H(X) - H(X|Z) = H(X) + H(Z) - H(X,Z) \qquad (12)$$

where $H(X)$ is the information entropy of X defined by equation (8), and $H(X,Z)$ is the joint entropy defined by the following equation:

$$H(X,Z) = -\sum_{x,z} p(x,z)\log_2 p(x|z) = -\sum_{x,z} p(x,z)\log_2 \frac{p(x,z)}{p(z)} \qquad (13)$$

with $p(x_i) = \frac{n_i}{\sum_i n_i}$, where n_i is the amount of vacant taxis in grid cell *i*.

We observe that a small amount of data (*Z*) is suffice to characterize the entire data set (*X*). The results have shown that $I(X;Z)$ converges around 40 days of historic data, therefore, since $I(X;X)$ is equal to $H(X)$, it implies that $I(X;Z) \approx I(X;X) = H(X)$ when at least the last 40 days of data are considered.

By computing the accuracy of the model considering data sets with different sizes (Figure 2-38) we can observe that after around 40 days of data, the model is able to achieve 90% of the highest performance (attained with 70 days of data).

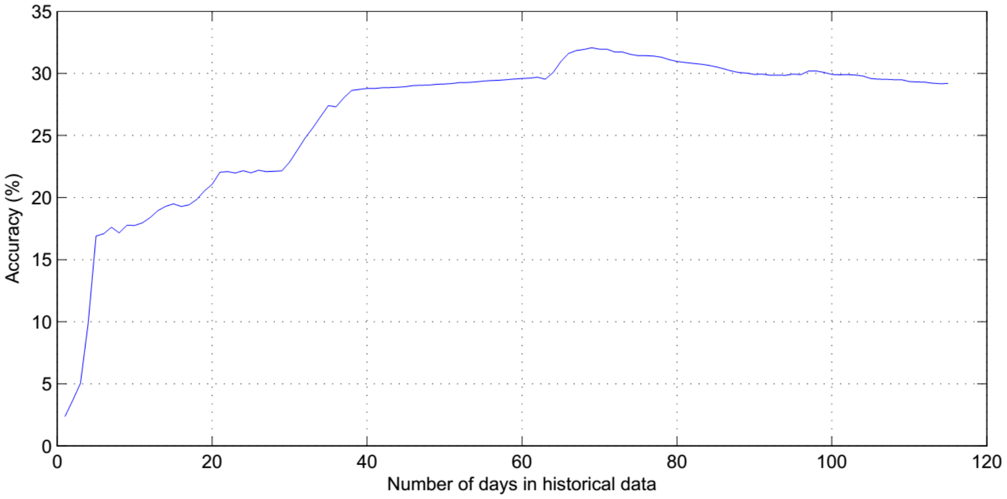


Figure 2-38 – Effect of data set size on the performance of the model: accuracy of the model considering the last *x* days of historic data.

From this result we can conclude that the inference engine does not need the entire data set to attain the highest accuracy possible. By reducing the size of historical data needed, we are also reducing the computing time and storage space.

2.7 Chapter summary and conclusions

This section summarizes the study of taxi traces in order to understand urban mobility. Main contributions and results are presented along with a discussion of the limitations of the study and future improvements of the work.

2.7.1 Overview and contributions

Taxi service is a flexible way of transportation, and dynamically adapts to the flow and need of passengers. However, the fast growth of urban areas complicates the process of efficiently searching for new customers. Therefore, taxi drivers pursue approaches to reduce waiting times and distance traveled to the next pick-up.

In this chapter we analyze a data set of taxi-GPS traces to (1) perform a exploratory spatiotemporal analysis of taxi patterns; (2) propose a recommendation system and its inference engine, based on a naïve Bayesian classifier, to assist the taxi driver in the task of picking-up new passengers; and (3) study the predictability of taxi activity and its sensibility to variations in the urban environment.

Using traces collected in Lisbon, Portugal, during a period of four months, we are able to capture the spatiotemporal variation and observe that trip distance and duration follow Gamma and Exponential distributions. We are also able to visualize the spatiotemporal patterns, identifying the main pick-up and drop-off locations and busy hours. These results were published in (Veloso, et al., 2011a).

Using the same traces, we are able to identify the relationship between pick-up and drop-off locations. The study shows strong links between public transportation hubs, where taxi service appears to be a bridge between different public transportation services. We analyze the taxi driver behavior during downtime – time spent searching for the next pick-ups - where taxis tend to avoid making long trips to suburbs for pick-ups. These results were published in (Veloso, et al., 2011b).

The analysis of top drivers' patterns shows specific strategies used to maximize their profit. It is preferable for a taxi driver to wait for passengers in locations related with main public transportation hubs (airport, train stations, ferry dock or main bus stops), during specific hours of the day, and avoid traveling great distances to the next pick-up location, unless it is to return to the aforementioned locations. Low performance drivers stay at the same location for long periods of time, show scattered pick-up locations, and are the major contributors to the apparent randomness of taxi flow. Therefore, the training process should rely on historical data from top drivers to retrieve their successful strategies. Moreover, when computing the best pick-up locations considering the current location, the historical data from the adjacent or neighbor cells should also be taken in consideration, since they all could represent a local cluster.

Our predictability analysis explores the possibility of estimating the next pick-up area (highest likelihood), given the current location (last drop-off area), day of the week, hour, weather conditions and area type. The inference engine, based on a naïve Bayesian classifier, achieves 56.3% of accuracy for specific conditions: weekdays, working hours and in areas with reasonable taxi activity (cells with low taxi activity were disregarded). Current location turned out to be the main contributor to the algorithm, contrarily to weather conditions which is the variable with less weight in the calculation. The impact of other variables is also examined: daily and weekly periods; taxi driver strategies; and cell size, alongside the study of the adequacy of historic data. These results were published in (Veloso, et al., 2011c) and (Phithakkitnukoon, et al., 2010a).

2.7.2 Limitations and future work

Nonetheless, the study presents some limitations. The highest accuracy (56.3%) is achieved under specific conditions: on weekdays; working hours; and in areas with reasonable taxi activity (cells with low taxi activity are disregarded). More importantly, the evaluation of the model considers that neighbor cells are also a positive prediction. The approach was needed due to insufficient data in some temporal and spatial slots. By removing these special conditions, the classifier achieves an accuracy of 31%. This is an expected lower performance since the search space is widened, and includes areas without sufficient data (low taxi activity cells). However, although this is a considerable lesser result compared to the first scenario, it is a reasonable outcome once we consider that the search space is composed of 370 possible locations. In this greater search space, each cell on average, has a probability of only 0.27% to randomly receive the next taxi pick-up, while the *a priori* probability for the best cells (those with very high taxi activity) is 2.4%. Since the lack of data is a major limitation in applying the algorithm to all of Lisbon municipality, a more comprehensive data set should be collected.

Additionally, the adoption of a Naïve Bayesian Classifier requires that the predictors are independent variables. However, POI may not be independent from day of the week and hour of the day. For instance, schools and commercial areas operate mostly on weekdays, from 9 AM to 6 PM, while bars and restaurants are attended mostly at night, and museums visited during weekends. This possible dependency between the variables could affect the performance of the Naïve Bayesian Classifier and other algorithms should also be explored. In that line, the study does not take into consideration urban events (e.g. sports, music concerts, cultural expositions, or even workers strikes from public transportation services), which could strongly affect the average patterns of taxi service.

Although some authors proposed the use of Artificial Neural Networks (KNN), in the form of multilayer perceptron with backpropagation, to tackle the possible randomness of taxi trajectories (Zhang, et al., 2012a), (Moreira-Matias, et al., 2014b), this approach did not improve the accuracy of estimating the likelihood of each possible pick-up location area. NBC and KNN showed similar performances (using a sigmoid activation function, and trained with different setups of hidden layers, from two to 20), however, KNN presented a significantly higher computational time.

Future work should deepen the analysis of top drivers, to uncover specific strategies and improve the classifier accuracy. Top drivers possess a cumulative knowledge from their own experience about traffic, the city topology and even passengers behaviors, being able to identify time periods and locations with higher likelihood for picking-up new passengers.

A contribution from our work is to model the entire city with a grid and consider every cell of the grid as a potential pick-up location. This approach contrasts with most of the authors, which define a set of restricted hotspots, thus reducing the search space. Although our approach encompasses the entire area of the city, it also produces a broader search space, which impacts negatively the performance of the classifier. This is a limitation in the original design of the problem, which results in a lower prediction accuracy.

The process for the grid creation can also be subject to discussion, since it doesn't take into consideration the location of particular areas of the city that affect the traffic conditions on time and space (e.g. main expressways and arterial road, stadiums, shopping centers, schools) or the original density of the taxi-GPS traces. Therefore, some hotspots could have been split between two (or more) grid cells, thus disturbing the outcome of the analysis. Castro et al. (Castro, et al., 2013) argue the benefits of a more adaptive grid decomposition, using the clusters of GPS traces to guide the size and placement of the grid. The authors proposed structures such as Binary Space Partitioning Trees or R-Trees in order to achieve that goal.

Some of the exploratory studies should be deepened in future work (e.g. gravity map of taxi activity), to analyze the effect of weather conditions or the hour of the day. Since the data set do not contain information about individual passengers, though daily patterns arise, we cannot confirm if they are based on the same share of passengers that keep daily routines or if they are produced by random passengers with random behaviors that collectively produce a defined pattern. Therefore, the nature of the passengers can have implications on the taxi service and the ability to make predictions. Moreover, this work does not explore the motivations as to why people use taxi services. Passengers could use taxis either because they do not have any other public transportation available or simply because it is convenient. A survey should be taken to clarify this issue since it can also affect taxi patterns (e.g. if passengers use taxis in a specific city's area because there is no other mean of transportation available, if a new bus or metro line is provided, taxi patterns on that area would be significantly affected).

Other limitations concern the recommendation system. A full recommendation system was not built, which is outside of the scope of this work. The study focuses instead on the inference engine. However, the development of a fully functional commercial framework would allow the system to be tested in real conditions and assess the true usefulness of the approach.

The study does not take into consideration if taxi drivers search for passengers independently or if they are being redirected by a central dispatching service. Some apparently random behaviors can be caused by specific and occasional customers' requests by phone. Additionally, the recommendation system does not take into consideration the current behavior of other taxi drivers. If the same information is provided to different taxi drivers, the recommendation system can lead to a scenario where taxi drivers compete for the same resource. A distributed and concurrent system should be explored, where estimations are performed taking in consideration the current status of other taxis, thus avoiding competition for the same passenger.

Due to the lack of data available for the same temporal window from other urban areas, the model was not tested in other cities. Although the data provider also made available a data set collected in second major city in Portugal (Porto), it did not contain enough samples to be analyzed. Therefore, our study is unable to perform any statement regarding the geographic replicability. Moreover, the absence of a complementary data set from a different year left us unable to also perform any statement regarding temporal replicability. Thus, as stated before, newer data set should be collected in the future, for the same urban region alongside with data set from different locations, to validate the temporal and geographic replicability of the model. Preferably, the newer data set should provide a wider temporal window to analyze seasonal effects of taxi service.

Finally, concerns around the quality of data. The data set was collected in 2009, representing a considerable temporal gap to the results now being presented. In between, the city under study has gone through several changes (e.g. administrative fusion of parishes, new urban development and policies, national crises changing inhabitants' habits). One can inquire if the results still hold true today. Additionally, although at the time of the data collection the data provider accounted for nearly 20% of taxi share in the city, the representativeness of the data can also be disputed. Once more, future work should encompass the collection of new data set, to assess if the procedure still holds true for different temporal periods, and if it could be expanded to other cities.

Chapter 3
The relationship between mobile phone activity and taxi traces

3.1 Introduction

The rapid growth of urbanization, the need for better services (e.g. public transportation, energy, communications) and urban planning (e.g. infrastructures, environments, policies) demands a better understanding of city dynamics. The development of pervasive technologies such as the global system for mobile communications (GSM) and the global positioning system (GPS) provide useful tools to sense social and traffic activities in the city. Analyzing GPS-enabled vehicle traces and mobile phone activity thus provides, to some extent, an overview of how the city functions.

Today's taxis are equipped with GPS devices for better monitoring and dispatching. Their traces have been used to study different aspects of the traffic network as they can provide data that reflect the state of traffic flow in the city (Liu, et al., 2009a) , (Yuan, et al., 2011a), (Castro, et al., 2012). Taxi traces typically carry occupancy information from which pick-up and drop-off location information can be inferred. Therefore, one can infer active spatiotemporal areas for taxi activity, as had been explored in the previous chapter.

Mobile phone call data, on the other hand, has been used to study the social aspect of the city (Candia, et al., 2008), (Becker, et al., 2011a), (Isaacman, et al., 2011), (Phithakkitnukoon, et al., 2014). With its high penetration rate, activity inferred from mobile phones can reveal the city's social characteristics.

By examining these two sources of data that describe the city from different perspectives, we aim to explore hidden relationships between them – particularly the inter-predictability: *can one data source be used to predict the other?* Although they explain the city in different ways, we believe that they are related in some way and we aim to explore the underlining relationship in the following sections. That being said, in this chapter we focus on the analysis of the relationship between mobile phone call activity and taxi-GPS traces acquired in the city of Lisbon, Portugal, to understand the inter-predictability between the two urban data sources. The contribution of this work lies in the following aspects:

1. a spatiotemporal analysis of a data set of mobile phone activity,
2. a study of the inter-predictability between mobile phone activity and taxi volume.

For the former, we analyze a historical data set of mobile phone activity to identify patterns in time and space; comparing the behavior between areas with high and low activity, and measure the patterns' proximity between taxi volume and mobile phone call activity. For the latter, we use linear regression to model the relationship between the two time series, explore different scenarios to improve the linear association, and identify the optimal temporal window that best fits the two urban data sources.

3.2 State of the art

With the advent of pervasive technologies (e.g. GPS, GSM, Wi-Fi), several works have been presented with the aim to explore and improve urban mobility. Among them mining taxi trajectories has recently attracted much attention. As described in Chapter 2, taxi-GPS traces have been used in a number of studies to develop better solutions and services in urban areas such as estimating optimal driving paths (Ziebart, et al., 2008), (Yuan, et al., 2010), and (Zheng, et al., 2010), predicting next taxi pick-up locations (Liu, et al., 2010b), (Ge, et al., 2010), and (Yuan, et al., 2011b), modeling driving strategies to improve taxis' profit (Ge, et al., 2010), and (Liu, et al., 2010a), identifying flaws and possible improvements in urban planning (Zheng, et al., 2011b), and (Chen, et al., 2013a), and developing models for urban mobility, social functions, and dynamics between different areas in the city (Qi, et al., 2011), and (Castro, et al., 2013).

In addition to understanding the dynamics of vehicular networks, so too the mobility of people at the individual level is important. With their ubiquity and high penetration rate, mobile phones and cellular phone networks have become probes used to sense human behavior and social dynamics. Therefore mobile phone data has been used increasingly in various studies aiming to develop general laws that govern human behavior, such as:

- *Analyze pedestrian movements predictability, identifying daily routines, commuting patterns and important places in peoples' lives* (Eagle & Pentland, 2006), (Sohn, et al., 2006), (Farrahi & Gatica-Perez, 2008), (Gonzalez, et al., 2008), (Eagle & Pentland, 2009), (Zhu, et al., 2009), (Calabrese, et al., 2010c), (Phithakkitnukoon, et al., 2010a), (Reddy, et al., 2010), (Song, et al., 2010a), (Song, et al., 2010b), (Calabrese, et al., 2011a), (Farrahi & Gatica-Perez, 2011), (Isaacman, et al., 2011), (Phithakkitnukoon & Ratti, 2011), (Altshuler, et al., 2012), (Frias-Martinez, et al., 2012), (Calabrese, et al., 2013), (Etter, et al., 2013), (Liu, et al., 2013), (Witayangkurn, et al., 2013), (Zheng, et al., 2013), (Do & Gatica-Perez, 2014), (Järv, et al., 2014), (Kim, et al., 2014), (Lin & Hsu, 2014), (Geurs, et al., 2015), (Zhao, et al., 2015b);

- *Study the social interactions in urban areas, monitoring travels to identify relationships between urban areas, and create profiles for the city organization and the population density* (Ratti, et al., 2005), (Calabrese, et al., 2007), (Hossain, et al., 2007), (Reades, et al., 2007), (Candia, et al., 2008), (Girardin, et al., 2008), (Miluzzo, et al., 2008), (Nickerson, et al., 2008), (Pulselli, et al., 2008), (Puntumapon & Pattara-atikom, 2008), (Eagle, et al., 2009a), (Eagle, et al., 2009b), (Krings, et al., 2009a), (Krings, et al., 2009b), (Li & Chen, 2009), (Hu, et al., 2009a), (Nobis & Lenz, 2009), (Ratti, et al., 2009), (Reades, et al., 2009), (Baron & Segerstad, 2010), (Blondel, et al., 2010), (Calabrese, et al., 2010b), (Eagle, et al., 2010), (Liu, 2010), (Quercia, et al., 2010), (Aharony, et al., 2011), (Becker, et al., 2011a), (Calabrese, et al., 2011b), (Onnela, et al., 2011), (Phithakkitnukoon, et al., 2011a), (Phithakkitnukoon, et al., 2011b), (Phithakkitnukoon & Dantu, 2011), (Stenneth, et al., 2011), (Traag, et al., 2011), (Demissie, et al., 2012b), (Xiao, et al., 2012), (Domenico, et al., 2013), (Pereira, et al., 2013), (Aguilera, et al., 2014), (Hoteit, et al., 2014), (Phithakkitnukoon, et al., 2014), (Can & Demirbas, 2015), (Steenbruggen, et al., 2015), (Trasart, et al., 2015), (Zhao, et al., 2015a);

- *Create Origin-Destination matrixes for urban patterns* (subset from the previous topic) (White & Wells, 2002), (Wideberg, et al., 2006), (Caceres, et al., 2007), (Bekhor, et al., 2011), (Iqbal, et al., 2014);

- *Traffic monitoring and estimation, congestion detection and route planning* (Sankar & Civil, 1997), (Bolla & Davoli, 2000), (Ygnace, et al., 2000), (Remy, 2001), (Ygnace, 2001), (Yim & Cayford, 2001), (Cayford & Johnson, 2003), (Sauret, 2003), (Thiessenhusen, et al., 2003), (Rutten, et al., 2004), (Saraydar, et al., 2004), (White, et al., 2004), (Hellinga, et al., 2005), (Hsiao & Chang, 2005), (Schneider & Mrakotsky, 2005), (Alger, et al., 2006), (Cayford & Yim, 2006), (Cheng, et al., 2006), (Geoff, 2006), (Gundlegard & Karlsson, 2006), (Hsiao & Chang, 2006), (Jin, et al., 2006), (Thajchayapong, et al., 2006), (Bar-Gera, 2007), (Birle, 2007), (Hellinga & Izadpanah, 2007), (Fontaine, et al., 2007), (Hopfner, et al., 2007), (Maerivoet & Logghe, 2007), (Qiu, et al., 2007), (Wunnava, et al., 2007), (Hellinga, et al., 2008), (Hongsakham, et al., 2008), (Leduc, 2008), (Liu, et al., 2008), (Qiu & Ran, 2008), (Gundlegard & Karlsson, 2009), (Valerio, et al., 2009a), (Valerio, et al., 2009b), (Valerio, 2009), , (Wang, et al., 2009b), (Chandrasekaran, et al., 2010), (Bazzi & Masini, 2011), (Lv, et al., 2011), (Aguilera, et al., 2012), (Demissie, et al., 2012a), (Hillson & Santis, 2012), (Tettamanti, et al., 2012), (Virtanen, 2012), (Wang, et al., 2012), (Cheng, et al., 2013), (Steenbruggen, et al., 2013a), (Steenbruggen, et al., 2013b), (Mathew & Xavier, 2014), (Tettamanti & Varga, 2014);

- *Assess the quality of road network and develop monitoring systems* (subset from the previous topic) (Caceres, et al., 2008), (Vaccari, et al., 2009), (Herrera, et al., 2010), (Becker, et al., 2011b), (Frutos & Castro, 2014);

- *Study and predict calls' profiles* (Phithakkitnukoon & Dantu, 2007), (Husna, et al., 2008), (Phithakkitnukoon & Dantu, 2008), (Melo, et al., 2010), (Phithakkitnukoon & Dantu, 2010a), (Phithakkitnukoon & Dantu, 2010b), (Yuan, et al., 2012), (Jiang, et al., 2013);

- *Explore the impact of weather conditions on mobile social interactions* (Phithakkitnukoon, et al., 2012), (Horanont, et al., 2013);

- *Perform indoor and outdoor location* (Wang, et al., 2009b), (Otsason, et al., 2005), (Otsasson, 2005), (Bento, et al., 2005), (Veloso, 2007), (Bento, et al., 2007);

- *Determine the adequacy of the mobile phone network and infrastructure* (Onnela, et al., 2007), (Hidalgo & Rodriguez-Sickert, 2008), (Chiang, et al., 2011), (Paul, et al., 2011), (Zhou, et al., 2012).

The analysis of mobile phone data explores essentially two sets of information: anonymized call detail records or handover patterns (also known as handoff patterns). This information is used mainly to study three sets of problems: user daily patterns and routines; urban social interactions or links between different areas of a city; and the use of mobile phones as probes for traffic monitoring. Some of the representative studies in each topic are described in the following sections.

3.2.1 Analysis of pedestrian movements, daily routines and commuting patterns

Farrahi et al. (Farrahi & Gatica-Perez, 2008) aimed to study the daily routines by mining mobile phone data. The authors presented a framework built from two Hierarchical Bayesian topic models to discover human location-driven routines: *Latent Dirichlet Allocation* and *Author Topic Model*. The former automatically discovers characteristic routines for all individuals in the study (e.g. going to work, returning home) while the latter finds routines characteristic of a selected groups of users, ranking users by their probability of conforming to certain daily routines. Farrahi et al. (Farrahi & Gatica-Perez, 2011) further investigated human routines which characterize both individual and group behavior in terms of location patterns, introducing the individual's entropy as a relevant parameter. The new study explored daily and weekly routines, and analyzed individual's behavior over time to determine regions with high variations in order to identify specific events.

Eagle & Pentland (Eagle & Pentland, 2006) started by analyzing the activity of 100 mobile phones to measure information access and use indifferent contexts, recognize social patterns in daily user activity, to infer relationships, and identify socially significant locations, in the Reality Mining project. Later, Eagle et al. (Eagle, et al., 2009b) demonstrated the possibility to infer friendship among mobile phone user with 95% of accuracy, even when pairs or friends users (termed *dyadic friendship*) show distinctive temporal and spatial patterns in their physical proximity and calling patterns. The model was based on a nonparametric multiple regression quadratic assignment procedure, a standard technique to analyze social network data. Finally, Eagle & Pentland (Eagle & Pentland, 2009) aimed to predict daily routines of mobile phone users. The authors were able to identify the structure inherent in daily behavior by representing daily routines in term of principal components, from which the authors extracted the primary eigenvectors (termed *eigenbehaviors*). This procedure allows the authors, at halfway through the day, to predict the remaining routine of the user with 79% accuracy. Moreover, the study used the dimensionality reduction technique to infer community affiliations within the subjects' social network by clustering individuals into sets termed *behavior spaces*. Elements of the same space share behavioral similarities, enabling a classification with 96% accuracy among community affiliations. The authors were also able to estimate relational ties such as friendship by measuring the distance between individuals in the behavioral space.

Phithakkitnukoon et al. (Phithakkitnukoon, et al., 2010a) developed the activity-aware map that describes the most probable activities associated with specific areas of a city, combining mobile phone-location traces and POI information. Results showed a strong correlation in daily activity patterns between groups of people who share common work area types, which will decrease as the distance between them increases. The analysis of around 130 million anonymous location estimations, from mobile phone data, allowed Phithakkitnukoon et al. (Phithakkitnukoon & Ratti, 2011) to identify non-symmetrical travel patterns, which accounts for 33% of all flows. High asymmetrical flows were observed in trips between low and high congested areas e.g. urban and suburban areas, as well as trips made to and from low populated areas. Finally, authors discussed the applications for Origin-Destination matrixes.

Altshuler et al. (Altshuler, et al., 2012) investigated the possibility of learning patterns from mobile phone data over time. The authors analyzed and proposed several models (C4.5, Decision Trees, Naive-Bayes, Rotation-Forest, Random-Forest, and AdaBoostM1) to predict daily routines and social relations based on mobile phone traces and activities, testing different input parameters. The models were able to detect life-partners, ethnicity, and whether or not a person is a student. Calabrese et al. (Calabrese, et al., 2010c) proposed a model to predict the location of mobile phone users based on their past behavior. The probabilistic model analyzes the user's individual patterns as well as the collective's habits, geographical features and points of interests. Experimental results, using a massive data set collected in Boston, showed good levels of accuracy.

Song et al. (Song, et al., 2010a) studied the randomness in human behavior and to what degree individual human actions are predictable by analyzing mobility patterns of mobile phone users. Results showed that a combination of an empirically determined user entropy and Fano's inequality indicated that there is a potential 93% average predictability in user mobility. This was a consequence of most individuals being localized in a finite neighborhood, but of few travelling widely. Similar observations were made by Gonzalez et al. (Gonzalez, et al., 2008), which explored an anonymized mobile phone data set of 100,000 users, collected during six months. The authors concluded that human trajectories show a high degree of temporal and spatial regularity, each individual being characterized by a time-independent characteristic travel distance and a significant probability to return to a few highly frequented locations, such as home or work. Moreover, the individual travel patterns follow a single spatial probability distribution, indicating that despite the diversity of their travel history, humans follow simple reproducible patterns.

Using cellular network data, Isaacman et al. (Isaacman, et al., 2011) proposed and evaluated three algorithms derived from logistic regression-based analysis, and described clustering techniques to identify important locations. The authors were able to detect home and work locations accurately, which was then used to perform an analysis of commuting distance and estimate commuting carbon footprints. Jiang et al. (Jiang, et al., 2013) explored a data set of 100,000 active cell phones in China, collected during four months, to identify that inter-call duration follows a power-law distribution with an exponential cutoff at the population level. However, the authors also found differences when focusing on individual users: the inter-call duration of less active users (around 73%) follow a Weibull distribution. On the other hand, Zhu et al. (Zhu, et al., 2009) explored the trajectories and behavior of mobile phone users. The authors observed that the staying time in each location followed a Zipf distribution (Powers, 1998). This observation led to a proposal for a runtime algorithm to mine the behavior patterns with less storage resources. Additionally, the concept of transient entropy was introduced, to identify the moving speed of users, and based on which, they define and mine four types of behavior patterns: frequent locations, frequent trajectories, meaningful locations, and travel modes.

3.2.2 Study of social interactions and relationships between urban areas

Ratti et al. (Ratti, et al., 2009) introduced the concept of collecting aggregated data in cities to identify hotspots of urban interactions and deploy Location Based Services. Following this line, the authors provided a geographical mapping of cell phone usage at different times of the day, from the metropolitan area of Milan, Italy, integrated in the project 'Mobile Landscapes'. A similar approach was previously used by Ratti et al. (Ratti, et al., 2005) to develop and show in real-time the mapping of cellphone traffic intensity, traffic migration (handovers) and traces of registered users as they move through the city of Graz, Austria. In the same direction, Reades et al. (Reades, et al., 2009) attempted to correlate the usage of cellular communications with the geography of human activity derived from data of commercial premises. The authors introduced the concept of *eigendecomposition* to identify and extract recurring patterns of mobile phone usage. The primary eigenvector was considered to indicate a common underlying pattern to mobile phone usage and was used to produce *eigenplaces*. By observing a spatial variation in time, the authors were also able to relate some patterns with specific activities.

To demonstrate the applicability of a real-time urban monitoring system, Calabrese et al. (Calabrese, et al., 2007) used the Localizing and Handling Network Event Systems (LocHNESs) platform, developed by Telecom Italia, to visualize and study the urban dynamics. The study was based on the anonymous monitoring of mobile cellular networks in the city of Rome, Italy. By combining information regarding public transportation (buses and taxis), the authors produced visualizations of the pedestrians movements and traffic conditions. Later, Calabrese et al. (Calabrese, et al., 2010b) presented an analysis of crowd mobility during special social events (e.g., sport game, concert) by analyzing mobile phone-location traces. Using data collected from about one million mobile phones, the authors were able to correlate social events that people attend with their home locations. For classification, a Multilayer Perceptron with one hidden layer was applied, and a K-Means algorithm was used for data clustering. To understand the relation between people's calls and their physical location, Calabrese et al. (Calabrese, et al., 2011b) explored a database of anonymized telecommunications of over one million customers, collected during 12 months in Portugal. The authors found that around 90% of users who called each other have also shared the same space (cell tower), and around 70% of users who call each other frequently (at least once per month on average) have shared the same space at the same time. Moreover, co-locations appear highly indicative of coordination calls occurring just before physical meetings. These observations allowed the authors to predict 61% of the number of co-locations from the number of calls, and users' homes distance (the number of co-locations decreases with the increase of distance between homes).

Phithakkitnukoon & Dantu (Phithakkitnukoon & Dantu, 2007) tackled the challenge of predicting future receiving calls, and proposed a call predictor. The behavior learning model is able to compute the probability of receiving calls during the next 24 hours, based on the caller's past history. The receiving call probability is based on the caller's behavior (caller's call arrival time and inter-arrival time) and reciprocity (number of outgoing calls per incoming call and inter-call time). The framework uses a nonparametric density estimation (the Parzen window estimator) with a Gaussian of zero mean as kernel, to estimate the probability model, assuming normal distributions for the call parameters. Based on data collected during 3 months from 20 phone users, results showed that the model performed reasonably well with false positive rate of 2.4416%, false negative rate of 2.9191%, and error rate of 5.3606%. Further improvements were made by Phithakkitnukoon & Dantu (Phithakkitnukoon & Dantu, 2008), which describes a Call Predicted List. Similarly with the former work, the model is based on the user's call history to build a probabilistic model of calling behavior based on the caller's calling patterns and reciprocity, adding a Naïve Bayesian Classifier. The authors were also able to infer the social closeness from the number of calls received.

Girardin et al. (Girardin, et al., 2008) explored a new methodology to identify tourists' behaviors in urban areas. The authors studied active and passive footprints of the city's visitors in the city of Rome, Italy. For passive footprints, the authors considered the interaction with the mobile phone network, while active footprints were composed of georeferenced photos, made publicly available on photo-sharing web sites, and aggregated records of wireless network events, generated by mobile phone users making calls and sending text messages. This work made visible the potentialities of this approach (a new data source that could be used to generate tourists' profiles) and difficulties (extracting information from EXIF metadata and the error associated with the manual geotagging) of the process. The automation of collecting daily activity data and publication on social frameworks was studied in the Dartmouth's CenceMe project (Miluzzo, et al., 2008). The authors proposed and tested an intelligent mobile sensor network capable of sensing nearby friends and their current activity.

Aharony et al. (Aharony, et al., 2011) studied the population behavior through a 15-month long data set which included information gathered from individual mobile phones. The authors investigated the relationship between human networks and personal decision making, where there is a statistically significant effect of social components on real-world in-situ physical activity levels. The study showed that individuals' social interaction diversity correlates with their current income level. Additionally, the study demonstrated a relationship between the number of mobile applications that two people share in common to the time they physically spend face-to-face. Candia et al. (Candia, et al., 2008) investigated the individual and collective behaviors. The authors were interested in the occurrence of anomalous events at large scales and patterns of calling activity at the individual level. The author showed that spatiotemporal anomalies could be described using standard percolation theory tools and that the inter-event time of consecutive calls is heavy-tailed. The analysis of individual records also showed the fraction of active traveling population and their average distance traveled.

Traag et al. (Traag, et al., 2011) described an approach to correlate human mobility patterns with social events using trajectories of mobile phone users. By detecting mobility behaviors that are different from daily routines, the probabilistic framework was able to determine which users participated in the event. A simple Bayesian location inference framework is proposed and validated with a smoothened Voronoi tessellation.

Bekhor et al. (Bekhor, et al., 2011) used passive location data from cellular phone systems in order to study long-distance travel patterns. As a result, the authors were able to construct the origin–destination tables directly from the cellular phone positions. Puntumapon & Pattara-atikom (Puntumapon & Pattara-atikom, 2008) applied a Naive Bayes model to a data set generated by cellular phones to classify two mobility modes (train and pedestrian). The model searches for key properties of each type of mobility, such as the repetition of pedestrian cell ID and the consistency of a train travel pattern. By using the number of the unique cell ID and the average cell dwell time of the unique cell ID, the authors were able to correctly predict the type of mobility with 93.1% of accuracy.

3.2.3 Traffic monitoring and estimation, congestion detection and route planning

Transportation authorities rely primarily on traditional road traffic data collection methods to monitor traffic conditions. However, these methods demand specific infrastructures, e.g. loop detectors, automatic video feed-based counts. Considering the increasing size of road networks and the costs of deploying and maintaining a dedicated monitoring infrastructure, a considerable amount of road segments is not covered by monitoring system. Therefore, transportation authorities have to rely on incomplete or erroneous data to support their decisions on urban planning. New approaches have been pursued to complete that information, or even replace the traditional infrastructure for traffic monitoring. The high penetration of mobile phone technologies makes them suitable candidates to be used as opportunistic probes for traffic conditions and has been explored with that goal. Changes in handover are usually used to identify traffic movement, while changes in calls patterns suggest the occurrence of incidents.

Valerio et al. (Valerio, et al., 2009a), (Valerio, 2009) discussed and proposed an architecture to identify traffic incidents (congestion or accidents) by monitoring deviations to the global pattern of cellular networks. Subsequently, the system was developed by Valerio et al. (Valerio, et al., 2009b), which explored the mobile phone network infrastructure to implement a road traffic estimation system, where certain traffic conditions or anomalies could be signaled by the mobile phone network patterns. The system is based on the idea that anomalous traffic events (e.g. accidents or congestion) produce abrupt changes in the mobile phone network patterns. The authors explored the traditional patterns of mobile phone activity and validated the anomalies with traffic data produced by road sensors and inductive loops. Although the system should be able to differentiate specific patterns of non-road mobile phone users, the authors tested the proposed framework on highways to have a more controlled environment. Additionally, the authors discussed the potential for predicting events, and exploring the deviation of daily patterns and routines.

Cayford & Johnson (Cayford & Johnson, 2003) explored the feasibility of a traffic monitoring system based on cell phone locations. They concluded that in order to use cell phone as probes a set of features must be verified: the accuracy of locations, the frequency with which the position is updated, and the total number of locations available. Hsiao & Chang (Hsiao & Chang, 2005) proposed a segment based approach, instead of the conventional distance based approach, to estimate traffic information from mobile phone signal data. The authors argued that the difference in the positions of the mobile phones, used to measure traffic information, has imprecise location accuracy, which leads to an unstable measurement and therefore causes variation or flutter in the positioning. This issue can be overcome by using areas or segments. Through simulation, the authors showed that the mean absolute percentage error of segment based method is only 1/3 of that from the distance based method. The study also showed that even a limited mobile phone penetration rate is enough to measure traffic information, however sample size and location accuracy are two critical factors for mobile phone location based traffic information systems.

The use of handover information has been widely adopted to identify patterns in traffic. Demisse et al. (Demissie, et al., 2012b) explored and identified a correlation between handover counts, from mobile phone cell sites near roads, and traffic levels (measured by traffic counts from the same roads). The authors proposed to build a multinomial logistic regression model and to train an artificial neural network to relate traffic volume and mobile phone handovers. The models confirmed the initial strong correlation (with a high accuracy of 72% and 81% respectively), which suggests a strong relation between mobile phone handover and traffic volumes. Also using handover information from mobile phone cell sites, Tettamanti & Varga (Tettamanti & Varga, 2014) proposed a methodology to produce origin-destination (O-D) matrixes and macroscopic traffic flow estimation. O-D matrixes were built using measurements and by filtering, signaling events occurring within the corresponding location area of the mobile phone network. On the other hand, traffic estimation is based on the aforementioned O-D matrixes and travel time data, obtained from handover sequences.

Cellular handover patterns from cellular phone networks were also studied by Becker et al. (Becker, et al., 2011b) to identify preferable routes taken by the inhabitants of urban areas. The authors showed that handover patterns are relatively stable across different routes, speeds, directions, phone models, and weather conditions. Additionally, they introduced a metric for measuring route variability, based on Earth Mover’s Distance, which was used to study the variability across repeated drives of the same route and between routes. Two algorithms were proposed to match handover patterns to routes: a nearest-neighbor classification and a probabilistic approach that uses the signal strength to compute the likelihood that a given handover pattern occurs on a particular route. Likewise, Chiang et al. (Chiang, et al., 2011) relied on multiple handover patterns to collect traffic information. However, instead of building the system over GSM technology, the authors explored the Universal Mobile Telecommunications System (UMTS). A proposal for a passive framework is presented, where instant traffic information estimation is composed of three phases: pattern matching, session speed calculation, and road speed calculation. Firstly, the signals are organized into clusters of the same definition with patterns using two proposed algorithms: *Cell Clustering Algorithm* and *Cluster Oscillation Filtering* (Chiang, et al., 2011). Traveling speed for each call session will be calculated by arrival time and distance between consecutive handovers (when during the travel, the caller moves from cell site A to B, and then from cell site B to C). Finally the framework merges the travel speed of all the session data into an average travel speed. The estimation errors were less than 20km/h in 82% of the measurements, and estimated speed values shared the same evolutionary trend with actual speed.

Wideberg et al. (Wideberg, et al., 2006) explored the notion that a GSM network has a constant estimation of the position of each terminal, referring to the location area of the base station that provides services, to acquire information for the origin-destination of traffic. Based on this principle, the authors simulated the path of several vehicles producing origin-destination matrixes. To prune the information regarding mobile phone users which are not traffic-related, the authors introduced an *adjustment factor* (or *Mobile Phones per Vehicle Equivalent*). The simulation was validated with information from traffic count sensors. Gundlegard et al. (Gundlegard & Karlsson, 2006), (Gundlegard & Karlsson, 2009) measured the accuracy for travel time estimation in both GSM (2G/2.5G) and UMTS (3G) systems. The study concluded that UMTS radio measurement data (higher data rate and shorter delay) and handover point data together can be used to predict travel time even more correctly than GSM, since the network reacts much faster to changes in the radio environment (better synchronization between base stations and mobile terminals). Therefore, the higher location accuracy in the UMTS network can be used to improve travel time accuracy, which is useful when detecting incidents.

In order to investigate the feasibility of using cellular phone data to estimate traffic volume, Qiu et al. (Qiu, et al., 2007) applied a linear regression approach. The authors were able to estimate travel speed and travel time. Furthermore, it was suggested to also integrate a macroscopic traffic flow theory such as Kalman Filtering and Particle Filtering. Bar-Gera (Bar-Gera, 2007) studied the use of information from a cellular phone to measure traffic speeds and travel times. The system focuses on handover events at which control of a phone is handed over from one cell to another.

The system matches this sequence of footprints generated by a moving vehicle to a route segment along the road network that appears to be the most likely. The algorithm also takes into account the possibility that not all observations are actually related to vehicles, but traveling together with the regular traffic along the designated road section. Similar results to previous works were achieved: correspondence between measurements from the cellular phone network and data collected from loop detectors.

3.2.4 Our approach

The aforementioned studies focus essentially on the analysis of anonymized call detail records to identify an individual's daily routines, or the examination of handover patterns to monitor traffic volume, which has been extensively explored. As a contribution, this work investigates the relationship between mobile phone call intensity and taxi volume, two data sources that describe the city in different ways. The study aims to identify in which extent the patterns of one data source can be used to estimate the other. Thus, our approach differs from those aforementioned by:

(1) performing an exploratory analysis to identify daily and weekly patterns of mobile phone activity; studying the potential for correlation between taxi volume and mobile phone activity; and identifying which variables are better predictors of mobile phone activity;

(2) performing an inferential analysis aiming to comprehend to what extent taxi volume can be used to estimate the mobile phone activity through linear regression; studying different scenarios and setups to improve the linear association; and apply a sliding window to identify the best fit.

To our knowledge, this is the first study exploring mobile phone call intensity and taxi volume altogether, and we hope that this study will pave the way for more in-depth investigations in this direction.

3.3 Methodology

Analogously to the previous chapter, our approach is based on the classic process of knowledge extraction from databases as described by Witten & Frank (Witten & Frank, 2005) and Santos & Azevedo (Santos & Azevedo, 2005). It comprises the following steps:

- Data collection;
- Data cleaning and transformation;
- Exploratory analysis;
- Inference analysis.

Taxi information was collected and provided by *Geotaxi*, corresponding to a database of taxi-GPS traces, while mobile phone information was collected and provided by *TMN* (currently rebranded to *MEO*). Both data sets were collected during the same time window (December 2009).

A cleaning and transformation process is applied in order to remove faulty, erroneous or missing samples, as well as to format the data set to a more suitable scheme. Additionally, data sets are aligned with the same sampling rate and expressed as time series. Finally, data is stored on a relational database where every sample is geo-referenced.

An exploratory study is performed to understand the distribution of the mobile phone activity in time and space (the analysis of the taxi data set was performed in the previous chapter). The aim of this step is to identify patterns and a possible correlation between taxi and mobile phone activity, to be further investigated during the inference analysis. Data exploration examines the spatiotemporal patterns, compares the behavior between areas with high and low mobile phone activity, and measures the patterns' proximity between taxi volume and mobile phone call intensity using Euclidean distance.

The main goal for the data analysis process is to study the interplay between mobile phone call intensity and taxi volume. More specifically, to which extent can one data source be used to predict the other. In order to achieve that goal, we extract the linear regression between both time series, using the method of the least squares

(Pallant, 2005), exploring different scenarios to improve the correlation (e.g. weekdays *versus* weekends, working hours *versus* night hours). To identify the best temporal fit we use a sliding window. The coefficient of determination (r^2) is used to attain the goodness of the linear function fitness (Kennedy, 2008). The significance of the regression was tested using the F test of ANOVA, which verifies the existence of a linear relation between the dependent variable and the predicting variables (Maroco, 2005).

Main findings and achievements were submitted to a peer-review international conference to validate our procedures and results (Veloso, et al., 2012).

3.4 Data description

This section describes the data on mobile phone call intensity and taxi volume in Lisbon, Portugal. The data was collected in December 2009 (a period of 31 days) for both data sets.

3.4.1 Mobile phone data

The mobile phone call intensity data set was provided by *TMN*[17] (currently rebranded as *MEO*), which is one of the main telecommunications operators in Portugal, with a market share of about 40%. The data set contains information from the traffic channel (TCH), which carries voice and data signals (time slot assignments), as illustrated in Table 3-1 and Table 3-2. The data was aggregated by hour, for each cell, with cleaning and transformation procedures performed by the data provider. Busy hour traffic (in *erlangs*) is considered for each cell, which is defined as *call intensity*. Busy hour traffic is a standard measure of carried load, widely used by mobile phone operators, which represents the average number of concurrent calls during an hour-period. An *erlang* is one person-hour of phone use, therefore, '1 *erlang*' could represent one person talking for an hour, two people talking for a half hour each, 30 people speaking for two minutes each, and so on (Reades, et al., 2007). Although the data set contains samples from December 2009 to March 2010, only a subset along December 2009 is used to align with the temporal window of taxi data set.

[17] MEO. http://www.meo.pt

Atribute	Type of data	Format	Description
cellid	Integer	xxxxxxxxxx	Cell ID (LAC + GSM ID)
cellname	String		Cell name
day	String (date)	mm-dd-yyyy	Date for data acquisition
hora	String (hour)	hh:mm:ss	Hour for data acquisition
attempts	Integer	xxx	Number of call attempts + handover
normatts	Integer	xxx	Number of call attempts
blocks	Integer	x	Blocked calls (due traffic congestion)
congtime	float	x.xx	Congestion time
seizures	Integer	xxx	Number of connections + handover
normseizs	Integer	xxx	Number of connections
traffic	float	x.xx	Busy hour traffic (erlang value)
maxbusy	Integer	xx	Max number of occupied channels
rflosses	Integer	xx	Number of dropped calls due to RF problems
dropped	Integer	x	Number of dropped calls
availch	float	x.xx	Number of available channels
definedch	Integer	xx	Number of defined channels (frequencies)

Table 3-1 - Attribute description from mobile phone’s raw log.

```
1110501643,FIL_2-01643,04/01/10,07:00:00,58,34,0,0.00,57,34,0.94,5,0,0,17.00,17
1110501643,FIL_2-01643,04/01/10,08:00:00,164,101,0,0.00,155,101,2.49,9,0,0,17.00,17
1110501643,FIL_2-01643,04/01/10,09:00:00,315,151,0,0.00,289,151,4.79,11,1,0,17.00,17
1110501643,FIL_2-01643,04/01/10,10:00:00,361,203,0,0.68,346,203,5.80,23,0,0,17.13,20
1110501643,FIL_2-01643,04/01/10,11:00:00,408,195,0,0.00,390,193,5.22,11,0,0,17.00,17
1110501643,FIL_2-01643,04/01/10,12:00:00,471,278,0,0.00,455,276,7.68,17,0,0,17.05,19
1110501643,FIL_2-01643,04/01/10,13:00:00,441,243,0,0.00,428,243,6.88,16,0,0,17.10,18
1110501643,FIL_2-01643,04/01/10,14:00:00,494,246,0,0.00,466,246,7.24,15,0,0,17.00,17
1110501643,FIL_2-01643,04/01/10,15:00:00,365,180,0,0.00,354,179,6.42,14,1,0,17.00,17
1110501643,FIL_2-01643,04/01/10,16:00:00,402,183,0,0.00,379,183,5.31,13,0,0,17.00,17
1110501643,FIL_2-01643,04/01/10,17:00:00,320,164,0,0.00,303,164,4.85,15,0,0,17.00,17
1110501643,FIL_2-01643,04/01/10,18:00:00,337,175,0,0.00,324,174,5.36,12,2,0,17.00,17
1110501643,FIL_2-01643,04/01/10,19:00:00,256,133,0,0.00,247,132,3.05,8,0,0,17.00,17
```

Table 3-2 - Sample of mobile phone’s raw log (10 records).

Figure 3-1 shows a spatial distribution of mobile phone call intensity in Lisbon, where each dot represents the location of a cell site and its size corresponds to the average amount of calls per hour. Areas with higher call intensity usually present higher taxi volume, as shown in Figure 3-2, where the spatial distribution of cell sites and corresponding mobile phone call intensity (represented by the radius of the circle) superimposes the taxi activity grid (where red cells represent an higher taxi activity).

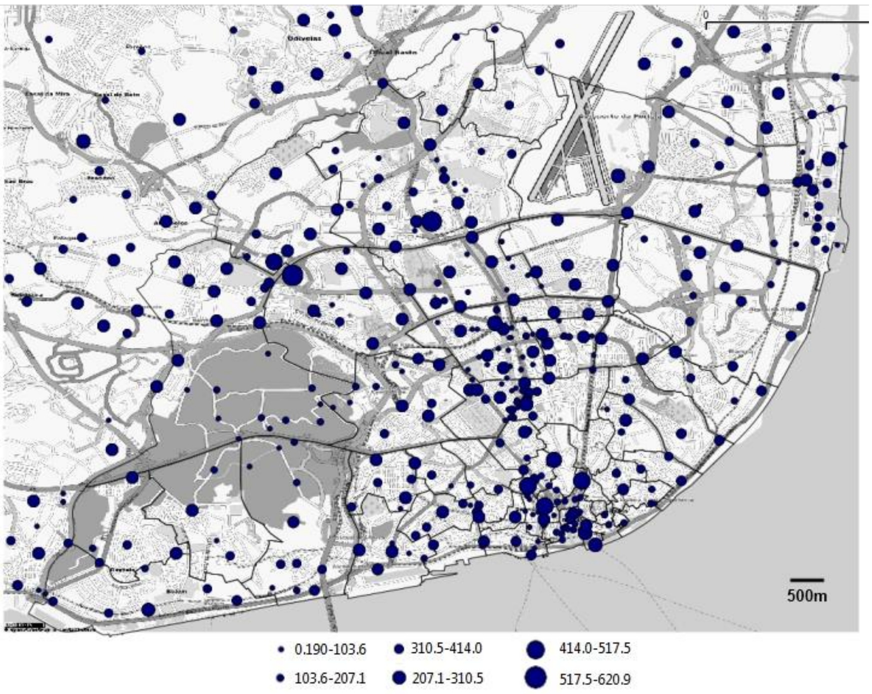


Figure 3-1 - Spatial distribution of cell sites and corresponding mobile phone call intensity in Lisbon (average amount of calls per hour on each site).

