University of the Basque Country
UPV/EHU

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Doctoral Thesis

Design and Validation of Novel Methods for Long-term Road Traffic Forecasting

---

Author:
Ibai Laña

Supervisors:
Prof. Dr. Javier Del Ser
Prof. Dr. Manuel Vélez

A Thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Department of Communications Engineering

September 24, 2018

"*We can only see a short distance ahead, but we can see plenty there that needs to be done.*"

Alan Turing

UNIVERSITY OF THE BASQUE COUNTRY UPV/EHU

Abstract

Engineering School of Bilbao
Department of Communications Engineering

Doctoral Degree

Design and Validation of Novel Methods for Long-term Road Traffic Forecasting

by Ibai Laña

Road traffic management is a critical aspect for the design and planning of complex urban transport networks, for which vehicle flow forecasting is an essential component. As a testimony of its paramount relevance in transport planning and logistics, thousands of scientific research works have covered the traffic forecasting topic during the last 50 years. Most seminal approaches relied on autoregressive models and other analysis methods suited for time series data. However, during the last two decades, the spotlight has shifted to data-driven procedures, as a result of the development of new technology, platforms and techniques for massive data processing under the Big Data umbrella, the availability of data from multiple sources fostered by the Open Data philosophy, and an ever-growing need of decision makers for accurate traffic predictions. Even in this convenient context, with abundance of open data to experiment and advanced techniques to exploit them, most predictive models reported in the literature aim at short-term forecasts, and their performance degrades when the prediction horizon is increased. Long-term forecasting strategies reported to date are more scarce, and commonly based on the detection and assignment to patterns. These approaches can perform reasonably well unless an unexpected event yields unpredictable changes, or if the allocation to a pattern is inaccurate for any reason.

The main contribution of the work presented in this Thesis revolves around data-driven traffic forecasting, ultimately pursuing long-term forecasts. This broadly entails a deep analysis and understanding of the state of the art, and dealing with incompleteness of data, among other issues. Besides, the second part of this dissertation presents an application outlook of the developed techniques, providing methods and unexpected insights of the local impact of traffic in pollution. The obtained results reveal that the impact of vehicular emissions on the pollution levels is overshadowed by the effects of stable meteorological conditions of this city.

Acknowledgements

Three years ago I walked the first steps of this Thesis and I was completely uncertain of what it could entail, but I was sure it was going to be a hard time. A few months before I had just changed my job (I barely knew the basics of road traffic then), and in that change, I met for the first time most of the people that are acknowledged in these lines. I was not aware then that all these people would play such an important role in what has become one of my most satisfying and enriching life experiences.

All my dedication, my efforts and a great part of the inspiration for this Thesis were always meticulously guided by the polar star of this project, Dr.^{2} Del Ser (yes, Doctor squared). I cannot begin to describe the privilege that it has been having such a passionate and dedicated supervisor, and without whose leading I would be probably wandering around and trying to figure out what to do in this PhD race. I am also thankful to my other supervisor Manolo, who guided me efficiently through the convoluted paperwork, and to conduct my efforts in the proper direction.

I am also indebted to Tecnalia and particularly to many of those people that I met three years ago. Iñaki has been tirelessly and selflessly supportive, inspiring and teaching me most of what I currently know about road traffic and its dynamics (and many other things, in fact). I also have to thank Isidoro, for long-term betting on the training of his team and fighting for the available hours that I have employed in developing this work (I cannot imagine how could I have done the work without this time). Joseba, Iñigo, Elena, and all the *council* of ICT division are also to be thanked for believing and facilitating a long-term research strategy. To all the people in OPTIMA, and specially those at SML, who had to share the work that I was not doing because I was dedicating my time to this PhD.

Joint Research Lab (a.k.a *El Aula*) and its collaborators, especially Izaskun and Txus (the *sacred cows* of JRL) have also played an essential role in my achievements. It has been a temple for research, a place to focus and meet people with the same passions. Precisely Txus deserves a distinctive acknowledge, as he has been my PhD companion since we started the Master degree in 2011, and we have experienced together the stay in New Zealand, during which I had the delight of living with his enchanting family.

The people at KEDRI made my stay in New Zealand more enjoyable and academically productive. Among them, I need to highlight the wise guidance provided by Professor Kasabov, the welcoming care that Joyce gave us, the amazing philosophical lunch conversations with Urtats and Lucien, and the data modeling debates with Israel. The stay would not have been the same without the incredible hospitality (and delicious food) that Elisa gave us. She was the cornerstone of our stay.

Beyond the people that have been there during these three years, I think that it is necessary to thank the Madrid City Council for publishing all those data that have been so helpful in my work. I have sought these kind of sources in many cities around the world, but none of them have so detailed and fine-grained catalogs of open data. I think they need to be

taken as an example of what should be done with data that belong to the citizens.

Finally, I have to thank my parents for their constant support and caring through all the stages of my life. I have saved the last word of acknowledgment for the Queen, Irantzu. She has been the moving force of all the important decisions in my life: she encouraged me in 2011 to start my researching activity, to change my job, to start my PhD, and to do the doctoral stay in NZ (I would have never gone so far for so long without her push). Her complicity and belief in me were essential not only to this Thesis but to my whole professional career. I only can hope that I can be to her as much as she is to me.

Contents

List of Figures

List of Tables

List of Abbreviations

Modeling Terms	
ANN	**A**rtificial **N**eural **N**etwork
AR	**A**uto **R**egression
ARIMA	**A**uto **R**egressive **I**ntegrated **M**oving **A**verage
GA	**G**enetic **A**lgorithm Optimization
HA	**H**istoric **A**verage
HS	**H**armony **S**earch Optimization
KF	**K**alman **F**iltering
KNN	**K** **N**earest **N**eighbors
LR	**L**inear **R**egression
ML	**M**achine **L**earning
PSO	**P**article **S**warm **O**ptimization
RF	**R**andom **F**orest
SVM	**S**upport **V**ector **M**achine
TSA	**T**ime **S**eries **A**nalysis

Traffic Terms	
ATIS	**A**dvanced **T**raveller **I**nformation **S**ystems
ATMS	**A**dvanced **T**raffic **M**anagement **S**ystems
ATR	**A**utomatic **T**raffic **R**eader
FCD	**F**loating **C**ar **D**ata
GPS	**G**lobal **P**ositioning **S**ystem
ITS	**I**ntelligent **T**ransportation **S**ystems
TT	**T**ravel **T**ime

Performance Metrics	
MAPE	**M**ean **A**bsolute **P**ercentage **E**rror
MFB	**M**ean **F**ractional **B**ias
NRMSE	**N**ormalized **R**oot **M**ean **S**quared **E**rror
$\mathbf{R^2}$	Coefficient of correlation
RMSE	**R**oot **M**ean **S**quared **E**rror

Chapter 3	
1NN	**1** **N**earest **N**eighbor
BASIC	**Basic** imputation technique
CL	**CL**ustering imputation technique
ELM	**E**xtreme **L**earning **M**achine
INT	Linear **INT**erpolation

MCAR	**M**issing **C**ompletely **A**t **R**andom
MAR	**M**issing **A**t **R**andom
MDV	**M**ean **D**ay **V**ariation
NMAR	**N**ot **M**issing **A**t **R**andom
PCC	**P**attern **C**lustering and **C**lassification
SSC	**S**patial **S**ensing by **C**ontext
SSO	**S**patial **S**ensing by context **O**ptimized

Chapter 4

SNN	**S**piking **N**eural **N**etwork
eSNN	**e**volving **S**piking **N**eural **N**etwork

Chapter 5

RS-EA	**R**oad**S**ide **E**scuelas **A**guirre
RS-BP	**R**oad**S**ide **B**arrio del **P**ilar
RS-FL	**R**oad**S**ide Plaza de **F**ernández **L**adreda
UB-PC	**U**rban **B**ackground **P**laza del **C**armen
UB-AS	**U**rban **B**ackground **A**rturo **S**oria
UB-FA	**U**rban **B**ackground **FA**rolillo
SU-CC	**S**ub**U**rban **C**asa de **C**ampo

Chapter 1

Introduction

Since the beginning of human civilizations history, transportation has been one of their foundations, bringing discovery, interaction and commerce among cultures. In the 19^{th} century, trains paved the way to the industrial revolution allowing not only industries but also cities to flourish and thrive. A few years later, and in a range of only five decades, the generalization of private vehicle owning and the upsurge of a new transportation mode, –the plane–, encompassed, along with other technological and scientific factors, an exponential population growth. People were able then, for the first time in history, to travel to distant countries in hours, to live outside big cities and commute to work, and to obtain almost any good produced anywhere in the world. Of course, all progress hides drawbacks. Nowadays, roughly 100 years after the first car models were produced in a industrial line, the road transportation is one of the main areas that public authorities, from local to continental, have to deal with. From a global point of view, international administrations are trying to reduce the environmental impact of the majority of means of transportation, which still rely in a century old technology, the combustion engine. According to the United States Environment Protection Agency, transportation is responsible of a 14% of the global greenhouse gas emissions[1]. When the scope is set at an urban level, the amount of vehicles traversing any them is massive, and keeps growing. Naturally, most cities are not prepared to keep up to this growing rate, which ultimately means congestion issues will develop.

Traffic congestion leads to social, economic and environmental problems. It degrades the urban landscape of cities, deteriorates the sleep and health of citizens and has a noticeable negative impact in local and regional industry and commerce. For this reason public and private organizations have attempted at addressing congestion for more than 50 years. Efforts devoted to this end have been conducted in three directions [1]: increasing infrastructures, promoting transport alternatives and managing traffic flows. The first direction is limited by topographical, budgetary and social factors. The roads and highways of a city can not keep growing indefinitely. Besides, citizens usually want to live in well-connected areas, but refuse, for obvious reasons, to live near major arterials that support high levels of traffic. The second direction is mainly a matter of developing

[1] https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions

policies (and financing them) that promote and facilitate the use of public transportation, multimodality, or alternative transport modes like car sharing, car pooling or cycling. Stimulating the use of electric vehicles is currently a popular measure slowly adopted in many European cities. Although this measure itself can not tackle congestion problems, it can help alleviating noise and pollution issues. The third direction involves traffic experts and engineering practitioners, and it is aimed at reducing congestion issues by managing the way in which infrastructures are used. To do so, the traffic particularities of each city, must be analyzed, understood and efficiently managed, for which technology has supported a fast growth of data acquisition and decision making systems.

When computers, sensors and communication technologies started to be applied to transportation management, the term Intelligent Transportation Systems (ITS) was coined [2]. ITS are a large set of methods and tools to efficiently manage traffic and make it more fluid and safer. They have two primary subfields: Advanced Traffic Management Systems (ATMS) and Advanced Traveller Information Systems (ATIS), setting the scope on traffic managers and road users respectively. Both ATMS and ATIS have been continuously improved in the last decades with the expansion of data provided by sensors in roads and vehicles, as well as the evolution of technologies required to exploit such data. This allows measuring, modeling and interpreting traffic features such as flow, occupancy, travel times or level of service, which are useful to manage lanes assignment, dynamically set the maximum speed of a road, automate warning and weather messages in information panels, or control traffic signals, among others. On the other hand, they are also a fundamental input for route planners, which allow road users to optimize their mobility.

In this context of sensorized road networks, road usage data started to become available decades ago, granting access to useful insights, and patterns that enabled informed decisions. Besides knowing how traffic behaves under different circumstances, these collected data also opened the door to model traffic and forecast its future behavior, which supposed a great breakthrough in traffic management. Data-driven traffic forecasting has been developing as a research topic since the late 1970s. The first attempts at predicting traffic flows consisted mainly of time-series approaches with different techniques [3]–[6], as well as early explorations of Kalman filtering (KF) methods [7]. Since those beginnings, data availability, analysis methods, and computational capacity have evolved and grown remarkably, along with the interest of the research community in this field. Nowadays, ITS conform a vivid area of research, policy making and technology development, for which one of its foundations is predicting traffic features [8], [9]. Data from road sensors are available with fine-grained resolution, not only posted in regularly updated public repositories, but also made accessible in the form of Floating Car Data (FCD), pedestrian mobility traces, bicycle traffic counts, or traffic lights operation conditions. Forecasting methods have evolved at a similar pace; although a great share of the latest research contributions still relies on time-series analysis, there is also a wide focus on Machine Learning (ML) methods. Simulation tools

have proliferated likewise, allowing for network-wide traffic flow forecasts among many other functionalities.

Regardless the huge advances that have taken place in this thriving field of research, specific forecasting aspects such as the prediction horizon have remained essentially the same for decades. Roughly all traffic prediction literature is oriented towards short-term horizons, despite the usefulness of long-term estimations for road administration purposes and/or for enhancing short-term models as an additional input feature [8], [10]. In works published in the last decade it is possible to find some timid efforts to obtain long-term predictions, but they confront an arduous challenge, due to the stochastic nature of traffic and all the external factors impacting on its dynamics.

1.1 Motivation and Objectives

As stated above, traffic state predictions constitute one of the essential tools of ATMS and a crucial input for ATIS and route planning services. From an intuitive point of view, as any other forecast, the further in the future a prediction is made, the more useful it will be; it is obvious that a traffic manager would have better options to plan and manage a congestion situation if it could be anticipated within 6 hours than if it was predicted only 15 minutes in advance [11]. However, and also as for any other kind of forecast, the prediction horizon is one of the main limitations of traffic forecasting. The degradation of forecast accuracy when the horizon is increased is an assumed common ground in all existing literature [1], [8], [12], which is often attributed to the stochastic nature of traffic itself. The vast majority of models proposed in past literature, from time-series analysis and auto-regressive methods to highly complex and deeply optimized ML models, are built upon previous traffic observations. The possible correlation of the state of traffic at a certain time t and its preceding instants $t-x$ vanishes when x is increased. There are also other factors that hinder this horizon extension, like unexpected events, incidents or changing weather [1]; the more distant a prediction is, the more likely it is to happen any of these traffic modifying events, making the prediction useless. Thus, short-term predictions (most operating under 60 minute horizons) constitute the preeminent body of traffic forecasting literature [13].

Regardless of the above, the pertinence of long-term predictions has been highlighted recurrently in the last decade [10], [14], [15]. The main objective of this Thesis focuses in this long-term forecasting challenge using ML techniques and modeling tools that have become popular among most research fields in the last 10 years, as well as a public traffic data source that provides a real environment to assess the performance of the developed methods. Building this long-term concept, however, requires a previous deep understanding of the traffic forecasting field and its complexities, which includes a profound command of short-term forecasting counterpart models and a broad grasp of data modeling and analysis. Hence, prior to the attainment of long-term forecasts, the following objectives are established:

- **Surveying the traffic forecasting field**: which implies a systematic review of the state of the art on the topic. The area has been subject of meta-studies in many occasions, which gives the opportunity not only to investigate the evolution of the matter, but also to define a taxonomy of the research, and to identify the considerations that are taken as common grounds by the majority of researchers.

- **Achieving state-of-the-art short-term predictions**: the path to accomplish long-term forecasts implies a skilled command of traditional short-term approaches, based on previous traffic variable readings. These are used as one of the key pieces of the adaptive long-term forecasting model, and they are also a fundamental part of the application case. Although this dissertation does not contemplate a specific section for it, the optimization of ML algorithm parameters has been a crucial element of all developed predictive models, so they can stand up to the quality levels found in literature.

- **Dealing with the inconveniences of real traffic data**: using real traffic data is challenging *per se*. One of its main issues is the abundant missing portions of data that can be encountered in the available sources, due to assorted reasons such as reader malfunctioning, network failures and others. These gaps can degrade substantially the outcomes of predictive models, thus dealing with them is indispensable. A whole chapter is devoted to this relevant topic, while other common real data derived troubles like noise are implicitly treated along the dissertation.

- **Taking advantage of the network relations of traffic**: network-level predictions were identified as a challenge of the traffic forecasting field a few years ago [13]. Since then, there has been a significant increase of works dealing with this kind of forecasts. However, most research devoted to this question is focused in the upstream-downstream correlations, obviating other kind of relations that can exist among traffic profiles in different nodes of the network. Finding these can be useful in urban contexts, where travel times between consecutive measuring nodes are considerably lower than the data capture frequency.

- **Including external factors to traffic prediction**: some circumstances independent of traffic like weather, planned events (such as roadworks or sports events) and unexpected ones (such traffic incidents), can have a great impact on it and the accuracy of its predictions. Still, these factors rarely are part of traffic prediction models. Blending data from different sources together into a single model can help obtaining more reliable forecasting methods.

- **Using the traffic predictions on an applied context**: a prescription case study where the traffic forecasting is at the service of a higher end. In this case, it is used to provide insights on the impact

of local sources of pollution (vehicular traffic) to the air quality of a city, helping authorities to take informed decisions about possible restrictions.

During the development of this Thesis, all of these objectives are addressed individually or jointly. They can be grouped into 4 main blocks that can be considered as the minimum workflow of any data science endeavor: **acquiring the knowledge of the field**, **preprocessing data**, **modeling** –in this case for prediction–, and **prescription**, or exploiting the insights provided by models in real life scenarios.

1.2 Outline and Contributions of the Thesis

This Thesis is structured by following the previously outlined scheme, with four chapters that address each of the main blocks and a concluding chapter that presents the final remarks and future lines of work. Chapters 3 to 5 will be referred to as *experimentation* chapters in the rest of this introduction, as they contain empirical experiments. A brief summary of each chapter is introduced below.

1.2.1 Chapter 2

This chapter aims to summarize the efforts made to date in previous related surveys towards extracting the main comparing criteria and challenges in this field. A review of the latest technical achievements in this topic is also provided, along with an insightful update of the main technical challenges that remain unsolved to date, with an emphasis placed on issues related to managing large sets of data. The ultimate goal of this chapter is to set an updated, thorough, rigorous compilation of prior literature around traffic prediction models so as to introduce the context for the rest of this dissertation, which indeed attempts at tackling some of the identified challenges. The provided analysis is also intended to be an interesting contribution to motivate and guide future research on this vibrant field.

1.2.2 Chapter 3

This chapter concentrates in the preprocessing stage required to build a forecasting model with real data. Traffic predictive models generally rely on data gathered by different types of sensors placed on roads, which occasionally produce faulty readings due to several causes, such as malfunctioning hardware or transmission errors. Filling in those gaps is relevant for constructing accurate models, a task which is engaged by diverse strategies, from a simple null value imputation to complex spatio-temporal context imputation models. The work presented in this chapter work elaborates on two ML approaches to update missing data with no gap length restrictions: a spatial context sensing model based on the information provided by surrounding sensors, and an automated clustering analysis tool that seeks optimal pattern clusters in order to impute values. Their performance is

assessed and compared to other common techniques and different missing data generation models over real data captured from the city of Madrid (Spain). The newly presented methods are found to be fairly superior when portions of missing data are large or very abundant, as it occurs in most practical cases. The contributions of this chapter are not restricted to the proposed imputation methods, but it also supplies valuable insights about the way in which imputation methods should be evaluated.

