Technical University of Denmark

PhD Thesis

Detection, analysis and prediction of traffic anomalies due to special events

Author:
Ioulia Markou

Supervisor:
Prof. Francisco C. Pereira
Asst. Prof. Filipe Rodrigues

A thesis submitted in fulfillment of the requirements for the degree of PhD

in the

Transport Modelling
Department or Management Engineering

May 30, 2019

Technical University of Denmark
Department of Management Engineering
Produktionstorvet, Building 426
2800 Kongens Lyngby, Denmark
Phone +45 4525 4800
www.man.dtu.dk

Declaration of Authorship

I, Ioulia Markou, declare that this thesis titled, “Detection, analysis and prediction of traffic anomalies
due to special events” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ____

Date: ____

"*Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.* "

Marie Curie

Summary (English)

The transport system consists of complex microscopic and macroscopic interactions that affect our transport choices, spatial planning, economic activities, safety, CO2 emissions and much more. This thesis focuses on unexpected and unwanted demand fluctuations that we often observe in the network and lead to system failures and cost implications. Significantly low speeds or excessively low flows at an unusual time are only some of the phenomena that may confuse a driver or transport authorities, since they are totally unexpected and frequently there is no obvious explanation for them. The term “anomalies” refers to those non-conforming patterns which appear into a well-defined notion of normal behavior. In the literature, similar phenomena can be described as outliers, exceptions or discordant observations.

With the knowledge of the need to understand traffic anomalies and eventually predict them after the localization of historical records spatially and temporally, we start by introducing a methodology that identifies traffic anomalies on traffic networks and correlates them with special events using internet data. We investigate why traffic congestion was occurring as well as why demand fluctuated on days when there were no apparent reasons for such phenomena. The system is evaluated by using Google’s public data set for taxi trips in New York City. A “normality” baseline is defined at the outset and then used in the subsequent study of the demand patterns of individual days to detect outliers. With the use of this approach it is possible to detect fluctuations in demand and to analyze and correlate them with disruptive event scenarios such as extreme weather conditions, public holidays, religious festivities, and parades. Kernel density analysis is used so that the affected areas, as well as the significance of the observed differences compared with the average day, can be depicted.

The search for possible explanations for the observed anomalies in the road network has highlighted the huge amount of information that is available on the internet, while stressing the difficulty of retrieving documents that are highly correlated with examined events’ details (location, time of the day, etc.). In the above contexts, we develop a framework that predicts transport demand with a supervised topic modeling algorithm by utilizing information about social events retrieved using various strategies, which made use of search aggregation, natural language processing, and query expansion. It is found that a two-step process produced the highest accuracy for transport demand prediction, where different (but related) queries are used to retrieve an initial set of documents, and then, based on these documents, a final query is constructed that obtains the set of predictive documents. These are then used to model the most discriminating topics related to the transport demand.

Having verified that the Internet can give a further insight into demand hotspots’ prediction, we explore time-series data and semantic information combinations using machine learning and deep learning techniques in the context of creating a prediction model that is able to capture in real-time future stressful situations of the studied transportation system. We apply the proposed approaches in event areas in New York using publicly available taxi data. We empirically show that the proposed models are able to significantly reduce the error in the forecasts. The importance of semantic information is highlighted in all presented methods.

In addition to the investigation of which types of data can positively contribute to the accuracy of our forecasts, the structure of the model that can perform better using the available information has also been studied more extensively. We mainly focus on the analysis, evaluation, and forecasting of prediction model's residuals in a real-time taxi demand forecasting framework. We comprise a deep learning architecture that is based on Fully-Connected dense layers. The analysis focuses on areas where significant fluctuations in demand are observed, due to popular venues located in the area. The performance of our proposed two-stage process with the inclusion of residuals' forecasts, is improved considerably.

Overall, the models proposed in this thesis highlight the value of data fusion of text and time series data, as well as the capabilities of information retrieval using query expansion methods. They can be of great value to a broad range of traffic incidents' management frameworks.

Summary (Danish)

Transportsystemet består af komplekse mikroskopiske og makroskopiske interaktioner, som påvirker vores transportvalg, fysisk planlægning, økonomiske aktiviteter, sikkerhed, CO2-emissioner og meget mere. Denne afhandling fokuserer på uventede og uønskede efterspørgselsmønstre, som vi dagligt observerer i netværket og som fører til systemfejl og omkostningsimplikationer. Signifikante lave hastigheder eller for lave volumirer på en usædvanlig tid er kun nogle af de fænomener, der kan forvirre en chauffør eller transportmyndigheder, fordi de er helt uventede, og ofte er der ingen åbenbar forklaring på dem. Udtrykket "anomalier" refererer til de ikke-overensstemmende mønstre, der fremgår af en veldefineret forestilling om normal adfærd. I litteraturen kan lignende fænomener betegnes som ekstreme observationer, undtagelser eller uoverensstemmende observationer.

Med viden om behovet for at forstå trafikanomalier og til sidst forudsige dem efter lokaliseringen af historiske registreringer rumligt og tidligt, begynder vi ved at indføre en metode, der identificerer trafikanomalier på trafiknet og korrelerer dem med særlige begivenheder ved brug af internetdata. Vi undersøger, hvorfor trafikbelastning forekom, samt hvorfor efterspørgslen svingede på dage, da der ikke var nogen åbenbare grunde til sådanne fænomener. Systemet evalueres ved at bruge Googles offentlige datasæt til taxiture i New York City. En "normalitet" basis scenarie er defineret i begyndelsen og derefter brugt i den efterfølgende undersøgelse for efterspørgselsmønstre af individuelle dage for at detektere ekstreme observationer. Ved hjælp af denne tilgang er det muligt at påvise fluktuationer i efterspørgslen og analysere og korrelere dem med forstyrrende begivenhedsscenarier som ekstreme vejrforhold, helligdage, religiøse festivaler og parader. Kerneldensitetsanalyse anvendes, således at de berørte områder samt betydningen af de observerede forskelle i forhold til den gennemsnitlige dag kan afbildes.

Søgningen efter mulige forklaringer til de observerede uregelmæssigheder i vejnetværket har fremhævet den enorme mængde information, der er tilgængelig på internettet, samtidig med at det er svært at hente dokumenter, der er stærkt korrelerede med undersøgte begivenheds detaljer (placering, tidspunkt på dagen , etc.). I de ovenstående sammenhænge udvikler vi en ramme, der forudsiger transportefterspørgslen med en overvåget "topic modeling“ algoritme ved at udnytte information om sociale begivenheder hentet ved hjælp af forskellige strategier, der benyttede søgning aggregation, naturlig sprogbehandling og forespørgselsudvidelse. Det er konstateret, at en to-trins proces producerede den højeste nøjagtighed for forudsigelse af transportefterspørgsel, hvor forskellige (men beslægtede) forespørgsler bruges til at hente et indledende sæt dokumenter, og derefter er der baseret på disse dokumenter en endelig forespørgsel konstrueret, der opnår sæt af prædiktive dokumenter. Disse bruges derefter til at modellere de mest diskriminerende emner i forbindelse med transportbehovet.

Efter at have kontrolleret, at internettet kan give et yderligere indblik i efterspørgsel hotspots forudsigelse, undersøger vi tidsserie data og semantiske information kombinationer ved hjælp af maskinindlæring og dybe indlæringsteknikker i sammenhæng med at skabe en forudsigelsesmodel, der er i stand til at fange i realtid fremtidige

stressfulde situationer i det studerede transportsystem. Vi anvender de foreslåede tilgange i begivenhedsområder i New York ved hjælp af offentligt tilgængelige taxa data. Vi viser empirisk, at de foreslåede modeller kan reducere fejlen i prognoserne betydeligt. Betydningen af semantiske oplysninger fremhæves i alle fremlagte metoder.

Foruden undersøgelsen af hvilke typer data der kan bidrage positivt til nøjagtigheden af vores prognoser, er strukturen af modellen, der kan udføre bedre ved hjælp af de tilgængelige oplysninger, også blevet undersøgt mere omfattende. Vi fokuserer hovedsagelig på analyse, evaluering og prognose af forudsigelsesmodellets residuals i en real-time tax demand forecast framework. Vi består af en dyb læringsarkitektur, der er baseret på fuldt forbundne tætte lag. Analysen fokuserer på områder, hvor der observeres betydelige udsving i efterspørgslen på grund af populære lokaliteter i området. Udførelsen af vores foreslåede to-trins proces med inddragelse af residuals prognoser forbedres betydeligt.

Samlet set fremhæver de modeller, der foreslås i denne afhandling, værdien af datafusion af tekst- og tidsseriedata samt mulighederne for informationsindhentning ved hjælp af forespørgselsudvidelsesmetoder. De kan være af stor værdi for en bred vifte af trafikhændelseres ledelsesrammer.

Acknowledgements

I would first like to thank my supervisor Francisco Pereira for his advice and guidance. I have learned a lot from him in many fronts, extending well beyond academics. I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly.

I would also like to thank my fellow researchers at the Machine Learning for Smart Mobility (MLSM) Group. Filipe has been a huge support in my research path, a great office-mate and a good friend. I wish all the best to my friend Meghdad; we started and finished together this great journey, and his support was invaluable.

I am also grateful to Costas Antoniou for hosting me at Transportation Systems Engineering (TSE) chair of the Technical University of Munich during my external stay. It was a pleasure to work and exchange ideas with everyone there.

I would like to thank my family, for their endless support and constant encouragement, especially my parents Babi and Panagiota, my sister Constantina and my boyfriend Costas.

Finally, the research that led to this thesis would not have been possible without the funding and support provided by Technical University of Denmark and Otto Mønsteds Foundation.

Contents

List of Figures

List of Tables

List of Abbreviations

API	**A**application **P**rogramming **I**nterface
AVs	**A**utonomous **V**ehicles
ARIMA	**A**uto**R**egressive **I**ntegrated **M**oving **A**verage
BOW	**B**ag-**O**f-**W**ords
DL	**D**eep **L**earning
EM	**E**xpectation-**M**aximization
FC	**F**ully-**C**onnected
GPs	**G**aussian **P**rocesses
GPS	**G**lobal **P**ositioning **S**ystem
KDE	**K**ernel **D**ensity **E**stimation
LDA	**L**atent **D**irichlet **A**llocation
LR	**L**inear **R**egression
LSTM	**L**ong **S**hort-**T**erm **M**emory neural (LSTM)
MAE	**M**ean **A**bsolute **E**rror
MedLDA	**M**aximum **e**ntropy **d**iscrimination
MLP	**M**ulti-**L**ayer **P**erceptron
NFC	**N**ear **F**ield **C**ommunication
POI	**P**oint **O**f **I**nterest
PCA	**P**rincipal **C**omponents **A**nalysis
RAKE	**R**apid **A**utomatic **K**eyword **E**xtraction
RBF	**R**adial **B**asis **F**unctions
RFID	**R**adio-**F**requency **I**dentification
RMSE	**R**oot **M**ean **S**quare **E**rror
SMOTE	**S**ynthetic **M**inority **O**ver-sampling **TE**chnique
SVM	**S**upport **V**ector **M**achine
TF-IDF	**T**erm **F**requency-**I**nverse **D**ocument **F**requency
TLC	NYC **T**axi and **L**imousine **C**ommission
URL	**U**niform **R**esource **L**ocator

Dedicated to my family

Chapter 1

Introduction

1.1 Motivation

Transport and logistics sectors are an integral part of the economic well-being of communities, municipalities, regions, countries and continents. The transport system consists of complex microscopic and macroscopic interactions that affect our transport choices, spatial planning, economic activities, safety, CO2 emissions and much more. Therefore, a structured and well-founded management of such a system is extremely important.

Requirements for improved and better integrated land use and transportation models have emerged. Traffic congestion remains a major problem in urban areas, since it has a significant adverse economic impact through deterioration of mobility, safety and air quality. According to a recent study (INRIX Research, 2017), the cities most affected by congestion are capital cities, known centers of commerce and politics. They typically have the densest road networks combined with large populations of residents, commuters and visitors. In Europe, even cities with lower overall levels of congestion, such as Zurich and Munich, show significantly higher demand in their network during peak hours, which consequently affects the quality of life of their residents. In United States, the situation is no better. New York drivers spent the second-most hours in congestion in North America and the third most globally, behind Los Angeles and Moscow, sitting in traffic 91 hours last year on average. As a result, the importance of better management of the road network to efficiently utilize existing capacity is increasing.

In general, mobility trends captured in complex transport systems consist of two basic components: utilitarian travel that mostly includes habitual behavior (e.g. commuting to work, weekly shopping) but also to a minor extent non-habitual needs (e.g. go to hospital, occasional shopping); and recreational travel, which comprises the human need for entertainment, social interaction and public expression. Efficient and effective intelligent transport systems should be able to take into consideration both of these factors for accurate demand predictions and better traffic management.

In the last few years, we have at our disposal a huge amount of sensory data for many cities in the world. Technologies, such as GPS, RFID, NFC, WiFi and Bluetooth, allow us to fully record all the phenomena observed in a traffic network at a very high spatial and temporal resolution. These technologies have become ubiquitous - we can find their use in transit smartcards, toll collection systems, floating car data, fleet management systems, car counters, mobile phones, wearable devices, etc.. Subsequently, we are able to capture in real-time the “pulse” of the city that we are focused on each time (both habitual and recreational travel).

Many urban areas build and operate modern Traffic Management Centers (TMCs), which perform several functions, including collection and warehousing of real-time traffic data, and utilization of this data for various dynamic traffic control and route guidance applications. Ride sharing services, such us Uber and Lift, that use internet-based mobile technology to match passengers and drivers, are interested in the maximization of their capacity utilization, namely the fraction of time that drivers have a fare-paying passenger in the car (Cramer and Krueger, 2016). Within this goal, they collect a plethora of data related to trips demand and car trajectories. Finally, Autonomous vehicles (AVs) declare a dynamic presence in the future of our transportation system as a new technology that has the potential to impact vehicle safety, congestion, and travel behavior. Researchers have already started developing ways for AV technology to reduce congestion and fuel consumption. All these applications require traffic models that provide, in real-time, estimation and prediction of traffic conditions. The complexity of transportation systems often dictates the use of detailed prediction models that can take into consideration as many parameters of our mobility trends as possible.

Current prediction approaches generally focus on capturing recurrent conditions, namely their seasonal spatial-temporal aspects (the “average" winter peak-hour Monday, in area X, with weather Y). The developed approaches can be successful for long-term planning applications or for modeling demand in non-eventful areas such as residential neighborhoods. However, in lively and dynamic areas where multiple special events take place, such as music concerts, sports games, festivals, parades and protests, these approaches fail to accurately model mobility demand precisely at times when it is needed - when the transport system of the area is under stress. The inability of the system to meet the new demand conditions emphasizes the need of good anticipatory capabilities which are capable to accept timely information on such phenomena.

Non-recurrent special events, such as concerts, sport games and demonstrations, are planned and largely advertised on the Web. An interesting fact is that it is much more likely to have citizens sharing their expectations/experiences about non-recurrent events than to talk about their daily commute. This plethora of information makes the Web an important tool for demand prediction and thus system's balance maintenance. Within this crowd-sourced data lay explanations for many of the mobility patterns that we observe. However, the correlation of the detected phenomena with special events is not a trivial process as there are many dimensions involved. Details such as the type of an event, popularity of the event's protagonists, size of the venue, tickets' price, etc. play a significant role and the necessity of computational methodologies that can learn the interactions of all these parameters in the past, and use them to forecast similar situations in the future is unquestionable.

From a data fusion perspective, combining time-series data with textual information for better understanding real-world phenomena is a very important, yet challenging, problem. The key intuition is that the textual information could contain clues that correlate with the time-series observations and, at least to some extent, explain its behaviour. Given the generality of this cross-domain data fusion problem, it is not surprising that different solutions arise from multiple research areas.

Previous studies have shown a strong correlation between number of public transport arrivals with the structured data mined from the Web (Pereira, Rodrigues, and Ben-Akiva, 2015; Pereira et al., 2015). Namely, semi-structured information about events from announcements websites can be used as features for public transport arrivals. However, information contained on these websites is usually incomplete, noisy or missing, which makes it difficult to generalize. Going beyond this approach raises two new challenges: what web-pages are relevant (information retrieval) and how to turn relevant information into model attributes (information extraction).

While investigating which types of information could be significant to be considered in the formulation of an accurate demand prediction model, the consequences of areas' interaction in the supply - demand traffic network equilibrium has been also under review. In our everyday life, many have noticed that what is happening in our nearby neighborhood does not necessarily depend on events that take place within it. There are many external factors that affect phenomena observed in close proximity, that we have to take them into account when trying to shape a valid predictive model of future similar situations.

It is common for example to see congested conditions in roads that are not in the vicinity of a special events venue, but they are the main route options for drivers wishing to leave a scheduled event. Another phenomenon that has also been observed is the increase in traffic around central train stations from people that select them as their transition point to other means of transport, after a football game or a concert in a venue located on the outskirts of the city. City's road network works as the circulatory system in our body. Multi-lane superhighways bring traffic to exits into small-state highways, and then to smaller arterials, collectors, and local roads. Signs of discomfort on the limbs of our body are often signs of a malfunction in the cardiovascular system of our body. Respectively in a road network, when we observe severe congestion phenomena in a specific segment, their cause can be identified in previous network segments that canalize traffic to the studied segment and generally strongly interact with it.

From the early 1990s, Machine Learning has started to have a significant and widespread practical impact, with the development of many successful applications in various research domains ranging from autonomous vehicles to computer vision, speech recognition, and natural language processing. Developers of artificial intelligence (AI) systems have recognized that in many cases it can be far easier to train a system by showing it examples of desired input-output behavior than to program it manually by anticipating the desired response for all possible inputs (Jordan and Mitchell, 2015). The field of AI rapidly tackled and solved problems that are intellectually difficult for human beings but relatively straightforward for computers—problems that can be described by a list of formal, mathematical rules. A person's everyday life requires an immense amount of knowledge about the world. Much of this knowledge is subjective and intuitive, and therefore difficult to articulate in a formal way. Computers need to capture this same knowledge in order to behave in an intelligent way, therefore one of the key challenges in AI is how to get this informal knowledge into a computer.

In order to capture the effects of events that we are interested in, we should exploit the vast amount of information that is shared online about what is planned to take place in the city. The growing sizes of relevant modern datasets make imperative the use of various algorithms through machine learning libraries, and more specifically the subclass of supervised learning algorithms. With their tools, our developed models are able to learn a mapping from inputs x to outputs y, given a labeled set of input-output pairs. Time-series data with textual information can be combined to build supervised predictive models that are able to understand the mapping from inputs x to outputs y and subsequently provide us a better understanding of demand patterns when extreme events are observed.

Many artificial intelligence tasks can be solved by designing the right set of features to extract for that task, then providing these features to a simple machine learning algorithm. For example, a useful feature for taxi demand estimation around a venue where a concert of Celine Dion is organized, is the number of people that informed the official Facebook page of the event that they will attend it. It therefore gives a strong clue as to whether the event is popular and subsequently whether there will be future taxi customers after its completion.

However, for this task and many other tasks, it is difficult to know what other features should be extracted. We don't know how significant some variables are in a problem where multiple factors are involved. One solution to this problem is to use machine learning to discover not only the mapping from representation to output but also the representation itself (Goodfellow et al., 2016). Deep learning, on the other hand, solves this central problem in representation learning by introducing representations that are expressed in terms of other, simpler representations. It allows the computer to build complex concepts out of simpler concepts thus achieving better forecasting results in a variety of applications compared to traditional approaches. In this thesis we explore several machine learning techniques to develop a reliable predictive transport demand model. A wealth of information from various sources is exploited while trying to identify the parameters that are correlated with our target variable, namely travel demand.

1.2 Contributions

This thesis aims at solving some of the research challenges described in the previous section by proposing a framework that takes into consideration observed anomalies on a traffic network and demand prediction solutions based on time-series data, automatic internet search queries generation methods and semantic information. In summary, the main contributions of this thesis are:

- A methodology that identifies anomalies on transport networks and correlates them with special events by using Internet data. Why congestion is happening as well as why there are demand fluctuations on days when there are no apparent reasons for such phenomena are some of the questions that are explored for framework's configuration. The building blocks for this endeavor are (i) automatic detection of the time (ii) location, and (iii) magnitude of special events for real-time explanation of traffic congestion or road closures that are highly correlated with them. By using the developed methodology, real-time surveillance of the state of the transportation network during non-recurrent scenarios becomes gradually possible.

- A demand prediction solution that is based on automatic query generation. We explore information available on the internet for future non-recurrent overcrowding (or hotspots) prediction. A particular emphasis is given to special events that are publicly disclosed on social media and attract many people. The proposed framework correlates unstructured textual information with time-series traffic data with the ultimate goal of promptly alerting stakeholders for potential upcoming overcrowding alarms. It combines the Maximum entropy discriminant Latent Dirichlet Allocation (MedLDA), a maximum margin supervised topic modeling algorithm, with textual data obtained from automatically generated queries. Special events' information, such as the title, its start time and location, that are obtained from event listing websites construct our queries.

- A real-time demand prediction model that is able to forecast taxi demand using special events' data around venues. Time-series data and semantic information combinations are explored using machine learning and deep learning techniques in the context of creating a prediction model that is able to capture in real-time future stressful situations of the studied transportation system. The proposed frameworks are applied in event areas in New York city using publicly available taxi data. Regarding the event data, it was extracted automatically from the Web using either screen scraping techniques or Application Programming Interfaces (APIs). We empirically show that the proposed models are able to significantly reduce the error in the forecasts. The importance of semantic information is highlighted in all presented methods. The comparison of this framework with our daily forecast approach emphasized the significant contribution of hourly real-time demand information to the ultimate accuracy of our results. Through the analysis, we have identified which parameters help model's more accurate demand peak positioning on the axis of time, as well as the extent of their contribution.

- A two-stage model for real-time taxi demand prediction, where the residuals of our predictions are taken into consideration. The proposed framework includes two main models, the demand prediction model, whose main objective is to predict taxi demand based on the available historical demand-supply and event data for the areas that we are interested in, and the model of residuals, whose main objective is the estimation of the demand prediction residuals based on the calculated residuals that the first model attributed to that particular day of the month/week in the past, where an event was or was not scheduled. In other words, we take into consideration the time-series of residuals together with the historical actual demand and supply data of the area and we manage to improve our predictions' accuracy.

1.3 Thesis Structure

As previously mentioned, the management of serious phenomena of traffic congestion, that are observed in unsolicited time and place, and adversely affect our daily routine, will be successful if the three basic steps of (i) identifying, (ii) analyzing and finally (iii) predicting similar future anomalies are processed with the best possible tools and data.

Chapter 2 of this thesis starts with the development of a methodology that can identify anomalies on traffic networks and correlate them with special events by using Internet data. The main subject of interest is the investigation of why traffic congestion was occurring as well as why demand fluctuated on days when there were no apparent reasons for such phenomena. A "normality" baseline was defined at the outset and then used in the subsequent study of the demand patterns of individual days to detect outliers.

In many cases at the previous study, it was difficult to identify related Web documents that include event details for anomalies' explanation. Some of the visualized demand hotspots remained unexplained. Therefore, it was considered valuable to formulate a methodology that automatically scans the internet for events using search queries and then associates them with noteworthy demand fluctuations. **Chapter** 3 addresses the challenges related to retrieving and analyzing web documents about real world events and using them for demand explanation (if related to a past event) and prediction (if a future one). A supervised topic modelling algorithm with some very popular classifiers is combined, for the identification of the most accurate framework for transport demand prediction. Search aggregation, natural language processing, and query expansion are some of the tools that were exploited.

Events' information and time-series data are thereafter combined for the the formulation of a taxi demand prediction model (**Chapter** 4). Taxi pick-ups and dropoffs, weather data and finally event data are gradually incorporated into the forecast model. Through a detailed analysis and evaluation of each type of information, the crucial importance of event information in model's final performance is emphasized. At the same time, the performance of different machine learning techniques is presented.

In **Chapter** 5 the idea of a two-stage process that focus on the analysis, evaluation, and forecasting of prediction model's residuals in a real-time taxi demand forecasting framework is presented. Two different models are formulated; one that is focused on the direct taxi demand prediction based on the available historical taxi data and a second one, whose objective is the estimation of demand prediction residuals based on the historical performance of the first model.

Finally, in Chapter 6 final conclusions regarding the developed models and the obtained results are drawn, and directions for future work are discussed.

Chapter 2

Demand pattern analysis of taxi trip data for anomalies detection and explanation

2.1 Introduction

Transport systems function generally well. In some occasions though, unexpected and unwanted performance patterns are noticed that lead to system failures and cost implications. Significantly low speeds or excessively low flows at an unusual time are only some of the phenomena that may confuse a driver or transport authorities, since they are totally unexpected and frequently there is no obvious explanation for them. The term “anomalies” refers to those non-conforming patterns which appear into a well-defined notion of normal behavior. In the literature, similar phenomena can be described as outliers, exceptions or discordant observations. The common feature of all these terminologies is that they represent critical information in a wide variety of application domains, which is particularly useful for momentous events identification and crisis management.

Anomaly detection is extensively used in a wide variety of applications. It is a crucial task in many safety-critical environments, such as fraud detection for credit cards, insurance or health care, intrusion detection, activity monitoring through mobile phones, etc. Anomalies could occur due to changes in the behavior of systems, human errors, natural deviations in populations or fraudulent behavior. The application area typically defines the anomaly detection system and the methodology tools that will effectively identify and collect the necessary information for events assessment.

Transportation networks present several anomalous situations of particular interest and merit. Phenomena of different scale and influence have attracted the interest of several researchers who try to monitor and explore their specifications. Accidents, protests, celebrations, concerts, sport events define crowds, disruptions, road closures, etc., which subsequently cost time, money and urban pollution. Therefore, several methodologies have been developed for the detection and analysis of location, time and the purpose of them, in order to provide improved guidance to the users of the traffic system and reduce the impact of the associated problems. The requirement of understanding why people are travelling on the other hand, is often abundant in natural language form, and has been largely neglected.

Sensors can detect and accurately measure traffic congestion, flow models can represent how it should evolve on a specific network area and time window, but they cannot find a parade organized nearby. Figure 2.1 shows how New York City Halloween Parade affects taxi demand at the surrounding areas. Regions which have been marked with red shades indicate areas with increased demand for taxi trips. Even when this context is captured manually, limited options exist to correlate the two phenomena: one described with rich semantic information, the other with traffic data. In general, travel choices are strongly context-dependent, but context has been considered difficult, if at all possible, to capture, let alone be included in behavior models. Travel surveys usually don't go further than multiple-choice questions (e.g. travel purpose, accompanying travelers, perceived comfort, weather) because of user burden and cognitive limitations (e.g. memory, context-awareness), which is constraining a whole field to a myopic view of travel behavior.

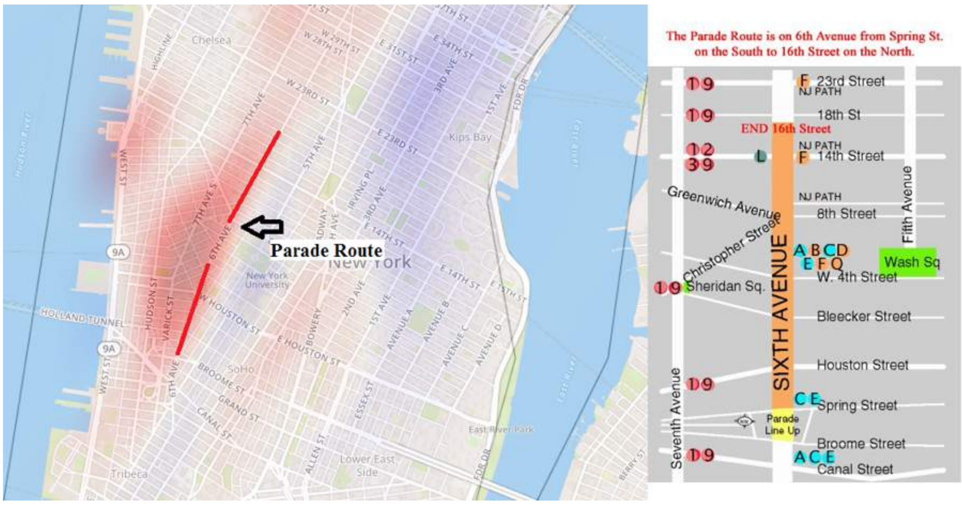


Figure 2.1: Taxi Demand for Halloween Parade NYC 2015

In previous works in collaboration with the Land and Transport Authority of Singapore, we showed that even simple event information (e.g. event category) collected from an online event directory, can be used to improve public transport arrival predictions (Pereira, Rodrigues, and Ben-Akiva, 2015). However, due to the complexity of the exploration of the open Web (e.g. using Google search), the use of internet data in transportation is currently limited to manually defined sources and highly fine-tuned processes. As mentioned in (Pereira, Bazzan, and Ben-Akiva, 2014), the grand challenge is to break this “barrier” and start jointly considering all kinds of contextual information by broadening the search space to the entire Web, instead of just focusing on a single type of contextual resource such as incident feeds or a manually build list of event websites.