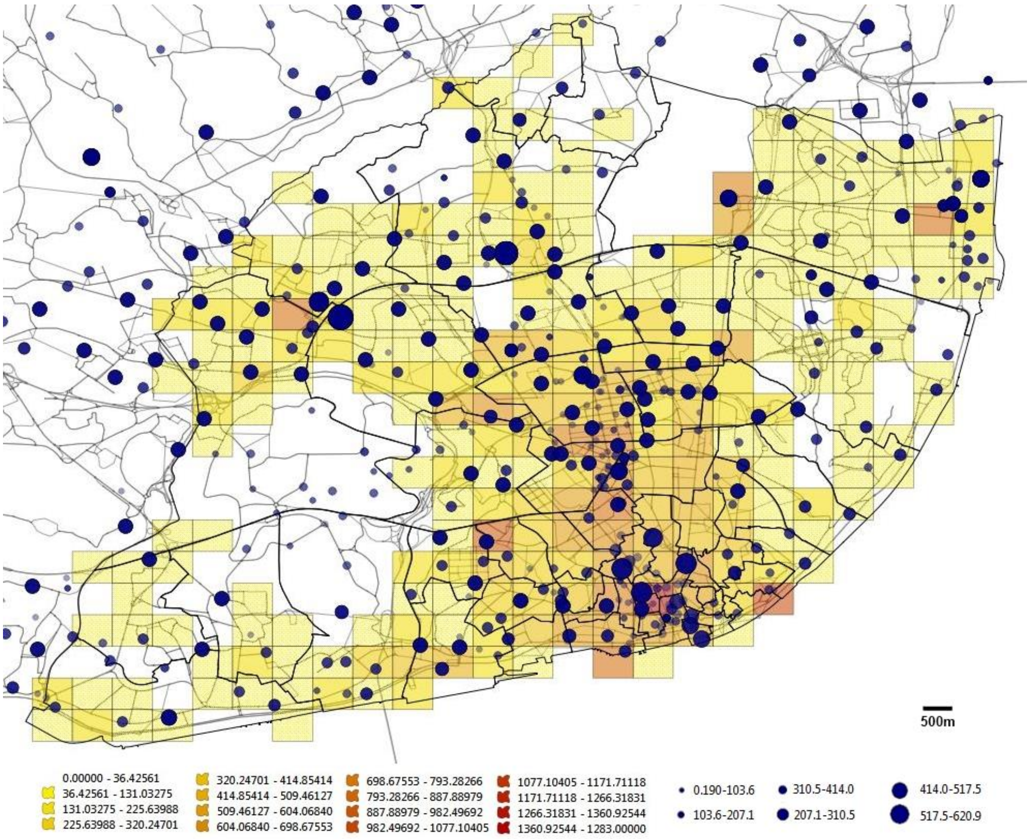


Figure 3-2 - Spatial distribution of mobile phone cell sites in Lisbon (average amount of calls per hour on each site) and taxi activity (number of pick-ups and drop-offs).

Mobile phone activity varies during the day. Figure 3-3 shows the hourly variation, with an increase of activity after 8 AM, high activity from 12 PM to 8 PM and a steady decrease until 4 AM when it reaches its minimum activity. A small decrease of activity is observable at 2 PM, followed by another increase, peaking at 8 PM. This pattern strongly relates with typical business hours. Figure 3-4 presents the daily variation, showing an higher activity on weekdays than on weekends.

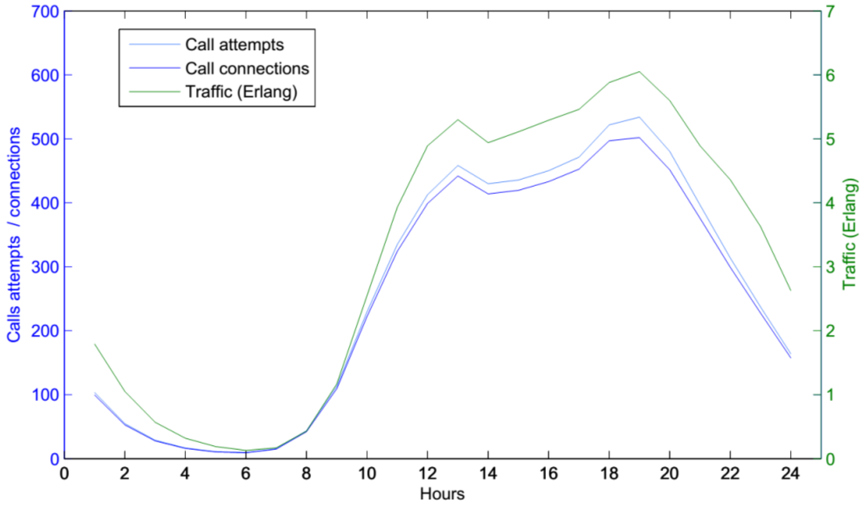


Figure 3-3 - Hourly variation of mobile phone call intensity in Lisbon (average of call attempts, successful calls and busy hour traffic).

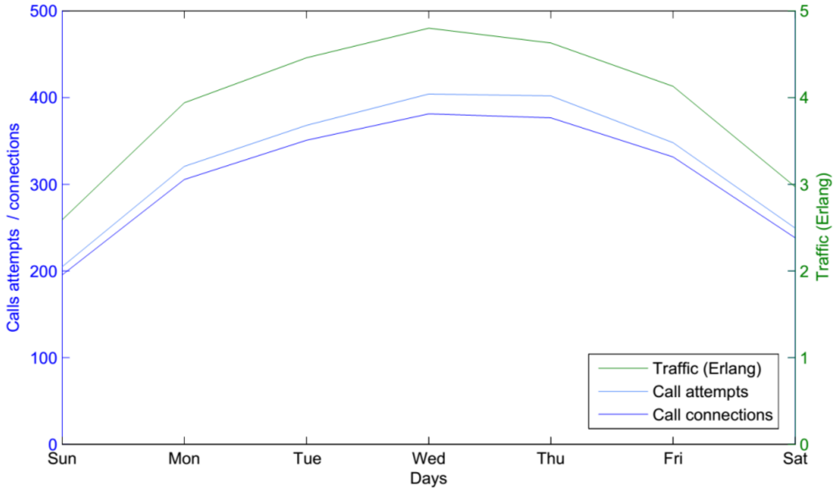


Figure 3-4 - Daily variation of mobile phone call intensity in Lisbon (average of call attempts, successful calls and busy hour traffic).

Three indicators are shown: the average amount of call attempts, the average amount of successfully started voice calls (which must be equal or lesser than the amount of call attempts), and the busy traffic hour (in *erlangs*). All indicators behave similarly, showing the same hourly and daily patterns. As stated before, for this study we analyze the busy hour traffic for each cell site. Nonetheless, at the end the same analysis was performed using all three indicators, with quite similar outcomes.

3.4.2 Taxi data

The taxi data set was provided by *GeoTaxi*. This corresponds to the data set described in the previous chapter, from which only a subset from December 2009 is used in this chapter. This subset comprises around 500,000 taxi-GPS location points, collected from 230 taxis. Along with the GPS location (latitude, longitude) information, it reports speed, bearing, engine status, and occupancy status. The amount of pick-ups and drop-offs were inferred, which accounted for 26,924 distinct trips and was termed *taxi volume* in this study.

The overall taxi volume's spatial distribution in Lisbon is shown in Figure 3-5 (on 500mx500m-grid cells, identical to Figure 2-4 however composed solely with data from December 2009), where the number of pick-ups on each cell during the period under study is represented by a color scale (red corresponds to cells with a higher number of pick-ups). City downtown (A), airport (B), train stations (C, D) and ferry dock (E). Different public transportation modalities (airport, train, ferry, bus) are well connected through taxi services.

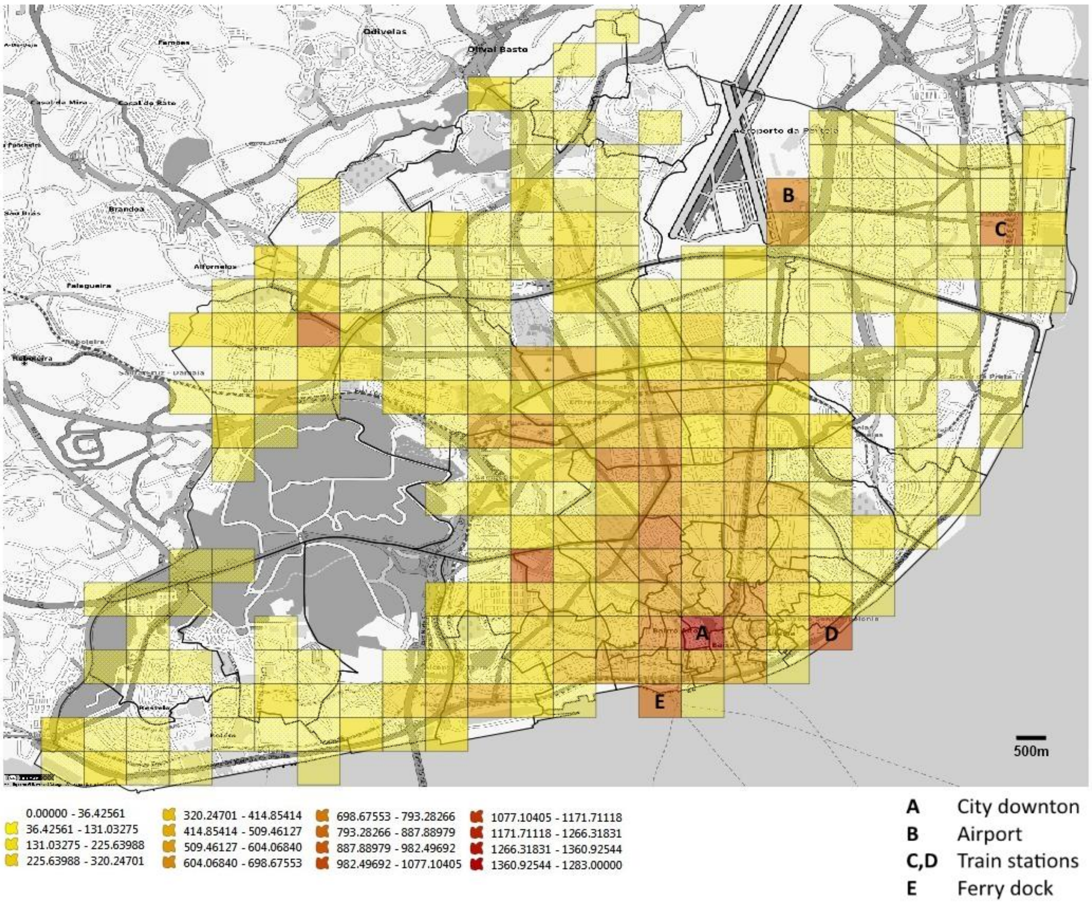


Figure 3-5 – Spatial distribution of taxi volume concentration in Lisbon.

Figure 3-6 compares the daily distribution of taxi service in December 2009 against a similar daily distribution of taxi service along the entire data set (from September to December 2009). The hourly pattern is similar in both periods. Taxi service is active throughout the day, but shows a peak of activity during business hours. It gradually increases from 7 AM, reaches its maximum between 11 AM and 1 PM, and slowly drops down in the afternoon. Values of taxi volume are normalized in the interval [0,1] to enable comparison.

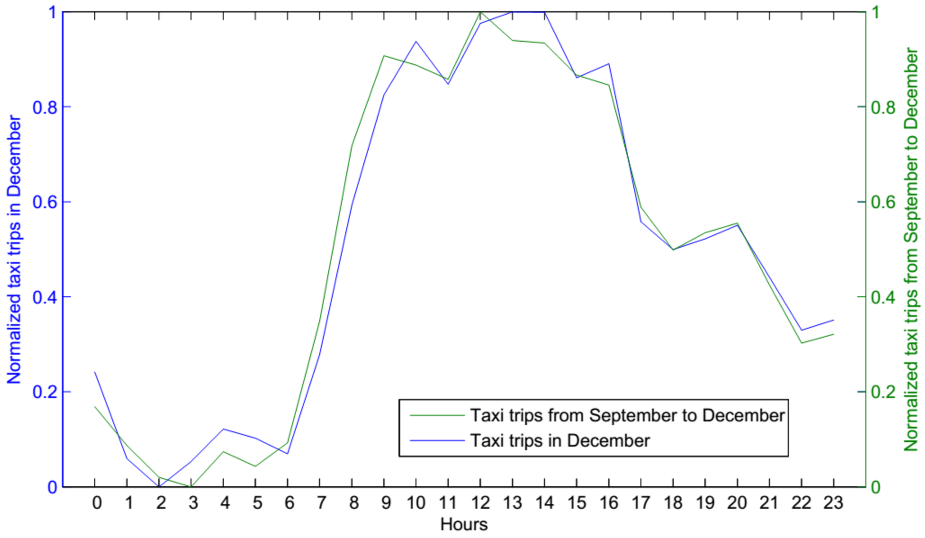


Figure 3-6 - Hourly variation of taxi volume in December 2009 (blue) and from September to December 2009 (green), corresponding to the entire data set in Lisbon.

The number of active taxi vehicles also varies during the day (Figure 3-7). It increases and decreases with the variation in taxi demand. This variation in the number of taxi vehicles precedes the variation in taxi demand.

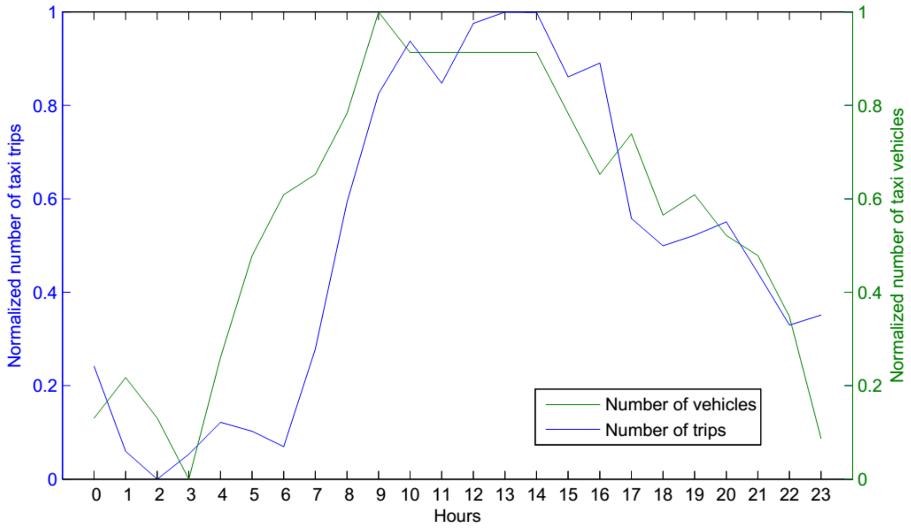


Figure 3-7 - Hourly variation of taxi volume (blue) and active taxi vehicles (green) in Lisbon.

3.5 Data exploration

This section characterizes the temporal and spatial patterns of mobile phone activity throughout the time series. Additionally, the correlation between mobile phone activity and taxi volume is studied, in order to identify relationships to be further examined during the inference analysis.

3.5.1 Mobile phone data

The mobile phone activity is fairly cyclic. By observing the time series (Figure 3-8), which represents the mobile phone call intensity on each hour, from December 1st, 2009 to March 31st, 2010, we notice cyclic components: daily and weekly patterns. A reduction in mobile phone activity on weekends is clearly visible, more pronounced on Sundays. Two small deviations are visible on the fourth and fifth week of December (marked in red) which corresponds to December 24th and December 31th, traditional periods of high intensity for mobile activity. Call intensity goes in an opposite direction on other national holidays (marked in grey): December 1st, 8th and 25th, January 1st, and February 16th ("*Carnaval*" or Shrove Tuesday). Although these events took place on weekdays (Tuesdays and Fridays) the pattern resembles a weekend, with low mobile phone activity. From winter colder months (December) to spring warmer months (March) the daily and weekly patterns are constant.

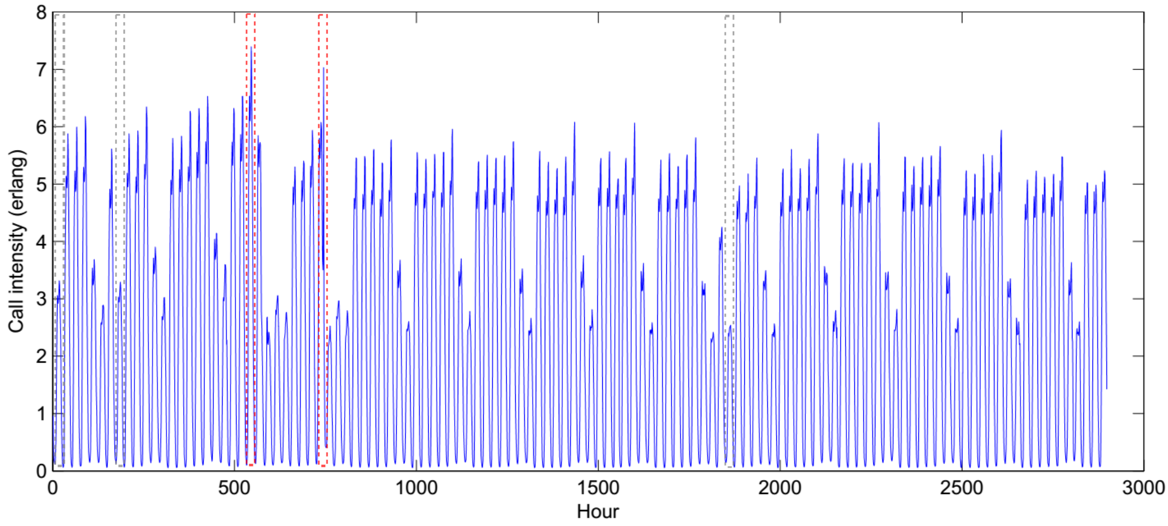


Figure 3-8 - Time series of average call intensity in Lisbon, for each hour, from December 2009 to March 2010.

The same conclusion can be obtained by observing the daily variation of the mobile phone call intensity in the month of December against the fairly similar variation of the mobile phone call intensity from December to March (Figure 3-9).

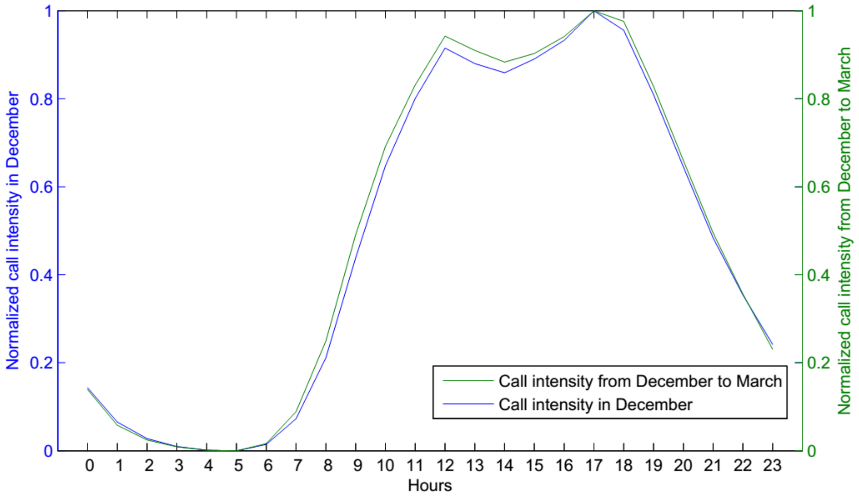


Figure 3-9 - Hourly variation of mobile phone call intensity in December 2009 (blue) and from December 2009 to March 2010 (green), corresponding to the entire data set in Lisbon.

Grouping cell sites into three categories (high activity, medium activity and low activity), based on the call intensity, suggests that the daily usage of the mobile phone service has a similar pattern across urban areas (Figure 3-10). Values of call intensity are normalized in the interval [0,1] to permit comparison.

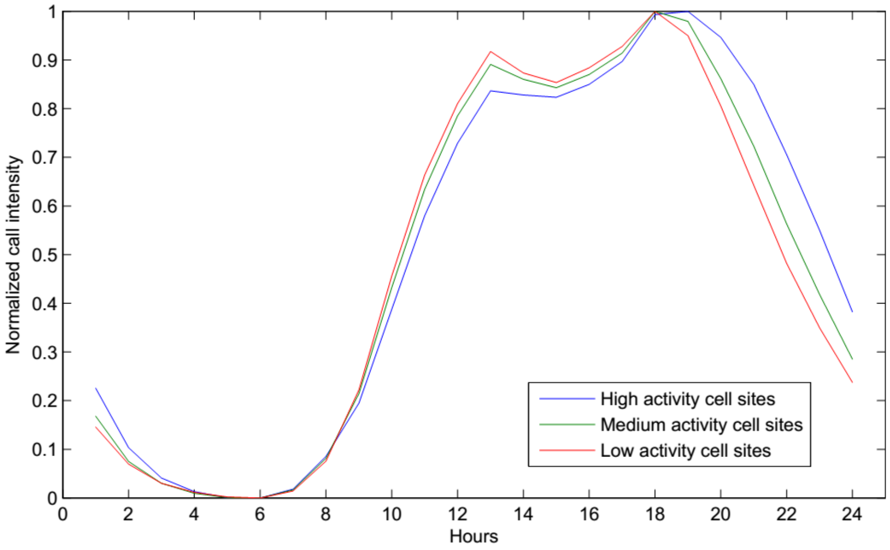


Figure 3-10 - Hourly variation of mobile phone call intensity on different cell sites (high, medium and low activity).

Some specific and individual cell sites can present some deviations from the average pattern. For instance, the cell site at the airport shows a third peak of activity around 9 AM, while at cell sites at *Oriente* (train station and commercial zone) the second daily peak of activity (around 9 PM) is significantly higher than the first peak of mobile phone activity (around 1 PM).

A similar scenario is observed when the call intensity is grouped and analyzed according to taxi activities (high, medium and low taxi activity cells) on the grid used to model the city (Figure 3-11). The same daily pattern arises in all groups of cells, with minor variances. The classification of cells on high, medium and low taxi activity follows the same procedure adopted for Figure 2-5 (section 2.4.2).

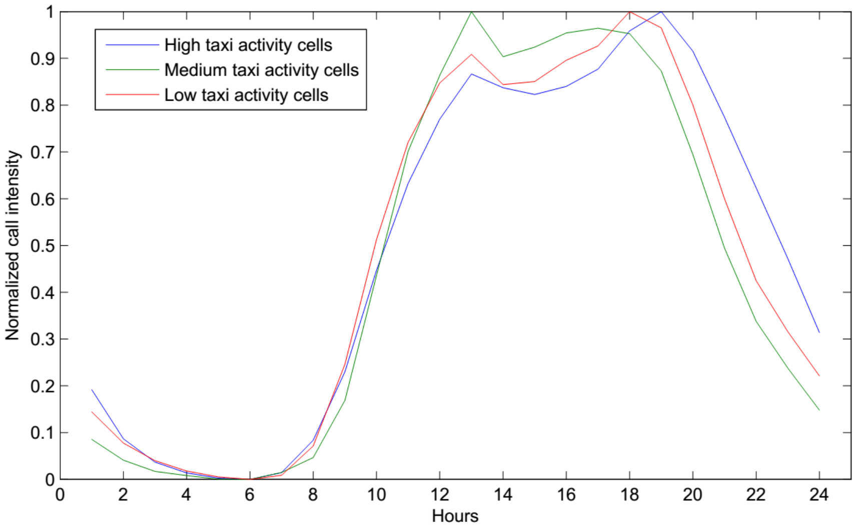


Figure 3-11 - Hourly variation of mobile phone call intensity on different taxi activity cells (high, medium and low taxy activity)

Furthermore, we use the classification for the predominant POI on each grid-cell (as described previously, in Figure 2-11, section 2.4.3) to analyze the call intensity in different areas of the city, as plotted in Figure 3-12. Call intensity maintains a fairly similar daily pattern regardless of the city area (cells where shopping is the most predominant POI deviate from the standard pattern by delaying the morning rising of the mobile phone activity).

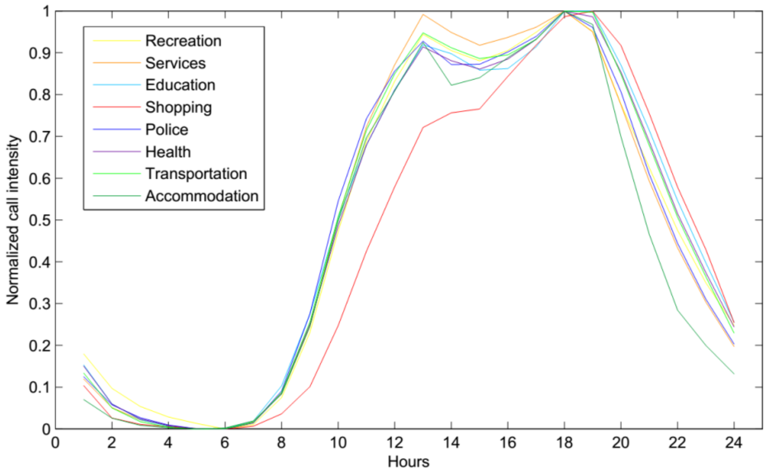


Figure 3-12 - Hourly variation of mobile phone call intensity in areas with different predominant POIs.

3.5.2 Correlation between mobile phone call intensity and taxi volume

By examining the daily temporal distributions of taxi volume and mobile phone call intensity as shown in Figure 3-13, we notice their similar patterns: both gradually increase in the morning around 7 AM, stay highly active, and then drop down slowly in the evening around 7 PM. In addition, we observe that mobile phone call intensity appears to follow taxi volume with an approximate gap of about 1-2 hours. Although Figure 3-13 represents the taxi volume and mobile phone call intensity for all Lisbon municipality, individual cells behave similarly, with minor deviations from the global pattern. The exception takes place on cells with low taxi activity where deviations from the global pattern can be accentuated.

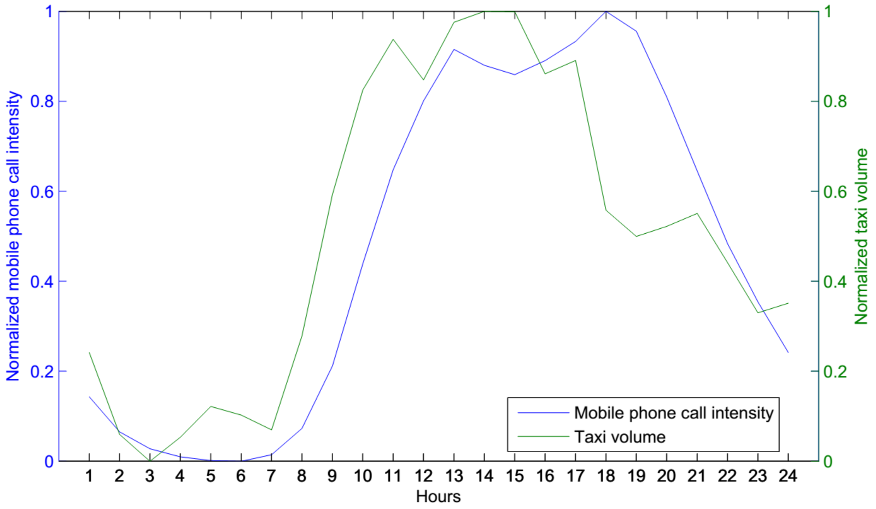


Figure 3-13 - Temporal distribution of mobile phone call intensity (blue) and taxi volume (green) across different hours of the day in Lisbon.

To further explore this relationship, we extract data as an hourly aggregated time series for all of Lisbon municipality. We define two variables: $G = \{g_1, g_2, \ldots, g_n\}$ represents the hourly time series of mobile phone call intensity, and $T = \{t_1, t_2, \ldots, t_n\}$ represents the hourly time series of taxi volume, both with length n (744 samples in total, corresponding to 24 daily samples during 31 days, in December). Since both variables have different units, the time series are normalized to the interval [0, 1], using the following equation:

$$z = \frac{x - min}{max - min} \qquad (14)$$

where *min* represents the minimum value of the time series and *max* the maximum value of the time series.

We overlay these time series on the same plot as shown in Figure 3-14 and observe similar temporal patterns. As observed before, both exhibit daily and weekly cycles. Mobile phone call intensity reaches almost zero (minimal activity) between midnight and 6 AM while high values appear around noon. Taxi volume time series appear to follow this pattern with low values emerging during off-peak hours (a short time after midnight up to early morning). A reduction of activity from both services is also observable on weekends (marked in grey) and on national holidays (December 1st, 8th, and 25th, marked in red).

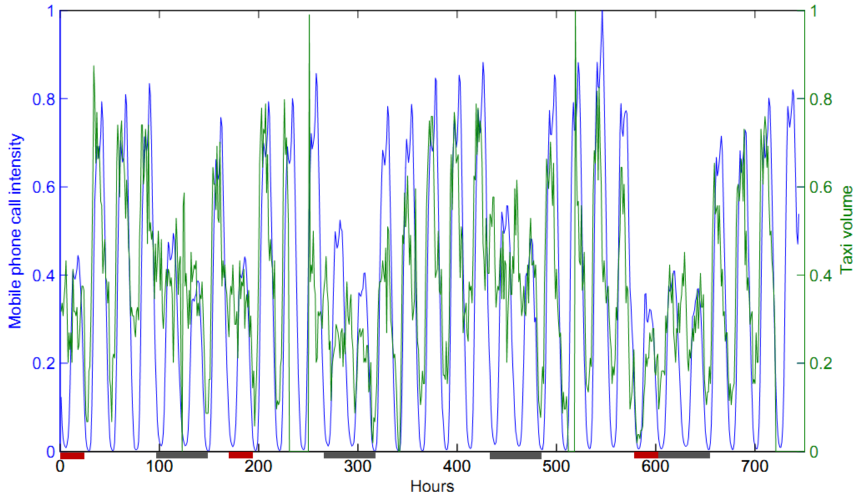


Figure 3-14 - Normalized time series of mobile phone call intensity (blue) and taxi volume (green) over 31 days of observation. The grey line on x-axis represents the weekend periods while the red line corresponds to national holidays (December 1st, 8th, and 25th).

Since both variables are normally distributed, we compute the *coefficient of correlation* of *Pearson* (*r*), which measures the strength of a linear relationship between normally distributes variables (Devore & Berk, 2012), defined as:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (15)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the arithmetic mean of variable *X* and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the arithmetic mean of variable *Y*, with $X = (x_1,..,x_n)$ and $Y = (y_1,..,y_n)$, two variables of size *n*.

By applying this coefficient we find out that mobile phone call intensity and taxi volume are highly correlated with a coefficient of correlation of *r* = 0.7559 (falls within the interval 0.7 to 0.89). Although this result considers data from all Lisbon municipality, it still holds true when individual grid cells are analyzed using the same procedure (summarized in Table 3-3).

	Coefficient of correlation of Pearson (r)
High taxi activity cells	0.7566
Medium taxi activity cells	0.7322
Low taxi activity cells	0.6293

Table 3-3 – Comparison of coefficients of correlation between mobile phone call intensity and taxi volume in different city areas.

On average, in grid cells with high or very high taxi activity (as defined in Figure 2-5) mobile phone call intensity correlates with taxi volume with *r* = 0.7566, in medium taxi activity cells the correlation is *r* = 0.7322 and in low taxi activity cells the correlation is *r* = 0.6293. However, for some cells with low taxi activity the process cannot be applied or presents a low correlation due to insufficient data. In these areas, taxi service is absent for some hours (especially at night) or there is a very low amount of taxi activity (less than 5 trips per hour).

3.6 Data analysis and results: monitoring the mobile phone activity through taxi traces

3.6.1 Quantifying the similarity between time series

To quantify the similarity between both mobile phone call intensity and taxi volume time series, for all Lisbon municipality, we compute the Euclidean distance (*ED*) as a measure of distance, as follows:

$$ED_i = \sqrt{(g_i - t_i)^2} = |g_i - t_i| \tag{16}$$

where g_i represents the mobile phone call intensity at hour *i* and t_i denotes taxi volume at hour *i*. Hence $G = \{g_1, g_2, \ldots, g_n\}$ and $T = \{t_1, t_2, \ldots, t_n\}$ represent respectively the normalized time series of mobile phone call intensity and taxi volume of length *n*.

Euclidean distance of these two time series turns out to be 5.6481 and its hourly distances are shown in Figure 3-15, where grey lines on x-axis represent the weekend periods. The smaller the Euclidean distance the higher similarity there is between the two time series.

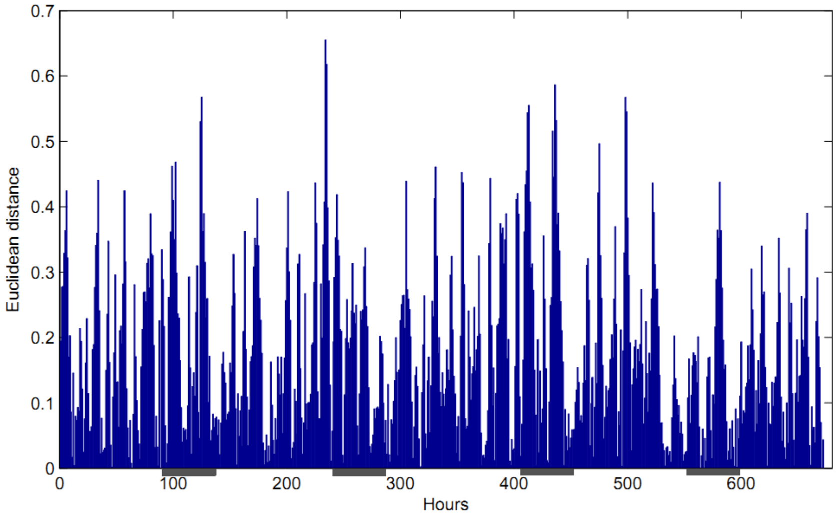


Figure 3-15 - Hourly Euclidean distance of the normalized time series of mobile phone call intensity and taxi volume. The grey line on x-axis represents the weekend periods.

Furthermore, we observe daily and weekly cycles. Through our examination of the data, we found that the highest similarity between these time series is from 8 AM to 10 PM (active hours) with the Euclidean distance of 4.4858. The hourly distance is shown in Figure 3-16.

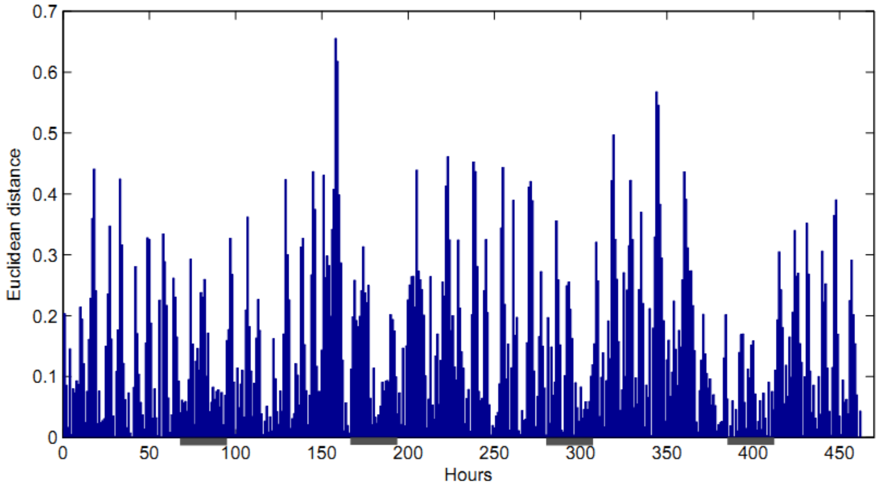


Figure 3-16 - Hourly Euclidean distance of the normalized time series of mobile phone call intensity and taxi volume from 8 AM to 11 PM for which the overall distance was found to be the lowest at 0.1917. The grey line on x-axis represents weekend periods.

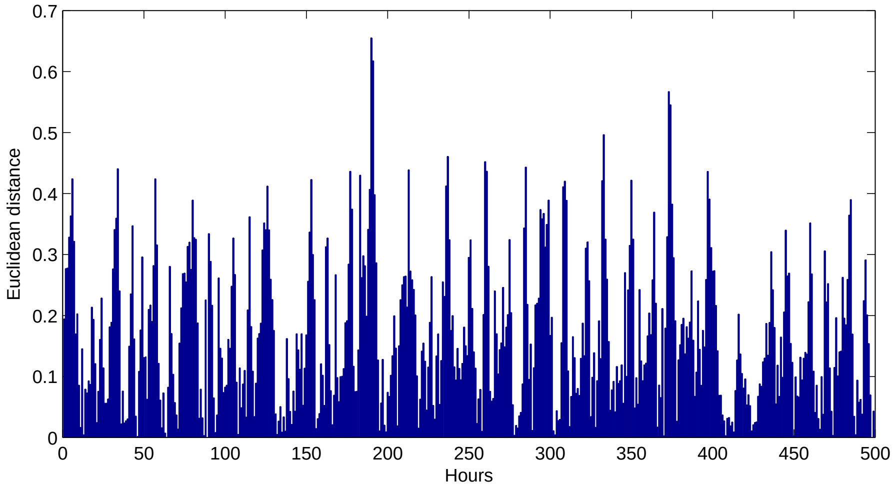


Figure 3-17 - Hourly Euclidean distance of the normalized time series of mobile phone call intensity and taxi volume during weekdays.

From a weekly cycle perspective, weekdays, which are associated with more activities (mostly repeated activities in temporal orders such as commuting to work, having lunch at the same time and at the same location, making a phone call before arriving at home, and so on) than weekends, unsurprisingly yield more correlated behaviors between mobile phone calls and the amount of taxis. The Euclidean distances are 4.7049 and 5.9581 respectively for weekdays and weekends. Figure 3-17 shows hourly Euclidean distance for weekdays.

3.6.2 Predictability between time series

We have so far observed that there is a correlation between the two time series, i.e., their values vary in a similar way, especially during active hours of the day (8 AM-10 PM) and active days of the week (weekdays). We want to investigate the predictability between them in more detail. More specifically, *can one data source be used to predict the other and to what extent?*

To do this, we apply a linear regression and employ the *coefficient of determination* (r^2 or *r*-squared, that is widely used for regression analysis) to measure the interdependency between these two urban data sources for different time shifts. The time shifting is used here to examine the predictability that one had on the other. For example, one-hour lag of *X* yields a high r^2 value with *Y* implies that *X* is likely a one-hour predictor of *Y*, i.e., the variation in values of *X* suggest a similar variation in values of *Y* of the next hour.

The coefficient of determination can be calculated as (Devore & Berk, 2012):

$$r^2 = \frac{\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \qquad (17)$$

where $\bar{y}$ is the mean and $\hat{y}$ denotes the predicted value of *y* (i.e., $\hat{y}_i = a + bx_i + \varepsilon_i$).

By fixing mobile phone time series and shifting taxi time series between -5 hours to +5 hours (e.g., -5 hours of time shift means considering mobile phone data at time *t* against taxi data at time *t*-5 hours), we discover that at time shift of -2 hours the two data sources have the highest correlation. As shown in Figure 3-18, for a time shift of -2 hours the Euclidean distance and coefficient of determination values are respectively 3.7431 and 0.7571. This suggests that generally the taxi volume is a 2-hour predictor of mobile phone intensity. In other words, *the variation in the amount of taxis is an indicative variable for the mobile phone call intensity for the next two hours.*

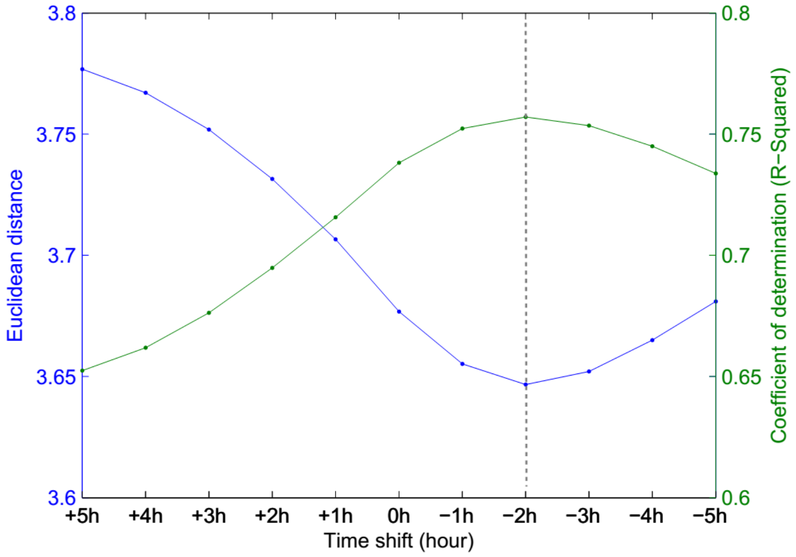


Figure 3-18 - Fitting results for the sliding window between GSM and taxi data.

The hourly Euclidean distance of this 2-hour difference comparison is shown in Figure 3-19. The plot of the normalized taxi volume against the normalized mobile phone call intensity is shown in Figure 3-20 along with the fitted linear function defined as:

$$y = p_1 x + p_2 = 0.5693x + 0.1437 \tag{18}$$

with a coefficient of determination (r^2) of 0.7571. The analysis of variance shows a *F*-statistic (ratio of the mean squared errors) of 13.85 and *p*-value of 0.0007 (highly significant), lower than the significance level (α) of 0.05.

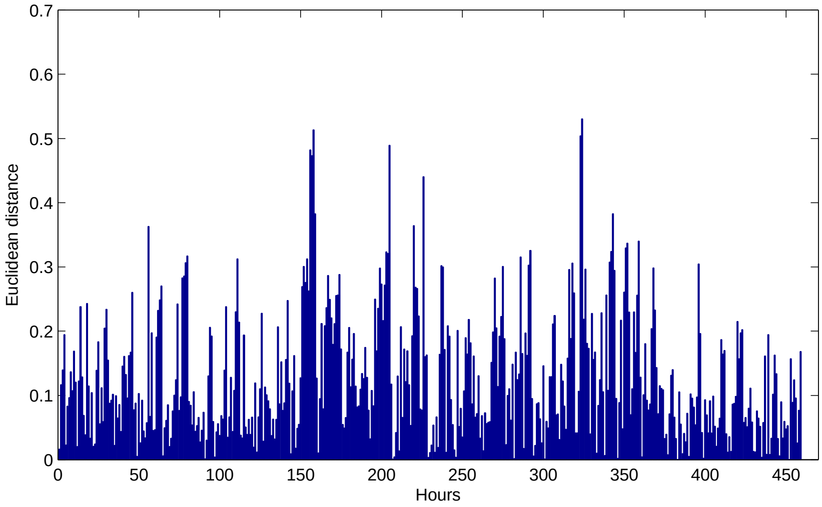


Figure 3-19 - Hourly Euclidean distance of the normalized time series of mobile phone call intensity and taxi volume of the time shift of -2 hours (i.e. comparing mobile phone data at time *t* with taxi data at time *t*-2).

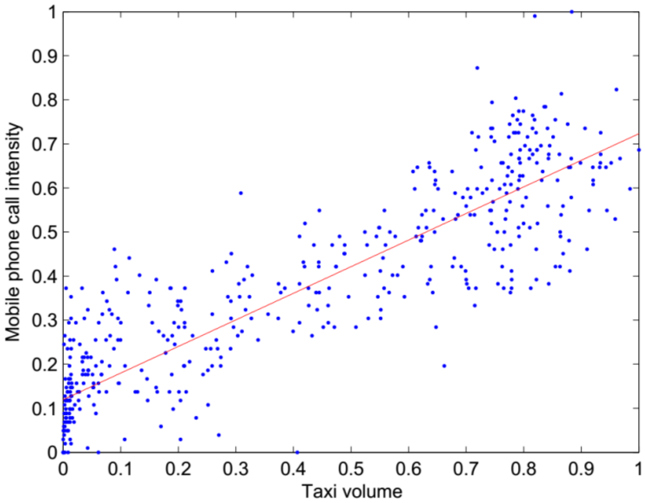


Figure 3-20 - Fitted linear function of the normalized taxi volume (at time *t*-2) against the normalized mobile phone call intensity (at time *t*).

As aforementioned, December 1st and 8th (2009) were national holidays, with a mobile phone activity distinct from the usual weekdays (instead resembling a weekend). During the analysis, we observed that these days contribute negatively for the model accuracy. The best coefficient of determination of 0.7571 previously presented is achieved by not taking into consideration the data from these two days. If we were to include data from these days in the model, the coefficient of determination would decrease to 0.7057. Interestingly, December 25th, also a national holiday on a weekday, did not significantly affect the model's accuracy. By removing data from December 25th the model improves slightly from r^2 = 0.7571 to r^2 = 0.7597.

3.6.3 Data selection to improve the model

During the exploratory analysis we observed that the mobile phone call intensity pattern was fairly consistent on all cell sites and all areas of Lisbon municipality. However, the pattern of taxi volume was less consistent on cells with low taxi activity. Although several factors can cause the low activity of taxi services (e.g. areas well served by public transportation, low income of the inhabitants in specific areas preventing the use of a more expensive mean of transport, areas outside the scope of the taxi company which provided the data), the low amount of data could influence the proximity of both time series. Therefore the same previous data analysis is performed, however discarding the data from the low taxi activity areas (considering only data from very high, high and medium taxi activity cells). By doing so, the relationship between mobile phone call intensity and taxi volume (at time shift of -2 hours, on weekdays and during business hours) is defined by the linear function:

$$y = p_1 x + p_2 = 0.599x - 0.1918 \tag{19}$$

with a coefficient of determination (r^2) equal to 0.8047 and a Euclidean distance of 3.1915, which is an improvement from the previous analysis, as well as a confirmation of the negative effect of the low taxi activity areas on the overall analysis. The analysis of variance shows a *F*-statistic (ratio of the mean squared errors) of 35.55 and *p*-value of 0.0009 (highly significant), lower than the significance level (α) of 0.05.

The hourly Euclidean distance between the two time series (comparing mobile phone data at time t with taxi data at time t-2) is plotted in Figure 3-21, while the normalized taxi volume against the normalized mobile phone call intensity is shown in Figure 3-22 along with the fitted linear function.

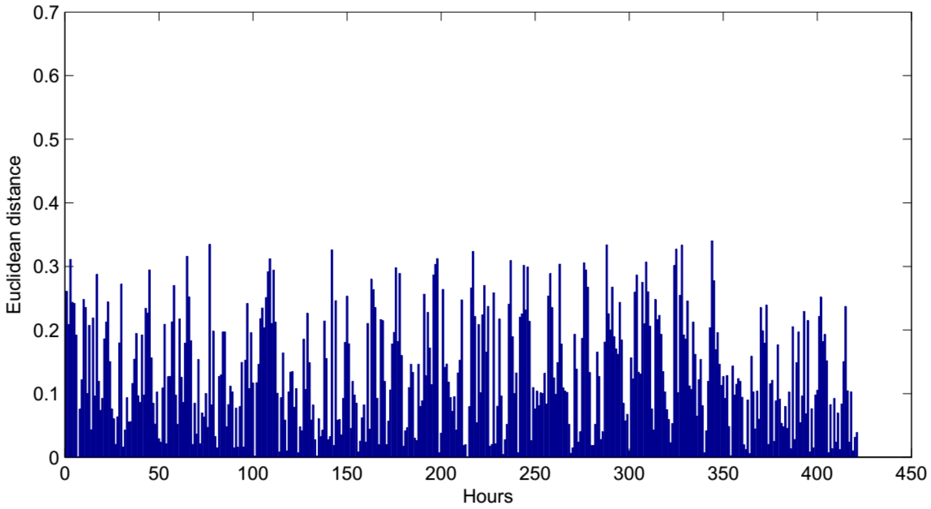


Figure 3-21 - Hourly Euclidean distance of the normalized time series of mobile phone call intensity and taxi volume of the time shift of -2 hours (i.e. comparing mobile phone data at time *t* with taxi data at time *t*-2) , excluding data from low taxi activity areas.

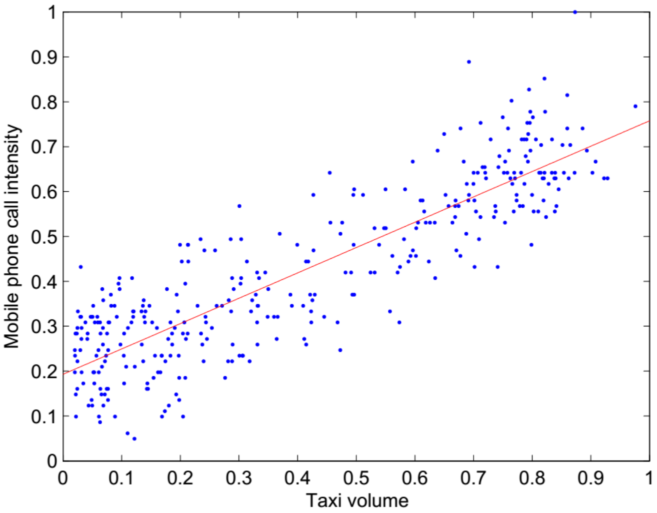


Figure 3-22 - Fitted linear function of the normalized taxi volume (at time *t*-2) against the normalized mobile phone call intensity (at time *t*), excluding data from low taxi activity areas.

Having observed significant correlations at active hours of the day and active days of the week, as well as a 2-hour time difference lead us to a further investigation on how this inter-predictability varies across different periods of the day.

3.6.4 Pursuing the best temporal fitting

Similarly to the previous approach, by keeping the normalized mobile phone time series fixed while shifting taxi time series between -5 and +5 hours, we compute the coefficient of determination values across varying time shifts for each different hour of the day. The result is shown in Figure 3-23 where this inter-predictability is observed to change over time. It turns out that there are strong inter-predictabilities (correlations) during active hours of the day, which are in line with our previous observations. Interestingly, we found that *during the active hours, mobile phone call intensity is a predictor for taxi volume in the AM hours and the relationship is reversed as the taxi volume becomes a predictor for mobile phone call intensity in the PM hours.* Hence at noon hours there is a strong correlation at 0 time shift. In other words, variations in both urban data sources are well synchronized around midday.