1.2.3 Chapter 4

Chapter 4 tackles the long-term estimation problem, presenting some of the existent long-term strategies and proposing new solutions to deal with the typical long-term estimation issues. Specially in urban contexts, this kind of forecasting approaches is scarce and commonly based on the detection and assignment to traffic patterns. These approaches can yield reasonably good results unless an unexpected event provokes unpredictable changes, or if the allocation to a pattern is inaccurate, due to the noise within the analyzed data. This chapter introduces a method to obtain long-term pattern forecasts and adapt them to real-time circumstances. The contributed method takes advantage of the architecture of evolving Spiking Neural Networks (eSNN) to perform online adaptations without retraining the model. Its performance is assessed over a real scenario with the same data of the center of Madrid, with observations taken each 5 minutes during a period of 6 months. Significant accuracy gains are obtained when applying the proposed online adaptation mechanism on days with special, non-predictable events that degrade the quality of their long-term traffic forecasts.

1.2.4 Chapter 5

This chapter presents an application case study in which traffic and its predictions are used to characterize the pollution registered in a city. Road traffic is one of the main sources of air pollutants, though topography characteristics and meteorological conditions can make pollution levels increase or diminish dramatically. In this context an upsurge of research has been conducted towards functionally linking variables of such domains to measured pollution data, with studies dealing with up to one-hour resolution meteorological data. However, the majority of such reported contributions do not deal with traffic data or, at most, simulate traffic conditions jointly with the consideration of different topographical features. The aim of the study presented in this chapter is to further explore this relationship by using high-resolution real traffic data. It describes a methodology based on the construction of regression models to predict levels of different pollutants (*i.e.* CO, NO, NO_2 , O_3 and PM_{10}) based on traffic data and meteorological conditions, from which an estimation of the predictive relevance (importance) of each utilized feature can be estimated by virtue of their particular training procedure. The study was made with one hour resolution meteorological, traffic and pollution historic data in

roadside and background locations captured over 2015. The obtained results reveal that the impact of vehicular emissions on the pollution levels is overshadowed by the effects of stable meteorological conditions of this city.

1.2.5 Chapter 6

The last chapter of this Thesis summarizes the concluding remarks that have crystallized after these years of research. This chapter also details the quantifiable results of this Doctoral Thesis: a list of contributions submitted to specialized journals and conferences. It also compiles some of the future lines of research that could be of great interest for the community.

1.2.6 Appendix A

The research procedures described in this dissertation, as well as in related research not explicitly included in it, have completely relied on a public open source of traffic data. This appendix is a reference guide of the infrastructure used to capture and provide these data, and the shape and granularity in which they are supplied.

1.3 Reading this Thesis

The contents of this Thesis can be read in a non sequential fashion. Although it has been arranged in a logical order that follows the writing rationale described above, the experimentation chapters (3, 4 and 5) have no interdependence and could be read in any order. Besides general conclusions, Chapter 6 summarizes any of the other chapters, so the reading could be done by alternating each main chapter and the last one. Figure 1.1 provides the logical dependencies among chapters and the possible alternatives for their convenient reading.

Reading Chapter 2 would be beneficial to the understanding of the remaining material, for the context it provides; however, this is not required, and a reader who is expert in the traffic domain could skip it or use it as a lookup reference. Appendix A is useful to understand the data of experimentation chapters, which can be also used as a reference if the reader finds it necessary.

1.3.1 Notes on the Formulation

Some notation conventions have been established to describe the formulation of experimentation chapters. A list of the chosen notational principles is provided below, in order to facilitate comprehension:

- **Indexes**: denoted with a small Latin letter, usually starting with i, or t for temporal indexes. Temporal indexes can have subscripts denoting the index of the day they belong to.

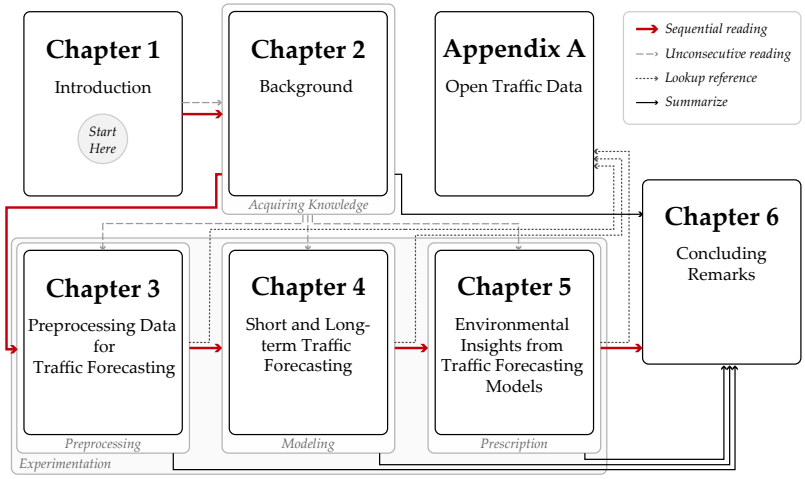

FIGURE 1.1: Block diagram of the relationships between chapters.

- **Observations**: individual traffic readings are denoted with letter o. Superscript denotes its index of sensor and subscript the temporal index (o_t^i).
- **Vectors**: vectors of traffic readings, usually representing one day of observations are denoted with small bold letter, *e.g.* $\mathbf{o}$. Superscripts specify index of sensor and index of day they belong to ($\mathbf{o}^{s,d}$). Vectors not related to traffic readings are denoted using bars over Latin letters, *e.g.* $\overline{W}$.
- **Datasets**: denoted with a calligraphic capital Latin letter, *e.g.* $\mathcal{H}$. They can have same superscripts and subscripts as individual observations, *e.g.* $\mathcal{G}^s$. When dataset instances are vectors, they are represented with bold calligraphic capital letters: $\boldsymbol{\mathcal{H}}$.
- **Functions**: denoted with small Greek letters.
- **Constants**: denoted with capital Latin letter, *e.g.* P.
- **Tildes ˜ and hats ˆ**: These modifiers are added to observations, vectors or datasets to specify a modification performed on them. For instance, in Chapter 3, $\mathcal{G}$ has two variants: $\tilde{\mathcal{G}}$ with artificially generated gaps, and $\hat{\mathcal{G}}$ with data imputed to those gaps.

1.3.2 Notes on the List of Abbreviations

The list of acronyms section does not compile all the acronyms used along this Thesis. The number of them is too large, partly due to the exploration of the state of the art on traffic forecasting, which yields a large number of methods, which in addition need to be put in a table, and thus their name must be abridged. Some of those methods are more common, and even used in the experimentation presented in this dissertation, while some

others are used only once. For this reason, and trying to provide a list of acronyms that can be useful as a quick reference, the following criteria have been considered to include acronyms in it:

- For machine learning modeling terms and traffic acronyms, only those terms that have been used more than once along the whole Thesis are included. Methods that are cited only once in a chapter are not included here, but can be found in the same chapter. As they are not ubiquitous, the reader will be able to find what they stand for in the neighboring paragraphs.

- For terms related to performance metrics, only those that are used in at least one of the experimentation chapters are included. If any term of this kind is used in other section, it is explained *in situ*.

- For each experimentation chapter there is a section that contains those acronyms that are used only in that particular chapter. In these cases, the acronyms are defined locally, but as they are used extensively across each chapter, they are included in the index for the reader's convenience.

Chapter 2

Background

For its applied interest, the field of traffic forecasting has been object of extensive research exploring its numerous dimensions. This chapter is intended to distillate the essence of the field by systematically examining the recent developments that gravitate on data-driven traffic forecasting methods. Among them, an emphasis is set on the technical advances made in recent years, the degree of achievement of the different technical challenges posed in the past, and a critical diagnosis of the unexplored areas of research that should be targeted by the community in forthcoming years. To this end the chapter capitalizes on recent surveys and enriches them with an analysis of the research trends, detected issues and identified challenges springing from the newest works in this field. The survey ends up by tracing the research niches that deserve further attention and efforts, solidly founded on the arrival of methodologies and tools related to Big Data, as well as on the characteristics of this particular data source. The discussions held through this chapter are intended to provide a comprehensive report of the current state of the art of traffic forecasting for early researchers and engineers, as well as to stimulate and steer future technical contributions towards directions of lacking maturity and reasoned potential.

2.1 Short-term Road Traffic Forecasting: A Historical Perspective

The research activity and contributions dedicated to the development of traffic forecasting methods during the last three decades are assorted and can be classified under very diverse criteria. Indeed, a selection of works in the last ten years have sorted and classified the existing literature on traffic forecasting by adopting very diverging perspectives. As such, Van Arem et al. in [1] explored applications of traffic forecasting to dynamic traffic management by analyzing the overall process from an economical demand and supply approach; the first can be represented by origin-destination flows and is considered a human behavior concern, and the latter stands for the ability of the road network to satisfy the demand. This human aspect of traffic processes is relevant for ATMS and the forecasting methods themselves, which need to have in account that larger demands imply lower

supplies (congested roads), which in turn leads to lower demands due to informed drivers who rearrange their planned routes. This noted impact of traffic conditions on their near-future selves and the stochastic nature of traffic [1], [12], [16] contribute to the fact that short-term forecasting is the main research focus of this and practically every other work on this subject.

Aside from the prediction horizon, this early review is concentrated on forecasting methodologies, performance evaluation techniques and real-world application examples. Besides, the authors provide a series of challenges that include representation, model validation, incorporation of human behavior to the model and design of the optimal monitoring network, among other aspects of relevance for the topic. As the authors suggested, the traffic forecasting field was "in its infancy" (*ad pedem literae*); from then on, a sprawl of research started and allowed [8] to review the subject in a more deep and principled manner. In this review the authors considered three main aspects of traffic forecasting: scope of application, output specification, and modeling features. The first aspect refers to the kind of roads for which the prediction is made and the type of application the prediction will be used for (mainly ATMS and ATIS). On the other hand, the output specification involves the prediction horizon and step concepts, and elaborates on which traffic parameters should be considered for developing the predictive model. The authors also provide a profound analysis of the modeling aspects of traffic forecasting. A classification of prediction models is presented, comparing their characteristics and performance. The authors contribute a methodological workflow to select and tune the model parameters that has been extensively referred [12], [17]–[22].

Although most of the reviewed literature is focused in road features forecasting, specially traffic flow or volume, travel time is an alternative variable to predict. It is more human-understandable than flow, occupancy or speed, where a given value may have different interpretations depending on the type of road under analysis. A survey on travel time prediction was published in [23], exposing the techniques and main drawbacks and difficulties to estimate this traffic feature. According to this study, in 2005 the main handicap for this specific traffic forecasting scenario was concluded to be the lack of data, which the authors proposed to overcome with simulation techniques. Nowadays, the widespread proliferation of GPS-enabled devices and connected vehicles that can supply FCD allows researchers to model and predict travel-times without resorting to simulation methods [24]–[27]. A more recent review on this field [28], delves in the latest travel-time prediction from FCD sources, and highlights the relevance of forecasting this feature for ATIS and operational planning of mass transit.

In 2007, a concise work on prediction modeling [12] delved into understanding and examining forecasting models proposed by previous research works. The authors considered a naïve category (predictions based in historic values or an average of them) in addition to the parametric/non-parametric category described by [8], and classified traffic simulation as a parametric method that can be used for traffic forecasting. The main

comparison features were prediction horizon, scope of application, computational speed and accuracy, most of them aligned with previous surveys. Interestingly, network-wide predictions and a comparison method between simulation and the rest of models were first suggested as open challenges in this work. More recently, Bolshinsky et al. in [18] have agreed on the latter, stating the difficulties to compare not only simulation to every other forecasting method, but all methods to each other.

In this chapter the main focus is set on techniques, adding a few not contemplated in previous reviews. Besides the forecasting method and the prediction horizon, the authors highlight the relevance of sources of data, as studies with different input data can be hardly compared. This work also ponders the pertinence of input data other than traffic data. An assortment of non-traffic factors affect traffic conditions, such as weather conditions, holiday periods, events, incidents and road works, or seasonal factors. Including these elements in a forecasting model can aid to refine predictions, but as reflected in [18], only a few previous authors have them into account. In the same year, [16] updated the taxonomy of forecasting methods provided by the same authors 5 years earlier [12] with more recent works. They conclude, as suggested by other authors before, that there is no universal method that fits every situation better than the rest. As a consequence, they propose as a challenge the development of model ensembles that outperform the existing ones, but more interestingly, the creation of a method to choose a technique, given the attributes of the forecast to make. A different view of the subject is proposed in [29], which offers a mathematical optimization perspective of the problem. This review summarizes the efforts made in this direction and states that all observability, estimation and prediction problems can be integrated and formulated as an optimization paradigm.

One of the most cited reviews on traffic forecasting is the one by Vlahogianni et al. in [13]. They studied literature on traffic forecasting since 2004, comparing scope of application, prediction horizon, input sources and methodological approach. According to all previous research contributions, they concluded that little effort had been dedicated to network-wide predictions, urban arterial forecasting, or multivariate models; besides, in most previous works predictions were made for traffic volume. Leaning on the ongoing development and expansion of data driven techniques and technologies, they gathered a series of challenges for future work in this area. Part of them coincide with previous surveys and the aforementioned conclusions; the key aspects to develop identified in these works are arterial and network-level predictions, shifting from traffic volume to travel-time prediction, spatio-temporal forecasts, model selection techniques and model comparison methodologies [8], [12], [16], [18], [23]. In addition to those recurrently formulated questions, the authors proposed new challenges related to data fusion, data aggregation, the explanatory capacity of variables and the responsiveness of the predictive models. A travel-time forecasting survey recently contributed in [30] portrays the main methods and issues in this field. Conclusions highlight similar relevant aspects: network-wide prediction, exogenous factors, data fusion, and the relevance

of congestion situations for this kind of forecasts. Oh et al. in [31] review the data-driven approach of travel-time prediction by focusing on methods and techniques that complement the previous references.

2.1.1 A Taxonomy of Research

Literature reviews have so far analyzed the state of the art around traffic forecasting under different criteria. Table 2.1 summarizes the key criteria considered by the aforementioned reviews. This table reveals that the most used criteria in literature reviews are those related to prediction methods, horizon, scale and output variables. Forecasting techniques are a very relevant part of the reviewing methodology, but comparing their performance is a more complex task, as usually each type of technique delivers different performance metrics. Criteria such as the *optimization type*, the *data resolution* or the *streaming of data for online learning* are addressed only in most recent surveys, which cover a wider spectrum of contributions concentrated on data-driven approaches. Prediction horizon and context are likewise significant criteria for most authors, and they are indeed widely used with comparing purposes. Nonetheless, prediction context allows for little comparative; as highlighted by [13], most models are built on freeway or highway contexts, while urban arterial traffic forecasting is less addressed and much more challenging [32]. Prediction horizon is kept under one hour in most works and is varying enough not to be considered as a fair comparison factor.

Accounting for which non-traffic inputs are taken into account to make predictions is an interesting criterion that was first proposed by [1], but it has been rarely used ever since. A recent review by [18] shows that very few works have addressed these factors, which might be the reason why other reviewers do not consider them as a comparative criterion. In addition, calendar-related aspects are implicit in time-series models, which constitute a great part of forecasting literature. However, these and the other factors listed in Table 2.1 are proven highly relevant in forecasting [32]–[34], with a considerable impact in real traffic conditions. Growing open-data initiatives facilitate access to a variety of inputs for future models. This also influences the *data sources* criterion, which is used as comparison element only in most recent reviews, implying a more data-centered corpus of literature. Prediction models are often built upon data from one source (mainly traffic loops, and road sensors), but it might be difficult to compare a model with ATR input data to another with traffic camera data. In this line, data fusion techniques allow to combine data from different sources in the same model, including traffic data from different sensors, or non-traffic inputs [35], and it is considered by [13], [36] as one of the main challenges in this field.

Computational effort aspect was only considered in [1], back in 1998; this aspect has become less relevant since then. Generalizability was deemed as a key feature of simulation models. Their need for parametrization renders them too specific; they improve if they are applicable to other

contexts. This is seldom analyzed for non-simulation models, but if standard test-beds and test data were to be available for testing and comparing algorithms, as suggested in [13], their results might be more easily generalizable. As data availability increases, more attention is paid on the type of learning process, which can rely on online streams of data [28].

TABLE 2.1: Prediction aspects assessed in previous reviews.

Criteria	**Types**	**Referenced in**
Prediction Method	Naïve (instant, historic average)	[12],[16],[18],[30],[37]
	Parametric (simulation, time series, Markov chains, Kalman filters)	[1],[8],[12],[16],[18],[13],[30],[37]
	Non-parametric (Neural Networks, bayesian, fuzzy logic, ATHENA)	[1],[8],[12],[16],[18],[13],[30],[37]
Prediction Horizon	Prediction horizon	[1],[8],[12],[16],[18],[13],[37],[38]
	Prediction Step	[8]
Prediction Scale	Single Location	[1],[12],[16],[18],[13],[30],[37]
	Road Segment	[1],[12],[13],[30],[38]
	Whole or part of the network	[1],[8],[12],[16],[13],[30],[38]
Prediction Context	Urban (arterial)	[1],[38],[8],[12],[13]
	Rural	[38],[12]
	Freeway	[1],[38],[8],[12],[13]
Data sources	Traffic management bureaus	[1],[18],[37]
	Automatic Traffic Recorders	[1],[18],[30],[37]
	Sensors (other than loops)	[18],[30],[37]
	Cameras	[1],[18],[30],[37]
	GPS-FCD	[18],[30],[37]
	Cellphone data	[18]
	Public transport information	[18]
	Crowd sourcing	[37]
	Social media	[37]
Exogenous factors	Calendar (Week days, weekends, bank holidays)	[1],[16],[18]
	Periods of the day	[1],[16],[18]
	Holidays	[1],[16],[18]
	Seasonal differences	[1],[16],[18]
	Weather	[1],[18],[13]
	Special events (demonstrations, parades)	[1],[18]
	Periodical events (sports, other social events)	[1],[18]
	Road Works	[1],[18]
	Traffic incidents	[1],[18],[13]
	Traffic source areas (malls, parkings, adjacent roads)	[1],[18]
Predicted variable	Traffic Flow (vehicles/hour)	[1],[8],[12],[16],[18],[13],[38]
	Traffic Density (vehicles/km)	[1],[8],[12],[16],[13],[38]
	Average speed	[1],[8],[12],[16],[13],[38]
	Travel time	[1],[8],[12],[16],[13],[30]
Uni/Multivariate	Not applicable	[1],[8],[18],[13]
Prediction performance	Not applicable	[1],[12],[38]
Optimization type	Not applicable	[13]
Computational effort	Not applicable	[12]
Generalizability	Not applicable	[38]
Scope of application	ATIS, AMTS, Logistic	[1],[13],[38]
Data resolution	Not applicable	[13]
Stream mining	Online, offline	[28]

Thus, plenty of aspects can be studied when evaluating and testing traffic forecasting methods, with different levels of relevance considering their abundance of use. The following section will describe the main issues found by researchers in traffic forecasting, in order to determine if the relevance of aspects is related to their impact in solving the issues, or some other aspects could be considered.