The main contributions of this chapter are the formulation of a proper methodology that identifies traffic anomalies on traffic networks and correlates them with special events using internet data. Our main subject of interest is the exploration of why traffic congestion is happening as well as why there are demand fluctuations in days were there are no apparent reasons for the occurrence of such phenomena. The present study is not yet about automatically searching from the web for a random event, instead it's about getting the first building blocks for this endeavor: automatically detect time, location, and magnitude of such events for real-time explanation of traffic congestion or road closures which are highly correlated with them. Utilizing the developed methodology, we are gradually led to the real-time surveillance of the state of the transportation network during non-recurrent scenarios, such as the events described in the research, and provide real-time information and guidance to travelers and transportation administrators.

2.2 Literature Review

During the last decades, research and development in intelligent transportation systems (ITS) have given mature tools for monitoring, estimation and control of traffic networks (Perallos et al., 2015). They are largely supported by the ubiquity of pervasive technologies, such as radio-frequency identification, GPS, Wi-Fi, NFC and mobile phone communications. All these tools enable researchers to understand the dynamics of a city and improve traffic management and decision making processes.

There is no well-defined threshold above which we identify anomalies on transport systems. They are caused by accidents, sport events, parades, demonstrations, extreme weather conditions etc. The intuition is that anomalies should happen whenever the supply (e.g. buses, trains, network capacity) is misaligned with the demand (e.g. travelers) in ways that are not common for that location and time.

Anomaly detection could be implemented using either macroscopic or microscopic traffic variables. Variables from the first category, such as flow and occupancy, have been extensively examined for traffic incident detection (Parkany and Xie, 2005). Speed variations incorporate useful information (Li et al., 2009) resulting in the sensitivity increase to traffic patterns deviation. The second category of variables describes individual vehicles behaviors which are also valuable for certain research areas. Lane changing fractions, relative speeds and inter-vehicle spacing are some of the parameters that have been studied (Sheu, 2004; Chen et al., 2006; Barria and Thajchayapong, 2011). Simulation results showed promising results on transient anomalies and incident detection with low false alarms rates (Barria and Thajchayapong, 2011). However, the scale of the analysis is not the only parameter that is taken into consideration. Some researchers choose to study the problem from a different perspective, namely to focus on the supply components of a traffic network, such as traffic flows, densities and routing behavior (Pan et al., 2013; Liu et al., 2011; Christoforou et al., 2016). Traffic dynamics through segment densities were tried to be understood and used for prediction (Castro, Zhang, and Li, 2012). Topological variation in traffic flow between points and the visualization of the affected road segments of the anomaly has also been studied (Pan et al., 2013).

Meanwhile, anomalies occurrence is also connected with demand components, and more specifically with non-habitual overcrowding scenarios, such as public special events (sports games, concerts, parades, sales, demonstrations, and festivals) that directly affect them. Transport systems are generally designed with reasonable spare capacity in order to cope with those demand fluctuations. However, in several cases high waiting times are observed due to a congested traffic network that is no longer able to serve increasing transportation and mobility needs. Relevant scenarios have been extensively studied (Lee and Sumiya, 2010; Becker, Naaman, and Gravano, 2010; Nichols, Mahmud, and Drews, 2012). The start time as well as the duration of an event are two of the most important parameters that will define how the demand around a certain area would be affected. As a result, several studies are oriented to new data sources that can provide transportation systems and models with that information. By better understanding why these crowds occur, transportation models could be improved and present better planning and prediction results (Pereira et al., 2015).

The exploitation of the information deducted from the Internet attracts great interest (Yardi and Boyd, 2010; Becker et al., 2012; Watanabe et al., 2011; Abdelhaq, Sengstock, and Gertz, 2013; Xu et al., 2016). Social Media (i.e. Facebook, Twitter, Google+ and Flickr) are rich in local-context information generated by large online crowds. Information about special public events from social networks and other platforms that have content with a dynamic context (e.g., news feeds) can help in finding explanations for real-world phenomena.

Generally, anomalies can be seen as a group of observations lying (considerably) outside a region of likely expected values, given the “normal” behavior of a system. Detecting and identifying these anomalies involves continuous estimation of models of normal system behaviors at specific areas or time of interest. This process requires finding the best methodology for the type of data available as well as for the computational capacity of the system. Consequently, researchers have developed several techniques for anomaly detection to meet these diverse needs. Classification-based techniques define abnormal sets of values by inspecting them to determine if they exceed a certain threshold. For example, Zhang computed shortest paths and compared the recorded distances with them (Zhang, 2012) and Castro et al. characterized a road segment as congested when the observed density was above a certain value (Castro, Zhang, and Li, 2012). Clustering-based techniques are based on the assumption that normal data instances belong to a cluster. Bu et al. monitored distance-based anomalies using data structures and algorithms employing local clustering (Bu et al., 2009) and Candia et al. showed that anomalous events give rise to spatially extended patterns. Statistical based techniques give low probability to anomalous situations (Candia et al., 2008). Parametric distributions, histograms, regression models etc., are included in this category. Finally, information theoretic techniques analyze the information content of a data set and try to minimize its subset size, by simultaneously aiming to the minimum possible information loss Chandola, Banerjee, and Kumar, 2009. Examples of popular algorithms include Latent Dirichlet Allocation, LDA (Blei, Ng, and Jordan, 2003), which re-represents each document as a linear combination of latent bag-of-words (BoW) vectors, or topics, that can be seen as the building blocks of all documents in the collection. These have had tremendous success, including in transportation applications (e.g. decomposition of an incident record into its probable constituents (Pereira et al., 2015); text analysis for special events (Pereira, Rodrigues, and Ben-Akiva, 2013), and urban planning (Quercia and Saez, 2014).

The above studies have shown that there is great potential for using Internet data for information about planned events and their popularity. However, none of the previous studies explored automatic ways of detecting the time, location, and magnitude of events within an area where serious traffic congestion or a road closure occurs. Additionally, the distribution of taxi demand has not yet been used for the identification and explanation of special events.

2.3 Methodology

2.3.1 Definition of normality

The range of an affected area varies during an event, and descriptive boundaries cannot be easily provided. A multitude of sources provide high-resolution information from alternative channels, such as cell phone and other communication data, social platforms, and media coverage combined with the more traditional transport sensor systems (loop detectors, traffic radars, cameras, Wi-Fi, and Bluetooth sensors). Depending on the scale of influence, the best processing tools, as well as the necessary data precision, could be defined for the explanation of abnormal observations.

Figure 2.2 shows an overview of the proposed methodology. The starting point is the definition of a baseline “normality” according to historical mobility data, which, in the case of the experiments carried out for this study, correspond to GPS data from taxi trips in New York City. The first stage of analysis is an exploration of the available data during an extended time interval, such as a year or several months, primarily so that the dynamics of the studied area can be understood. The knowledge of which areas present high daily demand or which days present fewer trips in total because of low demand from city residents (i.e., Sundays compared with the rest of the week) will restrict incorrect assessments of significantly high or low values for the parameters that indicate anomalies.

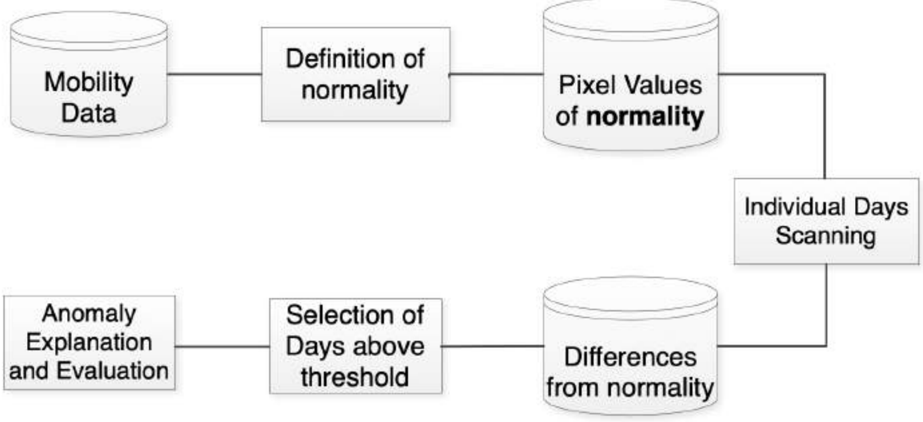

Mobility Data
Definition of normality
Pixel Values of normality
Individual Days Scanning
Differences from normality
Selection of Days above threshold
Anomaly Explanation and Evaluation

FIGURE 2.2: Developed Methodology

2.3.2 Estimation of Kernel Density

Special events, protests, visits by politicians, and so on attract many people in a certain area for a short or a long period. A theater, a stadium, or an exhibition center emerges as a point of interest for many citizens; therefore, these locations influence the overall heat map of a city's daily trips. For the detection of unusual motions or interactions, it is necessary to build representations of the selected areas of interest that are regions where the number of trips has significantly changed.

A general nonparametric technique that estimates the underlying density, thereby avoiding the need to store the complete data, is kernel density estimation.

Given a sample $S = \{x_i\}_{i=1,\ldots,N}$ from a distribution with density function $p(x)$, an estimate $\hat{p}(x)$ of the density at x can be calculated by using

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} K_\sigma(x - x_i) \tag{2.1}$$

where K_σ is a kernel function (sometimes called a "window" function) with a bandwidth (scale) σ such that $K_\sigma(x) = 1/\sigma K(x/\sigma)$. It is nonnegative, integrates to one, and has a mean of zero.

The Gaussian kernel function was chosen for this research because it is a function that weights included points. It is preferred because of its continuity, differentiability, and locality properties.

Data from one month at the initial stage, and for longer periods at a later stage, were analyzed. For each day, kernel density values were estimated and used for the calculation of the average day of the month to which they relate, according to the following equation:

$$\bar{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\hat{p}(x_i) \tag{2.2}$$

where N is the number of days included in the generalized average and $\hat{p}(x_i)$ are the kernel density values that correspond to 1 day. The kernel density values for a day are represented as a two-dimensional array. The step of the analysis grid is defined according to the scale of the studied area. In this analysis, each cell of the kernel grid is also called a "pixel“.

Figure 2.3a shows the monthly average demand for taxi pickups in the area of Manhattan in New York city, produced from the pixel-by pixel time series model used in this study. Blue shades denote areas with a high average demand, while lighter shades (yellow to white) denote areas with a lower demand. The figure merges analysis results from Python and a background image from OpenStreetMaps. This generalized depiction provides adequate information about the average dynamics of the studied area and was useful in evaluating, on a more concrete basis, the results obtained from analysis of individual day trips.

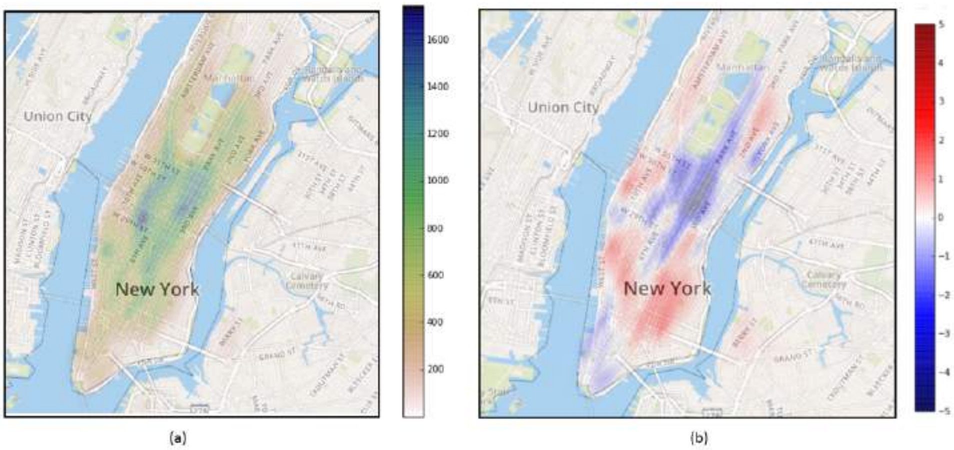


FIGURE 2.3: Kernel Density analysis depiction of (a) the average day (b) demand differences

2.3.3 Scanning Individual Days

After information was gathered on the distribution of demand in the area of interest for an average day, the next stage involves scanning individual days to find days during which the demand differed significantly from the "normality“ represented by an average day. The comparison phase only took into account the kernel density values and used the following Z-score formula:

$$Z_{diff,i} = \frac{\hat{p}(x_i) - \hat{p}(x)}{\sigma} \tag{2.3}$$

where i is the examined day and σ is the standard deviation of the average day. Through the Z-score data transformation, each kernel density value is given in units of how many standard deviations it is from the mean value and, consequently, how far it is from the demand levels of the average day. The maximum differences for each day were used for the next stage of analysis: determining the spatial localization of these maximum differences and finding an explanation for them. Differences in demand are depicted in Figure 2.3b, where red shades represent demand that is higher than the previous average analysis, whereas blue represents the opposite.

2.3.4 Explanation of anomalies

The Internet is a valuable resource for extracting information about special events, such as their location, duration, and popularity through Facebook likes or Google trends. Therefore, the Internet could be a very useful tool for the desired level of assessment for this research.

From the scanning procedure, it was possible to create a diagram that shows the most significant outliers captured for each day during the period of interest. There was no well-defined threshold above which a demand-side anomaly could be identified. Therefore, these observed demand fluctuations were dealt with by initially selecting for further analysis all days with a Z-score greater than two.

The information collected from the previous stage included not only the days that presented abnormal behavior, but also the location where the phenomenon of significantly high or low demand was noticed. The location was registered by identifying the pixel that showed the highest difference from the average picture. As a result, the above information could be used with Google search to determine if a special event had been held in that region. The generative procedure is presented in Figure 2.4.

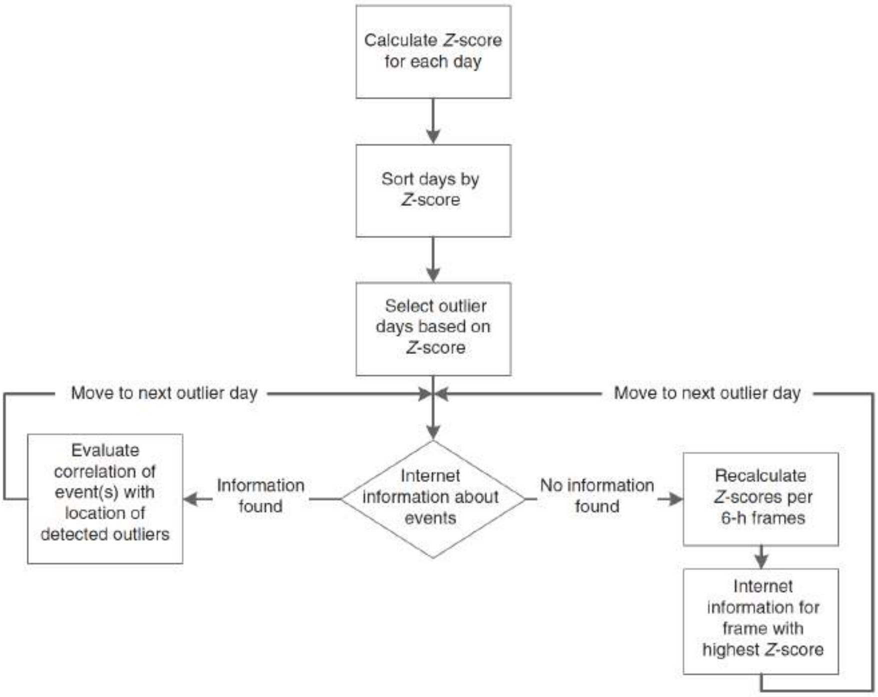

Calculate Z-score for each day
Sort days by Z-score
Select outlier days based on Z-score
Move to next outlier day
Move to next outlier day
Evaluate correlation of event(s) with location of detected outliers
Information found
Internet information about events
No information found
Recalculate Z-scores per 6-h frames
Internet information for frame with highest Z-score

FIGURE 2.4: Methodology for detection of anomalies and explanation.

2.4 Experiments

2.4.1 Description of Data

Google BigQuery public data sets (*BIG QUERY* 2016) were used as the main source of information for this study. In particular, the data set of trips made by Yellow Taxi in New York City since 2009 was explored. Records include fields capturing pickup and drop-off dates and times, pickup and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. Data structure as well as the level of detail enabled thorough understanding of the characteristics of taxi trips taken in a certain day. Through properly formatted SQL queries it was possible to select a period of interest as well as the variables to analyze in the context of this study.

2.4.2 Experimental Design

At the time of writing of this document, Google BigQuery limited the rate of incoming requests and enforced appropriate quotas on a per-project basis. There was also a maximum response size of 128 MB compressed; therefore, several experiments were implemented with a baseline of an average demand that corresponded to a 3-month interval. Each implemented query created a CSV file, which was saved and later exported from Google Cloud storage.

This research focused on monthly data from 2013 and 2015 around the area of Manhattan, the most densely populated borough of New York City. Each month, numerous events, festivals, and parades are organized. Several unexpected incidents further degrade the already low level of serviceability of the city's road network, such as accidents, road work, and so on. Road networks are also vulnerable to natural disasters such as floods and blizzards, which can adversely affect the travel on the network that remains intact after an event. The goal of this study was to detect demand-related anomalies and analyze and correlate them with scenarios like large sports events, concerts, religious festivities, and demonstrations. Studying and analyzing the vulnerabilities of road networks will help in prioritizing planning and budgeting and will also be useful in preparing emergency response plans.

2.4.3 Results

From the analysis of taxi-trip data sets from 2013 and 2015, several whole-day events and small-scale events were identified. The most characteristic examples are presented below.

Whole-Day Events

Large differences for a whole day were generally noticed when there was a public holiday or the city faced extreme weather conditions (Ferreira et al., 2013). By implementing the methodology described in detail in the previous section, it was possible to identify large deviations from the average day whenever a blizzard affected the city as well as on the Memorial Day.

Figure 2.5 shows a comparison of the Z-score values for each day in February 2013 with the average day for February. February 9 presents the greatest difference of the month. By using Google search, it was possible to easily determine that a blizzard in the northeast part of the United States affected the area of Manhattan on that day; therefore, taxi-trip counts and distribution were justifiably different from the average day.

But because extreme weather conditions are clearly noticeable and affect the entire region of a city, additional attention was given to events that mostly showed unexpected changes and traffic anomalies that transportation authorities and drivers could not easily explain and predict. Memorial Day falls in the category of all-day events. Citizens and authorities know in advance that taxi trips will be lower than on

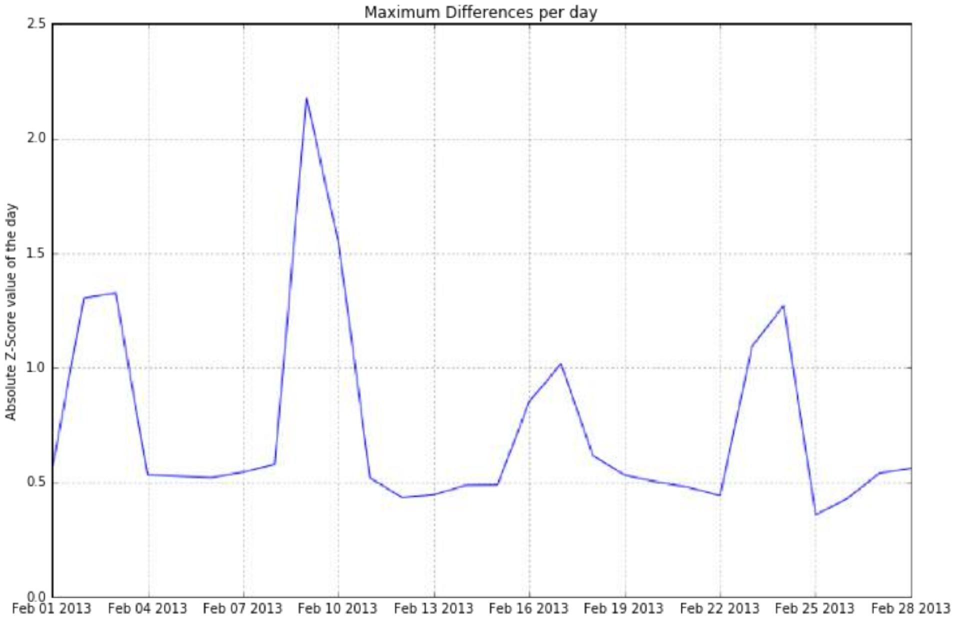


FIGURE 2.5: Absolute Z-Score values graph for February 2013.

an average Monday, because Memorial Day is a public holiday. This hypothesis was easily proved by an initial analysis of the data (Figure 2.6a). A decline in taxi-trip counts during this public holiday is evident.

A further examination of the differences in distribution of pickup points from the average day of May (Figure 2.6b) revealed a significant increase in traffic at Pennsylvania Station (known as Penn Station), one of New York City's train stations (Figure 2.6c). This result can be explained by the choice of many people to travel for the Memorial Day weekend. According to travel agency sales of the American Automobile Association and their website, AAA.com, some of the top destinations are Orlando, Florida; Myrtle Beach, South Carolina; Washington D.C.; and Miami, Florida. All these destinations are served by train trips departing from Penn Station; therefore, it is common for people to choose to get to the center of Manhattan at the end of their vacation by arriving at this station and then taking a taxi to their ultimate destination.

Small-Scale Events

Taxi pickup points for a whole day might not show significant changes in their distribution, because time intervals with higher fluctuations are combined with those that present low mobility patterns, such as night hours. Therefore, this study focused mostly on morning and afternoon pickup hours when people choose to participate in events, such as a sports event or a concert.

During the 3-month period from September to November 2015, Z-score values that significantly deviated from the average value were selected and, as a result, several traffic anomalies were observed.

October 31, 2015, New York City Halloween Parade

On Saturday, October 31, a Z-score value was observed that was significantly higher than that of the average Saturday for the 3-month period. By plotting the location of the highest value and then searching on the Internet, it was easy to

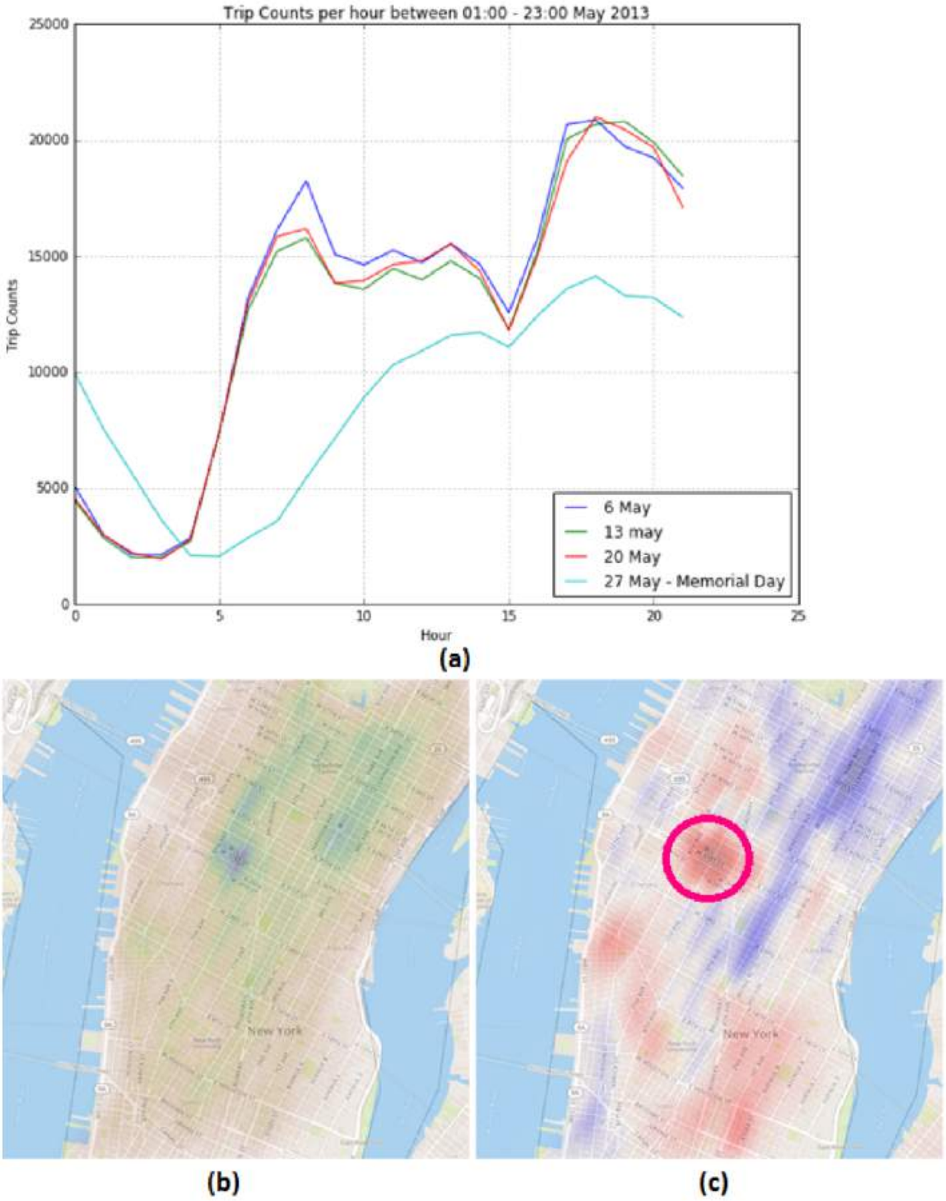


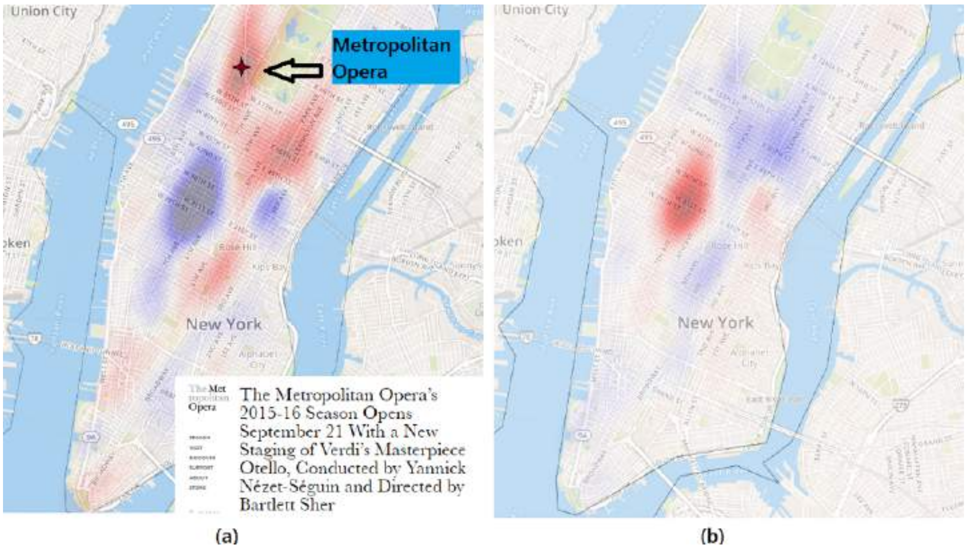
Figure 2.6: (a) Taxi Trip Counts per hour for each Monday of 2013 (b) Demand depiction on an average day in May, (c) demand on Memorial Day

determine that a Halloween parade took place on that day. Figure 2.1 shows how the distribution of taxi pickup points was developed after the end of the parade. The intense activity was transferred west of the parade route to walkways and 7th Avenue. The scale of the phenomenon around the route of the parade was consistent with the distance that someone might need to walk to find a taxi.

September 21, 2015, Metropolitan Opera Opening Night Gala

September was a month with several outliers from the average day of the 3-month period. On September 21, a Z-score value of 1.2 led to further analysis of the period from 21:00 to 23:30. The results are shown in Figure 7a. The location of the pixel with the highest Z-score (red star in Figure 7a) indicated that the Metropolitan Opera could be the explanation for this high demand. Eventually it was determined that the Metropolitan Opera's "Otello" Opening Night Gala, an event that attracts the spotlight and, therefore, the attendance of many prominent people, took place on September 21, 2015.

Figure 2.7a also shows very low demand (blue shading) close to Penn Station. Further investigation indicated that on September 7, the corresponding demand in this area was significantly high (red shading in Figure 2.7b), thus affecting the average corresponding day. Several people probably wanted to change modes after arriving at this central station from numerous events organized in the greater area of New York City (Indian carnival, the beginning of New York City Broadway Week, and the U.S. Open).



The Metropolitan Opera's 2015-16 Season Opens September 21 With a New Staging of Verdi's Masterpiece Otello, Conducted by Yannick Nézet-Séguin and Directed by Bartlett Sher

FIGURE 2.7: Differences in demand on (a) September 21 and (b) September 7.