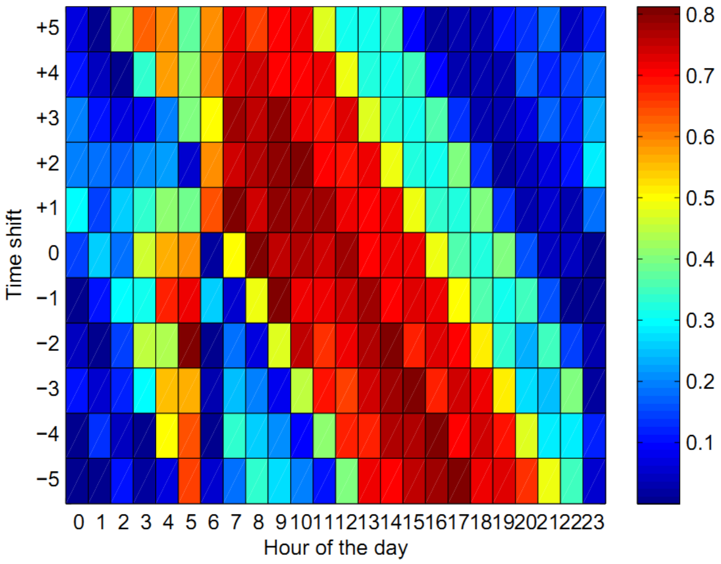


Figure 3-23 - Pseudocolor plot of coefficient of determination values across varying time shifts of different hours of the day.

3.6.5 Exploring the best indicator for mobile phone call intensity

During the exploratory analysis (Figure 3-3 replicated in this section as Figure 3-24) three indicators were presented: the average amount of call attempts, the average amount of successfully started voice calls, and the busy traffic hour. It was stated that for this study we analyze the busy traffic hour for each cell site, a widely used indicator of mobile phone activity, which we termed as *call intensity*. Thus, the previous results are achieved by using the busy traffic hour indicator. Seeing that all indicators behave similarly, showing the same daily and weekly patterns, we would like to explore which would be the best indicator for mobile phone activity in the relationship with taxi volume. Therefore, all previous procedures are repeated using the remaining two indicators for comparison.

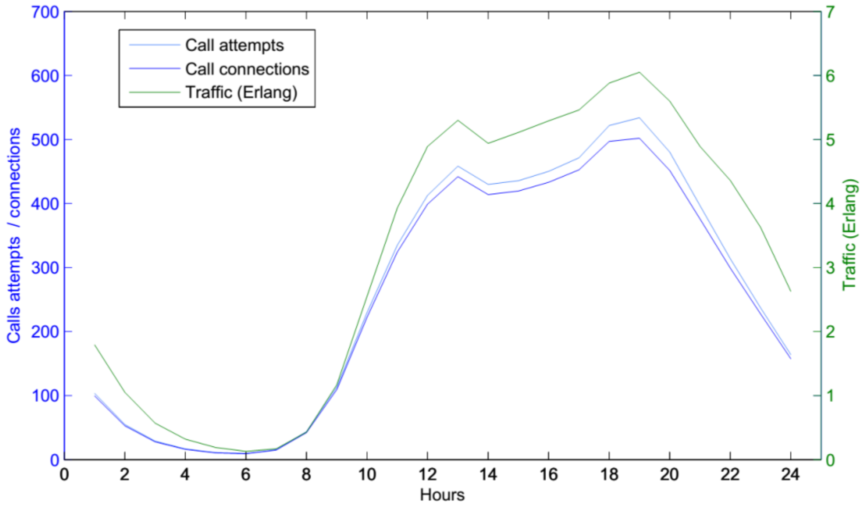


Figure 3-24 - Hourly variation of mobile phone call intensity in Lisbon (average of call attempts, successful calls and busy hour traffic).

Table 3-4 presents the coefficient of correlation between the three indicators. Each indicator was extracted as an averaged time series, aggregated by hour, of all cell sites from the mobile phone database. It is clear that all indicators are strongly correlated with each other, with a minimum value of linear association of 0.9850. As expected, the correlation between the average number of successfully started voice calls and the average number of call attempts is almost 1 (0.999998788), since the latter includes all attempts to start voice calls, both successful (the former) and unsuccessful.

	Average number of successfully started voice calls	**Average number of call attempts**	**Average busy hour traffic (*erlang*)**
Average number of successfully started voice calls	1	0.999998788	0.985021158
Average number of call attempts		1	0.985007737
Average busy hour traffic (*erlang*)			1

Table 3-4 – Coefficients of correlation between averaged time series, aggregated by hour, from three indicators for mobile phone activity.

Table 3-5 presents similar information as Table 3-4, however, instead of computing average values, the time series of the three indicators were extracted using the summation of values aggregated by hour. Similarly, a strong correlation is observed between all three indicators, with a minimum coefficient of correlation of 0.99465.

	Total number of successfully started voice calls	**Total number of call attempts**	**Total busy hour traffic (*erlang*)**
Total number of successfully started voice calls	1	0.999998064	0.994653286
Total number of call attempts		1	0.994646208
Total busy hour traffic (*erlang*)			1

Table 3-5 – Coefficients of correlation between time series using the summation of values aggregated by hour, from three indicators for mobile phone activity

Table 3-6 presents the coefficient of correlation between mobile phone activity and taxi volume, as explored in data analysis of previous sections, using different scenarios: initial state; during weekdays; during working hours; and the best correlation attained in both cases (during active hours on weekdays, with a 2-hour gap between taxi volume and mobile phone activity, without outliers). For the first row, the mobile phone activity is represented by the busy hour traffic time series, whilst for the second row the mobile phone activity is represented by the number of successfully started voice calls time series. The values were extracted aggregated by hour and represent the sum of all cell sites. No significant discrepancy is visible: the final (and best) correlation attained with both indicators differs only in 0.002667206 (0.897026 against 0.894359) and the maximum difference between both indicators is 0.010618897, observed during the weekday's analysis.

	Total busy hour traffic (*erlang*)	**Total number of successfully started voice calls**
Initial state	0.755874	0.750757
Weekdays	0.844133	0.833514
Working hours	0.761809	0.761593
Final (best)	0.897026	0.894359

Table 3-6 – Comparison of coefficients of correlation between taxi volume and mobile phone activity, in different scenarios. Mobile phone activity represented by total busy hour traffic and total number of successfully started voice calls.

The study presented in Table 3-7 computes de coefficient of correlation between taxi volume and mobile phone activity (represented by busy hour traffic). In the first row, the values for mobile phone activity were extracted aggregated by hour, and represent the sum of all cell sites, while for the second row the values for mobile phone activity were extracted aggregated by hour, and represent the average of all cell sites. The study analyses the linear association using different scenarios: all grid cells of Lisbon municipality; cells with high and very high taxi activity; cells with medium taxi activity; and cells with low taxi activity, in the initial state of the process (including weekends, night hours and outliers). Once again, no significant discrepancy is visible: the maximum difference is 0.001031515 between coefficients of correlation computed with data from low taxy activity cells.

	Total busy hour traffic (*erlang*)	
	Sum of all cell sites	**Average of all cell sites**
All cells	0.755874	0.755898
High and very high taxi activity cells	0.756638	0.756799
Medium taxi activity cells	0.732224	0.732127
Low taxi activity cells	0.620519	0.619558

Table 3-7 – Comparison of coefficients of correlation between time series with summation of values and averaged values.

This analysis showed that all three indicators are useful to represent mobile phone activity. No significant discrepancies were found in using any of the indicators to study the relation between the mobile phone activity and taxi volume in different scenarios. Moreover, there is no strong deviation in working with data as a summation of values or data as an average of values.

3.7 Chapter summary and conclusions

This section summarizes the work developed to understand the relationship between taxi volume and mobile phone activity. Main contributions and results are presented along with a discussion of the limitations of the study and future improvements of the work.

3.7.1 Overview and contributions

In this chapter, we explored the relationship between the taxi volume and mobile phone call intensity in Lisbon, Portugal. Particularly we were interested in the inter-predictability between these two urban data sources. Based on one-month of data (December), we found a strong correlation between the two time series during active hours of the day (8 AM-10 PM) and active days of the week (weekdays) in areas with medium and high taxi activity.

Moreover, we discovered that mobile phone call intensity has a significant correlation with taxi volume of the previous two hours, which means that the amount of taxis can be used to predict the intensity of mobile phone calls for the next two hours. Furthermore, we found that this inter-predictability could be modeled with a linear function. Intensity of mobile phone calls was a predictor of taxi volume in morning hours while the amount of taxi flow became a predictor of mobile phone calls in the afternoon and evening. These results have been published (Veloso, et al., 2012).

The exploratory analysis also showed a fairly regular pattern of the time series, especially for mobile phone call intensity. The use of mobile phone service appears to be consistent throughout the day and during the entire time series. Comparison between different urban areas showed only minor pattern deviations in mobile phone call intensity.

Finally, three indicators were identified to perfectly represent mobile phone activity. Busy hour traffic (in *erlangs*) and number of calls successfully started are among the main indicators used in research to represent mobile phone activity. Our study showed no significant discrepancy in the results produced using both indicators, in the analysis of the relation between taxi volume and mobile phone activity.

3.7.2 Limitations and future work

Nonetheless, there were a number of significant limitations to our study. The first of these is the limited amount of data used. Only one month of data was available to us at the time of this study, which limited our observations and results. Another potential limitation is the linear relation that was assumed between our two data sources in this study. Further investigation thus needs to be done to find the most suitable function for their relationship.

Despite the interesting results and the potential for improvement, the model only holds for specific conditions: working hours, weekdays, and cells with medium and high taxi activity. The study shows a considerable degradation of the model on weekends and at night. Additionally, the patterns on national holidays also weaken the model.

More evident is the low correlation between both time series on cells with low taxi activity, where the approach is not suitable. As expressed before, the mobile phone call intensity maintains a fairly regular pattern in all conditions. However, taxi activity shows a noticeable change in pattern, backed up by the significant decrease in correlation between both time series. Since the current taxi data set covers about 20% of the taxi service in Lisbon, a new and complete data set, with a longer temporal window, would be desirable and should be included in future work. It would support our claims of a significant correlation between mobile phone call intensity and taxi activity, as well as allow a better exploration of the patterns on the present low taxi activity cells.

A final limitation is related to the extent to which our findings are applicable beyond the city of Lisbon. The model was not tested on other urban areas, due to the lack of data available for the same temporal window for other regions. Although we believe that the findings are likely to be applicable to cities with broadly similar social, cultural, and economic profiles, the hypothesis was not tested. Thus, the geographic replicability can be disputed. Furthermore, although both time series show daily and weekly cyclic patterns, only one month of data is aligned in the same temporal window. A data set with a wider temporal window is needed to analyze the seasonal patterns throughout different months and weather conditions (e.g. summer patterns against winter patterns). A newer data set should be collected for the same urban region alongside with data set from different locations to validate the temporal and geographic replicability of the model.

As a closing remark, we believe that our findings to some extent, unveil the relationship between two different urban data sources; one describes sociality of the city while the other characterizes state of traffic flow. The findings are useful for developing efficient intelligent transportation systems as they provide the link between social and transportation networks. We hope that our findings suggest new ways to use multi-source data fusion to investigate the interplay between different urban entities. In the next chapter, the relation between urban data sources will be further explored, studying the relation between taxi volume, weather conditions and exhaust gases’ concentrations.

Chapter 3
The relationship between mobile phone activity and taxi traces

Chapter 4
Monitoring exhaust gases concentrations using taxi traces and meteorological conditions

4.1 Introduction

Urban areas are experiencing a fast growth in size and population, which demands more resources, namely improved mobility. As a non-desirable side effect of this growth, air pollution in cities is increasing due to anthropogenic emissions (Velasco & Roth, 2010). Traffic is one of the major sources of toxic compounds which are part of combustion gases that negatively impact the health of city inhabitants (Zavala, et al., 2006), (Karlsson, 2004), (Becker, et al., 2000). Understanding gas emission patterns and the ability to estimate their concentrations in urban areas are, thus essential in order to mitigate the problem.

The existing infrastructures are currently not adequately prepared for this emerging issue. Even in major cities, only a small number of monitoring stations is deployed for exhaust gas concentration, leaving vast areas of urban populations without scrutiny. With this growing urbanization, it has become more difficult and costly to measure and monitor exhaust gas concentration levels for the entire metropolitan area. Current urban projects are focused on complying with the demand for new transportation, dwelling, and energy, but there is still a lack of investment in monitoring systems. Therefore, there is an emerging need for solutions that allow monitoring the environment on a large scale, with improved spatial resolution, and at lower cost.

At the same time, the awareness of environmental and health issues is rising, prompted by recent governmental acts, scientific research, and concerns from individual citizens. Concerns about global warming, the increase in heatwaves, and the increase of toxic substances in the atmosphere is triggering alarms in society, which should be addressed.

Nowadays, following the developments in pervasive and ubiquitous computing technologies, taxis in various cities are equipped with GPS to improve their services with a better dispatching system. Using this inherent ubiquity, their traces have been used to study different aspects of the traffic network, as they provide data that reflects the state of traffic flow and can be used as a probe for traffic conditions (Castro, et al., 2012), (Gühnemann, et al., 2004).

By taking an opportunistic sensing approach, we use taxi-GPS traces collected in Lisbon, Portugal, to explore the ubiquitous data streams produced by taxi mobility patterns and meteorological information, to model the level of concentration of exhaust gases, focusing on nitrogen dioxide concentrations, and to improve spatial resolution of gas monitoring. This work makes the following contributions:

1. Analysis of the temporal and spatial characteristics of exhaust gas concentration, along with the effect of meteorological conditions (i.e., weather conditions, humidity, temperature and wind speed).

2. Exploration of techniques to estimate the concentration level of nitrogen dioxide without prior or historic information on exhaust gases' concentrations, taking into account other urban variables, namely taxi data and meteorological conditions.

For the former, we analyze a historical data set of exhaust gas concentration to identify patterns in time and space - comparing behaviors between traffic and background monitoring stations. The effect of different meteorological conditions on the concentrations of exhaust gases is also explored. For the latter, we study multi-linear regression and an artificial neural network to model the concentrations of nitrogen dioxide, using taxi activity (pick-ups, drop-offs and average speed), weather conditions, humidity, wind speed and temperature as predictors.

4.2 State of the art

Mining taxi trajectories has recently attracted much attention. As described in Chapter 2, Taxi-GPS traces have been used in a number of studies to develop better solutions and services in urban areas such as estimating optimal driving paths (Yuan, et al., 2010), (Zheng, et al., 2010), and (Ziebart, et al., 2008), predicting next taxi pick-up locations (Yuan, et al., 2011b), (Phithakkitnukoon, et al., 2010b), (Liu, et al., 2010b), and (Ge, et al., 2010), modeling driving strategies to improve taxis' profits (Ge, et al., 2010), and (Liu, et al., 2010a), identifying flaws and possible improvements in urban planning (Zheng, et al., 2011b), and developing models for urban mobility, social functions, and dynamics between different areas in the city (Qi, et al., 2011).

In addition to the dynamic in vehicular network, there is an extensive and broader list of studies focusing on the study of atmospheric pollutants' fluxes and the development of dispersion models:

- *Develop approaches to capture and monitor exhaust gases' concentrations* (Bukowiecki, et al., 2002), (Gühnemann, et al., 2004), (Pirjola, et al., 2004), (Velasco, et al., 2005), (Westerdahl, et al., 2005), (Zavala, et al., 2006), (Nemitz, et al., 2007), (Mak & Hung, 2008), (Schmidt, et al., 2008), (Zavala, et al., 2008), (Hu, et al., 2009b), (Parshall, et al., 2009), (Huang, et al., 2010), (Fuller, et al., 2012), (Hu, et al., 2012c), (Liu, et al., 2012), (Mao, et al., 2012), (Padró-Martínez, et al., 2012), (Carslaw & Rhys-Tyler, 2013), (Kousoulidou, et al., 2013), (Pattinson, et al., 2014), (Riley, et al., 2014), (Kumar, et al., 2015), (Moltchanov, et al., 2015);

- *Study models to forecast exhaust gases' concentrations* (Shi & Harrison, 1997), (Gardner & Dorling, 1998), (Gardner & Dorling, 1999), (Cobourn, et al., 2000), (Liley, et al., 2000), (Perez, et al., 2000), (Cogliani, 2001), (Grambsch, 2001), (Kolehmainen, et al., 2001), (Perez & Reyes, 2001), (Perez & Trier, 2001), (Dorling, et al., 2003), (Kukkonen, et al., 2003), (Schlink, et al., 2003), (Hooyberghs, et al., 2005), (Agirre-Basurko, et al., 2006), (Grivas & Chaloulakou, 2006), (Holmes & Morawska, 2006), (Kassomenos, et al., 2006), (Zou, et al., 2006), (Chattopadhyay & Bandyopadhyay, 2007), (Dutot, et al., 2007), (Venkatram, et al., 2007), (Faus-Kessler, et al., 2008), (Juhos, et al., 2008), (Saini, et al., 2008), (Agirre, et al., 2009), (Mukerjee, et al., 2009), (Su, et al., 2009), (Venkatram, et al., 2009), (Beelen, et al., 2010), (Johnson, et al., 2010), (Sfetsos & Vlachogiannis, 2010), (Allen, et al., 2011), (Csikos & Varga, 2011), (Madsen, et al., 2011), (Zwack, et al., 2011), (Merbitz, et al., 2012), (Misra, et al., 2013), (Russo & Soares, 2013), (Donnelly, et al., 2015);

- *Study the sources of atmospheric pollutants* (Cline, 1991), (Saville, 1993), (Small & Kazimi, 1995), (Becker, et al., 1999), (Becker, et al., 2000), (Borrego, et al., 2000), (Gilbert, et al., 2003), (Karlsson, 2004), (Pleijel, et al., 2004), (Kirchner, et al., 2005), (Ndoke & Jimoh, 2005), (Velasco, et al., 2005), (Yli-Tuomi, et al., 2005), (Borrego, et al., 2006), (Zavala, et al., 2006), (Gilbert, et al., 2007), (Zhou & Levy, 2007), (Beckerman, et al., 2008), (Matese, et al., 2009), (Oliveira, et al., 2010), (Smit, et al., 2010), (Velasco & Roth, 2010), (Donnelly, et al., 2011), (Ning, et al., 2012), (Pirjola, et al., 2012), (Dons, et al., 2013), (Franco, et al., 2013), (Ke, et al., 2013);

- *Analyze the profile and fluxes of exhaust gases and their effects* (Derwent, et al., 1995), (Uno, et al., 1996), (Hargreaves, et al., 2000), (Clapp & Jenkin, 2001), (Huang, et al., 2001), (Kodama, et al., 2002), (Tashiro & Taniyama, 2002), (Soegaard & Møller-Jensen, 2003), (Moriwaki & Kanda, 2004), (Krzyzanowski & Schneider, 2005), (Velasco, et al., 2005), (Pirjola, et al., 2006), (Vogt, et al., 2006), (Coutts, et al., 2007), (George, et al., 2007), (Matthes, et al., 2007), (Ojolo, et al., 2007), (Churkina, 2008), (Fruin, et al., 2008), (Pandey, et al., 2008), (Vesala, et al., 2008), (Carvalho, 2009), (Hu, et al., 2009c), (Westerdahl, et al., 2009), (Fontes, 2010), (Kordowski & Kuttler, 2010), (Hiller, et al., 2011), (Lin, et al., 2011), (Shon, et al., 2011), (Vann, 2011), (Crawford & Christen, 2012), (Gordon, et al., 2012), (Grimmond, et al., 2012), (Mavroidis & Ilia, 2012), (Melkonyan & Kuttler, 2012), (Baldauf, et al., 2013), (Barros, et al., 2013), (Kumar & Imam, 2013), (Venkatram, et al., 2013a), (Venkatram, et al., 2013b), (Wang, et al., 2013), (Patton, et al., 2014);

- *Simulation and spatial interpolation of atmospheric pollutants* (Borrego, et al., 2000), (Borrego, et al., 2001), (Borrego, et al., 2003), (Monteiro, et al., 2005), (Berkowicz, et al., 2006), (Monteiro, et al., 2007), (Stein, et al., 2007), (Pernigotti, et al., 2012), (Kota, et al., 2013).

Studies around atmospheric pollutants are extensive and well reported. Various fields of research have been created, from which we emphasize the following: measurement techniques and analysis of atmospheric pollutants; study of models to estimate exhaust gases' concentrations; and analysis of sources of atmospheric pollutants. Representative research works are summarized in the following sections.

4.2.1 Measurement techniques and analysis of atmospheric pollutants

Atmospheric pollutants are well characterized. The two daily peaks, where pollutants' concentrations increase in the atmosphere (morning and afternoon), have been observed and described by different studies, and linked to traffic commuting, alongside the patterns of atmospheric pollutants throughout the seasons. Long-term time series analysis have shown a reduction of NO (nitrogen monoxide) and NO_2 (nitrogen dioxide), but at different rates, associated, in some cases, with the increase of O_3 (ozone) due the rise of the average temperature. This long-term decrease is also linked to traffic sources, especially to the introduction of catalytic filters in the late 1990s, since traffic is considered to be one of the main sources of atmospheric pollutants. Wind is often observed as a major factor in atmospheric pollutants dispersion, while building disposition and the placement of green areas can also affect the patterns of pollutants.

Pirjola et al. (Pirjola, et al., 2006) studied traffic particles and pollutants (NO, NO_2 and CO), collecting data in the vicinity of a major highway in Finland, during summer and winter. The analysis of pollutant concentrations took into consideration air temperature and wind. Moreover, the authors analyzed the impact of wind with measurements from three sectors: wind perpendicular to the road; the opposite direction; and wind blowing parallel to the road. These observations indicated that wind is a major factor dispersing atmospheric pollutants. Seasonal analysis showed two to three times higher concentrations in winter than in the summer. Moreover, particles in the summer were smaller than 50nm while atmospheric pollutants' concentrations decreased around 35%, when sensors were located 65m from the roadside. To complement the work, a similar experiment was developed by Pirjola et al. (Pirjola, et al., 2012) in a dense urban area. The authors collected data from a road with intense traffic, in the city center of Helsinki, Finland. Concentrations of particles and pollutants (NO, NO_2 and CO) were collected, alongside meteorological data, although it was reported that weather conditions were stable. The authors studied the effect of surrounding buildings and wind fluxes in the dispersion of pollutants. The topography of the street led to a configuration where the upwind concentrations were higher on the sidewalk than on the road, affecting mostly pedestrians rather than drivers. It was reported that if buildings were built parallel to the street, the canyon effect would help better disperse atmospheric pollutants, dropping concentrations at ground-level. However, the geometry of nearby streets appears also to affect the local atmospheric pollutants' dispersion. The authors concluded that the surrounding built environment significantly affects the pollutant concentrations in urban areas, and should be taken into account for future urban designs.

Merbitz et al. (Merbitz, et al., 2012) studied the spatial variability of particulate matter (PM) in urban areas to produce a statistical model for spatial mapping of urban PM_{10} (particulate matter with dimension smaller than 10 µm) and $PM_{2.5}$ (particulate matter with a dimension smaller than 2.5 µm) concentrations. The authors performed mobile measurements in five distinct periods - from summer 2009 to spring 2011 - from 59 monitoring sites in Aachen, Germany. The authors observed a large spatial variability on a scale of tens of meters, mainly depending on traffic density and building structure, especially in inner city environments. Suburban concentrations exhibit the smallest spatial variability. Additionally, a regression model was used to estimate concentrations for the remaining urban areas. As input variables, the authors used an exponential particulate matter concentration from traffic emissions, building density, and green areas. The model showed a coefficient of determination of 0.79, with a tendency for underestimation (due to non-traffic sources), with better performance estimating PM_{10} than $PM_{2.5}$, especially in inner city areas. The authors argued that, the latter observation is due to the fact that coarse particles are more directly linked to local sources such as traffic, while $PM_{2.5}$ are more dependent on regional influences. Similar to Pirjola et al. (Pirjola, et al., 2012), the authors concluded that buildings in close vicinity have an impact on the dispersion of atmospheric pollutants, having a positive correlation with PM concentrations. Inversely, green areas present a negative correlation with PM concentrations, as an indication of the filtering effect of vegetation, removing particles by dry deposition.

Mavroidis & Ilia (Mavroidis & Ilia, 2012) analyzed the long-term trends of NO_2 and the ratio between NO_2 and NO_x (nitrogen oxides). Using data collected in Athens, Greece, from 1987 to 2008, the authors observed a decrease in NO_2 concentrations, but, at a slower rate than NO_x. The decrease of NO_2 concentrations is attributed to the increased use of three-way catalytic converters in gasoline fuelled vehicles in Greece, combined with the substitution of older vehicles, especially from the late 1990s. The authors attribute the slow decrease rate of NO_2 to the increase of secondary formation of NO_2 through photochemical oxidant reactions in the atmosphere. They have also explored the primary NO_2 concentrations, concluding that primary NO_2 concentration share has not altered significantly between 1998 and 2006. This is mainly attributed to the fact that in Athens, diesel passenger cars are not allowed and particle after-treatment technologies are not yet used in Greece. Melkonyan & Kuttler (Melkonyan & Kuttler, 2012) also analyzed long-term time series of NO, NO_2 and O_3. In this study, data was collected in Rhine-Westphalia, Germany, through six sensors (one industrial, two traffic, one urban and two rural) from 1981 to 2007. Similar to Mavroidis & Ilia, the authors observed a moderate decrease of NO_2 (10%), but much inferior to the decrease of NO (65%) in the same period. The same explanation was argued: catalytic filters in vehicles support emissions of NO_2 as a primary pollutant, especially from 1998 onwards. The increase of O_3 (20%, an indicator of summer smog) is related with changes in NO/NO_2 ratio, as observed by Mavroidis & Ilia, and due to the increase of average temperature. Daily and weekly patterns of NO, NO_2 and O_3 were analyzed, showing the traditional two daily peaks, with a steady reduction of maximum concentrations values from 1990 to 2007. This reduction is more visible in traffic stations than in background stations, confirming the idea that traffic is one of the main sources of atmospheric pollution.

To develop new approaches to capture and monitor exhaust gases' concentrations, Zavala et al. (Zavala, et al., 2006) used a mobile laboratory to measure on-road vehicle emission ratios in Mexico City. The authors observed that exhaust gases' emissions were strongly related to driving behaviors. Similarly, Velasco et al. (Velasco, et al., 2005) used an eddy covariance flux system to obtain direct measurements of CO_2 emissions in Mexico City. The analysis showed a clear diurnal pattern with the highest emissions during the morning and the lowest emissions during nighttime. The measured CO_2 fluxes were closely correlated to traffic patterns. Liu et al. (Liu, et al., 2012) applied a similar methodology to the city of Beijing, China, collecting data during a four-year period, with similar results. Daily and weekly cycles were observed, with a strong dependency with road traffic.

Allen et al. (Allen, et al., 2011) evaluated the geographical replicability of Land Use Regression models. The authors collected data from two similar cities in Canada, for 14 days, using identical geographic data sources and methods for site selection, data collection, and model development. The authors observed that the transferred models did not perform well as locally calibrated models. Moreover, better results were achieved when transferring NO_2 models (coefficient of determination 0.37-0.52) than with NO (coefficient of determination 0.24-0.41). Locally, the model was able to better explain the variations of NO_2 (0.81-0.84) than NO (0.55-0.56).

Donnelly et al. (Donnelly, et al., 2011) studied the effect of wind direction and wind speed in background concentrations of nitrogen dioxide (NO_2), using a non-parametric kernel regression method to quantify the effects. The authors verified that background NO_2 concentrations are significantly influenced by local winds and by sources located far apart. Venkatram et al. (Venkatram, et al., 2013a) also analyzed the wind impact in near-road pollutant concentrations. The authors observed an increase of pollutant concentrations with wind perpendicular to the road, at a distance of 100 m from the roadside. Therefore, the authors noted that pollutants are sensitive to wind speed and direction. Maximum concentrations occurred for short-life pollutants.

4.2.2 Models for estimating exhaust gases' concentrations

In order to forecast models of exhaust gases' concentrations, an extensive list of studies has been undertaken. Most of the research relied on Artificial Neural Networks (ANN), fed with historical data of gases' concentrations. Authors concluded that, generally ANN models perform better than statistical linear models. Additionally, nitrogen oxides (e.g. NO_2 and NO) predictions present a higher approximation with the measured data than particle predictions (e.g. $PM_{2.5}$ and PM_{10}) (Kukkonen, et al., 2003), (Juhos, et al., 2008).

Although first approaches relied on regression analysis - as studied by Shi & Harrison (Shi & Harrison, 1997) and Derwent et al. (Derwent, et al., 1995), Gardner et al. (Gardner & Dorling, 1999) and Shi et al. (Shi & Harrison, 1997) - were among the early authors to explore the use of ANN to build models for atmospheric studies. Gardner et al. (Gardner & Dorling, 1998) made an extensive study of the applicability of ANN (multilayer perceptron) to atmospheric studies. Later, the same authors explored the technique to train a model and estimate hourly concentrations of NO_x and CO_2 in Central London. The study showed the higher efficiency of ANN approaches against regression analysis (higher coefficient of correlation), stressing the ability of ANN models to solve complex patterns of source emissions (Gardner & Dorling, 1999). This result corroborated the work of Shi et al. (Shi & Harrison, 1997), which performed a similar experiment with data also collected in London.

Kolehmainen et al. (Kolehmainen, et al., 2001), using data from 1994-1998, collected in Stockholm, Finland, explored periodic regression and neural algorithms (self-organizing maps). The author concluded that the best results to predict NO_2 concentrations were attained by applying multi-layer perceptron to the original data. Kukkonen et al. (Kukkonen, et al., 2003) further explored the previous approach. The authors compared neural network models, a linear statistical model and a deterministic modelling system to predict NO_2 and PM_{10}, with data collected in two stations in Helsinki, Finland, from 1996 to 1999. Using three statistical criteria (index of agreement, squared coefficient of correlation and fractional bias), the authors concluded that results obtained with non-linear neural networks achieved a better approximation with the measured data. Both authors suggested that the inclusion of meteorological variables could improve the performance of the model, since it could describe the instability of the atmosphere.

Later, Niska et al. (Niskaa, et al., 2004) improved the work of Kolehmainen et al. (Kolehmainen, et al., 2001) by studying the use of genetic algorithms for selecting the inputs of a multi-layer perceptron. This approach was earlier used by Grivas & Chaloulakou (Grivas & Chaloulakou, 2006), which applied an ANN to provide predictions of PM_{10}, with a genetic algorithm for selecting the input variables. The work was based upon a data set collected between 2001 and 2002, in four measurement locations in the Greater Athens Area. In line with other works, the authors concluded that neural network models were superior in comparison to multiple linear regression models.

Agirre-Basurko et al. (Agirre-Basurko, et al., 2006) presented a model to estimate concentrations of ozone and nitrogen dioxide eight hours ahead, using historical data of the gases along with meteorological variables. The authors compared two multilayer perceptron and multiple linear regression models, observing a higher performance for the multilayer perceptron-based models over the multiple linear regression models. Based on these findings, the authors proposed the *airEsan* software to forecast ozone concentrations and assess the air quality (Agirre, et al., 2009). Further validation and tuning of the parameters was performed by (Agirre, et al., 2010).

Perez et al. (Perez, et al., 2000) (Perez & Reyes, 2001) compared three different methods: multilayer neural networks, linear regression, and persistence to model air particles from May to September (months with higher concentration) in 1994 and 1995, in Santiago, Chile, to conclude that neural networks provided the best results. Higher significance and smaller errors were attained in prediction up to six hours in advance. However, the authors pointed out the need for noise reduction prior to modelling. Later, the authors performed a similar experiment, in the same location, aiming to model NO concentrations with identical results for all monitoring stations (Perez & Trier, 2001).

Juhos et al. (Juhos, et al., 2008) compared multilayer perceptron models with support vector machine models to predict NO and NO_2 concentrations, four days in advance. The data set was composed by historical data of NO and NO_2 with meteorological variables (namely temperature, humidity and wind speed). Similar to previous studies, the authors observed a higher performance of ANN models. Kassomenos et al. (Kassomenos, et al., 2006) compared the COPERT methodology against ANN to predict five major pollutants (CO, Benzene, NO_x, PM_{10} and Volatile Organic Compounds or VOCs) in Athens, Greece, with data collected from seven major roads. Although no major difference in the results produced by the two techniques was found, the authors verified that the determinant parameters for CO emissions variation in a road, are traffic volume and traffic speed; for benzene and VOCs, the presence of motorcycles and passenger vehicles not equipped with catalytic converters in combination with traffic speed; and for PM_{10} and NO_x the percentage of diesel vehicles in the fleet.

Cogliani (Cogliani, 2001) studied the correlation between meteorological variables and air pollution. The author used an atmospheric pollution index provided by the Italian Ministries of Health and Environment and explored the relation with meteorological indicators using multiple linear correlation methods. The author discovered that, using an historic index of atmospheric pollution, the concentration of pollutants for the previous day, the day’s lowest temperature, and the forecast of the next day's highest temperature and wind speed, it was possible to forecast the day's air pollution. Highest correlations were achieved during a three month period (January to March).

The city under analysis was already the focus of several studies to estimate or simulate air quality and the dispersion of pollutants (Russo & Soares, 2013), (Borrego, et al., 2000), (Borrego, et al., 2003), and (Monteiro, et al., 2005).

Russo & Soares (Russo & Soares, 2013) aimed to forecast pollutant concentrations with high accuracy in time and space, using meteorological variables and pollution data. The authors propose a two-step methodology: firstly the use of a neural network to generate short-term temporal forecasts, and secondly, spatial stochastic simulations were performed for all area of the city of Lisbon.

Two dispersion models were studied by Borrego et al. (Borrego, et al., 2003) to assess the air pollution (carbon monoxide) in downtown Lisbon. The authors used TREM (Transport Emission model for Line Sources) to estimate traffic emissions and VADIS (Local Scale Dispersion Model, a computational fluid dynamic model for pollutants dispersion) to simulate the emissions flow and dispersion around obstacles, under variable wind conditions. Montero et al. (Monteiro, et al., 2005) (Monteiro, et al., 2007) examined the performance of the CHIMERE photochemical model, to simulate ozone and nitrogen dioxide in Portugal. The authors showed a decrease in errors and increase in correlation when the sum of photo oxidants were considered, instead of individual pollutants, pointing out the lack of monitoring stations as one of the possible causes for the model errors.

4.2.3 Analysis of sources of exhaust gases

Although atmospheric pollutants can have natural sources (due to microbial metabolism in the soil, chlorophyll decomposition, electrical discharges in the atmosphere from lightning, forest fires or volcanic eruptions) they also have anthropogenic sources (fuel burning at high temperatures, from domestic heating or internal combustion vehicles). Traffic is one of the main sources of pollutant in urban areas, responsible for 32%-98% of emissions of CO, volatile organic compounds (primarily hydrocarbons) and NO_x (Small & Kazimi, 1995), (Ndoke & Jimoh, 2005), and an important fraction of greenhouse gases emissions in USA (especially CO_2) (Cline, 1991). Nearly 50% of global CO, hydrocarbon, and NO_x emissions from fossil fuel combustion come from gasoline and diesel engines (Ndoke & Jimoh, 2005). Additionally, on highly congested streets, traffic can be responsible for as much as 90%–95% of the ambient CO levels, 80–90% of the NO_x and hydrocarbons, and a large portion of particulate matter (Saville, 1993). Therefore, traffic emissions are the focus of several studies, performing measurements on the roadsides of major traffic highways.

Becker et al. (Becker, et al., 1999) compared real world emissions of NO_2 and CO_2 from traffic and laboratory measurements, concluding that emissions from both scenarios are quite similar. To implement the experiments, the authors collected data from a road tunnel in Germany (Wuppertal tunnel) as real world emissions, and performed measurements using a chassis dynamometer emission in the laboratory, testing 26 different cars and trucks. Later, using data collected from the tunnel, the authors were able to extrapolate the results and calculate the global N20 emissions in the city. They achieved those results by assuming that the vehicle mix travelling through the tunnel was representative of the global vehicle population, and multiplied the measured emission factor by the global vehicle fuel consumption (Becker, et al., 2000).

Venkatram et al (Venkatram, et al., 2007) used a dispersion model to analyze and estimate the impact of traffic emissions, measured at distances of tens of meters from an eight-lane highway. Air quality measurements consisted of optical measurements of NO at distances of 7 and 17 m. Additionally, sonic anemometers were used to measure wind speed and turbulent velocities at 5 m and 20 m from the highway. The authors observed that NO concentrations near the road were governed by the emission rate, as represented by the traffic flow rate (monitored using traffic surveillance cameras). The effect of wind was also analyzed, concluding that the concentrations were relatively insensitive to the mean wind speed, except at distances from the roads that are comparable to the width of the road. Moreover, as long as the wind direction was within 45º from the normal to the road, the wind direction had little effect on near road concentrations. Venkatram et al. (Venkatram, et al., 2009) improved the analysis, exploring the AERMOD dispersion model to interpret concentrations of volatile organic compounds. Later, the authors further explored the effect of wind in the dispersion of atmospheric pollutants (Venkatram, et al., 2013b). The authors observed that light wind does not fully follow traditional models for vertical plume spread. Under light wind, stable, or transition periods, the boundary layer has a significant impact on near-road concentrations associated with roadway emissions.

Beckerman et al. (Beckerman, et al., 2008) observed that levels of NO_2 decay with increasing distance from the expressway, declining to background levels by 300 m. Additionally, the authors also observed moderate to high correlations between NO_2 measurements and NO_x and O_3. The authors stated that the variability of many traffic-related pollutants around an expressway could be characterized by measurements of NO_2. Moreover, experimental results showed that the distance decay gradients display differential characteristics between the upwind and the downwind sides of the expressway. On the upwind side, levels drop off to near background levels within 200 m and in the case of particles probably less than 100 m. On the downwind side, levels do not reach background until 300–500 m. These results are consistent with the observations of Zhow & Levy (Zhou & Levy, 2007), which analyzed the spatial extent of atmospheric pollutant concentrations, and the influence of local meteorology in the dispersion of pollutants. The authors observed that pollutants with higher background concentrations had the largest spatial extent, and pollutants formed in near-source chemical reactions (NO_2) had a larger spatial extent than pollutants depleted in near-source chemical reactions.

Pleijel et al. (Pleijel, et al., 2004), Gilbert et al. (Gilbert, et al., 2003), and Zou et al. (Zou, et al., 2006) aimed to define a mathematical formulation to describe the relation between atmospheric pollutants and the distance from highways, assumed to be the pollution sources. The former defined a model stating that NO_2 concentrations decreased significantly with the increasing logarithmic distance from the highway. Experimental results support the hypothesis with high correlation between observed and estimated values up to 10 m from the highway. The authors also noted that the regression slope is likely to be sensitive to wind speed, atmospheric stability, landscape roughness, and the background ozone concentrations in the area. The work proved the initial observations of (Gilbert, et al., 2003), which registered a strong negative correlation between NO levels and distance from the highway, stating that distance from the roadway may be a valid surrogate variable for at least some traffic-related air pollutants. However, Zou et al. (Zou, et al., 2006) observed a shifted power-law model to simulate concentrations of NO_2 with distance from a highway. Nevertheless, the authors also studied the wind profile, identifying some significant similarities between wind profile and air pollutants concentration near highways.

Ndoke & Jimoh (Ndoke & Jimoh, 2005) linked the growth of a city (Minna, Nigeria) with the increase of motor vehicles (400%) and the increase in traffic emissions. The authors observed that the concentration of CO decreases with an increase in distance from highways. By observing the social and health aspects of workers and inhabitants in the vicinity of highways, the authors linked the high values of CO_2 measured in the neighborhood of roads with the increase in respiratory diseases. However, level of pollutants (CO, CO_2, NO, NO_2 and SO_2) were below the maximum level stipulated by the Environmental Protection Agency. Measurements were made in a dry season to avoid the effect of rain on dispersion of atmospheric pollution.

Westerdahl et al. (Westerdahl, et al., 2009) characterized the on-road vehicle emissions in Beijing, China. In order to do that, the authors measured atmospheric pollutants (CO, particles and black carbon) in three distinct environments: on-road, roadside and open ambient. With the measurements, the authors were able to derive emission factor for on-road heavy-duty vehicles. A strong traffic impact is observed in the concentration of the pollutants in the three locations. The authors also noted the clear impact of diesel truck traffic activity in black carbon concentrations. However, the authors were unable to identify clear daily trends concerning the impact of meteorological factors (except at nighttime).

Matese et al. (Matese, et al., 2009) installed an eddy covariance station in the center of the city of Firenze, Italy, to measure carbon fluxes. The authors were able to correlate the CO_2 emissions with traffic and domestic sources, using estimations from detailed inventory of the city traffic and natural gas. Additionally, the effect of air turbulence on the dispersion of atmospheric pollutants was also observed. Finally, using data collected along 3.5 months, the authors were able to describe seasonal changes.

Measurements of exhaust gases' concentrations on the atmosphere are usually made using fluxes measured by eddy covariance (EC) technique. Mobile and wireless solutions have been explored to improve the spatial resolution of atmospheric pollutants monitoring, mainly based on vehicles equipped with sensors. To develop new approaches to capture exhaust gases' concentrations, Mao et al. (Mao, et al., 2012) presented *CitySee*, a real-time CO_2-monitoring system using wireless sensor networks for an urban area, in Wuxi, China, proposing a low-cost sensor deployment strategy. Moltchanov et al. (Moltchanov, et al., 2015) took advantage of advances in communication and sensory technologies to deploy a network of six wireless multi-sensor miniature nodes in three urban sites, about 150 m apart. The wireless distributed sensor networks was composed of metal oxide chemo-resistive sensors for O_3, NO_2, and volatile organic compounds, an optical (IR based) sensor for suspended particulate matter, an electret microphone (electrostatic capacitor-based microphone used as noise sensor), and a dual semiconductor for temperature and relative humidity. The measurements took place in the city of Haifa, Israel, in three distinct locations (low activity street, busy street and a main street), during 71 days in the summer of 2013. Measurements showed high correlations among the sensors. The authors demonstrated the network capability to capture spatiotemporal concentration variations with fine resolution. However, they also highlighted the need for a frequent in-situ calibration to maintain the consistency of some sensors, therefore a procedure for a field calibration is proposed.

Hu et al. (Hu, et al., 2009b) proposed a vehicular sensing system to collect CO_2 concentration in urban areas, based on GSM short messages and GPS information from vehicles. Vehicles were used as carriers of sensing devices to monitor CO_2 concentrations while driving through the city. The concept was tested using the ZigBee platform. Kumar et al. (Kumar, et al., 2015) reviewed the state of the art for low cost sensing of atmospheric pollution. The authors argued that new developments in sensor technology are able to provide low-cost and sensible devices, with the ability to communicate and store information. However, they also warn of the challenge in managing widespread sensor networks. As wider networks provide higher spatial resolution that implies there will be more data to be processed and stored, more communication bandwidth, and maintenance issues. An equilibrium between simulation models and the number of sensors should be pondered, in order to reduce the costs of maintaining a growing infra-structure and still be able to accurately estimate atmospheric pollutants. However, the focus should rely on the quality of sensors, since the data collected directly affects the accuracy of the models.

4.2.4 Simulation of exhaust gases dispersion in the atmosphere

Various studies are based on the simulation of scenarios, which includes models for atmospheric pollutants. To model particle dispersion in the atmosphere, several models are available, as described by Holes & Morawska (Holmes & Morawska, 2006): box models (AURORA, CPB and PBM), Gaussian models (CALINE4, HIWAY2, CAR-FMI, OSPM, CALPUFF, AEROPOL, AERMOD, UK-ADMS and SCREEN3), Lagrangian/Eulerian Models (GRAL, TAPM, ARIA Regional), CFD models (ARIA Local, MISKAM, MICRO-CALGRID) and models which include aerosol dynamics (GATOR, MONO32, UHMA, CIT, AERO, RPM, AEROFOR2, URM-1ATM, MADRID, CALGRID and UNI-AERO).

Additional modelling alternatives can also be observed, such as AUSPLUME (Hurley, 2006); CAMx (Comprehensive Air Quality Model with Extensions) (Holmes & Morawska, 2006); ISCST3 (Industrial Source Complex - Short Term) (Hurley, 2006); ROM (Regional Oxidant Model) (Russo & Soares, 2013); SAPRC99 (Kota, et al., 2013), (Stein, et al., 2007); TAMNROM-3D (Kota, et al., 2013); TAPM (The Air Pollution Model) (Carvalho, 2009); TREM (Transport Emission model for Line Sources) (Borrego, et al., 2003); or VADIS (Local Scale Dispersion Model) (Borrego, et al., 2003).

More common and recommended models are:

- ADMS (Atmospheric Dispersion Modelling System) (Venkatram, et al., 2007), (Venkatram, et al., 2009), (Hurley, 2006);

- AERMOD (AERmic MODel) (Holmes & Morawska, 2006), (Johnson, et al., 2010), (Kota, et al., 2013), (Misra, et al., 2013), (Stein, et al., 2007), (Venkatram, et al., 2007), (Venkatram, et al., 2009), (Venkatram, et al., 2013a);

- CALINE4 (Kota, et al., 2013), (Beelen, et al., 2010), (Colls & Tiwary, 2009), (Krzyzanowski & Schneider, 2005), (Misra, et al., 2013), (Venkatram, et al., 2007), (Venkatram, et al., 2009);

- CHIMERE (Chemistry-Transport Model Simulation), a multi-scale deterministic model for air quality forecasting and simulation (Dutot, et al., 2007), (Monteiro, et al., 2005) (Monteiro, et al., 2007), (Pernigotti, et al., 2012), (Russo & Soares, 2013);

- CMAQ (Community Multiscale Air Quality Model) (Beelen, et al., 2010), (Johnson, et al., 2010), (Kota, et al., 2013), (Russo & Soares, 2013), (Stein, et al., 2007);

- QUIC (Quick Urban & Industrial Complex) (Misra, et al., 2013);

- UAM (Urban Airshed Model) (Agirre-Basurko, et al., 2006), (Borrego, et al., 2003), (Russo & Soares, 2013);

- URBIS (URBis Information System) (Beelen, et al., 2010), (Krzyzanowski & Schneider, 2005).

Finally, Land Use Regression (LUR) modeling is a statistical technique used to determine exposure to air pollutants in epidemiological studies (Dons, et al., 2013), widely used to model the dispersion of atmospheric pollutants as explored by Allen et al. (Allen, et al., 2011), Beelen et al. (Beelen, et al., 2010), Johnson et al. (Johnson, et al., 2010), Madsen et al. (Madsen, et al., 2011), Mukerjee et al. (Mukerjee, et al., 2009), Su et al. (Su, et al., 2009), and Wang et al. (Wang, et al., 2013).

Nevertheless, it is not the intent of this work to study or compare different atmospheric models for atmospheric pollutant dispersion, as performed by Holes & Morawska (Holmes & Morawska, 2006), Stein et al. (Stein, et al., 2007), (Beelen, et al., 2010), Misra et al (Misra, et al., 2013), or Russo & Soares (Russo & Soares, 2013), but to analyze the temporal and spatial characteristics of exhaust gas concentration, and explore techniques to estimate the concentration level of nitrogen dioxide without prior or historical information on exhaust gases' concentrations, taking into account other urban variables.

4.2.5 Our approach

Most of the described works focus on two main topics: forecasts based on historical data and inventories of atmospheric pollutants, or a process of data collection and analysis. Our approach differs from those aforementioned in the sense that (1) we aim to estimate exhaust gases' concentrations without using prior or historical information about atmospheric pollutants, and instead rely on opportunistic data provided from distinct sources (meteorological conditions and taxi activity); and (2) we do not propose a new procedure to collect data, and instead rely on the present sources and infra-structures already deployed in the city.

4.3 Methodology

Following the methodology of previous chapters, our approach is based on the classic process of knowledge extraction from databases as described by Witten & Frank (Witten & Frank, 2005) and Santos & Azevedo (Santos & Azevedo, 2005). It comprises the following steps:

- Data collection;

- Data cleaning and transformation;
- Exploratory analysis;
- Inference analysis;
- Validation.

Data about the concentrations of exhaust gases was compiled and provided by '*Comissão de Coordenação e Desenvolvimento Regional de Lisboa e Vale do Tejo*' (CCDR-LVT) and '*Agência Portuguesa do Ambiente*' (APA). Information about meteorological conditions was retrieved from the Weather Underground. Both data sets were retrieved during the same time window to match taxi data (from September to December, 2009).

Although data providers prepared data beforehand, a cleaning and transformation process is applied in order to remove faulty, erroneous or missing samples, as well as to format the data set into a more suitable scheme. Finally, data is stored in a relational database system with support for geographical objects.

The exploratory analysis studies the exhaust gases time series to identify temporal trends, which will be examined during the inferential analysis. The variables are characterized and daily and seasonal patterns are extracted. The exhaust gases profiles are studied under the influence of different meteorological conditions (weather conditions, temperature, and wind). The correlation between the main atmospheric pollutants is investigated in order to attain which exhaust gases should be analyzed during the next step.

Data analysis investigates different models to estimate exhaust gases' concentrations. More specifically, we explore techniques to estimate the concentration of nitrogen dioxide (NO_2) based on meteorological conditions (humidity, temperature, wind, and weather conditions) and taxi activity (pick-ups, drop-offs and average speed).

We use regression as a statistical technique to model relations between variables. Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Moreover, this technique allows identifying which proportion of the variance of the dependent variable is explained by the explanatory variables (also designated predictor or independent variables), and the contribution of each explanatory variable (Pallant, 2005).

To estimate the coefficients, the method of ordinary least squares (OLS) is proposed, consisting in minimizing the sum of squares of residuals (Pallant, 2005). The model coefficients are tested using the coefficient of correlation of Pearson (r) and the coefficient of determination (r^2). The coefficient of correlation of Pearson is a measure of linear association between variables, while the coefficient of determination (the square of the coefficient of correlation) explains the percentage of variation of the dependent variable determined by the independent variables, i.e. the goodness of the fitting of the model to data (Kennedy, 2008).