2.1.2 Some Well-Established Considerations

Reviewing traffic forecasting literature has lead preceding researchers to diverse conclusions, future directions and challenges, which are next reviewed and analyzed. A fair share of them is maintained through the years, despite the advances in some of the fields involved, e.g. computer processing capacity, machine learning algorithms, simulation tools and access to data.

2.1.2.1 Stochastic Nature of Traffic

This feature is frequently stated in all kinds of traffic forecasting literature, and its effects have been considered by transport modelers since the first works. Traffic predictions have two natural limits: one is related to the randomness of events that can affect traffic; the other is related to the effect predictions themselves can have on drivers' decisions and habits [1].

2.1.2.2 Network Application

A conclusion common to all reviews is the lack of network-wide prediction models, compared to single-point or road segment predictions. The latter are useful for ATMS, but the first help building more efficient ATIS [16], and thus can reach to the general public. Since the earlier reviews, when this was considered a network sensor coverage design problem [1], the subject has arisen in every review. This issue is related to the urban traffic prediction problem, for network predictions are more useful in urban environments. Network-wide predictions have been explored mainly via simulation [12], [16], although other approaches can be found in literature [39], [40]. There is an increasing number of this kind of predictions in recent works (as later shown in Tables 2.2 and 2.3), but they are usually referred to compact areas. Network-wide prediction models still remains a challenge in this field.

2.1.2.3 Urban Traffic

Predicting urban traffic is defined in previous reviews as a very complex task. Signals, interactions with close links, sources of traffic, and a more complex origin-destination relation have made some researchers state that traffic flow in urban arterials -even single location traffic- cannot be predicted as accurately as in freeways [32]. Again, simulation models are the best suited to address this issue, parametrizing preceding inputs; notwithstanding, some researchers have concentrated on this matter by incorporating spatial information to time series [32], neural networks [41]–[43] or other non-parametric approaches [19], [44]. Arterial traffic prediction has grown in interest, yet it constitutes a slight portion of works on traffic forecasting [13], and embodies along with network prediction one of the main challenges of forecasting.

2.1.2.4 Applicability and Model Selection

Or the ability of a model to adapt to different contexts. This is a typical simulation problem [38]; the more parameters are set, the more difficult it is to apply the simulation model to other circumstances. Non-simulation methods, and specifically non-parametric methods adapt better to different contexts [18], but all previous reviews coincide in asserting there is no best method that suits all situations [8], [12], [13], [16], [18], which implies an applicability at a higher level, not of the model, but of the method to choose the most suitable model given the characteristics of the forecasting problem [16]. A recent work [45] proposes a meta-modeling technique to tackle the model selection and parameter tuning. A lot of research is yet to be made to take advantage of these techniques, and even extend them to greater decision making tools.

2.1.2.5 Metrics of Performance and Comparison

Abundance of methods for traffic forecasting makes establishing a unified comparing metric an intricate task [1], [8]. Although Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) are usual metrics for model performance measuring [12], [16], [46], they do not provide comparable measures when the complexities of compared models are too divergent [9] – e.g. a neural network and an Autorregresive Integrated Moving Average (ARIMA) model – or when the input datasets are completely different. Also, for network-wide models, errors can propagate through time and space along the network, so a spatio-temporal correlation between successive predictions can help measuring the performance. Although contemporary studies keep using the Root Mean Square Error (RMSE) and the Mean Absolute Percentage Error (MAPE) to assess performance, the definition of benchmark datasets, environments and metrics is currently identified as a necessity in the area [13].

2.1.2.6 Hybridization of Methods

Combining prediction techniques is a tendency that was explored in the first place with combination of ARIMA models with other methods to improve accuracy, and more recently combining the prediction method with techniques to preprocess data like clustering [33], [34] and to optimize the model [13]. Current works frequently combine methods with different purposes, and in recent research is common to find several types of optimizations for tuning the models (see Tables 2.2 and 2.3). Tselentis et al. [47] showed that combining models with different degrees of spatio-temporal complexity and exogeneities is most likely to be the best choice in terms of accuracy. Moreover, the risk of combining forecasts is lower than the risk of choosing a single model with increased spatio-temporal complexity.

2.2 New and Revisited Challenges

Existing background of traffic forecasting literature and reviews present countless efforts devoted to tackle foregoing challenges. This section is aims to broaden the scope of previous questions and present the most recent developments in them. The latest survey in [13] presented a vast literature review up to 2014, and compiled challenges in ten global categories. We delve into some of those challenge categories, and propose future lines of development in the field of traffic forecasting, having in mind the shift to data driven machine-learning techniques that is also mentioned in [13].

2.2.1 The Context of Forecasting

By *context* we refer to the forecasting setting and the traffic parameters used. As can be observed in Tables 2.2 and 2.3, scope-context column displays the context of application (urban or freeway) and the scope (point, scattered points, segment, network). A significant increase in network predictions is revealed since previous state of the art, which considered it a challenge. Around 1 out of 20 of works examined in [13] had network-wide coverage. Since their review, the proportion has risen considerably. Network-wide predictions usually require the model to know the influence of surrounding road links. This spatio-temporal correlation is mirrored in corresponding column (ST), and it shows that which was considered a challenge in 2014 has started to develop in current works. Data driven approaches facilitate the inclusion of spatial and temporal correlations in the model, and k-Nearest Neighbor (kNN) models are widely used [48]–[51]. Origin destination matrices are estimated to make predictions in [52], and context-aware models that modify forecasts depending on surrounding link predictions are employed in [53]–[56].

Travel-time replacing flow as predicted value was also considered a relevant challenge, as the first is a more helpful metric of traffic. While predicting volume can rely only in traffic counts, estimating travel-time is more complex [27]; although many attempts have been made to estimate travel-time from traffic counts [57], [58], availability of other data such as FCD [28], [59]–[61] or vehicle identification [54], allows to build trajectories and perform improved estimations [62]. The relative amount of works centered in travel-time prediction is maintained similar to that found in previous reviews. Data sources column shows precisely that even when input data fusion is considered a challenge, and data can be obtained in many ways, most of works are still concentrated on traffic loops. It is possible to observe, though, some studies that fuse inductive loop readings with camera information, FCD or automatic number plate recognition (ANPR) [63]–[65] and an upsurge of the use of FCD, more reachable, and more exploitable with data driven methods.

2.2.2 Long-term Prediction Horizons

Increasing the prediction horizon is not usually regarded as a challenge. It is commonly assumed that predictions degrade when this horizon is

extended [1], [8], [12], but long-term predictions can be useful from the ATMS perspective [11], [66], or for scheduling in logistics planning [23]. The relevance of long-term predictions is claimed also for macroscopic network planning [14] for infrastructure development [67]. Besides, although long-term predictions cannot provide accurate outputs, they have been proposed as another input to short-term prediction models [10], [15], [68]. Notwithstanding, this characteristic is kept in almost every work before 2014 below 60 minutes [13] in the future. Greater accurate forecasting horizons might assist the progress of traffic management systems, and their achievement is connected to better prediction models, more and more linked data sources and spatial context aware predictions.

In terms of forecasting horizon, Tables 2.2 and 2.3 show that most of recent works also predict traffic features under a 60 minute extent. Nonetheless, an increment in larger horizons or non-horizon models is noticeable, paired with a decreasing trend in ARIMA (and variants) models. Data driven models based on large data bases of readings are broadening in recent years, allowing researchers to make heuristic predictions at any point in the future [33], [34], [53], [69], [70], and data extension reaches to 5 years of data [65]. Recent literature shows that forecasts further than 60 minutes are possible, in the data-driven context, and can perform as good as short-term.

2.2.3 Exogenous Factors in Multi-Input Models

The main obstacles for accurate long-term and even short-term prediction are factors that affect traffic but are not part of its seasonal behavior and confer traffic its stochastic nature [1], [12], [16]: road works, incidents, events, weather, proximity to traffic affecting facilities (parking lots, shopping areas, work/study centers), and calendar matters (bank holidays, weekends). Although incidents or weather changes can happen suddenly and they can be difficult to predict, other traffic affecting factors like road works or events are usually foreseeable. Feeding these kind of inputs to data-driven prediction models can enhance their performance, enriching the provided forecasts [18]. Anticipated by [12], mobility data sources and availability have been increasing since then. This entails a chance and a challenge regarding data fusion and integration [1], [13], [36]. Integrating exogenous factors from different data sources is a direction that should be considered in future works.

Exogenous factors continue to be barely addressed in recent research, and calendar information is the most considered one [10], [34], [48], [50], [59], [65], [70], [77], [88], [89], [94], [105], [108]. Weather is only used in three works [64], [79], [83], with less predictive relevance in the model than expected. Weather, pollutants [115], noise [79] and incidents [53], are variables that need to be predicted too. A model can learn how traffic behaves when an incident happens, but it will need a forecast of future incidents to elaborate the output. Incident forecasting is addressed in [116] via Bayesian Neural Networks and detection through other traffic features. This work arises the many facets involved in incident prediction.

Table 2.2: Literature of the 2014-2016 period (1/2).

Reference	Scope context	Predicts	ST	Max. Horizon	Data extension	Step	Data Source	Exogenous Factors	Prediction Model	Comparative models	Ageing
[68]	U-Sgm	Speed	×	2 min.	1 month	2 min.	RTMS	-	LSTM-NN	SVM, ARIMA, NN, KF	Memory Block
[63]	F-Sgm	Speed, TT	×	1 hour	1 month	15 min.	Camera, FCD, Loop	-	NN	GAM, ARIMA	×
[34]	U-Pts	Flow	×	NH	1 year	15 min.	Loop	Calendar	Cluster+RF	HA	×
[71]	U-Sgm	Volume, TT	×	2 days	15 days	-	Loop	-	NN	LR, ARIMA	×
[48]	U-Ntw	Speed	✓	1 hour	1 month	5 min.	FCD	Calendar	KNN	HA, LSSVM, NN	×
[10]	F-Sgm	Flow	×	1 week	1 week	10 min.	RTMS	Calendar	FNR	SARIMA, NN, LSSVM	×
[72]	U-Pts	Flow	×	3 days	15 days	1 hour	Loops	-	Block Regression	LR, SARIMA	×
[52]	U-Ntw	Flow	✓	30 min.	-	5 min.	Loops, Simulation	-	LLDR	⟲	×
[53]	U-Ntw	Speed	✓	15 min.	5 years	1min.	Loops	Incidents	Context Aware Ensemble	⟲	Reward-regret
[73]	U-Pts	Flow	×	5 min.	1 month	5 min.	Loops	-	MVLR	⟲	×
[74]	U-Sgm	Congestion	✓	-	-	-	Simulation	-	MCS	⟲	×
[75]	U-Pts	Flow	×	1 hour	3 months	1 hour	Loops	-	NN Ensemble	Basic, HA	×
[76]	U-Pts	Speed	✓	30 min.	-	30 min.	Simulation	-	NN	⟲	×
[77]	F-Pts	Flow	×	5 min.	5 days	5 min.	Manual	Calendar, Type of Vehicle, Speed	NN	⟲	×
[78]	U-Pt	Flow	×	1 day	3 days	10 min.	Camera	-	SARIMA	Basic, HA	×
[64]	U-Pts	Flow	×	NH	1 year	15 min.	CCTV, Laser, Loops	GIS, Weather	Cluster + NN	Basic, HA, SVR, MVLR, KNN, NLR	×
[79]	U-Ntw	Flow	✓	1 hour	3 months	4.5 min.	Camera	Weather, pollutants, noise, FCD	CRF	⟲	×
[80]	F-Pt	Flow	×	2 min.	1 day	2 min.	RTMS	-	AKNN-AVL	KNN, ARIMA	AVL Case Database
[81]	F-Pt	Flow	×	4 hours	7 days	5 min.	Loops	-	Volterra	RFBNN	×
[82]	F-Sgm	Flow	×	1 day	1 day	1 min.	RTMS	-	MLPNN	RBFNN, WavenetNN	RL-Difference
[83]	F-Pt	Flow	×	10 min.	1 month	10 min.	Radar	Weather	CCGA-GSVR	SVR, WASVR, PSO-BP, ARIMA	×
[49]	F-Sgm	Flow	✓	30 min.	2 months	-	Loops	-	KNN	NN	×
[84]	F-Ntw	Flow	✓	2 min.	1 day	2 min.	RTMS	-	Custom TSA	ARIMA, LR	×
[55]	U-Ntw	Flow	✓	1 hour	1 month	5 min.	Loops	-	STRE	ARMA, STARMA, NN	×
[85]	U-Ntw	Flow	✓	15 min.	-	5 min.	Simulation	Speed	LCG-BN	ARIMA, NN, BN	×
[86]	U-Pts	Flow	×	15 min.	25 days	15 min.	Loops	-	GPDM	KNN, RW	MTLNN
[87]	U-Pts	Flow	×	1 hour	7 days	-	Loops+manual	-	Fuzzy Logic	NN, LR	×
[88]	F-Pt	Speed	×	10 min.	2 months	2 min.	Loops	Calendar	HPSO-NN	ARMA	Error compensation
[89]	U-Ntw	TT	×	3 days	15 months	-	Public transport info	Calendar	EnsemblePPR, SVM, RF	-	×
[59]	U-Ntw	TT	×	NH	-	-	FCD	Calendar, time	Markov chain		×
[90]	F-Sgm	TT	×	NH	15 days	5 min.	Loops	Simulation, occupation, volume, speed	Evolving Fuzzy NN	MLR, Basic, LM, NN, CP	×
[60]	F-Pts	TT, Speed	×	1 min.	24 days	1 min.	Loops	FCD	Grey Model EFGM	TSA	×
[91]	F-Pts	Flow	×	4 hours	3 months	5 min.	Loops	-	ML-KNN	ARIMA, KNN, Basic	×
[56]	F-Sgm	TT	✓	25 min.	93 days	1 min.	Loops	-	KF	-	×
[92]	F-Ntw	TT	✓	1 hour	383 days	10 min.	Loops	-	NN, LR, RTree, RF	⟲	×
[93]	F-Sgm	Flow	×	1 day	15 months	0.5 min.	Loops	Workzones data	LLR	-	×
[94]	F-Pt	Flow	×	1 day	7 days	5 min.	Loops	Calendar	Wavelet NN	⟲	×
[95]	F-Sgm	Flow	×	5 min.	1 month	5 min.	Radar	-	GRNN	Vector Regression, HA	×
[96]	F-Pt	Flow	×	5 min.	7 days	6 min.	Loops	-	Cluster + NN	-	×
[97]	F-Pt	Flow	✓	5 min.	-	5 min.	RTMS	-	Fuzzy NN	-	Real Time correction
[50]	F-Sgm	Flow	✓	1 hour	6 weeks	15 min.	Loops	Calendar	MKNN	HA, SARIMA, NBR, KNN	×
[70]	F-Pts	Flow	×	NH	82 days	5 min.	Loops	Calendar	Similarity, RBLTF	ARIMA, GRNN	×

TABLE 2.3: Literature of the 2014-2016 period (2/2).

Reference	Scope context	Predicts	ST	Max. Horizon	Data extension	Step	Data Source	Exogenous Factors	Prediction Model	Comparative models	Ageing
[98]	F-Pt	Flow	×	10 min.	6 days	15 min.	Loops	-	ElmanNN+Bayes	BPNN, Wavelet NN	×
[99]	F-Pt	Flow	×	1 hour	9 days	5 min.	Loops	-	ieRSPOP	-	ieRSPOP
[100]	F-Ntw	Flow	×	30 min.	2 months	5 min.	Loops	-	DTT1, DTTinc	ARIMA, SVR	DTT
[101]	F-Pts	Flow	×	2 days	1 month	1 hour	Loops	Speed, occupancy	TS-TVEC (TSA)	MLPNN	Error correction
[102]	F-Pts	Incidents	×	15 min.	3 weeks	5 min.	Loops	Flow, occupancy	Drift3Flow	ARIMA, ETS	Concept drift
[103]	F-Sgm	Flow	×	30 min.	1 month	5 min.	Loops	Density	PHFRBS + GA	⟲	×
[104]	U-Ntw	Speed	×	15 min.	6 days	5 min.	FCD	-	BN, NN, SARIMA+BN	HA	×
[105]	F-Pts	Speed	×	10 min.	2 days	-	Loops	Calendar	LSSVM+FfO	LSSVM	×
[51]	F-Sgm	Flow	✓	5 min.	15 days	5 min.	FCD	-	STW-KNN	KNN, NN, NB, RF, C4.5	×
[106]	U-Pt	Flow	×	10 min.	4 days	10 min.	Loops	-	LSSVM+FfO	LSSVM, RBFNN, LSSVM+PSO	×
[107]	U-Ntw	Speed	×	5 min.	15 months	5 min.	Loops	-	Custom TSA	⟲	×
[54]	U-Ntw	TT	✓	30 min.	166 days	5 min.	ANPR	-	STNN	HA, ARIMA, STARIMA	×
[108]	F-Sgm	TT	✓	15 min.	1 year	5 min.	FCD	Calendar, Lat-lon	GradientBoost	ARIMA, RF	×
[109]	U-Pt	Flow	×	1 day	4 days	15 min.	Loops	-	WaveletNN+GA	WaveletNN	×
[65]	F-Ntw	Flow	×	15 min.	5 years	15 min.	Loops, ANPR, Camera, FCD	Speed, TT, Calendar, Lat-lon	FIMT-DD	⟲	Drift Detect
[110]	U-Pts	Flow	×	2 hours	1 year	15 min.	Loops	-	MLP + HS	⟲	×
[111]	U-Pts	Flow	×	5 min.	4 years	5min.	Loops	-	Extreme Learning	⟲	Forgetting mechanism
[112]	F-Pt	TT	×	1 min.	15 hours	-	Loops	-	Cluster+ARIMA	⟲	×
[113]	U-Ntw	TT	✓	30 Seconds	1 year	Online	AVL	Calendar	RF	⟲	✓
[61]	U-Ntw	TT	✓	30 min.	15 days / 2 months	5 min.	FCD	-	Graph Based Lag STARIMA	Basic, HA, kNN, RF, SVR	×
[114]	U-Ntw	Flow	✓	50 min.	1 month	10 min.	Loops	-	SVR	ARIMA, MARS, Spatio-Temporal Bayesian MARS, AR	×

- **Context**: U→Urban; F→Freeway or highway; Ntw→network; Pt→single point; Pts→diverse points; Sgm→points in a road segment.
- **ST**→Spatial-temporal prediction
- **Horizon, Extension, Step**: -→No data available; NH→no horizon
- **Source**: RTMS→Remote traffic microwave sensor.; Lat-lon→Latitude and longitude; AVL→Automatic Vehicle Location
- **Models**: ⟲→Model tested against itself; AKNN-AVL→K-Nearest Neighbor combined with balanced binary tree; CCGA→Cloud Chaos Genetic Algorithm; CRF→Conditional Random Field; DTT→Dynamic Topology-Aware Temporal traffic model; EFGM→Grey Verhulst model with Fourier Error corrections; FIMT-DD→Fast Incremental Model Trees with Drift Detection;FNR→Functional Nonparametric Regression; FfO→Firefly Optimization; GAM→Generalized Additive Model; GPDM→Gaussian Process Dynamical Models; GSVR→Gaussian Support Vector Regression; ieRSPOP→incremental Rough Set-based Pseudo Outer-product with ensemble learning; LCG-BN→Linear conditional Gaussian Bayesian network; LLDR→Link-to-link dividing ratio; LSSVM→Least Squares Support Vector Machine; LSTM-NN→Long Short-Term Memory Neural Network; MARS→Multivariate Adaptive Regression Splines; MCS→Mobile crowd sensing; MKNN→Multivariate k-Nearest Neighbors; MVLR→Multi-variable linear regression; PHFRBS→Parallel Hierarchical Fuzzy Rule-Based System;PPR→Projection Pursuit Regression; RL→Reinforcement learning; STRE→Spatio-temporal random effects; STW-KNN→Spatio-Temporal Weighted k-Nearest Neighbors;

A long as it is possible to build models of these inputs, it is also possible to encompass them in traffic forecasting models. Information about sport events, parades or in general any human-organized happenings is scarcely used [93], [117]. These events presumably have similar effects in traffic, if they are repeated over time, and combined with calendar information, they might provide a valuable input to prediction models. Road works are predictable, but they are not usually repeated over time in the same place, and each location is affected in a different way. However, it is possible to model the impact of roadworks depending on their location, affected lanes, and other factors, and include this in traffic conditions prediction model [118]–[120].