September 24, 2015, Visit by Pope Francis

Large differences in the distribution of taxi pickup points were noticed on September 24, between 15:00 and 21:00. The Z-score value for that day was –2.155. An online search showed that Pope Francis visited New York City that day and, more specifically, Saint Patrick's Cathedral for an evening prayer. The negative sign of the Z-score indicates an overall reduction of taxi pickups in the hotspots compared with the average day; this finding is explained by the scheduled road closures in many parts of the city for security reasons related to the Pope's visit. Taxi pickup points increased in the area around the cathedral, but not close to it because, as Figure 2.8 shows, all transit through the roads around it was forbidden.

2.5 Conclusion and future work

Transportation networks present several anomalous situations of particular interest and merit. A methodology was presented here that identifies traffic anomalies on traffic networks and correlates them with special events by using Internet data. The system was evaluated by using the data set of trips made by Yellow Taxi in New York City during 2013 and 2015. A "normality" baseline was defined at the outset and used in studying the demand patterns for individual days to detect outliers. By using this approach it was possible to detect fluctuations in demand and to analyze and

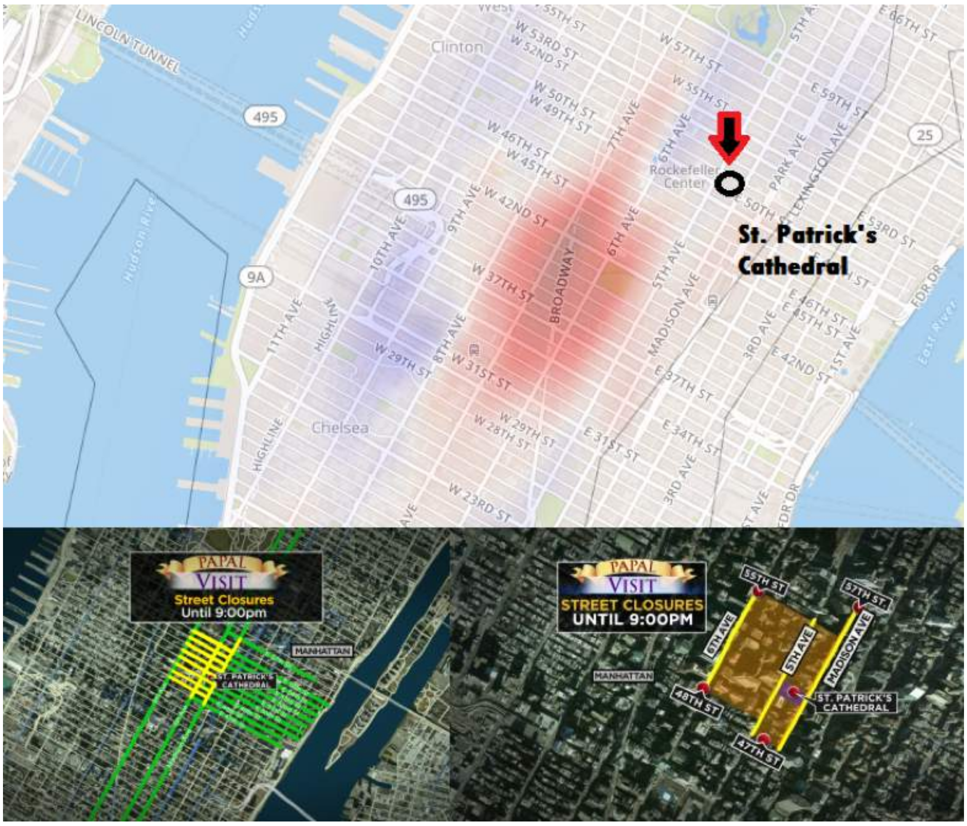


Figure 2.8: High demand during pope's presence at Saint Patrick's Cathedral.

correlate them with disruptive event scenarios like extreme weather conditions, public holidays, religious festivities, and parades. Kernel density analysis was used so that the affected areas, as well as the significance of the observed differences compared with the average day, could be depicted. By using the results of this investigation, it is possible to distinguish how a special event affects the spatial and temporal traffic flow in a studied region.

The focus of this study was not about automatically searching the web for a random event; instead, the goal was to establish the first building blocks for this endeavor: automatic detection of the time, location, and magnitude of such events for real-time explanations of traffic congestion or road closures that are highly correlated with them. By using prior knowledge, the goal is to monitor and predict in real time the state of the transportation network in nonrecurrent scenarios, such as the events described in this research, and to provide real-time information and guidance to travelers and transportation administrators. Future research includes (a) the application of information retrieval techniques to automatically capture relevant documents that explain abnormal conditions of the transport network identified by anomaly detection algorithms and (b) the use of natural language processing and popularity estimation techniques to extract contextual features that can be incorporated in transport prediction models, thereby making them context aware and more adaptive to demand.

Chapter 3

Predicting taxi demand hotspots using automated Internet Search Queries

3.1 Introduction

In Chapter 2 we discussed how to localize significant demand fluctuations, by taking into consideration the transport system's conditions on a normal day. Using kernel density visualizations, we analyzed the observed anomalies by emphasizing the date and time of the day that they were detected, as well as their extend. All the above features, but also some Web documents that we manually retrieved using internet search queries, led us to specific explanations for their appearance.

In this chapter, we shall take a closer look into the internet search queries that we used for traffic anomalies' explanation inspection. Well-defined queries can help us collect comprehensive and reliable information early enough to prepare mitigation measures. They are therefore a very important part of a sound and credible policy of dealing with some of the traffic anomalies presented in the previous chapter.

In general, mobility trends captured in complex transport systems consist of two basic components: utilitarian travel that mostly includes habitual behavior (e.g. commuting to work, weekly shopping) but also to a minor extent non-habitual needs (e.g. go to hospital, occasional shopping); and recreational travel, which comprises the human need for entertainment, social interaction and public expression. Efficient and effective intelligent transport systems should be able to take into consideration both of these factors for accurate demand predictions and better traffic management.

Current prediction approaches generally focus on capturing recurrent conditions, namely their seasonal spatial-temporal aspects (the "average" winter peak-hour Monday, in area X, with weather Y). The developed approaches can be successful for long-term planning applications or for modeling demand in non-eventful areas such as residential neighborhoods. However, in lively and dynamic areas where multiple special events take place, such as music concerts, sports games, festivals, parades and protests, these approaches fail to accurately model mobility demand precisely at times when it is needed - when the transport system of the area is under stress. The inability of the system to meet the new demand conditions emphasizes the need of good anticipatory capabilities which are capable to accept timely information on such phenomena.

Non-recurrent special events, such as concerts, sport games and demonstrations, are planned and largely advertised on the Web. An interesting fact is that it is much more likely to have citizens sharing their expectations/experiences about non-recurrent events than to talk about their daily commute. This plethora of information makes the Web an important tool for demand prediction and thus system's balance maintenance.

Previous studies have shown a strong correlation between number of public transport arrivals with the structured data mined from the Web Pereira, Rodrigues, and Ben-Akiva, 2015; Pereira et al., 2015. Namely, semi-structured information about events from announcements websites can be used as features for public transport arrivals. It has also been highlighted that they can be incorporated into a prediction model using topic modeling Markou, Rodrigues, and C. Pereira, 2018 or embeddings Rodrigues, Markou, and Pereira, 2019 and significantly reduce the error in the forecasts. However, information contained on these websites is usually incomplete, noisy or missing, which makes it difficult to generalize. Going beyond this approach raises two new challenges: what web-pages are relevant (information retrieval) and how to turn relevant information into model attributes (information extraction).

The aim of this chapter is the exploitation of information available on the internet for future non-recurrent overcrowding (or "hotspots") prediction. A particular emphasis will be given to special events that are publicly disclosed on social media and attract many people. The proposed framework will be able to correlate unstructured textual information with time-series traffic data with the ultimate goal of promptly alerting stakeholders for potential upcoming overcrowding alarms. It combines the Maximum entropy discriminant LDA (MedLDA) Zhu, Ahmed, and Xing, 2009, a maximum margin supervised topic modeling algorithm, with textual data obtained from automatically generated queries. These queries are constructed from basic event information (title, location, time) obtained from event listing websites.

3.2 Literature Review

The available spatial datasets, as well as the potential of events information and topic modeling for transportation problems, should be taken into consideration for an accurate demand prediction model formulation.

3.2.1 Demand Prediction for special events

Special events have a huge impact in urban mobility, regardless of their scale and type. Understanding their influence on the balance of a transport system is crucial for the development of reliable traffic management operations. For large-scale events (e.g. World cup, Formula One and Olympic games), best practices are already available for authorities to follow in order to manage these events and prepare for them well in advance Dunn Jr, Latoski, and Bedsole, 2006; Coutroubas and Tzivelou, 2003. However, these manual approaches do not scale to the vast amount of smaller and medium-sized events that take place on large metropolitan areas on a daily basis. Despite their reduced scale, these events still have a significant impact in the transportation system Pereira et al., 2015, especially when multiple co-occur. In these scenarios, common practice relies on reactive approaches rather than on planning Fuhs and Brinckerhoff, 2010; Kuppam et al., 2011. The demand prediction solution that we propose in this paper, takes into consideration event information that is automatically mined from the Web, and present itself with the potential for anticipating the effects of events and showing reliable tools for hotspot predictions in eventful areas.

Traffic demand modeling can benefit to a significant extend from earlier stage predictions. Developed methodologies can be grouped into two general classes: disaggregate response, typically with discrete choice models, where individual behavior choices are represented as a function of individual's characteristics (e.g. gender, age) and alternative choice properties (e.g. cost, duration); and aggregate response, based on machine learning or classical statistics, where a response variable (e.g. travel delay, total attendance) is modeled as a function of available data (e.g. location, time of day, event category).

The research of Shahin, Hüseyin, and Kemal, 2014 is included in the first framework class. They implemented an analysis of surveys conducted at three Turkish stadiums in advance and after matches, for the estimation of a binary logit model of mode choice (private car or public transport). A four-step model approach was proposed by Kuppam et al., 2011; Chang and Lu, 2013 to predict the number of trips, trip origin/destination, mode and vehicle miles traveled or transit boardings related to events. Despite being behaviorally sound and providing plenty of detail, these works highly depend on survey response and usually consider event features on a very superficial way. For instance, they rarely go deeper than general event category (e.g. sports, concerts).

The second framework includes several studies on forecasting demand. Some of the methods that have been proposed include Gaussian Processes Markou, Rodrigues, and C. Pereira, 2018, probabilistic graphical models Yuan et al., 2011; Rodrigues et al., 2017, neural networks Xu et al., 2017 and time series modeling Davis, Raina, and Jagannathan, 2016; Moreira-Matias et al., 2013. Markou et al. Markou, Rodrigues, and C. Pereira, 2018 highlighted the importance of semantic information through the formulation of a real-time taxi demand prediction model. The incorporation of popular events' information using topic modeling resulted in a noticeable increase of forecasts' accuracy. Similar conclusions were presented through the implementation of two deep learning architectures that leverage word embeddings and convolutional layers for combining text information with time-series data by Rodrigues et al. Rodrigues, Markou, and Pereira, 2019. A unified linear regression model that outperforms other popular non-linear models in the prediction accuracy is proposed by Tong et al. Tong et al., 2017. Particular emphasis was given to the conclusion that a simple model structure that eliminates the need for repeated model redesign proves to be able to behave better in prediction scenarios with high-dimensional features. Finally, over the last decade, deep learning has enabled many practical applications of machine learning in the fields of transportation and urban mobility Lv et al., 2015; Ma et al., 2015; Zhang, Zheng, and Qi, 2017a.

Taxi demand has been the subject of several applications, since the related datasets are sufficiently detailed. The yellow and green taxi public dataset of New York City in particular, has been the subject of a lot of research. Morgul and Ozbay Morgul and Ozbay, 2015 present an empirical assessment of taxicab drivers' labor supply. Yang and Gonzales Yang and Gonzales, 2017 identify locations and times of day where there is a mismatch between the availability of taxicabs and taxi service demand. Zhao et al. Zhao et al., 2016 use entropy and the temporal correlation of human mobility to measure the demand uncertainty at the building block level. They implemented three prediction algorithms to validate their maximum predictability theory. We Markou, Rodrigues, and Pereira, 2017 used kernel density analysis for demand fluctuations detection and analysis. Significant deviations from the average day were correlated with disruptive event scenarios such as extreme weather conditions, public holidays, religious festivities, and parades. Finally, some other research studies used this taxicab data to explore taxicab driver's airport pick-up decisions Yazici, Kamga, and Singhal, 2013, or travel time variability analysis Kamga and Yazıcı, 2014.

3.2.2 Online sensing and information retrieval

Along with the evolution of the Internet, the information contributed publicly by all of us keeps increasing substantially. Through popular websites and social platforms such as Facebook, Twitter, Wikipedia, eventful.com Foursquare, etc., it is possible nowadays to collect information about popular events that happened in the past as well as information for those planned in the near future. The anticipation of the consequences of popular events in the transport system will allow us to better prepare for such scenarios.

Based on the spatio-temporal information that defines a special event, such as its title (what), starting and ending time (when), and venue (where), queries can be automatically constructed for retrieving relevant documents using web search engine Application Programming Interfaces (APIs). However, the construction of such questions or *queries* is a nontrivial task Hölscher and Strube, 2000, because it contains a high risk of getting very broad matches. One trivial solution would be to just constrain to a limited number of online event announcement websites. But doing so risk to miss important information that could be elsewhere (e.g. venues own homepages, social media comments).

Another challenge is the fact that most of the retrieved documents contain details about an event in unstructured form and in a limited volume Schütze, Manning, and Raghavan, 2008. At the same time, many of the documents may be irrelevant to the matter of interest. Therefore, scoring metrics that determine the rank of each document with respect to many deciding factors have been developed and widely used. Google, for instance, indicates that they presently use over 200 factors to determine the ranking of the presented search results.

The most common metric for text ranking is based on weighting the importance of a term in a document, based on the statistics of occurrence of the term. The "Term Frequency-Inverse Document Frequency" (**TF-IDF**) is one of the most popular term-weighting schemes, that is also used in this research Jones, 1972. *tf* correlates to the term's frequency, defined as the number of times term t appears in the currently scored document d. Documents with higher score have more occurrences of a given term compared to the rest.

For the scores calculation, each set of documents is firstly represented in numerical form as a vector:

$$d_j = (t_{1,j}, t_{2,j}, ..., t_{n,j}) \tag{3.1}$$

where $t_{n,j}$ corresponds to a separate term in the document j (also known as state-space model). If a term occurs in the document, its value in the vector is non-zero. If we denote the raw count by $f_{t,d}$, then the simplest *tf* scheme is $tf(t,d) = f_{t,d}$. Other scheme possibilities from the literature are the Boolean frequency, where $tf(t,d) = 1$ if t occurs in d and 0 otherwise, the logarithmically scaled frequency and the augmented frequency. In this research the latter is preferred, as it prevents a bias towards longer documents. The raw frequency is divided by the raw frequency of the most occurring term in the document:

$$tf(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{max\left\{f_{t',d} : t' \in d\right\}} \tag{3.2}$$

where t' represents the most occurring term.

The inverse document frequency *idf* is a measure of how much information the word provides. This value correlates to the inverse of the number of documents in which the term t appears. Its default computation is:

$$idf(t, D) = log \frac{N}{\left|\left\{d \in D : t \in d\right\}\right|} \tag{3.3}$$

where N is the total amount of documents in the corpus. The denominator represents the number of documents where the term t appears.

Finally, the **tf-idf** is calculated as:

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \tag{3.4}$$

The calculated weights tend to filter out common terms, because a high tf-idf score is reached by a high term frequency in the given document and a low document frequency of the term in the whole collection of documents. Having therefore performed all the above steps, we can successfully detect documents' significant words and efficiently expand our new queries in the second stage.

The goal of query expansion is to ultimately arrive to an “ideal query vector", that gives us the best possible, ranked, set of documents. We start with the basic query, and refine it with each new result. Assuming we have an oracle for relevance of a retrieved document (i.e. a mechanism that tells which documents are relevant from the list), we can incorporate this information using the method proposed by Rocchio in 1971, therefore known as Rocchio algorithm (Rocchio, 1971):

$$\vec{q_m} = \alpha \vec{q_0} + \beta \frac{1}{|D_r|} \sum_{\vec{d_j} \in D_r} \vec{d_j} - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d_j} \in D_{nr}} \vec{d_j} \tag{3.5}$$

where $\vec{q_0}$ represents the original query, D_r the set of relevant documents and D_{nr} the set of irrelevant documents. All terms can be weighted separately in their respective document sets, using e.g. the TF-IDF statistic. The weights α, β, and γ are used to steer the importance of each set of documents. Some systems also ignore negative feedback ($\gamma = 0$) or leave the initial query part $\vec{q_0}$ unaltered, so that it is not taken into consideration in the equation. With the equation 3.5 updated questions can be formed for the web search engines, thus increasing the relevance of newly retrieved documents.

3.2.3 Topic Modeling

A considerable amount of important information about a planned event is in textual form. Adding to other structured information, such as date, time and location, we can find useful details concerning its content in the description, title or comments on the website hosting the announcement. To obtain an automated system, we still need to convert such data into a proper representation that a machine learning algorithm can understand. However, the dimensionality of the machine learning model will be increased beyond reasonable if we explicitly include the text, word by word. Natural language is rich in synonymy and polysemy, different announcers and locations may use different words, besides it is not always obvious which words are more “relevant". Topic modeling is the research topic that focuses on covering these weaknesses.

The approach of topic modeling is to represent a text document as a finite set of *topics*. These topics correspond to sets of words that tend to co-occur together rather than a single word associated with a specific topic. For example, a rock festival textual description could have a weight w_1 assigned to topic 1 (e.g. words related to concerts in general), w_2 of topic 2 (e.g. words related to festivals), w_3 of topic 3 (e.g. words related to the venue descriptions) and so on. In particular, we use a specific technique that is called Latent Dirichlet Allocation (LDA). For the readers that are familiar with Principal Components Analysis (PCA), there is a simple analogy: PCA re-represents a signal as a linear combination of its eigenvectors, while LDA re-represents a text as a linear combination of topics. In this way, we reduce the dimensionality from the total number of different words of a text to the number of topics, typically very low. Each document is represented as a distribution over topics, and each topic is a distribution over words. For further details concerning LDA's generative process please refer to the original article of David Blei and colleagues Blei, Ng, and Jordan, 2003.

In the particular domain of urban computing, Pereira et al. Pereira et al., 2015 studied the problem of using event data to help predict public transport demand. Their approach consists of using LDA to learn a topic model, and using the topic assignments as features in probabilistic graphical model. Other recent research efforts have focused on tools dealing with event identification in real time using micro-blog services. Kireyev et al. Kireyev, Palen, and Anderson, 2009 suggest that the family of topics models is a particularly promising tool for disaster response agencies, as Twitter often provides critical up-to-date and on-location updates about an unfolding crisis. Ramage et al. Ramage, Dumais, and Liebling, 2010 present a partially supervised learning model (Labeled LDA) that maps the content of the twitter feed into different dimensions, including substance, style, status, and social characteristics of posts. Ma H. et al. Ma, Wang, and Li, 2012 present a scalable implementation of a topic modeling (Adaptive Link-IPLSA) based method for online event analysis, which summarize the gist of the massive amount of changing tweets. All these studies further confirm the feasibility of topic modeling for microblog representation.

After finishing the task of finding the most representative topics of available event data Pereira et al., 2015, tweets Kireyev, Palen, and Anderson, 2009; Ramage, Dumais, and Liebling, 2010; Ma, Wang, and Li, 2012, mobile Web usage logs Yuan et al., 2015, etc., follows the classification stage, where algorithms such as the linear Support Vector Machines (SVMs) categorize available records into the classes that we are interested in. The integration of the mechanism behind the max-margin prediction models (e.g., SVMs) with the mechanism behind the hierarchical Bayesian topic models (e.g., LDA) under a unified constrained optimization framework can be also implemented using the maximum entropy discrimination latent Dirichlet allocation (MedLDA) model Zhu, Ahmed, and Xing, 2009. While LDA is unsupervised and just gives us the topics that best represent the corpus of documents, MedLDA looks for the topics that maximize the performance of a certain classification task. Such topics may in fact be quite poor to represent the corpus of documents itself, but they are better to solve a certain classification task. Intuitively speaking, MedLDA tries to find the topics that enable the maximum possible margin between the classes in question. As a consequence, in our case, topics will either strongly support or strongly oppose the likelihood of an event to be a hotspot or not.

The classification rule for a given document is determined by:

$$c^* = argmax\mathbb{E}[\eta_c^T \bar{z}|\alpha, \beta]$$

where c^* is the predicted class, $\bar{z}$ is a vector with the topic proportions of the document, η_c is a class-specific set of weights, α is a prior hyperparameter for the LDA component of the model, and β contains K vectors, β_k, each one being the word distribution for the topic k. The estimation η_c and $\bar{z}$, together with the best values of α and β is accomplished using an Expectation–Maximization algorithm (EM). In the E-step, we assume α and β to be fixed and estimate the posterior distributions for η_c and $\bar{z}$; in the M-step, we use those distributions to estimate the best α and β. The process is repeated until convergence. For further details please refer to Zhu, Ahmed, and Xing, 2009.

3.3 Data Description and Preparation

In the context of this research, two main categories of datasets were prepared and analyzed. The first category corresponds to taxi data, which are distributed by technology providers under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP) and were made publicly available by the NYC Taxi and Limousine Commission (TLC) *TLC Trip Record Data* 2018. The raw dataset includes fields capturing pick-up and drop-off dates/times, pickup and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

The second category corresponds to event data, obtained from event listing sites, such as Facebook, Timeoutworld, Eventful and Ticketmaster. They were extracted automatically from the Web using either screen scraping techniques or API's. Table 3.1 shows an example of all event attributes collected using Facebook's API. A prerequisite for including an identified event in our database is the correct knowledge of its start time, its location as well as the title of the event. Very important elements of an event record were also considered its detailed description and the number of people attending the event (using mainly Facebook's API). However, it was not feasible to have the last two features in all our records.

TABLE 3.1: Event data excerpt.

name	example
sid	facebook_235808830094292
title	ZHU - Neon City Tour: Terminal 5 - New York, N...
venue	TERMINAL 5
description	In an ever shifting social landscape of elusiv...
latitude	40.769615
longitude	-73.992770
start	2016-05-12 21:00:00
end	2016-05-12 23:00:00
url*	https://www.facebook.com/events/235808830094292
attending*	978
category*	MUSIC
address*	610 W 56th St, New York, NY 10019, USA

3.3.1 Hotspots identification using Kernel Density maps

Special events, protests, visits by politicians, and so on attract many people in a certain area for a short or a long time period. A theater, a stadium, or an exhibition center emerges as a point of interest for many citizens; therefore, these locations influence the overall heat map of a city's daily trips. For the detection of unusual motions or interactions (hotspots), it is necessary to build representations of the selected areas of interest that are regions where the number of trips has significantly changed.

A general non-parametric technique that estimates the underlying density of historical taxi data, thereby avoiding the need to store the complete data, is kernel density estimation (KDE). Given a sample $S = \{x_i\}_{i=1,\ldots,N}$ from a distribution with density function $p(x)$, an estimate $\hat{p}(x)$ of the density at x can be calculated by using:

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} K_\sigma(x - x_i)$$

where K_σ is a kernel function with a bandwidth (scale) σ such as $K_\sigma(x) = \frac{1}{\sigma} K(\frac{x}{\sigma})$. It is non-negative, integrates to one, and has a mean of zero.

The kernel density values are represented as a two-dimensional array. For the Manhattan area, which is our main study area, a grid step was chosen with a value of 100, which results in a space of approximately 93 meters on the longitudinal axis and 200 meters on the latitudinal axis between each grid point. The selected grid size gives us a fairly informative picture of taxi demand in the area of interest, since we believe that a popular venue can easily influence the demand conditions at a distance of up to 200 meters around it.

We calculated the kernel density values of an average day for the first six months of 2016 (six different heat maps, one for each month). At a later stage, we compared each day with the average demand day of the corresponding month. Some example heat maps are depicted in Figure 3.1. We standardized the final output using the Z-score formula:

$$Z_{diff,i} = \frac{\hat{p}(x_i) - \hat{p}(x)}{\sigma}$$

where i is the examined day and σ is the standard deviation of the average day. Through the Z-score data transformation, each kernel density value is given in units of how many standard deviations it is from the mean value and, consequently, how far it is from the demand levels of the average day. Z-score values above 2 will also emphasize the areas where a hotspot has been located and should be taken into account in the next stages of our analysis.

For more details on the average and individual day kernel density analysis, please refer to Markou, Rodrigues, and Pereira, 2017.

Our objective is to develop an architecture that is able to classify potential hotspots for future events. An approximate average value for an event duration could be four hours, based on the existing list of events with a clear start and end time that we have at our disposal. Therefore, the concluding output of this stage is 6 heat maps per day so that our final Kernel density maps, and subsequently the located hotspots (Z-score value $>$ 2) on each one of them, correspond to a 4-hour time window. A characteristic example of a heat map with multiple events ending at the same time is depicted in Figure 3.2. The white “x"s define the events that took place on a Thursday in May. Through those heat-maps we are in a position to better visualize and evaluate the observed demand fluctuations. The number of taxi journeys recorded daily is enormous, so through the kernel density analysis we are able to make clearer results in the appropriate spatial resolution and avoid the chaotic depiction of thousands of points on the map.

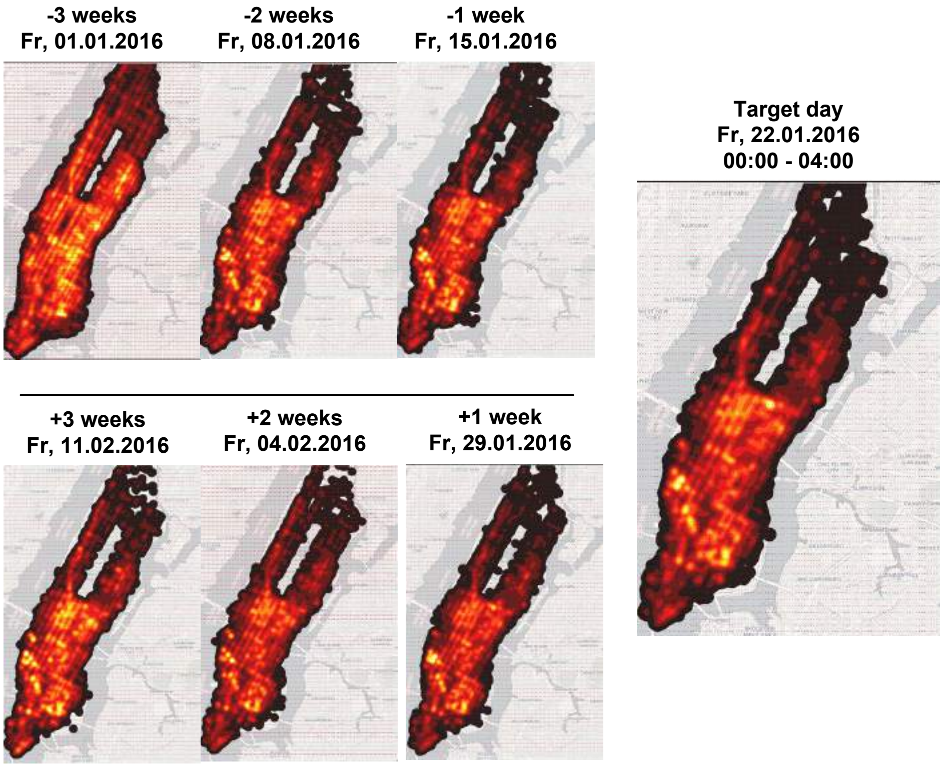


FIGURE 3.1: KDE of taxi pickups several weeks from each other.

3.3.2 Event records filtering

The event database is a list of event titles, locations, start times and occasionally short descriptions. Since our study analysis is limited to the time period January - June 2016 and each heat map gives us possible hotspots during a 4-hour window, we:

- Discard events without end time information.
- Discard events without description. At a later stage, this information is necessary as it will be used for each event's attribute "category" (e.g. music, sports) determination.
- Remove events where the venue is specified as “online” or “New York.”
- Discard duplicates after records collapse. Their identification was mainly accomplished through the intersection identification of similar tokens in title and description.

At a second stage, event records are also correlated with the Z-score values of the corresponding heat maps. As described earlier, for each cell of the grid we have a Z-score value. Knowing the exact coordinates of each cell, as well as the event location, attempts were made to identify areas with a high Z-score within a radius of 300m around the event.