The regression significance is tested using the *F* test of ANOVA, which analyses the existence of a linear relation between the dependent variable and some of the explanatory variables (Maroco, 2005). By rejecting the null hypothesis, we are assuming that, at least one of the explanatory variables contributes to the model. Therefore, the null hypothesis with significance level (*p*-value) lower than 0.05 should be rejected (Clemente, 2013).

To select the most efficient regression model, the Stepwise method is applied. In this method, the model starts with no variables, and gradually those that are statistically significant are added, until all the variables of the model are significant and no further improvement is possible (Demuth, et al., 2008). Additionally, multicollinearity is verified using the VIF (Variance Inflation Factor) method. The absence of multicollinearity is achieved if all VIF values are below a critical level (10, ideally near 1), assuring that the explanatory variables are not correlated (Pallant, 2005).

An alternative model is explored due to the apparent complexity of the problem, and the uncertainty about whether linear relations were an adequate fit to the problem. Suggested by different authors to model air pollution (Perez & Reyes, 2001), (Kukkonen, et al., 2003), (Agirre-Basurko, et al., 2006), (Juhos, et al., 2008), (Ahmed, et al., 2010), a multilayer perceptron with backpropagation was additionally studied for comparison with the linear regression approach, in this way exploring different parameter configurations.

As previously described, during the linear regression analysis, the Stepwise method with the multicollinearity analysis is used to select the most significant predictor variables. Nevertheless, to deal with the amount of possible predictor variables and identify which ones could better explain the variation of the dependent variable (exhaust gas), an alternative approach is explored. A factorial analysis - Principal Component Analysis - is considered (Pallant, 2005), a method which presents no *a priori* requirements, using the Kaiser's rule for the eigenvalues (Larsen & Warne, 2010) and the scree plot approach to retain the most significant factors. Additionally, the three-sigma rule of thumb (or 68–95–99.7 rule), a conventional heuristic that states almost all values lie within three standard deviations of the mean (Kazmier, 2003), is also applied to help improve the model.

The experiments are conducted using one representative exhaust gas (NO_2) in one predominant monitoring station ('*Av. Liberdade*'). Finally the approach is applied to the remaining monitoring stations for comparison. Main findings and achievements were submitted to a peer-review international conference (Veloso, et al., 2013), (Veloso, et al., 2015).

4.4 Data description

This section describes the data set used for the exploratory study and data analysis. The data set comprises information concerning exhaust gases' concentrations, meteorological conditions (temperature, humidity, wind speed and weather conditions), and taxi activity (described in Chapter 2). The data was collected from September to December 2009.

4.4.1 Exhaust gases

4.4.1.1 Data set content

The exhaust gases' data set was provided by both the '*Comissão de Coordenação e Desenvolvimento Regional de Lisboa e Vale do Tejo*' (CCDR-LVT)[18], and the '*Agência Portuguesa do Ambiente*'[19], which are governmental institutions

[18] CCDR-LVT. http://www.ccdr-lvt.pt/pt/ .

[19] Agência Portuguesa do Ambiente. http://www.qualar.org .

responsible for monitoring atmospheric pollutants. The data set was composed of hourly readings of different gases' concentrations in seven monitoring stations (shown in Table 4-1). Every station monitors nitrogen monoxide or nitric oxide (NO), nitrogen dioxide (NO_2), nitrogen oxides (NO_x, composed by NO and NO_2), and carbon monoxide (CO) concentrations, measured in $\mu g/m^3$ (micro gram per cubic meters), which are flue combustion gases (from natural or anthropogenic sources, such as traffic, industry or house heating), also termed *exhaust gases*. Additionally, some stations also monitor the amount of particles (with dimensions smaller than 10 µm) in the atmosphere or particulate matter (PM), Sulphur Dioxide (SO_2), Benzene (C_6H_6) and Ozone (O_3) concentrations, as described in Table 4-1. The data were obtained with a one-hour sampling rate.

	Type	**PM < 10 µm**	**NO**	**NO_2**	**CO**	**NO_x**	**SO_2**	**C_6H_6**	**O_3**
Monitoring Station / Units			**µg/m3**	**µg/m3**	**µg/m3**	**µg/m3**	**µg/m3**	**µg/m3**	**µg/m3**
(A) Olivais	B	✓	✓	✓	✓	✓	✓		✓
(B) Chelas	B		✓	✓	✓	✓	✓		
(C) Beato	B		✓	✓	✓	✓	✓	✓	✓
(D) Entrecampos	T	✓	✓	✓	✓	✓	✓	✓	✓
(E) Avenida da Liberdade	T	✓	✓	✓	✓	✓			
(F) Santa Cruz de Benfica	T	✓	✓	✓	✓	✓	✓		
(G) Restelo	B	✓	✓	✓	✓	✓			✓

T = Traffic; B = Background

Table 4-1 - Monitored gases and particles for each station.

Although the current work focuses on a common window of observation including taxi data, from September to December 2009, the exhaust gases' database contains data from 2008 to 2013, which is explored in this section.

The goal of the monitoring stations is to collect information about atmospheric noxious or toxic gases that pose a direct danger to human health when above a certain threshold. Carbon dioxide (CO_2) is not measured by these stations since it is a natural component of the atmosphere and does not pose a direct or immediate danger for human health (although it can generate danger indirectly, for instance increasing the greenhouse effect) (APA, 2012).

4.4.1.2 Characterization of exhaust gases

NO, NO_2, CO, O_3 and SO_2 are molecular, chemical compounds. They are colorless and odorless gases at low concentrations, with the exception of NO_2, a brown gas. NO_2 and SO_2 produce a distinctive scent at high concentrations.

Nitrogen Oxides (NO_x): Nitrogen Monoxide (NO) and Nitrogen Dioxide (NO_2)

NO, NO_2, NO_x and SO_2 are primary pollutants; those emitted directly from the source to the atmosphere (e.g. exhaust gases, from internal combustion motors in vehicles). NO and NO_2 have both natural sources (due to microbial metabolism in the soil, electrical discharges from lightning in the atmosphere) and anthropogenic sources (fuel burning at high temperatures, from domestic heating, or internal combustion vehicles, the latter being the main source of pollutants in urban areas). Although NO is the result of natural bacterial action and an important cellular signalling molecule in humans (and in mammals in general), their concentration in urban environments is between 10 and 100 times greater than in non-urban areas (Ahrens & Henson, 2014). Moreover, NO emitted to the atmosphere can produce a photochemical oxidation forming NO_2 in the troposphere (ground-level), during daylight (Clapp & Jenkin, 2001). Short episodes of high concentrations of NO_2, lasting from a few hours to several days, can occur due to different mechanisms. In summer, photochemistry can rapidly increase the concentrations of NO_2, therefore decreasing the concentrations of NO. In winter, the inversion of temperatures traps emissions. In this scenario, NO concentrations are also high. Nitrogen oxides are highly reactive gases that play a key role in producing ozone and other ingredients in photochemical smog. Additionally, in moist air, NO_2 reacts with water vapour to form corrosive nitric acid (HNO3) (Colls & Tiwary, 2009).

Ozone (O_3)

O_3 is a secondary pollutant, the result of a slow photochemical reaction between NO_x, CO and Volatile Organic Compounds (VOC, emitted during incomplete combustions and fuel volatilization), in the presence of solar radiation and high temperatures (higher concentrations in the summer) (Ahrens & Henson, 2014). It is also an essential gas in the stratosphere, since it is able to reduce ultra-violet radiation. However, when located in the troposphere (ground-level), it is considered a noxious pollutant, the main constituent of photochemical smog. In areas directly influenced by traffic emissions, solar radiation induces NO_2 photolysis to form O_3 during daylight.

However, overnight, O_3 is destroyed, while NO_2 builds up until the next day, when it can generate new O_3 (Lameiras & Povoas, 2005).

Carbon Monoxide (CO)

CO is naturally produced by volcanic eruptions, forest fires and chlorophyll decomposition. Anthropogenic sources are related to incomplete combustions and other organic materials. The concentration decreases with the increase of distance to the source, as it is spatially variable and short-lived (can quickly be removed from the atmosphere by microorganisms in the soil), having a role in the formation of ground-level ozone. Nevertheless, it is a very dangerous gas, even in small concentrations (Lameiras & Povoas, 2005) (Ahrens & Henson, 2014).

Sulphur Dioxide (SO_2)

SO_2 can also appear naturally in the atmosphere as a result of volcanic eruptions. The main anthropogenic sources are fossil fuel burning in the energy industry, although diesel vehicles can also produce this gas. The resulting SO_2 from fuel burning can be transformed into Trioxide Sulfur (SO_3), which in contact with atmospheric humidity, produces Sulfuric Acid (H_2SO_4) (Lameiras & Povoas, 2005).

Particles

Atmospheric particles can be naturally produced by volcanic eruptions, forest fires, and the effect of wind on the soil. Anthropogenic sources are diverse and include traffic, burning of fossil fuels, and industrial processes. In urban areas, most of the particles are derived from primary pollutants, such as NO_2 and SO_2. While particles with higher dimensions settle, low dimension particles are long lived in the atmosphere and can be transported along large distances (APA, 2012), (Lameiras & Povoas, 2005).

Impact on human health

The atmospheric pollutants have a considerable impact on human health. While CO affects essentially the cardiovascular system (inhibits the O_2 exchange between blood vessels and vital tissues), NO_2, O_3, and SO_2 affect the respiratory system. Particles should also be taken into consideration, not only due to their size and the immediate effect on the human respiratory tract, but also due to the ability to absorb hydrocarbons and heavy metals, transport them to the lungs and from there to the blood vessels (if size < 2.5 µm) (APA, 2012), (Lameiras & Povoas, 2005).

4.4.1.3 Characterization of monitoring stations

The monitoring stations were classified by CCDR-LVT into two groups: traffic stations (*D, E* and *F* in Figure 4-1) and background stations (*A*, *B*, *C* and *G* in Figure 4-1). The traffic stations are located near traffic roads while the background stations are located away from main roads. The former are used to monitor exhaust gas emission from traffic vehicles around 100m radius, whereas the latter are used to monitor domestic or industrial sources of exhaust gases, and can sense emission within a 3000m radius (Fontes, 2010). On average, traffic stations perceive higher concentrations of exhaust gases (57.4 μg/m^3 for NO_2, during the four months of observations) than background stations (33.7 μg/m^3 for NO_2). These observations are in line with Pleijel et al. (Pleijel, et al., 2004), Gilbert et al. (Gilbert, et al., 2003), Zou et al. (Zou, et al., 2006) and Ndoke & Jimoh (Ndoke & Jimoh, 2005), which observed that concentrations of exhaust gases decreased with the increasing distance from roads.

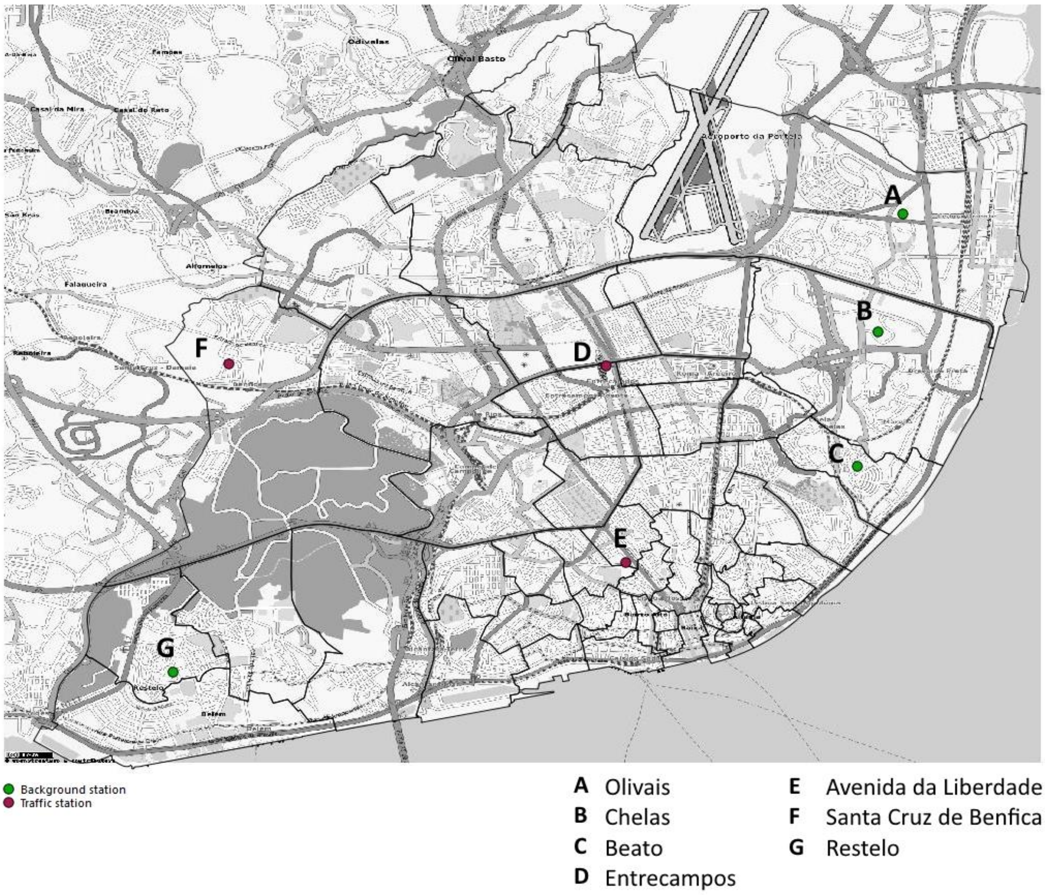
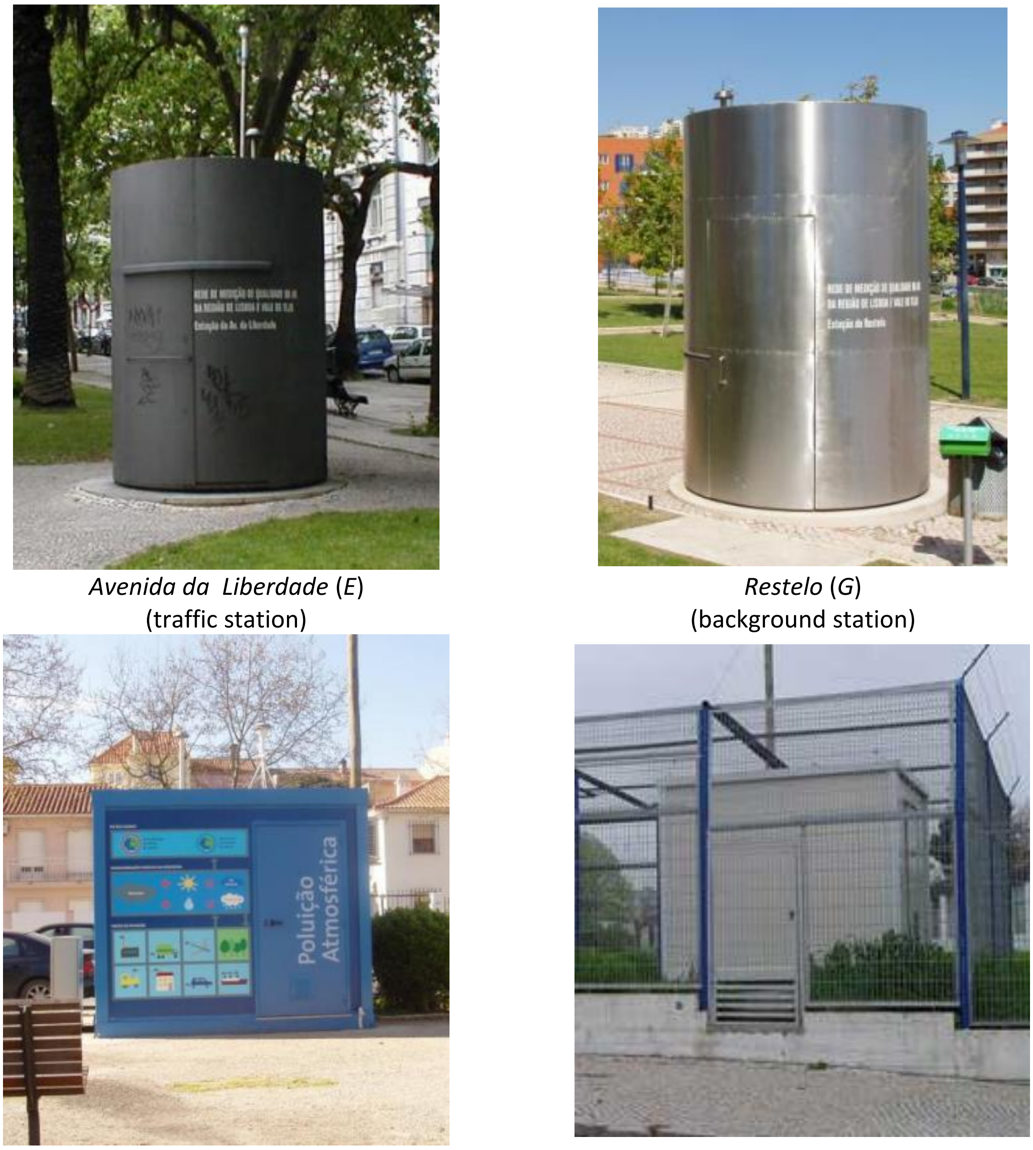


Figure 4-1 - Location of monitoring stations.



Avenida da Liberdade (*E*)
(traffic station)

Restelo (*G*)
(background station)

Santa Cruz de Benfica (*F*)
(traffic station)

Beato (*C*)
(brackground station)

Figure 4-2 – Images of different types of monitoring stations.

"*Av. Liberdade*" (*E*), "*Entrecampos*" (*D*) and "*St. Cruz Benfica*" (*F*) are important monitoring stations for the study, since they are located near main roads, to sense exhaust gas emissions from traffic. Usually, these stations detect high levels of short-lived pollutants. "*Av. Liberdade*" is a major road in Lisbon, with 10 lanes for traffic and wide pedestrian areas on each side of the avenue. "*Entrecampos*" is a crossroad (major roundabout) between "*Campo Grande - Av. da República*" and "*Av. das Forças Armadas - Av. Estados Unidos da América*", all main roads with intense traffic (APA, 2012). "*St. Cruz Benfica*" is located 500 m away from the intersection of two main highways in the city (A36 and A37), and 2 km apart from two other major highways (A5 and IC16). The remaining monitoring stations (background) are located in residential areas, with low intensity of traffic and away from major sources of atmospheric pollution. Figure 4-2 represents some of the monitoring stations (images provided by CCDR-LVT).

4.4.2 Meteorological conditions

Information on weather conditions was retrieved from the Weather Underground[20] - an online weather information service provider. Much like the exhaust gas data set, only samples from September to December 2009 were analyzed.

The data set contains 20 types of weather conditions, which were grouped into three sets (Clear; Cloudy; Rainy), as shown in Figure 4-3. As summer season gives way to autumn and then winter, the weather conditions tend to worsen. The number of days with clear weather decreases and at the same time the number of rainy days increases.

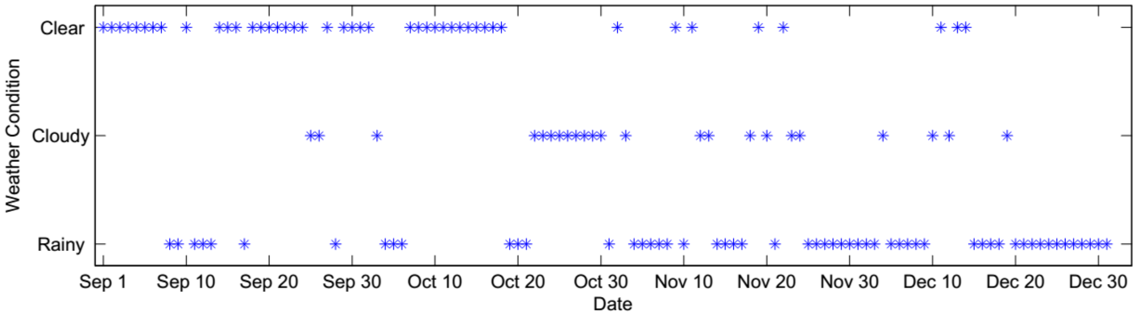


Figure 4-3 - Weather conditions for each day from September 1st to December 31st, 2009.

In addition to weather conditions, temperature, humidity, and wind speed were also acquired hourly. The corresponding time series can be observed in Figure 4-4. Moving from September to December we can identify a decrease in temperature and an increase in humidity. Daily profile of wind speed appears to be comparable between hotter and colder months, with a stronger increase at the end of the time series (last two weeks of December).

[20] Weather Underground. http://www.wunderground.com/history/

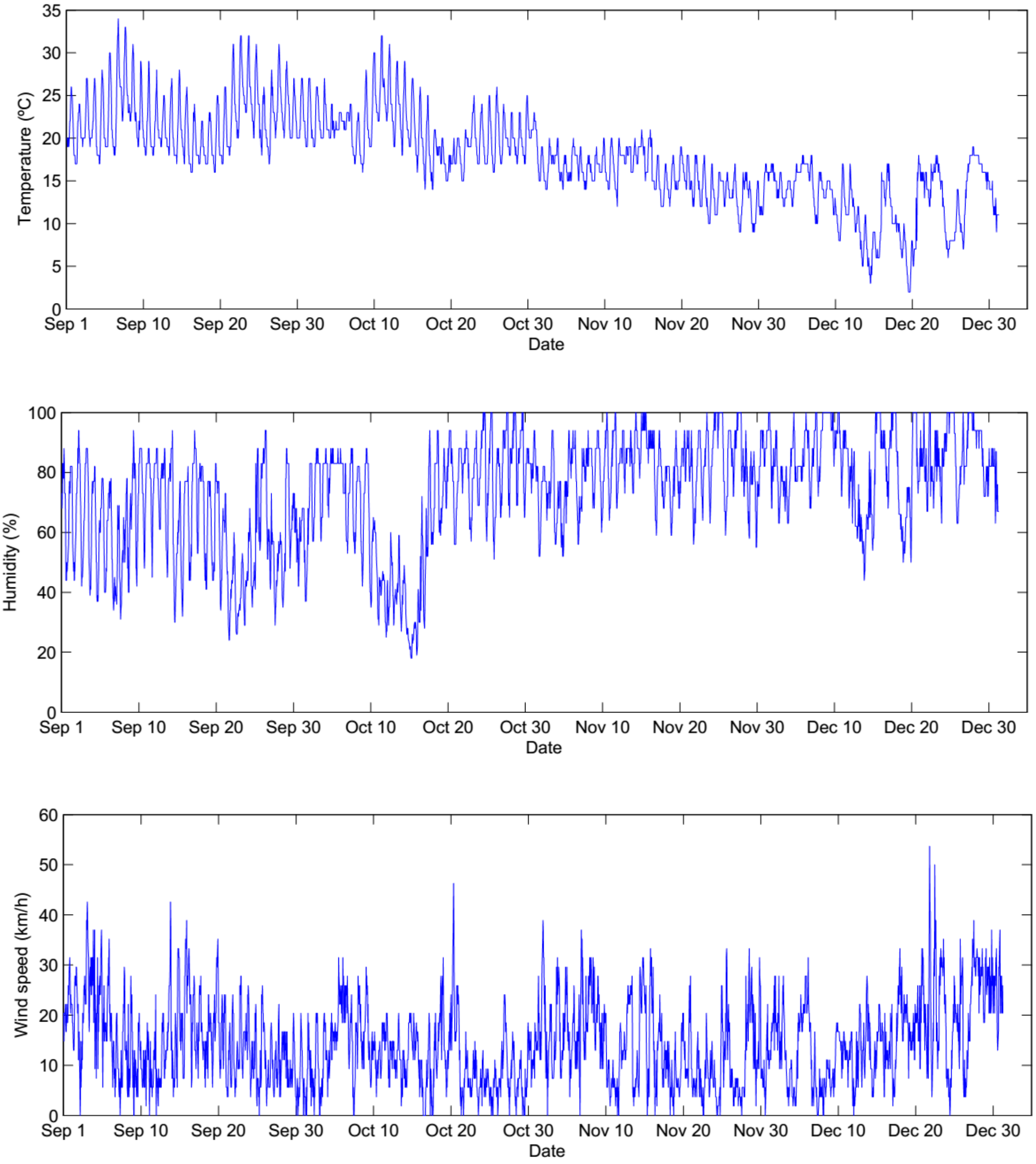


Figure 4-4 - Time series of temperature (top), humidity (middle) and wind speed (bottom) in Lisbon, from September 1 to December 31, 2009.

Figure 4-5 shows the average daily profile of temperature, humidity and wind speed (data collected from September to December). Both temperature and wind speed start increasing around 8 AM onwards, peaking at 5 PM. Humidity presents the opposite behavior. It decreases from 7 AM, reaching the lowest value at 4 PM.

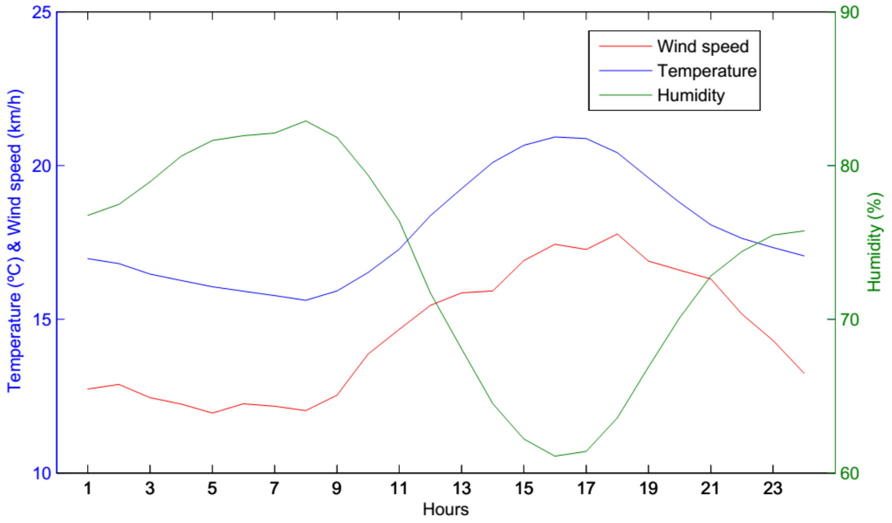


Figure 4-5 - Variation of temperature, wind speed and humidity during a 24h-period in Lisbon (data collected from September 1, 2009, to December 31, 2009).

4.5 Data exploration

This section characterizes the temporal and spatial patterns of exhaust gases and meteorological conditions throughout the four months of observations (from September to December 2009). Additionally, the correlation between exhaust gases is studied in order to identify a suitable marker for exhaust-pipe emissions.

4.5.1 Temporal analysis

4.5.1.1 Time series of exhaust gases

The concentration of atmospheric pollutants can vary according to seasons. Lower temperatures slow the rate of dispersion of gases, therefore, higher concentrations of atmospheric pollutants are usually perceived in winter, although the formation and dispersion of exhaust gases is the result of a complex interaction between several variables and not exclusive of a single factor. Moreover, wintertime is more prone to episodes of inversion layer, which traps atmospheric pollutants in the lower levels of the atmosphere (Clapp & Jenkin, 2001). Additionally, road traffic increases (Hu, et al., 2009c) and the larger number of pollution sources (e.g. domestic heating) (Colls & Tiwary, 2009) also contributes to the increase in atmospheric pollutants on winter. This scenario can be observed in Figure 4-6, which plots the time series for average concentrations of NO, NO_2, NO_x and CO during 2009. The data is aggregated per hour, as an average measure of all monitoring stations.

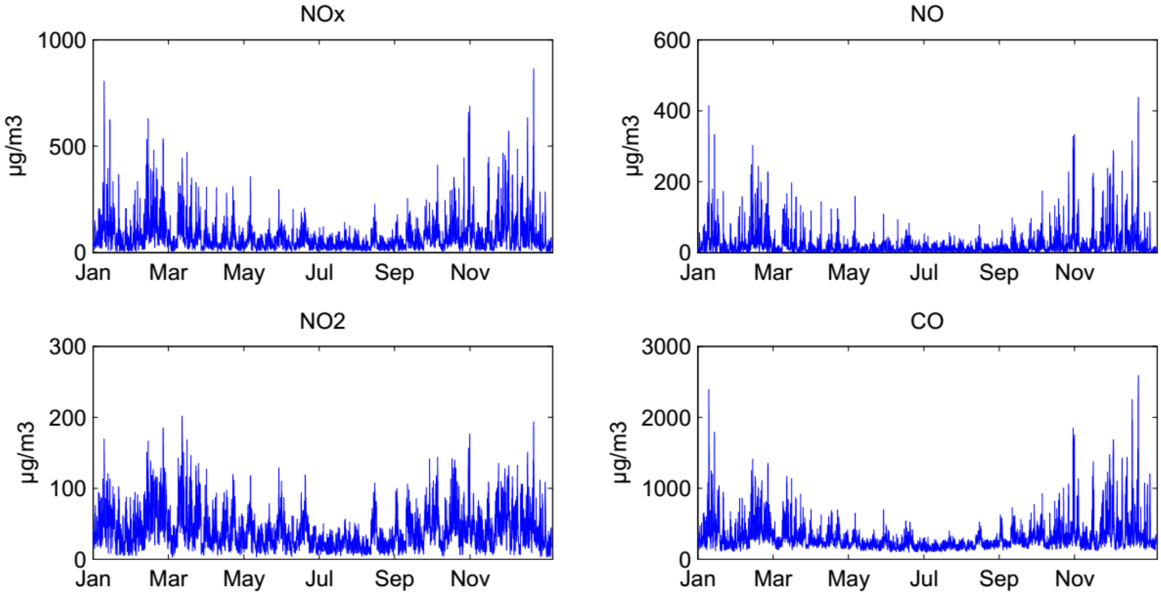


Figure 4-6 - Time series of gases common to all monitoring stations during a 12 months period (from January to December 2009) for Lisbon Municipality (all monitoring stations considered).

Annual mean concentration of NO_2 is 39.88 µg/m^3, while NO is 22.09 µg/m^3 and NO_x is 73.74 µg/m^3. The concentration decreases from winter to summer - it is higher in NO (from 33.12 µg/m^3 to 12.58 µg/m^3) than in NO_2 (from 46.42 µg/m^3 to 30.87 µg/m^3). Although in summer higher temperatures facilitate the dispersion of atmospheric pollutants, higher temperatures in daylight can also enable a photochemical oxidation of NO forming NO_2 (Clapp & Jenkin, 2001). Table 4-3 summarizes the results:

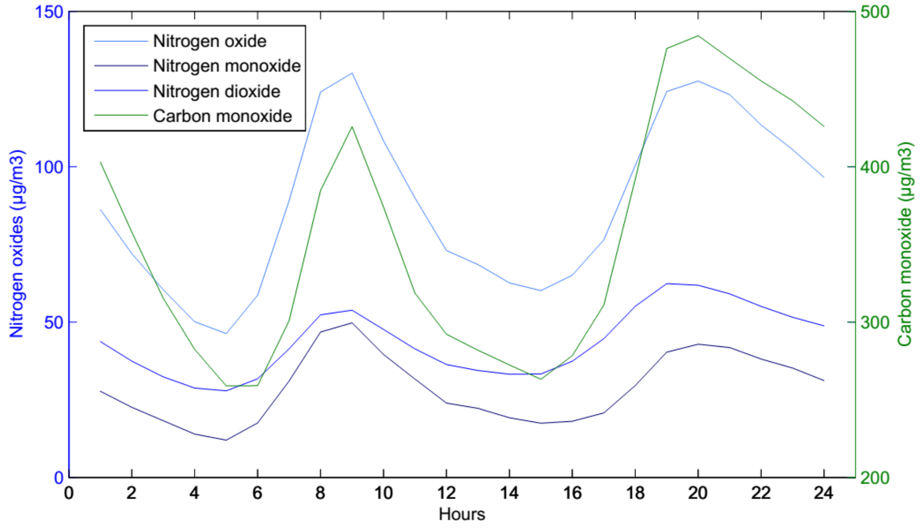
	Winter	Summer	Annual
NO	33.12	12.58	22.09
NO_2	46.42	30.87	39.88
NO_x	97.20	50.15	73.74
CO	386.80	230.81	304.64

Table 4-2 – Average concentrations of exhaust gases in winter and summer months. Measurements in µg/m^3.

4.5.1.2 Daily variation

Figure 4-7 shows the average variation of exhaust gas over the course of one day (on top, the data was collected during one year, from January to December 2009; on bottom the data was collected during four months, from September to December 2009, considering the contribution of all monitoring stations). Two daily peaks of gas concentration, which are related to traffic congestion, were also observed by Colls & Tiwary (Colls & Tiwary, 2009), Melkonyan & Kuttler (Melkonyan & Kuttler, 2012) and described in (APA, 2012). The morning peak quickly increases from 5 AM, reaches the maximum around 9 AM and quickly drops, corresponding to the inbound traffic to the city. In the afternoon, gas concentration gradually rises around 3 PM, reaches the maximum around 8 PM and then slowly drops, corresponding to the outbound traffic from the city. When atmospheric stability is high, we can observe a gradual decline during the remainder of the day, after the rush-hour (Colls & Tiwary, 2009). on average, the second daily peak reaches higher concentrations (15.85% for NO_2) than the first daily peak and takes longer to disperse, as observed by Uno et al. (Uno, et al., 1996). Remaining atmospheric pollutants resultant from the morning commuting add to the afternoon emissions, causing a higher concentration of atmospheric pollutants. Additionally, the drop of temperature at the beginning of the night slows the dispersion of gases. This scenario can be reversed on warmer months, with higher temperatures at the beginning of the night helping to disperse atmospheric pollutants, as observed by Huang et al. (Huang, et al., 2001).

As stated before, the rate of gas dispersion is also affected by temperature. Heated gas expands its' volume as higher temperatures increase molecules' speed and hence disperses more quickly. The opposite occurs when faced with cold, as gas responds by contracting and by slowly dispersing (Beychok, 2005). However, at night, although there is a significant decrease in temperature, the production of atmospheric pollutants drops considerably. The number of moving vehicles reduces to a minimum, traffic being one of the major sources of exhaust gases. On average, we observed a reduction of exhaust gas concentrations of about 19.1% at night (from 10 PM to 7 AM) and 23.1% on weekends.



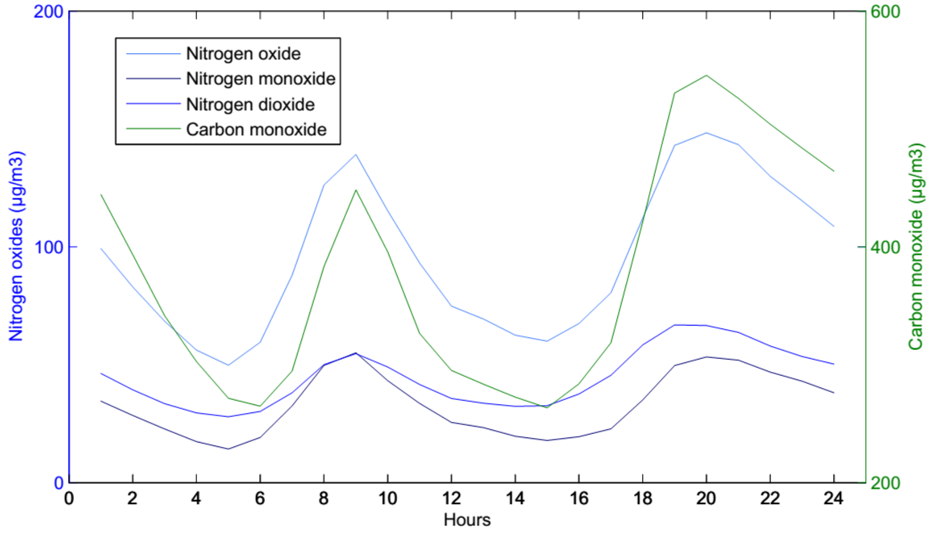


Figure 4-7 - Average variation of exhaust gases over the course of a day for Lisbon Municipality.
Top: January to December 2009 (12 months period).
Bottom: September to December 2009 (4 months period).

A similar scenario can be observed on individual monitoring stations. Figure 4-8 shows the average variation of exhaust gas over the course of one day by considering only the "*Av. Liberdade*" monitoring station. At the top, the data was collected during one year, from January to December 2009, whilst at the bottom the data was collected during four months, from September to December 2009. Similar to the previous figure, two daily peaks are present. Overall gas concentrations are higher on this monitoring station since it is located in the vicinity of a main road, thus sensing the direct impact of traffic exhaust gases. Additionally, the nearby road network can also influence the concentrations of atmospheric pollutants (Hu, et al., 2012c). Roads with a higher amount of crosswalks, stop signs, and/or traffic signals require frequent accelerations from vehicles. The same goes for roads with short segments and roads with compact traffic (Hu, et al., 2012c). Moreover, in this station, the second daily peak shows a 29.67% increase on NO_2 concentrations compared to the first daily peak, which was higher than the previous observations.

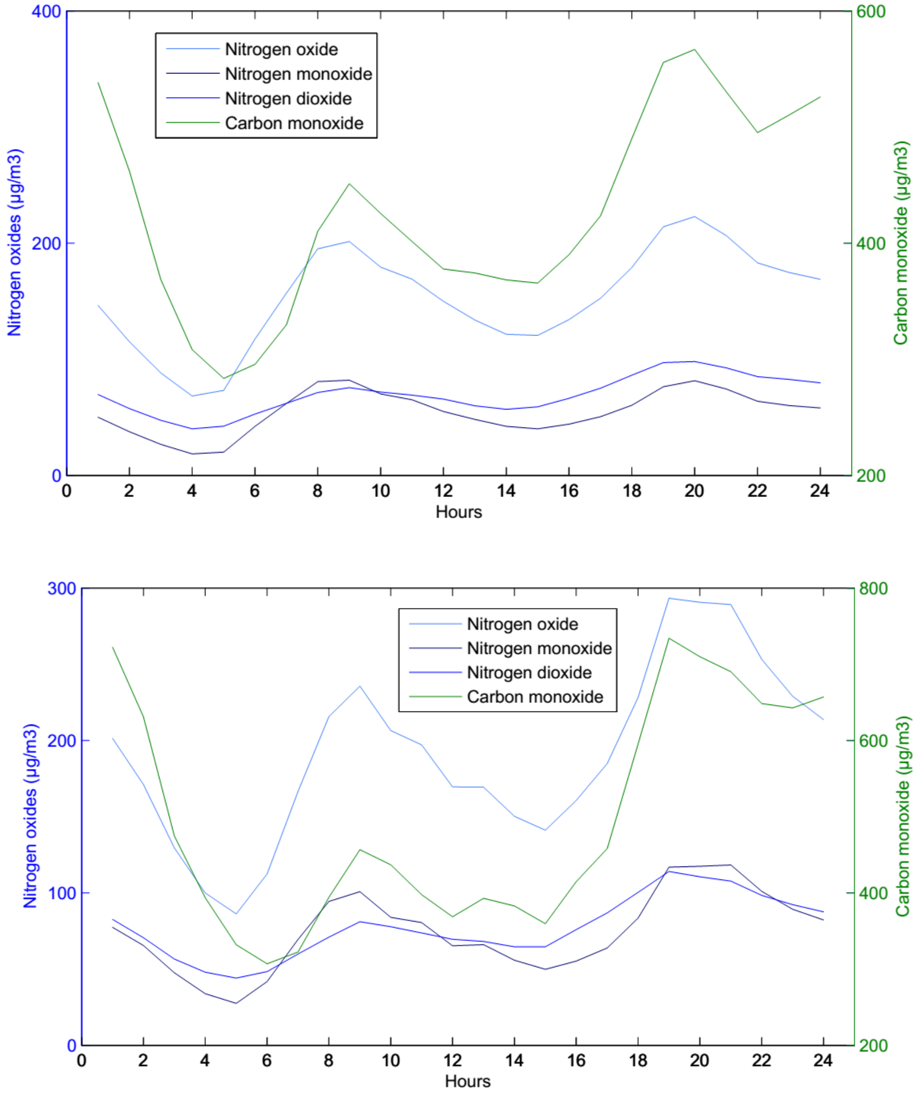


Figure 4-8 - Average variation of exhaust gases over the course of a day for "*Av. Liberdade*" station.
Top: January to December 2009 (12 months period).
Bottom: September to December 2009 (4 months period).

Generally, traffic stations perceive higher concentrations of atmospheric pollutants than background stations, as can be observed in Table 4-3. Moreover, the difference between the first and the second peak in daily concentrations of atmospheric pollutants is also higher in traffic stations (on average, an increase of

5.22% in the second daily peak) than in background stations (on average, an increase of 1.34% in the second daily peak). As mentioned, this is due to the nature of these sensors, which are located near main roads, and subject to the direct impact of nearby traffic.

	4 months (Sep – Dec, 2009)		
	Traffic	Background	All stations
NO	48.85	13.87	28.83
NO_2	57.43	33.65	43.82
NO_x	132.34	54.92	88.02
CO	477.65	262.88	355.41

	1 year (Jan – Dec, 2009)		
	Traffic	Background	All stations
NO	39.21	9.36	22.09
NO_2	53.99	29.39	39.88
NO_x	114.11	43.74	73.74
CO	408.67	227.08	304.64

Table 4-3 – Average concentrations of exhaust gases for traffic and background stations, using data collected from September to December 2009 (left) and data collected from January to December 2009 (right). Measurements in µg/m^3.

4.5.1.3 Seasonal variation

Likewise, warmer months (June, July and August) have, on average, lower exhaust gas concentration (25.7 µg/m^3 for NO_2) than colder months (44.8 µg/m^3 for NO_2, in October, November and December), which can be observed in Figure 4-9 (data as an average measure of all monitoring stations).

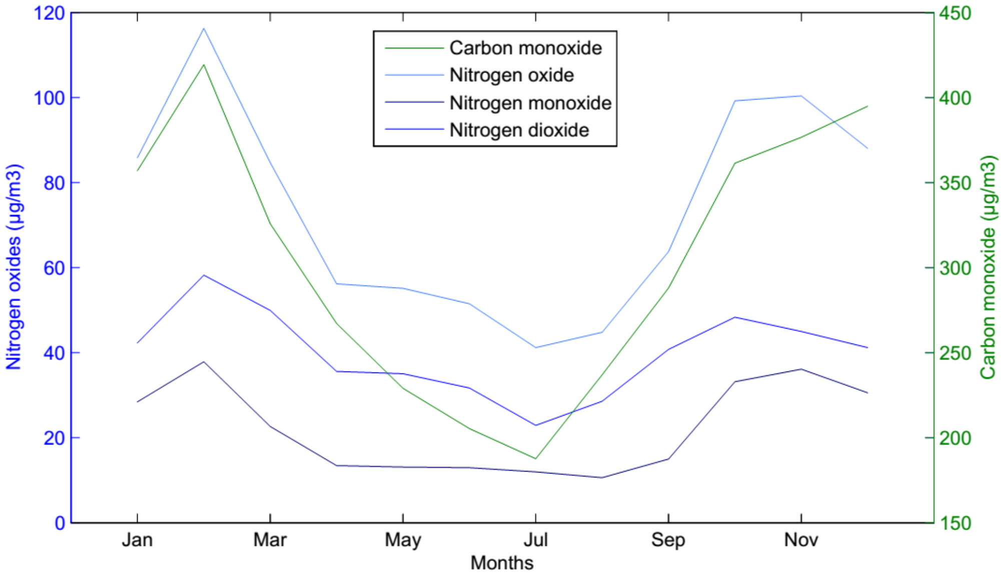


Figure 4-9 - Average variation of exhaust gases across every month in 2009.

In cooler months, the afternoon peak attains higher values of gas concentrations than the morning peak (67.8 µg/m^3 against 55.9 µg/m^3 for NO_2), while in warmer months, the morning peak reaches higher values than the afternoon peak

(37.8 μg/m^3 against 29.8 μg/m^3 for NO_2). These observations can be explained by the higher temperatures at the beginning of the night in warmer months and the occurrence of the inversion layer effect in colder months. Moreover, there is a narrower gap between the maximum and minimum average concentrations of exhaust gases in warmer months (23.3 μg/m^3 for NO_2) when compared with cooler months (40.2 μg/m^3 for NO_2).

Similar profiles can be observed throughout different months (Figure 4-10, with data as an average measure of all monitoring stations). As described before, in September (warmer month) the morning peak is more pronounced than in December (cooler month); however the afternoon peak in December is stronger than that of September, which is similar to the observations of Huang et al. (Huang, et al., 2001) and Uno et al. (Uno, et al., 1996).

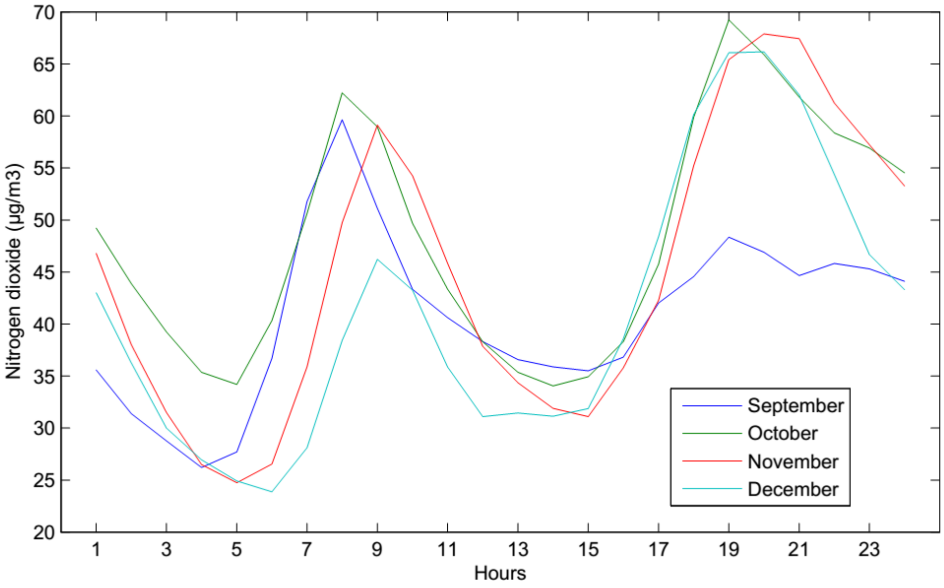


Figure 4-10 - Comparison of average variation of nitrogen dioxide, during a 24h period in different months (from September to December).

Similar patterns were observed when exploring data from different years (between 2008 and 2011) as presented in Figure 4-11 (data as an average measure of all monitoring stations, collected during four months, from September to December, each year). Two daily peaks of NO_2 concentrations are observable, with the second peak, on average, reaching higher values than the first peak.

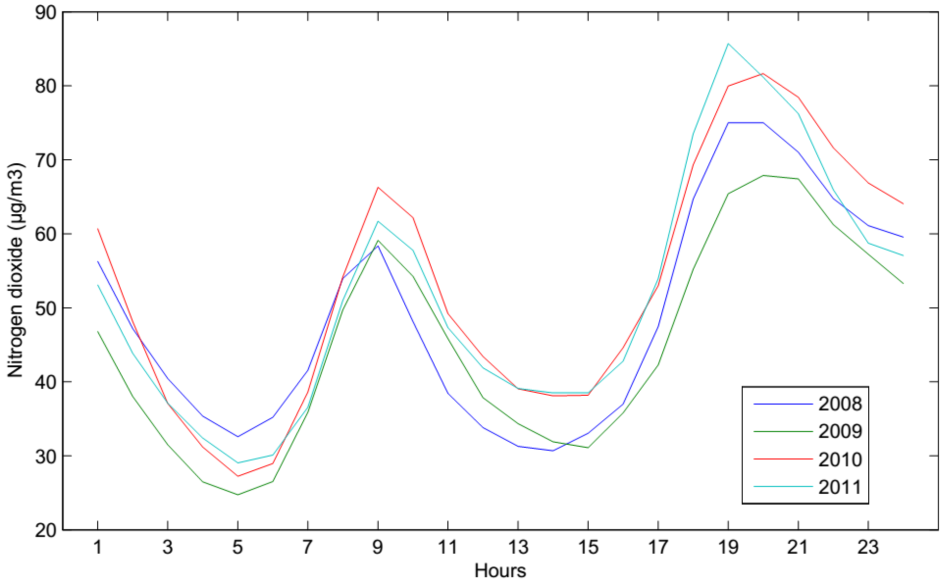


Figure 4-11 – Comparison of average variation of nitrogen dioxide, during a 24h period from September to December, over a period of four years.

4.5.1.4 Weekly analysis

Weekly variation of exhaust gases concentrations is shown in Figure 4-12 (data as an average measure of all monitoring stations). On weekends, the average concentration of exhaust gases is lower than that on weekdays due to the overall reduction in traffic.