2.2.4 Ageing and Concept Drift in Models

A shift to data-driven approaches is observable in most recent literature about traffic variables forecasting [13]. These models use large databases with plentiful records from which they learn and produce predictions. As data extension increases, it is more probable that knowledge extracted from those data is less factual; road networks are constantly changing, specially in large urban areas. The usage of those roads is also variable within long periods of time, increasing in thriving metropolis and decreasing in economic crises affected areas. If a prediction model learns from a 5 year database, like [65], a change in flow direction, a closed or new lane, or a change in usage patterns during those years are possible, and depending on the urban area modeled, they can be highly probable.

Responsive models that adapt to unexpected short-term factors, like accidents, congestion situations or weather conditions, were proposed in [13]. An adaptation to long-term factors is feasible through data ageing. Weighing down old measurements can be a naïve approach to provide a forgetting mechanism that gives a greater relevance to current input values of the model [111]. Concept drift techniques [121] allow for an adaptive learning strategy that has recently started being applied to traffic prediction. The application of these techniques spans from adapting a model to detecting anomalies and recalling the inferred traffic patterns prior to the occurrence of the anomaly, which can be also exploited under incident or atypical road congestion [122].

Prediction models have experienced a considerable changeover. ARIMA models are in clear decline, and only three of the analyzed studies elaborate predictions with this kind of methods. They still are, though, usual models to compare to.

The shift to ML techniques and Artificial Intelligence (AI) has roots to the nature of traffic datasets. Traffic data are usually messy, extremely irregular. The non-stationary and nonlinear nature of traffic has been frequently observed in forecasting literature [123]–[126]. ML and AI approaches may overcome the parametric nature of statistical models and the relevant modeling constraints. They are capable of mining information from messy and multi-dimensional traffic datasets. To this end, neural networks still maintain a hefty presence, with several variations, such as

[75], which uses an ensemble of neural networks that cooperate and aggregate predictions obtaining better results. Similarly, [64] and [96] cluster data prior to perform a neural network regression. Likewise, [90] and [97] use fuzzy rules with clustering purposes. In [111] an On-line Sequential Extreme Learning Machine (OSELM), a type of neural network, is used with a forgetting mechanism that allows the author perform adaptive learning. Particle swarm optimization is used in [88] to optimize the model parameters and in [109] a genetic algorithm is used to optimize a wavelet neural network. Aside from neural networks, fruit fly optimization [106] and firefly optimization [105] solvers are also used to optimize hyper parameters of support vector machine (SVM) algorithms. KNN models are also widely used [49]–[51], [80], [91], specially related to spatio-temporal forecasts.

A few works include any sort of concept drift techniques currently, but the proliferation of Big Data technologies and long-term prediction methods should lead to a more generalized usage of data ageing mechanisms. Adaptive learning techniques are applied in some of the works. In [99], an incremental learning algorithm (coined as *ieRSPOP*) is tested on three datasets, one of them with only 9 days of traffic flow. Incremental learning implies any instance of data can only be used once for training [127], so older instances become less relevant when the algorithm evolves. A more traffic-specific study with incremental learning is made in [100], obtaining smaller MAPE values than the same method without incremental learning. Incident detection is achieved by [102] using *Drift3Flow*, an online incremental learning method that detects changes in traffic flow and occupancy and infers when they are caused by incidents. Same authors develop a prediction model to mitigate the effect of bus bunching by using information gathered by the vehicles, and requires the model to be constantly updated with online data [113]. Drift detection is also used in [65] to improve the prediction efficiency with 5 years of data.

2.2.5 Big Data and Architecture Implementation

Cloud and parallel computing big data paradigms can provide the means for macrosimulation, computationally inexpensive entire network predictions, traffic deep learning and more explanatory power of models [128]. The importance of developing Big Data approaches well suited for dealing with traffic forecasting will imminently become paramount and gain momentum in the research community.

To this end, the efficient representation for geospatial big data will play a decisive role. Most traffic information – mainly due to the crowdsourcing initiatives – have a clear location dimension (e.g. GPS data), which may hinder significant information on the recurrent and non-recurrent traffic patterns. The above along with the need to visualize and quantitatively process geospatial big data for decision support (e.g. real-time traffic management) with methods that conceptually defy the classical statistical modeling constrains of optimal design or model based sampling opens new opportunities to modeling and forecasting.

It should be noted that leveraging big data to improve short-term traffic predictions is strongly related to the availability of relevant data that are altogether linked. Traffic and mobility data openness is of extreme importance not only to the accuracy and adaptability of traffic forecasts, but also to the ability to provide users and planners with timely traffic information. Nevertheless, open data initiatives come with certain challenges, such as the economic cost of openness, issues of data reliability, access and control, security and legislative framework and so on. In any case, the concepts of opened and linked data are crucial and will have significant socio-economic perspectives. In the future, those that possess big data and can analyze them will most likely have a significant competitive advantage.

2.2.6 Computer Traffic versus Road Traffic

Computer networks share some relevant features with road networks. Researchers have established analogies specially at microscopic levels, being data packages the counterpart to individual vehicles [129], [130]. At macroscopic – defining the origin and destination of a route – and mesoscopic – controlling the route and making decisions – levels differences are larger, as computer traffic is entirely controlled by the infrastructure [130]. Nevertheless, traffic in both kind of networks needs to be managed: optimal routes, congestion situations, demand administration, priority control and alterations in the level of service, among others. In computer networks packages are identified and pinpointed at all times. Nodes, links, their status, features and in general the complete map of the network are available, while in road networks drivers are ultimately in control of routing. Despite these essential differences the diversity of ITS technologies developed during the last decade (from adaptive cruise control to incipient autonomous driving and cooperative intelligent vehicles [131]) are converging at a microscopic level to computer networks, in an steady evolution towards fully managed road networks.

In the field of forecasting, computer network management usually requires predictions of future demand at different parts of the network, which are frequently based on the same principles than those used for road traffic predictions. In computer networks predictive information is used to rearrange their topology or route packages through the most cost-effective links (with *cost* often defined in familiar terms, e.g. end-to-end delay), while in road network traffic forecasts are used to inform drivers – who have the last decision – and to inform road managers for their traffic management duties. A more effectively and controlled infrastructure would make forecasts more relevant, hence the way traffic predictions are obtained and used in computer networks should motivate new forecasting models and applications in road networks. For instance, ARIMA and Artificial Neural Networks (ANN) are also mainly used methods in network traffic prediction [132]–[134]. Prediction methods and techniques have evolved in similar ways, and it can be possible to take some the more advanced models in network traffic prediction and apply them to road traffic prediction,

subsequently triggering data-based traffic management strategies such as traffic rerouting [135].

2.3 Conclusions

Traffic variables have been object of analysis and predictions for more than 40 years. Several surveys have examined the subject with different contrasting criteria and field challenge assessments. The most relevant development in the field of traffic prediction in recent years is related to a shift in the prediction modeling paradigm. Advances in data oriented techniques and technologies, explosion of Big Data and machine learning, along with growing availability of traffic related data from plentiful sources have contributed to leave behind time-series analysis methods and given a boost to data driven models. This shift has compelled most recent researchers to lay out new horizons and challenges for the field. This chapter has intended to compile and recapitulate previous work, to propose a comparing framework and to review most recent literature with respect to the updated criteria.

Traffic variables prediction literature has been studied thoroughly in previous surveys, so this chapter has focused in recent works. Updated reference criteria have been used to study new literature. In our inspection, the aforementioned shift to data driven models is clear: the use of ARIMA and other time-series analysis methods is lessening, and first works with prediction horizons longer than 60 minutes appear, laying the foundations for future work in this line. Models with concept drift or adaptive learning are also found in a small share of the reviewed works, but the progressive incorporation of models based in large databases which initial knowledge changes in time and requires adaptation, might precipitate this technique to be dominant. Exogenous factors are yet scantly introduced in traffic forecasting models. Their convenience has been specially proven when using calendar and time of day information as model parameters, but many other data are more available every day, which could boost traffic prediction performance of future models.

Chapter 3

Preprocessing Data for Road Traffic Forecasting

The history of road traffic forecasting has hitherto involved time-series analysis and prediction models with a wide diversity of algorithmic variants and processing enhancements. As evinced in the previous chapter, ML techniques have acquired momentum by virtue of the large amount of successful methodologies, algorithms and optimization procedures [12], [13], [136], [137], further propelled by the advent of Big Data technologies [37], [138]. In this context, the most relevant traffic variables (*i.e.* flow, speed, travel time, occupancy) have been predicted using data captured by magnetic loops, cameras, plate readers and floating car data, among many other sources. Within them, inductive loops or ATRs are one of the most frequently selected data sources for traffic forecasting [13]. ATRs count each vehicle passing through a particular point in the network, but they often undergo situations in which the output data are faulty, to the extreme of existing long periods of time with no captured data due to prolonged reading, recording or transmission errors. In some cases, organizations that manage the sensors and provide data remove measurements that are considered to be samples with invalid values, like miscounts, sensor calibration errors or round-off errors [139]. In other cases, the same managers aggregate or process data before publishing, a mechanism that sometimes entails errors [140]. These eventualities result in data streams with missing portions of data of diverse sizes, having a negative effect on the forecasting models [139], [141]–[143].

Evidently, missing data unchain problems not only in traffic forecasting, but in any prediction, regression or data analysis based on data obtained from diverse sources [144]. Thus, researchers from many fields have devoted significant efforts towards new imputation methods for missing data. As such, one of the most straightforward approaches is to fill in the gaps with artificially created data [145]–[150]. Although these fields are related to atmospheric, meteorological or geophysical variables, they relate to time series and some of their typical issues are common to traffic time series. For instance, a thorough review of imputation techniques for CO_2 flux time series is contributed in [145], most of which are applicable to a traffic context. Strategies for imputing missing data can be of paramount relevance also in traffic datasets. Indeed, the quality of data, defined as the

fullness of data, has been lately identified as one of the major challenges of road traffic forecasting, including data-driven methods [13].

3.1 Related Work

In the traffic forecasting domain, elaborated missing data imputing methods were first reported in the early 2000s, when a few approaches were introduced in [141] and later categorized by [151] in two main groups: 1) statistical, considering Expectation Maximization [152] and Data Augmentation algorithms; and 2) heuristic methods, comprising various averaging techniques over historic data. A more recent classification by [143] divides imputation strategies into those based on prediction, interpolation and statistical learning. The inclusion of a prediction category brings many more methods based on considering missing data as values to be predicted. Among the representative literature related to this category it is worth to highlight the seminal work in [141], where a simple historical mean imputation was shown to outperform *no-substitution* and *substitution-by-zero* methods when used in combination with an ARIMA and ANN as prediction models. Remarkably for the scope of this Thesis, this early study considered missing data densities of up to 30%, generated uniformly at random. Authors also showed that ARIMA models are more sensitive to missing values than their ANN counterparts.

In general, a model that relies on the time dimension of a dataset is prone to be sensitive to missing data, as these models typically require an uninterrupted time series as their input. On the other hand, when a dataset has a substantial extension with very few corrupted/missing data entries, a simple strategy of removing instances affected by gaps or imputing a constant value to them may suffice for the forecasting method to model the traffic conditions [13]. Van Lint et al. [139] consider null imputation, linear interpolation and ARIMA as filling methods prior to a State Space Neural Network predictive model, dealing with up to 40% of randomly located missing data occurring successively in intervals of length up to 30 samples. In their scenario, simple, non-parametric imputation methods were shown to handle missing data efficiently. Henrickson et al. in [153] introduce a statistical approach that performs successfully even with 1-month-long missing data. Their so-called predictive mean matching method draws random values to impute from a distribution obtained from the present values, considering one measuring station. Probabilistic Principal Component Analysis (PPCA) method was also proposed in [154], addressing some commonly made assumptions about missing data. Methods relying on component analysis have been widely used ever since [143], [155]–[159] and, to the date of this Thesis, they embody one of the most popular processing approaches for imputing missing data. In a comparison among 6 methods performed by [160] authors conclude that PPCA is the most efficient imputing technique within their sample not only in terms of performance, but also in ease of implementation and speed. Other numerical approaches include 1) Bie et al.[161], where an online imputation method is proposed consisting of a multiple linear regression based

on data from loops that are part of the same measuring station; and 2) the similarity-based imputation technique proposed by Zhong et al. [162], where daily curves with gaps are compared to candidate curves without gaps, using the closest one – under a measure of similarity – to impute. The missing intervals reached 12 hour length, but they only considered one type of day pertaining to a particular season of the year. Tensor based methods have been exploited recently to deal with missing data introducing spatial context relations [157], [158], [163]. These methods model the interactions between multiple traffic variables into multi-dimensional arrays (tensors), thus allowing for the combination of multiple correlations between the different variables to impute missing data.

ML methods are also becoming prominent in recent years, most of them falling in the aforementioned *prediction* category. Kernel regression in combination with KNN was used in [164] to obtain forecasts of missing values using information from neighboring stations. The study only covered input data generated on Tuesdays, but they performed an analysis of the missing data characteristics present in the dataset in order to generate gaps that realistically mimic the real ones. Imputation of missing data was also tackled as predictions in [140], [165], which proposed to build ANNs optimized via genetic algorithms to obtain missing data estimations of up to 1 hour. Clustering approaches have been recently explored in [166] and [167]. The former introduces the widely neglected distinction between days of the week, representing the input data as values taken on a time step of a certain day of the week. This helps the model to distinguish patterns in different days. A Fuzzy C-means algorithm is then used to group known days, and a genetic algorithm to estimate missing data by minimizing errors between imputation and actual values of clusters. Likewise, [167] considers a large group of sensors of a network and uses a K-means algorithm to cluster them based on their average daily traffic; then they use a deep learning method – specifically, a Stacked Denoising Autoencoder (SDAE) – to model relationships between sensors of each cluster. Once built, the model is able to impute missing values to all the sensors simultaneously. The performance of the model is tested over 6 days of data with 10% to 90% missing values. In a similar direction, [168] presented a SDAE that considers weekdays and non-weekdays, different selections of sensors, and up to 50% of missing data.

Along with all the above imputing methods, some authors derive robust models to cope with missing data and obtain forecasts without considering any imputation mechanism [169]. Sun et al. [142] introduced a sampling Markov chain method to carry out short-term traffic forecasting with incomplete data with no previous imputation phase. Later, this work was extended in [170] by using a Bayesian inference mechanism to obtain robust predictions with incomplete data, and complemented in [171] by a selective random subspace predictor that leans on the information supplied by surrounding sensors that are correlated to the one under study. By exploiting this augmented and redundant information subsets with missing data can be dismissed.

3.1.1 Contribution

Despite these approaches, incomplete data can become a problem – even for data-oriented robust models – when the amount of missing values is high and spans long periods for which no useful information can be considered to obtain a model [143]. Surprisingly, despite this widely acknowledged circumstance, the literature so far is scarce in what regards to empirical evidences of the comparative performance of imputation strategies under different yet realistically modeled distributions for missing data. Moreover, the implications of imputed data in the performance of predictive models for traffic forecasting have not been deeply studied and analyzed. This manuscript aims at presenting and discussing strategies to deal with missing data, as well as to obtain new insights on the main relevant aspects of data imputation. Specifically, the main goals of this chapter can be summarized as follows:

- A review of the techniques for generating synthetic missing points and intervals (missing data), numerically exploring their implications on the quality of imputed data.
- An analysis of the impact of the distribution of missing data and the imputing methods on the performance of forecasting methods.
- Two novel imputing strategies to tackle long periods of missing data from two different perspectives: 1) a clustering-classification algorithm which incorporates external data that that are always available, such as days of the week, months or holiday information, and 2) an ELM [172] model optimized with a genetic algorithm, that builds upon information obtained only from surrounding sensors.
- The use of 2-year worth of data obtained from the sensor network of Madrid.

The rest of the chapter is organized as follows: Section 3.2 describes the input data, the different artificial missing data generation techniques, the proposed imputing methods, and the results evaluation and comparison methodology. Section 3.3 presents and analyses the performance of the proposed methods in different missing data scenarios. Finally, Section 3.4 draws concluding remarks inferred from the obtained results and prescribes future research lines related to this work.

3.2 Materials and Methods

In order to extract informed conclusions from empirical findings this research work uses traffic data obtained from a public source. Over them, artificial missing data are created and our proposed imputation methods are applied. The following subsections describe the source and selection criteria for the input data, the missing data generation methods, our imputation models and the performance evaluation procedures.

3.2.1 Input Data Selection

Input data for this research have been collected from a public source maintained by the City Council of Madrid (Spain). The details and characteristics of the data can be found in Appendix A. Using one-minute resolution data would require a collecting process that would take as long as the time span of the desired dataset. To overcome this issue, the focus has been set on historic 15-minute data of complete years, which provides enough information to consider seasonality in a data driven approach: one year can be used as training data for the developed models, and any other as test data. This seasonality can be of great relevance depending on the traffic profile of a certain location: in a business area, a model trained with data collected in March would intuitively perform poorly when predicting values for the month of August. On the contrary, in a residential area with less fluctuating traffic profiles, winter data might be useful to obtain summer forecasts. A model trained with whole-year data can, on the other hand, learn seasonal patterns and apply them for the prediction.