After all, only events with an attending count over 500 people or a Z-score above ”2“ were kept for further exploration. Of all 36 000 events that our initial database has, only 858 events met these criteria. High Z-Score values showed only 279 events, therefore the number of "abnormal" events is smaller compared to those events that did not affect transport system's balance considerably. The attributes of those 858

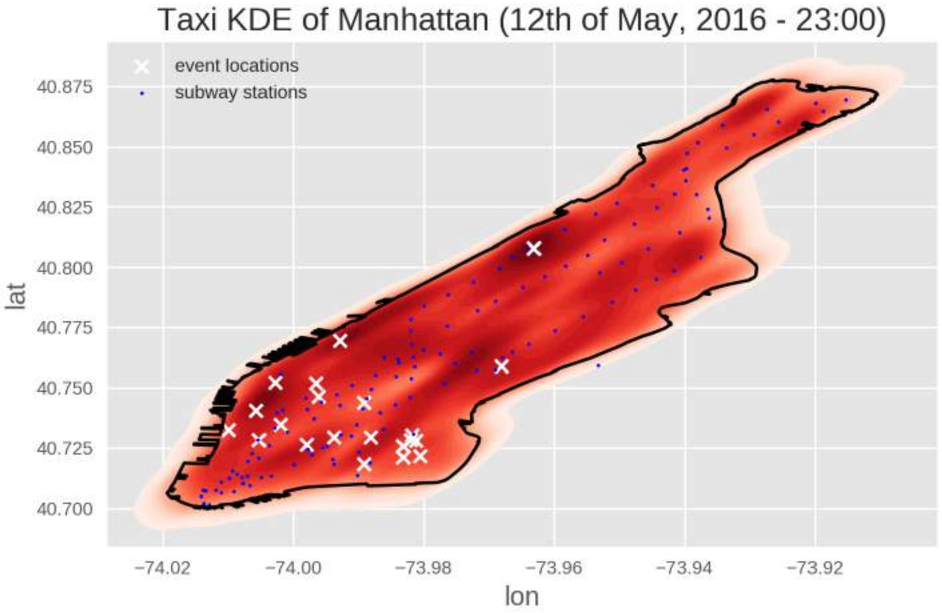


FIGURE 3.2: Heat map of taxi pickups in Manhattan

events are going to be used for training and testing the proposed methodology described in the following section.

3.4 Methodology

From a previous study Markou, Rodrigues, and Pereira, 2017 it is already known that there are several categories of events that can significantly affect the equilibrium of demand-supply of a transportation system. The degree of influence varies according to the day and time that the event takes place, its location, and of course its popularity. In this research, we are interested in predicting a demand hotspot considering the degree of influence of past events. Through machine learning techniques, several queries generation and expansion practices we will examine the prediction performance, of a classifier that determines whether an upcoming event is a hotspot or not.

The methodology presented in Figure 3.3 was developed to achieve the above objectives. It depicts the training stage which consists of two parts, applied in sequence: (a) the supervised regression model that will be responsible for characterizing texts as relevant or not, an essential component in the query expansion process (see Rocchio's formulation, eq. 3.5); and (b) the topic model that groups all retrieved documents based on specific events and gives us the word terms that increase hotspot's identification accuracy.

3.4.1 Ranking framework development

For the first part of the training mode, a detailed framework is formulated on the calculation of the relevance score between each event and the corresponding documents retrieved from the Web using search queries. The final scores will allow us to perform query expansion at a later stage.

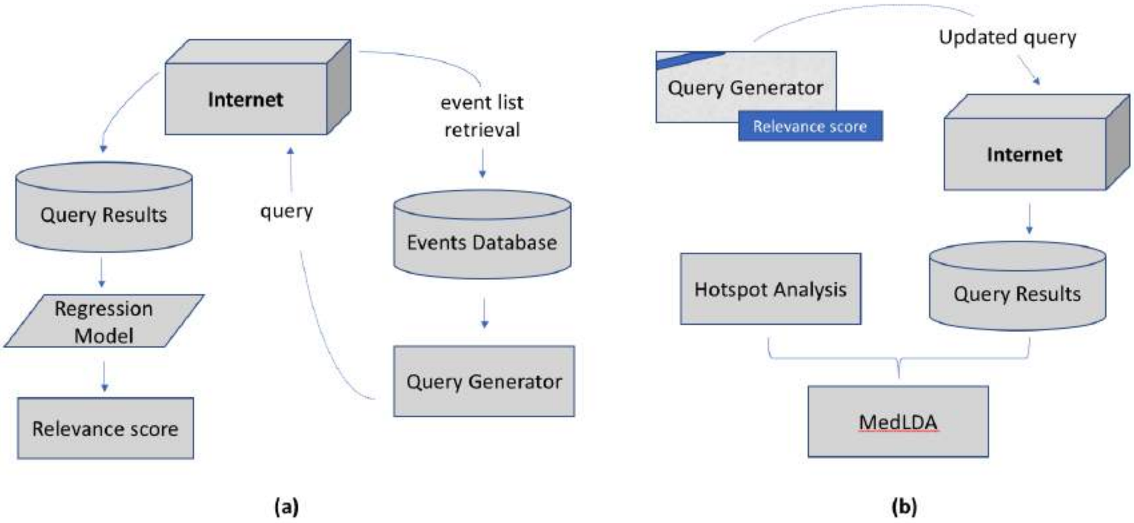


Figure 3.3: Methodology architecture

Web search engines allow to ask questions in the form of queries to hopefully get some relevant information. In this study, the events database of 858 records is used for the query generation step. Three query sets were developed (see also Fig. 3.4):

- **Query-Set A - Baseline structure** : Only one query per event using the attributes "title“ and "venue“.
- **Query-Set B - Varied structure** : It includes three different queries per event using the attributes "start time“ and "venue“. The first query involves an unrestricted search of the terms, while the second and third query seeks exact matches of the titles using quoted phrases.
- **Query-Set C - Enhanced structure** : That set is formulated after the exploitation of **Query-Set A** and **B** and the receiving of the corresponding Web documents; for this and the characterization "enhanced“. Our final Query-Set includes only one query per event using the attributes "start time“ and "venue“ and two expansion terms from the **Query Set B**.

Although there are plenty of the other possibilities, like extracted named entities from the event description, we assume that the selected attributes are the most informative for the event and its context. We also removed non-alphanumeric symbols from queries and added the phrase “New York” to narrow down the search results.

All queries in the respective sets are sent through different Searx instances to the web search engines Bing and Google and an aggregated list of search results is returned per query. The total number of retrieved documents per query is restricted to 50 for the first and third query set and to 30 for the second query set.

The **Query-Set B - Varied structure** was formed to be particularly restrictive. As mentioned earlier, it includes 3 different queries and their rules are further explained below:

- 1^{st} query: Exact match search of venue name and start time using quotes.
- 2^{nd} query: Exact match search of only the venue name. The date was not restricted. Three letter abbreviations of weekday and month were also added.

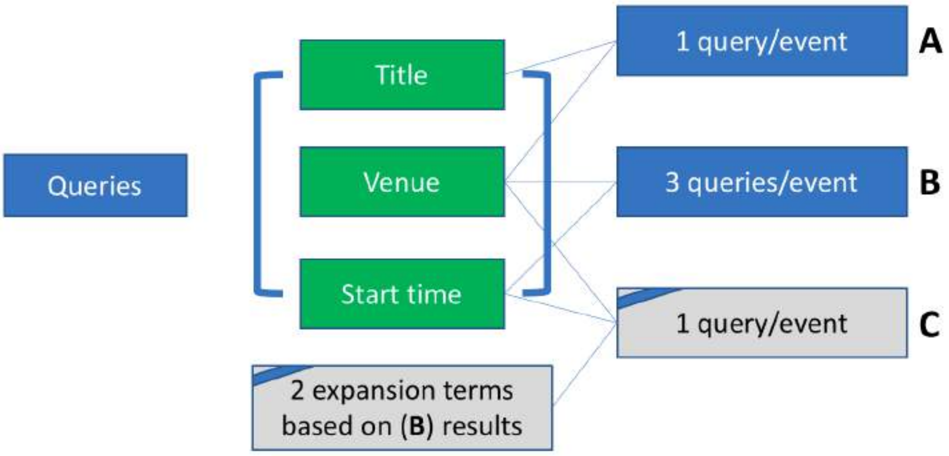


FIGURE 3.4: Methodology diagram for queries generation

- 3^{rd} query: Venue and start time are included without restrictions. Time is inserted after 12-hour time convention.

All response attributes from our queries are listed in Table 3.2. The shown example is returned using the 3^{rd} query of **Query-Set B**. It is worth mentioning that the retrieved URL can be used for document aggregation as it can only refer to a unique web page.

TABLE 3.2: Searx data excerpt of one query result.

name	example
title	ZHU - Tickets - Terminal 5 - New York, NY - May 12th, 2016
url	http://www.terminal5nyc.com/event/1107229-zhu-new-york/
snippet	May 12, 2016 ... With sold out debuts in Los Angeles and New York City [...]
positions	[1]
score	1
engines	[google]
parsed_url	[“http”, “www.terminal5nyc.com”, “/event/1107229-zhu-new-york/”]
queries	[3]

3.4.2 Framework enhancement using event’s category

An attribute that we considered fairly important for a decent performance of our classifier, is the "category“ of the event. With this term we denote the type of event that each database entry represents, e.g. music concert, sport game, demonstration, etc. This important feature is not feasible to always retrieve using APIs, so in many records this information was missing. Therefore, we applied the Term Frequency Inverse Document Frequency (TF-IDF) representation on the attribute "event description" to create an event category classifier.

According to J. Ramos et al. Ramos, 2003, TF-IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with high TF-IDF numbers imply a strong relationship with the document they appear in, suggesting that if that word were to appear in a query, the document could be of interest to the user. In our case, using the description of the event we will try to find the key words that have the highest weights and are highly correlated with our predefined categories. Before the calculation of TF-IDF weights, stopwords and terms whose frequency in documents exceeds 50% are removed. We have also developed a list of all possible event categories encoded as ordinal numbers. After TF-IDF implementation, the final weight matrix and the event categories were used as inputs for our stochastic gradient descent (SGD) classifier. The specific model is chosen because it is suited for handling sparse data with ease. From the initial dataset, 20% of events had the "category“ attribute empty, and using the above tools, we managed to associate 75% of our event records with one of listed event categories. Table 3.3 summarizes the results of the classifier for 14 indicative categories, as well as its average performance for all included categories.

Table 3.3: SGD Classifier Performance

Category	**Precision**	**Recall**	**F1-Score**
Art	0.76	0.84	0.80
Book	0.86	0.83	0.85
Business	0.74	0.78	0.76
Comedy	0.94	0.88	0.91
Community	O.82	0.68	0.74
Education	0.74	0.71	0.72
Fitness	0.82	0.79	0.81
Food	0.80	0.81	0.80
Hobby	0.83	0.75	0.79
Movie	0.85	0.80	0.82
Music	0.89	0.93	0.91
Nightlife	0.84	0.87	0.85
Religion	0.74	0.81	0.77
Volunteering	0.72	0.64	0.68
Average scores	0.82	0.83	0.82

3.4.3 Relevance score estimation

For the estimation of the relevance score, we will create a database of relevant documents for the target events. The criteria that we took into consideration for the identification of those documents are:

- Direct matches on time information, and URLs
- String distance measurements
- Keywords matching

For the first criterion, the scores are based on the elements presented in Table 3.4. Great emphasis is given to the duration of the event, because those matches are rare yet very useful.

The second criterion implies the identification of similarity between two specified string groups. Four distance algorithms are used for that purpose: Levenshtein l, Ratcliff/Obershelp r, Sørensen/Dice d and Jaccard j. The end return of each one of them is a value between 0 and 1. All four algorithms were applied in the title and description attributes of event records and document contents respectively. Then, each score is aggregated and to the final matching score calculation using the equation: $s = \frac{2(l+r+d+j)_{title}+(l+r+d+j)_{desc}}{2}$. Double weight was given to the "title“ string group as it contains more relevant information about the event. The highest possible score on complete accordance of all groups has therefore chosen to be "6“.

Table 3.4: Relevance Score Estimation

Event Attribute	Document Content	Score
Date	Date	+0.25
Start Time	Start Time	+0.25
End Time	End Time	+0.25
Duration	Duration	+1
URL	URL	+2
Domain Name	Domain Name	+1

The final criterion for the relevance score calculation is the keywords matching. Exact keyword matches will increase our relevance score by 2, since it can be considered a relatively strong indicator.The Rapid Automatic Keyword Extraction (RAKE) algorithm Rose et al., 2010, an unsupervised, domain–independent, and language–independent method for extracting keywords from individual documents was used for each text. It is also worth noting that, since the keyword extraction only happens on equally long snippets, there is no bias towards larger texts. Finally, besides the exact keyword matching identification, we also considered word similarities. When very similar keywords are detected using the Ratcliff/Obershelp similarity measurement ($r > 0.85$), the total relevance score increases by 1.

The threshold value between relevant and non-relevant documents discrimination was the score “6". In other scenarios, when more queries are used, this score can be adjusted accordingly. Of the 858 events, 717 were associated with at least one relevant retrieved document. For the rest 141 events, it is interesting to note that 36% (52) of them were correlated with a demand hotspot at a previous stage (see subsection 3.3.2) and we were expecting to find effortlessly relevant documents, since they looked fairly popular only by the heat-map inspection.

Thereafter follows the **query expansion** stage. Using the the Rocchio algorithm that was presented in section 3.2, updated queries were formulated with the venue name, the start time and two new expansion terms. The 3^{rd} Query from the 2^{nd} Query Set was used as the initial base query q_0 and the numerical statistic TF-IDF was applied as a weighting scheme for the respective document set with $\beta = 0.75$ and $\gamma = 0.15$. Terms that are finally chosen to be used in the new query, but already exist in the base query q_0 terms, are ultimately rejected. The first 50 documents that the web search engine returns using the expanded query are then saved for the supervised classification model training and Query Sets' evaluation and comparison.

3.4.4 Supervised Topic Modeling algorithm Application

The performance of our final model is not only based on the Query Sets' application but also on the selection of the classification mechanism and its parameters. MedLDA has several hyperparameters to tune, namely α, l (penalty term for misclassifications) and C (penalty for soft margin slack variables). Following the results of a prior 10-fold cross-validation grid search methodology implementation Borysov et al., 2016, we set $\alpha = 0.001/K$, $l = 10$, and $C = 30$. The topic number K was also set to 10, based on the same previous research.

Apart from the input constants mentioned above, MedLDA receives as input the retrieved documents from our search queries. All event documents were placed in chronological order since they are not independent and identically distributed. The same keywords presented in the training set, were also used in the test set.

As mentioned in 3.3.2, in total 279 events met the hotspot criteria from the whole dataset of 878 events. Subsequently, our dataset is imbalanced, since the classification categories are not approximately equally represented. Therefore, undersampling of the majority (normal) class has been implemented as a good means of increasing the sensitivity of a classifier to the minority class (hotspots). Using the synthetic minority over-sampling technique (SMOTE), the correct equilibrium of hotspot-correlated events between training and test set was achieved.

We will compare MedLDA's performance with three other baseline models, all based on a typical two-stage process that firstly determines the topics using the standard unsupervised LDA, and secondly characterizes each event as a demand hotspot or not using a classifier. The classifiers that we chose are the Support Vector Machine (SVM) classifier, because MedLDA integrates the same mechanism behind the max-margin prediction, Logistic Regression (LR) and the Multi-layer Perceptron (MLP) classifier, as two very popular benchmark algorithms. The output of all these frameworks is the prediction of demand hotspots' presence, based on the characteristics of each event, as represented through the topics.

The comparison will be based on three measures: *accuracy*, *F1-score*, *standard deviation (σ)* and *Cohen's kappa statistic (κ)*. Since the *accuracy* is just the percentage of correctly classified instances, it could be misleading because the test set is imbalanced. Therefore, the *F1-score* is been taken into consideration as it relates the number of correctly classified positive instances with the quantity of the missed ones. Furthermore, κ quantifies the chances that the agreement between the model's results and the perfect classifier is random. The value of $\kappa = 0$ corresponds to the performance of a random classifier, while $\kappa = 1$ corresponds to the perfect classification without random coincidence. Finally, the standard deviation was also included as an additional accuracy indicator.

3.5 Results

The final results of MedLDA, LDA+SVM, LDA+LR and LDA+MLP choosing $K = 10$ as the preferred number of topics, are presented in Table 3.6. It can be concluded that the performance of MedLDA is much better than all the other approaches. It is also worth noting that there is no noticeable performance difference between the three Query sets. Extracted topics using the aggregated documents of **Query-Set B** were as useful as the ones extracted from the baseline queries for hotspot prediction.

For a more detailed analysis of the results, the most discriminative terms of each topic as well as their assigned η are illustrated in Fig. 3.5. Table 3.5 additionally summarizes the number of hotspots and scheduled events that are correlated with twelve of the most popular venues in NYC. Looking thoroughly the **Query-Set A** (Fig. 3.5), the term "Hirschfeld" appears on the top topic (10a). The specific topic has also a high η value ($\eta = 12.45$) which indicates that the Hirschfeld Theater, which is located in midtown Manhattan, is very often associated with a hotspot presence. More generally, most of the stemmed terms in this query set can be separated into the following abstract classes: (i) the type of venue (e.g. “restaurant” in topic 7a and “gallery” in topic 10a), (ii) artist name or performance title words (e.g. the name “eric prydz” of a musician in topic 2a and the title of a theater performance: “waitress” in topic 5) and (iii) event categories (e.g. “music” in 2a and “family” in 6a).

Table 3.5: Unique event venues and their hotspots.

	hotspot counts	total event counts
Central Park	7	20
TERMINAL 5	6	15
SOBs Sounds of Brazil	5	7
Cornelia Street Cafe	4	9
Al Hirschfeld Theatre	4	5
Intrepid Sea Air and Space Museum	4	5
Le Bain at The Standard High Line	4	5
BB King Blues Club and Grill	3	19
City Winery New York	0	24
Madison Square Garden	0	11
Webster Hall	0	11

For the **Query-Set C**, a clear contribution to the hotspot classification seems to make the first (10b) and last topic (1b). The keywords “madison” and “garden” that appear in topic 1b (and 1a from the **Query-Set A**) give us a clear indication that special events in Madison Square Garden are not related to high taxi demand existence. This conclusion can be also justified by the fact that very good public transport services are offered in the vicinity, and there may be already a significant number of taxi demand regardless of events (notice that a hotspot is a relative measure, i.e. a strong deviation from the norm).

Other conclusions about the total performance of query sets are:

- The **Query-Set C** has more time-related keywords in its topics, since a different date format was used in its corresponding queries that was apparently more useful in searching relevant documents.
- The **Query-Set C** has a wider keyword variety, which allows for clearer classification results.
- Fund-raising events that took place in Central Park (e.g. "AIDS Walk“ and "9/11 Memorial Walk and Run“), as well as street festivals that were very popular in June, are well represented by topic 9b.
- The **Query-Set C** has the highest rate of false positives, due to the fact that after the query expansion stage, the results returned by the search engine are not filtered, and documents with only time information and no accurate event correlation are included in the topic model.
- According to the proportions of event categories, there was an above average number of hotspot events in the category “nightlife” with over 30% of them ending at 4 a.m..

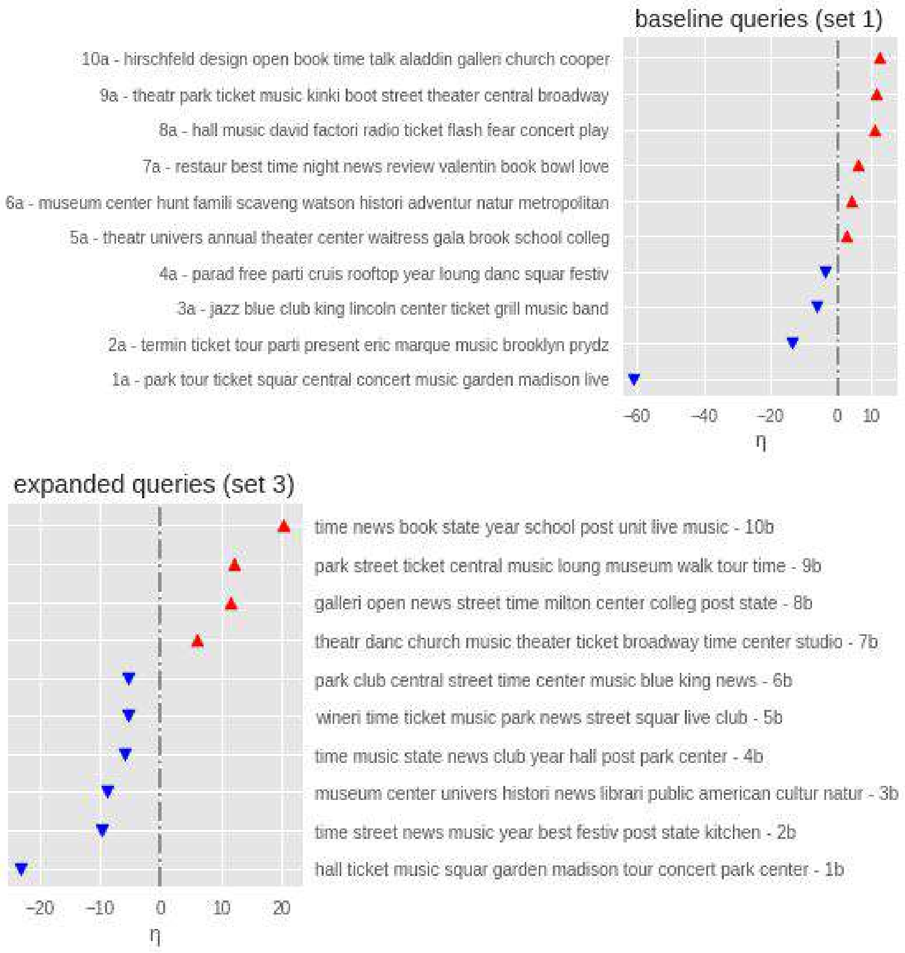


FIGURE 3.5: Topics for baseline queries with assigned η.

3.6 Conclusion

We proposed a framework for real-world phenomena dynamics prediction, and more specifically for taxi demand prediction during non-recurrent events. We combined information extracted from the Web with time-series data to build a predictive model of taxi demand hotspots around special event venue areas. This is typically a challenging case for transport planning since special events originate high variance in demand. Taxi demand is correlated with many parameters of underlying information and, currently, most taxi centers rely on formal processes and manual work for a fleet organization and taxi distribution. Even the more advanced new services, like Uber or Lyft, still face great challenges in terms of demand prediction. Our results show that information retrieval using query expansion methods outperforms other baseline methods that rely solely on the basic attributes of an event.

It is concluded that there are a lot of compromises that need to be made in real-world applications. A methodology that relies on commercial search engines has its shortcomings since there is limited access to their full search index. Structural data from transport services and event listing sites do not always contain enough details for an adequate correlation of different sources of information.

In future work, we aim at exploring at a deeper level the potentials of the proposed machine-learned ranking model. More complicated forms of query expansion

Table 3.6: Comparison of classification performance of different Query sets.

	Accuracy	F_1-score	κ
LDA + SVM - Query-Set A	0.501	0.525	0.141
LDA + SVM - Query-Set B	0.512	0.517	0.135
LDA + SVM - Query-Set C	0.622	0.564	0.265
LDA + LR - Query-Set A	0.535	0.497	0.126
LDA + LR - Query-Set B	0.559	0.501	0.203
LDA + LR - Query-Set C	0.600	0.532	0.214
LDA + MLP - Query-Set A	0.534	0.500	0.133
LDA + MLP - Query-Set B	0.494	0.485	0.088
LDA + MLP - Query-Set C	0.600	0.543	0.226
MedLDA - Query-Set A	0.715	0.484	0.296
MedLDA - Query-Set B	0.732	0.5	0.329
MedLDA - Query-Set C	0.756	0.618	0.439

frameworks will be examined and the contribution of events' descriptions will be considered, since the implemented analysis showed that they contain very important keywords.

Chapter 4

Is travel demand actually deep? An application in event areas using semantic information

4.1 Introduction

In the previous chapter we saw how we can predict demand hotspots emergence around venues, using historical information about similar events, as well as well-defined strategies, which made use of search aggregation, natural language processing, and query expansion. Motivated by the outcomes of that work, in this chapter we explore the contribution of events information in real-time demand forecasting, along with the use of time-series data.

Together with the evolution of the Internet, the information contributed publicly by all of us keeps increasing substantially. Through popular websites and social platforms such as Facebook, Twitter, Wikipedia etc. that we frequently visit and update, it is possible nowadays to collect information about popular events that happened in the past as well as information for those planned in the near future. In most cases, this information is typically in the form of unstructured natural-language text. Nevertheless, using screen scraping techniques or Application Programming Interfaces (APIs) we are able to retrieve accurate spatio-temporal details that define a special event, such as its title (what), starting and ending time (when), and venue (where) in the area of our interest. This information may prove to be reasonably valuable for the understanding of observed phenomena in a transport system that are directly related to scheduled events and to the foresight of similar situations in the future.

Disruptions due to special events are a well-known challenge in transport operations, since the transport system is typically designed for habitual demand. Large events (e.g. music concerts, sport games, political rallies) do not receive any special treatment or attention, which often creates non-recurring congestion and overcrowding. Taxi-calling platforms, such as Uber (Wikipedia, 2017b), Grab (Wikipedia, 2017a) and Beat (*Beat* 2017) are becoming increasingly popular, especially in situations of traffic congestion, because they can efficiently facilitate resource allocation. Through their application, passengers are able to call or pre-order a taxi, even when they are located in an area where it is very hard to find a driver. This trend, therefore, proves that there is a tremendous need for better taxi fleet organization and taxi distribution from a taxi center, according to the demand of an entire city (Chan et al., 2016).

The prediction of taxi demand is challenging because it is correlated with many parameters of underlying information. Currently, the general practice is to rely on formal processes and manual work. For very big events, such as the Olympic games or football world cup matches, the event organizers engage with operators and authorities to meet the enormous demand. For smaller events though, this task is labour-intensive and even with a list of events, their impact is hard to estimate. A timely and accurate notion of demand impact is accordingly needed in order to design adequate system changes and to disseminate appropriate information to the public.

One of the main contributions of this paper is the development of a real-time demand prediction model that is able to forecast taxi demand using special events' data around venues. The time window of the study comprises 4 years (2013-2016) and our work is focused on New York City (NYC). Using a large-scale public dataset of 1.1 billion taxi trips and event data from the Web, we empirically show the value of modeling textual information associated with the events, and that the proposed machine learning approaches are able to outperform other methods from the state of the art by combining information from different sources and formats.

In the context of new forecasting models' formulation, the use of tools gaining ground day by day has also been attempted. More specifically, along with the machine learning techniques we will explore deep learning architectures that over the last decade have made major advances in solving artificial intelligence problems in different domains such as speech recognition, visual object recognition, and video processing (Schmidhuber, 2015). It is precisely this success of deep learning in handling different types of data, such as images, audio and text from different domains, that makes it particularly well-suited for the data fusion problem of combining time-series and textual data.

4.2 Literature Review

The available techniques for demand forecasting, the accessible spatial datasets, as well as the potential of events information and topic modeling for transportation problems, should be taken into consideration for an accurate demand prediction model formulation.

4.2.1 Prediction Applications using Taxi Data

In urban systems, nature, economy, environment, and many other settings, there are multiple simultaneous phenomena happening that are of interest to model and predict. Taxi demand prediction is one of the non-trivial research subjects that attracts particular interest due to its inherent complexity. Taxi centers need to better organize their fleet, so that they can maximize their profits, as well as the satisfaction of their employees and customers. Therefore, a better demand prediction can be beneficial for all interested parties. It should be also noted that a good prediction model will be necessary in the future, because self-driving taxis will need to decide where to roam before picking up passengers.

Several methods have been proposed to predict taxi demand, including probabilistic models (Yuan et al., 2011), neural networks (Xu et al., 2017) and time series modeling (Davis, Raina, and Jagannathan, 2016; Moreira-Matias et al., 2013). A unified linear regression model that outperforms other popular non-linear models in the prediction accuracy of the Unit Original Taxi Demand (UOTD) is proposed by Tong et al. (Tong et al., 2017). A simple model structure that eliminates the need for repeated model redesign proves to be able to behave better in prediction scenarios with high-dimensional features.

One of the main pillars of research using taxi data is the modeling of the dispatching center. Zhang et al. propose a taxi drivers' recommendation system which is based on the combination of drivers' location and demand hotspot's hotness (Zhang et al., 2016). They focus on the travel requirements' understanding as well as on the reduction of cruising time and wasted energy. Miao et al. propose a dispatching framework for balancing taxi supply in a city (Miao et al., 2016) . Their objectives include matching spatio-temporal ratio between demand and supply for service quality with minimum current and anticipated future taxi idle driving distance. New York City has been a subject of several studies, since its yellow and green taxi public dataset is easily accessible and sufficiently detailed. Markou et al. used kernel density analysis for demand fluctuations detection and analysis (Markou, Rodrigues, and Pereira, 2017). Significant deviations from the average day were correlated with disruptive event scenarios such as extreme weather conditions, public holidays, religious festivities, and parades. Morgul and Ozbay present an empirical assessment of taxicab drivers' labor supply (Morgul and Ozbay, 2015). Yang and Gonzales identify locations and times of day where there is a mismatch between the availability of taxicabs and taxi service demand (Yang and Gonzales, 2017). Zhao et al. use entropy and the temporal correlation of human mobility to measure the demand uncertainty at the building block level (Zhao et al., 2016). They implemented three prediction algorithms to validate their maximum predictability theory. Some other research studies used this taxicab data to explore taxicab driver's airport pick-up decisions (Yazici, Kamga, and Singhal, 2013), or travel time variability analysis (Kamga and Yazıcı, 2014).