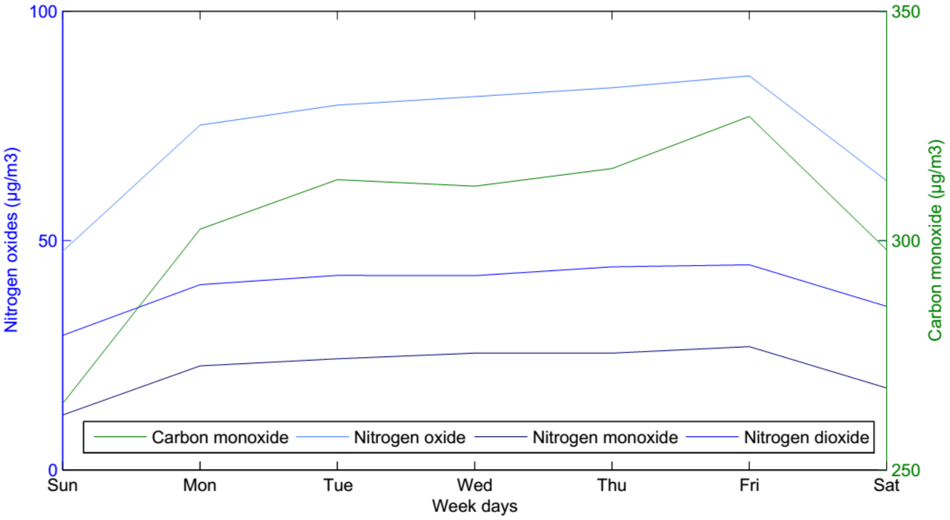


Figure 4-12 – Exhaust gases distribution during the seven days of the week (data collected during 12 months, from January to December 2009) for Lisbon Municipality (all monitoring stations considered).

The reduction of traffic on weekends is visible on the concentration of atmospheric pollutants (Figure 4-13, data as an average measure of all monitoring stations, collected during four months, from September to December). Some consequences can be observed: the average concentration of NO_2 is much lower on weekends (36.08 µg/m^3) than on weekdays (46.85 µg/m^3); the two daily peaks are less emphasized on weekends; and the difference between the maximum and minimum concentration of NO_2 is higher on weekdays (maximum value is 144.7% higher than minimum value) than on weekends (maximum value is 86.5% higher than the minimum value). Table 4-4 summarizes the results:

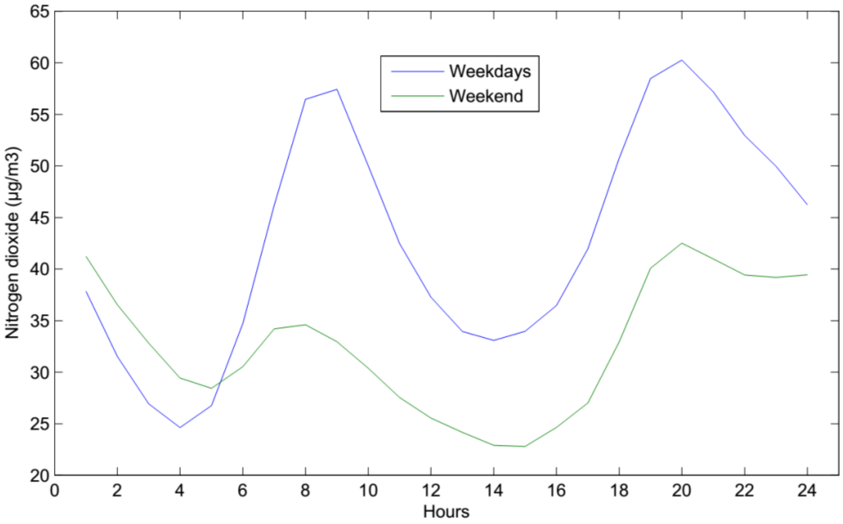


Figure 4-13 - Comparison of average variation of nitrogen dioxide, during a 24h period, on weekdays and weekends.

	4 months (Sep – Dec, 2009)	
	Weekdays	Weekends
NO	32.05	20.58
NO_2	46.85	36.08
NO_x	96.00	67.64
CO	366.77	326.21

	1 year (Jan – Dec, 2009)	
	Weekdays	Weekends
NO	24.95	14.92
NO_2	42.83	32.50
NO_x	81.08	55.38
CO	314.11	281.22

Table 4-4 – Average concentrations of exhaust gases on weekdays and weekends. Measurements in µg/m^3.

Moreover, on Sundays the variation is quite different (Figure 4-14). The average NO_2 concentration is lower than on Saturdays (and weekdays) due to the overall reduction in traffic. The first daily peak is less accentuated and occurs in the early morning, possibly as people return home from Saturday night activities (Colls & Tiwary, 2009).

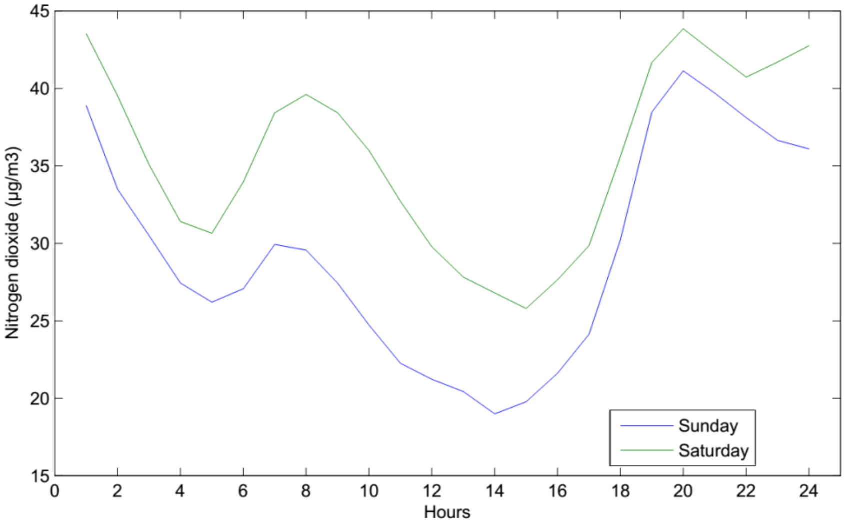


Figure 4-14 - Comparison of average variation of nitrogen dioxide, during a 24h period, on weekends (Sundays and Saturdays).

4.5.2 Characterization of meteorological events

Besides topography (the shape of the landscape), meteorological conditions are the fundamental factor for the dispersion of atmospheric pollutants, therefore, wind, temperature, humidity, air pressure, and weather conditions affect the concentrations of atmospheric pollutants (Kolehmainen, et al., 2001), (Kukkonen, et al., 2003), (Agirre-Basurko, et al., 2006), (Juhos, et al., 2008). There are two main components: *vertical component*, generated by the turbulence of the vertical temperature gradient (thermal gradient) between different low layers of the atmosphere; and *horizontal component*, where the wind is the main agent in transporting and gas mixing (APA, 2012).

4.5.2.1 Temperature

Temperature affects the rate of diffusion of a gas. High temperatures (or heat) causes gases to expand (increase in molecules' speed or kinetic energy), making it less dense, so molecules move faster and there will be more spontaneous spreading of the material, which in turn causes gases to diffuse more quickly. Low temperatures cause gases to contract, making them denser, thus diffusing more slowly (Miller, 2009). In the atmosphere, air temperature usually decreases with the increase in altitude (roughly 1ºC per 100m). This variation is termed *lapse rate,* and it is due to the fact that ascending dry air is subject to lower pressure, which increases the volume and decreases the temperature (APA, 2012). Also to be considered is the event of temperature inversion, or inversion layer. This takes place when the surface air is cooler than the air above. The cooler air near the surface is heavier and will not ascend due to the warmer air above. Pollutants released near the surface will get trapped and build up in the cooler layer of air, near the surface. This event occurs mostly on mountain valleys in clear nights, with light wind, especially in winter (Colls & Tiwary, 2009) (Clapp & Jenkin, 2001).

4.5.2.2 Wind

Wind can be represented by a vector with both magnitude and direction. The vertical component of the wind (Z axle) is responsible for the turbulence, while the remaining components (X and Y axles) define the transport and dilution of pollutants (APA, 2012). Wind speed increases with height, directly affecting the pollutants' dispersion from chimney flues at high altitude (e.g. factory chimneys), especially in the initial mixing of expelling gases with the atmosphere. With a stable atmosphere, gas plumes can be transported long distances, depositing gas concentrations in soil level at locations far away from pollutant sources (Miller, 2009). Therefore, wind speed determines how quickly the pollutants mix with the surrounding air and how fast they move away from their source. Strong winds tend to lower the concentration of pollutants by spreading them apart as they move downstream. Moreover, the stronger the wind, the more turbulent the air and the more diluted the pollutants are (Ahrens & Henson, 2014), (Venkatram, et al., 2013a).

As with wind, so too can breezes disperse pollutants. They are the result of differences in temperatures. Costal sea-shore areas are subject to morning *marine breezes*. This movement of air is directed from the ocean towards inland, beginning at noon and stopping at nightfall. It can be stronger on warmer days and weaker on cloudy days (APA, 2012). On clear, still nights, a thermal low-pressure area may form over urban areas due to the accumulated heat of the city infrastructure, which is much warmer than the cooler rural areas. This can generate *country breezes*, which move from the countryside to the city. If industrial areas on the outskirts of the city are located in the path of the breezes, pollutants may be carried to the city center, thereby increasing their concentrations (Ahrens & Henson, 2014).

4.5.2.3 Air pressure

High pressure centers in the atmosphere (also known as *anticyclone areas*) are defined by air descending (in spirals) from the higher layers. The air expands at the surface, which warms and stays stable, suppressing ascending movements (necessary for cloud formation and precipitation). Because high pressure centers are characterized by a considerable stability, there is a low vertical mixture, and therefore a weak dispersion of pollutants. On the other hand, low pressure centers are ascending movements, associated with instability conditions and high turbulence, which promote pollutant dispersion (APA, 2012), (Ahrens & Henson, 2014).

4.5.2.4 Weather conditions

Rain is able to wash out water-soluble pollutants and particulate matter in the atmosphere, reducing the concentrations. On the other hand, clear weather (especially under sunlight) can prompt chemical reactions in air pollutants and produce smog. Without wind to disperse pollutants, areas under the influence of clear weather and direct sunlight are subject to high concentrations of pollutants (Ahrens & Henson, 2014). Cloudy or overcast conditions tend to increase pollutants' concentrations, if the wind is calm. Ozone is the exception, since the best conditions for the formation of the chemical compound encompass sunlight and higher temperatures (Grambsch, 2001).

It is important to note that although each individual meteorological variable has some impact on the formation and dispersion of gases, atmospheric pollutants are the result of a complex combination of all variables. None of the variables work in isolation, therefore none of the variables is able to individually explain the dispersion patterns of atmospheric pollutants. Thus, during inferential analysis all the variables should be considered.

4.5.3 Effect of meteorological conditions on exhaust gases concentrations

4.5.3.1 Weather conditions

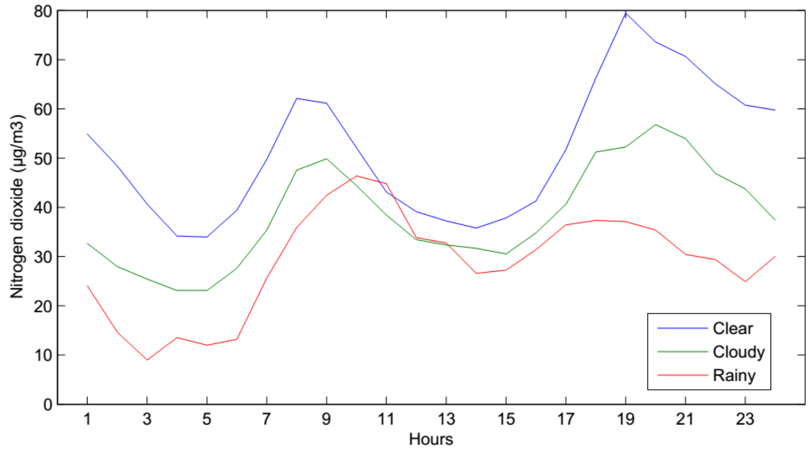
During the four months of study (from September to December 2009), as the weather conditions aggravate (i.e. change from clear to cloudy and from cloudy to rainy); the temperature drops (from 18.8 ºC to 15.5 ºC on average); the wind speed increases (from 11.5 km/h to 19.1 km/h on average); and the humidity also increases (from 65.4% to 89.6% on average). These observations are summarized in Table 4-5:

Weather condition	Temperature (ºC)	Humidity (%)	Wind Speed (km/h)
Clear	18.74	65.58	11.54
Cloudy	17.55	77.80	12.77
Rainy	15.46	89.66	19.02

Table 4-5 - Average temperature, humidity and wind speed with different weather conditions, in Lisbon Municipality.

Figure 4-15 (top) represents the daily NO_2 variation for different weather conditions. During clear weather, we observe a higher concentration of NO_2, compared to cloudy or rainy conditions. One could expect to observe the lowest NO_2 concentrations under clear and stable weather conditions. However, the patterns for exhaust gases dispersion are the result of a complex interaction among several variables (e.g. temperature, wind speed, humidity, weather conditions). The analysis on Table 4-5 indicates that in clear weather, a higher temperature can also be observed. Although higher temperature is an important factor in gas dispersion, NO_2 can be formed in these conditions through a photochemical oxidation of NO in the troposphere during daylight (Clapp & Jenkin, 2001). Additionally, the average low wind speed may not be sufficient to disperse the atmospheric pollutants, resulting in high concentrations of pollutants under direct sunlight (Ahrens & Henson, 2014). A lower concentration of NO_2 was observed during rainy days, when the second concentration peak of the day is less noticeable.

When specific monitoring stations are analyzed, a similar outcome is observed. Figure 4-15 (bottom) represents the average concentrations of NO_2 in "*Av. Liberdade*" station. Higher concentration values are perceived, since this is a traffic station.



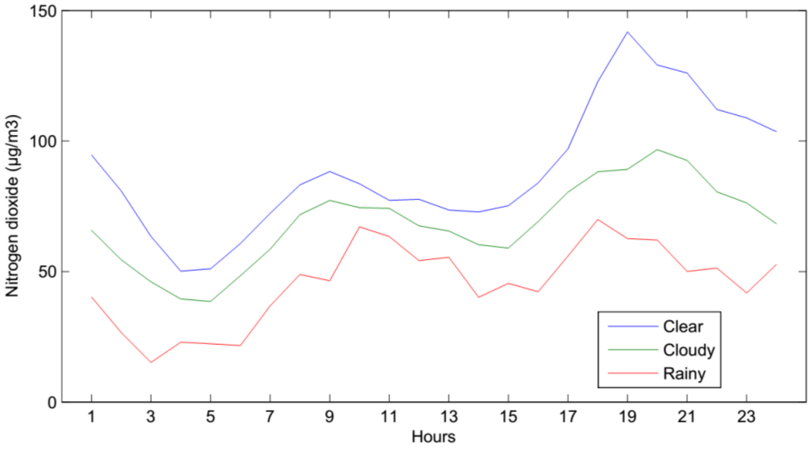


Figure 4-15 - Effect of weather conditions on daily NO_2 concentrations.
Top: considering data from all monitoring stations.
Bottom: considering data from "*Av. Liberdade*" station.

Table 4-6 summarizes the effect of different weather conditions in concentrations of exhaust gases. The same trend is visible: higher concentrations in clear weather, lower concentrations under rainy days. Once again, "*Av. Liberdade*" (bottom), a traffic station, presents higher concentration values for all four exhaust gases.

Weather condition	Clear	Cloudy	Rainy
Nitrogen oxides (µg/m3)	103.53	76.97	49.47
Nitrogen monoxide (µg/m3)	33.96	24.73	13.69
Nitrogen dioxide (µg/m3)	51.45	39.04	28.48
Carbon monoxide (µg/m3)	393.62	326.76	263.48

Weather condition	Clear	Cloudy	Rainy
Nitrogen oxides (µg/m3)	212.08	159.48	87.21
Nitrogen monoxide (µg/m3)	80.65	58.66	27.47
Nitrogen dioxide (µg/m3)	88.43	69.54	45.10
Carbon monoxide (µg/m3)	559.87	432.40	312.82

Table 4-6 - Average concentration of exhaust gases in different weather conditions.
Top: considering data from all monitoring stations.
Bottom: considering data from "*Av. Liberdade*" station.

Also important to note is that the previous analysis is based on data collected between September and December and not during a full year. Relevant and extreme weather periods are missing from the data set, such as July and August, the peak of summer with highest temperatures and best weather conditions, and January and February, the peak of winter with the lowest temperatures and rougher weather conditions. Therefore, the temporal sample (from September to December) could not be representative for the complex patterns taking places along a full year.

4.5.3.2 Wind speed

Wind speed is a major factor which affects atmospheric pollutants. Figure 4-16 (top) plots the variation of exhaust gases' concentrations according to the wind speed variation, considering the average measures from all monitoring stations in Lisbon municipality, from September to December. A clear trend is observable: with the increase of wind speed, there is a steady decrease of all exhaust gases' concentrations. A similar scenario is observable in individual monitoring stations, from which "*Av. Liberdade*" station is an example Figure 4-16 (bottom).

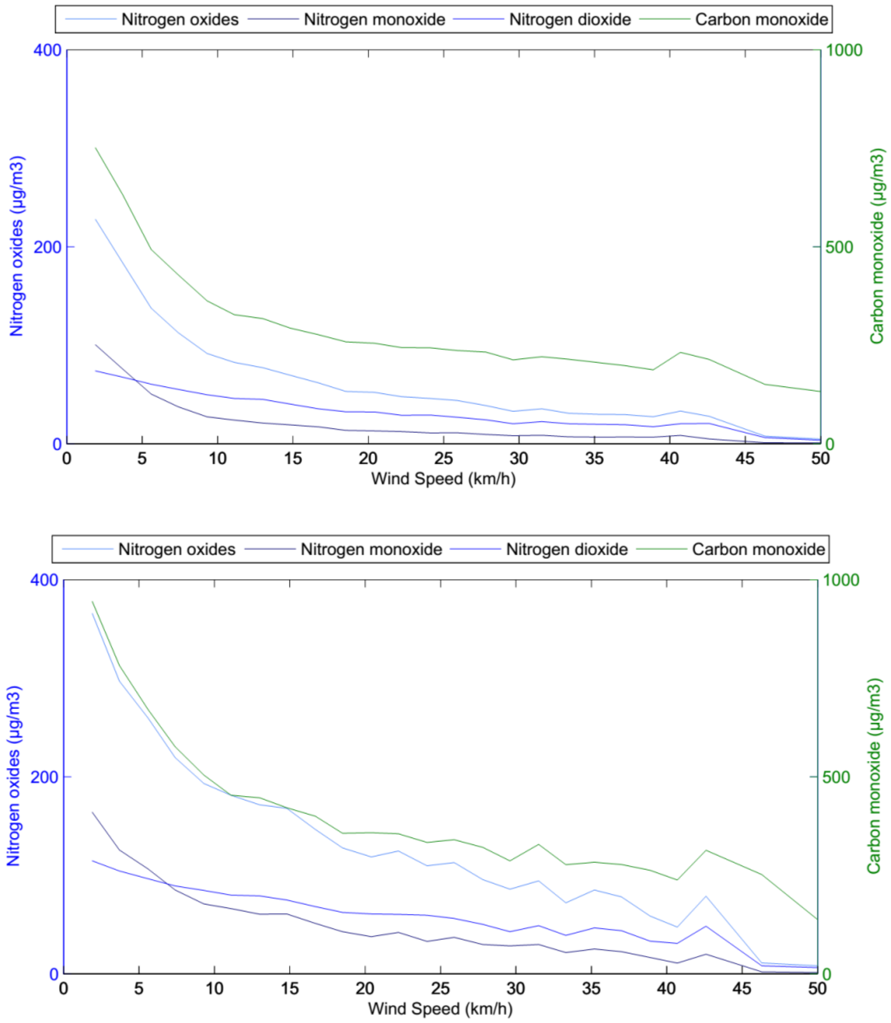


Figure 4-16 - Variation of exhaust gases' concentrations according to wind speed in Lisbon municipality (top) and "*Av. Liberdade*" station (bottom).

This decrease of exhaust gases' concentration with the increase of wind speed, is consistent along all months analyzed. Figure 4-17 plots the effect of wind speed on exhaust gases' concentrations in September (top) and December (bottom). In both months there is a clear decrease of concentrations with the increase of wind speed. However, in September the rate is constant, while in December there is an initial substantial drop of concentrations. Note that in September the average wind speed is lower, with higher average temperature and stable weather conditions, compared to December.

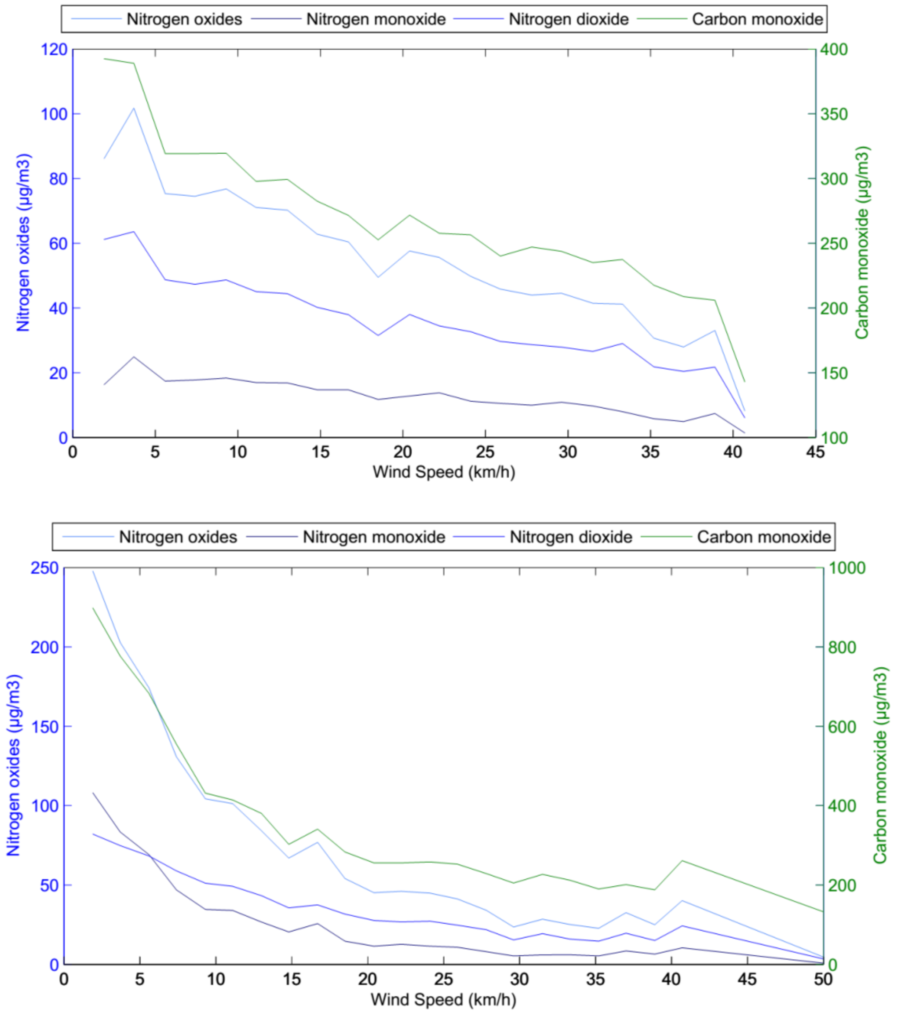


Figure 4-17 - Variation of exhaust gas concentrations according to wind speed in Lisbon municipality, in September (top) and December (bottom)

4.5.3.3 Humidity and Temperature

The effect of temperature and humidity in the variation of exhaust gases' concentrations was also analyzed. However, exhaust gases' concentrations do not appear to have a clear behavior against temperature or humidity, since other variables are simultaneously affecting the atmospheric pollutants.

As pointed out before, a single variable is unable to explain the variation of a gas when monitoring atmospheric pollution, due to the complex interactions involved with other variables. This may justify the absence of a clear and regular pattern of exhaust gases' concentrations when the humidity or temperature varies.

4.5.4 Correlation between exhaust gases

As discussed before, NO_2, NO and CO have natural sources, but also anthropogenic sources, usually from fuel burning at high temperatures (from domestic heating or internal combustion vehicles, traffic being one of the main sources of pollutant in urban areas (Zavala, et al., 2006), (Karlsson, 2004)). Moreover, NO_x is composed by NO_2 and NO, and NO emitted to the atmosphere can form NO_2. Therefore, in urban areas, these atmospheric pollutants appear to have similar anthropogenic sources.

To understand to what extent they could be related, the Pearson's coefficient of correlation is computed between them. Table 4-7 presents the correlation among NO, NO_2, NO_x and CO, considering data collected from all monitoring stations. The coefficient of correlation of Pearson (r) between two variables (X,Y) is defined as (Devore & Berk, 2012):

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (20)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the arithmetic mean of variable X and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the arithmetic mean of variable Y, with $X = (x_1,..,x_n)$ and $Y = (y_1,..,y_n)$, two variables of size n.

Overall, exhaust gases are strongly correlated. CO and NO_x present the highest correlation with other exhausted gases, while NO_2 is more divergent, especially the correlation with NO. These observations are in line with Lin et al. (Lin, et al., 2011). The authors observed that NO_x, CO, and SO_2 are strongly correlated among each other, hypothesizing that these exhaust gases are emitted by some common sources.

Exhaust Gases	**NO_2**	**NO**	**NO_x**	**CO**
NO_2	1	0.7563	0.8722	0.8314
NO		1	0.9797	0.9360
NO_x			1	0.9548
CO				1

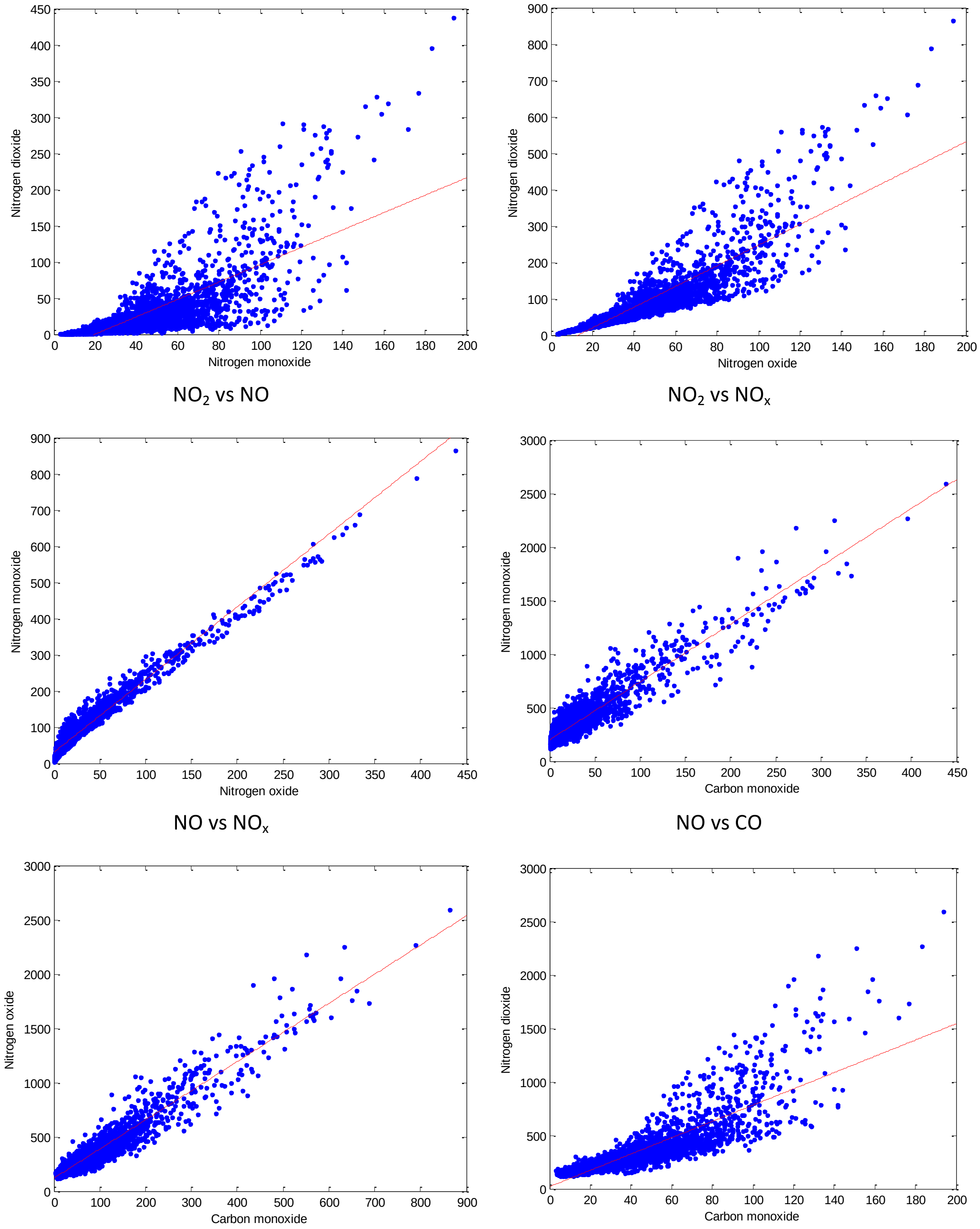
Table 4-7 - Coefficient of correlation of Pearson between exhaust gases in Lisbon municipality.

Table 4-8 shows a similar experiment but only using data from "*Av. Liberdade*" station. This is a traffic station, located in the vicinity of intensive traffic. In this station, the correlation of CO with the remaining exhaust gases decreases in strength from the previous scenario, while NO_2, NO and NO_x strengthen the correlations. Overall, the exhaust gases are strongly correlated in this traffic station, as observed using data from all monitoring stations. Moreover, the strong correlation between NO_2 and other exhaust gases is an indication that NO_2 can be used to estimate the concentrations of other exhaust gases. This is a relevant observation since NO_2 is usually used as a marker for traffic emissions (Krzyzanowski & Schneider, 2005).

Exhaust Gases	**NO_2**	**NO**	**NO_x**	**CO**
NO_2	1	0.8492	0.9187	0.8245
NO		1	0.9887	0.8605
NO_x			1	0.8774
CO				1

Table 4-8 - Coefficient of correlation of Pearson between exhaust gases in "*Av. Liberdade*" station.

Figure 4-18 represents the correlation among exhaust gases, using a scatter plot and fitted linear function.



NO_2 vs NO

NO_2 vs NO_x

NO vs NO_x

NO vs CO

NO_x vs CO

NO_2 vs CO

Figure 4-18 - Scatter plot and fitted linear function between different exhaust gases (in µg/m^3).

4.5.5 Nitrogen dioxide profile

According to Krzyzanowski et al. (Krzyzanowski & Schneider, 2005), NO_2 is a marker for combustion processes and an indicator of fresh exhaust-pipe emissions near roads. As expressed before, NO emitted to the atmosphere can produce a rapid photochemical oxidation forming NO_2 in the troposphere during daylight (Clapp & Jenkin, 2001). Because NO can potentially be converted to NO_2 in this process, it is typical to express NO_2 when making emission estimates (Colls & Tiwary, 2009). In our analysis, only nitrogen dioxide (NO_2) is considered, since it is often considered a marker for transport-related air pollution (Becker, et al., 2000).

From NO_2, one can estimate the concentrations of the other remaining exhaust gases. In our study, we found a strong correlations between the NO_2 concentration and other exhaust gases' concentration; NO (r = 0.8492), NO_x (*r* = 0.9187), and CO (*r* = 0.8245); as discussed in the previous section.

NO_2 behaves similarly in every station, as shown in Figure 4-19. The same two daily peaks are observed, although with different intensities. On average, traffic stations sense higher concentration of exhaust gases (57.4 μg/m^3 for NO_2, during four months of data) than background stations (33.7 μg/m^3 for NO_2, during four months of data), which is in line with Pleijel et al. (Pleijel, et al., 2004), Gilbert et al. (Gilbert, et al., 2003), Zou et al. (Zou, et al., 2006) and Ndoke & Jimoh (Ndoke & Jimoh, 2005). The authors observed that concentrations of exhaust gas decrease when moving away from the roads. Colls & Tiwary (Colls & Tiwary, 2009) stated that NO_2 concentrations on traffic stations can be extremely high in comparison to background levels. This scenario can be observed in Table 4-9, where on average, NO_2 concentrations at the traffic stations are 42% higher than in background stations.

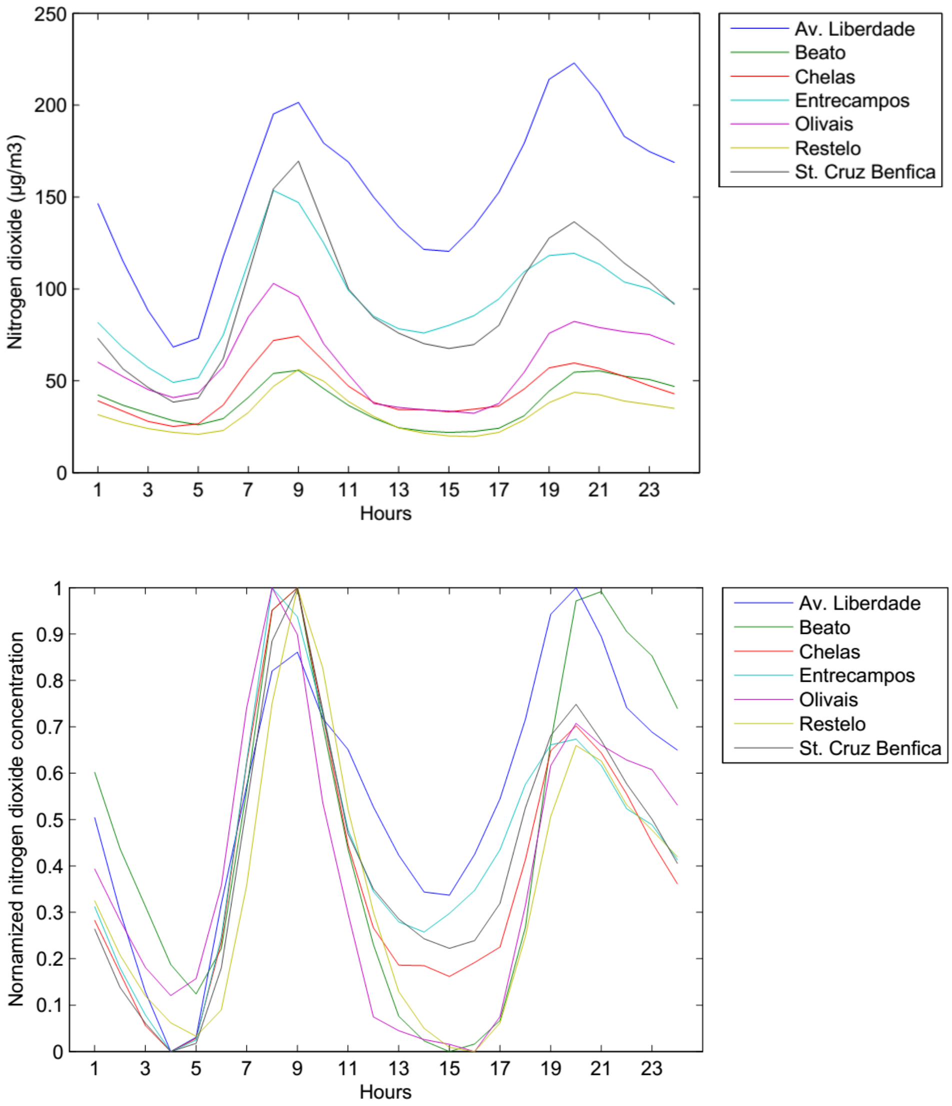


Figure 4-19 – Average variation of NO_2 over the course of one day from different stations (top) and normalized data (bottom), during four months of data.

By observing the normalized plot (Figure 4-19, bottom) one can verify that NO_2 reaches the minimum value in traffic stations around 4 AM, while in some background stations the minimum NO_2 value is achieved between 3 PM and 4 PM (e.g. “*Olivais*”, “*Beato*” and “*Restelo*”). As discussed before, traffic stations are deployed to sense exhaust gases directly from traffic, and during business hours there is a continuous emission of atmospheric pollutants from that source. Although traffic starts to reduce the activity in the evening, only between 1 AM and 5 AM does it reach minimum activity. On the other hand, background stations are deployed to perceive exhaust gases from different sources (e.g. domestic heating or industrial burning), other than traffic. In that sense, some of those sources could be active even during the night.

Station	Type	Avg	Max	Min	Std Dev
Avenida Liberdade	Traffic	77.27	390.15	5.26	49.64
Beato	Background	32.88	156.83	1.91	25.39
Chelas	Background	36.45	149.18	1.91	26.21
Entrecampos	Traffic	53.27	266.32	1.43	36.91
Olivais	Background	40.15	180.73	1.43	29.27
Restelo	Background	27.81	180.06	2.86	22.43
Santa Cruz de Benfica	Traffic	45.28	205.59	1.91	29.37

Table 4-9 - Nitrogen dioxide average, maximum and minimum concentrations (μg/m^3) in each monitoring station (data collected during a four months period, from September to December 2009).

The seasonal variation of NO_2 concentrations follow the patterns observed with other exhaust gases: lower concentrations in summer and higher concentrations in winter (Figure 4-20). This behavior is similar at all monitoring stations. Once again, traffic stations show higher concentration values than background stations, with larger amplitude between maximum and minimum values achieved on traffic stations.

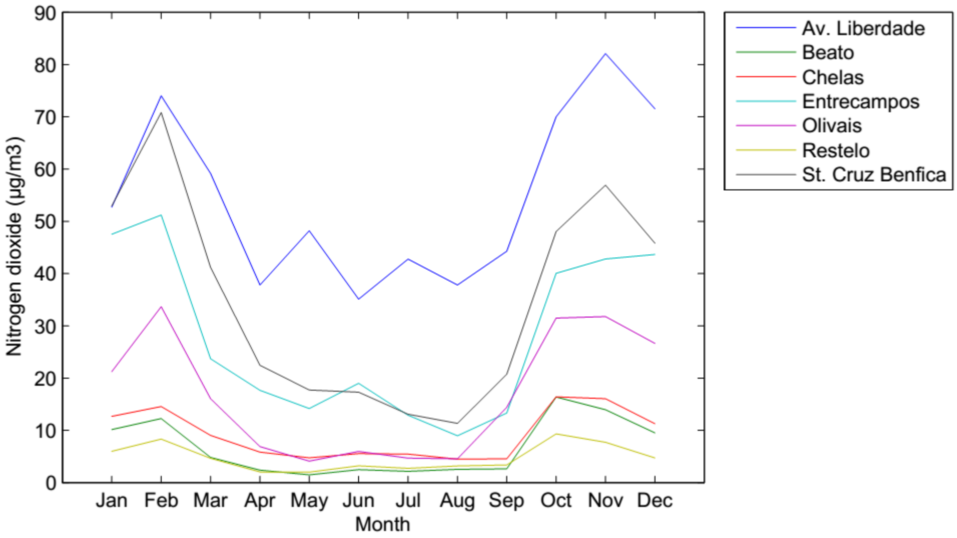


Figure 4-20 - Variation of nitrogen dioxide concentration on each monitoring during a year (2009).

Figure 4-21 represents the location of each monitoring station and the corresponding average concentration of NO_2 (during four months of data, from September to December 2009).

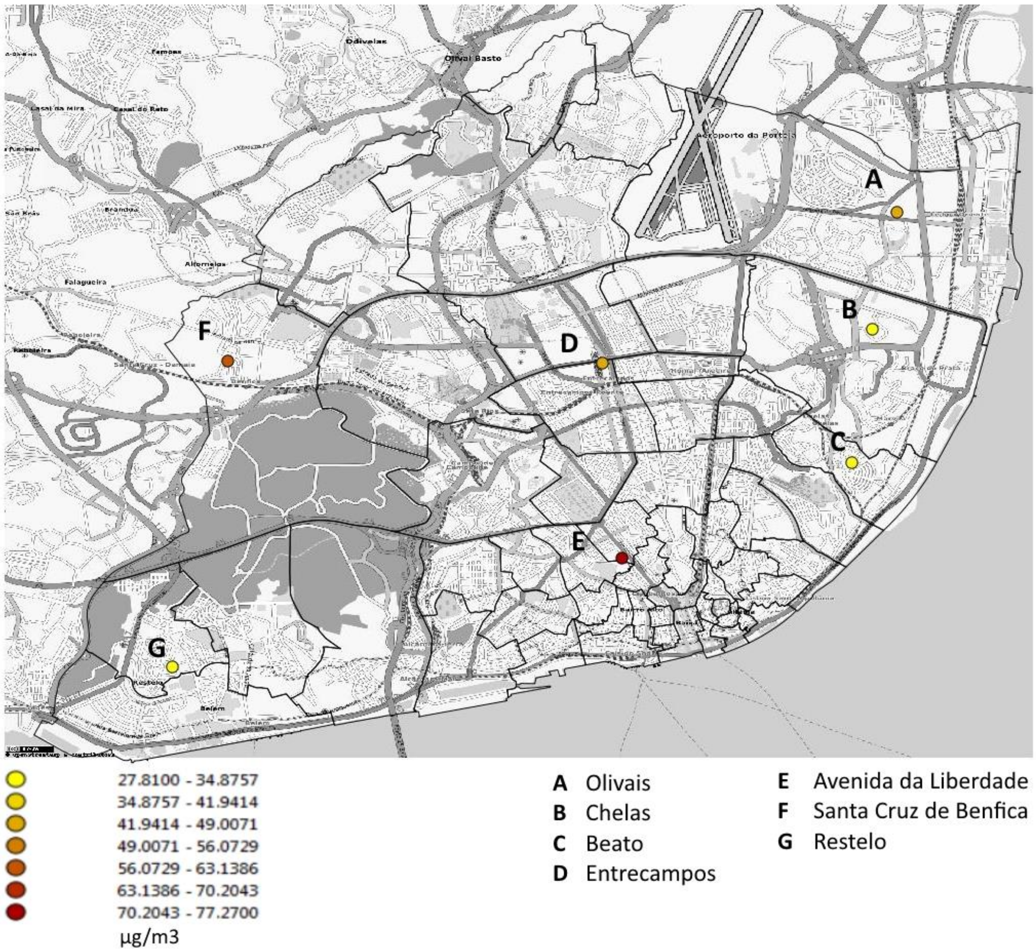


Figure 4-21 - Average concentration of nitrogen dioxide (µg/m^3) in each monitoring station.

"*Av. Liberdade*" is one of the most important stations, since it is located in the middle of the city, adjacent to a major road. By observing the data from this station, we were able to strongly correlate the concentration of NO_2 with the other stations. The coefficient of correlation value ranges from r = 0.923 ("*Entrecampos*", a traffic station) and r = 0.758 ("*Beato*", a background station). Therefore, the following inference analysis will explore techniques to model NO_2 concentrations (as a representative exhaust gas) with data collected at '*Av. Liberdade*' monitoring station (as a predominant monitoring station), from September to December 2009. At the end of the analysis, the same procedure will be applied to the remaining monitoring stations.

4.6 Data analysis and results: using taxis as probe for exhaust gases

The previous section has shown that distinct exhaust gases could be related among themselves and are affected by other variables. However, this interplay may not be straightforward due to their complex interdependencies.

Since NO_2 can be used as a marker for exhaust-pipe emissions and it is strongly correlated with other exhaust gases (NO, NO_x and CO), our goal was to estimate the NO_2 concentration in each hour given:

- hour of the day (*TH*= {1, 2, …, 24});
- day of the week (*D* = {Sunday, …, Saturday});
- weather conditions (*W* = {Clear, Cloudy, Rainy});
- temperature (*T* = [0,34]);
- humidity (*H* = [0,100]);
- wind speed (*WS*=[0,50]);
- wind direction (*WD* = [0,360]);
- number of taxi pick-ups (*TP* = [0,25]);
- number of taxi drop-offs (*TD* = [0,25]);
- number of distinct taxi vehicles during pick-ups (*TPV* = [0,20]) ;
- number of distinct taxi vehicles during drop-off (*TDV*=[0,20]);
- average taxi speed (*TS* = [1,120]); and
- number of taxi-GPS samples (*TG* = [1,605]).

These are our 13 independent explanatory variables. The NO_2 (or any other exhaust gas) does not have a strong correlation with any of the other aforementioned variables individually. Among the highest correlations found, NO_2 correlates best with wind speed (r = 0.3601) and humidity (r = 0.3368).

The procedure was, firstly, applied to data from "*Av. Liberdade*" station, from which the results that are presented in the next subsections were obtained. (This is one of the most important monitoring stations, since it is located right in the center of the city, near an important road). At the end of the experiment, the same procedure was applied to the data from the remaining monitoring stations, from which a similar outcome was observed and it is summarized in the last section of this chapter.

4.6.1 Searching for a linear relationship

Aiming for a simple model, we start by exploring a linear relationship between the dependent variable *Y* (represented by NO_2) and the explanatory variable *X* (represented by the vector {*TH, D, W, T, H, WS, WD, TP, TD, TPV, TDV, TS, TG*}), as suggested by Donelly et al. (Donnelly, et al., 2015). As shown before, simple linear regression is not suitable for the current setup, since, individually, the independent variables are unable to explain the variation or behavior of NO_2. Therefore, we will explore the multiple linear regression, which attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. This technique allows the identification of the percentage of the dependent variable explained by the explanatory variables and the contribution of each explanatory variable. The estimation of the coefficients is based on the method of ordinary least squares, consisting of minimizing the sum of squares of residuals (Pallant, 2005).

Given a data set $\{y_i, x_{i1},\dots,x_{ip}\}$, $i=1,\dots,n$, a linear regression model assumes that the relationship between the dependent variable y_i and the vector of independent variables x_i is linear taking the form (Freedman, 2009):

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad (21)$$

with *i* = 1,…,n (the number of instances) and *εi* an error random variable that adds noise to the linear relationship between the dependent variable and the independent variables. This approach assumes linear combination of the regression coefficients and the predictor variables, and constant variance in the errors (*homoscedasticity*) (Freedman, 2009).

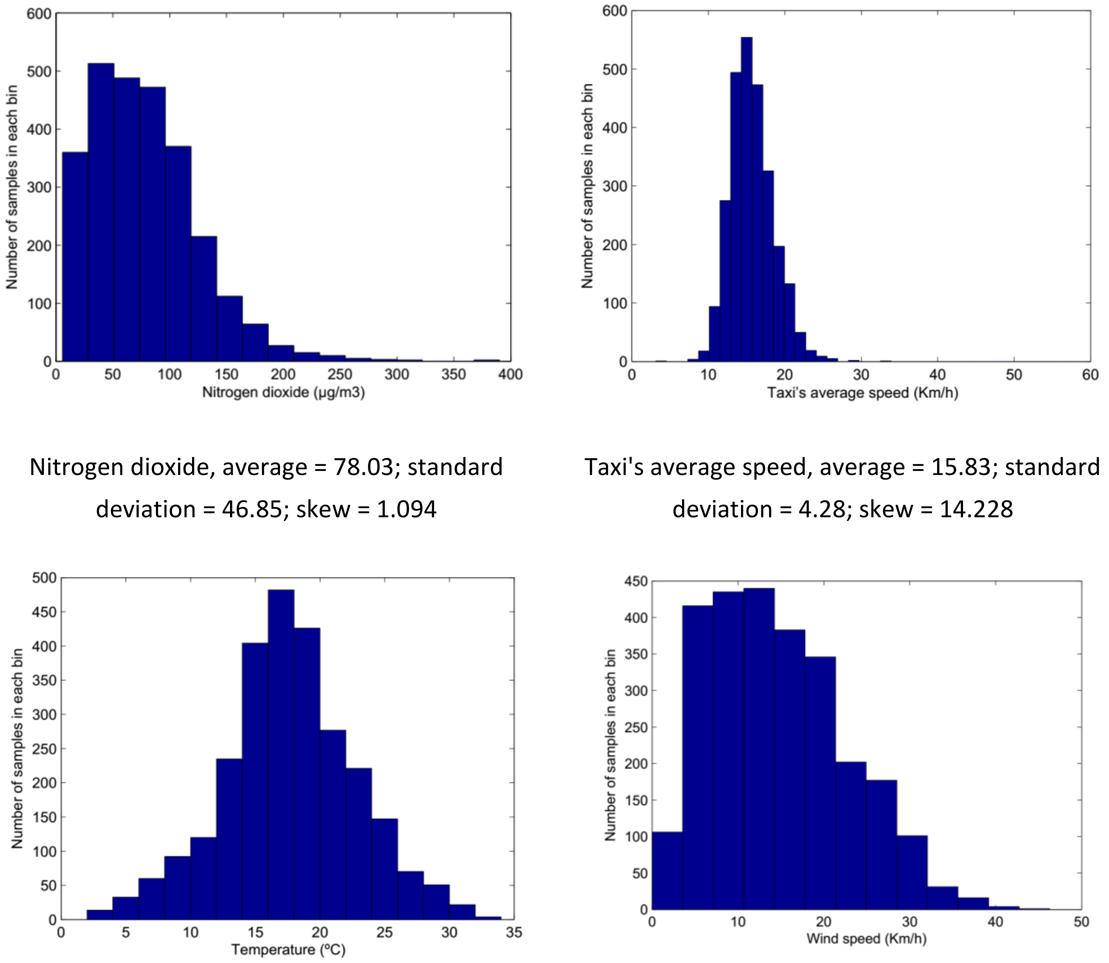
Based on a holdout method (the oldest 2/3 forming the training set and the latest 1/3 forming the testing set), the model takes the form:

$$\begin{aligned} NO2 = 2.090 \times TH + 0.771 \times D - 5.887 \times W - 2.813 \times T \\ - 0.8269 \times H - 2.741 \times WS + 0.05437 \times WD \\ + 2.137 \times TP + 2.08 \times TD - 1.434 \times TPV \\ - 1.264 \times TDV + 0.0801 \times TS - 0.865 \times TG \\ + 204.5 \end{aligned} \tag{22}$$

Using this model, we are able to estimate the value of NO_2 with the coefficient of correlation (*r*) of 0.6120. By removing error samples with empty values or anomalous data (e.g. negative values of exhaust gas, humidity, or wind speed), we observe a slight increase of the coefficient of correlation to 0.6215.