By the time this research line was started aggregated published data were just available for 2014 and 2015, and three months of 2016. Therefore, input data are taken from a subset of sensors for 2014 and 2015. The choice of the sensors for further analysis was made under the following criteria: a location close to the city center, avoiding flat traffic profiles of residential areas (for which imputing missing data would be more straightforward, potentially misleading our conclusions); and the availability of data, required to assess the imputing performance after artificially generating missing data points and intervals. Figure 3.1 shows all ATRs located within a 2 kilometer radius of Puerta de Alcalá, one of the main business areas of the city, which represents a first filter for our imputing model. The color code portrays the available percentage of the total 35040 annual readings for each magnetic loop during 2014, which will be subsequently used as training data for our models. A considerable amount of sensors have less than 50% of data available in this year, and from 186 loops accessible in this area, only 21 have served data for more than 98% of the period. This noted fact emphasizes the actual need for robust imputation methods in this particular context of application.

The introduced spatial context imputing strategy is built upon past information of the studied ATR and past and current data coming from neighboring sensors. In an application context, our spatio-temporal strategy would rely on the neighboring sensors with the most complete information available. Hence, we have taken into account only those locations with more than 34500 observations available (more than 98%) for 2014 and consider them as training data, yielding the set of 21 loops depicted in Figure 3.1 as the first of the categories. On the other hand, the testing of our spatial context model requires 2015 complete data from surrounding loops. Thus, data from aforementioned locations is examined for 2015, seeking the longest series of consecutive correct readings common to all locations. A shared subsequence of 8463 consecutive observations (ca. three months of data) has been found for 13 of the sensors. One of these 13 sensors has been randomly selected as the target ATR, while the rest $N = 12$ are used

FIGURE 3.1: Automatic traffic recorders (ATRs) in the center of Madrid, colored by their data availability during 2014 (in % of valid 15-minute intervals over the year).

as context sensors. In the test data from that sensor, artificially generated missing points and intervals will be introduced (modeled as later explained in the following Subsection), and imputation will be performed on those synthetic missing data. The rest will act as surrounding loops. These sensors are shown in Figure 3.1 highlighted with a star marker, while the loop under study is annotated as the *target*.

We denote the observation obtained from the i-th ATR at time index t as o_t^i where the time index t spans over years 2014 and 2015, and i takes integer values from the range $[0, N]$. The selected target ATR for which the imputation process will be performed corresponds to $i = 0$, and for notational convenience will be labeled henceforth as s. When referring only to the context ATRs index $j \in [1, N]$ is used. The subset of observations used for training the models, *i.e.* with time indexes corresponding to 2014 historic data, are denoted as $\mathcal{H}^i = \{o_t^i : t \in 2014\}$, while the subset of observations with 2015 time indexes, used for test, are denoted as $\mathcal{G}^i = \{o_t^i : t \in 2015\}$. As with the individual observations, these sets are instanced subsequently as $\mathcal{H}^s$ and $\mathcal{G}^s$ to specify observations taken from the target loop s, and $\mathcal{H}^j$ and $\mathcal{G}^j$ to refer to those of the context sensors.

3.2.2 Generative Models for Missing Data

Before delving into the models used for generating missing entries in the considered test dataset $\mathcal{G}^s$, it is insightful to note that some authors deal with incomplete datasets from a prediction perspective: instead of presenting a strategy to fill in the gaps, they rather propose models to obtain forecasts overcoming gaps [143], [164], [173], [174]. Consequently, their score to measure the effectiveness of their methods hinges on the performance of the chosen prediction models regardless of which data were declared as missing. On the contrary, this work focuses on comparing among imputing models by using a defined set of synthetically generated missing data, as well as by determining to which extent an improvement of the imputed value yields an enhanced accuracy of subsequent traffic forecasting models.

This being said, three broad families of generative models for missing data can be found in the literature [175]: the so-called *Missing Completely At Random* (MCAR), *Missing At Random* (MAR) and *Not Missing At Random* (NMAR). The first two imply that there is no mechanism underneath for generating the missing data, whereas the latter assumes a dependence of the distribution of missing data on the complete dataset. This general classification has been used as a reference by most researchers in the traffic context [153], [154], [166]. Van Lint et al. [139] describes three types of data failures: random, structural and intrinsic, where the first represents stochastic reading or transmitting errors, the second consists of gaps resulting of a sensor being offline, and the latter refers to noise, bias or errors caused by processing the data. A similar classification is proposed in [164], with two random (one with all independent and one with related missing points) and one structural model for missing data generation. Chiou et al. [156] alludes to the unlikelihood of distinguishing the source or kind of the missing data, reducing them to two practical categories: *point-wise* and *interval-wise* missing data, representing MCAR and MAR respectively, and considering that intervals are groups that occur randomly. They also contemplate a mixture of both types in their datasets. Although with different names, these two approaches for creating missing data are common to most related contributions: some authors consider only point-wise random generation expecting that high percentages of missing data will create long intervals, whereas the rest tend to consider both methods, either combined or in isolation.

In this line of work, artificially generated gaps ranging from 25% to 65% of the dataset are considered in [176], using the rest of data as an input for their imputing methods. In [141] gaps are generated for 10%, 20% and 30%, and in [139] the percentage increases up to 40%. Moffat et al. [145] produced up to 50 gap scenarios, defining 4 sizes of gaps and making 10 combinations of each size, plus 10 scenarios with mixed sizes, although in all cases the total amount of missing data amounted up to 10% of the entire dataset. In [162] 12 successive hour gaps were introduced in different days, which allowed studying their effect and the effectiveness of their considered imputation techniques depending on the type and hour of the day. In [167] a clustering approach was applied to randomly generated gaps for up to 90% of the dataset, obtaining satisfactory results even with large portions

of missing data. Interval-wise generation ranges from 24 consecutive points as in [156] to one month as in [153]. These extended range gaps can be regarded as a representative application of NMAR generative models for missing data, where a failure in the sensor or the communication hampers the proper collection of data for a long period.

The experiments in this work use a dataset $\tilde{\mathcal{G}}^s$, which is the result of artificially removing data from $\mathcal{G}^s$ by both of the approaches detailed below. The elements of $\tilde{\mathcal{G}}^s$ are denoted as $\tilde{o}_t^s$, namely $\tilde{\mathcal{G}}^s = \{\tilde{o}_t^s\}$. Each of these values, $\tilde{o}_t^s$, is either *well defined*, *i.e.* equal to the observation o_t^s, or is an artificially generated blank. For a given value of t a function is defined such that

$$\delta(t) = \begin{cases} 1, & \text{if } \tilde{o}_t^s \text{ is well defined,} \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

In the following subsections a detailed explanation of the artificially generated missing data methods is provided.

3.2.2.1 Point-wise Generation

We have defined percentages ($\xi \in \{1, 10, 25, 50, 80, 100\}$ [%]) of missing observations in the whole test $\mathcal{G}^s$ sequence of data entries. The missing data points or *blanks* are placed individually uniformly at random. When the percentage is low, missing points are separated from each other naturally, *i.e.* consecutive blanks are rarely obtained for low values of ξ. When ξ is increased, groups of gaps emerge and it results easier to find sequences of missing data. Due to the uniform distribution of holes, even for $\xi = 50\%$ and $\xi = 80\%$, there are no completely *empty* days (*i.e.* days with all-blank entries). The case when $\xi = 100\%$ is uncommon in previous works; its purpose is to test the effectiveness of methods introduced in this work under these circumstances (3 complete months of missing data).

3.2.2.2 Interval-wise Generation

Traffic flow observations posted in the open data portal for the urban network of Madrid have been preprocessed beforehand and, as in many other cases [140], missing points could have already been imputed by the entity managing this repository. This means that a 15-minute reading integrates multiple shorter-term observations and also that, in this particular case, missing data could be mainly due to errors in the aggregation or processing stages. In order to generate intervals of missing data that actually reflect the behavior of gaps in our dataset, we have assessed the real distribution of missing data for all the year 2015, which contains the test set of our experiments, for each of the considered measuring points (Table 3.1).

Despite their sparsity, the similarities found in all the considered locations, such as the almost identical percentage of missing data or the number of gaps, suggest that errors are probably produced in the aggregation stage and affect similarly to groups of sensors. The most frequent gap length is 96 positions, which corresponds exactly to one day worth of data;

TABLE 3.1: Analysis of *actual* missing data distribution. Column names stand for loop ID (as per the naming convention of the repository), yearly average flow of cars measured in vehicles/hour, total number of missing points, percentage of total missing data (considering 35040 samples), number of intervals grouping missing points, average gap length measured in samples $\langle L \rangle$, statistical mode of the length of the gaps L.

ID	Avg. flow	# Missing entries	Missing data	# intervals	$\langle L \rangle$	mode(L)
10006	420	4626	13.2%	27	135.70	96
10018	134	4576	13.1%	26	139.12	21
10023	143	4550	13.0%	25	143.83	96
10030	82	4713	13.5%	28	133.89	96
13026	131	4559	13.0%	26	138.44	96
13032	474	4575	13.1%	26	139.08	21
18018	332	4622	13.2%	27	135.54	96
19011	683	4791	13.7%	28	136.78	21
21007	204	4556	13.0%	26	138.32	96
90033	317	4640	13.2%	28	131.19	96
90034	236	4640	13.2%	28	131.19	96
90035	198	4639	13.2%	28	131.15	96
90041	233	4636	13.2%	28	131.04	96

a further inspection of the data at hand reveals that these gaps usually match natural days, starting at 0:00 AM and ending at 11:45 PM. Also the distribution of gap lengths has been examined for the target loop: besides the 96 length gap, the most frequent lengths are 48 and 192 positions (half a day and two days of data records, respectively).

According to these characteristics of the input data, it is expected that any missing data generation strategy not producing entirely empty days (e.g. any of the random point-wise generation percentages) will not properly represent the statistical distribution of real gaps in this particular scenario. Consequently we have defined 6 sets of target data, each of them with gaps of 24, 48, 72, 96, 144 and 196 consecutive positions respectively. For each set, gaps are placed randomly and amount up to 13% of the total test data.

3.2.3 Imputing Data Methods

In this chapter two new approaches for imputing missing data are introduced. One that depends on information gathered from other sensors and other approach depending on external factors that define clusters of days. The following subsections describe the details of these two methods.

3.2.3.1 Spatial Context Sensing

Traffic state data gathered by a sensor network supply spatio-temporal information, as vehicles often navigate through several detectors along their trajectories. Intuitively, the traffic profile at a road segment should be very similar to that in an upstream segment a τ before, whenever τ equals the average travel time between both segments. Nonetheless, there are two main factors that put into question this intuitive statement, e.g. the lack of continuity due to road bifurcations or parking areas, and the speed dispersion. Features of gathered data, such as the distance between points

of collection, or the location of sensors in an urban context can make the effect of previously mentioned factors more noticeable. Moreover, the available temporal resolution, renders it impractical to establish a direct relationship between the measurements taken by two neighboring sensors, even when they are in two adjacent segments. To illustrate this, we hypothesize two sensors placed in an urban street with synchronized readings at intervals of $\Delta T = 15$ minutes. In this scenario, any direct correlation of their traffic profile would be most probably spurious, as in 15 minutes great variations may occur in an urban context. Despite this noted relational uncertainty, plenty of contributions dealing with traffic prediction and missing data imputation [139], [143], [164], [167], [168], [170], [177] have relied on spatio-temporal relationships, even in urban contexts and with coarse-grained data, on account of different techniques that allow researchers to find interrelation models among nearby located sensors [178].

This being said, it is noticeable in Figure 3.1 that distance between the location under study and the others is not necessarily short, *i.e.* they are not so *closely neighboring*. Our first proposed imputing method leans on the relationships between measurements of different, not necessarily nearby, sensors at the center area of a city. Missing data entries are imputed by means of a forecasting model that predicts values for a sensor by learning from the information provided by other sensors. As exposed in [171], correlations between the traffic among two separate links produce better forecasting results disregarding the distance [167]. Conceptually, the model retrieves data from locations (where available), being defined initially by a great deal of observations collected from each loop. This does not necessarily produce a good model, as some of the loops can be placed in locations with very different traffic profiles, and would constitute noise for the imputation procedure. For this reason, an optimization step is added to the predictive model to adjust the amount of information that each sensor contributes to the training dataset. Figure 3.2 displays the overall operation of this model.

For a specific reading o_t^s of the selected loop s at time t, a number w_j of observations prior to t are taken from each surrounding loop $j \in [1, N]$ towards defining a vector of features that ultimately constitutes the dataset with the observation o_t^s as target variable. The window size w_j of each loop can be different, suggesting a level of influence of the surrounding of loop j in the prediction of o_t^s. The forecast horizon h is defined as the number of time steps in the future for which the prediction is made, *i.e.* the difference between t and the most recent time of the samples collected for the surrounding sensors. For instance, by setting $h = 1$ forecasts of values taking place ΔT in the future.

It must be noted that this method requires a certain degree of completeness within the historical training data $\{\mathcal{H}^i\}$ from which the model is constructed (indeed the selection of surrounding loops has been made accordingly), but its training phase is robust to sporadic missing data in the historic dataset. Real missing data in the train time series are flagged and after the train dataset is built, instances containing flags are removed, still resulting in a relatively large dataset for the scenario in hands (more

than 32000 training samples available out of the initial 34500 entries in the retrieved repository). This is an important feature, as most imputing methods require complete historic data [157], becoming a practical issue in the majority of real life scenarios.

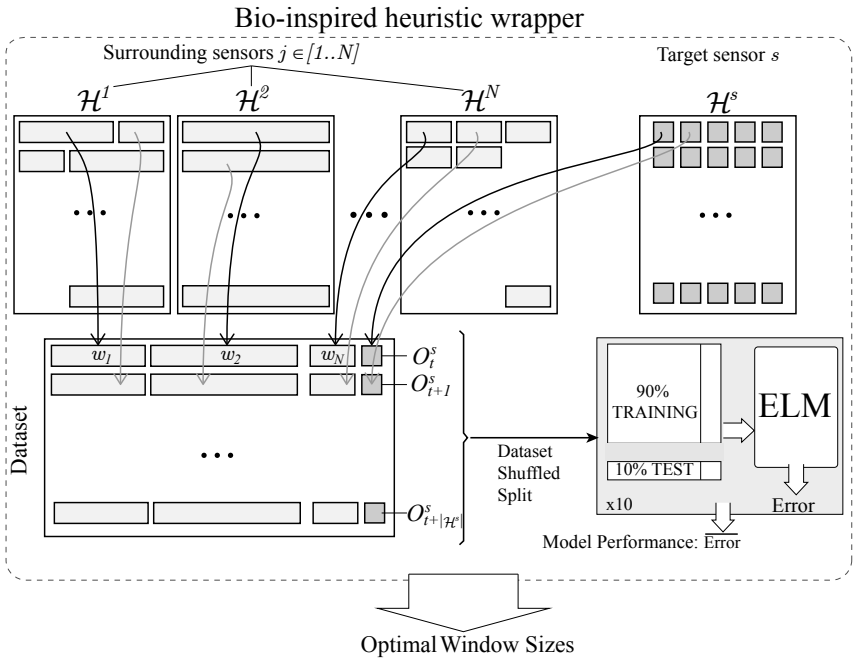

FIGURE 3.2: Training of a forecasting model through window-size optimization and ELM regression models. Optimized parameters of this model are used after to obtain predictions that act as imputed values.

The combination of diverse kinds of ANNs and otherML methods with heuristic optimization algorithms has been extensively explored in previous works [13], often yielding more responsive results to changes in data than time-series forecasting. In this scenario, our imputing method is built upon the predictions obtained by an ELM model, which is trained with a dataset built by following the scheme depicted in Figure 3.2. The general operation of the model is similar to that in previous works [110], [179], introducing in this case the inputs from surrounding sensors. A bio-inspired heuristic solver is introduced to find the optimal window sizes w_j of each sensor j. This procedure can reduce considerably the processing time, and provide insights on the importance of some of the sensors to predict and impute the missing values of the target location, if any of the optimized window sizes equal 0.

Initially, a maximum value of 50 steps (12.5 hours) is defined for the window size w_j of all 12 surrounding loops around the target loop, rendering a dataset of 600 features and around 32000 instances. The population of the bio-inspired heuristic solver is composed of the window sizes of each surrounding loop, and in each generation of the optimization algorithm,

the prediction model is built, trained and validated, obtaining an RMSE performance metric. When the optimization process ends, the window sizes found in the best generation are the ones that potentially yield the best RMSE score. The ELM model is then trained considering these windows on the $\{\mathcal{H}^i\}$ dataset, and tested on the $\{\mathcal{G}^i\}$ dataset as a single hold-out. This produces forecasts for all values of the $\mathcal{G}^s$ series, as if the missing data were the 100%, making this method an interesting option under such circumstances. Values obtained for positions where a gap was generated are then compared to the actual observation, and assessed via the metrics discussed below.

3.2.3.2 Pattern Clustering and Classification

The second method proposed in this chapter involves only data from the target loop. As in the previous technique, a set of samples prior to the period at hand is required for training. This method is designed to produce data in a complete day fashion, as opposed to point-wise filling counterparts. Several schemes in the literature [140], [156], [158], [160], [167] involve splitting the series of data in lots of data per day. The pattern clustering and classification imputing method, as well as two of the proposed comparison methods, perform this splitting of incoming data into day-wise vectors:

$$\mathbf{o}^{s,d} = \left[o^s_{t_d}, o^s_{t_d+1}, \ldots, o^s_{t_d+P-1}\right], \tag{3.2}$$

where $o^s_{t_d}$ is the value of the observation captured at sensor s and time t_d, being t_d the first time index of day d; and $P = 96$ is the number of observations obtained within a day for a capture period of $\Delta T = 15$ min. Following the dataset division criterion explained in Subsection 3.2.1, we have defined a training dataset with $\mathcal{H}^s$ data, $\boldsymbol{\mathcal{H}} = \left\{\mathbf{o}^{s,d} : t_d \in 2014\right\}$ and a dataset with $\tilde{\mathcal{G}}^s$ data $\tilde{\boldsymbol{\mathcal{G}}} = \left\{\tilde{\mathbf{o}}^{s,d} : t_d \in 2015\right\}$.

In general, as $\tilde{\mathcal{G}}^s$ includes artificially generated blanks, each vector $\tilde{\mathbf{o}}^{s,d}$ has $P^{s,d} \leq P$ valid (non blank) values given by

$$P^{s,d} \doteq \sum_{p=0}^{P-1} \delta(t_d + p). \tag{3.3}$$

Based on this definition we establish a metric of similarity between any vector from $\boldsymbol{\mathcal{H}}$ and any vector from $\tilde{\boldsymbol{\mathcal{G}}}$:

$$S(d, d') \doteq \frac{1}{P^{s,d'}} \sqrt{\sum_{p=0}^{P-1} \delta(t_{d'} + p) \left(o^s_{t_d+p} - \tilde{o}^s_{t_{d'}+p}\right)^2}, \tag{3.4}$$

where $d \in \boldsymbol{\mathcal{H}}$ and $d' \in \tilde{\boldsymbol{\mathcal{G}}}$.