4.2.2 Deep models in transportation

Deep learning is evolving rapidly in solving problems that have resisted the best attempts of the artificial intelligence community for many years. It has proven to be able to find intricate structures in high-dimensional data, and thus it is an important tool in various applications in the domain of science (LeCun, Bengio, and Hinton, 2015). In the field of transportation and urban mobility there are already studies showing deep learning's successfullness.

Lv et al. proposed a deep-learning-based traffic flow prediction method that takes into consideration the traffic flow features as learned by a stacked autoencoder model (SAE) (Lv et al., 2015). Results comparison with more traditional approaches based on Support Vector Regression (SVR) and radial basis functions (RBFs) showed proposed method's superiority. Ma et al. proposed a long short-term memory neural (LSTM) network for travel speed prediction (Ma et al., 2015). Their empirical results on data from Beijing indicate that LSTMs outperform other methods such ARIMA and SVR, which the authors justify with the ability of LSTMs to capture long-term dependencies over the time-series. A model with Mixture Density Networks (MDN) on top of LSTM was proposed by Xu et al. (Xu et al., 2017). In their approach, the city is previously divided in smaller areas and then the LSTM-based model is used to jointly predict the taxi demand for the next time-step in all the areas. Finally, the prediction of crowds' traffic in city's regions using a deep-learning based approach, called ST-ResNet, is presented by Zhang et al. (Zhang, Zheng, and Qi, 2017b). Experiments on two types of crowd flows in Beijing and New York City (NYC) demonstrate that the proposed method outperforms standard approaches such as ARIMA and vector auto-regressive models.

While the approaches described above demonstrate the potential of deep learning for transportation problems, none of these approaches consider the effect of events in order to improve their predictions. The deep learning approaches proposed in this paper aim at bridging this gap by focusing on event areas and showing that data fusion techniques that combine text data about events and time-series observations of mobility demand can significantly improve predictions.

Finally, in addition to the methodologies presented above, this research gives a further insight into another deep framework presented by Damianou and Laurence, the deep Gaussian process models (Damianou and Lawrence, 2013). Gaussian processes (GPs) are already a well-known technique for their flexibility and their non-parametric, probabilistic approaches to function estimation in an analytically tractable manner. In the deep GP framework, the data is modeled as the output of a multivariate GP whose inputs are governed by another GP. Experiments with simulated and real data showed that always DGPs exceed or match the performance of a single layer GP. Salimbeni and Deisenroth have also shown that the DGPs often exceed the single layer significantly, even when the quality of the approximation to the single layer is improved (Salimbeni and Deisenroth, 2017). For mobility trends where we have two basic components, namely the habitual behaviour of drivers and their provisional behavior which comprises the human need for entertainment, social interaction and public expression, we believe that DGPs can provide the tools for capturing this multi-factorial nature. To the best of our knowledge they have not been tested yet as a traffic prediction method, and we will try to investigate their utility through this research.

4.2.3 Internet as a data source for special events

Internet, and more specifically the several social networking services that exist, has become a popular distribution outlet for users looking to share their experiences and interests on the Web. Taking as an example the Facebook, which has over 1.86 billion monthly active Facebook users (Facebook MAUs) worldwide, it is clearly understood that the information derived from the above platforms, can undeniably help discerning explanations about observed real-world phenomena, such as non-habitual overcrowding scenarios.

Due to the importance of special events' impact in urban mobility, it is not surprising that they are a predominant part of transportation research. Fortunately, the Internet is rich in information about public special events. In an earlier work, using public transport data Pereira et al. compared an origin/destination (OD) prediction model with and without simple information obtained from the Internet, such as event type or whether the performer/event had a Wikipedia page (Pereira, Rodrigues, and Ben-Akiva, 2012). It was verified that such information could reduce the root mean squared error (RMSE) by more than 50% in each OD. In another study, Pereira et al. presented a machine learning model that classifies aggregated crowd observations into explanatory components (Pereira et al., 2015). After the identification of overcrowding hotspots in the city-state of Singapore, potential explanations from several event announcements websites were retrieved. It was observed that the model is able to recover observed total impacts with an RMSE between 55% and 85%.

The internet is also a valuable source for other aspects of mobility research. For example, Twitter has been used for crisis management (Thom et al., 2012; Sakaki, Okazaki, and Matsuo, 2010), urban management and planning (Frias-Martinez et al., 2012), the analysis of different aspects of mobility (Cheng et al., 2011) and the mobility characteristics of different nations (Hawelka et al., 2014). Due to the complexity of the exploration of the open Web (e.g. using Google search), the use of internet data in transportation, however, is currently limited to manually defined sources and highly fine-tuned processes.

4.2.4 Topic models

A considerable amount of important information about a planned event is in textual form. Adding to other structured information, such as date, time and location, we can find useful details concerning its content in the description, title, comments on the website hosting the announcement. To obtain an automated system, we still need to convert such data into a proper representation that a machine learning can understand. However, the dimensionality of the machine learning model will be increased beyond reasonable if we explicitly include the text, word by word. Natural language is rich in synonymy and polysemy, different announcers and locations may use different words, besides it is not always obvious which words are more “relevant". Topic modeling is the research topic that focuses on covering these weaknesses.

The approach of topic modeling is to represent a text document as a finite set of *topics*. These topics correspond to sets of words that tend to co-occur together rather than a single word associated with a specific topic. For example, a rock festival textual description could have a weight w_1 assigned to topic 1 (e.g. words related to concerts in general), w_2 of topic 2 (e.g. words related to festivals), w_3 of topic 3 (e.g. words related to the venue descriptions) and so on. In particular, we use a specific technique that is called Latent Dirichlet Allocation (LDA). For the readers that are familiar with Principal Components Analysis (PCA), there is a simple analogy: PCA re-represents a signal as a linear combination of its eigenvectors, while LDA re-represents a text as a linear combination of topics. In this way, we reduce the dimensionality from the total number of different words of a text to the number of topics, typically very low. Each document is represented as a distribution over topics, and each topic is a distribution over words. For further details concerning LDA's generative process please refer to the original article of David Blei and colleagues (Blei, Ng, and Jordan, 2003).

In the particular domain of urban computing, Pereira et al. (Pereira, Rodrigues, and Ben-Akiva, 2015) studied the problem of using event data to help predict public transport demand. Their approach consists of using LDA to learn a topic model, and using the topic assignments as features in a shallow neural network model.

Other recent research efforts have focused on tools dealing with event identification in real time using micro-blog services. Kireyev et al. (Kireyev, Palen, and Anderson, 2009) suggest that the family of topics models is a particularly promising tool for disaster response agencies, as Twitter often provides critical up-to-date and on-location updates about an unfolding crisis. Ramage et al. (Ramage, Dumais, and Liebling, 2010) present a partially supervised learning model (Labeled LDA) that maps the content of the Twitter feed into different dimensions, including substance, style, status, and social characteristics of posts. Ma H. et al. (Ma, Wang, and Li, 2012) present a scalable implementation of a topic modeling (Adaptive Link-IPLSA) based method for online event analysis, which summarize the gist of the massive amount of changing tweets. All these studies further confirm the practicability of topic modeling for microblog representation.

While the approaches described above demonstrate the potential of events information and topic modeling for transportation problems, none of them work on a real-time basis. The machine learning techniques proposed in this paper take into consideration the advantage of short-term time-series trends and predict taxi demand having knowledge of traffic conditions in the near past. The developed model can be

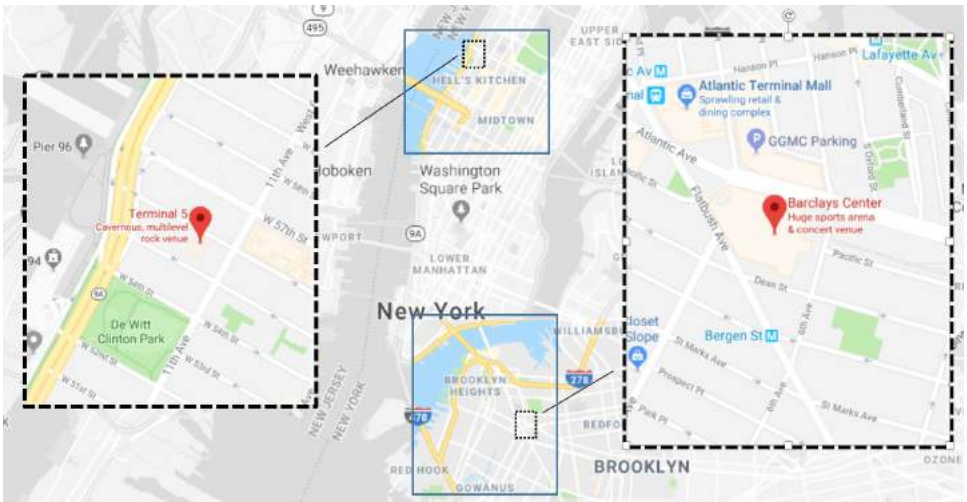


Figure 4.1: Map of the two study areas.

used to instantaneously inform interested parties about possible increased demand around venues and better critical fleet distribution for system requirements fulfillment.

4.3 Data Description and Preparation

In this research, we work with two major datasets: events and taxi data. The latter is distributed by technology providers of authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP) and were made publicly available by the NYC Taxi and Limousine Commission (TLC) (*TLC Trip Record Data* 2018). We use taxi data from 1/1/2013 through 6/30/2016, which includes around 600 millions taxi trips after data filtering. The dataset specifies for each drop-off and pick-up event the GPS location and the time-stamp.

Based on this data, we looked at a list of the top venues in NYC (*Best concert venues in NYC* 2017) and selected the two venues for which more complete event records were available online: the Barclays Center and Terminal 5. The first venue is located in the heart of Brooklyn and it is the state-of-the-art home of the NBA's Brooklyn Nets and the NHL's New York Islanders. It is one of the most popular facilities in the New York metropolitan area because it hosts many sold-out concerts, conventions and other sporting and entertainment events. It is ranked top five globally in 2015 for gross revenue and attendance by Billboard and Venues Today. On the other hand, the Terminal 5 is a 3-floor venue that regularly hosts concerts with many different audiences and that is located in the heart of Manhattan. Given the geographical coordinates of these two venues, we selected all the taxi pickups that took place within a bounding box of ±0.003 decimal degrees (roughly 500 meters) to be our study areas. Figure 4.1 shows a map of these areas.

Regarding the event data, it was extracted automatically from the Web using either screen scraping techniques or API's. For the Barclays Center, the event information was scrapped from its official website, since it maintains a very accurate and detailed calendar. We collected a total of 751 events since its inauguration in late 2012 until June 2016. As for the Terminal 5, we used the Facebook API to extract 315 events for a similar time period. In both cases, the event data includes event's title, date, time and description. In Table 4.1, we show an example of an advertised event and its data, as obtained from the official website of the venue.

TABLE 4.1: Event data example

Field	Content
start time	12/04/2015 20:00:00
title	Stevie Wonder
url	http://www.barclayscenter.com/events/detail/stevie-wonder-songs-in-the-key-of-life-performance
description	Legendary singer, songwriter, musician and producer Stevie Wonder is bringing his SONGS IN THE KEY OF LIFE PERFORMANCE tour to Barclays Center on April 12. Rolling Stone declared that the show is possibly 2014s greatest testament to the limitless potential of American music itself.

4.3.1 Preliminary Analysis of time-series data

The raw dataset that we obtained from TLC (*TLC Trip Record Data* 2018) includes fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. Within the framework of the forecasting model that we want to formulate for Barclays Center and Terminal 5, the filtered taxi trip records are additionally aggregated by hour.

For most regions, the taxi demand is governed by a certain amount of randomness (e.g., unexpected events) and some degree of regularity (e.g., weekly patterns), which can be exploited for prediction. Before the configuration of initial modeling structure, we decided to remove any deterministic trends and focus our analysis on the remaining fluctuations. A simple, yet very effective way, of identifying these daily or weekly recurring patterns is by constructing a historical averages model, which computes the individual averages for each (hour of day, day of the week) pair based on historical data (from the train set only). The historical averages then represent a fixed recurring trend, which can be easily removed from the data.

Based on our experience with various time-series forecasting problems with urban mobility data, by removing the burden of capturing these well-known recurrent trends from the model, this simple detrending procedure significantly improves the predictive performance. With our focus on the demand component that is directly correlated with events scheduled on that day, we managed to see more accurate predictions. In fact, as our experiments demonstrate, it makes even the simplest linear models quite competitive baselines to outperform.

4.3.2 Text data pre-processing

Generally, textual data mined from the Web is difficult to process in its original state. Specific pre-processing steps are usually required in order to make it more amenable to learning methods, and more specifically to the topic modelling stage that will follow. Therefore, we follow a simple conventional text-processing pipeline consisting of:

- HTML tag removal
- Lowercase transformation for words' variability restriction purposes
- Tokenization, a tool that divides a sequence of characters into pieces of tokens

- Lemmatization for inflectional endings removal, and words return to their base form (lemma)
- Stopwords and very frequent words removal, which typically do not bring any additional useful information
- Removal of words that appear only once in the whole dataset

4.4 Demand Prediction Model

Our prediction model should demonstrate the hypothesis that contextual information is significant for real-time taxi demand prediction in the vicinity of special event venues. Generalizing to other cases, and upon available event data, this should be valid to any area in which demand can be somehow associated with available contextual information, as for example school areas and information about school holidays, shopping areas and large sales, and main governmental buildings and public demonstrations.

4.4.1 Model Selection and Comparison

The proposed approach is focused on hourly short-term predictions. The specific time-step was chosen, as it represents the average time that a driver needs to cover the maximum distance within Manhattan. Consequently, demand predictions for the following hour can be very beneficial for a taxi company or a share-mobility application that aims at optimizing its fleet exploitation.

For the implementation of our short-term predictions we consider first two simple approaches (i) a linear regression (LR) model and (ii) a Gaussian process (GP) regression model using the scikit-learn machine learning library in Python (Pedregosa et al., 2011). The former is chosen for its simplicity and interpretability and the latter because it is flexible enough to represent a wide variety of interesting model structures. GPs have shown to achieve state-of-the-art results for various tasks, such as travel-time prediction (Idé and Kato, 2009), traffic volume forecasting (Xie et al., 2010) and public transportation trips predictions around special event areas (Rodrigues et al., 2016).

Thereinafter, we consider more complex models: a neural network architecture based on fully-connected (FC) dense layers and a Deep GPs architecture. The motivation for using FC layers results from already existing research studies demonstrating their good performance for certain time-series forecasting problems (Gers, Eck, and Schmidhuber, 2002; Cheng et al., 2016; Rodrigues, Markou, and Pereira, 2019). Deep GPs on the other hand have shown great efficiency on modeling complex data by automatically discovering useful structures and encoding abstract information (Damianou and Lawrence, 2013).

For models' performance validation and comparison we will use the mean absolute error (MAE), the root mean square error (RMSE) and the coefficient of determination (R^2), computed as follows:

$$MAE = \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n| \tag{4.1}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2} \tag{4.2}$$

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y_n})^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} \tag{4.3}$$

where N denotes the number of instances in the dataset, $\hat{y}_n$ is the predicted taxi pick-ups count for the n_{th} instance, y_n is the corresponding true pick-ups count and $\bar{y}$ is the mean of the observed counts.

4.4.2 Model formulation

The final hourly taxi pick-ups and drop-offs will be used for the establishment of our first two basic models: an elemental model with only pick-ups information; and a second model with both pick-up and drop-off information. For the initial models' formulation and evaluation we focused on Barclays Center.

The starting point of our forecasting model is to use a 12^{th} Order Autoregressive Model, namely Y_t is regressed against $Y_{t-1}, Y_{t-2}, ..., Y_{t-11}$, where Y represents the detrended taxi pickups counts.

$$\hat{Y}_{t+1} = \hat{\beta_0} + \hat{\beta_1} Y_t + ... + \hat{\beta_{12}} Y_{t-11} \tag{4.4}$$

where $\hat{\beta_0}, \hat{\beta_1}, ..., \hat{\beta_{12}}$ are estimated using data through period t. The order of the autoregressive model was chosen after several experiments with different number of lags. The best performance was achieved using information from the previous 12 hours only.

The first three years of our dataset (2013-2015) are model's training set and the first six months of 2016 (January 2016 - June 2016) our test set. The results of this simple model form the baseline of our analysis (see Table 4.4).

Then, taxi drop-off counts were introduced. We insert the same number of detrended drop-off counts, thus increasing the number of model's dependent parameters to 24.

$$\begin{aligned}\hat{Y}_{t+1} = \hat{\beta_0} &+ \hat{\beta_{1,1}} Y_t + ... + \hat{\beta_{1,12}} Y_{t-11} \\ &+ \hat{\beta_{2,1}} D_t + ... + \hat{\beta_{2,12}} D_{t-11}\end{aligned} \tag{4.5}$$

where D represents the detrended hourly drop-off counts.

The performance of our enriched model is significantly improved (Table 4.4). The R^2 score has increased by almost 7%, thus demonstrating how important it is to introduce information about taxi arrivals at the same area, at an earlier time. Through a sensitivity analysis of models' parameters, it was also found that the more recent the pickup lag is, the greater its significance to the final model performance. The final parameter vector $\beta = (\beta_0, \beta_{1,1}, ..., \beta_{2,12})$ showed us that the drop-off lags D_{t-3} and D_{t-4} appear to contribute more effectively than the rest of drop-off lags. This result allows us to understand that there is a strong correlation between taxi drop-offs three to four hours before our calculated pickups; a conclusion that seems to be related ultimately to the average duration of an event that Barclays Center hosts, such as a basketball game or a concert.

4.4.3 Weather data

Before including information about events that took place in the selected venues, we decided to evaluate the contribution of weather data in the hourly demand prediction. The dataset was obtained from the National Oceanic and Atmospheric Administration (NOAA) and corresponds to daily observations from a weather station located in the Central Park in NYC. It contains information about daily minimum and maximum temperatures, daily precipitation, wind, presence of snow, fog, snow depth, etc.. The performance of the enhanced model showed us that weather does not contribute positively to our forecast's accuracy. All three error statistics remain almost unaffected (see Table 4.4). This result may be due to the fact that we have only daily weather information at our disposal and not for shorter time periods. Therefore, for an hourly demand prediction model, information about the weather of an entire day is probably too aggregated. Additionally, our model can "capture“ changes in demand and supply due to weather conditions *indirectly* from the pick-up and drop-off lags. If extreme weather conditions, such as a blizzard or a thunderstorm, occur on a particular day, then the reduced taxi demand in the previous hour will reflect this phenomenon into our model.

4.4.4 Event information

The selected area around Barclays Center shows significant changes in taxi demand mainly after afternoon hours. The cause of most of these intense fluctuations is possibly several events that take place in this popular venue. As mentioned in Section 4.3, the events' dataset includes all records' details that could be scrapped from the official Barclays' s Center website; title, date, and event description. Utilizing the above information, we choose to introduce into our enhanced forecasting model six new parameters that indicate the time position of each hourly pick-up count record in relation to the start time of an event.

More specifically, our new parameters are binary identifiers that indicate if there is an event "*3 hours before*", "*2 hours before*", "*1 hour before*", "*1 hour after*", "*2 hours after*" or "*3 hours after*" the current dataset instance. The start time of venue's events and the date-time information of each entry were used to complete the identifiers' values.

The structure of the updated linear regression model is:

$$\begin{aligned}\hat{Y}_{t+1} = \hat{\beta}_0 &+ \hat{\beta_{1,1}} Y_t + ... + \hat{\beta_{1,12}} Y_{t-12} \\ &+ \hat{\beta_{2,1}} D_t + ... + \hat{\beta_{2,12}} D_{t-12} \\ &+ \hat{\beta_{3,1}} Event_{3h\_bef,t} + ... + \hat{\beta_{3,6}} Event_{3h\_after,t} \end{aligned} \tag{4.6}$$

where $Event_{3h\_bef,t}$, $Event_{3h\_after,t}$ represent the binary variables described above. We excluded variables related with weather, since their contribution was proven insignificant. The same training and test were used for the evaluation and comparison of the updated model structure (see Table 4.4). The improvement rate is small but important in an already well-behaved model. The specific parameters allowed our model to predict demand peaks in a more accurate time position. This conclusion was obtained after visualizing the previous and updated model results in daily plots. A characteristic example is presented in Fig. 4.2. The historical demand average (HA) is represented by the blue dashed line. The hourly predictions of our previous model are depicted with the green dashed line, and the new predictions, enhanced by adding the binary event identifiers, with the red thick dashed line. It is obvious, that the updated values are better placed in time.

4.4.5 Introduction of Topics

In the fourth and final stage of our existing linear regression model optimization we decided to implement a Latent Dirichlet Allocation (LDA) process on the events'

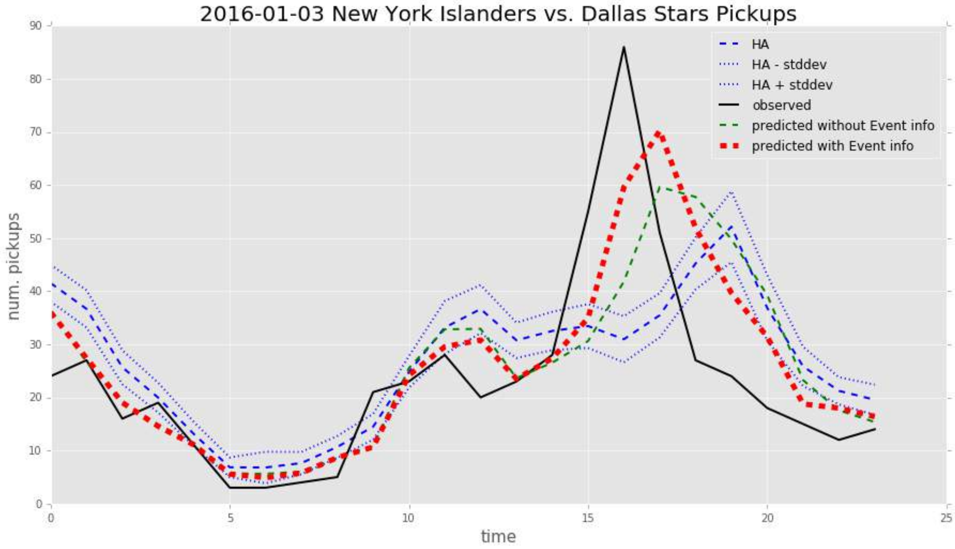


FIGURE 4.2: Prediction results for an event day using only taxi data (green line) and event information (red thicker line).

description. LDA assumes the following generative process:

1. Draw a topic β_k from $\beta_k \sim Dirichlet(\eta)$ for $k = 1...K$
2. For each document d:
   (a) Draw topics proportions θ_d such that $\theta_d \sim Dirichlet(\alpha)$
   (b) For each word $w_{d,n}$:
      i. Draw topic assignment $z_{d,n} \sim \text{Multinomial}(\theta_d)$
      ii. Draw word $w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$

The parameters α and η are hyperparameters that indicate respectively the priors on per-document topic distribution and per-topic word distribution, respectively. Thus, $w_{d,n}$ are the only observable variables, all the others are latent in this model. For a set of D documents, given the parameters α and η, the joint distribution of a topic mixture θ, word-topic mixtures β, topics z, and a set of N words is given by:

$$p(\theta, \beta, z, w|\alpha, \eta) = \prod_{k=1}^{K} p(\beta_k|n) \prod_{d=1}^{D} p(\theta_d|a) =$$

$$= \prod_{n=1}^{N} (p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_k, k = z_{d,n})) \quad (4.7)$$

Broadly speaking, the training task is to find the posterior distribution of the latent variables (the per-document topic proportions θ_d, the per-word topic assignments $z_{d,n}$ and the topics β_k) that maximize this probability.

The parameter that we mainly focused in this study is the number of topics. We tested a range of values between 5 and 30, and we empirically concluded that the value of 10 yielded the best model results. With 10 topics we are able to capture all kinds of events included in our event database, and we also narrow down possible equivocal topics that could deteriorate our results. The other parameters, the α and η priors, were kept as default (1.0/(number of topics)). To confirm this was a safe

Table 4.2: LDA Results

Topic	No. Events	Popular Words
Topic_1	24	ice, disney, present, magic, new
Topic_2	72	basketball, championship, atlantic, game, tournament
Topic_3	10	show, artist, box, office, special
Topic_4	32	music, atlantic, championship, basketball, game
Topic_5	22	game, marriot, corporate, bridge, hotel
Topic_6	10	train, service, islander, view, time
Topic_7	34	tour, album, show, meet, up
Topic_8	12	circus, family, out, space, earth
Topic_9	12	dinner, reservation, jay, menu, restaurant
Topic_10	42	champion, game, group, boxing, hoop

choice, we ran several iterations with different initial α and η priors and they generally converged to similar outcomes. The LDA results are presented in Table 4.2.

Each event record now corresponds to a unique topic assignment, namely a vector of 10 values. Building on the previous 6 binary event identifiers, and for each hourly taxi demand count we added the topic assignment of the event that was planned “*3 hours before*", “*2 hours before*", “*1 hour before*", “*1 hour after*", “*2 hours after*" or “*3 hours after*". Therefore, each aggregated taxi record is enriched with $6 * 10$ new variables, where 6 is the number of the previously defined event identifiers.

For example, if the taxi pick-up count refers to 6pm and from our event dataset we know that there will be an event at 8pm, whose topic assignment vector is $[0, 0, 0.87, 0, 0.01, 0.05, 0, 0, 0.07, 0]$ then the 60 new variables that our model is going to receive are shown in Table 4.3.

Table 4.3: Example of topic modeling introduction

Event identifier	Value [0 or 1]	Topic Variables Assignment
3 hours before	0	[0,0,0,0,0,0,0,0,0,0]
2 hours before	0	[0,0,0,0,0,0,0,0,0,0]
1 hour before	0	[0,0,0,0,0,0,0,0,0,0
1 hour after	0	[0,0,0,0,0,0,0,0,0,0]
2 hours after	1	[0,0,0.87,0,0.01,0.05,0,0,0.07,0]
3 hours after	0	[0,0,0,0,0,0,0,0,0,0]

We provide these apparently redundant features in the context of experimenting with the best model representation and thus with our final best predictor. Through the implementation of many tests, it was found that the optimal performance of the model is achieved by introducing only the last 30 parameters that correspond to the time period after event's start time, namely the identifiers “*1 hour after*", “*2 hours after*" and “*3 hours after*". This is probably justified by the fact that our model already contains information about the time window before the start of the event (by using drop-off lags) and not sufficient information for the time period after its start time. Our model is considerably improved, and the updated values of MAE, RMSE, and R^2 prove it (see Table 4.4).

TABLE 4.4: Linear Regression Results - 6 Months Period

Input Data	MAE	RMSE	R^2
Pick-up Lags	9.733	19.073	0.750
Pick-up & Drop-off Lags	8.772	16.298	0.817
Pick-up & Drop-off Lags & Weather	8.824	16.379	0.815
Pick-up & Drop-off Lags & Event info	8.689	16.024	0.823
Pick-up & Drop-off Lags & Event info & Topics	8.301	14.932	0.847

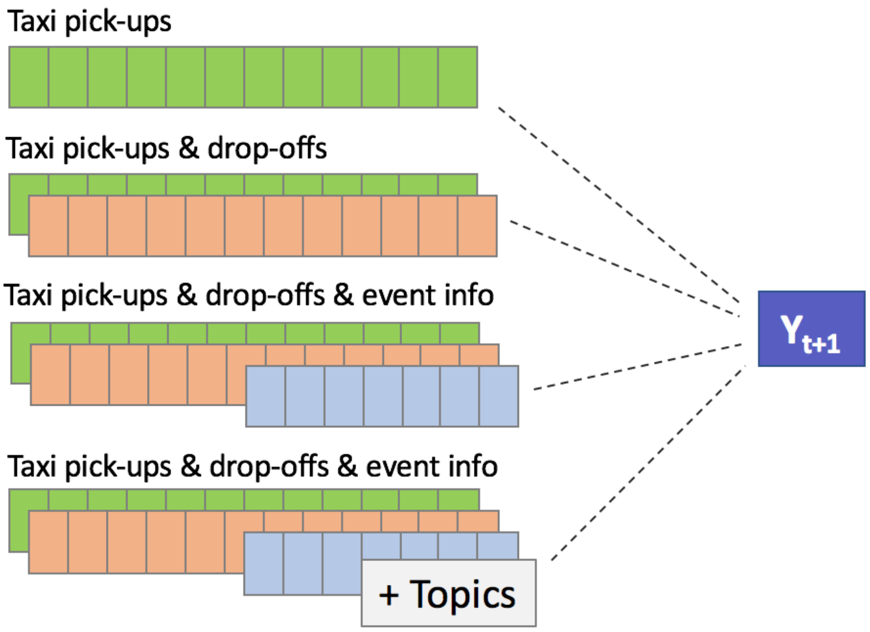

Taxi pick-ups
Taxi pick-ups & drop-offs
Taxi pick-ups & drop-offs & event info
Taxi pick-ups & drop-offs & event info
+ Topics
Y_{t+1}

FIGURE 4.3: Proposed models

4.4.6 Model evaluation using only event days

To summarize, from the previous stage of our analysis, four models were developed (see also Fig. 4.3) :

- The **baseline model** using only pick-up lags
- The **pick-up – drop-off model** which is enriched with drop-off lags
- The **event identifier model**, where six binary event identifiers were introduced
- The **topic model** with additional 30 topic variables based on LDA's results.