Furthermore, we can also observe that the majority of the variables (NO_2, temperature, wind speed and taxi-related parameters) followed a Gaussian distribution, usually with a positive skew (Figure 4-22). Therefore, we apply the three-sigma rule of thumb (or 68–95–99.7 rule), a conventional heuristic that states almost all values lie within three standard deviations of the mean (Kazmier, 2003). By doing so, we slightly improve the coefficient of correlation to 0.6289, with a corresponding coefficient of determination (r^2) of 0.3955.

Moreover, hour of the day, temperature, humidity, wind speed, and number of pick-ups are the dominant variables, with a higher contribution to the process. By removing one of them, the coefficient of correlation drops considerably - 14% on average, while wind speed was responsible for a drop of 23%.



Nitrogen dioxide, average = 78.03; standard deviation = 46.85; skew = 1.094

Taxi's average speed, average = 15.83; standard deviation = 4.28; skew = 14.228

Temperature, average = 18.15; standard deviation = 5.16; skew = 0.041

Wind speed. average = 14.35; standard deviation = 8.23; skew = 0.534

Figure 4-22 – Histogram representation of the distribution of samples into groups of ranges, for nitrogen dioxide, taxis' average speed, temperature, and wind speed.

4.6.2 Selecting predictor variables

4.6.2.1 Overfitting evaluation

Overfitting takes place in complex models, due to excess of data, when the regression model adapts to random errors or noise instead of the underlying unbiased estimates of properties and relationship within the population, thus not reflecting the overall population (Meyer & Krueger, 2004). Excess of predictors can lead to overfitting, which prevents our model to fit new samples. Although our data set safely fulfills the rule of thumb of 10-15 observations for each term (Meyer & Krueger, 2004), some techniques can be used to prevent overfitting. The analysis of multicollinearity and predicted r^2 are adequate to verify if the model is suitable. Additionally, the adjusted r^2 should also be considered (Meyer & Krueger, 2004).

Adjusted r^2 compares the explanatory power of regression models which contain different numbers of predictors. It is a modified version of the regular r^2, adjusted for the number of predictors in the model. The adjusted r^2 is always lower than regular r^2 and can present negative values. It increases only if the new term improves the model more than would be expected by chance, thus contributing to the model. It decreases when a predictor improves the model by less than expected by chance, thus not contributing significantly to the model (Meyer & Krueger, 2004), (Hogg & Ledolter, 1987).

Predicted r^2 indicates how well a regression model predicts responses for new observations, and can prevent an overfitting model. It is usually calculated by systematically removing each observation from the data set, estimating the regression equation, and determining how well the model predicts the removed observation. This can be viewed as a cross-validation. Predicted r^2 is always lower than r^2 and can present negative values. A predicted r^2 value much lower than the regular predicted r^2 is an indication of model overfitting and an excess of predictors (Meyer & Krueger, 2004), (Hogg & Ledolter, 1987).

Multicollinearity in multiple linear regression is a condition that takes place when one or more predictor variables in the model are correlated with other predictor variables. As a result, this situation can increase the variance of the regression coefficients, and is an indication of redundant variables (Meyer & Krueger, 2004). To measure multicollinearity, we can examine the variance inflation factors (VIF). It measures how much the variance of an estimated regression coefficient increases if the predictor variables are correlated. Absence of multicollinearity is achieved if all VIFs are equal to one. VIFs between 5 and 10 are a clear indication of correlation among predictor variables. VIFs above 10 are evidence of multicollinearity among predictor variables, thus affecting the estimation of the regression coefficients (Meyer & Krueger, 2004).

4.6.2.2 Inspecting predictor variables

Table 4-10 shows the coefficient statistics for the regression model. The variance inflation factor row shows the correlation for each coefficient. Most of the coefficients present a VIF value around the unit, which is desirable. However, four coefficients present VIF values well above 10, meaning a strong correlation among them: Number of distinct Taxi Vehicles during Pick-ups (*TPV*); Number of Taxi Pick-ups (*TP*); Number of distinct Taxi Vehicles during Drop-off (*TDV*); and Number of Taxi Drop-offs (*TD*). The *p*-value above 0.05 also indicates a low significance of the coefficients. Therefore, predictor variables Number of distinct Taxi Vehicles during Pick-ups (*TPV*) and Number of distinct Taxi Vehicles during Drop-off (*TDV*) do not contribute to the model and should be removed.

Term	**Coefficient**	**Standard Error of the Coefficient**	***T*-Value**	***P*-Value**	**Variance Inflation Factor**
Constant	204.50	7.51	27.23	0.000	
Hour (TH)	2.0900	0.11	19	0.000	1.13
Day of the Week (D)	0.7710	0.356	2.16	0.031	1.02
Weather Conditions (W)	-5.8870	0.967	-6.09	0.000	1.33
Temperature (T)	-2.8130	0.181	-15.58	0.000	1.72
Humidity (H)	-0.8269	0.0551	-15.02	0.000	1.94
Wind Speed (WS)	-2.7410	0.1	-27.38	0.000	1.35
Wind Direction (WD)	0.0543	0.00623	8.73	0.000	1.27
Number of distinct Taxi Vehicles during Pick-ups (TPV)	-1.4340	0.774	-1.85	0.064	17.45
Number of Taxi Pick-ups (TP)	2.1370	0.916	2.33	0.020	17.33
Number of distinct Taxi Vehicles during Drop-off (TDV)	-1.2640	0.872	-1.45	0.147	21.80
Number of Taxi Drop-offs (TD)	2.0800	1	2.08	0.038	21.41
Number of Taxi-GPS Samples (TG)	0.0810	0.0225	3.57	0.000	1.11
Average Taxi Speed (TS)	-0.8650	0.172	-5.02	0.000	1.08

Table 4-10 – Coefficient statistics for multiple linear regression.

4.6.2.3 Excluding non-significant variables

In order to assure that only the significant predictor variables are kept in the regression model, the Stepwise regression method is applied, with a forward selection. In this method the model starts with no variables. Each variable is tested to verify if it improves the model, based on partial *F*-tests (i.e., the *T*-tests). The process stops when those variables statistically significant are added to the model and no further improvement is possible (Simon & Kwanisai, 2003). The alpha significance level to add

(Alpha-to-Enter, α_E) and to remove (Alpha-to-Remove, α_R) variables to the model was set to 0.15 (Simon & Kwanisai, 2003).

Term	Coefficient	Standard Error of the Coefficient	*T*-Value	*P*-Value	Variance Inflation Factor
Constant	206.23	7.50	27.48	0.000	
Hour (TH)	2.126	0.109	19.46	0.000	1.11
Day of the Week (D)	0.853	0.356	2.39	0.017	1.02
Weather Conditions (W)	-5.832	0.969	-6.02	0.000	1.33
Temperature (T)	-2.792	0.180	-15.50	0.000	1.71
Humidity (H)	-0.8341	0.0550	-15.16	0.000	1.93
Wind Speed (WS)	-2.735	0.100	-27.32	0.000	1.34
Wind Direction (WD)	0.05381	0.00623	8.63	0.000	1.27
Number of Taxi Pick-ups (TP)	0.690	0.205	3.36	0.001	1.20
Number of Taxi-GPS Samples (TG)	0.0849	0.0223	3.80	0.000	1.10
Average Taxi Speed (TS)	-0.887	0.172	-5.14	0.000	1.08

Table 4-11 – Coefficient statistics for multiple linear regression after Stepwise method was applied.

After the use of the Stepwise method (Table 4-11), all the coefficients show VIF values around the unit, an indication of no correlation among the predictor variables, as expected. Additionally, the *p*-value is always lower than 0.05 (the highest achieved being 0.017), which confirms the significance of the result. With this new line-up of predictor variables, a new model is produced, defined by

$$\begin{aligned} NO2 = 2.126 \times TH + 0.85 \times D - 5.832 \times W - 2.792 \times T \\ - 0.8341 \times H - 2.735 \times WS + 0.05381 \times WD \\ + 0.690 \times TP + 0.0849 \times TS - 0.887 \times TG + 206.23 \end{aligned} \tag{23}$$

The coefficient of correlation (*r*) is 0.6254, with the correspondent coefficient of determination (r^2) of 0.3912. The adjusted r^2 is 0.3889, while the predicted r^2 is 0.3790. Since the adjusted r^2 and predicted r^2 don't diverge significantly from the regular r^2 (a maximum of 0.0122), it is reasonable to assume the absence of overfitting in the model.

4.6.2.4 Exploring improvements for variables' selection

In the previous section, we observed that the dominant variables were hour of the day (*TH*), temperature (*T*), humidity (*H*), wind speed (*WS*), and number of pick-ups (*TP*). By using only these five predictor variables, we are able to model NO_2 concentrations with a coefficient of correlation of 0.5937 and corresponding coefficient of determination of 0.3525, with an adjusted r^2 of 0.3513 and a predicted r^2 of 0.3494. The model is defined by:

$$NO2 = 2.242 \times TH - 2.650 \times T - 0.9325 \times H - 2.5155 \times WS + 0.933 \times TP + 197.35 \qquad (24)$$

Table 4-12 shows the coefficients statistics using the five dominant predictor variables. As expected, VIF values for all coefficients approach the unit and *p*-values are zero, thus assuring the absence of correlation between predictor variables, and ensuring the significance of the coefficient of determination.

Term	Coefficient	Standard Error of the Coefficient	*T*-Value	*P*-Value	Variance Inflation Factor
Constant	197,35	6.670	29.57	0.000	
Hour (TH)	2.242	0.111	20.15	0.000	1.09
Temperature (T)	-2.65	0.183	-14.51	0.000	1.66
Humidity (H)	-0.9325	0.0512	-18.20	0.000	1.66
Wind Speed (WS)	-25.155	0.0907	-27.73	0.000	1.04
Number of Taxi Pick-ups (TP)	0.933	0.198	4.71	0.000	1.06

Table 4-12 – Coefficient statistics for multiple linear regression using five dominant predictor variables.

Although the value of coefficient of correlation using the dominant five variables is slightly lower than the value obtained from the previous approach (using 10 uncorrelated variables), it shows that with fewer variables it is possible to achieve almost similar accuracy.

Table 4-13 summarizes the results from the experiments up until now. It presents the comparison of performances for multiple linear regression using different sets of predictor variables. Highest coefficient of correlation was achieved using all predictor variables (13), however this produced overfitting, since several predictor variables were correlated. By removing the correlated variables, the coefficient of correlation drops, but assures the absence of overfitting of the model. Finally, by using just the five predominant variables, the performance of the model is reasonably close to the previous two scenarios. Nevertheless, the goodness of the fitting (coefficient of determination) from the predicted model is poor in all cases. The next sections will explore possible improvements.

Number of predictor variables	**Coefficient of Correlation (*r*)**	**Coefficient of Determination (*r^2*)**
13 predictor variables	0.6289	0.3955
10 predictor variables	0.6254	0.3912
5 predictor variables	0.5937	0.3525

Table 4-13 – Comparison of performances using different setups for predictor variables.

4.6.3 Searching for model improvements

Moving further along in improving our understanding of the correlation between the variables, we applied the principal component analysis (PCA) to all 13 initial predictor variables. This technique allows us to identify patterns in data, emphasizing the similarities and differences between sets of observations, through orthogonal transformation, expressed in terms of principal components with the highest possible variance. These orthogonal components are the eigenvectors of the covariance matrix. An additional advantage is the possibility to reduce dimensionality and complexity without losing too much data, thus reducing the size of the data set (Jolliffe, 2002).

As a first step, a correlation matrix for our non-normalized data is constructed to standardize the data (since the correlation matrix is the standardized covariance matrix). After computing the eigenvalues and eigenvectors, we keep those components with an eigenvalue greater than one (Figure 4-23), following the Kaiser's rule (Larsen & Warne, 2010) – thus assuring components with higher variance (assuming each original variable has variance 1), noting that eigenvalues are a measure of variance.

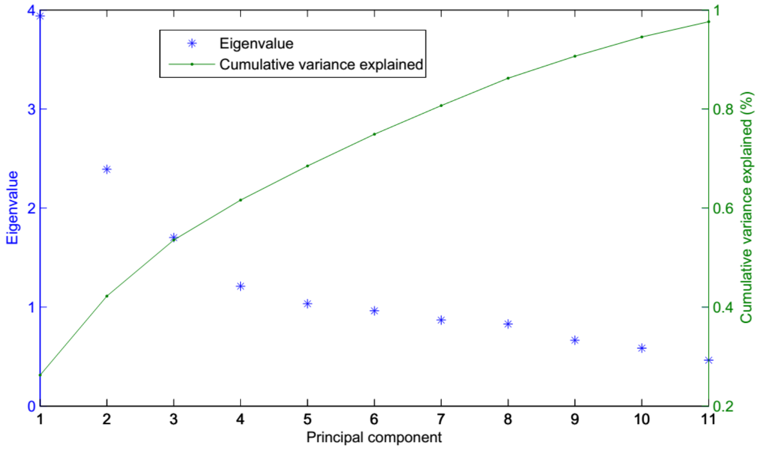


Figure 4-23 - Scree plot of eigenvalues and cumulative variance explained for each principal component.

The first four principal components (with eigenvalues greater than 1) account for 69% of the variance, i.e. can explain 69% of the variance of the data. The first two components alone (with eigenvalues 3.939 and 2.391) account for nearly 50% of the variance. Also, noticeably, the dominant variables of the first eigenvector (with absolute value of the loadings greater than 0.3) are taxi-related (taxi pick-up, taxi drop-off, and number of distinct vehicles), while the dominant variables of the second eigenvector are temperature, humidity, and wind speed.

Using the two principal components, we estimate the value of the NO_2, using a linear relation with a coefficient of correlation of 0.5144. Although the resulting correlation values are not significant (and are noticeably inferior to the previous approach), the explored procedure holds some interest: a) demonstrates that there is an interplay between the exhaust gas and other urban variables; b) the data can be reduced without compromising too much information. Nevertheless, this procedure does not improve on the original model. Therefore, new techniques should be explored.

4.6.4 Finding a better approach

The previous section discusses the relation between exhaust gases and other urban variables. However, the resulting model from multiple linear regression does not show a significant correlation between them. Therefore, we pursue another approach to this problem. Several authors suggested that a multilayer perceptron with backpropagation is a fit and robust technique to model variables related with exhaust gases or meteorological events (Rosenblatt, 1958), (Shi & Harrison, 1997), (Gardner & Dorling, 1999), (Kolehmainen, et al., 2001), (Perez & Reyes, 2001), (Kukkonen, et al., 2003), (Agirre-Basurko, et al., 2006), (Juhos, et al., 2008), (Ahmed, et al., 2010). Artificial Neural Networks (ANN) can provide a robust approach to approximate discrete-valued target functions, which has been inspired by the complex web of interconnected natural neurons (MacKay, 2003).

A multilayer perceptron system is based on a unit (perceptron) that calculates a linear combination of an input of real-values and outputs a Boolean value (1,-1) if the result is greater or lower than a certain threshold. The backpropagation algorithm learns the weights of a multilayer network using the gradient descent to minimize the squared error between the network outputs and the target values (Mitchell, 1997). Ahead we present the variable selection for this process.

4.6.4.1 Variables selection

Previous sections have shown that some of the predictor variables were correlated among them: Number of distinct Taxi Vehicles during Pick-ups (*TPV*); Number of Taxi Pick-ups (*TP*); Number of distinct Taxi Vehicles during Drop-off (*TDV*); and Number of Taxi Drop-offs (*TD*). This subset is composed by taxi-related variables. May et al. (May, et al., 2011) and Fernando et al. (Fernando, et al., 2005) recommend avoiding using correlated input variables in artificial neural networks, as the performance of the final model is heavily dependent on the input variables used to develop the model. In multidimensional systems, correlated variables create redundancy and affect the efficiency of the model to obtain good generalization with finite data (Fernando, et al., 2005). Therefore, only Number of Taxi Pick-ups (*TP*) is kept, removing the remaining correlated variables from the original set. Number of Taxi Pick-ups (*TP*) was selected since previous experiments have shown it is a better predictor than the other taxi-related variables.

This step produces a set of 10 non-correlated variables (from an original set of 13) to be used as predictors for NO_2 concentrations: hour of the day (*TH*); day of the week (*D*); weather conditions (*W*); temperature (*T*); humidity (H); wind speed (*WS*); wind direction (*WD*); number of taxi pick-ups (*TP*); average taxi speed (*TS*); and number of taxi-GPS samples (*TG*).

4.6.4.2 Artificial neural network configuration

In our experiment, we set up a multilayer perceptron with backpropagation and 15 hidden layers; a sigmoid activation function; training time of 500 epochs; and a learning rate to update weights of 0.3. The inputs correspond to the 10 non-correlated predictor variables and the output is an estimation of NO_2 concentrations. Each sigmoid node has 10 inputs, corresponding to the predictor variables.

Considering the size of the data set (four months of data) and the existence of a timestamp, the samples are organized into training and testing subsets, following a holdout configuration (the oldest 2/3 forming the training set and the latest 1/3 forming the testing set).

We are able to estimate the value of NO_2 with a coefficient of correlation of 0.7869 and the corresponding coefficient of determination of 0.6192. Compared with the previous approaches (multiple linear regression), which produced a model with a coefficient of correlation of 0.6254 and the correspondent coefficient of determination of 0.3912, this is a significant improvement. Figure 4-24 shows the fitted linear function (defined by $Y = 0.7063X + 50.304$) of real value of NO_2 against the estimated value of NO_2.

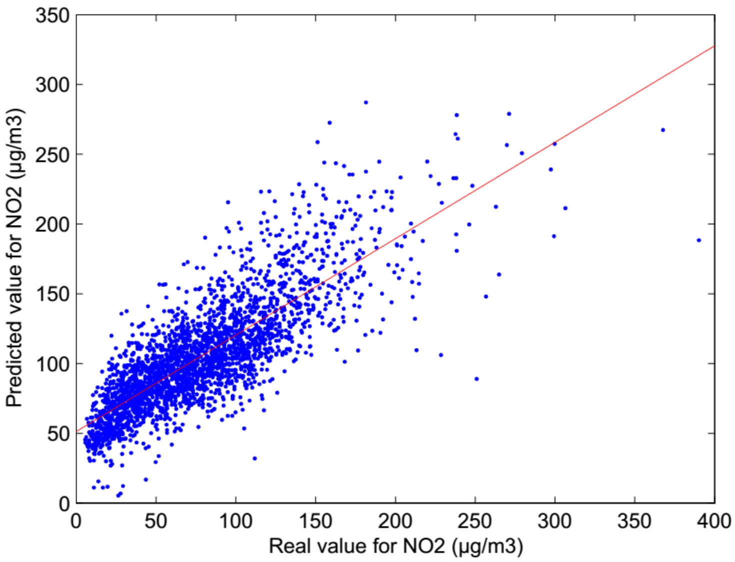


Figure 4-24 - The fitted linear function of real value of NO_2 against the estimated value of NO_2.

4.6.4.3 Quality assessment

The model produces a Root Mean Squared Error (RMSE) of 33.7418, while the Normalized RMSE (NRMSE) was 0.1172, and a Mean Absolute Error (MAE) of 27.2792. The low divergence between RSME and MAE are an evidence of small variances in the individual errors.

Root Mean Squared Error (RMSE) or Root Mean Squared Deviation (RMSD) measures the difference between the real value and the estimate value (quadratic scoring), and it is defined by:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \tag{25}$$

where $\hat{y}_t$ represents the predicted value and y_t represents the observed value, in this case the measured value of NO_2. Since errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors (Devore & Berk, 2012). The Normalized Root Mean Squared Error (NRMSE) is defined by:

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \tag{26}$$

Mean Absolute Error (MAE) measures the average magnitude of the errors (linear scoring), to analyse how close predictions are to the eventual outcomes, and it is defined by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_t - y_t| \tag{27}$$

where, as in the RMSE, $\hat{y}_t$ represents the predicted value and y_t represents the observed value. The RMSE is expected to be larger or equal to the MAE, and the greater difference between them, the greater the variance in the individual errors in the sample. Both MAE and RMSE can range from zero to infinity, where lower values are preferable since they represent smaller errors between forecast and measured values (Devore & Berk, 2012).

4.6.4.4 Performance under different environments

Impact of hidden layers

Several experiments were performed in order to derive the values for these parameter configurations. To identify the optimum number of hidden layers, the highest coefficient of determination was obtained with 33 hidden layers. However, with just 15 hidden layers we are able to achieve 99.17% of the highest coefficient of determination, hence the computational cost is reduced (Figure 4-25).

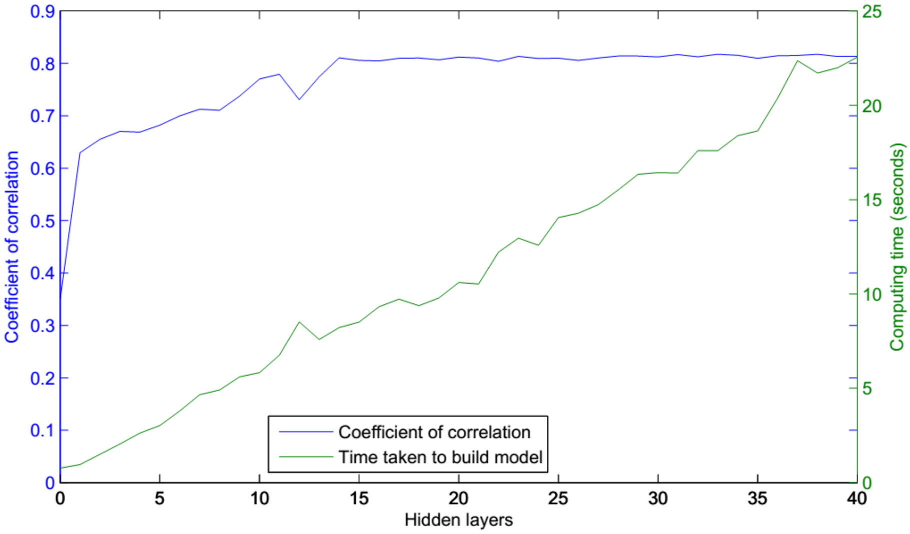


Figure 4-25 - Effect of increasing the number of hidden layers on the efficiency of a multilayer perceptron.

Contribution of predictor variables

Additionally, different setups for input variables using artificial neural networks are explored, in the same line with the previous study using multiple linear regression. The results are summarized in Table 4-14. The difference in the coefficient of determination between the executions with 10 or 13 input variables is almost negligible, however, it is preferable to adopt the set-up with 10 variables, since it assures non-correlation among the predictors. Using only the five dominant variables (observed during the study on multiple linear regression) considerably drops the model performance produced by the ANN. Nevertheless, the setup with five predictor variables is able to achieve 90% of the best correlation.

Number of predictor variables	Coefficient of Correlation (*r*)	Coefficient of Determination (r^2)
13 predictor variables	0.7893	0.6229
10 predictor variables	0.7869	0.6192
5 predictor variables	0.7329	0.5371

Table 4-14 – Comparison of ANN performance using different predictor variables.

Temporal behaviour

NO_2 concentrations (and atmospheric pollutants in general) are directly affected by meteorological conditions. Wind, temperature, weather conditions and air pressure are fundamental factors for atmospheric pollutants dispersion (Kukkonen, et al., 2003), (Agirre-Basurko, et al., 2006), (Juhos, et al., 2008). These meteorological conditions can change from month to month and between different hours of the day (as observed in section 4.5.3), thus affecting the atmospheric pollutants concentrations. In order to capture these changes, the previous artificial neural network, based on a multilayer perceptron with 15 hidden layers, is applied to data sets split by month and daily periods, summarized in Table 4-15:

Month	Coefficient of Correlation
September	0.6867
October	0.7879
November	0.7134
December	0.7721

Daily period	Coefficient of Correlation
Morning	0.6804
Afternoon	0.7666
Night	0.6985

Table 4-15 – Comparison of ANN performance on different temporal periods.

October, closely followed by December, is the month when the ANN performed best. On the other side, data from September shows the lowest performance. The behavior of individual variables is unable to explain this scenario. For instance, the average temperature gradually decreases from September to December, while the weather conditions gradually worsen in the same period, as described in section 4.4.2. However, the performance of the ANN neither gradually improves nor degrades from September to October. This is an indication of the interplay of different variables on the process of pollutants concentration and dispersion.

By analyzing the performance during different daily periods, we can observe that afternoon (4 PM – 12 AM) is the period with the highest performance from the ANN. This is also the period with the highest temperature and wind speed. Although these two meteorological conditions should help disperse atmospheric pollutants, the daily profile of NO_2 (and other pollutants, observed on section 4.5.1) shows a peak of concentrations between 6 PM and 10 PM, especially during cold months. As a possible explanation, we should also consider that the afternoon and nightfall register an increase of traffic (commuting from work to home) and on winter the increase of domestic heating. These emissions will add to existing atmospheric pollutants from the morning period. The performance during morning (8 AM – 4 PM) and night (12 AM – 8 AM) periods are reasonably similar.

Finally, the performance of the ANN on weekdays (*r* = 0.7641) is quite similar to the performance on weekends (*r* = 0.7694). This observation is surprising, since previous chapters have shown that traffic, on weekends, is less regular and has a decrease in activity (weekdays are characterized by repeated activities in temporal orders, such as commuting to work, lunch time at same similar places, school activities, etc.). However, in a way, it is a possible indication that atmospheric pollutants follow the traffic patterns to some extent.

Effect of meteorological conditions

A similar study is performed using wind speed and temperature, as described in Table 4-16:

Wind speed (km/h)	Coefficient of Correlation
1 - 10	0.6473
11 - 20	0.6415
21 - 30	0.7708
31 - 40	0.7692

Temperature (ºC)	Coefficient of Correlation
1 - 10	0.6144
11 - 20	0.6795
21 - 30	0.6290

Table 4-16 – Comparison of ANN performance on different meteorological conditions.

With the increase of wind speed, the performance of the model seems to improve. Nevertheless, during the period under observation, there is no register of extreme wind speeds (maximum wind speed registered of 45 km/h, with an average of 14.4km/h). However, the same effect is not observed with the variation in temperature. Temperatures between 11ºC and 20ºC (the average) enable the highest performance, while lower performance is observed during temperature extremes (very low or very high temperatures). Similar to the wind speed, no extreme temperatures were registered during the period under observation (a minimum of 2ºC, maximum of 33ºC and average of 18.2ºC). On the other hand, clear or cloudy weather conditions enable a better performance of the model, as described in Table 4-17.

Weather conditions	Coefficient of Correlation
Clear	0.7122
Cloudy	0.7187
Rain	0.6612

Table 4-17 – Comparison of ANN performance on different weather conditions.

As a final note, the model seems to perform best on stable meteorological conditions during the afternoon and nightfall: stable weather conditions (clear or cloudy), with average temperatures (11ºC to 20ºC) and reasonable wind speed (21-30 km/h).

It is important to stress again that the behavior of atmospheric pollutants is the result of a complex interaction between different meteorological variables (wind, temperature, humidity, and weather conditions). Individually, none of the variables is able to completely explain the behavior of exhaust gases in the atmosphere. Therefore, unexpected behaviors from exhaust gases can be observed when individual meteorological variables are analyzed.

4.6.5 Reproduce the experiment on the remaining monitoring stations

Up until now, the inference analysis was performed using NO_2 - a representative exhaust gas - with data collected at '*Av. Liberdade*' monitoring station, a predominant monitoring station. At the end of the analysis, the same procedure is applied to the remaining monitoring stations. Although NO_2 is strongly correlated with NO, NO_x and CO, this subsection also studies how suitable are the models to estimate the remaining exhaust gases.

4.6.5.1 Overall performance of NO_2 estimation

Monitoring stations are located in urban areas with specific profiles. We are interested to observe the performance of the models in different scenarios, especially between traffic and background stations. Results for comparison of different techniques are summarized in Table 4-18:

Monitoring Station	Multilayer Perceptron	Linear Regression
Av. Liberdade (T)	0.7869	0.6254
Beato (B)	0.6811	0.6218
Chelas (B)	0.6531	0.6102
Entrecampos (T)	0.7172	0.5880
Olivais (B)	0.6706	0.6095
Restelo (B)	0.6069	0.5688
S. C. Benfica (T)	0.6996	0.6473

Table 4-18 - Values of coefficient of correlation obtained from different techniques to model NO_2, applied on different monitoring stations.

The monitoring station with the highest correlation values is "*Av. Liberdade*", followed by "*Entrecampos*" and "*S. C. Benfica*", both traffic stations (*T*). Considering that the experiment aimed to estimate the value of NO_2, and traffic stations are positioned adjacent to main roads to perceive traffic emissions, this is an expected outcome. The vicinity to highway infrastructures, packed with traffic, could intuitively explain the observed results. Coincidentally, these three stations present the highest average concentrations of NO_2, as observed during data exploration (Section 4.5.5, Table 4-9). We can conclude that the highest correlation between NO_2 and the predictor variables is achieved in traffic stations.

These observations are in line with Pleijel et al. (Pleijel, et al., 2004), Gilbert et al. (Gilbert, et al., 2003), Zou et al. (Zou, et al., 2006) and Ndoke & Jimoh (Ndoke & Jimoh, 2005). The authors observed that concentrations of exhaust gases decrease with distance from roads. Therefore, concentrations of NO_2 sensed in traffic stations can be higher in comparison to background stations, as stated by Colls & Tiwary (Colls & Tiwary, 2009). Moreover, since traffic is one of the main anthropogenic sources of air pollutants - namely NO_2 - through the photochemical oxidation at ground-level of NO produced during internal combustion (Zavala, et al., 2006), (Karlsson, 2004), traffic stations can better perceive the concentrations of NO_2 due to their location (near traffic). Since traffic stations are deployed to mainly perceive emissions from vehicles, our model seems to perform better in monitoring pollutants emitted by traffic. Moreover, the KNN model outperforms linear regression model, which is in line with other authors (Perez & Reyes, 2001), (Kukkonen, et al., 2003), (Agirre-Basurko, et al., 2006), (Grivas & Chaloulakou, 2006).

4.6.5.2 Overall performance of NO estimation

As stated before, NO_2 is not directly emitted to the atmosphere by most vehicles, it is instead a byproduct of a photochemical oxidation at ground-level of NO, with NO being directly emitted during fuel burning (Clapp & Jenkin, 2001). In that sense, the same experiences are performed to estimate the concentrations of NO. Table 4-19 summarizes the results, using the 10 non-correlated predictor variables, following a holdout method (the oldest 2/3 forming the training set and the latest 1/3 used for testing):

Monitoring Station	**Multilayer Perceptron**	**Linear Regression**
Av. Liberdade (T)	0.6350	0.4964
Beato (B)	0.4733	0.3739
Chelas (B)	0.4911	0.4241
Entrecampos (T)	0.6431	0.4809
Olivais (B)	0.5324	0.3632
Restelo (B)	0.4211	0.3739
S. C. Benfica (T)	0.5504	0.4988

Table 4-19 - Values of coefficient of correlation obtained from different techniques to model NO, applied on different monitoring stations.

In the same line with the estimation of NO_2 concentrations, higher values of correlation are observed in traffic stations ("*Av. Liberdade*", "*Entrecampos*", and "*S. C. Benfica*"). Similarly, the artificial neural network shows consistently a better performance than the linear regression approach. However, the best results produced by ANN are achieved using 5 hidden layers (to estimate NO_2 concentrations best results were achieved with 15 hidden layers).

Most significantly, the correlation values between NO and the predictor variables are noticeably lower than the correlation values between NO_2 and the same predictor variables, similar to the observations of Allen et al. (Allen, et al., 2011). Furthermore, this difference is higher in background stations (on average 26%, against 16% in traffic stations). Therefore, in our analysis, NO_2 appears to be a better marker for exhaust-pipe emissions on road-adjacent locations, as observed by other authors (Krzyzanowski & Schneider, 2005), (Becker, et al., 2000).

4.6.5.3 Overall performance of NO_x estimation

NO_x is composed of NO and NO_2, thus it should reflect the performance of the latter exhaust gases. Table 4-20 summarizes a similar experience, performed to estimate the concentrations of NO_x. As expected, ANN outperforms the linear regression. Results are, on average, lower than those obtained with NO_2 but higher than those obtained with NO. The "*Av. Liberdade*" station is once again the monitoring station with the best performance, using both ANN and linear regression. Traffic stations also outperform background stations in both approaches.

Monitoring Station	Multilayer Perceptron	Linear Regression
Av. Liberdade (T)	0.6394	0.5354
Beato (B)	0.5799	0.5108
Chelas (B)	0.6043	0.5222
Entrecampos (T)	0.6209	0.5199
Olivais (B)	0.6009	0.4900
Restelo (B)	0.5285	0.4931
S. C. Benfica (T)	0.6112	0.5420

Table 4-20 - Values of coefficient of correlation obtained from different techniques to model NO_x, applied on different monitoring stations.

4.6.5.4 Overall performance of CO estimation

Finally, Table 4-21 summarizes a similar experience, performed to estimate the concentrations of CO. Although, as expected, ANN also outperforms linear regression, the best performance is not achieved by "*Av. Liberdade*" station, like in the previous scenarios. The best performance using ANN is achieved by "*Chelas*" station, a background station, while the best performance using linear regression is achieved by "*Santa Cruz de Benfica*", a traffic station. Interestingly, the estimation of CO achieves better results, on average, in background stations with both techniques. The performance of ANN outperforms the same technique applied to NO_x, however, the overall performance of linear regression is quite similar to the performance achieved with NO_x. Nevertheless, NO, NO_x, and CO underperformed in both techniques against NO_2.

Monitoring Station	Multilayer Perceptron	Linear Regression
Av. Liberdade (T)	0.6141	0.4866
Beato (B)	0.6445	0.5108
Chelas (B)	0.6792	0.5196
Entrecampos (T)	0.6357	0.4919
Olivais (B)	0.6237	0.4972
Restelo (B)	0.6168	0.5599
S. C. Benfica (T)	0.6214	0.5824

Table 4-21 - Values of coefficient of correlation obtained from different techniques to model CO, applied on different monitoring stations.

4.7 Chapter summary and conclusions

This section summarizes the work developed to estimate the concentration of exhaust gases using taxi data and meteorological conditions. Main contributions and results are presented along with a discussion of the limitations of the study and future improvements to the work.

4.7.1 Overview and contributions

With the rapid increase in size and population of urban areas, it becomes important to understand urban environmental influencers, so that better informed decisions can be made for more sustainable urban environments. Taxis represent one of the urban mobility modes from which city planners can gain a better understanding of mobility in general, as well as its relationship with other environmental elements.

Although air pollutants can be generated naturally, they have been strongly linked to anthropogenic sources. Traffic is one of the main urban sources of exhaust gases (together with industrial burning and domestic heating). Additionally, taxis can be used as a probe for traffic conditions (Castro, et al., 2012), (Yuan, et al., 2011a), (Gühnemann, et al., 2004), (Liu, et al., 2009a), while meteorological conditions (wind, temperature, weather conditions) are the fundamental factor to atmospheric pollutants dispersion (Kolehmainen, et al., 2001), (Kukkonen, et al., 2003), (Agirre-Basurko, et al., 2006), (Juhos, et al., 2008). Moreover, NO_2 appears to be a good marker for exhaust-pipe emissions (Krzyzanowski & Schneider, 2005), (Becker, et al., 2000). Therefore, a combination of these predictor variables can be used to estimate NO_2 concentrations in an urban area.

In this chapter, we studied the exhaust gas concentration patterns in Lisbon, Portugal, and explored techniques to estimate the levels of NO_2 concentrations through the analysis of other related urban variables, such as taxi location and meteorological conditions.

Based on four months of data we revealed the daily and seasonal patterns of exhaust gases, how they are correlated with weather conditions (humidity, temperature, and wind speed), and how NO_2 is strongly correlated with other exhaust gases (NO, NO_x, and CO).

This study has shown a relationship between exhaust gas concentration and other urban variables. Using a multilayer perceptron, with 15 hidden layers and a sigmoid activation function, we were able to estimate the NO_2 concentrations, with a coefficient of correlation of 0.7869 (“*Av. Liberdade*” monitoring station). Linear regression was only able to provide a correlation of 0.6254. Therefore, we concluded that KNN models outperform linear regression models, which is in line with other authors (Perez & Reyes, 2001), (Kukkonen, et al., 2003), (Agirre-Basurko, et al., 2006), (Grivas & Chaloulakou, 2006).

The multicollinearity analysis identified correlations among some predictor variables, which could lead to overfitting. Therefore, the initial data set of 13 predictor variables was reduced to 10. Moreover, we were able to obtain 90% of the highest correlation using just five variables (hour, temperature, humidity, wind speed and taxi pick-up). This result is in line with studies of literature, stating that wind and temperature are among the most important factors in the dilution and spreading of atmospheric pollution.

Monitoring stations located near main roads (traffic stations) are able to sense exhaust gas concentrations with a higher coefficient correlation. The performance of the model considerably drops on stations located distantly from roads (background stations).

The performance of the model was tested with different meteorological scenarios. The model seems to perform best on stable meteorological conditions during the afternoon and nightfall: stable weather conditions (clear or cloudy), with average temperatures (11ºC to 20ºC) and reasonable wind speed (21-30 km/h). Moreover, the model is more suitable to estimate NO_2 concentrations than NO. These results were published in (Veloso, et al., 2013) and submitted to (Veloso, et al., 2015).

The proposed approach is based on existing infra-structures and does not demand a deployment of new hardware. It takes advantage of opportunistic sensing, using data collected by in-board taxi-GPS devices, and meteorological stations already installed and covering most of urban areas.

This approach can be used to improve spatial resolution of exhaust gas monitoring. Since few monitoring stations are deployed in urban areas, the model could be used to overcome the absence of local measurements. Moreover, the proposed approach relies on existing equipment, and does not demand for a dedicated infrastructure. We hope that this work shed some new light on the complex interrelationships of urban system variables.

4.7.2 Limitations and future work

The current work also shows weaknesses and limitations. The low performance from applying linear regression is an indication of the complex interplay between the predictor variables. Although the artificial neural network presents an improvement, the achieved coefficient of determination is far from being an optimal output. As a result, our work cannot conclude that the use of this set of predictor variables is able to accurately estimate NO_2 concentrations and other atmospheric pollutants in all conditions. Therefore, further studies are needed to explore alternative sets of variables available and models produced using different techniques.

The ANN model presents a fair performance (it is able to estimate the NO_2 concentrations with a coefficient of correlation of 0.7869). Although the model cannot be used to accurately predict concentration values of NO_2, it can be used to estimate areas with dangerous levels of atmospheric pollutants, where no current monitoring is available, especially on ground-level, near roads. This could be the first level of a warning system to assess air quality. If the model estimates values over a safety threshold, a warning should be automatically triggered and further actions should be taken, namely the confirmation of the prediction with local measurements.

Additionally, using historical data, the system could be able to anticipate areas where exhaust gases' concentrations will overreach a safety threshold and become dangerous for human health. However, although taxi activity follows a daily and weekly pattern, meteorological variables are more instable and can change quickly. To note that although hour is a predominant predictor variable in the model, the same is not true for day of the week.

The complex dynamics of atmospheric pollutants' dispersion prevents the existence of models based on a single predictor variable. In consequence, taxi data alone is unable to explain the behavior of NO_2 concentrations, and its contribution to the model is, in fact, inferior to that from other variables (e.g. wind or temperature). However, major concerns are related to the representativeness of taxi data. As mentioned in previous chapters, although at the time of collection the data provider accounted for nearly 20% of taxi share in the city, the representativeness of the data can also be disputed. Therefore, the argument that taxis can be used as a probe for traffic conditions (Castro, et al., 2012) can also be disputed. Moreover, the quality of taxi data could also have a negative impact on the performance of the model. Newer data sets should be gathered to confirm the validity of the findings.

Besides temperature, wind speed, and weather conditions, other variables play an important role in the formation and dispersion of atmospheric pollutants. Among them, the topography, air pressure, and wind direction are relevant factors. However, our work does not explore their contribution to the model. This absence could explain why the proposed model is unable to completely describe the behavior of NO_2. Spatial patterns of atmospheric pollution should be discussed and explored in future work, providing inputs for the model to capture spatial effects.

Besides NO_2, NO, NO_x, and CO_2, other toxic gases pose a danger to human health, namely SO_2 (Sulphur Dioxide) and C_6H_6 (Benzene). These gases were not analyzed in this study due to the lack of data in most of the monitoring stations. Therefore, future work should consider collecting newer data set with a broader set of toxic gases.

Although the approach is based on presently deployed infra-structures, for optimal performance it needs data in real-time, at the very least provided in one-hour intervals. It is not clear if current systems are able to transmit and provide data within that timeframe. All the analysis performed in this study is based on an offline database. Moreover, although present meteorological stations are able to survey all (or most) urban areas, some variables - essential to the model - may present changes between small spatial areas (e.g. between neighbor block of houses) that are not captured (e.g. humidity, wind speed).

Finally, the model was not tested in other urban areas. Therefore the geographic replicability can be disputed. This limitation is due to the lack of data available for the same temporal window for other regions. As stated before, newer data set should be collected for the same urban region alongside with data sets from different locations, with a wider temporal window, in order to validate the temporal and geographic replicability of the model.

We hope our findings are a relevant contribution to the complex urban dynamics, especially to improve spatial atmospheric pollution monitoring. We aim to continue our investigation as aforementioned. Our future work will explore more robust approaches, improve the size and quality of the data set, and validate the results on different urban areas.

Chapter 5
Conclusions

5.1 Overview

As stated by Castro et al. (Castro, et al., 2013), GPS-equipped taxis can be viewed as pervasive sensors, and the large quantity of GPS traces produced, allow us to reveal facts about social urban life. In this work, we explored the potential for historic taxi-GPS traces to represent the city's dynamics and its relation with other urban data sources. In order to do that, we studied different data sets containing information about taxi activity, mobile phone activity, exhaust gases' concentrations, points of interest, and weather conditions, collected in Lisbon, Portugal, during 2009.

The study was able to produce the following three main contributions: (a) developed a model to estimate taxi demand and explored the spatiotemporal distribution of taxi activity; (b) analysed the relationship between taxi and mobile phone activity and studied the spatiotemporal patterns of mobile phone call intensity; and (c) explored models that use taxi activity information and meteorological conditions in order to estimate atmospheric pollutants, and analyse the exhaust gases patterns. These three main contributions will be detailed, alongside their limitations, in the following sections.

5.2 Contributions

The main three contributions of this work can be further detailed, as follows. For better comprehension the contributions are grouped by section.

A. Development of a model to estimate taxi demand and explore the spatiotemporal distribution of taxi activity

Using taxi-GPS traces collected during a period of four months (September – December, 2009) the study was able to visualize the spatiotemporal variation of taxi activity, identified the main pick-up and drop-off locations and busy hours, and observed that trip distance and duration follow Gamma and Exponential distributions. The study was also able to identify the link between pick-up and drop-off locations, observed strong links between public transportation hubs, where taxi service appears to operate as a bridge between different public transportation services. Additionally, an analysis of taxi driver behavior during downtime was performed – time spent searching for next pick-ups - when taxis tend to avoid making long trips to suburban areas for pick-ups.

The analysis of taxi-GPS traces from top drivers showed specific strategies used to maximize the profit. Either by waiting for passengers in locations related with main public transportation hubs (airport, train stations, ferry dock or main bus stops), during specific hours of the day, or by avoiding traveling great distances to the next pick-up location, unless it was to return to the aforementioned locations. Low performance drivers showed no such specific strategy and were the major contributors to the apparently randomness of taxi flow.

Our inference analysis explored the possibility to estimate the next pick-up area, given the current location (last drop-off), day of the week, hour, weather conditions, and area type (predominant POI). The inference engine is based on a naïve Bayesian classifier, achieving 56.3% of accuracy of the training samples, on specific conditions: weekdays, working hours, and in areas with reasonable taxi activity (cells with low taxi activity were disregarded). Current location turned out to be the main contributor to the algorithm, contrary to weather conditions which was the variable with less weight in the calculation. Several effects to the classifier were explored, namely, the influence of daily and weekly periods and the impact of a cell's size.

The results achieved in this study show that to some extent, taxi volume follows daily and weekly patterns that can be modeled to infer the likelihood of the next pick-up location, especially when the behavior of top drivers is taken into consideration.

B. Analysis of the relationship between taxi and mobile phone activity and studied the spatiotemporal patterns of mobile phone call intensity

Based on one-month of data (December 2009), the study performed an exploratory analysis of the mobile phone call intensity, which showed a fairly regular pattern, consistent throughout the day and during the entire time series. Comparisons between different urban areas showed only minor pattern deviations of mobile phone call intensity. Additionally, several indicators were identified that perfectly represent mobile phone activity, namely busy hour traffic (in *erlangs*) and number of calls successfully started.

During data analysis, a significant correlation between the taxi volume and mobile phone call intensity was found, over active hours of the day (8 AM-10 PM) and active days of the week (weekdays), in areas with medium and high taxi activity. Moreover, mobile phone call intensity had a significant correlation with taxi volume of the previous two hours, which means that the amount of taxis could be used to predict the intensity of mobile phone calls along the next two hours.

Furthermore, we have found that this inter-predictability could be modeled with a linear function and varied across different times of the day. Intensity of mobile phone calls was a predictor of taxi volume in morning hours, while the amount of taxi flow became a predictor of mobile phone calls in the afternoon and evening.

The results achieved in this study support the hypothesis of an inter-predictability between taxi volume and mobile phone call intensity.

C. Explored models that use taxi activity information and meteorological conditions in order to estimate atmospheric pollutants, and analysed the exhaust gases patterns

Based on four months of data, we studied the exhaust gas concentration patterns in Lisbon, Portugal, and explored techniques to estimate the levels of NO_2 concentrations through the analysis of other related urban variables, such as taxi activity and meteorological conditions.

The study revealed the daily and seasonal patterns of exhaust gases, how they were correlated with meteorological conditions (humidity, temperature, wind speed, and weather conditions), and how NO_2 strongly correlates with other exhaust gases (NO, NO_x, and CO).

The study has shown a relationship between the exhaust gas concentration and other urban variables. Using a multilayer perceptron, with 15 hidden layers and a sigmoid activation function, we were able to estimate the NO_2 concentrations, with a coefficient of correlation of 0.7869. Linear regression was able to provide a maximum correlation of 0.6254. Therefore, KNN model outperformed linear regression model, which is in line with the work of other authors (Perez & Reyes, 2001), (Kukkonen, et al., 2003), (Agirre-Basurko, et al., 2006), (Grivas & Chaloulakou, 2006).

The multicollinearity analysis identified correlations among some predictor variables, which could lead to overfitting. Therefore, the initial data set of 13 predictor variables was reduced to 10. Moreover, we were able to obtain 90% of the highest correlation using just five variables (hour, temperature, humidity, wind speed, and number of taxi pick-ups). This result is in line with other studies in the literature, stating that wind and temperature are among the most important factors in the dilution and spreading of atmospheric pollution.

Monitoring stations located near main roads (traffic stations) were able to sense exhaust gas concentration with a higher coefficient correlation. The performance of the model considerably dropped in background stations (located distant from roads).

The performance of the model was tested with different meteorological scenarios. The model seemed to perform best on stable meteorological conditions, during the afternoon and nightfall: stable weather conditions (clear or cloudy), with average temperatures (11ºC to 20ºC) and reasonable wind speed (21-30 km/h). Moreover, the model was more suitable to estimate NO_2 concentrations than NO.

This approach can be used to improve spatial resolution of exhaust gas monitoring. Since few monitoring stations are deployed in urban areas, the model could be used to overcome the absence of local measurements. Moreover, the proposed approach relies on existing equipment, and does not demand a dedicated infrastructure.

5.3 Limitations

Several limitations can be pointed out in the current study. For better legibility, the contributions are grouped by sections.

A. Development of a model to estimate taxi demand and explore the spatiotemporal distribution of taxi activity

The highest accuracy of the inference engine (56.3%) is achieved under specific conditions: on weekdays; working hours; and in areas with reasonable taxi activity (cells with low taxi activity were disregarded), and considering neighbor cells as positive predictions. The approach was needed due to insufficient data in some temporal and spatial slots. By removing these special conditions, the classifier achieved a lower accuracy of 31%, as a result of a larger search space, composed by 370 possible locations, which includes areas without sufficient data (low taxi activity cells). This lower performance is an expected result considering the search space is widened, and includes areas without sufficient data. Nevertheless, it was a fair outcome once we consider that *a priori* probability for the best cells was 2.4%. Since the lack of data is a major limitation in applying the algorithm to all Lisbon municipality, a more comprehensive data set should collected.

The adoption of a naïve Bayesian classifier to estimate the likelihood of the next pick-up location requires that the predictors are independent variables. However, POI may not be independent from day of the week and hour of the day. This possible dependency between the variables could affect the performance of the naïve Bayesian classifier. Therefore, other approaches should also be explored along with an alternative set of predictor variables.

The study did not take into consideration urban events (e.g. sports, music concerts, cultural expositions, or even workers strikes from public transportation services), which could affect the patterns of taxi service. The occurrence of events and the impact on taxi service should be explored in the future. Since a possible bridge between public transportation hubs was identified, the timetables of these transportation services should also be considered.