Clustering

Once the input data are separated in days, a clustering algorithm is performed over $\boldsymbol{\mathcal{H}}$, obtaining groups of days with similar set of measurements.

Performing a clustering process over a space with such a large number of dimensions requires large computational resources. Furthermore, the overall process could be biased by localized, high-frequency noise, producing too many groups for the overall cluster space to be useful. To overcome this issue the dataset is preprocessed by averaging every K samples. This averaging process not only reduces the number of dimensions of the space over which to perform the clustering process from P down to $\lceil P/K \rceil$, but also smooths out any local disturbance the measurements may undergo, reducing the chances of producing too many clusters (more than the necessary to represent the actual traffic patterns).

Two clustering algorithms have been considered to produce groups within the feature space based on the above similarity metric: DBSCAN [180] and Affinity Propagation [181]. DBSCAN is a density based clustering algorithm which delimits clusters by regions where the density of samples is high, labeling points located in low-density regions as *outliers*. Affinity Propagation is a clustering algorithm based on exchanging messages between the different data points. It finds *exemplars*, members of the input set that are representative of clusters. Neither of them require the number of clusters as an input, as opposed to other clustering techniques such as K-Means. Both methods produce similar results when their parameters are chosen appropriately, therefore only one of them (DBSCAN) has been used for the experiments.

The result of the clustering algorithm produces a partitioning of $\boldsymbol{\mathcal{H}}$ into C clusters $\{\boldsymbol{\mathcal{H}}_c\}_{c=1}^{C}$. Cluster $\boldsymbol{\mathcal{H}}_c$ is represented by its centroid $\mathbf{o}^{s,\odot_c} = [o_{t_d}^{s,\odot_c}, \ldots, o_{t_d+P-1}^{s,\odot_c}]$, which is computed by taking the average of its member observations, *i.e.* the p-th element of the centroid is computed by taking the average of the p-th elements of all the members of that cluster:

$$o_p^{s,\odot_c} = \frac{1}{|\boldsymbol{\mathcal{H}}_c|} \sum_d o_{t_d+p}^{s} \ , \tag{3.5}$$

where $o^{s,d} \in \boldsymbol{\mathcal{H}}_c$ and $|\cdot|$ denotes cardinality of a set.

Classification

The previously defined clustering process would suffice for imputing missing values, and in fact it will be used as a comparison method in the experiments later discussed, choosing the closest $\mathbf{o}^{s,\odot_c}$ to the element of $\tilde{\boldsymbol{\mathcal{G}}}$ where the imputation is needed. However, when there is a particular day for which all measurements are missing – *i.e.* $P^{s,d} = 0$, the clustering process is not able to assign it to any of the clusters. In order to overcome this shortcoming, external information independent from the traffic data is incorporated to the dataset, and an algorithm is built over the C clusters obtained in the method explained in the previous subsection. A supervised learning classifier is trained with cluster indexes as classes, and over those features that do not depend on the actual traffic observations. We have designated as features the day of the week D, the month M and a binary feature bH to indicate whether a day is a bank holiday [34]. These time-related features are very relevant to group traffic by days, as traffic patterns are mostly daily cyclical [169]. Other external features such as the weather

or the celebration of regular events could also be included to obtain a more precise classification. Thus, a dataset with 3 features and C classes is composed from $\boldsymbol{\mathcal{H}}$, which is used to train a supervised classifier to estimate the cluster assignment of a day belonging to $\tilde{\boldsymbol{\mathcal{G}}}$.

The supervised learning model utilized for the regression problem posed in this chapter is Random Forest, which relies on the *bagging* concept [182], [183] to create a diverse set of regressors by introducing randomness in the construction of an ensemble of tree learners. This procedure has been shown to decrease the variance of the model without increasing its bias, as weak learners are fed with different training sets that consequently decorrelate their structure and provide diversity to the ensemble. Imputation is finally done by equaling missing entries of the tested day to those of the centroid $\mathbf{o}^{s,\odot_c}$ of the cluster to which it is predicted to belong. Specifically, if $\tilde{\mathbf{o}}^{s,d}$ denotes a test day for which $P^{s,d}$ missing entries are to be imputed, the proposed method creates a vector with components:

$$\hat{o}^s_{t_d+p} = \begin{cases} o^{s,\odot_c}_p & \text{if } \delta\,(t_d+p)=0, \\ \tilde{o}^s_{t_d+p} = o^s_{t_d+p} & \text{otherwise.} \end{cases} \tag{3.6}$$

The whole clustering classification process is graphically summarized in Figure 3.3.

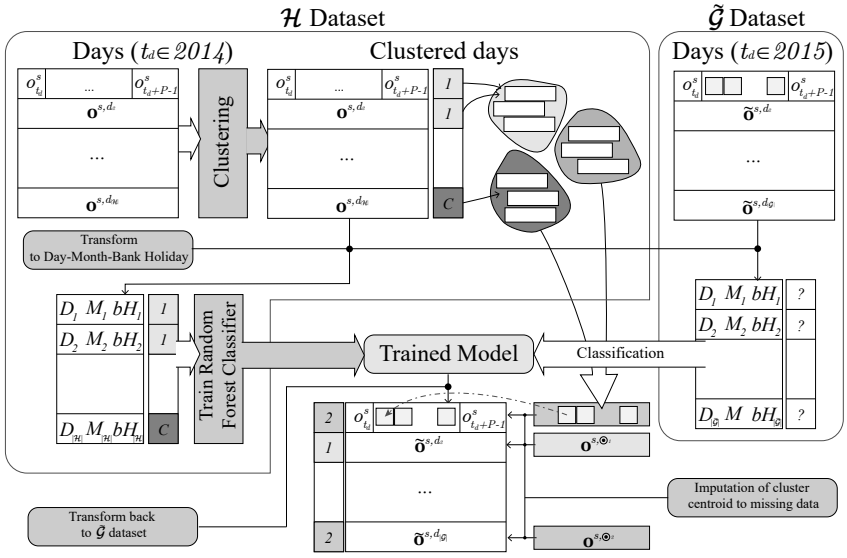

FIGURE 3.3: Clustering-Classification process.

3.2.4 Methods for Comparison

A selection of the most common methods have been used to appraise the performance of the ones here proposed. Early research in this field [141], [151] adopted some of the basic imputing methods that have been used ever

since for comparison: historical average, average over surrounding locations, average over close timestamps, or Expectation-Maximization (EM) methods. The diversity of imputation methods reported in the literature has grown lately, achieving high levels of complexity, but in general they continue to be benchmarked against the portfolio of imputing techniques mentioned above on account of their good performance when missing data entries are not profuse. Following this common practice, we have compared our proposed methods to 5 techniques of increasing complexity:

- *Basic Imputation (BASIC):* this naïve approach consists of imputing a constant value for all missing data, usually 0 [141] or a value based on statistical characteristics of the dataset, commonly the average of non-missing observations [139], [141], [142], [157]. Although imputed values would probably differ from the actual ones, this method provides effortlessly a dataset without missing data, allowing for the application of forecasting techniques in a straightforward manner.

- *Linear Interpolation (INT):* a reliable technique when missing data are scarce and individual [139]. Imputing gaps of a small amount of positions with linear interpolation is fast, easy, produces fairly accurate values, and provides a smooth traffic profile. However, it degrades severely when the length of intervals with missing data increases.

- *Mean Day Variation (MDV):* this is one of the most common techniques to impute missing data, which resorts to averages of the available observations at the same time index of the day to compute the value to fill in the missing entry [145]. In order to quantify the impact of the ratio ξ of missing values in the performance of this method, we have considered 2 possible input datasets: $\mathcal{H}^s$ dataset (corresponding to data captured in 2014 without any missing values) (MDV14) and $\tilde{\mathcal{G}}^s$ (corr., 2015 with artificial missing data) (MDV). The latter case depends on the quantity of missing data, and the performance of MDV is expected to degrade as the ratio ξ of missing values increases.

- *1-Nearest Neighbor (1NN):* this method, similarly to the clustering-classification scheme proposed in this chapter, relies on the day-splitting paradigm described in Subsection 3.2.3.2. Conceptually similar to the method proposed in [162], days with missing data from $\tilde{\boldsymbol{\mathcal{G}}}$ are compared to those in the dataset with complete days $\boldsymbol{\mathcal{H}}$, looking for the closest one under the measure of similarity given in Expression (3.4). Missing values within the incomplete sample $\tilde{\mathbf{o}}^{s,d}$ are filled with those of its closest instance in $\boldsymbol{\mathcal{H}}$.

  This procedure is simple to implement and computationally efficient, but presents several potential problems: 1) when $\boldsymbol{\mathcal{H}}$ is large finding the minimum through a exhaustive search can be time demanding; 2) when trying to impute values for a day with completely different measurements than any of the days in $\boldsymbol{\mathcal{H}}$ the process will produce values far from the *real* ones; 3) this procedure might be highly influenced by high frequency noise in the data sample.

- *Clustering (CL):* the use of clustering techniques has become frequent in the field of missing data imputation [166], [167]. For comparison purposes we consider a clustering algorithm defined analogously to the one described in Subsection 3.2.3.2. In this standard clustering, instances from the $\tilde{\mathcal{G}}$ dataset are mapped directly to the clusters defined with the $\mathcal{H}$ dataset, instead of creating a proxy classifier, which in turn represents the core contribution of our clustering method. To this end, clusters are selected based on the minimum distance – as per (3.4) – between the instance to be imputed and the cluster centroids. Missing data of a particular day are filled with the averaged values of the cluster it belongs to. As other methods that rely on partitions of the dataset on a per day basis, this technique is expected to fail when entire days of data are missing.

3.2.5 Quantifying the Imputation Performance

Missing data imputation should not be regarded as an end in itself, but a necessary step to reconstruct data and perform forecasts. In contrast with some authors that develop robust techniques to predict traffic regardless the missing data [142], [169]–[171], most authors validate the imputing results by measuring their distance to real data, but they do not test the prediction accuracy of methods that use imputed data as inputs. In some cases, an imputation strategy can provide a marginal improvement over traditional approaches, but at a high computational cost. This efficiency trade-off could be worthless in practice should the differences between prediction performances with and without imputed data be negligible.

For this reason, besides the usual imputation error analysis, we propose an alternative methodology for assessing the prediction performance for each method. Once missing observations have been imputed over $\tilde{\mathcal{G}}^s$, an *imputed* dataset $\hat{\mathcal{G}}^s$ is produced for each value of ξ (ratio of missing observations) and L (interval length). Each dataset $\hat{\mathcal{G}}^s$ is split in two chunks: one containing the first 80% of observations ($\hat{\mathcal{G}}^{trn}$, *training*) and the second with the remaining 20% (corr. $\hat{\mathcal{G}}^{tst}$, *test*), for which imputed data $\hat{o}_t^s$ are replaced with their respective real values o_t^s. Thus, $\hat{\mathcal{G}}^{tst}$ is for all cases the same chunk of real test data, whose observations belong to $\mathcal{G}^s$. A Random Forest (RF) regression model is then trained on each of the $\hat{\mathcal{G}}^{tst}$ datasets, after which predictions are obtained and tested against $\hat{\mathcal{G}}^{tst}$. The prediction scores achieved with each input set of data (with original or imputed values) are compared to each other, providing insights on how the accuracy of every technique propagates to the prediction score of predictive models when the imputation is used to reconstruct datasets for traffic forecasting.

As depicted in Figure 3.4 the model is built analogously to the one presented in Subsection 3.2.3.1: a window of w observations (fixed to 20) predicts the value of the observation $h = 1$ steps in the future.

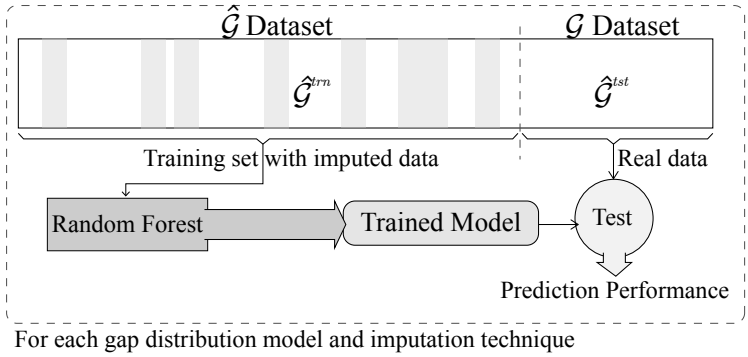

FIGURE 3.4: Proposed method for evaluating the performance and quality of the values imputed by every technique in a context of traffic forecasting.

3.3 Results and Discussion

After conducting the experiments described in previous sections, we examine the performance of the proposed missing data imputing techniques in compared with the methods enumerated in Subsection 3.2.4. For the sake of statistical characterization, 20 independent runs have been completed for each percentage and missing interval distribution with different random positions of the missing points and intervals. Thus, each missing data imputation method is evaluated against 20 different sets of missing data ratios ξ and length interval L, except for $\xi = 100$, as no different combinations of missing data can be performed when all the data are missing. The score utilized to evaluate results is the Root Mean Squared Error (RMSE), defined as:

$$\text{RMSE} \doteq \sqrt{\frac{1}{N} \sum_{\forall t:\delta(t)=0} (o_t^s - \hat{o}_t^s)^2}, \tag{3.7}$$

where $N \doteq \sum_{t \in 2015} (1 - \delta(t))$ denotes the number of imputed observations in $\tilde{\mathcal{G}}^s$.

Also, the coefficient of determination R^2, which shows how approximate are the predicted values to the real ones, is calculated according to:

$$R^2 \doteq 1.0 - \frac{\sum_{\forall t:\delta(t)=0} (o_t^s - \hat{o}_t^s)}{\sum_{\forall t:\delta(t)=0} (o_t^s - \bar{o}_t^s)} \ . \tag{3.8}$$

These two evaluation metrics are averaged over the 20 experiments, obtaining averages and standard deviations reported in Tables 3.2 and 3.3 for RMSE and Tables 3.4 and 3.5 for R^2. The imputing methods are identified as defined in Subsection 3.2.4, using also SSC for Spatial context sensing complete (without optimization of window size), SSO for Spatial context sensing optimized and PCC for Pattern clustering and

classification. Estimations are shown considering one significant figure, following the criteria described in [184].

TABLE 3.2: RMSE results for different percentages ξ of point-wise missing data. In this and following tables, results are shown as *mean±standard deviation*, and statistically best results (determined by a Wilcoxon test with 95% confidence interval) are highlighted in bold.

	ξ					
Method	**1 %**	**10 %**	**25 %**	**50 %**	**80 %**	**100 %**
BASIC	420 ± 30	420 ± 10	438 ± 9	489 ± 5	586 ± 2	661
INT	**65 ± 6**	**70 ± 5**	**75 ± 4**	**92 ± 5**	175 ± 9	500
MDV	260 ± 30	250 ± 10	248 ± 6	250 ± 5	251 ± 3	661
MDV14	260 ± 30	250 ± 10	249 ± 6	249 ± 3	248 ± 1	249
1NN	110 ± 20	107 ± 6	107 ± 4	109 ± 3	**121 ± 3**	390
CL	110 ± 20	117 ± 8	117 ± 5	116 ± 2	**121 ± 3**	400
SSC	140 ± 20	133 ± 7	134 ± 4	133 ± 2	132 ± 1	132
SSO	130 ± 20	128 ± 6	127 ± 3	126 ± 2	126.6 ± 0.6	126
PCC	120 ± 20	122 ± 8	124 ± 5	122 ± 2	122 ± 1	**122**

TABLE 3.3: RMSE results for different length L intervals of missing data.

	L					
Method	**24**	**48**	**72**	**96**	**144**	**192**
BASIC	420 ± 40	410 ± 20	420 ± 30	440 ± 30	420 ± 40	420 ± 30
INT	300 ± 50	430 ± 50	470 ± 70	560 ± 80	490 ± 60	530 ± 70
MDV	260 ± 30	240 ± 30	240 ± 30	240 ± 20	260 ± 20	240 ± 20
MDV14	250 ± 20	240 ± 20	240 ± 20	250 ± 10	260 ± 20	250 ± 20
1NN	130 ± 20	130 ± 10	170 ± 30	230 ± 50	260 ± 50	310 ± 60
CL	120 ± 20	140 ± 10	160 ± 30	240 ± 60	270 ± 40	320 ± 50
SSC	130 ± 20	130 ± 20	**120 ± 10**	**121 ± 6**	122 ± 8	**120 ± 8**
SSO	123 ± 7	124 ± 4	124 ± 5	123 ± 5	123 ± 6	124 ± 7
PCC	**120 ± 10**	**120 ± 10**	**120 ± 10**	122 ± 9	**120 ± 10**	122 ± 9

Results displayed in both sets of tables lead to similar conclusions. When missing data consist of percentages of random missing points, most methods perform reasonably well even when percentage of gaps ξ reaches 80%. As expected, when all data are missing (ξ = 100%), any technique relying on the availability of these data fails. Basic imputing does not yield good results in any case, as we are comparing values obtained from the distribution defined by real traffic to a constant. Mean-based methods produce an outcome with stability along different values of ξ and L, due to the averaging process that uses all the measurements available for a certain time frame of the day, disregarding the differences among types of days, which are remarkable in these central locations of the city. Linear interpolation outperforms the rest of methods in almost any scenario, which reflects the main inconvenient of this random missing data generation method: empty positions are distributed uniformly, and rarely in lengthy groups, allowing a simple linear interpolation method to produce good estimations. But this randomly uniform distribution does not commonly represent the real disposition of gaps.

TABLE 3.4: R^2 results for different percentages ξ of point-wise missing data.

	ξ					
Method	**1 %**	**10 %**	**25 %**	**50 %**	**80 %**	**100 %**
BASIC	-0.009 ± 0.007	-0.018 ± 0.008	-0.1 ± 0.01	-0.38 ± 0.02	-0.98 ± 0.01	-1.52
INT	0.975 ± 0.006	0.972 ± 0.004	0.968 ± 0.003	0.951 ± 0.005	0.82 ± 0.02	-0.17
MDV	0.60 ± 0.09	0.65 ± 0.02	0.65 ± 0.02	0.64 ± 0.02	0.635 ± 0.009	-1.52
MDV14	0.61 ± 0.06	0.65 ± 0.02	0.64 ± 0.01	0.64 ± 0.009	0.644 ± 0.003	0.64
1NN	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.935 \pm 0.005}$	$\mathbf{0.934 \pm 0.004}$	$\mathbf{0.931 \pm 0.003}$	$\mathbf{0.916 \pm 0.005}$	0.11
CL	0.92 ± 0.03	0.923 ± 0.008	0.922 ± 0.006	0.922 ± 0.003	0.915 ± 0.004	0.10
SSC	0.89 ± 0.04	0.899 ± 0.009	0.896 ± 0.007	0.898 ± 0.004	0.899 ± 0.002	0.90
SSO	0.90 ± 0.03	0.91 ± 0.01	0.907 ± 0.005	0.908 ± 0.003	0.907 ± 0.001	**0.91**
PCC	0.91 ± 0.04	0.915 ± 0.008	0.912 ± 0.006	0.914 ± 0.004	0.914 ± 0.002	**0.91**

TABLE 3.5: R^2 results for different length L intervals of missing data.