Fig. 4.4 shows how the $R^2 Score$ is changing for the first 20 days of January 2016. The red bars correspond to the performance of our baseline model, the yellow bars to the third model with event information and the green bars corresponds to our full model with topics. The black dots note on which days we have an event ("0.5“ = event day, "0“ no-event day). Our final model is able to improve the quality of its predictions quite significantly. On days where there is no event, the difference between our third and fourth (full) model is minimal, since our model has all the

Table 4.5: Linear Regression Results - Event Period

Input Data	MAE	RMSE	R^2
Pick-up Lags	11.037	22.480	0.742
Pick-up & Drop-off Lags	9.840	18.939	0.817
Pick-up & Drop-off Lags & Weather	9.819	18.923	0.817
Pick-up & Drop-off Lags & Event info	9.597	18.230	0.830
Pick-up & Drop-off Lags & Event info & Topics	8.585	15.837	0.872

necessary information concerning events from our binary variables.

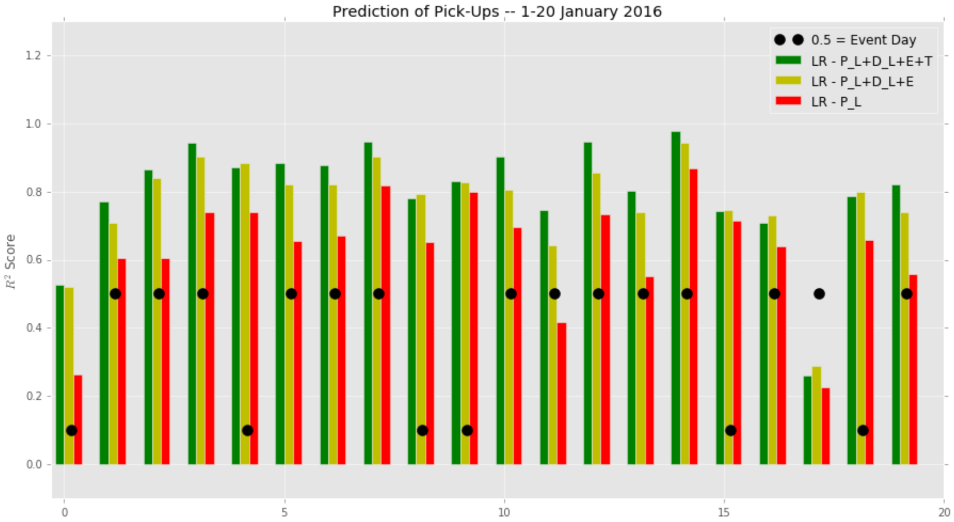


Figure 4.4: LR performance in January 2016

Each model is an enhanced version of the previous one. The proposed architecture performs respectively well on event days. Table 4.5 summarizes those results. The exploitation of drop-off parameters improves demand predictions by 10.11%. This significant improvement proves once again the importance of monitoring traffic network's conditions before an event starts. The dummy variables of the third model have small but positive effect on the final result, while topics' impact is once again significant (5% further improvement on an already fairly reliable model).

4.4.7 Gaussian Processes for demand prediction

Besides these baselines, the proposed approach is further compared with another popular method from the state of the art for time-series forecasting, the Gaussian processes (GPs) (Rasmussen, 2004). We chose new independent training, validation and test sets, since GPs implementation is a computational demanding process. The training set consists of 6 months of observations (January - June 2015), the validation set contains the first two months of 2016 (January-February) and the remaining four months of the first half of 2016 were used for testing. The hyper-parameters of GPs, namely the length-scale parameter of the Radial-basis function (RBF) kernel, and the Tikhonov regularization of the assumed covariance between the training points ($alpha$) (Seeger, 2004), were tuned based on their performance on the validation set.

An exhaustive search procedure of all possible combinations within a predefined set of possible value parameters was used.

Table 4.6: LR and GP Comparison - 4 months period

	Barclays Center						Terminal 5					
	Linear Regression			Gaussian Processes			Linear Regression			Gaussian Processes		
Input Data	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
Pick-up Lags	9.523	18.733	0.756	9.522	18.675	0.757	10.606	19.215	0.750	10.771	18.905	0.758
Pick-up & Drop-off Lags	8.620	16.280	0.815	8.669	16.378	0.813	8.928	15.409	0.840	9.095	15.452	0.839
Pick-up & Drop-off Lags & Weather	8.623	16.289	0.815	8.670	16.377	0.813	8.938	15.421	0.838	9.086	15.448	0.839
Pick-up & Drop-off Lags & Weather & Event info	8,577	16.369	0.813	8.571	16.396	0.813	8.828	14.969	0.849	9.070	15.380	0.840
Pick-up & Drop-off Lags & Weather & Event info & Topics	8.258	15.656	0.829	8.100	14.979	0.844	8.935	15.254	0.843	8.206	13.095	0.864

Along with GPs implementation, the corresponding forecasts of linear regression models were also evaluated. Both machine learning techniques use the same time intervals for training and testing. Furthermore, more experiments were implemented using Terminal 5's datasets. The final results are presented in Table 4.6.

There are several very interesting aspects that are worth discussing. From the first three models, it is concluded that the performance of LR and GPs is similar. A remarkable change appears though with the introduction of topics. For Barclays Center both forecasts are improved, something that does not seem to be confirmed for Terminal 5. More specifically, for the first venue the LR with topics gives a final R^2 of 0.829, and an improvement of almost 2%, while for the second venue the final R^2 is 0.843, with a deterioration of 0.5%. On the other hand, the GPs seem to respond positively to all imported parameters. With the introduction of topics, the R^2 is increased by 3% for Barclays Center and by 2.9% for Terminal 5. The GP model emerges as the ideal technique for our real-time forecasting model. Once again, the weather did not make a significant contribution.

The same comparison is implemented using the event periods. Table 4.7 summarizes the final results. GPs prove to be the best performing approach for our predictions. By using topics, the proposed forecasting model reduces its error for both venues. These results clearly highlight how crucial data fusion of time-series data and semantic information can be, in particular for the problem of predicting taxi demand in event areas considered in this paper.

Lastly, from the perspective of transportation practitioners, it is important to note that by exploiting event information automatically extracted from the Web and by developing two regression models for combining this information with historical time-series data, we were able to reduce prediction error in event areas quite dramatically in both study areas. In the case of the Barclays Center area, we started with an initial R^2 of 0.757 using pick-up lags and we were able to obtain a R^2 of 0.844 by using the full GP model. Likewise, in the case of the Terminal 5 area, we started with a MAE of 10.77 using pick-up lags and reached a MAE of 8.21 for the full model. These are very significant improvements that emphasize the importance of accounting for the effect of special events when forecasting mobility demand in dynamic and lively urban areas.

Following the performance results presented above, a comparison of our real-time approach with a previous study on daily demand forecasts (Rodrigues, Markou, and Pereira, 2019) is implemented and presented in Table 4.8. The comparison refers to the Barclays Center, and the scores presented in the left part of the table under the title “Daily Demand Prediction", were included as documented in the stated article. GPs were evaluated in both cases, as well as the incorporation of pickups, drop-offs and events' occurrence information.

It is quite obvious that our real-time approach performs significantly better than the daily demand prediction framework. The noteworthy deviation of accuracy is mainly due to the significant enhancement of forecast precision that the hourly pick-up and drop-off lags offer to our real-time approach. In those cases where the real-time model was not able to predict a significant change in demand at a particular time, then it will somehow adjust its prediction at the next time-step. Therefore the overall aggregated daily demand forecast will not deviate significantly from the actual value. Additionally, the significance of weather data in the first daily model is noteworthy. As we mentioned earlier, weather information is available only per day and not per hour of the day. Therefore, daily demand and supply fluctuations can be explained, and model's forecast accuracy can be enhanced due to this parameter, since we do not have any other information about the state of the transport system throughout the day (as for the real-time model). Finally, it is worth mentioning that for the daily demand prediction, information about events' occurrence on that particular day plays a decisive role in the accuracy of the model, since we get a better anticipation of the increased demand, while for the real-time model, indications of high demand can be better perceived by the hourly pick-up and drop-off lags.

Table 4.7: LR and GP Comparison - Events' period

	Barclays Center						**Terminal 5**					
	Linear Regression			**Gaussian Processes**			**Linear Regression**			**Gaussian Processes**		
Input Data	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
Pick-up Lags	10.395	20.797	0.753	11.546	22.864	0.702	13.469	23.997	0.684	13.654	23.909	0.686
Pick-up & Drop-off Lags	9.425	18.246	0.810	10.065	19.357	0.786	11.458	19.667	0.788	11.270	19.650	0.788
Pick-up & Drop-off Lags & Weather	9.393	18.265	0.810	9.858	18.801	0.799	11.480	19.615	0.789	11.328	19.718	0.787
Pick-up & Drop-off Lags & Weather & Event info	9.569	18.875	0.797	9.844	18.789	0.799	11.554	19.487	0.792	11.214	19.552	0.790
Pick-up & Drop-off Lags & Weather & Event info & Topics	9.236	18.141	0.812	8.945	17.235	0.830	11.539	19.462	0.793	10.912	18.996	0.799

Table 4.8: Real-Time and Daily Demand Prediction Comparison

	Daily Demand Prediction						**Aggregated Daily Demand Prediction**					
	SVR			**Gaussian Processes**			**Linear Regression**			**Gaussian Processes**		
Input Data	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
Pick-up & Drop-off Lags	120.7	167.5	0.428	145.2	199.4	0.189	48.012	61.763	0.936	71.1	102.2	0.826
Pick-up & Drop-off Lags & Weather	120.1	166.4	0.436	127.3	176.9	0.362	48.0	61.9	0.936	70.9	102.0	0.827
Pick-up & Drop-off Lags & Weather & Event info	97.9	139.1	0.605	100.0	141.1	0.594	45.5	59.2	0.942	71.1	102.2	0.826

4.4.8 Deep Gaussian Processes for demand prediction

The architecture of deep Gaussian process that Damianou and Lawrence Damianou and Lawrence, 2013 formulated is presented in Fig. 4.5. It corresponds to a graphical model with three kinds of nodes: the leaf nodes $Y \in R^{N \times D}$ which are observed, the intermediate latent spaces $X_h \in R^{N \times Q_h}$, $h = 1, ..., H-1$, where H is the number of hidden layers, Q is the number of latent dimensions, and the parent latent node $Z = X_H \in R^{N \times Q_z}$. In the proposed deep architecture, all intermediate nodes X_h act as inputs for the right layer and as outputs for the left layer. So, when we have a simple structure with only two hidden units, the generative method is formulated as follows:

$$y_{nd} = f_d^Y(x_n) + \epsilon_{nd}^Y, \quad d = 1, ..., D, \quad x_n \in R^Q \tag{4.8}$$

$$x_{nq} = f_q^X(z_n) + \epsilon_{nq}^X, \quad q = 1, ..., Q, \quad z_n \in R^{Q_z} \tag{4.9}$$

and the intermediate node is involved in two Gaussian processes, f^Y and f^X, playing the role of an input and an output respectively: $f^Y \sim \boldsymbol{GP}(0, k^Y(X, X))$, and $f^X \sim \boldsymbol{GP}(0, k^X(Z, Z))$. More layers could be added, thus increasing the number of significant model parameters, and system's complexity. For further details concerning Deep GPs' generative process please refer to the original article of Damianou and Lawrence (Damianou and Lawrence, 2013).

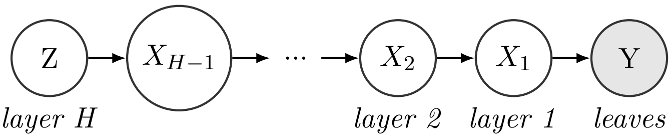


FIGURE 4.5: Representation of a Deep GP model with a cascade of H hidden layers

It has been argued that deep networks have more representational power compared to the *simpler/shallow* ones. However, Duvenaud et al. noted that many hidden layers will not always result in a better model (Duvenaud et al., 2014). An extreme expression of this effect can result in a pathology and, therefore, very deep networks are not always preferable. In our case, the demand fluctuations do not have very complicated patterns. In the venue areas studied in this case, smooth demand changes are observed during the morning and afternoon hours and in days when we have one or two big events, one or two clear demand peaks are detected (see for example Fig. 4.2). Therefore, our deep network will not necessarily perform well with many Gaussian processes involved.

After a sufficient number of tests, it is concluded that if our developed model has more than two hidden layers, the prediction results worsened considerably. Therefore, to compare Deep GPs with the previously described model, we developed the architecture depicted in Figure 4.6. We implemented a simple architecture with two Gaussian processes both with Radial Basis Function kernels in the context of supervised learning. We have a single hidden layer, since the top layer is observed. The kernel length values, as well as the number of latent dimensions of the hidden layer are defined using Bayesian Optimization, a sequential design strategy for global optimization of functions' hyperparameters (Snoek, Larochelle, and Adams, 2012).

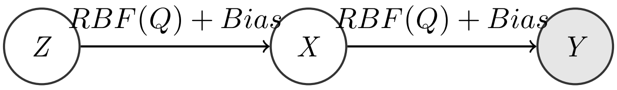


FIGURE 4.6: Implemented Deep GPs architecture

We thoroughly studied travel demand predictions for the month of June 2016, since we had distinct demand differences between weeks and we wanted to see how each methodology responds. More specifically, for Terminal 5 we had a week with no events scheduled and a week with four very popular concerts in its calendar. We therefore considered these cases to be representative for the evaluation purposes of this study. The previous 7 months (5000 instances) were chosen as the training set for each method. We wanted to keep the training set's size manageable due to the computational requirements that GPs and subsequently Deep GPs have.

Table 4.9: LR GPs and DeepGPs Comparison - Terminal 5 - No events

	Deep GPs			Linear Regression			Gaussian Processes		
Input Data	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
Pick-up Lags	10.631	13.987	0.705	10.150	13.341	0.732	7.273	10.372	0.838
Pick-up & Drop-off Lags	9.517	12.912	0.749	6.496	8.606	0.888	6.059	8.153	0.900
Pick-up & Drop-off Lags & Weather	9.612	13,041	0.742	6.558	8.672	0.879	6.120	8.235	0.892
Pick-up & Drop-off Lags & Weather & Event info	9.001	12.229	0.774	6.490	8.607	0.888	6.334	8.598	0.888
Pick-up & Drop-off Lags & Weather & Event info & Topics	8.972	12.163	0.777	6.497	8.632	0.888	6.336	8.599	0.888

In Tables 4.9 and 4.10 we summarize the prediction results for a week with no events and four events respectively. In the first case, it is clear that LR and GPs have similar performance. GPs are able to provide acceptable predictions with just the pick-up lags, which highlights their ability to better adapt to fluctuating demand. However, Deep GPs performance is considerably lower than the other two methods. However, by observing their performance in Table 4.9, it appears that adding more parameters and subsequently more information about events and their characteristics helps them identify the different patterns between days with events and days with no events and improve their performance by 10.21% compared to the baseline model with only pickup lags. In the second case, where we have a week with four events, it appears that Deep GPs are not able to identify correctly the demand peaks. In Figure 4.7 we can see that Deep GPs predict days with average demand even though an event is organized, and in two cases there is a forecast for high demand during the afternoon hours with no event scheduled. On the other hand, LR and GPs performance is better as shown in Table 4.10. The introduction of event and topic parameters shows once again demand peaks' effective recognition enhancement.

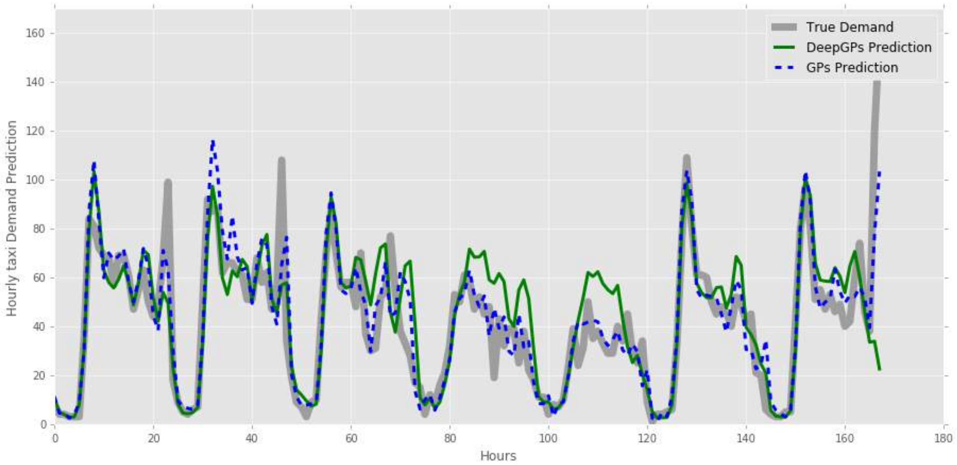


Figure 4.7: Demand Prediction using GPs and DeepGPs

4.4.9 Demand forecasting using Fully-Connected Layers

The data fusion architecture that makes use of fully-connected (FC) layers for modeling the time-series data is depicted in Figure 4.8. All the time-series information is provided as a flat input vector to the network in the form of lagged information, following the implemented methodology steps presented in the previous methods. The network is fed with the values for the observations at times $\{t, t-1, ..., t-L\}$ in a

TABLE 4.10: LR GPs and DeepGPs Comparison - Terminal 5 - Four events

	Deep GPs			Linear Regression			Gaussian Processes		
Input Data	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
Pick-up Lags	11.211	18.694	0.544	11.700	17.804	0.586	9.164	15.932	0.698
Pick-up & Drop-off Lags	11.279	18.739	0.542	8.332	13.462	0.764	8.701	13.265	0.770
Pick-up & Drop-off Lags & Weather	11.392	18.926	0.539	8.415	13.597	0.757	8.788	13.398	0.763
Pick-up & Drop-off Lags & Weather & Event info	11.203	18.706	0.544	8.499	13.407	0.766	8.698	13.259	0.771
Pick-up & Drop-off Lags & Weather & Event info & Topics	11.233	18.717	0.543	8.504	13.277	0.770	8.473	12.413	0.799

vector of size $L+1$, where $L+1$ corresponds to the number of lags. This vector is fed into a FC layer with 100-200 hidden units and hyperbolic tangent (tanh) activations, which also can receive additional inputs with other relevant information, such as the event details described in previous subsections. The output of this FC layer is then passed to a second FC layer with 50 units and tanh activations. We apply Batch-Normalization (Ioffe and Szegedy, 2015) before every FC layer, Dropout between FC layers and we use regularization whenever necessary.

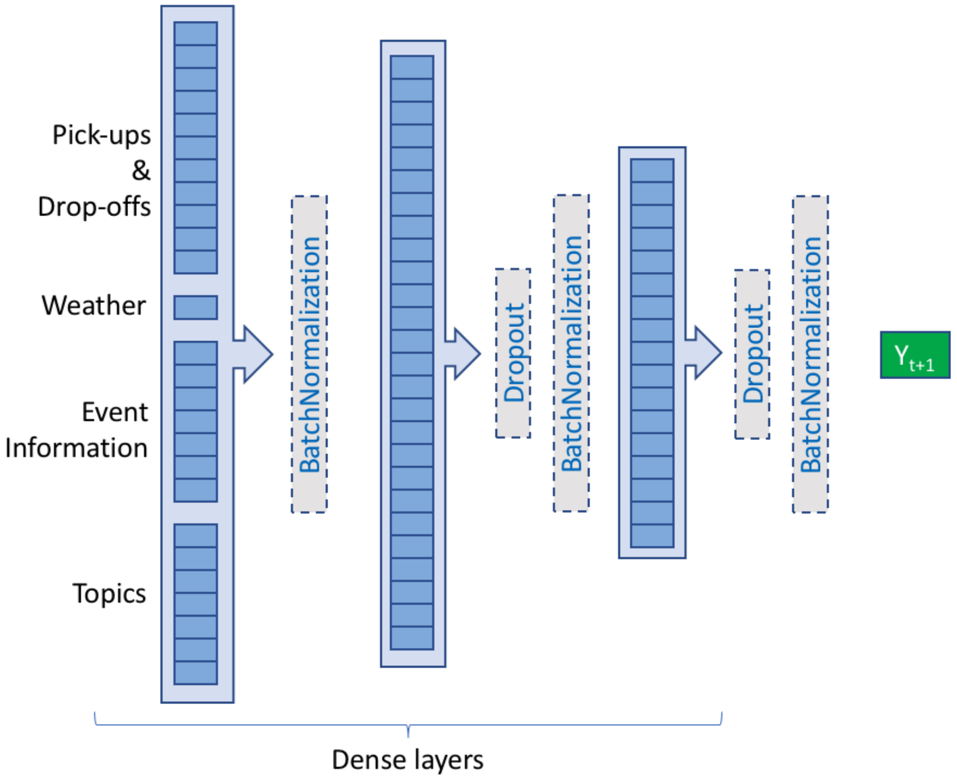


FIGURE 4.8: Proposed neural network architecture with FC layers

The idea is that the output of the last FC layer corresponds to a latent vector representation that encodes all the necessary information form the time-series and other relevant inputs. From this latent vector representation we will finally produce a prediction for $t+1$ using a dense layer. The final prediction is obtained by adding back the removed recurrent trend (based on the historical average) to the output of the neural network.

We kept the same training set of instances as in Deep GPs. For the validation

TABLE 4.11: LR, GPs, Deep GPs and DL-FC Model Comparison - Terminal 5 - No events

	Linear Regression			**Gaussian Processes**			**Deep GPs**			**DL-FC**		
Input Data	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
Pick-up Lags	10.150	13.341	0.732	7.273	10.372	0.838	10.631	13.987	0.705	7.131	9.826	0.854
Pick-up & Drop-off Lags	6.496	8.606	0.888	6.059	8.153	0.900	9.517	12.912	0.749	6.323	8.399	0.894
Pick-up & Drop-off Lags & Weather	6.558	8.672	0.879	6.120	8.235	0.892	9.612	13.041	0.742	6.450	8.567	0.876
Pick-up & Drop-off Lags & Weather & Event info	6.490	8.607	0.888	6.334	8.598	0.888	9.001	12.229	0.774	7.306	10.155	0.850
Pick-up & Drop-off Lags & Weather & Event info & Topics	6.497	8.632	0.888	6.336	8.599	0.888	8.972	12.163	0.777	7.298	10.123	0.851

TABLE 4.12: LR, GPs and DL-FC Model Comparison - Terminal 5 - Four events

	Linear Regression			**Gaussian Processes**			**Deep GPs**			**DL-FC**		
Input Data	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
Pick-up Lags	11.700	17.804	0.586	9.164	15.932	0.698	11.211	18.694	0.544	9.406	15.629	0.681
Pick-up & Drop-off Lags	8.332	13.462	0.764	8.701	13.265	0.770	11.279	18.739	0.542	8.133	12.160	0.774
Pick-up & Drop-off Lags & Weather	8.415	13.597	0.757	8.788	13.398	0.763	11.392	18.926	0.539	8.296	12.403	0.759
Pick-up & Drop-off Lags & Weather & Event info	8.499	13.407	0.766	8.698	13.259	0.771	11.203	18.706	0.544	7.071	10.971	0.822
Pick-up & Drop-off Lags & Weather & Event info & Topics	8.504	13.277	0.770	8.473	12.413	0.799	11.233	18.717	0.543	6.744	10.563	0.835

of the proposed architecture, we isolate a separate set of instances. Through this procedure, we keep track of the best performing model during training and we adjust accordingly the majority of the design and hyperparameter choices that the proposed deep learning architectures reflect. The presented architecture was implemented in Keras (Chollet, 2015).

The comparison of the proposed methodologies - architectures is concluded with the results of our DL-FC model in Tables 4.11 and 4.12. Interestingly, the results show an apparent superiority of the DL-FC model's predictions for the week with four events. With an R^2 of 0.835, we conclude that the FC layers can provide us with fairly satisfactory forecasts of days with high demand peaks as well as forecasts of days with demand close the typical daily patterns. Finally, we can once again verify that including information about the events leads to a MAE reduction of 13.1% and 10.8% in RMSE. Including the topics in the DL-FC model allows it to further reduce MAE by another 4.6% and to increase the average R^2 form 0.822 to 0.835.

4.5 Conclusion

We demonstrated that using online information, we can improve the quality of taxi demand prediction even in scenarios where the transport system is under stress. We combined information extracted from the web with time-series data to build a predictive model of taxi demand around special event venue areas. This is typically a challenging case for transport planning since special events originate high variance in demand. Taxi demand is correlated with many parameters of underlying information and currently, most taxi centers rely on formal processes and manual work for a fleet organization and taxi distribution. Even the more advanced new services, like Uber or Lyft, still face great challenges in terms of demand prediction (shown by *price surges* and variations thereof). Our results show that by making use of event information from the Web, the proposed models are able to improve the quality of their predictions quite dramatically, thus significantly outperforming other popular time-series forecasting methods from the state of the art that do not account for event information. The proposed approaches make use of the topic modeling advantages that allow the introduction of useful details concerning events' description in our model and from our empirical results the need for accounting events' effect when modeling mobility demand is highlighted.

The initial structure of the model was presented and evaluated using machine learning techniques. Then, each method's performance was compared with a deep learning architecture that is based on fully-connected dense layers, as well as with Deep GPs architecture. In most cases, Gaussian Processes managed to better adapt to the studied venues' demand patterns. Every piece of information, except weather data, contributed positively to the results, thus emphasizing their importance for the formulation of a hourly demand prediction model around a venue.

Noteworthy was also the accuracy of the deep learning architecture. Their predictions' quality improvement is remarkable by making use of event information. At time periods with various events, the presented methodology manages to overcome the alternative models.

In future work, we aim at exploring the impact of spatio-temporal interactions on taxi demand prediction. The development of a city-wide spatio-temporal model that accounts for information about all the events that take place across the city could be a generalization potential of this methodology.

Chapter 5

A two-stage model for real-time taxi demand prediction using data from the Web

5.1 Introduction

In Chapter 4 a real-time taxi demand prediction model was formulated, utilizing historical time-series data and topic modelling based variables. Model's performance was evaluated through several machine learning techniques, and the results were very promising. In this chapter, we will explore how we will examine how we can further improve the performance of the model, changing its structure, and making it a little more complex.

In general, mobility trends captured in complex transport systems consist of two basic components: utilitarian travel that mostly includes habitual behavior (e.g. commuting to work, weekly shopping) but also to a minor extent non-habitual needs (e.g. go to hospital, occasional shopping); and recreational travel, which comprises the human need for entertainment, social interaction and public expression. Efficient and effective intelligent transport systems should be able to take into consideration both of these factors for accurate demand predictions and better traffic management.

Current prediction approaches generally focus on capturing recurrent conditions, namely their seasonal spatial-temporal aspects (the “average” winter peak-hour Monday, in area X, with weather Y). The developed approaches can be successful for long-term planning applications or for modeling demand in non-eventful areas such as residential neighborhoods. However, in lively and dynamic areas where multiple special events take place, such as music concerts, sports games, festivals, parades and protests, these approaches fail to accurately model mobility demand precisely at times when it is needed - when the transport system of the area is under stress. The inability of the system to meet the new demand conditions emphasizes the need of good anticipatory capabilities which are capable to accept timely information on such phenomena.

Non-recurrent special events, such as concerts, sport games and demonstrations, are planned and largely advertised on the Web. An interesting fact is that it is much more likely to have citizens sharing their expectations/experiences about non-recurrent events than to talk about their daily commute. This plethora of information makes the Web an important tool for demand prediction and thus system's balance maintenance.

Previous studies have shown a strong correlation between number of public transport arrivals with the structured data mined from the Web (Pereira, Rodrigues, and Ben-Akiva, 2012; Pereira et al., 2015). Namely, semi-structured information about events from announcements websites can be used as features for public transport arrivals. However, information contained on these websites is usually incomplete, noisy or missing, which makes it difficult to generalize. Going beyond this approach raises two challenges: which details about a scheduled event (time, type of event) are useful and how relevant information can be turned into model.