Additionally, future investigation should also deepen the exploratory study, to analyze the effect of different factors on taxi service (e.g. weather conditions or the hour of the day). In the same line, the reasons that drive a passenger to choose to use a taxi are not perfectly clear, and that knowledge could improve the performance of the inference engine. Therefore, it is advisable to perform a survey among taxi passengers.

The process of grid creation can also be subject to discussion, since it doesn't take into consideration the location of particular areas of the city that affect the traffic conditions, or the original density of the taxi-GPS traces. Thus, some hotspots could have been split between several grid cells, thus disturbing the outcome of the analysis. The clusters of GPS traces could be used to guide the size and placement of the grid as suggested by Castro et al. (Castro, et al., 2013).

A full recommendation system was not built, focusing only at the inference engine. However, the development of a fully functional commercial framework would allow it to be tested on real conditions and to assess the true usefulness of the approach. Additionally, the recommendation system does not take into consideration the current behavior of other taxi drivers. Therefore, they could be competing for the same resource (passenger). To improve the efficiency of the recommendation system, future work should consider a concurrent approach, where estimations are performed taking in consideration the current status of other taxis, thus avoiding competition for the same passenger.

B. Analysis of the relationship between taxi and mobile phone activity and studied the spatiotemporal patterns of mobile phone call intensity

The reduced amount of data used (only one month) and the absence of data (grid-cells with low taxi activity) could have limited the analysis. To validate the observations a newer data set should be collected, that encompass a broader temporal window. Additionally, an alternative data set from a different urban area would allow the validation of the geographical replicability of the approach.

Another potential limitation was the linear relation that was assumed between the two data sources in this study. Further investigation should be done in finding a possibly more suitable function to model the relationship between taxi volume and mobile phone call intensity.

Finally, the model achieved higher performance when applied to specific conditions: working hours, weekdays, and cells with medium and high taxi activity. The study showed a considerable degradation of the model on weekends, at night, and on national holidays. More evident was the low correlation between both time series on cells with low taxi activity, where the approach was not suitable, since there was insufficient data. This can be an indication of the possible low representativeness of taxi volume and the absence of data in certain areas (e.g. grid-cells with low taxi activity). As stated before, a new and broader data set should be attained in the future to confirm the results obtained in this study.

C. Explored models that use taxi activity information and meteorological conditions in order to estimate atmospheric pollutants, and analysed the exhaust gases patterns

The ANN model does not present a significant performance (estimates of NO_2 concentrations, with a coefficient of correlation of 0.7869). However, although the model cannot be used to accurately predict concentration values of NO_2, it can be used to estimate areas with dangerous levels of atmospheric pollutants, where no current monitoring is available. This could be the first level of a warning system for air quality. If the model estimates values higher than a safety threshold, a warning should be triggered and further actions should be taken, namely the confirmation of the prediction with local measurements.

The low performance from the linear regression was an indication of the complex interplay between the predictor variables. Although the artificial neural network presented an improvement, the achieved coefficient of determination is not an optimal output. As a result, our work could not conclude that the use of this set of predictor variables is able to accurately estimate NO_2 concentrations and other atmospheric pollutants in all conditions. In order to improve the performance of the model further studies are needed, to explore alternative sets of variables available and techniques.

The complex dynamics of atmospheric pollutants' dispersion prevent the existence of models based on a single predictor variable. Therefore, taxi data alone was unable to explain the behavior of NO_2 concentrations. Besides temperature, wind speed and weather conditions - which were explored in this study - other variables play an important role in the formation and dispersion of atmospheric pollutants. Among then, topography, air pressure, and wind direction are also relevant factors. However, our work did not explore their contribution to the model. Spatial patterns of atmospheric pollution should be discussed and explored in future investigation, providing inputs for the model to capture spatial effects.

Global limitations

A set of limitations were common throughout the work, narrowing the analysis performed:

- Geographic replicability;
- Quality of the data set;
- Conditions for optimal performance of the models;
- Significance of the results.

The analysis focused on a single city (Lisbon). This is a consequence of a lack of data sets from other cities. The outcome could be strengthened if data from different urban areas were available to apply the same procedures, analyses, and to compare results. Hence, future work should encompass the acquisition of data from different urban areas.

Equally important are the concerns around the quality of data. Although the exploratory analysis showed that patterns are somewhat consistent between months and years (when data was available), the data set was collected in 2009, representing a considerable temporal gap to the results now being presented. Since then, the city under study has experienced several changes. Moreover, the limitation of the size of the time series (e.g. inter-predictability study between taxi volume and mobile phone call intensity) or the lack of data in certain areas of the target city (e.g. grid-cells with low taxi activity) were visible throughout the analysis. Additionally, although at the time of the data collection the data provider accounted for nearly 20% of taxi share in the city, the representativeness of the data can now be disputed. Therefore, as stated before, a newer data set should be collected to validate and improve our findings.

The best results in each model were achieved under a specific set of conditions. For instance, to estimate the best location for a pick-up, the model performs better on weekdays, working hours, and in areas with reasonable taxi activity; the relationship between mobile phone activity and taxi volume is higher on active hours of the day (8 AM-10 PM) and active days of the week (weekdays), in areas with medium and high taxi activity; and the estimation of NO_2 concentrations are better perceived in traffic stations, located near main roads. This is due the aforementioned insufficient data in some temporal and spatial slots. As result, the study was unable to identify a universal model to perform under all conditions. Further investigation is planned to improve these results, putting efforts in the acquisition of newer data sets.

Considering the aforementioned special conditions, the models can be described as fair in estimating or predicting the value of the dependent variables. Although taxis are related to and have influence in urban mobility (e.g. they can be used to estimate how inhabitants move, mobile phone usage, or concentrations of NO_2), it is not the dominant predictor variable. This is especially true when estimating air pollutants concentrations. Despite taxis being a relevant and useful variable to estimate NO_2 concentrations, the formation and dispersion of air pollutants are mainly affected and explained by meteorological conditions (temperature, wind speed, humidity and weather conditions). Nevertheless, taxis are an important player in urban mobility.

5.4 Future Work

From the aforementioned limitations, it is clear that a new and broader data set should be gathered in the future, from the same municipality as well as other urban areas, to assess if the procedure still holds true for different temporal periods and spatial locations. Although each city has its own topography and culture, similar patterns can be observed; therefore the study should be expanded to other urban areas.

The recommendation system should be fully developed. The experience learned from this study would be valuable to produce a complete and useful tool to assist taxi drivers with strategies to identify their next passenger, reducing the time, distance, and fuel necessary to accomplish that goal. Additionally, a survey on the motivations of the passenger to use (or avoid using) taxi service could shed light on certain patterns and help improve the inference engine.

Other variables should also be explored in the future to improve the algorithms. For instance, social events affect urban mobility, thus influencing taxi patterns. In a similar way, the topography affects the dispersion of gases, thus influencing the concentration of exhaust gases in urban areas.

Final remarks

We hope that our work could contribute to a better comprehension of the complex interactions between the diversity of urban processes. Our findings, to some extent, unveil the relationships between different urban data sources, which describe the city from different perspectives, and can produce an aggregated and collective view of urban areas.

We expect that our findings could suggest new ways to use multi-source data fusion to investigate the interplay between different urban entities. Our observations and results need further improvements, but nevertheless, could contribute to the intense research on urban dynamics that is currently ongoing, in order to assist the developing of more efficient intelligent transportation systems.

Urban studies are growing in importance with the growth and expansion of urban areas. The need for a more efficient use of resources allied with the requirement for sustainable communities demand new approaches and models, and the adaptation of current infra-structures for further activities and usage. Our work is a small contribution to this overall and embracing goal.

References

Agirre-Basurko, E., Ibarra-Berastegi, G. & Madariaga, I., 2006. Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area. *Journal Environmental Modelling & Software,* 21(4), pp. 430-446.

Agirre, E., Anta, A. & Barron, L., 2009. *Development of a Computational Software to Forecast Ozone Levels.* s.l.:InTech.

Agirre, E., Anta, A. & Barron, L., 2010. Forecasting ozone levels using artificial neural networks. In: *Forecasting Models: Methods & Applications.* s.l.:CreateSpace Press, pp. 207-2018.

Aguilera, V. et al., 2012. Estimating the Quality Service of Underground Transit Systems with Cellular Network Data. *Social and Behavioral Sciences,* Volume 48, p. 2262–2271.

Aguilera, V., Milion, C. & Allio, S., 2014. *Territory analysis using cell-phone data,* Paris, France: Transport Research Arena.

Aharony, N. et al., 2011. Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing,* Volume 7, p. 643–659.

Ahmed, N., Atiya, A., El Gayar, N. & El-Shishiny, H., 2010. An Empirical Comparison of Machine Learning Models for Time Series Forecasting. *Econometric Reviews,* 29(5-6), pp. 594-621 .

Ahrens, C. & Henson, R., 2014. *Meteorology Today - An Introduction to Weather, Climate, and the Environment.* 11th ed. s.l.:Cengage Learning.

Akaike, H., 1974. A new look at the statistical model identification. *Transactions on Automatic Control,* 19(6), p. 716–723.

Alger, M., Wilson, E. & Gould, T., 2006. *Real-time traffic monitoring using mobile phone data,* s.l.: Vodafone Pilotentwicklung GmbH.

Allen, R., Amram, O., Wheeler, A. & Brauer, M., 2011. The transferability of NO and NO2 land use regression models between cities and pollutants. *Atmospheric Environment,* Volume 45, pp. 369-378.

Alshamsi, A., Abdallah, S. & Rahwan, I., 2009. *Multiagent Self-organization for a Taxi Dispatch System.* Budapest, Hungary, s.n.

Altshuler, Y. et al., 2012. *Incremental learning with accuracy predictions of social and individual properties from mobile-phone data.* s.l., IEEE.

Amat, C., Ortigosa, J. & Estrada, M., 2014. *Assessment of the taxi sector efficiency and profitability continuous monitoring and methodology to review fares.* Washington, D.C., USA, s.n.

APA, 2012. *Evolução da qualidade do ar em Portugal,* Lisbon: Agência Portuguesa do Ambiente.

Aslam, J., Lim, S. & Rus, D., 2012. *Congestion-aware Traffic Routing System Using Sensor Data.* Anchorage, Alaska, USA, IEEE, pp. 1006-1014.

Austin, D. & Zegras, P., 2011. *The taxicab as public transportation in Boston,* Boston, MA, USA: Massachusetts Institute of Technology.

Balan, R., Khoa, N. & Jiang, L., 2011. *Real-Time Trip Information Service for a Large Taxi Fleet.* Bethesda, Maryland, USA, ACM.

Baldauf, R. et al., 2013. Air quality variability near a highway in a complex urban environment. *Atmospheric Environment,* Volume 64, pp. 169-178.

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,* Volume 15, p. 380–391.

Baron, N. & Segerstad, Y., 2010. Cross-cultural patterns in mobile phone use: Public space and reachability in Sweden, the US, and Japan. *New Media & Society,* 12(1), pp. 13-34.

Barros, N., Fontes, T., Silva, M. & Manso, M., 2013. How wide should be the adjacent area to an urban motorway to prevent potential health impacts from traffic emissions?. *Transportation Research - Part A,* Volume 50, p. 113–128.

Bastani, F., Huang, Y., Xie, X. & Powell, J., 2011. *A greener transportation mode: Flexible route discovery from GPS trajectory data.* s.l., s.n., p. 405–408.

Bazzani, A. et al., 2010. Statistical Laws in Urban Mobility from microscopic GPS data in the area of Florence. *Journal of Statistical Mechanics: Theory and Experiment,* Volume 2010.

Bazzi, A. & Masini, B., 2011. *T aking Advantage of V2V Communications for T raffic Management.* Baden-Baden, Germany, IEEE, pp. 504-509.

Becker, K. et al., 2000. Contribution of vehicle exhaust to the global N2O budget. *Chemosphere ± Global Change Science,* Volume 2, pp. 387-395.

Becker, K. et al., 1999. Nitrous oxide emission from vehicles. *Environmental Science and Technology,* Volume 33, pp. 4134-4139.

Beckerman, B. et al., 2008. Correlation of nitrogen dioxide with other traffic pollutants near a major expressway. *Atmospheric Environment,* Volume 42, p. 275–290.

Becker, R. et al., 2011a. A tale of one city: using cellular network data for urban planning. *Pervasive Computing,* pp. 18-27.

Becker, R. et al., 2011b. *Route Classification using Cellular Handoff Patterns.* Beijing, China, ACM.

Beelen, R. et al., 2010. Comparison of the performances of land use regression modelling and dispersion modelling in estimating small-scale variations in long-term air pollution concentrations in a Dutch urban area. *Atmospheric Environment,* Volume 44, pp. 4614-4621.

Bekhor, S., Cohen, Y. & Solomon, C., 2011. Evaluating long-distance travel patterns in Israel by tracking cellular phone positions. *Journal of Advance Transportation,* 47(4), p. 435–446.

Bento, C., Soares, T., Veloso, M. & Baptista, B., 2007. *A Study on the Suitability of GSM Signatures for Indoor Location.* Darmstadt, Germany, Springer, pp. 108-123.

Bento, C., Veloso, M. & Peixoto, J., 2005. *A Case-Based Approach for Indoor Location.* Chicago, IL, USA, Springer, pp. 78-90.

Berkowicz, R., Winther, M. & Ketzel, M., 2006. Traffic pollution modelling and emission data. *Environmental Modelling & Software,* 21(4), p. 454–460.

Beychok, M. R., 2005. *Fundamentals of Stack Gas Dispersion.* s.l.:Milton R. Beychok Publisher.

Bi, J., Embrechts, M., Breneman, C. & Song, M., 2003. Dimensionality Reduction via Sparse Support Vector Machines. *Journal of Machine Learning Research,* Volume 3, pp. 1229-1243.

Birle, C., 2007. *Traffic Online-Via GSM network data to traffic information in real time in real time,* Munich, Germany: Intelligent Transport Systems.

Blondel, V., Krings, G. & Thomas, I., 2010. Regions andborders of mobile telephony in Belgium and in the Brussels metropolitan zone. *Brussels Studies,* 42(4), p. 13.

Bolla, R. & Davoli, F., 2000. *Road traffic estimation from location tracking data in the mobile cellular network.* Chicago, IL, USA, IEEE, pp. 1107-1112.

Borrego, C. et al., 2001. *The impact of road traffic on urban air quality: a modelling and an experimental approach.* Syros, Greece, s.n., pp. 90-97.

Borrego, C., Tchepel, O., Barros, N. & Miranda, A. I., 2000. Impact of road traffic emissions on air quality of the Lisbon region. *Atmospheric Environment ,* Volume 34, p. 4683–4690.

Borrego, C. et al., 2006. Traffic-related particulate air pollution exposure in urban areas. *Atmospheric Environment,* 40(37), p. 7205–7214.

Borrego, C. et al., 2003. Emission and dispersion modelling of Lisbon air quality at local scale. *Atmospheric Environment ,* Volume 37, pp. 5197-5205.

Box, G., Jenkins, G. & Reinsel, G., 2008. *Time Series Analysis - Forecasting and Control.* 4th ed. New Jersey: John Wiley & Sons.

Brummelen, G., 2012. *Heavenly Mathematics: The Forgotten Art of Spherical Trigonometry.* 1st ed. s.l.:Princeton University Press.

Bukowiecki, N. et al., 2002. A mobile pollutant measurement laboratory—measuring gas phase and aerosol ambient concentrations with high spatial and temporal resolution. *Atmospheric Environment,* Volume 36, p. 5569–5579.

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

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

Calabrese, F. et al., 2007. Real-Time Urban Monitoring Using Cellular Phones: a Case-Study in Rome. *Transactions on Intelligent Transportation Systems.*

Calabrese, F. et al., 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C,* Volume 26, pp. 301-313.

Calabrese, F., Lorenzo, G., Liu, L. & Ratti, C., 2011a. Estimating Origin-Destination Flows using Mobile phone Location Data. *Pervasive Computing,* pp. 36-45.

Calabrese, F., Lorenzo, G. & Ratti, C., 2010c. *Human Mobility Prediction based on Individual and Collective Geographical Preferences.* s.l., IEEE, pp. 312-317.

Calabrese, F. et al., 2010b. *The geography of taste: analyzing cell-phone mobility and social events.* s.l., Springer.

Calabrese, F., Reades, J. & Ratti, C., 2010a. Eigenplaces: Segmenting Space through Digital Signatures. *Pervasive Computing,* pp. 78-87.

Calabrese, F., Smoreda, Z., Blondel, V. & Ratti, C., 2011b. Interplay between Telecommunications and Face-to-Face Interactions: A Study Using Mobile Phone Data. *PLoS ONE,* Volume 6, pp. 1-6.

Candia, J. et al., 2008. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical,* Volume 41.

Can, Z. & Demirbas, M., 2015. Smartphone-based data collection from wireless sensor networks in an urban environment. *Journal of Network and Computer Applications,* Volume 58, p. 208–216.

Carslaw, D. & Rhys-Tyler, G., 2013. New insights from comprehensive on-road measurements of NOx, NO2 and NH3 from vehicle emission remote sensing in London, UK. *Atmospheric Environment,* Volume 81, pp. 339-347.

Carvalho, R., 2009. *Avaliação do efeito da Grande Área Metropolitana Porto nas concentrações de CO2 sobe a zona costeira adjacente.* Porto: Universidade Fernando Pessoa.

Castro, P. et al., 2013. From Taxi GPS Traces to Social and Community Dynamics: A Survey. *ACM Computing Surveys,* 46(2).

Castro, P., Zhang, D. & Li, S., 2012. *Urban traffic modelling and prediction using large scale taxi GPS traces.* s.l., s.n.

Cayford, R. & Johnson, T., 2003. *Operational Parameters Affecting the Use of Anonymous Cell Phone Tracking for Generating Traffic Information.* Washington, D.C., s.n., p. 20.

Cayford, R. & Yim, Y., 2006. *Field Operational Test using Anonymous Cell Phone Tracking for Generating Traffic Information.* Washington D.C., USA, s.n.

CellInt TrafficSense, 2007. *Performance and Limitations of Cellular Based Traffic Monitoring Systems,* s.l.: European Congress and Exhibition on Intelligent Transport System and Services.

Chandrasekaran, G. et al., 2010. *Vehicular Speed Estimation using Received Signal Strength from Mobile Phones.* Copenhagen, Denmark, ACM.

Chang, H., Tai, Y., Chen, H. & Hsu, J., 2008. *iTaxi: Context-aware taxi demand hotspots prediction using ontology and data mining approaches.* s.l., s.n.

Chang, H., Tai, Y. & Hsu, J., 2010. *Context-aware taxi demand hotspots prediction.* s.l., s.n., pp. 3-18.

Chattopadhyay, S. & Bandyopadhyay, G., 2007. Artificial neural network with backpropagation learning to predict mean monthly total ozone in Arosa, Switzerland. *International Journal of Remote Sensing,* 28(20), pp. 4471-4482.

Chen, C. et al., 2011. *Real-time Detection of Anomalous Taxi Trajectories from GPS Traces.* s.l., s.n.

Chen, C. et al., 2013b. *TaxiExp: A Novel Framework for City-wide Package Express Shipping via Taxi CrowdSourcing.* s.l., s.n.

Chen, C. et al., 2013a. *B-Planner: Night Bus Route Planning using Large-scale Taxi GPS Traces.* San Diego, USA, IEEE, pp. 224-233.

Chen, G., Chen, B. & Yu, Y., 2010b. *Mining frequent trajectory patterns from GPS tracks.* s.l., s.n., pp. 1-6.

Cheng, D. et al., 2013. The Optimal Sampling Period of a Fingerprint Positioning Algorithm for Vehicle Speed Estimation. *Mathematical Problems in Engineering,* Volume 2013, p. 12.

Chen, G., Jin, X. & Yang, J., 2010a. *Study on spatial and temporal mobility pattern of urban taxi services.* s.l., s.n., p. 422–425.

Cheng, P., Qiu, Z. & Ran, B., 2006. *Particle Filter Based Traffic State Estimation Using Cell Phone Network Data.* Toronto, Canada, IEEE, pp. 1047-1052.

Cheng, S. & Qu, X., 2009. *A Service Choice Model for Optimizing Taxi Service Delivery.* St. Louis, USA, IEEE.

Chen, P., Liu, J. & Chen, W., 2010. *A Fuel-Saving and Pollution-Reducing Dynamic Taxi-Sharing Protocol in VANETs.* Ottawa, ON, IEEE, pp. 1 - 5.

Chiang, C. et al., 2011. *Estimating Instant T raffic Information by Identifying Handover Patterns of UMTS Signals.* Washington, DC, USA, IEEE, pp. 390-396.

Churkina, G., 2008. Modeling the carbon cycle of urban systems. *Ecological Modelling,* Volume 216, pp. 107-113.

Clapp, L. & Jenkin, M., 2001. Analysis of the relationship between ambient levels of O3, NO2 and NO as a function of NOx in the UK. *Atmospheric Environment,* 35(36), p. 6391–6405.

Clemente, C., 2013. *O Amor à Marca e Seus Determinantes - Um Estudo Comparativo Entre Marcas.* Coimbra, Portugal: University of Coimbra.

Cline, W., 1991. Scientific Basis for the Greenhouse Effect. *Economic Journal,* Volume 101, p. 904–919.

Cobourn, W., Dolcine, L., French, M. & Hubbard, M., 2000. A comparison of nonlinear regression and neural network models for ground-level ozone forecasting. *Air Waste Manage Assoc,* Volume 50, p. 1999–2009.

Cogliani, E., 2001. Air pollution forecast in cities by an air pollution index highly correlated with meteorological variables. *Atmospheric Environment,* Volume 35, pp. 2871-2877.

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20(1), p. 37–46.

Colls, J. & Tiwary, A., 2009. *Air Pollution: Measurement, Modelling and Mitigation.* 3rd ed. s.l.:CRC Press.

Cooper, J., Mundy, R. & Nelson, J., 2010. *Taxi! Urban Economies and the Social and Transport Impacts of the Taxicab.* Farnham, UK: Ashgate Publishing.

Coutts, A., Beringer, J. & Tapper, N., 2007. Characteristics influencing the variability of urban CO2 fluxes in Melbourne, Australia. *Atmospheric Environment,* Volume 41, p. 51–62.

Cover, T. & Thomas, J., 2006. *Elements of information theory.* 2nd ed. New York, NY, USA: John Wiley & Sons.

Crawford, B. & Christen, A., 2012. *Temporal and spatial partitioning of measured urban carbon dioxide fluxes.* Dublin, Ireland, s.n.

Csikos, A. & Varga, I., 2011. Real-Time Estimation of Emissions Emerging from Motorways Based on Macroscopic Traffic Data. *Acta Polytechnica Hungarica,* 8(6), pp. 95-110.

Daniels, R. & Mulley, C., 2011. *Explaining walking distance to public transport: the dominance of public transport supply.* Whistler Canada, s.n.

Demissie, M., Correia, G. & Bento, C., 2012a. *Exploring cellular network handover information for urban mobility analysis using GIS and statistical analysis.* s.l., 92nd Annual Meeting of the Transportation Research Board.

Demissie, M., Correia, G. & Bento, C., 2012b. *Intelligent road traffic status detection system through cellular networks handover information: An exploratory study.* s.l., s.n.

Demuth, H., Beale, M. & Hagan, M., 2008. *Neural Network Toolbox - User's Guide.* Natick, MA, USA: The MathWorks.

Derwent, R. et al., 1995. Analysis and interpretation of air quality data from an urban roadside location in Central London. *Atmospheric Environment,* 29(8), p. 923–946.

Devore, J. & Berk, K., 2012. *Modern Mathematical Statistics with Applications.* 2nd ed. New York: Springer-Verlag.

Ding, B., Yu, J. & Qin, L., 2008. *Finding time-dependent shortest paths over large graphs.* Nantes, France, ACM, pp. 205-216.

Ding, L., Fan, H. & Meng, L., 2015. *Understanding taxi driving behaviors from movement data.* s.l., s.n.

Domenico, M., Lima, A. & Musolesi, M., 2013. Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing,* Volume 9, p. 798–807.

Donnelly, A., Misstear, B. & Broderick, B., 2011. Application of nonparametric regression methods to study the relationship between NO2 concentrations and local wind direction and speed at background sites. *Science of The Total Environment,* 409(6), p. 1134–1144.

Donnelly, A., Misstear, B. & Broderick, B., 2015. Real time air quality forecasting using integrated parametric and non-parametric regression techniques. *Atmospheric Environment,* Volume 103, pp. 53-65.

Dons, E. et al., 2013. Modeling temporal and spatial variability of traffic-related air pollution: Hourly land use regression models for black carbon. *Atmospheric Environment,* Volume 74, pp. 237-246.

Dorling, S., Foxall, R., Mandic, D. & Cawley, G., 2003. Maximum likelihood cost functions for neural network models. *Atmospheric Environment,* 37(24), p. 3435–3443.

Do, T. & Gatica-Perez, D., 2014. Where and what: Using smartphones to predict next locations and applications in daily life. *Pervasive and Mobile Computing,* Volume 12, p. 79–91.

Dunning, B. & Ford, I., 2003. *Personal automated transportation: status and potential of personal rapid transit,* s.l.: Advanced Transit Association.

Dutot, A., Rynkiewicz, J., Steiner, F. & Rude, J., 2007. A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions. *Environmental Modelling & Software,* 22(9), p. 1261–1269.

Eagle, N., Macy, M. & Claxton, R., 2010. Network Diversity and Economic Development. *Science,* Volume 328, pp. 1029-1031.

Eagle, N., Montjoye, Y. & Bettencourt, L., 2009a. *Community Computing: Comparisons between Rural and Urban Societies using Mobile Phone Data.* s.l., International Conference on Computational Science and Engineering, pp. 144-151.

Eagle, N. & Pentland, A., 2006. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing,* Volume 10, p. 255–268.

Eagle, N. & Pentland, A., 2009. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology,* Volume 63, p. 1057–1066.

Eagle, N., Pentland, A. & Lazer, D., 2009b. Inferring friendship network structure by using mobile phone data. *Porceedings of the National Academy of Science,* 106(36), p. 15274–15278.

EEA, 2011. *Air quality in Europe (Technical Report 12/2011),* s.l.: European Environment Agency.

El-Rabbany, A., 2006. *Inroduction to GPS: The Global Positioning System.* 2nd ed. s.l.:Artech House Publishers.

Etter, V. et al., 2013. Where to go from here? Mobility prediction from instantaneous information. *Pervasive and Mobile Computing,* Volume 9, p. 784–797.

Farber, H., 2014. *Why You Can't Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers,* Boston, MA: National Bureau of Economic Research.

Farkas, C. & Dan, A., 2014. Stochastic Modeling of Electric Car Charging Station for a Taxi Fleet. *Periodica Polytechnica Electrical Engineering and Computer Science,* 58(4), pp. 175-181.

Farrahi, K. & Gatica-Perez, D., 2008. *What Did You Do Today? Discovering Daily Routines from Large-Scale Mobile Data.* Vancouver, British Columbia, Canada, ACM, pp. 849-852.

Farrahi, K. & Gatica-Perez, D., 2011. Discovering routines from large-scale human locations using probabilistic topic models. *Transactions on Intelligent Transportation Systems,* 2(1), p. 3:1–3:27.

Faus-Kessler, T., Kirchner, M. & Jakobi, G., 2008. Modelling the decay of concentrations of nitrogenous compounds with distance from roads. *Atmospheric Environment,* Volume 42, p. 4589–4600.

Fernando, T., Maier, H., Dandy, D. & May, R., 2005. *Efficient Selection of Inputs for Artificial Neural Network Models.* s.l., s.n., pp. 1806-1813.

Fleiss, J., 1981. *Statistical methods for rates and proportions.* 2nd ed. New York: John Wiley.

Fontaine, M., Yakkala, A. & Smith, B., 2007. *Probe Sampling Strategies for Traffic Monitoring Systems Based on Wireless Location Technology,* s.l.: Virginia Transportation Research Council.

Fontes, T., 2010. *Impacte da qualidade do ar urbana na saúde pública: O caso da cidade do Porto.* Aveiro, Portugal: Universidade de Aveiro, Departamento de Ambiente e Ordenamento, PhD Thesis.

Franco, V. et al., 2013. Road vehicle emission factors development: A review. *Atmospheric Environment,* Volume 70, pp. 84-97.

Freedman, D., 2009. *Statistical Models Theory and Practice.* 2nd ed. s.l.:Cambridge University Press.

Frias-Martinez, V., Soguero, C. & Frias-Martinez, E., 2012. *Estimation of urban commuting patterns using cellphone network data.* Beijing, China, ACM.

Fruin, S. et al., 2008. Measurements and predictors of on-road ultrafine particle concentrations and associated pollutants in Los Angeles. *Atmospheric Environment,* Volume 42, p. 207–219.

Frutos, S. & Castro, M., 2014. Using smartphones as a very low-cost tool for road inventories. *Transportation Research Part C,* Volume 38, p. 136–145.

Fuller, C. et al., 2012. Estimation of ultrafine particle concentrations at near-highway residences using data from local and central monitors. *Atmospheric Environment,* Volume 57, pp. 257-265.

Fusco, G., 2003. Looking for sustainable urban mobility through bayesian network. *SR Scienze Regionali / Italian Journal of Regional Science,* Volume 3, pp. 87-106.

Gama, J. et al., 2012. *Extração de Conhecimento de Dados - Data Mining.* Lisbon: Edições Sílabo.

Gardner, M. & Dorling, S., 1998. Artificial neural networks (the multi-layer perceptron) – a review of applications in the atmospheric sciences. *Atmos. Environ. ,* Volume 32, p. 2627–2636.

Gardner, M. & Dorling, S., 1999. Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. *Atmos. Environ.,* Volume 33, p. 709–719.

Geoff, R., 2006. Mobile Phones as Traffic Probes: Practices, Prospects and Issues. *Transport Reviews,* 26(3), pp. 275-291.

George, K., Ziska, L., Bunce, J. & Quebedeaux, B., 2007. Elevated atmospheric CO2concentration and temperature across an urban–rural transect. *Atmospheric Environment,* Volume 41, p. 7654–7665.

Geurs, K., Thomas, T., Bijlsma, M. & Douhou, S., 2015. Automatic trip and mode detection with MoveSmarter: first results from the Dutch Mobile Mobility Panel. *Transportation Research Procedia,* Volume 11, p. 247–262.

Ge, Y., Xiong, H., Liu, C. & Zhou, Z., 2011. *A Taxi Driving Fraud Detection System.* s.l., s.n.

Ge, Y. et al., 2010. *An Energy-Efficient Mobile Recommender System.* s.l., ACM Press, pp. 899-908.

Gilbert, N., Goldberg, M., Brook, J. & Jerrett, M., 2007. The influence of highway traffic on ambient nitrogen dioxide concentrations beyond the immediate vicinity of highways. *Atmospheric Environment,* Volume 41, p. 2670–2673.

Gilbert, N., Woodhouse, S., Stieb, D. & Brook, J., 2003. Ambient nitrogen dioxide and distance from a major highway. *Science of the Total Environment,* Volume 312, p. 43–46.

Girardin, F. & Blat, J., 2010. *The co-evolution of taxi drivers and their in-car navigation systems.* s.l., s.n.

Girardin, F. et al., 2008. Digital footprinting: Uncovering tourists with User-Generated Content. *Pervasive Computing,* 7(4), pp. 36-43.

Gonzales, E., Yang, C., Morgul, E. & Ozbay, K., 2014. *Modeling Taxi Demand with GPS Data from Taxis and Transit,* San José, CA, USA: Mineta National Transit Research Consortium.

Gonzalez, H. et al., 2007. *Adaptive Fastest Path Computation on a Road Network: A Traffic Mining Approach.* Vienna, Austria, ACM.

Gonzalez, M., Hidalgo, C. & Barabasi, A., 2008. Understanding individual human mobility patterns. *Nature,* Volume 453, pp. 779-783.

Gordon, M. et al., 2012. Measured and modeled variation in pollutant concentration near roadways. *Atmospheric Environment,* Volume 57, pp. 138-145.

Grambsch, A., 2001. *Climate change and air quality,* Washington DC, USA: Department of Tansportation Centre for Climate Change and Environmental Forecasting.

Grau, J., Romeu, M., Mitsakis, E. & Stamos, I., 2013. Agent Based Modeling for Simulation of Taxi Services. *Journal of Traffic and Logistics Engineering,* 1(2), pp. 159-163.

Grimmond, C. et al., 2012. Local-scale fluxes of carbon dioxide in urban environments: methodological challenges and results from Chicago. *Environmental Pollution,* Volume 116, p. S243–S254.

Grivas, G. & Chaloulakou, A., 2006. Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. *Atmospheric Environment,* 40(7), p. 1216–1229.

Gühnemann, A., Schäfer, R., Thiessenhusen, K. & Wagner, P., 2004. *Monitoring Traffic and Emissions by Floating Car Data,* Sydney: Institute of Transport Studies, Australian Key Centre in Transport Management, University of Sydney.

Gundlegard, D. & Karlsson, J., 2006. *Generating Road Traffic Information from Cellular Networks - New Possibilities in UMTS.* s.l., IEEE, pp. 1128-1133.

Gundlegard, D. & Karlsson, J., 2009. Road Traffic Estimation using Cellular Network Signaling in Intelligent Transportation Systems. In: M. Zhou, Y. Zhang & L. Yan, eds. *Wireless Technologies in Intelligent Transportation Systems.* s.l.:Nova Science Publishers.

Hao, W., 2004. *Improving Taxi Dispatch Services with Real-Time Traffic and Customer Information,* s.l.: National University of Singapore, PhD Thesis.

Hargreaves, P. et al., 2000. Local and seasonal variations in atmospheric nitrogen dioxide levels at Rothamsted, UK, and relationships with meteorological conditions. *Atmospheric Environment,* Volume 34, pp. 843-853.

Hellinga, B., Fu, L. & Takada, H., 2005. *Traffic Network Condition Monitoring via Mobile Phone Location Referencing.* Toronto, Ontario,Canada, s.n.

Hellinga, B. & Izadpanah, P., 2007. *An Opportunity Assessment of Wireless Monitoring of Network-Wide Road Traffic Conditions,* Ontario, Canada: Ministry of Transportation of Ontario.

Hellinga, B., Izadpanah, P., Takada, H. & Fu, L., 2008. Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments. *Transportation Research Part C,* Volume 16, p. 768–782.

Herrera, J. et al., 2010. Evaluation of traffic data obtained via GPSenabled mobile phones: The Mobile Century field experiment. *Transportation Research Part C: Emerging Technologies,* Volume 18, pp. 568-583.

Hidalgo, C. & Rodriguez-Sickert, C., 2008. The dynamics of a mobile phone network. *Elsevier Physica A,* Volume 387, p. 3017–3024.

Hiller, R., McFadden, J. & Kljun, N., 2011. Interpreting CO2Fluxes Over a Suburban Lawn: The Influence of Traffic Emissions. *Boundary-Layer Meteorol,* Volume 138, p. 215–230.

Hillson, S. & Santis, M., 2012. *Traffic Monitoring Application of Cellular Positioning Technology: Proof of Concept,* Calgary, Canada: Cell-Loc Inc..

Hogg, R. & Ledolter, J., 1987. *Engineering Statistics.* New York, USA: MacMillan.

Holmes, N. & Morawska, L., 2006. A review of dispersion modelling and its application to the dispersion of particles: An overview of different dispersion models available. *Atmospheric Environment,* Volume 40, p. 5902–5928.

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.* Krabi, Thailand, IEEE, pp. 13-16.

Hooyberghs, J. et al., 2005. A neural network forecast for dailyaverage PM10 concentrations in Belgium. *Atmospheric Environment,* Volume 39, p. 3279–3289.

Hopfner, M., Lemmer, K. & Ehrenpfordt, I., 2007. *Cellular Data for Traffic Management.* s.l., s.n.

Horanont, T. et al., 2013. Weather Effects on the Patterns of People's Everyday Activities: A Study Using GPS Traces of Mobile Phone Users. *Plos One*, 18 December.

Hossain, L., Chung, K. & Murshed, S., 2007. Exploring Temporal Communication Through Social Networks. In: C. Baranauskas, P. Palanque, J. Abascal & S. Barbosa, eds. *Human-Computer Interaction – INTERACT 2007 .* Rio de Janeiro, Brazil: Springer, p. 19–30.

Hoteit, S. et al., 2014. Estimating human trajectories and hotspots through mobile phone data. *Computer Networks,* Volume 64, p. 296–307.

Hou, Y. et al., 2013. *Towards Efficient Vacant Taxis Cruising Guidance.* s.l., IEEE, pp. 54-60.

Hsiao, W. & Chang, S., 2005. *Segment based Traffic Information Estimation Method Using Cellular Network Data.* Vienna, Austria, IEEE, pp. 44-49.

Hsiao, W. & Chang, S., 2006. *The Optimal Location Update Strategy of Cellular Network Based Traffic Information System.* Toronto, Canada, IEEE, pp. 248-253.

Huang, H., Akustu, Y., Ari, M. & Tamura, M., 2001. Analysis of photochemical pollution in summer and winter using a photochemical box model in the center of Tokyo, Japan. *Chemosphere,* Volume 44, pp. 223-230.

Huang, H. et al., 2010. *META: a Mobility Model of MEtropolitan TAxis Extracted from GPS Traces.* s.l., IEEE.

Hu, H. et al., 2012a. *Pick-up tree based route recommendation from Taxi Trajectories.* s.l., s.n., p. 471–483.

Hu, J., Cao, W., Luo, J. & Yu, X., 2009a. *Dynamic modeling of urban population travel behavior based on data fusion of mobile phone positioning data and FCD.* s.l., s.n., pp. 1-5.

Hurley, P., 2006. An Evaluation and Inter-comparison of AUSPLUME, AERMOD and TAPM for Seven Field Datasets of Point Source Dispersion. *Clean Air Environmental Quality,* 40(1), pp. 45-50.

Hu, S. et al., 2009c. A wide area of air pollutant impact downwind of a freeway during pre-sunrise hours. *Atmospheric Environment,* Volume 43, p. 2541–2549.

Husna, H., Phithakkitnukoon, S., Baatarjav, E. & Dantu, D., 2008. *Quantifying Presence using Calling Patterns.* s.l., s.n.

Hu, S. et al., 2012c. Observation of elevated air pollutant concentrations in a residential neighborhood of Los Angeles California using a mobile platform. *Atmospheric Environment 51,* Volume 51, pp. 311-319.

Hu, S. et al., 2009b. *Vehicular Sensing System for CO2 Monitoring Applications.* s.l., s.n.

Hu, X., Sao, G., Chiu, Y. & Lin, D., 2012b. Modeling Routing Behavior for Vacant Taxicabs in Urban Traffic Networks. *Journal of the Transportation Research Board,* Volume 2284, p. 81–88.

INFRAS, 1999. *Handbook of Emission Factors for Road Transport, HBEFA, v1.2,* Bern: s.n.

Ioannides, Y. & Overman, G., 2003. Zipf's law for cities: an empirical examination. *Regional Science and Urban Economics,* Volume 33, p. 127–137.

Iqbal, S., Choudhury, C., Wang, P. & González, M., 2014. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C,* Volume 40, p. 63–74.

Isaacman, S. et al., 2011. *Identifying Important Places in People's Lives from Cellular Network Data.* San Francisco, California, USA, s.n.

Ivan, C. & Popa, R., 2015. A Cloud based Mobile Dispatching System with Built-in Social CRM Component: Design and Implementation. *Computers,* Volume 4.

Järv, O., Ahas, R. & Witlox, F., 2014. Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C,* Volume 38, p. 122–135.

Jiang, Z. et al., 2013. Calling patterns in human communication dynamics. *Proceedings of the National Academy of Sciences (PNAS),* 110(5), p. 1600–1605 .

Jianqin, Z. et al., 2015. A space-time visualization analysis method for taxi operation in Beijing. *Journal of Visual Languages and Computing,* Volume 31, p. 1–8.

Jin, J., Qiu, Z. & Ran, B., 2006. *Intelligent Route-based Speed Estimation using Timing Advance.* Toronto, Canada, Toronto, pp. 194-197.

Johnson, M. et al., 2010. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. *Atmospheric Environment,* Volume 44, pp. 3660-3668.

Jolliffe, I., 2002. *Principal Component Analysis.* 2nd ed. s.l.:Springer.

Juhos, I., Makra, L. & Tóth, B., 2008. Forecasting of traffic origin NO and NO2 concentrations by Support Vector Machines and neural networks using Principal Component Analysis. *Simulation Modelling Practice and Theory,* Volume 16, p. 1488–1502.

Karlsson, H., 2004. Ammonia, nitrous oxide and hydrogen cyanide emissions from five passenger vehicles. *Science of the Total Environment,* Volume 334/335, pp. 125-132.

Kartika, C., 2015. *Visual Exploration of Spatial-Temporal Traffic Congestion Patterns Using Floating Car Data,* München, Germany: Technische Universität München, MSc Thesis.

Kassomenos, P., Karakitsios, S. & Papaloukas, C., 2006. Estimation of daily traffic emissions in a South-European urban agglomeration during a workday.

Evaluation of several “what if” scenarios. *Science of The Total Environment,* 270(2-3), p. 480–490.

Kazmier, L., 2003. *Schaum's Outline of Business Statistics.* s.l.:McGraw Hill Professional.

Ke, H., Ondov, J. & Rogge, W., 2013. Detailed emission profiles for on-road vehicles derived from ambient measurements during a windless traffic episode in Baltimore using a multi-model approach. *Atmospheric Environment,* Volume 81, pp. 280-287.

Kennedy, P., 2008. *A Guide to Econometrics.* Massachusetts: Blackwell Publishing.

Kim, Y. et al., 2014. *Activity Recognition for a Smartphone Based Travel Survey Based on Cross-User History Data.* Stockholm, 22nd International Conference on Pattern Recognition (ICPR), pp. 432-437.

Kirchner, M. et al., 2005. Elevated NH3 and NO2 air concentrations and nitrogen deposition rates in the vicinity of a highway in Southern Bavaria. *Atmospheric Environment,* Volume 39, p. 4531–4542.

Kodama, Y. et al., 2002. Environmental NO2 Concentration and Exposure in Daily Life along Main Roads in Tokyo. *Environmental Research,* Volume 89, pp. 236-244.

Kolehmainen, M., Martikainen, H. & Ruuskanen, J., 2001. Neural networks and periodic components used in air quality forecasting. *Atmos. Environ.,* Volume 35, p. 815–825.

Kordowski, K. & Kuttler, W., 2010. Carbon dioxidefluxes over an urban park area. *Atmospheric Environment,* pp. 1-9.

Kota, S., Ying, Q. & Zhang, Y., 2013. Simulating near-road reactive dispersion of gaseous air pollutants using a three-dimensional Eulerian model. *Science of the Total Environment,* Volume 454–455, p. 348–357.

Kousoulidou, M. et al., 2013. Use of portable emissions measurement system (PEMS) for the development and validation of passenger car emission factors. *Atmospheric Environment,* Volume 64, pp. 329-338.

Krings, G., Calabrese, F., Ratti, C. & Blondel, V., 2009a. *Scaling Behaviors in the Communication Network Between Cities.* s.l., International Conference on Computational Science and Engineering, pp. 936-941.

Krings, G., Calabrese, F., Ratti, C. & Blondel, V., 2009b. Urban Gravity: a Model for Intercity Telecommunication Flows. *Journal of Statistical Mechanics: Theory and Experiment,* pp. 1-8.

Krumm, J. & Horvitz, E., 2006. *Predestination: Inferring Destinations from Partial Trajectories.* Orange County, CA, USA, s.n.

Krzyzanowski, M. K.-D. B. & Schneider, J., 2005. *Health effects of transport-related air pollution.* Copenhagen Ø, Denmark: World Health Organization - Regional Office for Europe.

Kukkonen, J. et al., 2003. Extensive evaluation of neural network models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. *Atmospheric Environment,* Volume 37, p. 4539–4550.

Kumar, P. & Imam, B., 2013. Footprints of air pollution and changing environment on the sustainability of built infrastructure. *Science of the Total Environment,* Volume 444, p. 85–101.

Kumar, P. et al., 2015. The rise of low-cost sensing for managing air pollution in cities. *Environment International,* Volume 75, p. 199–205.

Lameiras, H. & Povoas, F., 2005. *Relatório da Qualidade do Ar na Região Centro,* s.l.: Comissão de Coordenação e Desenvolvimento do Ar na Regiao Centro.

Larsen, R. & Warne, R., 2010. Estimating confidence intervals for eigenvalues in exploratory factor analysis. *Behavior Research Methods,* pp. 871-876.

Leduc, G., 2008. *Road Traffic Data: Collection Methods and Applications,* Luxembourg: European Commission.

Lee, D., Park, M., J. & Cho, W., 2014. Estimating the number of taxi passengers depending on time zones and weather conditions. *Advanced Science and Technology Letters,* Volume 79, pp. 92-96.

Lee, D., Wang, H., Cheu, R. & Teo, S., 2004. Taxi dispatch system based on current demands and real-time traffic conditions. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 1882, p. 193–200.

Lee, D. & Wu, X., 2013. *Dispatching Strategies for the Taxi-Customer Searching Problem the Booking Taxi Service.* Washington D.C., s.n.

Lee, J., Shin, I. & Park, G., 2008. *Analysis of the passenger pick-up pattern for taxi location recommendation.* Gyeongju, China, IEEE, p. 199–204.

Liang, X., Zhao, J. & Xu, K., 2015. A general law of human mobility. *Information Sciences - Science China,* Volume 58, p. 14.

Liang, X. et al., 2011. The scaling of human mobility by taxis is exponential. *Physica A.*

Liao, Z., 2003. Real-time taxi dispatching using Global Positioning Systems. *Communications of the ACM,* 46(5), p. 81–83.

Li, B. et al., 2011a. *Hunting or Waiting? Discovering Passenger-Finding Strategies from a Large-scale Real-world Taxi Dataset.* s.l., IEEE, pp. 63-69.

Liley, J. et al., 2000. Stratospheric NO2 variations from a long time series at Lauder, New Zealand. *Geophysical Research,* 105(9), pp. 11633-11640.

Li, N. & Chen, G., 2009. *Multi-layered friendship modeling for location-based mobile social networks.* s.l., s.n., pp. 1-10.

Lin, M. & Hsu, W., 2014. Mining GPS data for mobility patterns: A survey. *Pervasive and Mobile Computing,* Volume 12, p. 1–16.

Linoff, G. & Berry, M., 2011. *Data Mining Techniques.* 3rd ed. Indianapolis: Wiley.

Lin, W., Xu, X., Ge, B. & Liu, X., 2011. Gaseous pollutants in Beijing urban area during the heating period 2007–2008: variability, sources, meteorological, and chemical impacts. *Atmospheric Chemistry and Physics,* Volume 11, p. 8157–8170.

Lin, Y., Li, W., Qiu, F. & Xu, H., 2012. Research on Optimization of Vehicle Routing Problem for Ride-sharing Taxi. *Procedia - Social and Behavioral Sciences,* Volume 43, p. 494 – 502.

Li, Q., Zeng, Z., Yang, B. & Zhang, T., 2009a. *Hierarchical route planning based on taxi gps-trajectories.* s.l., IEEE.

Li, Q. et al., 2011c. Path-finding through flexible hierarchical road networks: An experiential approach using taxi trajectory data. *International Journal of Applied Earth Observation and Geoinformation,* Volume 13, p. 110–119.

Li, S., 2006. *Multi-Attribute Taxi Logistics Optimization.* s.l.:Massachusetts Institute of Technology.

Liu, C., 2010. Measuring and prioritising value of mobile phone usage. *International Journal of Mobile Communications ,* 8(1), pp. 41-52 .

Liu, F., Janssens, D., Wets, G. & Cools, M., 2013. Annotating mobile phone location data with activity purposes using machine learning algorithms. *Expert Systems with Applications,* Volume 40, p. 3299–3311.

Liu, H., Danczyk, A., Brewer, R. & Starr, R., 2008. Evaluation of Cell Phone Traffic Data in Minnesota. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 2086, pp. 1-7.

Liu, H. Z., Feng, J. W., Jarvi, L. & Vesala, T., 2012. Eddy covariance measurements of CO2 and energy fluxes in the city of Beijing. *Journal Atmospheric Chemistry and Physics,* Volume 12, pp. 7677-7704.

Liu, K., Yamamoto, T. & Morikawa, T., 2009a. Feasibility of using taxi dispatch system as probes for collecting traffic information. *Intelligent Transportation Systems,* 13(1), p. 16–27.

Liu, L., Andris, C., Bidderman, A. & Ratti, C., 2010a. *Revealing taxi drivers mobility intelligence through his trace.* s.l., s.n., pp. 105-120.

Liu, L., Andris, C., Bidderman, A. & Ratti, C., 2010b. Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems.*

Liu, L., Biderman, A. & Ratti, C., 2009b. *Urban mobility landscape: Real Time Monitoring of Urban Mobility Patterns.* s.l., s.n.

Liu, S. et al., 2010c. *Towards Mobility-based Clustering.* Washington, DC, USA, ACM, pp. 919-928.

Liu, X. et al., 2012a. *Mining Large-Scale, Sparse GPS Traces for Map Inference: Comparison of Approaches.* Beijing, China, ACM.

Liu, X. et al., 2012b. *Road Recognition using Coarse-grained Vehicular Traces.* s.l., Hewlett-Packard Laboratories.

Liu, Y. et al., 2012d. Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems,* 14(4), pp. 463-483.

Liu, Y., Sui, Z., Kang, C. & Gao, Y., 2014. Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-In Data. *PLoS ONE,* 9(1).

Liu, Y., Wang, F., Xiao, Y. & Gao, S., 2012c. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning,* Volume 106, p. 73–87.