	L					
Method	**24**	**48**	**72**	**96**	**144**	**192**
BASIC	-0.03 ± 0.04	-0.03 ± 0.04	-0.04 ± 0.03	-0.04 ± 0.02	-0.03 ± 0.04	-0.03 ± 0.02
INT	0.4 ± 0.1	-0.2 ± 0.3	-0.4 ± 0.4	-0.7 ± 0.4	-0.4 ± 0.3	-0.7 ± 0.4
MDV	0.6 ± 0.1	0.6 ± 0.1	0.6 ± 0.1	0.7 ± 0.1	0.6 ± 0.1	0.64 ± 0.09
MDV14	0.61 ± 0.08	0.64 ± 0.07	0.64 ± 0.08	0.66 ± 0.08	0.60 ± 0.08	0.64 ± 0.06
1NN	0.90 ± 0.02	0.89 ± 0.02	0.82 ± 0.07	0.7 ± 0.1	0.6 ± 0.2	0.4 ± 0.2
CL	0.91 ± 0.02	0.89 ± 0.02	0.83 ± 0.07	0.7 ± 0.1	0.6 ± 0.1	0.4 ± 0.20
SSC	0.90 ± 0.03	0.88 ± 0.04	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.91 \pm 0.01}$
SSO	0.91 ± 0.02	0.90 ± 0.01	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.91 \pm 0.01}$
PCC	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.91 \pm 0.01}$

This issue is observed more clearly when inspecting Tables 3.3 and 3.5, where results for any of the comparison methods (except for the mean-based ones and the basic imputing) degrade severely as the length of gaps L is increased. With linear interpolation, this effect is specially noticeable: having 48 missing entries (half a day), the interpolation is made between two points within which great traffic variations may occur, hence traffic data in between cannot be modeled by a line. Above that size of gap, this technique is unable to produce acceptable values. Clustering and 1NN similarity interpolations behave similarly in both situations: with point-wise missing data generation, they are able to model fairly well traffic with the data they have available for each day, even with a 80% of missing data. When interval-wise gaps are 1 day long ($L = 96$ positions) their performance decays; no data are available to establish their similarity, thus they are assigned to random days. Figure 3.5 shows a boxplot for both kinds of gap generation scenarios, considering an interval of $L = 96$ positions, being this the most common, and a 10% of missing data, being this percentage the most similar to the total amount of missing data in the interval-wise mode (13%). This figure displays clearly that for a very similar volume of missing data, the way in which gaps are generated affects severely the RMSE results, except for the more robust techniques, such the ones proposed in our work (SSC, SSO, PCC).

In contrast with the performance deterioration that all comparison methods suffer when long intervals or complete absence of data are introduced, our proposed methods achieve a stable operation independent of the abundance or size of data gaps. Among them, spatial context sensing, based in measurements from surrounding sensors is more resilient to the unavailability of data during a certain time frame. Two versions of

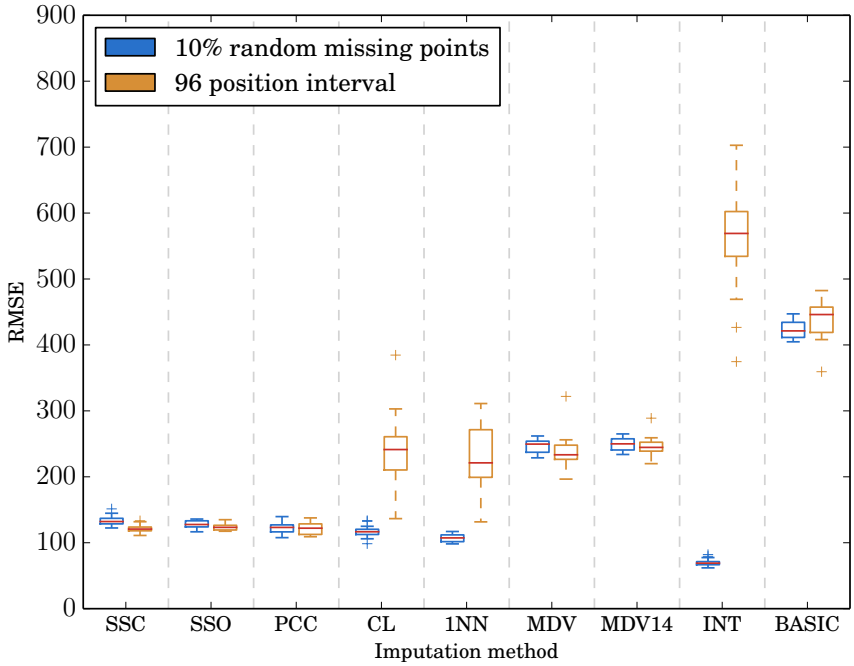

FIGURE 3.5: Imputation performances for both gap generation models.

this scheme have been tested: one with "complete" sets of measurements (50 sized time windows for each surrounding loop, resulting in a dataset of 600 features), and one optimized by using the bio-inspired heuristic wrapper described in Subsection 3.2.3.1. In the latter, optimized window sizes are never greater than 6, and for some of the loops are 0, indicating in these cases the null importance of those sensors for the imputation and subsequent prediction procedures. This reduces the number of features to 40 on average, and speeds up the prediction algorithm used for imputing data: in an Intel i7 Linux machine with 16GB of RAM, the running time of SSO is 50% of the time required to run SSC.

Besides this lower computational cost, the outcomes of both versions of the method in terms of imputing performance are very similar; therefore, a Wilcoxon test has been run in order to examine if such a difference is statistically significant. The p-values of this test are shown in Table 3.6. They demonstrate that within a 95% confidence interval, only in some of the point-wise percentage scenarios the results provided by SSO are significantly different from those obtained with SSC. Even in those cases, maximum error differences amount up to only 6 vehicles per hour. Thus, the complexities involved in the optimization of the algorithm could be avoided when the computation capacity is not a practical limitation.

PCC also performs robustly in all considered situation. Moreover, its operation is more efficient in terms of computational complexity, with runtimes in the order of 94% relative to the time taken by SSC for the same

TABLE 3.6: Wilcoxon test p-values showing statistical significance of differences between RMSE results of SSC and SSO.

Methods	**1%**	**10%**	**25%**	**50%**	**80%**	**100%**	**24**	**48**	**72**	**96**	**144**	**192**
SSC vs SSO	0.09	0.02	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.41	0.1	0.05	0.11	0.63	0.09

scenario. A Wilcoxon test has been also performed to compare its results to those of spatial context sensing without optimization, rendering p-values shown in Table 3.7. As for the previous two methods, the statistically significant differences are found for the point-wise missing data, for which PCC performs better. Aside from this particular performance gain, it is remarkable that this method can be further improved by adding more traffic-agnostic features that in theory could improve the classification process.

TABLE 3.7: Wilcoxon test p-values showing statistical significance of differences between RMSE results of SSC and PCC.

Methods	**1%**	**10%**	**25%**	**50%**	**80%**	**100%**	**24**	**48**	**72**	**96**	**144**	**192**
SSC-PCC	0.01	$< 10^{-3}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.07	0.09	0.15	0.58	0.82	0.28

3.3.1 Prediction-wise Imputation Performance

According to the results discussed above, naïve similarity or interpolation methods should be used when missing data consist of short gaps, while the longer sized gaps would require alternative methods. Nonetheless, when the imputed data are analysed from the perspective of their ability to obtain accurate predictions, this intuition may change as discussed next:

TABLE 3.8: RMSE prediction results for different percentages ξ of point-wise missing data.

	ξ					
Method	**1 %**	**10 %**	**25 %**	**50 %**	**80 %**	**100 %**
BASIC	80 ± 2	103 ± 4	154 ± 10	240 ± 20	350 ± 20	618
INT	**77 ± 2**	**78 ± 2**	**76 ± 2**	**78 ± 1**	85 ± 1	400
MDV	81 ± 3	91 ± 4	107 ± 4	123 ± 10	130 ± 10	618
MDV14	82 ± 5	92 ± 3	107 ± 4	128 ± 7	135 ± 6	171
1NN	78 ± 2	82 ± 4	85 ± 4	87 ± 3	**83 ± 3**	**76**
CL	78 ± 2	81 ± 5	81 ± 3	84 ± 3	86 ± 3	95
SSC	80 ± 3	86 ± 4	91 ± 4	95 ± 3	97 ± 3	89
SSO	79 ± 2	81 ± 3	85 ± 3	87 ± 3	87 ± 2	86
PCC	78 ± 2	79 ± 4	81 ± 3	86 ± 4	94 ± 8	89

Tables 3.8 and 3.9 shows the averaged RMSE and its standard deviation for 20 forecasts performed for each method and missing data model, following the process described in Section 3.2.5. With this evaluation approach, the most relevant matter is the abundance of missing data, as for low percentages the rest of data are enough to build a good forecasting model, regardless the quality of imputed values. This is clearly observed

TABLE 3.9: RMSE prediction results for different length L intervals of missing data.

	L					
Method	**24**	**48**	**72**	**96**	**144**	**192**
BASIC	78 ± 2	78 ± 2	78 ± 2	77 ± 2	76 ± 2	78 ± 3
INT	75 ± 2	75 ± 2	76 ± 3	77 ± 2	77 ± 2	77 ± 2
MDV	76 ± 2	75 ± 1	75 ± 2	75.0 ± 0.8	76 ± 1	76 ± 1
MDV14	$\mathbf{74 \pm 1}$	$\mathbf{74 \pm 1}$	74 ± 1	$\mathbf{74 \pm 1}$	$\mathbf{73.7 \pm 0.8}$	$\mathbf{74 \pm 1}$
1NN	79 ± 3	79 ± 2	76 ± 2	76 ± 2	76 ± 2	76 ± 3
CL	77 ± 1	76 ± 2	$\mathbf{73 \pm 2}$	74 ± 2	74 ± 2	73 ± 2
SSC	76 ± 2	77 ± 1	77 ± 1	77 ± 1	77 ± 1	76 ± 1
SSO	78 ± 2	78 ± 3	78 ± 2	78 ± 2	78 ± 2	77 ± 2
PCC	78 ± 2	78 ± 2	78 ± 2	78 ± 2	78 ± 2	78 ± 2

in Table 3.9 , which represents 13% missing data for all cases. Although up to 2-day long gaps are created in this experiment, and even if we know that imputed values are inaccurate for some of the methods, all of them perform well, because the 87% remaining real data are sufficient to train properly the regression model. The same algorithm applied to a dataset with no missing data returns a score equal to RMSE = 74 ± 2, very similar to those presented in Tables 3.8 and 3.9 . Undoubtedly, this good performance owes much to the way in which ML algorithms operate and model data. When imputed data are "bad", they are essentially noise for the training, hence the forecasting performance will also depend on the algorithm's ability to deal with noise. Similarity methods (1NN and Clustering) impute a default day when they are not able to find a similarity pattern. This default day comes from the same loop and it is probable that its profile is similar to the missing data, therefore the acceptable results of completely missing dataset. On the contrary, for great amounts of missing data, performance degrades for naïve methods, but our proposed algorithms generate fairly robust predictions.

3.4 Conclusions

In this chapter we have investigated into missing data imputation strategies for traffic data, covering all stages of the procedure, from the creation of datasets with artificial gaps to the evaluation of the obtained results. We have also presented two novel missing data imputing methods based in contextual and historic information of a certain traffic measuring point, aimed to obtain accurately imputed values when missing data are abundant or presented in long intervals. Real data obtained from the traffic sensor network deployed in the city of Madrid (Spain) reveal that a noticeable proportion of the locations present lengthy intervals of missing data, making it necessary to design an efficient imputation strategy for scenarios with these missing data characteristics.

First, the missing data model has been approached from the perspective of the allocation of artificially created missing points. Missing data

distribution can vary substantially among different ATRs in a city, and much more among cities, thence a proper examination of the real gap distribution of a dataset should be performed prior to the generation of artificial gaps. For this particular dataset we have found out that simple random point-wise generation strategies fail to represent the true nature of the gaps. In fact, some of the evaluated imputation techniques perform reasonably good with high missing data percentages, but decay quickly when the length of missing data is increased, even if the total amount of gaps is low. A fair amount of the literature reviewed in this manuscript does not test validate models with long intervals, so it is not possible to know how they would perform if that were the case.

Furthermore, the evaluation of imputation results has been discussed. Beyond the similarity measurement of performance that is found in most of the literature, (based on comparing the imputed value to the actual one), we have gauged the imputation performance by considering the forecasting ability of a dataset built with those imputed data. As a result, we have concluded that if the prediction technique is based on a machine learning approach capable (to a certain extent) of dealing with noisy data, and the amount of imputed data is low (less than 20%), the most basic and inexpensive imputing method can perform as good as the most complex one, probably due to the capacity of the forecasting algorithm to properly model the data despite the gaps. Such techniques are currently popular within traffic forecasting, so a previous analysis of the distribution and profusion of missing data could save a significant processing time in future contributions dealing with this topic.

The presented imputation methods have produced steady outcomes regardless the distribution and size of missing data, and indifferently to the evaluation procedure. Our spatial context sensing algorithm has proven that with enough contextual information, it is possible to impute missing values even when no data is available. This reinforces the notion of relations among traffic profiles registered over an entire city, even though they are not directly upstream or downstream correlated. On the other hand, our clustering approach highlights again the relevance of external features to obtain traffic patterns, an easily obtained input that is scarcely used in traffic forecasting and imputing research. Three simple temporal characteristics have been enough to obtain a good performance of the clustering-classification algorithm, but in theory, other external features such as sports events, demonstrations or spatial information about locations of interest could enhance further the predictive power of these patterns. Both introduced techniques require big amounts of previous or spatial context data that are not always available, but their operation is flexible enough to take data from any existing sources.

In light of the experiments, adaptive techniques should be implemented for data imputing, using a mix of computationally efficient methods, such as linear interpolations for individual or short interval missing values, and more complex algorithms for filling in long intervals of missing data. This hybridization will lie at the core of future research lines stemming from this work.

Chapter 4

Long-term Road Traffic State Estimation

4.1 Related Work

Short-term forecasting methods have involved time-series analysis and a wide diversity of variants and enhancements of machine learning approaches; even Big Data technologies have provided new angles to tackle the traffic forecasting problem, almost always in the short prediction horizons. Although there are evident obstacles to achieve long-term forecasts, recent data driven approaches, and the availability of big volumes of data and contextual information have pushed forward new advances in long-term forecasting methodologies. These methodologies are mainly based on defining patterns that represent typical traffic profiles in different circumstances, while trying to assign future situations to one of such patterns. The ways in which the patterns are defined and the assignation performed comprise an assortment of methods and techniques. In this line of work, [69] defined the long-term predictions as trends over which Principal Component Analysis is applied to detect abnormal cases. In [70], context information from surrounding measuring stations is used to build similarity patterns, and then short-term forecasting is performed with previously defined ground truth. Statistical models are used in [185] and [186]. Both use B-splines to estimate traffic flow and characterize types of days. A clustering and proxy dataset approach was presented in [34], and will be used as an starting point in this chapter, introducing some improvements.

Online learning strategies are also a trend, seeking models that can operate with a few observations of current traffic. In [53], authors proposed an online traffic prediction algorithm that predicts with real time readings leaning on ensembles of weak predictors. In the wide range of ML techniques, neural networks and their variations are particularly popular [13], [110] for traffic forecasting; however, finding contributions that make use of the potential of the so-called third generation of neural networks is challenging.

Spiking Neural Networks (SNNs) [187] were originally developed to obtain more accurate representations of biological neural networks in mammals [188]. This technique and its variants simulate the operation of actual neurons, by communicating among them with sequences of spikes, and representing accurately the operation of synapses [189] as learning rules. Nevertheless, their modeling capacity goes beyond the representation of complex spatio-temporal patterns: they have been proven an efficient tool for a range of engineering problems and other fields [190]–[194]. However, it is surprising that no works on traffic forecasting have so far relied on SNNs, despite their specialized capability to represent spatio-temporal data [195], and being this one of the hot topics of traffic modeling [13]. Besides this relevant feature of the general SNN model, one of its variants, evolving Spiking Neural Networks (eSNN) [196], presents a certainly appealing characteristic for online traffic forecasting: an eSNN model can grow and learn new information without retraining it, by evolving (i.e. incrementally adding) spiking neurons [189]. This particular trait provides eSNNs with a fast updating structure, of utmost utility for the adaptive online learning proposed in what follows.

The aforementioned trends, regarding the exploitation of large amounts of traffic data for long-term predictions in an online fashion, confront relevant challenges associated to its stochastic nature and evolution during long periods of time. Indeed, the analysis of large streams of data that may cover temporal ranges of years has become the focus of an exhaustive research area, in which *online learning* and *concept drift* detection and adaptation are central topics. Concept drift – namely, a change in the underlying distribution of streaming data [197] – is present in many fields where data streams are generated, such as computer and sensor networks, financial markets, mobile phones, intelligent user interfaces, and of course, traffic [198]. The data generation process may be affected by non-stationary phenomena (i.e. seasonality, periodicity, sensor errors, and the like), causing models trained over these data to become obsolete and consequently unable to adapt to new data distributions. In order to overcome these issues, concept drift detection, characterization and adaptation has gained popularity in recent literature [199], [200], essentially due to the necessity for adaptive prediction techniques that blend together drift detection and adaptation for these changing environments [201].

Aside from drifts provoked by the passing of time and inherent evolution of traffic, works that use big traffic data [65] have capitalized on other issues with more immediate effects that can be also addressed with similar change detection and adaptation tools. Among them, some very frequent in long-term traffic prediction are unplanned incidents or events, the planned ones that were not contemplated in the original model, or even a simple bad pattern assignment due to the classifier fallibility. All of them can render the long-term pattern model useless in different degrees, and detecting them and adapting quickly to new circumstances is crucial for a proper operation of the forecasting tools.

4.1.1 Contributions

This chapter finds its motivation in this noted need for adaptation to changes in traffic forecasting. Specifically, we propose a novel methodology that yields long-term forecasts using similarity-based clustering of daily traffic volume data, and monitors them in real time to adapt them in case of a mismatch or contingency. All of this is achieved by means of eSNN techniques and a change detection and adaptation mechanism. The main contributions of this research work can be summarized as follows:

1. The implementation of an improved method to obtain traffic patterns for any location and date that can be used by ATIS and ATMS.

2. The use of eSNNs in the traffic domain and the implementation of a method to encode traffic data into spikes over the time domain that feed an eSNN.

3. The development of an online detection and adaptation mechanism that constantly evaluates and updates pattern-based forecasts, taking into consideration current traffic volume observations.