The aim of this chapter is the exploitation of information available on the internet for real-time demand prediction using a two-stage model. A particular emphasis will be given to venue areas, where several special events that are publicly disclosed on social media are hosted and attract many people. The proposed framework will be able to predict intervals of high demand that the average supply of the studied transport system (taxi services, Uber etc.) cannot easily cover.

5.2 Literature Review

5.2.1 Internet as a data source for special events

Internet, and more specifically the several social networking services that exist, has become a popular distribution outlet for users looking to share their experiences and interests on the Web. Taking as an example the Facebook, which has over 2.19 billion monthly active Facebook users (Facebook MAUs) worldwide, it is clearly understood that the information derived from the above platforms, can undeniably help discerning explanations about observed real-world phenomena, such as non-habitual overcrowding scenarios. Due to the importance of special events' impact in urban mobility, it is not surprising that they are a predominant part of transportation research. Fortunately, the Internet is rich in information about public special events. In an earlier work, Pereira et al. compared an origin/destination (OD) prediction model based on public transport data with and without simple information obtained from the Internet, such as event type or whether the performer/event had a Wikipedia page (Pereira, Rodrigues, and Ben-Akiva, 2012). It was verified that such information could reduce the root mean squared error (RMSE) by more than 50% in each OD. In another study, Pereira et al. presented a machine learning model that classifies aggregated crowd observations into explanatory components (Pereira et al., 2015). After the identification of overcrowding hotspots in the city-state of Singapore, potential explanations from several event announcements websites were retrieved. The internet is also a valuable source for other aspects of mobility research. For example, Twitter has been used for crisis management (Thom et al., 2012; Sakaki, Okazaki, and Matsuo, 2010), urban management and planning (Frias-Martinez et al., 2012), the analysis of different aspects of mobility (Cheng et al., 2011) and the mobility characteristics of different nations (Hawelka et al., 2014). Due to the complexity of the exploration of the open Web (e.g. using Google search), the use of internet data in transportation, however, is currently limited to manually defined sources and highly fine-tuned processes.

5.2.2 Demand Prediction for special events

Special events have a huge impact in urban mobility, regardless of their scale and type. Understanding their influence on the balance of a transport system is crucial for the development of reliable traffic management operations. For large-scale events (e.g. World cup, Formula One and Olympic games), best practices are already available for authorities to follow in order to manage these events and prepare for them well in advance (Dunn Jr, Latoski, and Bedsole, 2006; Coutroubas and Tzivelou, 2003). However, these manual approaches do not scale to the vast amount of smaller and medium-sized events that take place on large metropolitan areas on a daily basis. Despite their reduced scale, these events still have a significant impact in the transportation system (Pereira et al., 2015), especially when multiple co-occur. In these scenarios, common practice relies on reactive approaches rather than on planning (Fuhs and Brinckerhoff, 2010; Kuppam et al., 2011). The demand prediction solution that we propose in this paper, takes into consideration event information that is automatically mined from the Web, and present itself with the potential for anticipating the effects of events and showing reliable tools for hotspot predictions in eventful areas.

Taxi demand has been the subject of several applications, since the related datasets are sufficiently detailed. The yellow and green taxi public dataset of New York City in particular, has been the subject of a lot of research. Morgul and Ozbay present an empirical assessment of taxicab drivers' labor supply (Morgul and Ozbay, 2015). Yang and Gonzales identify locations and times of day where there is a mismatch between the availability of taxicabs and taxi service demand (Yang and Gonzales, 2017). Zhao et al. use entropy and the temporal correlation of human mobility to measure the demand uncertainty at the building block level (Zhao et al., 2016). They implemented three prediction algorithms to validate their maximum predictability theory. The importance of identifying hotspots, where demand is expected to be higher than the expected average demand is highlighted in the research of Markou et al. (Markou, Rodrigues, and Pereira, 2017). Through kernel density analysis, demand fluctuations were detected and analysed and significant deviations from the average day were correlated with disruptive event scenarios such as extreme weather conditions, public holidays, religious festivities, and parades. Finally, some other research studies used this taxicab data to explore taxicab driver's airport pick-up decisions (Yazici, Kamga, and Singhal, 2013) or travel time variability analysis (Kamga and Yazıcı, 2014).

5.3 Deep models in transportation

Deep learning is evolving rapidly in solving problems that have resisted the best attempts of the artificial intelligence community for many years. It has proven to be able to find intricate structures in high-dimensional data, and thus it is an important tool in various applications in the domain of science (18). In the field of transportation and urban mobility there are already studies showing deep learning's successfullness.

Lv et al. proposed a deep-learning-based traffic flow prediction method that takes into consideration the traffic flow features as learned by a stacked autoencoder model (SAE) (Lv et al., 2015). Results comparison with more traditional approaches based on Support Vector Regression (SVR) and radial basis functions (RBFs) showed proposed method's superiority. Ma et al. proposed a long short-term memory neural (LSTM) network for travel speed prediction (Ma et al., 2015). Their empirical results on data from Beijing indicate that LSTMs outperform other methods such ARIMA and SVR, which the authors justify with the ability of LSTMs to capture long-term dependencies over the time-series. A model with Mixture Density Networks (MDN) on top of LSTM was proposed by Xu et al. (Xu et al., 2017). In their approach, the city is previously divided in smaller areas and then the LSTM-based model is used to jointly predict the taxi demand for the next time-step in all the areas. Finally, the prediction of crowds' traffic in city's regions using a deep-learning based approach, called ST-ResNet, is presented by Zhang et al. (Zhang, Zheng, and Qi, 2017b). Experiments on two types of crowd flows in Beijing and New York City (NYC) demonstrate that the proposed method outperforms standard approaches such as ARIMA and vector auto-regressive models.

While the approaches described above demonstrate the potential of deep learning for transportation problems, none of these approaches consider the effect of events in order to improve their predictions. Markou et al. took into consideration these details using topic modeling and formulated a real-time taxi demand prediction model that successfully predicts significant demand peaks around venues (Markou, Rodrigues, and Pereira, 2019). Linear Regression, Gaussian Processes, Fully-Connected dense layers and Deep Gaussian Processes were evaluated and compared in the context of finding the most appropriate methodology that satisfies immediate congestion management requirements. Finally, the combination of time-series and textual data using word embeddings and convolutional layers for daily demand prediction has also proven to be significant (Rodrigues, Markou, and Pereira, 2019).

This study aims at presenting a two-stage framework that recognizes conditions of greater prediction uncertainty, and thus improves the final demand forecasts by taking into consideration the time-series of demand forecast's residuals, that were observed the near past.

5.4 Methodology

From previous research, we have already highlighted the importance of textual data for more accurate daily forecasts in event areas (Rodrigues, Markou, and Pereira, 2019). The proposed neural network architectures lead to significant reductions in forecasting error using event information extracted from the Web. In this study, having at our disposal all possible information for future events, we present how data fusion can also be very useful on forecasting the error of our neural network architecture and thus on the even greater performance of our final model. Our focus is taxi demand prediction in real-time.

5.4.1 Model formulation

The proposed structure of the prediction model architecture includes two phases, (a) the training phase and (b) the test phase. In the first phase, only the first forecasting model (referred to as "Demand Prediction Model" – "**DP-Model**") is used, whose architecture is presented in the next subsection. The main objective of our DP-Model is to predict taxi demand based on the available historical data for the areas that we are interested in.

At the second phase, we use the forecasts we received at the end of the training phase, for the calculation of the **DP-Model**'s forecast deviation from the actual demand. The obtained residuals that correspond to the previous timeframe are used as the training dataset of our second deep learning model (referred to as "Error of Demand Prediction" - "**EoDP-Model**"), whose independent variables include the day of the week, day of month, topics and dummy variables that represent information about the presence of events before or after the predicted hourly demand. The objective of the **EoDP-Model** is the estimation of the demand prediction residuals based on the calculated residuals that the **DP-Model** attributed to that particular day of the month/week in the past, where an event was or was not scheduled.

At each stage we use separate training, validation and test sets and there are no time periods overlap between the two phases.

To elaborate:

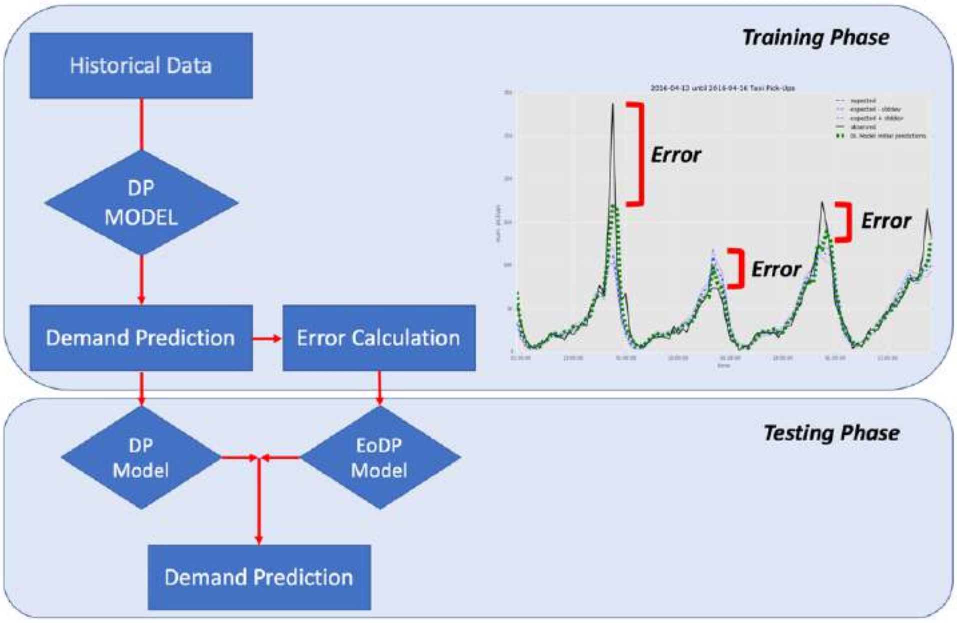


FIGURE 5.1: Proposed methodology

- The training phase is subdivided into two time periods. During the first time period (A Period) the **DP-Model** runs independently and gives demand predictions for the second time period (B Period). These predictions are evaluated based on true demand values that were observed during the B period, and the vector of residuals (predicted − true_values) is obtained. That vector will be the training dataset for our **EoDP-Model**.

- the testing phase refers to a new time period (C Period), for which the **DP-Model** gives demand predictions and the **EoDP-Model** gives future residuals estimations. The final outcome is the sum of those predictions.

5.4.2 Time series detrending

One of the most important steps in preparing time-series data for analysis is detrending (25). For the particular case of urban mobility applications, such as traffic flow forecasting and taxi demand prediction, there are obvious cyclic trends that result from daily commuting and other recurrent behaviors.

Before the configuration of initial modeling structure, we decided to remove any deterministic trends and focus our analysis on the remaining fluctuations. A simple, yet very effective way, of identifying these daily or weekly recurring patterns is by constructing a historical averages model, which computes the individual averages for each (hour of day, day of the week) pair based on historical data (from the train set only). The historical averages then represent a fixed recurring trend, which can be easily removed from the data.

5.4.3 Deep Learning Architecture

The data fusion architecture that makes use of fully-connected (FC) layers for modeling the time-series data is depicted in Figure 5.2. All the time-series information is provided as a flat input vector to the network in the form of lagged information. The network is fed with the values for the observations at $\{t, t-1, ..., t-L\}$ in a vector of size $L + 1$, where $L + 1$ corresponds to the number of lags. This vector is fed into a FC layer with 100-200 hidden units and hyperbolic tangent (*tanh*) activations, which also can receive additional inputs with other relevant information, such as the event details described in previous subsections. The output of this FC layer is then passed to a second FC layer with 50 units and *tanh* activations. We apply *BatchNormalization (24)* before every FC layer, Dropout between FC layers and we use regularization whenever necessary.

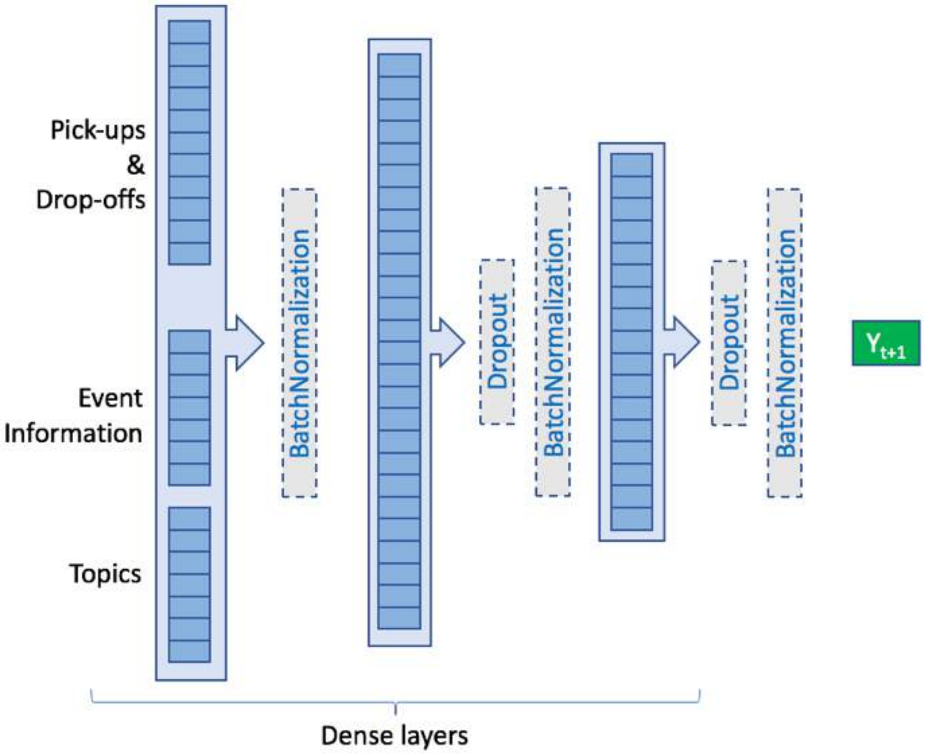


FIGURE 5.2: Proposed neural network architecture with FC layers (1^{st} DL Model).

The idea is that the output of the last FC layer corresponds to a latent vector representation that encodes all the necessary information form the time-series and other relevant inputs. From this latent vector representation, we will finally produce a prediction for $t + 1$ using a dense layer. The final prediction is obtained by adding back the removed recurrent trend (based on the historical average) to the output of the neural network.

5.4.4 Text data pre-processing

Generally, textual data mined from the Web is difficult to process in its original state. Specific pre-processing steps are usually required in order to make it more amenable to learning methods, and more specifically to the topic modelling stage that will follow. Therefore, we follow a simple conventional text-processing pipeline consisting of:

- HTML tag removal
- Lowercase transformation for words' variability restriction purposes

- Tokenization, a tool that divides a sequence of characters into pieces of tokens
- Lemmatization for inflectional endings removal, and words return to their base form (lemma)
- Stopwords and very frequent words removal, which typically do not bring any additional useful information
- Removal of words that appear only once in the whole dataset

5.4.5 Topic Modeling

A considerable amount of important information about a planned event is in textual form. Adding to other structured information, such as date, time and location, we can find useful details concerning its content in the description, title, comments on the website hosting the announcement. To obtain an automated system, we still need to convert such data into a proper representation that a machine learning can understand. However, the dimensionality of the machine learning model will be increased beyond reasonable if we explicitly include the text, word by word. Natural language is rich in synonymy and polysemy, different announcers and locations may use different words, besides it is not always obvious which words are more “relevant". Topic modeling is the research topic that focuses on covering these weaknesses.

The approach of topic modeling is to represent a text document as a finite set of *topics*. These topics correspond to sets of words that tend to co-occur together rather than a single word associated with a specific topic. For example, a rock festival textual description could have a weight w_1 assigned to topic 1 (e.g. words related to concerts in general), w_2 of topic 2 (e.g. words related to festivals), w_3 of topic 3 (e.g. words related to the venue descriptions) and so on. In particular, we use a specific technique that is called Latent Dirichlet Allocation (LDA). For the readers that are familiar with Principal Components Analysis (PCA), there is a simple analogy: PCA re-represents a signal as a linear combination of its eigenvectors, while LDA re-represents a text as a linear combination of topics. In this way, we reduce the dimensionality from the total number of different words of a text to the number of topics, typically very low. Each document is represented as a distribution over topics, and each topic is a distribution over words.

LDA’s generative process includes the following steps:

1. Draw a topic β_k from $\beta_k \sim Dirichlet(\eta)$ for $k = 1...K$
2. For each document d:
   (a) Draw topics proportions θ_d such that $\theta_d \sim Dirichlet(\alpha)$
   (b) For each word $w_{d,n}$:
      i. Draw topic assignment $z_{d,n} \sim \text{Multinomial}(\theta_d)$
      ii. Draw word $w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$

The parameters α and η are hyperparameters that indicate respectively the priors on per-document topic distribution and per-topic word distribution, respectively. Thus, $w_{d,n}$ are the only observable variables, all the others are latent in this model. For a set of D documents, given the parameters α and η, the joint distribution of a topic mixture θ, word-topic mixtures β, topics z, and a set of N words is given by:

$$p(\theta, \beta, z, w|\alpha, \eta) = \prod_{k=1}^{K} p(\beta_k|n) \prod_{d=1}^{D} p(\theta_d|a) =$$
$$= \prod_{n=1}^{N} (p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_k, k = z_{d,n})) \quad (5.1)$$

Broadly speaking, the training task is to find the posterior distribution of the latent variables (the per-document topic proportions θ_d, the per-word topic assignments $z_{d,n}$ and the topics β_k) that maximize this probability.

The parameter that we mainly focused in this study is the number of topics. We tested a range of values between 5 and 30, and we empirically concluded that the value of 10 yielded the best model results. With 10 topics we are able to capture all kinds of events included in our event database, and we also narrow down possible equivocal topics that could deteriorate our results. The other parameters, the α and η priors, were kept as default (1.0/(number of topics)). To confirm this was a safe choice, we ran several iterations with different initial α and η priors and they generally converged to similar outcomes. The LDA results are presented in Table 5.1.

TABLE 5.1: LDA Results

Topic	No. Events	Popular Words
Topic_1	24	ice, disney, present, magic, new
Topic_2	72	basketball, championship, atlantic, game, tournament
Topic_3	10	show, artist, box, office, special
Topic_4	32	music, atlantic, championship, basketball, game
Topic_5	22	game, marriot, corporate, bridge, hotel
Topic_6	10	train, service, islander, view, time
Topic_7	34	tour, album, show, meet, up
Topic_8	12	circus, family, out, space, earth
Topic_9	12	dinner, reservation, jay, menu, restaurant
Topic_10	42	champion, game, group, boxing, hoop

5.5 Experiments

In this section, we demonstrate the hypothesis that information about events is significant in real-time taxi demand prediction in the vicinity of special event venues. The inclusion of information about the occurrence of planned event, allows a better understanding of demand fluctuations, as well as the restriction of final forecasts' margin of error. Our approach is evaluated in two event areas in New York City (NYC) and the proposed data fusion methodology was implemented in Keras (Chollet, 2015).

5.5.1 Dataset and case studies

Our base dataset consists of 1.1 billion taxi trips from New York, distributed by technology providers of authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP) and were made publicly available by the NYC Taxi and Limousine Commission (TLC). We use taxi data from 1/1/2013 through 6/30/2016, which includes around 600 million taxi trips after data filtering. The dataset specifies for each drop-off and pick-up event the GPS location and the time-stamp.

Based on this data, we looked at a list of the top venues in NYC and selected the two venues for which more complete event records were available online: the Barclays Center and Terminal 5. The first venue is located in the heart of Brooklyn and it is the state-of-the-art home of the NBA's Brooklyn Nets and the NHL's New York Islanders. It is one of the most popular facilities in the New York metropolitan area because it hosts many sold-out concerts, conventions and other sporting and entertainment events. It is ranked top five globally in 2015 for gross revenue and attendance by Billboard and Venues Today. On the other hand, the Terminal 5 is a 3-floor venue that regularly hosts concerts with many different audiences and that is located in the heart of Manhattan. Given the geographical coordinates of these two venues, we selected all the taxi pickups that took place within a bounding box of ±0.003 decimal degrees (roughly 500 meters) to be our study areas (Figure 5.3).

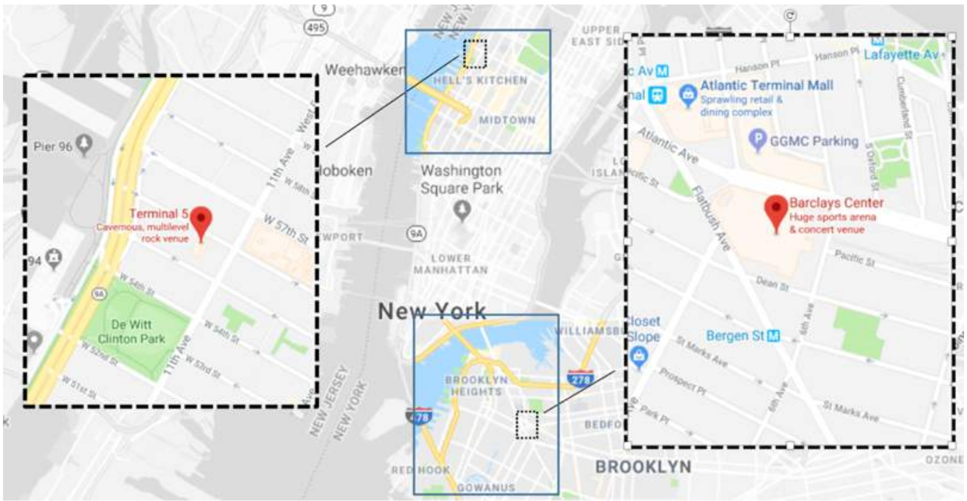


FIGURE 5.3: Map of the two studied areas. Map data © OpenStreetMap contributors

The individual records that fall within the boundaries described above were grouped in a time-series of hourly counts. Our goal is to predict the taxi demand of the area at the next hour, considering the demand from previous records, as well as event information extracted from the Web. In this way, stakeholders, such as companies like Uber and taxi operators, can have a clear image of demand in the near future, thus allowing them to better organize their fleet. Precise next-hour demand forecasts allow those companies' fleet to become more efficient as routes become targeted and balanced with the demand.

Regarding the event data, it was extracted automatically from the Web using either screen scrapping techniques or Application Programming Interfaces (API's). For the Barclays Center, the event information was scrapped from its official website, since it maintains a very accurate and detailed calendar. We collected a total of 751 events since its inauguration in late 2012 until June 2016. As for the Terminal 5, we used the Facebook API to extract 315 events from its official page, for a similar time period. In both cases, the event data includes event's title, date, time and description.

5.5.2 Experimental Setup

The taxi dataset includes records of trips from 2013 and we created separate training and test sets for each stage of our methodology (Figure 5.4). More specifically, we selected the first three years (January 2013 – December 2015) as our training set and the first 5 months of 2016 as our test set for the first phase of our methodology with the "**DP-Model**". For the validation process, we separated 20% of our training set using the automatic tools of Keras.

At the second stage, where the "**EoDP-Model**" is introduced, we extend the training set of the "DP-Model" to May 2016, and we use the last month of our dataset (June 2016) for testing. The first five months of 2016 are used as the training dataset of "**EoDP-Model**", since the calculated error values from the first stage correspond to this timeframe.

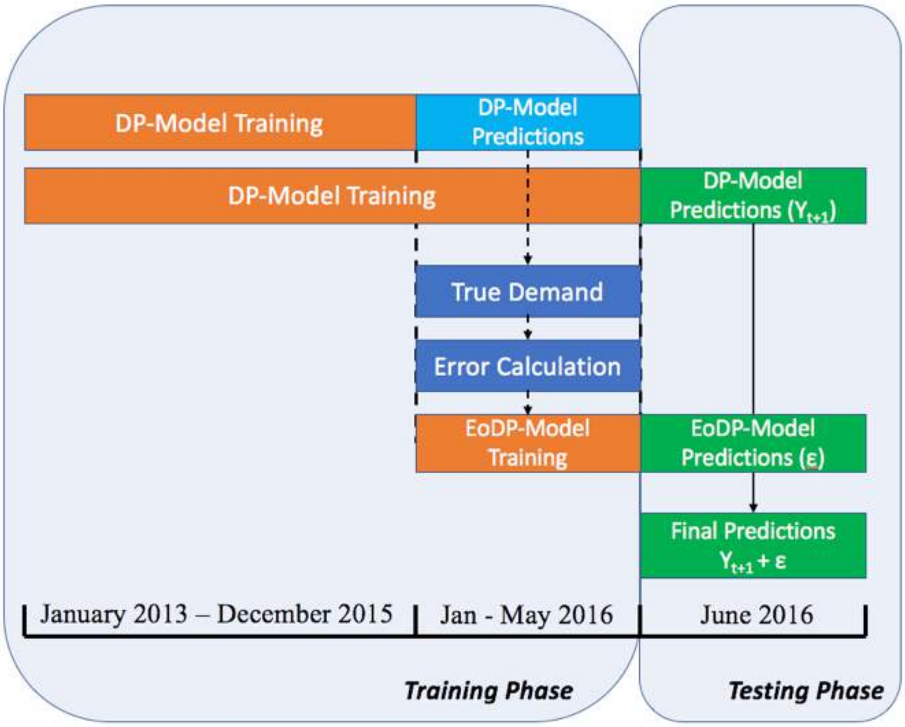


FIGURE 5.4: Experimental Setup Depiction.

Training Phase

In order to evaluate the contribution of the different sources of information, we perform an incremental analysis of the proposed deep learning architecture. We start with only the part of the network that is responsible for modeling the time-series data and we keep adding components to the network until the full model depicted in Figure 5.2 is obtained. Therefore, we start with a model that only takes the lagged pickup observations (referred to as "P") as input and move to models that also include: drop-off lags (denoted "P+D"), information about the presence of events ("P+D+E") and finally, the full model that also considers events' topics ("P+D+E+T").

For models' performance validation and comparison, we will use the mean absolute error (MAE), the root-mean-square error (RMSE) and the coefficient of determination (R2), computed as follows:

$$MAE = \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n| \tag{5.2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2} \tag{5.3}$$

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y_n})^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} \tag{5.4}$$

where N denotes the number of instances in the dataset, $\hat{y}_n$ is the predicted taxi pick-ups count for the n_{th} instance, y_n is the corresponding true pick-ups count and $\bar{y}$ is the mean of the observed counts.

After the forecasts are obtained, the deviation of each measurement from the true value is calculated. The new vector will be used as the dependent variable of the **EoDP-Model** at the testing phase of the proposed methodology.

Testing Phase

At this stage, we implement a parallel training of our models. For the DP-Model, the training dataset is expanded by 5 months (Jan-May 2016) and for the **EoDP-Model** we use the vector of residuals that correspond to the same time period Jan-May 2016.

The evaluation of our approach starts with the day of month and day of week as **EoDP-Model**'s independent variables (denoted “d+m“) and we add event information (“d+m+E“) and events' topics (“d+m+E+T“) in the subsequent steps. The **EoDP-Model**'s architecture makes use of Fully-Connected (FC) layers as as in the DP-Model model. The only difference is the composition of the first dense layer, since we do not have as inputs the pickups and drop-offs lags, but the day of week and day of month.

For the step-by-step and detailed analysis of the contribution of each parameter, we begin with the study of the basic models and gradually add more parameters. Therefore, at first, we check the accuracy of the models using only pickup lags (for the DR-Model) and the day of week and day of month variables (for the **EoDP-Model**), and afterwards we add more variables that could improve method's performance.

Besides these baselines, the proposed approach is further compared with another popular method from the state of the art for time-series forecasting, the Linear Regression (LR). Its simplicity and interpretability were the criterias for its selection.

5.6 Results

Table 5.2 shows the results of demand predictions using only the DP-Model for June 2016. They will form our baseline for the evaluation and comparison of that simple architecture of a single model with the proposed two-step approach that we will implement afterwards.

From the initial results we understand that in the case of linear regression, information about scheduled events plays an influential role. For Barclays center, the type of event contributes more to the results' improvement, while for Terminal 5 the start and end time of an event seems to be more determinant. This conclusion can be also justified by the fact that the popular venue of Brooklyn hosts concerts, conventions and other sporting and entertainment events, which attract a different number of people each time, and demand fluctuations can also be different because of that. Topic modeling captures the event categories that the venue hosts, therefore its contribution is clear on the final accuracy of the model. On the other hand, Terminal 5 hosts mostly music events and and audience attendance can be estimated satisfactorily with the start and end time of each event.