Li, X., Li, M., Shu, W. & Wu, M., 2007. *A practical map-matching algorithm for GPS-based vehicular networks in Shanghai urban area.* s.l., s.n., p. 454–457.

Li, X. et al., 2011b. Prediction of urban human mobility using large-scale taxi traces and its applications. *Front. Comput. Sci.,* 6(1), p. 111–121.

Li, X. et al., 2009b. Performance Evaluation of Vehicle-Based Mobile Sensor Networks for Traffic Monitoring. *Transactions on Vehicular Technology,* 58(4), pp. 1647-1653.

Lou, Y., Zhang, C. & Zheng, Y., 2009. *Map-Matching for Low-Sampling-Rate GPS Trajectories.* Seattle, WA, USA, ACM.

Lv, W., Ma, S., Liang, C. & Zhu, T., 2011. *Effective data identification in travel time estimation based on cellular network signaling.* Toulouse, IEEE, pp. 1-5.

Maciejewski, M. & Nagel, K., 2013. Simulation and dynamic optimization of taxi services in MATSim. *Transportation Science,* p. 34.

MacKay, D., 2003. *Information Theory, Inference, and Learning Algorithms.* 4th ed. s.l.:Cambridge University Press.

Madsen, C. et al., 2011. Comparison of land-use regression models for predicting spatial NOx contrasts over a three year period in Oslo, Norway. *Atmospheric Environment,* Volume 45, pp. 3576-3583.

Maerivoet, S. & Logghe, S., 2007. *Validation of Travel Times based on Cellular Floating Vehicle Data.* s.l., s.n., p. 9.

Mak, K. & Hung, W., 2008. Developing air pollutant profiles using routine monitoring data in road tunnels: A note. *Transportation Research - Part D,* Volume 13, p. 404–411.

Mao, X. et al., 2012. *CitySee: Urban CO2 Monitoring with Sensors.* s.l., s.n.

Maroco, J., 2005. *Análise Estatística - Com Utilização do SPSS.* Lisbon, Portugal: Edições Silabo.

Martinez, L., Correia, G. & Viegas, J., 2013. *Modeling the taxi mark et for the Lisbon municipality.* Rio de Janeiro, Brazil, s.n.

Ma, S., Zheng, Y. & Wolfson, O., 2013. *T-Share: A large-scale dynamic taxi ridesharing service.* Brisbane, IEEE, pp. 410 - 421 .

Matese, A. et al., 2009. Carbon Dioxide Emissions of the City Center of Firenze, Italy: Measurement, Evaluation, and Source Partitioning. *Journal of Applied Meteorology and Climatology,* Volume 48, pp. 1940-1948.

Mathew, J. & Xavier, P., 2014. A Survey on Using Wireless Signs for Road Traffic Detection. *International Journal of Research in Engineering and Technology,* 3(1), pp. 97-102.

Matsushima, K. & Kobayashi, K., 2007. *Differentiation of Taxi Spot Markets and Social Welfare.* Berkeley, California, USA, s.n.

Matthes, S., Grewe, V., Sausen, R. & Roelofs, G., 2007. Global impact of road traffic emissions on tropospheric ozone. *Atmospheric Chemistry and Physics,* Volume 7, p. 1707–1718.

Mavroidis, I. & Ilia, M., 2012. Trends of NOx, NO2 and O3 concentrations at three different types of air quality monitoring stations in Athens, Greece. *Atmospheric Environment,* Volume 63, pp. 135-147.

May, R., Dandy, G. & Maier, H., 2011. Review of Input Variable Selection Methods for Artificial Neural Networks. In: *Artificial Neural Networks - Methodological Advances and Biomedical Applications.* Shanghai, China: InTech, pp. 20-44.

Melkonyan, A. & Kuttler, W., 2012. Long-term analysis of NO, NO2 and O3 concentrations in North Rhine-Westphalia, Germany. *Atmospheric Environment,* Volume 60, pp. 316-326.

Melo, P., Akoglu, L., Faloutsos, C. & Loureiro, A., 2010. *Surprising Patterns for the Call Duration Distribution of Mobile Phone Users.* s.l., s.n.

Merbitz, H., Fritz, S. & Schneider, C., 2012. Mobile measurements and regression modeling of the spatial particulate matter variability in an urban area. *Science of the Total Environment,* Volume 438, p. 389–403.

Meyer, R. & Krueger, D., 2004. *Minitab Guide to Statistics.* 3rd ed. s.l.:Pearson.

Miao, F. et al., 2015. *Taxi dispatch with real-time sensing data in metropolitan areas: a receding horizon control approach.* s.l., ACM/IEEE, pp. 100-109 .

Miller, G., 2009. *Living in the Environment: Principles, Connections, and Solutions.* 17th ed. s.l.:Brooks Cole.

Miluzzo, E. et al., 2008. *Sensing meets mobile social networks: The design, implementation and evaluation.* Raleigh, North Carolina, USA, s.n., p. 337–350.

Misra, A., Roorda, M. & MacLean, H., 2013. An integrated modelling approach to estimate urban traffic emissions. *Atmospheric Environment,* Volume 73, pp. 81-91.

Mitchell, T. M., 1997. *Machine Learning.* New York: McGraw-Hill.

Moltchanov, S. et al., 2015. On the feasibility of measuring urban air pollution by wireless distributed sensor networks. *Science of the Total Environment,* Volume 502, p. 537–547.

Monreale, A., Pinelli, F., Trasarti, R. & Giannotti, F., 2009. *Where Next: A location predictor on trajectory pattern mining.* Paris, ACM, pp. 637-645.

Monteiro, A., Miranda, A., Borrego, C. & Vautard, R., 2007. Air quality assessment for Portugal. *Science of The Total Environment,* 373(1), p. Pages 22–31.

Monteiro, A., Vautard, R., Borrego, C. & Miranda, A., 2005. Long-term simulations of photo oxidant pollution over Portugal using the CHIMERE model. *Atmos Environ,* 39(17), p. 3089–3101.

Moreira-Matias, L., Gama, J., Ferreira, M. & Damas, L., 2012a. *A Predictive Model for the Passenger Demand on a Taxi Network.* Anchorage, Alaska, USA, IEEE, pp. 1014-1020.

Moreira-Matias, L. et al., 2014b. On Predicting the Taxi-Passenger Demand: A Real-Time Approach. *Progress in Artificial Intelligence,* Volume 8154, p. 54–65.

Moreira-Matias, L. et al., 2016. Time-evolving O-D matrix estimation using high-speed GPS data streams. *Expert Systems With Applications,* Volume 44, p. 275–288.

Moreira-Matias, L. et al., 2012b. *Online Predictive Model for Taxi Services.* s.l., Springer-Verlag, p. 230–240.

Moreira-Matias, L. et al., 2013. Predicting Taxi–Passenger Demand Using Streaming Data. *IEEE Transactions on Intelligent Transportation Systems,* 14(3), pp. 1393-2403.

Moreira-Matias, L. et al., 2014a. *An Online Learning Framework for Predicting the Taxi Stand's Profitability.* s.l., IEEE.

Moriwaki, R. & Kanda, M., 2004. Seasonal and Diurnal Fluxes of Radiation, Heat, Water Vapor, and Carbon Dioxide over a Suburban Area. *Journal of Applied Meteorology,* Volume 43, pp. 1700-1711.

Mukerjee, S. et al., 2009. Spatial analysis and land use regression of VOCs and NO2 from school-based urban air monitoring in Detroit/Dearborn, USA. *Science of the Total Environment,* Volume 407, p. 4642–4651.

Ndoke, P. N. & Jimoh, O. D., 2005. Impacts of Traffic Emission on Air Quality in a Developing City of Nigeria. *AU Journal of Technology,* Volume 8, pp. 222-228.

Nemitz, E. et al., 2007. *Micrometeorological measurements of anthropogenic VOC emissions from urban areas.* s.l., s.n., pp. 68-71.

Nickerson, R., Isaac, H. & Mak, B., 2008. A multi-national study of attitudes abou mobile phone use in social settings. *International Journal of Mobile Communications,* 6(5), p. 541–563.

Ning, Z., Wubulihairen, M. & Yang, F., 2012. PM, NOx and butane emissions from on-road vehicle fleets in Hong Kong and their implications on emission control policy. *Atmospheric Environment,* Volume 61, pp. 265-274.

Niskaa, H. et al., 2004. Evolving the neural network model for forecasting air pollution time series. *Engineering Applications of Artificial Intelligence,* 17(2), p. Engineering Applications of Artificial Intelligence.

Nobis, C. & Lenz, B., 2009. Communication and mobility behaviour – a trend and panel analysis of the correlation between mobile phone use and mobility. *Journal of Transport Geography,* Volume 17, p. 93–103.

Ojolo, S., Oke, S., Dinrifo, R. & Eboda, F., 2007. A survey on the effects of vehicle emissions on human health in Nigeria. *Journal of Rural and TropicalPublic Health,* Volume 6, pp. 16-23.

Oliveira, C. et al., 2010. Road traffic impact on urban atmospheric aerosol loading at Oporto, Portugal. *Atmospheric Environment,* Volume 44, pp. 3147-3158.

Onnela, J. et al., 2011. Geographic Constraints on Social Network Groups. *PLoS ONE (e16939),* 6(4), p. 7.

Onnela, J. et al., 2007. Structure and tie strengths in mobile communication networks. *PNAS,* 104(18), p. 7332–7336.

Orey, P., Fernandes, R. & Ferreira, M., 2012. *Empirical Evaluation of a Dynamic and Distributed Taxi-Sharing System.* Anchorage, Alaska, USA, IEEE, pp. 140-147.

Otsason, V., Varshavsky, A., LaMarca, A. & Lara, E., 2005. *Accurate GSM Indoor Localization.* s.l., Springer Berlin / Heidelberg, p. 141–158.

Otsasson, V., 2005. *Accurate GSM Indoor Localization Using Wide GSM Fingerprinting,* Tartu, Estonia: University of Tartu, MSc Thesis.

Padró-Martínez, L. et al., 2012. Mobile monitoring of particle number concentration and other traffic-related air pollutants in a near-highway neighborhood over the course of a year. *Atmospheric Environment,* Volume 61, pp. 253-264.

Pallant, J., 2005. *SPSS survival manual: a step by step guide to data analysis using SPSS for Windows.* Maidenhead(Berkshire): Open University Press.

Pandey, S. et al., 2008. Long-term study of NOx behavior at urban roadside and background locations in Seoul, Korea. *Atmospheric Environment,* Volume 42, p. 607–622.

Pan, G. et al., 2013. Land-Use Classification Using Taxi GPS Traces. *IEEE Transactions on Intelligent Transportation Systems,* 14(1), pp. 113-124.

Parshall, L. et al., 2009. Modeling energy consumption and CO2 emissions at the urban scale: Methodological challenges and insights from the United States. *Energy Policy.*

Pattinson, W., Longley, I. & Kingham, S., 2014. Using mobile monitoring to visualise diurnal variation of traffic pollutants across two near-highway neighbourhoods. *Atmospheric Environment,* Volume 94, pp. 782-792.

Patton, A. et al., 2014. Spatial and temporal differences in traffic-related air pollution in three urban neighborhoods near an interstate highway. *Atmospheric Environment,* Volume 99, pp. 309-321.

Paul, U., Subramanian, S., Buddhikot, M. & Das, S., 2011. *Understanding traffic dynamics in cellular data networks.* Shanghai, IEEE, pp. 882-890 .

Peng, C. et al., 2012. Collective Human Mobility Pattern from Taxi Trips in Urban Area. *PLoS ONE,* 7(4).

Pereira, F. et al., 2013. The Future Mobility Survey: Overview and Preliminary Evaluation. *Eastern Asia Society for Transportation Studies,* Volume 9.

Perez, P. & Reyes, J., 2001. Prediction of Particlulate Air Pollution using Neural Techniques. *Neural Comput. Appl.,* 10(2), p. 165–171.

Perez, P. & Trier, A., 2001. Prediction of NO and NO2 concentrations near a street with heavy traffic in Santiago, Chile. *Atmos. Environ. ,* Volume 35, p. 1783–1789.

Perez, P., Trier, A. & Reyes, J., 2000. Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. *Atmos. Environ. ,* Volume 34, p. 1189–1196.

Pernigotti, D., Georgieva, E., Thunis, P. & Bessagnet, B., 2012. Impact of meteorology on air quality modeling over the Po valley in northern Italy. *Atmospheric Environment,* Volume 51, pp. 303-310.

Phithakkitnukoon, S., Calabrese, F., Smoreda, Z. & Ratti, C., 2011a. *Out of Sight Out of Mind – How our mobile social network changes during migration,* s.l.: s.n.

Phithakkitnukoon, S. & Dantu, R., 2007. *Predicting Calls – New Service for an Intelligent Phone.* Berlin, Heidelberg: Springer-Verlag, p. 26–37.

Phithakkitnukoon, S. & Dantu, R., 2008. *CPL: Enhancing Mobile Phone Functionality by Call Predicted List.* s.l., s.n., pp. 571-581.

Phithakkitnukoon, S. & Dantu, R., 2010a. *Towards Ubiquitous Computing with Call Prediction.* s.l., s.n.

Phithakkitnukoon, S. & Dantu, R., 2010b. *Mobile Social Closeness and Communication Patterns.* s.l., s.n.

Phithakkitnukoon, S. & Dantu, R., 2011. Mobile social group sizes and scaling ratio. *Springer AI & Soc,* Volume 26, p. 71–85.

Phithakkitnukoon, S. et al., 2010a. *Activity-aware map: Identifying human daily activity pattern using mobile phone data.* s.l., IEEE.

Phithakkitnukoon, S. et al., 2014. Understanding Tourist Behavior using Large-scale Mobile Sensing Approach: A case study of mobile phone users in Japan. *Elsevier Journal on Pervasive and Mobile Computing.*

Phithakkitnukoon, S., Leong, T., Smoreda, Z. & Olivier, P., 2012. *Weather Effects on Mobile Social Interaction: A case study of mobile phone users in Lisbon, Portugal,* Newcastle, UK: Newcastle University.

Phithakkitnukoon, S. & Ratti, C., 2011. *Inferring Asymmetry of Inhabitant Flow using Call Detail Records.* s.l., s.n.

Phithakkitnukoon, S., Smoreda, Z. & Olivier, P., 2011b. *Socio-geography of Human Mobility: A study using longitudinal mobile phone data,* Newcastle, UK: Newcastle University.

Phithakkitnukoon, S. et al., 2010b. *Taxi-Aware Map: Identifying and predicting vacant taxis in the city.* Malaga, Spain, s.n., pp. 86-95.

Pirjola, L. et al., 2012. Spatial and temporal characterization of traffic emissions in urban microenvironments with a mobile laboratory. *Atmospheric Environment,* Volume 63, pp. 156-167.

Pirjola, L. et al., 2006. Dispersion of particles and trace gases nearby a city highway: Mobile laboratory measurements in Finland. *Atmospheric Environment,* Volume 40, p. 867–879.

Pirjola, L. et al., 2004. Sniffer - A novel tool for chasing vehicles and measuring traffic pollutants. *Atmospheric Environment,* Volume 38, p. 3625–3635.

Pleijel, H., Karlsson, G. & Gerdin, E., 2004. On the logarithmic relationship between NO2 concentration and the distance from a highroad. *Science of the Total Environment,* Volume 332, p. 261–264.

Powell, J., Huang, Y., Bastani, F. & Ji, M., 2011. *Towards reducing taxicab cruising time using spatiotemporal profitability maps.* s.l., s.n., pp. 242-260 .

Powers, D., 1998. *Applications and explanations of Zipf's law.* s.l., Association for Computational Linguistics, pp. 151-160.

Public Transport Authority, 2003. *Design and Planning Guidelines for Public Transport Infrastructure: Bus Route Planning and Transit Streets,* Western Australia: Public Transport Authority.

Pulselli, R., 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 & Nature and Ecodynamics,* 3(2), pp. 121-134.

Puntumapon, K. & Pattara-atikom, W., 2008. *Classification of Cellular Phone Mobility using Naive Bayes Model.* Singapore, IEEE, pp. 3021-3025.

Qian, S., Zhu, Y. & Li, M., 2012. *Smart recommendation by mining large-scale GPS traces.* Shanghai, IEEE, pp. 3267-3272.

Qian, X. & Ukkusuri, S., 2015. Spatial variation of the urban taxi ridership using GPS data. *Applied Geography,* Volume 31-42, p. 59.

Qi, G. et al., 2011. *Measuring Social Functions of City Regions from Large-scale Taxi Behaviors.* Seattle, USA, s.n., pp. 21-25.

Qi, G. et al., 2013. *How Long a Passenger Waits for a Vacant Taxi.* Beijing, China, IEEE, pp. 1029-1036 .

Qiu, Z., Cheng, P. & Ran, B., 2007. Investigate the feasibility of traffic speed estimation using cell phones as probes. *International Journal of Services Operations ans Informatics,* 2(1), pp. 53-64.

Qiu, Z. et al., 2014. *Finding Vacant Taxis Using Large Scale GPS Traces.* Macau, China, Springer, pp. 793-804.

Qiu, Z. & Ran, B., 2008. *Evaluation Method Investigation for Non-Traditional Traffic Monitoring Technology,* s.l.: s.n.

Quercia, D. et al., 2010. *Recommending Social Events from Mobile Phone Location Data.* s.l., s.n.

Ratti, C., Pulselli, R., Williams, S. & Frenchman, D., 2009. Mobile landscapes: Using location data from cell phones for urban analysis. *Environ. Plann. B,* 33(5), p. 727–748.

Ratti, C., Sevtsuk, A., Huang, S. & Pailer, R., 2005. *Mobile Landscapes: Graz in Real Time.* Vienna, Austria, s.n.

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,* Volume 36, pp. 824 - 836.

Reades, J., Calabrese, F., Setvsuk, A. & Ratti, C., 2007. Cellular Census: Explorations in Urban Data Collection. *IEEE Computer Society - Pervasive Computing,* 6(3), pp. 30-41.

Reddy, S. et al., 2010 . Using Mobile Phones to Determine Transportation Modes. *Transactions on Sensor Networks (TOSN),* 6(2), p. 27.

Remy, J., 2001. *Computing travel time estimates from GSM signalling messages: the STRIP project.* Oakland, CA, USA, IEEE, p. 4.

Riley, E. et al., 2014. Multi-pollutant mobile platform measurements of air pollutants adjacent to a major roadway. *Atmospheric Environment,* Volume 98, pp. 492-499.

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review,* 65(6), pp. 386-408.

Ross, S., 2010. *Introduction to Probability Models.* 10th ed. Salt Lake City, Utah, USA: Academic press.

Russo, A. & Soares, A., 2013. Hybrid Model for Urban Air Pollution Forecasting: A Stochastic Spatio-Temporal Approach. *Mathematical Geosciences.*

Rutten, B., Wolff, P. & Vlist, M., 2004. *GSM as the Source for Traffic Information.* s.l., s.n.

Sagarra, O. & Diaz-Guilera, A., 2012. *Statistical Complex Analysis of Taxi Mobility in San Francisco.* s.l., s.n.

Saini, R., Satsangi, G. & Taneja, A., 2008. Concentrations of Surface O3, NO2 and CO During Winter Seaons at a Semi-Arid Region. *Indian Journal of Radio and Space Physics,* Volume 37, pp. 121-130.

Salanova, J., 2013. *Taxi services modeling for decision making support.* s.l., s.n.

Salanova, J., Estrada, M., Aifadopoulou, G. & Mitsakis, M., 2011. A review of the modeling of taxi services. *Elsevier Procedia Social and Behavioral Sciences,* Volume 20, p. 150–161.

Salanova, J., Romeu, M. & Amat, C., 2014. Aggregated Modeling of Urban Taxi Services. *Procedia - Social and Behavioral Sciences,* Volume 160, p. 352–361.

Sankar, R. & Civil, L., 1997. *Traffic monitoring and congestion prediction using handoffs in wireless cellular communications.* Phoenix, AZ, USA, IEEE, pp. 520-524.

Santani, D., Balan, B. & Woodard, C., 2008a. *Understanding and Improving a GPS-based Taxi System.* s.l., s.n.

Santani, D., Balan, R. & Woodard, C., 2008b. *Spatio-temporal Efficiency in a Taxi Dispatch System,* Singapore: Research Collection School Of Information Systems - Singapore Management University.

Santi, P. et al., 2013. Taxi pooling in New York City: a network-based approach to social sharing problems.

Santos, M. & Azevedo, C., 2005. *Data Mining - Descoberta de Conhecimento em Bases de Dados.* Lisbon: FCA.

Saraydar, C., Tekinay, S. & Choi, W., 2004. *Efficient vehicular traffic monitoring using mobility management in cellular networks.* s.l., IEEE, pp. 40-45.

Sauret, P., 2003. *Pilot Project for Travel Time Measurements with Mobile Phones in Spain.* s.l., ITS World Congress.

Saville, S., 1993. Automotive options and air quality management in developing countries. *Industry and Environment,* 16(1-2), p. 32.

Schaller Consulting, 2006. *The New York City Taxicab Fact Book,* New York, USA: Schaller Consulting.

Schlink, U. et al., 2003. A rigorous inter-comparison of ground-level ozone predictions. *Atmospheric Environment,* 37(23), p. 3237–3253.

Schmidt, A., Wrzesinsky, T. & Klemm, O., 2008. Gap Filling and Quality Assessment of CO 2and Water Vapour Fluxes above an Urban Area with Radial Basis Function Neural Networks. *Boundary-Layer Meteorol,* Volume 126, p. 389–413.

Schneider, W. & Mrakotsky, E., 2005. *Mobile Phones as a Basis for Traffic State Information.* Vienna, Austria, IEEE, pp. 782-784.

Schroedl, D. et al., 2004. Mining GPS traces for map refinement. *Data Mining and Knowledge Discovery,* Volume 9, p. 59–87.

Seow, K., Dang, N. & Lee, D., 2010. A Collaborative Multiagent Taxi-Dispatch System. *Transactions on Automation Science and Engineering,* 7(3), pp. 607 - 616 .

Sfetsos, A. & Vlachogiannis, D., 2010. Time series forecasting of hourly PM10 using localized linear models. *Journal of Software Engineering and Applications ,* 3(4), pp. 374-384.

Shannon, C. E., 1948. A mathematical theory of communication. *Bell System Technical Journal,* Volume 27, p. 379–423; 623–656.

Shao, H., Lam, W., Sumalee, A. & Hazelton, M., 2015. Estimation of mean and covariance of stochastic multi-class OD demands from classified traffic counts. *Transportation Research Part C,* Volume 59, p. 92–110.

Shi, P. & Harrison, R., 1997. Regression modelling of hourly NOx and NO2 concentrations in urban air in London. *Atmos. Environ. ,* 31(24), p. 4081–4094.

Shon, Z., Kim, K. & Song, S., 2011. Long-term trend in NO2 and NOx levels and their emission ratio in relation to road traffic activities in East Asia. *Atmospheric Environment,* Volume 45, pp. 3120-3131.

Silva, A. & Balassiano, R., 2011. *Global taxi schemes and their integration in sustainable urban transport systems.* Rio de Janeiro, Brazil, s.n.

Simon, L. & Kwanisai, T., 2003. *Applied Regression Analysis,* Pennsylvania, USA: Penn State University. Department of Statistics.

Small, K. & Kazimi, C., 1995. On the Costs of Air Pollution from Motor Vehicles. *Journal of Transportation and Economics Policy,* Volume 29, pp. 7-32.

Smit, R., Ntziachristos, L. & Boulter, P., 2010. Validation of road vehicle and traffic emission models – A review and meta-analysis. *Atmospheric Environment,* Volume 44, pp. 2943-2953.

Soegaard, H. & Møller-Jensen, L., 2003. Towards a spatial CO2 budget of a metropolitan region based on textural image classification and flux measurements. *Remote Sensing of Environment,* Volume 87, p. 283 – 294.

Sohn, T. et al., 2006. *Mobility Detection Using Everyday GSM Traces.* s.l., s.n., p. 212–224.

Song, C., Koren, T., Wang, P. & Barabási, A., 2010b. Modeling the scaling properties of human mobility. *Nature Physics,* Volume 6, pp. 818-823.

Song, C., Qu, Z., Blumm, N. & Barabási, A., 2010a. Limits of Predictability in Human Mobility. *Science,* 327(5968), pp. 1018-1021.

Song, C., Qu, Z., Blumm, N. & Barabási, A., 2010. Limits of Predictability in Human Mobility. *Science,* 327(5968), pp. 1018-1021.

Song, L. et al., 2008. Analysis of Taxi Operation Characteristics with Traffic Control. *Journal of Transportation Systems Engineering and Information Technology,* 8(6), pp. 127-131.

Steenbruggen, J., Borzacchiello, M., Nijkamp, P. & Scholten, H., 2013a. Data from telecommunication networks for incident management: An exploratory review on transport safety and security. *Transport Policy,* Volume 28, p. 86–102.

Steenbruggen, J., Borzacchiello, M., Nijkamp, P. & Scholten, H., 2013a. Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal,* 78(2), pp. 223-243.

Steenbruggen, J., Borzacchiello, M., Nijkamp, P. & Scholten, H., 2013b. Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal,* 78(2), p. 223–243.

Steenbruggen, J., Tranos, E. & Nijkamp, P., 2015. Data from mobile phone operators: A tool for smarter cities?. *Telecommunications Policy,* Volume 39, p. 335–346.

Stein, A., Isakov, V., Godowitch, J. & Draxler, R., 2007. A hybrid modeling approach to resolve pollutant concentrations in an urban area. *Atmospheric Environment,* Volume 41, p. 9410–9426.

Stenneth, L., Wolfson, O. & Yu, P. X. B., 2011. *Transportation Mode Detection using Mobile Phones and GIS Information.* Chicago, IL, USA, ACM, pp. 54-63.

Stover, V. & McCormack, E., 2012. The Impact of Weather on Bus Ridership in Pierce County, Washington. *Journal Of Public Transportation,* 15(1), p. 16.

Su, J. et al., 2009. Predicting traffic-related air pollution in Los Angeles using a distance decay regression selection strategy. *Environmental Research,* Volume 109, p. 657–670.

Sun, G., Wang, J. & Hu, X., 2013. A Bi-level Programming Model and Solution Algorithms for Taxi Fare in Taxi Market of China. *Journal of Information & Computational Science,* 10(17), p. 5787–5803.

Sun, L. et al., 2012. Real Time Anomalous Trajectory Detection and Analysis. *Journal Mobile Networks and Applications,* 18(3), pp. 341-356.

Takayama, T. et al., 2011. Waiting/cruising location recommendation for efficient taxi business. *International Journal of Systems Applications, Engineering & Development,* 5(2), pp. 224-236.

Tang, H., Kerber, M., Huang, Q. & Guibas, L., 2013. *Locating Lucrative Passengers for Taxicab Drivers.* Orlando, FL, USA, ACM.

Tang, J., Liu, F., Wang, Y. & Wang, H., 2015. Uncovering urban human mobility from large scale taxi GPS. *Physica A: Statistical Mechanics and its applications,* Volume 438, p. 140–153.

Tao, C., 2007. *Dynamic taxi-sharing service using intelligent transportation system technologies.* Shanghai, IEEE, pp. 3209 - 3212 .

Tashiro, Y. & Taniyama, T., 2002. Atmospheric NO2 and CO concentration in Lima, Peru. *Environment International,* Volume 28, p. 227–233.

Tettamanti, T., Demeter, H. & Varga, I., 2012. Route Choice Estimation Based on Cellular Signaling Data. *Acta Polytechnica Hungarica,* 9(4), pp. 207-220.

Tettamanti, T. & Varga, I., 2014. Mobile Phone Location Area Based Traffic Flow Estimation in Urban Road Traffic. *Advances in Civil and Environmental Engineering,* 1(1), pp. 1-15.

Thajchayapong, S., Pattara-atikom, W., Chadil, N. & Mitrpant, C., 2006. *Enhanced Detection of Road Traffic Congestion Areas using Cell Dwell Times.* Toronto, Canada, IEEE.

Thiessenhusen, K., Schäfer, R. & Lang, T., 2003. *Traffic data from cell phones: a comparison with loops and probe vehicle,* Berlin, Germany: Institute of Transport Research German Aerospace Center.

Thompson, M. & Bae, H., 2014. *A Functional Thinking Approachto the Design of Future Transportation Systems: Taxis as a Proxy for Personal Rapid Transit in South Korea.* Denmark, s.n., pp. 47-63.

Traag, V. A., Browet, A., Calabrese, F. & Morlot, F., 2011. *Social Event Detection in Massive Mobile Phone Data Using Probabilistic Location Inference.* s.l., IEEE.

Trasart, T. et al., 2015. Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. *Telecommunications Policy,* Volume 39, p. 347–362.

United Nations, 2008. *World Urbanization Prospects: The 2008 Revision,* United Nations: United Nations publication - Department of Economic and Social Affairs (DESA).

United Nations, 2012. *Population Distribution, Urbanization, Internal Migration and Development: An International Perspective,* New York: United Nations publication - Department of Economic and Social Affairs (DESA).

Uno, I., Ohara, T. & Wakamatsu, S., 1996. Analysis of wintertime NO2 pollution in the Tokyo metropolitan area. *Atmospheric Environment,* 30(5), pp. 703-713.

Vaccari, A. et al., 2009. *A holistic framework for the study of urban traces and the profiling of urban processes and dynamics.* St. Louis, MO, USA, IEEE, pp. 273-278.

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

Valerio, D., Alconzo, A., Ricciato, F. & Wiedermann, W., 2009a. *Exploiting Cellular Networks for Road Traffic Estimation: A Survey and a Research Roadmap.* Barcelona, IEEE, pp. 1-5.

Valerio, D. et al., 2009b. *Road traffic estimation from cellular network monitoring: a hands-on investigation.* Tokyo, IEEE, pp. 3035-3039.

Vann, B., 2011. Intensive Study of Ambient Carbon Dioxide Variability in Urban Atlanta. *Geosciences Theses.*

Velasco, E. et al., 2005. Measurements of CO2 fluxes from the Mexico City urban landscape. *Journal of Atmospheric Environment,* Volume 39, pp. 7433-7446.

Velasco, E. & Roth, M., 2010. Cities as Net Sources of CO2: Review of Atmospheric CO2 Exchange in Urban Environments Measured by Eddy Covariance Technique. *Geography Compass,* 4(9), p. 1238–1259.

Veloso, M., 2007. *Localização no Interior de Edificios Recorrendo a Assinaturas GSM,* Coimbra: Universidade de Coimbra, MSc Thesis.

Veloso, M., Bento, C. & Pereira, F., 2009. *Multi-Sensor Data Fusion on Intelligent Transport Systems.* s.l., MIT Portugal, Transportation Systems, Working Paper Series.

Veloso, M., Phithakkitnukoon, S. & Bento, 2011b. *Sensing Urban Mobility with Taxi Flow.* Chicago, Illinois, USA, ACM Digital Library.

Veloso, M., Phithakkitnukoon, S. & Bento, 2012. *Exploring the Relationship between Mobile Phone Call Intensity and Taxi Volume in Urban Area.* Anchorage, Alaska, IEEE.

Veloso, M., Phithakkitnukoon, S. & Bento, 2013. *Exploring Relationship Between Taxi Volume and Flue Gases' Concentrations.* Zurich, Switzerland, ACM.

Veloso, M., Phithakkitnukoon, S. & Bento, C., 2011c. *Urban Mobility Study using Taxi Traces.* Beijing, China, ACM Digital Library and UbiComp Extended Proceedings.

Veloso, M., Phithakkitnukoon, S. & Bento, C., 2015. Monitoring Urban Flue Gas Concentration Using Taxi Location and Weather Information. *Transactions on Intelligent Public Transportation.*

Veloso, M., Phithakkitnukoon, S., Bento, C. & Olivier, P., 2011a. *Exploratory Study of Urban Flow using Taxi Traces.* San Francisco, California, USA, s.n.

Veloso, M. P. S. & C., B., 2016a. Towards Recommendation System for Taxi Drivers. *Journal of Urban Technology.*

Venkatram, A., Isakov, V., Seila, R. & Baldauf, R., 2009. Modeling the impacts of traffic emissions on air toxics concentrations near roadways. *Atmospheric Environment,* Volume 43, p. 3191–3199.

Venkatram, A., Isakov, V., Thoma, E. & Baldauf, R., 2007. Analysis of air quality data near roadways using a dispersion model. *Atmospheric Environment,* Volume 41, p. 9481–9497.

Venkatram, A., Snyder, M. & Isakov, V., 2013b. Modeling the impact of roadway emissions in light wind, stable and transition conditions. *Transportation Research - Part D,* Volume 24, p. 110–119.

Venkatram, A., Snyder, M., Isakov, V. & Kimbrough, S., 2013a. Impact of wind direction on near-road pollutant concentrations. *Atmospheric Environment,* Volume 80, pp. 248-258.

Vesala, T. et al., 2008. Surface–atmosphere interactions over complex urban terrain in Helsinki, Finland. *Tellus,* Volume 60B, p. 188–199.

Virtanen, J., 2012. *Mobile phones as probes in travel time monitoring,* Helsinki, Finland: Finnish Road Administration.

Vogt, R. et al., 2006. Temporal dynamics of CO2fluxes and profiles over a Central European city. *Theor. Appl. Climatol.,* Volume 84, p. 117–126.

Wang, C., Chen, H. & Ng, W., 2012. *From Data to Knowledge to Action: A Taxi Business Intelligence System.* Singapore, s.n., pp. 1623-1628.

Wang, H., Zou, H., Yue, Y. & Li, Q., 2009a. *Visualizing Hot Spot Analysis Result Based on Mashup.* Seattle, WA, USA, ACM, p. 45–48.

Wang, L., Hu, K., Ku, T. & Wu, J., 2014. Urban Mobility Dynamics Based on Flexible Discrete Region Partition. *International Journal of Distributed Sensor Networks,* Volume 2014, p. 10.

Wang, M., Mulinazzi, T. & Schrock, S., 2009b. *Feasibility of Using Cellular Telephone Data to Determine the Truckshed of Intermodal Facilities.* Ames, Iowa, USA, s.n., p. 13.

Wang, M., Schrock, S., Broek, N. & Mulinazzi, T., 2012. *The Use of Cell Phone Network Data in Traffic Data Collection and Long-Haul Truckshed (Geographic Extent) Tracking,* Lincoln, NE, USA: Mid-America Transportation Center.

Wang, R. et al., 2013. Temporal stability of land use regression models for traffic-related air pollution. *Atmospheric Environment,* Volume 64, pp. 312-319.

Wang, Y. et al., 2015. *Regularity and Conformity: Location Prediction Using Heterogeneous Mobility Data.* Sydney, ACM.

Westerdahl, D. et al., 2005. Mobile platform measurements of ultrafine particles and associated pollutant concentrations on freeways and residential streets in Los Angeles. *Atmospheric Environment,* Volume 39, p. 3597–3610.

Westerdahl, D., Wang, X., Pan, X. & Zhang, K., 2009. Characterization of on-road vehicle emission factors and microenvironmental air quality in Beijing, China. *Atmospheric Environment,* Volume 43, p. 697–705.

White, J., Quick, J. & Philippou, P., 2004. *The use of mobile phone location data for traffic information.* s.l., IEEE, pp. 321-325 .

White, J. & Wells, I., 2002. *Extracting origin destination information from mobile phone data.* London, UK, s.n., pp. 30-24.

Wideberg, J., Caceres, N. & Benitez, F., 2006. *Deriving Traffic Data from a Cellular Network.* s.l., s.n., p. 8.

Witayangkurn, A. et al., 2013. *Trip reconstruction and transportation mode extraction on low data rate gps data from mobile phone.* s.l., s.n., p. 1–19.

Witten, I. & Frank, E., 2005. *Data Mining - Pratical Machine Learning Tools and Techniques.* 2nd ed. San Francisco(CA): Morgan Kaufman.

Wong, K. & Bell, M., 2006. The optimal dispatching of taxis under congestion: a rolling horizon approach. *Journal of Advanced Transportation.*

Wong, K., Wong, S. & Yang, H., 2001. Modeling urban taxi services in congested road networks with elastic demand. *Elsevier Transportation Research Part B,* Volume 35, pp. 819-842.

Wong, K., Wong, S., Yang, H. & Tong, C., 2003. The Effect of Perceived Profitability on the Level of Taxi Service in Remote Areas. *Journal of the Eastern Asia Society for Transportation Studies,* Volume 5, pp. 79-94.

Wong, K., Wong, S., Yang, H. & Wu, J., 2008. Modeling urban taxi services with multiple user classes and vehicle modes. *Transportation Research Part B,* Volume 42, p. 985–1007.

Wong, R., Szeto, W. & Wong, S., 2015. A Two-Stage Approach to Modeling Vacant Taxi Movements. *Transportation Research Procedia,* Volume 7, p. 254 – 275.

Wunnava, S. et al., 2007. *Travel Time Estimation USing Cell Phones for Highways and Roadways,* Miami, Florida, USA: Florida Department of Transportation.

Wu, W., Siong, W., Krishnaswamy, S. & Sinha, A., 2012. *To Taxi or Not to Taxi? - Enabling Personalised and Real-Time Transportation Decisions for Mobile Users.* Bengaluru, Karnataka, IEEE, pp. 320-323.

Wu, X. & Lee, D., 2013. An Integrated Taxi Dispatching Strategy Handling both Current and Advance Bookings. *Eastern Asia Society for Transportation Studies,* Volume 9.

Xiang, W., 2013. *The Modelling of State of the Art Taxi Operations and Dispatching Approaches,* Singapore: National University of Singapore.

Xiao, Y. et al., 2012. *Transportation Activity Analysis Using Smartphones.* Washington, DC, USA, s.n.

Xu, J. & Huang, Z., 2009. An Intelligent Model for Urban Demand-responsive Transport System Control. *Journal of Software,* Volume 4, pp. 766-776.

Yamamoto, K., Uesugi, K. & Watanabe, T., 2008. *Adaptive Routing of Cruising Taxis by Mutual Exchange of Pathways.* s.l., s.n., p. 559–566.

Yang, H., Fung, C., Wong, K. & Wong, S., 2010b. Nonlinear pricing of taxi services. *Transportation Research Part A,* Volume 44, p. 337–348.

Yang, H., Lau, Y., Wong, S. & Lo, H., 2000. A macroscopic taxi model for passenger demand, taxi utilization and level of services. *Transportation,* 27(3), pp. 317-340.

Yang, H., Leung, C., Wong, S. & Bell, M., 2010. Equilibria of bilateral taxi–customer searching and meeting on networks. *Elsevier Transportation Research Part B: Methodological,* Volume 44, p. 1067–1083.

Yang, H. & Wong, S., 1998. A network model of urban taxi services. *Elsevier Transportation Research Part B,* Volume 34, pp. 235-246.

Yang, H., Wong, S. & Wong, K., 2002. Demand–supply equilibrium of taxi services in a network under competition and regulation. *Elsevier Transportation Research Part B,* Volume 36, p. 799–819.

Yang, H., Ye, M., Tang, W. & Wong, S., 2005. Regulating taxi services in the presence of congestion externality. *Elsevier Transportation Research Part A,* Volume 39, p. 17–40.

Yang, Q. et al., 2015. *Taxi Operation Optimization Based on Big Traffic Data.* s.l., s.n.

Yang, T., Yang, H. & Wong, S., 2009. *Modeling Taxi Services with a Bilateral Taxi-Customer Searching and Meeting Function.* s.l., s.n.

Yao, B. et al., 2015. Fleet size and fare optimization for taxi under dynamic demand. *Journal of Transport Literature,* 10(4), pp. 45-49.

Yao, Z. & Cheng, G., 2012. Clustering Taxi Quantity Regulatory Levels of Chinese Capital Cities. *Journal of Transportation Systems Engineering and Information Technology,* 12(5).

Ygnace, J., 2001. *Travel timespeed estimates on the French Rhone Comdor Network using cellular phones as probes,* Lyon, France: INRETS.

Ygnace, J., Drane, C., Yim, Y. & Lacvivier, R., 2000. *Travel Time Estimation on the San Francisco Bay Area Network Using Cellular Phones as Probes,* s.l.: California Partners for Advanced Transit and Highways (PATH).

Yim, Y. & Cayford, R., 2001. *Investigation of Vehicles as Probes Using Global Positioning System and Cellular Phone Tracking: Field Operational Test,* Berkeley, CA, USA: California PATH Program, University of California.

Yli-Tuomi, T. et al., 2005. Emissions of fine particles, NOx, and CO from on-road vehicles in Finland. *Atmospheric Environment,* Volume 39, p. 6696–6706.

Yuan, J., Zheng, Y. & Xie, X., 2012b. *Discovering Regions of Different Functions in a City Using Human Mobility and POIs.* s.l., ACM.

Yuan, J., Zheng, Y., Xie, X. & Sun, G., 2011a. *Driving with Knowledge from the Physical World.* San Diego, California, USA, s.n.

Yuan, J., Zheng, Y., Xie, X. & Sun, G., 2013. T-Drive: Enhancing driving directions with taxi drivers' intelligence. *Transactions on Knowledge and Data Engineering,* 25(1), pp. 220-232.

Yuan, J. et al., 2010. *T-Drive: Driving Directions Based on Taxi Trajectories.* s.l., ACM Association for Computing Machinery, pp. 99-108.

Yuan, J., Zheng, Y., Zhang, C. & Xie, X., 2009. *An interactive-voting based map matching algorithm.* s.l., IEEE, pp. 43-52.

Yuan, J. et al., 2011b. *Where to Find My Next Passenger?.* China, s.n.

Yuan, J., Zheng, Y., Zhang, L. & Xie, X., 2012a. *T-Finder: A Recommender System for Finding Passengers and Vacant Taxis.* Beijing, China, ACM 13th International Conference on Ubiquitous computing (UbiComp'11).

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,* Volume 36, p. 118–130.

Yue, Y., Wang, H., Hu, B. & Li, L., 2011. *Identifying Shopping Center Attractiveness Using Taxi Trajectory Data.* Beijing, China, ACM TDMA'11.

Yue, Y. et al., 2012. Exploratory calibration of a spatial interaction model using taxi GPS trajectories. *Computers, Environment and Urban Systems,* Volume 36, p. 140–153.

Yue, Y., Zhuang, Y., Li, Q. & Mao, Q., 2009. *Mining Time-dependent Attractive Areas and Movement Patterns from Taxi Trajectory Data.* s.l., s.n.

Zavala, M. et al., 2006. Characterization of on-road vehicle emissions in the Mexico City Metropolitan Area using a mobile laboratory in chase and fleet average measurement modes during the MCMA-2003 field campaign. *Journal of Atmospheric Chemistry and Physics,* Volume 6, pp. 5129-5142.

Zavala, M. et al., 2008. Comparison of emission ratios from on-road sources using a mobile laboratory under various driving and operational sampling modes. *Atmos. Chem. Phys. Discuss.,* Volume 8, p. 8059–8090.

Zhang, D. & He, T., 2012. *pCruise: Reducing Cruising Miles for Taxicab Networks.* s.l., IEEE 33rd Real-Time Systems Symposium.

Zhang, D. et al., 2011a. *iBAT: Detecting Anomalous Taxi Trajectories from GPS Traces.* Beijing, China, ACM UbiComp'11.

Zhang, D., Peng, Z. & Sun, D., 2014b. *A Comprehensive Taxi Assessment Index Using Floating Car Data.* Washington, D.C., s.n.

Zhang, D. et al., 2014a. *Understanding Taxi Service Strategies from Taxi GPS Traces.* s.l., s.n.

Zhang, H., 2004. *The optimality of naive Bayes.* s.l., AAAI Press, p. 562–567.

Zhang, W., Li, S. & Pan, G., 2012a. *Mining the Semantics of Origin-Destination Flows using Taxi Traces.* Pittsburgh, USA, ACM UbiComp'12.

Zhang, W., Xu, J. & Wang, H., 2007. Urban Traffic Situation Calculation Methods Based on Probe Vehicle Data. *Journal of Transportation SystemsEngineering and Information Technology,* 7(1), p. 43−49.

Zhang, Y., 2014d. *How Do Taxis Work in Beijing? An Exploratory Study of Spatio-Temporal Taxi Travel Pattern Using GPS Data,* Los Angeles, CA, USA: University of California, MSc Thesis.

Zhan, X., Hasan, S., Ukkusuri, S. & Kamga, C., 2013. Urban link travel time estimation using large-scale taxi data with partial information. *Transportation Research Part C,* Volume 33, p. 37–49.

Zhan, X., Qian, X. & Ukkusuri, S., 2014c. *Measuring the Efficiency of Urban Taxi Service System.* New York, NY, USA, s.n.

Zhao, F. et al., 2015b. Stop Detection in Smartphone-based Travel Surveys. *Transportation Research Procedia,* Volume 11, p. 218–226.

Zhao, F., Pereira, F., Ball, R. & Kim, Y., 2015a. *Exploratory Analysis of a Smartphone-Based Travel Survey in Singapore.* s.l., Transportation Research Board of the National Academies.

Zheng, F. & Zuylen, H., 2012. Urban link travel time estimation based on sparse probe vehicle data. *Transportation Research - Part C,* p. 14.

Zheng, J., Liu, S. & Ni, L., 2013. *Effective Routine Behavior Pattern Discovery from Sparse Mobile Phone Data via Collaborative Filtering.* San Diego, IEEE, pp. 29-37.

Zheng, X., Liang, X. & Xu, K., 2012b. *Where to wait for a taxi?.* s.l., ACM, p. 149–156.

Zheng, Y., Liu, Y., Yuan, J. & Xie, X., 2011b. *Urban Computing with Taxicabs.* China, s.n.

Zheng, Y. et al., 2010. *Drive Smartly as a Taxi Driver.* s.l., s.n., pp. 484-486.

Zheng, Y., Zhang, L., Xie, X. & Ma, W., 2009. *Mining Interesting Locations and Travel Sequences from GPS Trajectories.* Madrid, Spain, ACM International World Wide Web Conference Committee (IW3C2).

Zheng, Z., Rasouli, S. & Timmermans, H., 2014. Evaluating the accuracy of GPS-based taxi trajectory records. *Procedia Environmental Sciences - Design and Decision Support Systems in Architecture and Urban Planning,* Volume 22, p. 186 – 198.

Zhou, X. et al., 2012. *The Predictability of Cellular Networks Traffic.* Gold Coast, QLD, IEEE, pp. 973-978 .

Zhou, Y. et al., 2015. Functionally critical locations in an urban transportation network: Identification and space–time analysis using taxi trajectories. *Computers, Environment and Urban Systems,* Volume 52, p. 34–47.

Zhou, Y. & Levy, J., 2007. Factors influencing the spatial extent of mobile source air pollution impacts: a meta-analysis. *BMC Public Health,* 7(89), p. 11.

Zhuang, L., Gong, J., He, Z. & Xu, F., 2012. *Framework of Experienced Route Planning.* Anchorage, Alaska, USA, 15th International IEEE Conference on Intelligent Transportation Systems.

Zhu, J., Shuai, B., Z., H. & Sun, C., 2013. The Optimal Taxi Fleet Size Structure under Various Market Regimes When Charging Taxis with Link-Based Toll. *Journal of Applied Mathematics,* p. 11.

Zhu, Y. et al., 2009. *Trajectory Enabled Service Support Platform for Mobile Users' Behavior Pattern Mining.* Toronto, Canada, s.n., pp. 1-10.

Ziebart, B., Maas, A., Dey, A. & Bagnell, J., 2008. *Navigate like a cabbie: probabilistic reasoning from observed context-aware behavior.* New York, NY, USA, ACM, pp. 322-331.

Zou, X. et al., 2006. Shifted power-law relationship between NO2 concentration and the distance from a highway: A new dispersion model based on the wind profile model. *Atmospheric Environment,* Volume 40, p. 8068–8073.

Zwack, L., Hanna, S., Spengler, J. & Levy, J., 2011. Using advanced dispersion models and mobile monitoring to characterize spatial patterns of ultrafine particles in an urban area. *Atmospheric Environment,* Volume 45, pp. 4822-4829.

Appendix

Throughout the study, several tools were used to assist the analysis:

- PostgreSQL (with PostGIS module)[21], a Database Management System (DBMS) to store data using an Entity-Relationship model;

- Quantum GIS[22], a Geographic Information System (GIS) used to visualize spatial data models;

- Waikato Environment for Knowledge Analysis (WEKA)[23], a data mining environment, used to clean, transform, and mine data;

- Konstanz Information Miner (KNIME)[24], a data mining environment, used to clean, transform and mine data;

- MathWorks Matrix Laboratory (MATLAB)[25], a numerical computing environment, used for data mining and data visualization;

- Minitab Statistical Software[26], a statistical environment, used for statistical analysis and data mining;

- IBM Statistical Package for the Social Sciences (SPSS)[27], a statistical environment, used for statistical analysis and data mining;

- Programming Language Java[28] with IDE Eclipse, used to program any additional features needed to clean, transform, or analyze the data.

[21] PostgreSQL. http://www.postgresql.org/ .

[22] Quantum GIS. http://www.qgis.org/ .

[23] Waikato Environment for Knowledge Analysis. http://www.cs.waikato.ac.nz/ml/weka/ .

[24] Konstanz Information Miner. http://www.knime.org/ .

[25] MathWorks Matrix Laboratory. http://www.mathworks.com/products/matlab/ .

[26] Minitab Statistical Software. https://www.minitab.com/ .

[27] Statistical Package for the Social Sciences. http://www-01.ibm.com/software/analytics/spss/ .

[28] Programming Language Java. http://www.java.com/ .