In the case of deep learning, it is obvious that event information does not appear to have any significant effect on the results, when it is included directly to the demand

Table 5.2: Demand prediction using only the DP-Model

	BARCLAYS CENTER			TERMINAL 5		
	MAE	RMSE	R2	MAE	RMSE	R2
LR-P	7,989	13,266	0,762	7,966	12,203	0,770
LR-P+D	7,437	11,817	0,811	7,159	10,755	0,821
LR-P+D+E	7,527	12,265	0,796	**7,134**	**10,487**	**0,830**
LR-P+D+E+T	**7,187**	**11,188**	**0,831**	7,163	10,558	0,828
DL-P	8,081	13,393	0,757	7,954	12,252	0,768
DL-P+D	7,671	12,121	0,801	7,079	10,556	**0,828**
DL-P+D+E	7,630	12,202	0,799	7,094	10,743	0,822
DL-P+D+E+T	7,625	12,106	**0,802**	7,151	10,736	0,825

Table 5.3: Demand prediction using DP-Model and EoDP-Model (Barclays Center)

	MAE			RMSE			R2		
	d+m	d+m+E	d+m+E+T	d+m	d+m+E	d+m+E+T	d+m	d+m+E	d+m+E+T
LR-P	8,080	**7,955**	7,902	13,288	**12,496**	12,049	0,761	**0,789**	0,804
LR-P+D	7,689	**7,365**	7,598	12,212	**11,152**	11,711	0,798	**0,832**	0,814
LR-P+D+E	7,672	**7,293**	7,506	12,157	**11,133**	11,884	0,800	**0,832**	0,809
LR-P+D+E+T	7,632	**7,355**	7,717	12,080	**11,302**	12,194	0,803	**0,827**	0,799
DL-P	8,034	**7,743**	8,067	13,293	**12,177**	14,250	0,761	**0,799**	0,725
DL-P+D	7,439	**7,072**	7,379	11,817	**10,714**	12,577	0,811	**0,845**	0.786
DL-P+D+E	7,528	**7,074**	7,417	12,260	**10,798**	12,719	0,797	**0,842**	0,781
DL-P+D+E+T	7,185	**7,105**	7,484	11,168	**10,775**	13,023	0,831	**0,843**	0,770

forecasting model. Pickup and drop-off lags seem enough for model's best possible performance. It remains to be seen, if by using the proposed architecture with the **EoDP-Model**, the results will be changed.

Table 5.3 shows the performance of our two-step architecture for Barclays Center in June 2016, namely the same time period that Table 5.1 results refer to. In these measurements, both models are used, based on the methodology described in the methodology section.

We can see from the final scores that the introduction of a demand error forecasting model contributes significantly to the reduction of the final forecasting error. It is noteworthy to mention that the **EoDP-Model** with event information has the greatest impact. In both cases (DL and LR models) the error forecasting model contributes positively.

For Linear Regression, significant differences appear only using the **"d+m+E" EoDP-Model**. The MAE is decreased by 1,6%, which is also considered important, since the previous model (Table 5.2 results) was already fairly accurate. Moving to the Terminal 5 study area, Table 5.4 shows the obtained results. In this case, the results are not as clear as in the previous study area. The positive contribution of the **EoDP-Model** using the DL method appears only in the "d+m+E" case, where the R2 score of an already good forecasting model is still increased by 1,7%.

It seems that the mobility patterns around this venue, when a special event is organized, are less predictable than in Barclays Center. This is probably due to its location, which is in a very central area of Manhattan, where we can also locate some other popular venues, bars and restaurants that citizens prefer to visit daily. Consequently, the observed demand fluctuations in this area are directly affected by other parameters which are not considered in this study, and therefore can not be predicted.

Table 5.4: Demand prediction using DP-Model and EoDP-Model (Terminal 5)

	MAE			RMSE			R2		
	d+m	d+m+E	d+m+E+T	d+m	d+m+E	d+m+E+T	d+m	d+m+E	d+m+E+T
LR-P	8,162	**8,004**	8,229	12,444	**11,989**	13,681	0,761	**0,778**	0,711
LR-P+D	7,280	**7,133**	7,194	10,695	**10,358**	10,827	0,823	**0,834**	0,819
LR-P+D+E	7,067	**6,975**	7,624	10,613	**10,316**	10,035	0,826	**0,836**	0,834
LR-P+D+E+T	7,360	7,319	7,438	10,986	11,056	11,703	0,814	0,811	0.788
DL-P	8,072	8,072	8,112	12,257	12,257	12,385	0,768	0,768	0,763
DL-P+D	7,188	7,188	7,246	10,752	10,752	10,884	0,821	0,821	0,817
DL-P+D+E	7,158	**7,158**	7,241	10,474	**10,474**	10,91	0,830	**0,831**	0,816
DL-P+D+E+T	7,186	7,186	7,306	10,544	10,544	11,258	0,828	0,828	0,804

5.7 Conclusion

We demonstrated that using online information, we can improve the quality of taxi demand prediction even in scenarios where the transport system is under stress. We combined information extracted from the web with time-series data to formulate our two-step approach with tow predictive models that capture in real-time future demand in event areas. This is typically a challenging case for transport planning since special events originate high variance in demand. Taxi demand is correlated with many parameters of underlying information and currently, most taxi centers rely on formal processes and manual work for a fleet organization and taxi distribution. Even the more advanced new services, like Uber or Lyft, still face great challenges in terms of demand prediction (shown by price surges and variations thereof). Our results show a second model that predicts the forecasting error of the main model is able to further improve the final predictions. Information about events from the Web contributes decisively to the ultimate accuracy of the proposed methodology. Hence, besides the value of event information and time series data, our empirical results also highlight the need for accounting for the effect of events when modeling mobility demand.

In future work, we aim at exploring the impact of spatio-temporal interactions on taxi demand prediction. The development of a city-wide spatio-temporal model that accounts for information about all the events that take place across the city could be a generalization potential of this methodology

Chapter 6

Conclusions and future work

The understanding and combating of congestion phenomena that we experience in the transport system of our city due to special events, was the main pillar of research in this thesis. It is already well-known that our high rhythm, demanding everyday life requires us to spend a lot of time in a car or on public transport in order to accomplish all tasks that we have in our calendars. Therefore, it is quite unpleasant when we have to spend additional time on a congested road due to a popular concert that has just ended nearby. It's been many times that we wished we had a intelligent system that would inform us on time to change our route because an event, an accident or a public demonstration has fully constrained a major road axis. This thesis presented a framework that detects traffic anomalies happened in the past, analyzes them by identifying which parameters are the most influential, collects valuable data about critical events using internet search queries, and finally predicts demand hotspots utilizing the outcome of all previous tasks.

The detection of traffic anomalies was crucial for understanding each event's degree of influence. Some of events' characteristics that emerged as the most critical for demand hotspots presence are the day of the week and the time of day that the event is organized, as well as its location. Other details, such as the type of event and the number of people that expressed their willingness to attend the event through social media, did not give us clear conclusions.

The kernel density maps allowed us to investigate which of the events that we were interested in, disrupted transport system's balance, as well as the extent of it. This visualization tool helped us examine an area of interest more generously than to every single trip separately. We correlated observed hotspots with events that took place in the area around the same time and we came to the conclusion that in order to make a noteworthy prediction model, that will be able to escape from daily traffic patterns and predict extreme phenomena too, it is fundamental to incorporate events' information.

Having clearly demonstrated the importance of events, we moved on to the information retrieval task. We developed a framework that sequentially handled all stages of data gathering, enrichment, and prediction with the intention of generating automated search queries. Using machine learning techniques, several queries generation and expansion practices were examined for the accomplishment of the best possible prediction performance of a classifier that determines whether an upcoming event is a hotspot or not.

The outcome of several experiments using LDA and three popular classifiers, namely the Support Vector Machine (SVM) classifier, Logistic Regression (LR) and the Multi-layer Perceptron (MPL) classifier, showed that MedLDA successfully maximized the performance of our classification task. Demand hotspots were predicted with good accuracy using retrieved Web documents that include details about the location of the event, as well as time-related keywords.

Through these experiments, the significance of temporal and spatial information of events for the prediction of extreme phenomena was emphasized once again. The location of an event seems to play a decisive role, and one of the reasons is that there are plenty of venues with a very large capacity, and even if a small percentage of people takes a taxi on departure, the average demand of the area will significantly be affected. Therefore, in future cases, it is considered very important to categorize the venues according to their capacity and the completeness that they generally present at their events.

Along with the classification evaluation, we studied the structure of the queries that we need to formulate in order to get the most useful outcome from the search engines. It has been found that the query expansion stage is very significant, since the variety of keywords that is incorporated, based on the evaluation of the initial query's results, influence the final outcome quite positively. Additionally, the right selection of documents was a challenging problem due to the heterogeneous and noisy nature of the data. We specifically noticed that in most cases it was necessary to import time information in various formats. This requirement gave us an insight into the enormous volume of information that is available on websites, and how difficult it is to explore it directly (without requiring queries reformulation).

Results from the information retrieval and anomalies detection tasks that we previously met, were finally combined for the explanation of specific events occurrence. It was very interesting to see each dataset's contribution to the model's final accuracy; distinctly important among those datasets, was the one with topic modelling based variables. Each category (e.g. family event, basketball game etc.) seems to affect demand peaks differently. The reason behind this indication would be interesting to investigate later in the future.

Model's performance was evaluated through several machine learning techniques. Linear Regression, Gaussian Processes, Fully-Connected dense layers and Deep Gaussian Processes were tested, in order to find the most appropriate method for our real-time prediction goals. It was concluded that, Gaussian Processes and the deep learning architecture with FC layers were the predominant ones. Gaussian processes though have the advantage of publishing the desired estimations faster compared to the other three, consequently we are able to inform the stakeholders more promptly.

In the context of further improving the accuracy of our forecast results, the idea of a two-stage process was introduced. The proposed framework is focused on the analysis, evaluation, and forecasting of prediction model's residuals. More specifically, two different models were formulated; one that is focused on the direct taxi demand prediction based on the available historical taxi data and a second one, whose objective is the estimation of demand prediction residuals based on the historical performance of the first model. The results look promising, because the introduction of a demand error forecasting model contributed significantly to the reduction of the final forecasting error. The deep learning architecture with pickups, drop-off lags and event information leads to an overall MAE reduction of 6% and 12% in RMSE. The above results helped us to conclude that there is still room for further improvement of an already satisfactory predictive model.

In summary, this thesis proposed a detailed framework for dealing with anomalies on a traffic network that are highly correlated with special events. The proposed approaches explore machine learning techniques for time-series observations' analysis and textual data incorporation. The overall research was based on well-defined areas in NYC and each model was tuned using trip records from the area that refers to. Hence, future work could explore the utility of information from other areas for model's taxi demand prediction optimization. This collaboration can be defined through a correlation structure that is strongly dependent on domain, and itself potentially dynamic.

It is true that many of us are not confined to a single neighborhood for our shopping, sport activities or our entertainment. The internet, and more intensively social media, inform as continuously about events that are going to take place in another distinct. Therefore, we eventually end up having more choices and greater need to transit to other regions for our entertainment, family or social activities. Our daily program affects the supply and demand equilibrium of more than one area, and since our relocations are highly related to the services that each region offers (restaurants, bars, schools, theatres etc.), respectively can be related with the demand and supply forecast models. These components would be very useful and interesting to include into our future prediction frameworks.

The correlation of models can be implemented using only historical time-series data, but also with the exploitation of Points of Interest (POIs). We can define new areas based on the characteristics of the POIs that they contain (late-night entertainment zone, sport entertainment zone, business zone etc.), create new forecasting models for each one of them, and finally see if they could be correlated based on the time-schedule that people usually follow. In this way, we could be able to explain special traffic patterns observed in our study area as a consequence of an event scheduled in another area.

During the last years so-called free-floating Car Sharing Systems became very popular. The spatial distribution of their vehicles is in a few cases manually controlled by system operators or self-organized, which means it is only dependent on the customer's demand (Weikl and Bogenberger, 2013). The developed models that indicate areas where higher demand needs to be met, do not yet account for special events' influence, a component that is extremely important for model's good performance. Hence, in future work, we would like to explore the contribution of events' information in car sharing systems fleet distribution, but also in autonomous vehicles when we have a high penetration rate of them on our streets.

The models presented in this thesis are intended to predict extreme demand fluctuations due to special events. In future work, we would like to extend our research to other types of events, such as crisis scenarios, incidents, demonstrations or religious events. The process of crisis management comprises of a complete life cycle of activities that is carried out as soon as a situation is identified as a crisis. Therefore, prior knowledge of such events' influence on a transport system, which is based on historical observations, will be the most important tool for their successful management.

Bibliography

Abdelhaq, Hamed, Christian Sengstock, and Michael Gertz (2013). “Eventweet: Online localized event detection from twitter”. In: *Proceedings of the VLDB Endowment* 6.12, pp. 1326–1329.

Barria, Javier A and Suttipong Thajchayapong (2011). “Detection and classification of traffic anomalies using microscopic traffic variables”. In: *IEEE Transactions on Intelligent Transportation Systems* 12.3, pp. 695–704.

Beat (2017). https://thebeat.co/en/. [Online; accessed 21-November-2017].

Becker, Hila, Mor Naaman, and Luis Gravano (2010). “Learning similarity metrics for event identification in social media”. In: *Proceedings of the third ACM international conference on Web search and data mining*. ACM, pp. 291–300.

Becker, Hila et al. (2012). “Identifying content for planned events across social media sites”. In: *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, pp. 533–542.

Best concert venues in NYC (2017). https://www.nyc.com/nyc-guides/best_concert_venues_in_nyc.308/. [Online; accessed 14-August-2017].

BIG QUERY (2016). https://cloud.google.com/bigquery/. [Online; accessed 1-August-2016].

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Borysov, Stanislav et al. (2016). “Using internet search queries to predict human mobility in social events”. In: *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE, pp. 1342–1347.

Bu, Yingyi et al. (2009). “Efficient anomaly monitoring over moving object trajectory streams”. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 159–168.

Candia, Julián et al. (2008). “Uncovering individual and collective human dynamics from mobile phone records”. In: *Journal of physics A: mathematical and theoretical* 41.22, p. 224015.

Castro, Pablo Samuel, Daqing Zhang, and Shijian Li (2012). “Urban traffic modelling and prediction using large scale taxi GPS traces”. In: *International Conference on Pervasive Computing*. Springer, pp. 57–72.

Chan, Jacky WY et al. (2016). “Taxi App Market Analysis in Hong Kong”. In: *Journal of Economics, Business and Management* 4.3.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar (2009). “Anomaly detection: A survey”. In: *ACM computing surveys (CSUR)* 41.3, p. 15.

Chang, Mei-Shiang and Pei-Rong Lu (2013). “A multinomial logit model of mode and arrival time choices for planned special events”. In: *Journal of the Eastern Asia Society for Transportation Studies* 10, pp. 710–727.

Chen, Andrew et al. (2006). “Smoothing vehicular traffic flow using vehicular-based ad hoc networking & computing grid (VGrid)”. In: *Intelligent Transportation Systems Conference, 2006. ITSC’06. IEEE*. IEEE, pp. 349–354.

Cheng, Qixiu et al. (2016). *Analysis and forecasting of the day-to-day travel demand variations for large-scale transportation networks: A deep learning approach.*

Cheng, Zhiyuan et al. (2011). "Exploring millions of footprints in location sharing services." In: *ICWSM* 2011, pp. 81–88.

Chollet, François et al. (2015). "Keras". In:

Christoforou, Zoi et al. (2016). "Managing Planned Disruptions of Mass Transit Systems". In: *Transportation Research Record: Journal of the Transportation Research Board* 2541, pp. 46–55.

Coutroubas, F and N Tzivelou (2003). "Public transport planning for the greatest event: the 2004 Olympic Games". In: *Proceedings of the ETC, Strasbourg, France.*

Cramer, Judd and Alan B Krueger (2016). "Disruptive change in the taxi business: The case of Uber". In: *American Economic Review* 106.5, pp. 177–82.

Damianou, Andreas and Neil Lawrence (2013). "Deep Gaussian Processes". In: *Artificial Intelligence and Statistics*, pp. 207–215.

Davis, Neema, Gaurav Raina, and Krishna Jagannathan (2016). "A multi-level clustering approach for forecasting taxi travel demand". In: *ITSC, 2016.* IEEE, pp. 223–228.

Dunn Jr, Walter M, Steven P Latoski, and Elizabeth Bedsole (2006). *Planned Special Events: Checklists for Practitioners.* Tech. rep.

Duvenaud, David et al. (2014). "Avoiding pathologies in very deep networks". In: *Artificial Intelligence and Statistics*, pp. 202–210.

Ferreira, Nivan et al. (2013). "Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12, pp. 2149–2158.

Frias-Martinez, Vanessa et al. (2012). "Characterizing urban landscapes using geolocated tweets". In: *Privacy, Security, Risk and Trust (PASSAT), 2012 International Confernece on Social Computing (SocialCom).* IEEE, pp. 239–248.

Fuhs, Chuck and Parsons Brinckerhoff (2010). *Synthesis of active traffic management experiences in Europe and the United States.* Tech. rep. United States. Federal Highway Administration.

Gers, Felix A, Douglas Eck, and Jürgen Schmidhuber (2002). "Applying LSTM to time series predictable through time-window approaches". In: *Neural Nets WIRN Vietri-01.* Springer, pp. 193–200.

Goodfellow, Ian et al. (2016). *Deep learning.* Vol. 1. MIT press Cambridge.

Hawelka, Bartosz et al. (2014). "Geo-located Twitter as proxy for global mobility patterns". In: *Cartography and Geographic Information Science* 41.3, pp. 260–271.

Hölscher, Christoph and Gerhard Strube (2000). "Web search behavior of Internet experts and newbies". In: *Computer networks* 33.1-6, pp. 337–346.

Idé, Tsuyoshi and Sei Kato (2009). "Travel-time prediction using Gaussian Process Regression: A trajectory-based approach". In: *Proceedings of the 2009 SIAM Int. Conf. on Data Mining.* SIAM, pp. 1185–1196.

INRIX Research (2017). *INRIX Global Traffic Scoreboard.*

Ioffe, Sergey and Christian Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167.*

Jones, Karen Spärck (1972). "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of Documentation* 28, pp. 11–21.

Jordan, Michael I and Tom M Mitchell (2015). "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245, pp. 255–260.

Kamga, Camille and M Anıl Yazıcı (2014). "Temporal and weather related variation patterns of urban travel time: Considerations and caveats for value of travel time, value of variability, and mode choice studies". In: *TR-C: Emerging Technologies* 45, pp. 4–16.

Kireyev, Kirill, Leysia Palen, and Kenneth Anderson (2009). "Applications of topics models to analysis of disaster-related twitter data". In: *NIPS Workshop on Applications for Topic Models: Text and Beyond.* Vol. 1. Canada: Whistler.

Kuppam, Arun et al. (2011). "Innovative methods for collecting data and for modeling travel related to special events". In: *Transportation Research Record: Journal of the Transportation Research Board* 2246, pp. 24–31.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep Learning". In: *nature* 521.7553, p. 436.

Lee, Ryong and Kazutoshi Sumiya (2010). "Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection". In: *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks.* ACM, pp. 1–10.

Li, Xiaolei et al. (2009). "Temporal outlier detection in vehicle traffic data". In: *IEEE International Conference on Data Engineering.* IEEE, pp. 1319–1322.

Liu, Wei et al. (2011). "Discovering spatio-temporal causal interactions in traffic data streams". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, pp. 1010–1018.

Lv, Yisheng et al. (2015). "Traffic flow prediction with big data: a deep learning approach". In: *IEEE Transactions on Intelligent Transportation Systems* 16.2, pp. 865–873.

Ma, Huifang, Bo Wang, and Ning Li (2012). "A novel online event analysis framework for micro-blog based on incremental topic modeling". In: *Software Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing (SNPD), 2012 13th ACIS International Conference on.* IEEE, pp. 73–76.

Ma, Xiaolei et al. (2015). "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data". In: *Transportation Research Part C: Emerging Technologies* 54, pp. 187–197.

Markou, Ioulia, Filipe Rodrigues, and Francisco C. Pereira (2018). "Real-Time Taxi Demand Prediction using data from the web". In: *ITSC, 2018.* IEEE.

Markou, Ioulia, Filipe Rodrigues, and Francisco C Pereira (2017). "Use of Taxi-Trip Data in Analysis of Demand Patterns for Detection and Explanation of Anomalies". In: *Transportation Research Record: Journal of the Transportation Research Board* 2643, pp. 129–138.

— (2019). "Is travel demand actually deep? An application in event areas using semantic information". In: *IEEE Transactions on Intelligent Transportation.*

Miao, Fei et al. (2016). "Taxi dispatch with real-time sensing data in metropolitan areas: A receding horizon control approach". In: *IEEE Transactions on Automation Science and Engineering* 13.2, pp. 463–478.

Moreira-Matias, Luis et al. (2013). "Predicting taxi–passenger demand using streaming data". In: *IEEE Transactions on Intelligent Transportation Systems* 14.3, pp. 1393–1402.

Morgul, Ender Faruk and Kaan Ozbay (2015). "Revisiting labor supply of new york city taxi drivers: Empirical evidence from large-scale taxi data". In: *TRB 94th Annual Meeting.* 15-3331.

Nichols, Jeffrey, Jalal Mahmud, and Clemens Drews (2012). "Summarizing sporting events using twitter". In: *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces.* ACM, pp. 189–198.

Pan, Bei et al. (2013). "Crowd sensing of traffic anomalies based on human mobility and social media". In: *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.* ACM, pp. 344–353.

Parkany, Emily and Chi Xie (2005). *A complete review of incident detection algorithms & their deployment: what works and what doesn't.* Tech. rep.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Perallos, Asier et al. (2015). *Intelligent Transport Systems: Technologies and Applications.* John Wiley & Sons.

Pereira, Francisco C, Ana LC Bazzan, and Moshe Ben-Akiva (2014). "The role of context in transport prediction". In: *IEEE Intelligent Systems* 29.1, pp. 76–80.

Pereira, Francisco C, Filipe Rodrigues, and Moshe Ben-Akiva (2012). *Internet as a sensor: a case study with special events.* Tech. rep.

— (2013). "Text analysis in incident duration prediction". In: *Transportation Research Part C: Emerging Technologies* 37, pp. 177–192.

— (2015). "Using data from the web to predict public transport arrivals under special events scenarios". In: *Journal of Intelligent Transportation Systems* 19.3, pp. 273–288.

Pereira, Francisco C et al. (2015). "Why so many people? explaining nonhabitual transport overcrowding with internet data". In: *IEEE Transactions on Intelligent Transportation Systems* 16.3, pp. 1370–1379.

Quercia, Daniele and Diego Saez (2014). "Mining urban deprivation from foursquare: Implicit crowdsourcing of city land use". In: *IEEE Pervasive Computing* 13.2, pp. 30–36.

Ramage, Daniel, Susan T Dumais, and Daniel J Liebling (2010). "Characterizing microblogs with topic models." In: *ICWSM* 10, pp. 1–1.

Ramos, Juan et al. (2003). "Using tf-idf to determine word relevance in document queries". In: *Proceedings of the first instructional conference on machine learning.* Vol. 242, pp. 133–142.

Rasmussen, Carl Edward (2004). "Gaussian processes in machine learning". In: *Advanced lectures on machine learning.* Springer, pp. 63–71.

Rocchio, J. J. (1971). "Relevance feedback in information retrieval". In: *The Smart retrieval system - experiments in automatic document processing.* Ed. by G. Salton. Englewood Cliffs, NJ: Prentice-Hall, pp. 313–323.

Rodrigues, Filipe, Ioulia Markou, and Francisco C Pereira (2019). "Combining timeseries and textual data for taxi demand prediction in event areas: a deep learning approach". In: *Information Fusion* 49, pp. 120–129.

Rodrigues, Filipe et al. (2016). "A Bayesian additive model for understanding public transport usage in special events". In: *IEEE transactions on pattern analysis and machine intelligence.*

Rodrigues, Filipe et al. (2017). "Learning supervised topic models for classification and regression from crowds". In: *IEEE transactions on pattern analysis and machine intelligence* 39.12, pp. 2409–2422.

Rose, Stuart et al. (2010). "Automatic keyword extraction from individual documents". In: *Text Mining: Applications and Theory*, pp. 1–20.

Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo (2010). "Earthquake shakes Twitter users: real-time event detection by social sensors". In: *Proceedings of the 19th int. conference on World wide web.* ACM, pp. 851–860.

Salimbeni, Hugh and Marc Deisenroth (2017). "Doubly stochastic variational inference for deep gaussian processes". In: *Advances in Neural Information Processing Systems*, pp. 4591–4602.

Schmidhuber, Jürgen (2015). "Deep learning in neural networks: An overview". In: *Neural networks* 61, pp. 85–117.

Schütze, Hinrich, Christopher D Manning, and Prabhakar Raghavan (2008). *Introduction to information retrieval*. Vol. 39. Cambridge University Press.

Seeger, Matthias (2004). “Gaussian processes for machine learning”. In: *International journal of neural systems* 14.02, pp. 69–106.

Shahin, Shakibaei, Tezcan Onur Hüseyin, and Öğüt Selçuk Kemal (2014). “Evaluating transportation preferences for special events: A case study for a megacity, Istanbul”. In: *Procedia-Social and Behavioral Sciences* 111, pp. 98–106.

Sheu, Jiuh-Biing (2004). “A sequential detection approach to real-time freeway incident detection and characterization”. In: *European Journal of Operational Research* 157.2, pp. 471–485.

Snoek, Jasper, Hugo Larochelle, and Ryan P Adams (2012). “Practical bayesian optimization of machine learning algorithms”. In: *Advances in neural information processing systems*, pp. 2951–2959.

Thom, Dennis et al. (2012). “Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages”. In: *Pacific visualization symposium (PacificVis), 2012 IEEE*. IEEE, pp. 41–48.

TLC Trip Record Data (2018). `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`. [Online; accessed 21-March-2018].

Tong, Yongxin et al. (2017). “The simpler the better: a unified approach to predicting original taxi demands based on large-scale online platforms”. In: *Proceedings of the 23rd ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining*. ACM, pp. 1653–1662.

Watanabe, Kazufumi et al. (2011). “Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs”. In: *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, pp. 2541–2544.

Weikl, Simone and Klaus Bogenberger (2013). “Relocation strategies and algorithms for free-floating car sharing systems”. In: *IEEE Intelligent Transportation Systems Magazine* 5.4, pp. 100–111.

Wikipedia (2017a). *Grab (application)*. `https://en.wikipedia.org/wiki/Grab_(application)/`. [Online; accessed 21-November-2017].

— (2017b). *Uber (company)*. `https://en.wikipedia.org/wiki/Uber_(company)/`. [Online; accessed 21-November-2017].

Xie, Yuanchang et al. (2010). “Gaussian Processes for short-term traffic volume forecasting”. In: *Transportation Research Record* 2165, pp. 69–78.

Xu, Jun et al. (2017). “Real-Time Prediction of Taxi Demand Using Recurrent Neural Networks”. In: *IEEE Transactions on ITS*.

Xu, Zheng et al. (2016). “Crowdsourcing based description of urban emergency events using social media big data”. In: *IEEE Transactions on Cloud Computing* 10, p. 1109.

Yang, Ci and Eric J Gonzales (2017). “Modeling taxi demand and supply in New York City using large-scale taxi GPS data”. In: *Seeing Cities Through Big Data*. Springer, pp. 405–425.

Yardi, Sarita and Danah Boyd (2010). “Tweeting from the Town Square: Measuring Geographic Local Networks.” In: *ICWSM*, pp. 194–201.

Yazici, M Anil, Camille Kamga, and Abhishek Singhal (2013). “A big data driven model for taxi drivers’ airport pick-up decisions in new york city”. In: *2013 IEEE Intern. Conf. on Big Data*. IEEE, pp. 37–44.

Yuan, Jing et al. (2011). “Where to find my next passenger”. In: *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, pp. 109–118.

Yuan, Jinhui et al. (2015). “Lightlda: Big topic models on modest computer clusters”. In: *Proceedings of the 24th International Conference on World Wide Web.* International World Wide Web Conferences Steering Committee, pp. 1351–1361.

Zhang, J, Y Zheng, and D Qi (2017a). “Deep spatio-temporal residual networks for citywide crowd flows prediction.” In: *AAAI*, pp. 1655–1661.

Zhang, Jianting (2012). “Smarter outlier detection and deeper understanding of large-scale taxi trip records: a case study of NYC”. In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing.* ACM, pp. 157–162.

Zhang, Junbo, Yu Zheng, and Dekang Qi (2017b). “Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction.” In: *AAAI*, pp. 1655–1661.

Zhang, Kai et al. (2016). “A Framework for Passengers Demand Prediction and Recommendation”. In: *Int. Conference on SCC.* IEEE, pp. 340–347.

Zhao, Kai et al. (2016). “Predicting taxi demand at high spatial resolution: Approaching the limit of predictability”. In: *2016 IEEE Intern. Conf. on Big Data.* IEEE, pp. 833–842.

Zhu, Jun, Amr Ahmed, and Eric P Xing (2009). “MedLDA: maximum margin supervised topic models for regression and classification”. In: *Proceedings of the 26th annual international conference on machine learning.* ACM, pp. 1257–1264.