4. The proposal of a framework to effectively maintain long-term predictions with incremental short-term adaptations which can be self updated with new knowledge embedded in newly arriving data samples.

The rest of the chapter is organized as follows: Section 4.2 describes the input data, the two step forecasting model and delves into the encoding of traffic volume to spikes of an eSNN, as well as the adaptation process. Section 4.3 presents and analyses the performance of the proposed methods in different missing data scenarios. Finally, Section 4.4 draws concluding remarks inferred from the obtained results and prescribes future research lines related to this work.

4.2 Materials and Methods

This research work introduces a long-term forecasting model with a change detection and adaptation mechanism. The general operation of the model is shown in Figure 4.1. The method consists of two phases, each relying on a different subset of traffic data. A first subset is used in an offline learning phase, obtaining clusters that define daily traffic patterns and subsequently fed to an eSNN classifier. This classifier is then used in the online detection and adaptation phase, which is fed by the second subset of data. The classifier assigns a pattern to a day, and the change detection and adaptation mechanism allows to reassign the day to another pattern if it was originally incorrect. The source and selection criteria for the input data, as well as the hybrid forecasting method is detailed in the following subsections. Nevertheless, the proposed framework can process any other data source without any loss of generality.

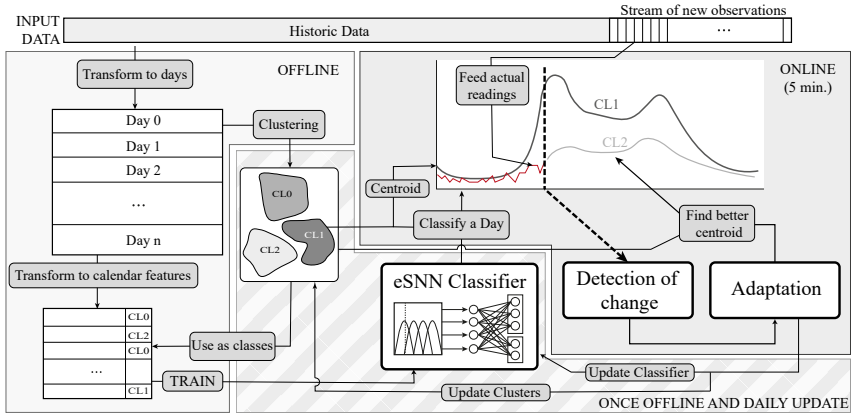

FIGURE 4.1: Offline-online model for pattern classification and adaptation.

4.2.1 Input Data

As in Chapter 3, real traffic data were collected from the public source described in Appendix A. In this case, aiming for a greater granularity, a collection tool was implemented to capture readings from a set of sensors in a set of locations in the one of the business areas of the city. Data from November 2016 to June 2017 were gathered and aggregated into $\Delta T = 5$ minutes intervals. The selection of the sensors for the experiments was made by attending to a data completeness criterion, as for most locations, reading errors are produced, resulting in frequent gaps in the data streams [202]. In a real life scenario, a certain amount of available historic data are used to build a model that will be applied later to new data as they are received. In order to mimic this scheme, our stream of data consists of 59328 consecutive observations o_t, which amounts up to 206 days. These data were then divided into two datasets, $\mathcal{H}$, for historic data that allow detecting and defining patterns, and $\mathcal{P}$ for new data that will be assigned to patterns. The subset $\mathcal{H}$ contains 75% of observations o_t, that are rounded in terms of full days, to $|\mathcal{H}| = 154$ days or 44352 observations. The remaining 25% constitute the subset $\mathcal{P}$, adding up to $|\mathcal{P}| = 52$ days.

The division of the dataset is strictly chronological and not stratified, thus leaving out the first offline phase a model that is trained with less information than we have available. For instance, in our case, with data starting in November and ending in June, the partition is made in the middle of April. This means that the offline model will have learned particularities of the Christmas period, but it will not be trained for special holiday days happening in May. This intends to mirror the behavior of the model in real life, where beyond its best efforts in the offline modeling phase, it could find circumstances producing traffic profiles very different from the assigned pattern in the online phase. With this division, a need for adaptation is forced into the test data $\mathcal{P}$ in order to assess properly the change detection and adaptation mechanism. If a non predictable circumstance has happened before, the model will be able to recognise it and provide a pattern, having potential to deal with other unpredictable events

like accidents or abrupt weather changes.

4.2.2 Offline Processing: Clustering Traffic Data and Building a Classifier

In the online phase of our forecasting method, traffic patterns are assigned to days for which there is no known traffic profile. Thus, the offline phase is an initialization procedure that produces:

- A well defined set of patterns that suffice for representing the variety of kinds of days for a given location.
- A classification model capable of predicting the most suitable pattern for a day without having any of the traffic observations for that day, thus relying on features not related to traffic that are available for future days.

In the Chapter 3, and also in related previous research [34], [202], we have successfully performed traffic pattern discovery and classification with different purposes. The initial processing phase of our proposed model builds upon these previous works, introducing further changes that allow for a fast online classification and improvements in the modeling of a dataset with features not based in traffic data. As in these previously developed methods, and in other works [140], [156], [158], [160], [167], the continuous sequence of traffic volume observations is split into day-wise vectors given by:

$$\mathbf{o}^d = \left[o_{t_d}, o_{t_d+1}, \ldots, o_{t_d+P-1}\right], \tag{4.1}$$

where o_{t_d} is the traffic volume captured at time t_d, being t_d the first observation of day d; and $P = 288$ is the number of observations obtained within a day for a capture period of $\Delta T = 5$ min. These vectors conform a dataset $\boldsymbol{\mathcal{H}}$ over which a clustering procedure is performed, in order to find groups of similar days, aiming to label each day with the cluster it belongs to. The high dimensionality of $\boldsymbol{\mathcal{H}}$, with 288 features per sample, can hinder the clustering performance and its results could be distorted by the noise, which is inherent to individual samples. Therefore, the original $P = 288$ dimensions of the dataset are reduced to $\lceil P/K \rceil$ by averaging every K samples, smoothing each sample without loosing their ability to represent the traffic profile of every day.

4.2.2.1 Clustering

The clustering process has been performed with DBSCAN algorithm [180]. This clustering method defines areas in the data space by grouping samples delimited by their density in each area. Its suitability to our study is supported by two of its main features: it does not require a predefined number of clusters, a parameter that is not known *a priori*; and grouping noise samples in a separate cluster, otherwise they would be assigned to another existing cluster, producing biases when obtaining the centroid. The selection of the DBSCAN main parameters (i.e. the maximum distance

between samples to belong to the same cluster, and the minimum number of instances to be considered a cluster) can lead the algorithm to perform between two extremes: 1) an overly high number of clusters composed by very few data examples, and many noise instances; or 2) a low number of clusters, as a result of a high value of distance, to the point that if the distance parameter is high enough to fit all data in the same cluster, there is only one pattern, and no sample is left outside. Therefore, a balance between noise instances and number of clusters must be met in order to avoid such extremes and extract relevant patterns from the traffic data. To this end, an iterative process was conducted aimed to both reduce the number of noise instances and to maximize the number of clusters. No right amount of clusters was considered, but it was assumed that more clusters represent more types of days and more of their particularities.

A centroid $\mathbf{o}^{\odot_c} = [o_{t_d}^{\odot_c}, \ldots, o_{t_d+P-1}^{\odot_c}]$ is obtained for each of the resulting C clusters $\{\boldsymbol{\mathcal{H}}_c\}_{c=1}^{C}$, by averaging all the cluster members in an element-wise fashion:

$$o_p^{\odot_c} = \frac{1}{|\boldsymbol{\mathcal{H}}_c|} \sum_d o_{t_d+p} \ , \tag{4.2}$$

where $o_{t_d} \in \boldsymbol{\mathcal{H}}_c$ and $|\cdot|$ denotes cardinality of a set.

4.2.2.2 Proxy Dataset

After clustering, the dataset $\boldsymbol{\mathcal{H}}$ is defined with 288 features and a class, *i.e.* the cluster identifier. However, it is not suitable to train a classifier aimed to assign patterns to future days, because no readings will be available those days to build the samples. Consequently, a proxy dataset is built with the same number instances and classes as $\boldsymbol{\mathcal{H}}$, but a set of features that are known for the days to come. An analysis of the clusters obtained in $\boldsymbol{\mathcal{H}}$ can lead, along with field expert knowledge, to the extraction of features that can define each cluster. The choice of features is a crucial step of the presented modeling scheme. A feature selection process should be conducted individually for each location, like in any other classification or regression problem. In this work we consider a set of calendar features, which are among the most frequently used for this purpose [65], [70], [108], [203]. Although there are some other variables that are commonly very relevant for traffic, such as weather or traffic incidents, they present the same problem for this scenario as traffic features: they are unknown for future days, so the model would work with estimations. Disregarding them isolates the results from a noise inherent to a bad input prediction.

The selected calendar features are described below. They will be initially used for all locations, although some of them could be removed if they are irrelevant for a certain location:

- Day of week: Represented by a number between 1 and 7
- Month: Represented by a number between 1 and 12
- Public holiday: a number between 0 and 3 representing respectively: normal day, local holiday, region holiday, national holiday. Different types of public holidays impact the traffic in different ways [204].

- Academic holiday: this feature takes value 1 if the day is within a period of academic holidays and 0 otherwise. During these periods, not only the traffic is affected by a reduction of scholar transport, but it is also frequent that entire families go on holidays, affecting traffic considerably.

- Bridging days: Binary feature that represents when a working Monday precedes a holiday Tuesday or a working Friday is after a holiday Friday. Depending on the location, this days tend to have a different traffic pattern than normal Mondays or Fridays.

- Proximity to holiday: Traffic volume is also affected when a public holiday is close in time, as part of the population may decide to take vacation days to extend the break [204]. This has been encoded in the proxy dataset with this feature that takes values from 0 to 5, representing 0 no proximity, 5 holiday day, and 1 to 4, the inverse of the time distance to the holiday of the 4 surrounding days, before and after the holiday day (e.g., the next and previous days to the holiday would have value 4, and so forth).

The combinations of these features are able to portray most of the situations and types of days that can happen during a year and affect the traffic profile (without considering other unpredictable incidents, special events or meteorology extremes that can cause even more severe changes in traffic). With them, a dataset $\tilde{\boldsymbol{\mathcal{H}}}$ is formed with 6 features is built and used to train the classifier described in the following Section.

4.2.2.3 Classification with Evolving Spiking Neural Networks

Evolving Spiking Neural Networks are intelligent machines able to apply incremental learning rules to adapt their structure to the data. This feature of eSNNs makes them efficient to handle online classification problems and, since they were first proposed by [196], they have been applied to different types of data [205]–[207]. To the best of our knowledge, no applications have been found within the online traffic forecasting domain and only in [195] traffic data are used as a benchmark of their spatio-temporal representation capabilities. However, an eSNN classifier is well-suited to build a fast online learning model, as neurons are generated incrementally to allow the system to self-grow and learn new information using only one-pass data propagation.

As shown in Figure 4.2, an eSNN model is structured into different layers: firstly, every feature in a sample is transformed into a train of spikes using a number G of Gaussian receptive fields, as proposed by [208]. Centers and widths of Gaussian curves are defined for every feature, and each field represents a pre-synaptic neuron. This method uses overlapping Gaussian activation functions to populate continuous inputs over multiple neurons. This is a biologically inspired approach which simulates cortical neural processing of external inputs [209]–[211], and has been successfully applied to represent real-valued data [208], [212]–[214]. The points in

which each curve is cut by the real value define the times in which spikes are produced, as depicted in Figure. 4.3.

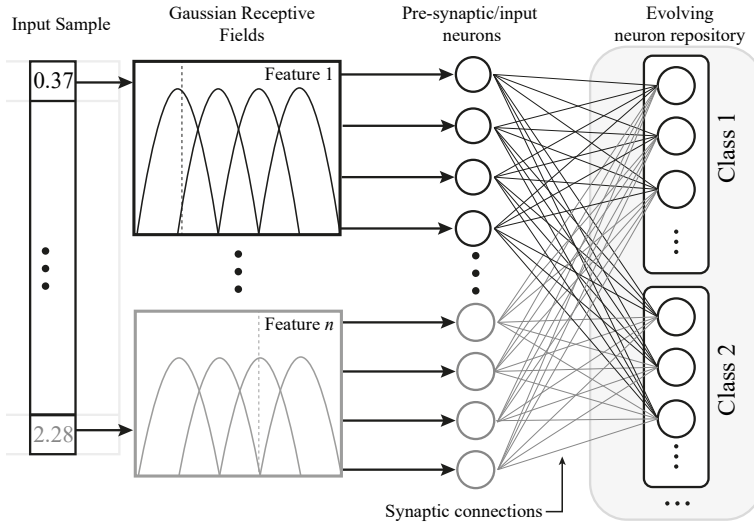

FIGURE 4.2: An eSNN architecture with its three main layers: input layer, a pre-synaptic layer and an output layer.

After encoding input data, a repository of trained output neurons is created for every class, and connections to all pre-synaptic neurons are established through the computation of a vector of weights that depends on the order of spike transmission, as defined by [215]. Each neuron i has a firing threshold $\vartheta^{(i)}$ that is obtained through a model parameter c and the maximal potential of the neuron, defined by its weights, the order of spike transmission and a modulation parameter m. In this way, a reservoir of trained output neurons is generated during supervised learning. The total weight value of each trained neuron is then compared with the weight value of each stored neuron and the minimal Euclidean distance between them is calculated. If its value is less than a set similarity threshold s, the two neurons are considered "similar" and they are merged by averaging their weights and firing thresholds $\vartheta^{(i)}$; otherwise, if a new neuron is added (evolved) as a new output neuron of the SNN. When training is performed and output neurons are defined, classification is made by propagating a sample through the network; the assigned class is that of the output neuron with the shortest response time. The details of the operation of eSNN model are presented in [216]. Thus, adding new samples to the trained model only implies merging them with their corresponding neuron repository.

With the operation scheme presented in Figure 4.1, clusters are obtained from traffic data in $\boldsymbol{\mathcal{H}}$ and the proxy dataset $\tilde{\boldsymbol{\mathcal{H}}}$ is built. The eSNN is then initially trained with this proxy dataset based on historical observations. In order to obtain the optimal values for the main eSNN parameters, i.e., the similarity threshold s, parameters c and m, and the number

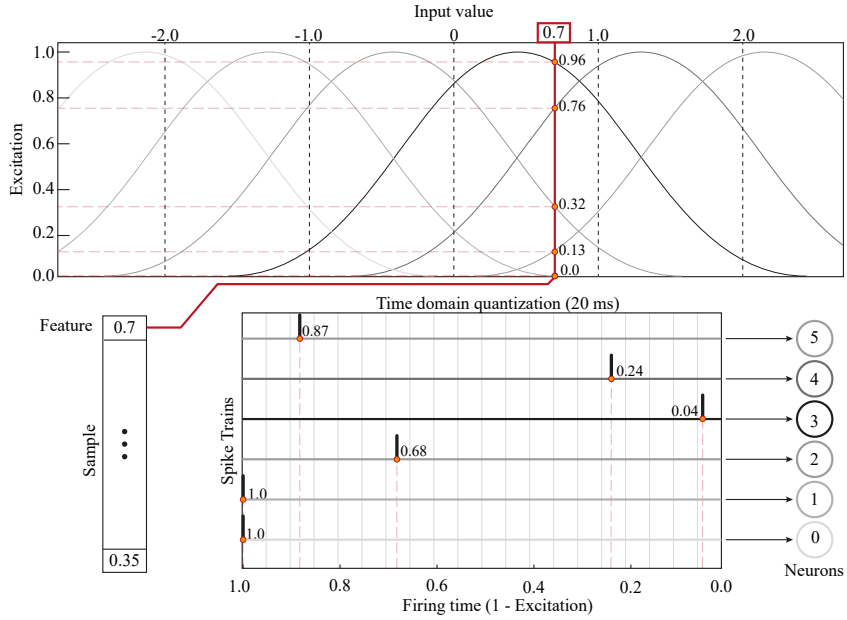

FIGURE 4.3: Example of how a random continuous value is encoded into spike trains for a number of input neurons.Value 0.7 activates six overlapping receptive fields, which excite six corresponding input neurons at different firing times.

of Gaussian receptive fields G, the classifier is run through a parameter-tuning optimization procedure based on a genetic algorithm to efficiently search over the parametric configuration space for each location under study. The trained network will then be used online for classification, and evolved daily, by virtue of the incremental evolution capability of output neurons [216]. New knowledge can be easily introduced to the network by encoding it and processing it to the neuron repository, with the above described mechanism.

4.2.3 Online Processing: Adaptation to Change

The main contribution of this work is the online phase depicted in the right part of Figure 4.1. Predictions obtained in the offline phase, with the discovery of patterns and assignment of new days to them are assumed to be representative of the diversity of traffic profiles in the data of the loop under study to the moment a prediction is queried. After the pattern clustering and classification procedure is executed, most future days will be classified accurately and the assigned traffic profile will match the actual one within a certain tolerance [34]. However, there are circumstances for which this condition is not satisfied. The most immediate obstacle for it to be met is the clustering and classification accuracy. In the first place, it can be expected that the clustering process does not infer all possible patterns that occur in the data at hand. In addition, even under the assumption that it is possible to adjust the clustering process to yield a

set of clusters enough to describe all the existing kinds of days in the available training dataset, it is likely that the classification model fails to classify some of the days. Other circumstances are less immediate, but are encompassed in these two. For instance, an accident will probably change completely the traffic profile of a certain location, but if the clustering process contemplates a cluster for the pattern of that kind of day including this kind of event, the method will have an option to adapt and change to that pattern. In the same way, if there is a drift in the long-term traffic, and if the clustering and classification models are updated accordingly they will be able to effectively cope with the new traffic profiles.

The theoretical framework proposed in this work deals with the update of the clustering and classification models without considering any incident or event that alters traffic. Detection of change and adaptation are hence oriented to detect and correct misclassification problems and to update the clustering model with information that it has not received before. Including external modifying factors like traffic accidents or weather conditions would extend the model capabilities, but would not entail an essential change in its concept. It would require to increase the number of clusters and to add new features to the proxy dataset with data paired to traffic observations. These data are not currently available, hence the conceptual model is developed just for the calendar features, which pose themselves a challenge.

Dealing with changes in the traffic profile in an online fashion has two facets: the detection and adaptation to change of the ongoing traffic observations stream, and the update of classification and clustering algorithms. Both parts are described in following subsections.

4.2.3.1 Detection and Adaptation to Change

The first part of the online phase consists of detecting if the received traffic observations are excessively deviating from the predicted baseline. In that case, the model must adapt the predicted traffic profile by finding a better baseline from the available ones, being the baselines the long-term predictions obtained in the offline phase. The whole detection and adaptation process is graphically summarized in Figure 4.4.

Anomaly detection is a wide field of study applicable to a variety of contexts, specially to those where a network of sensors gathers data [217]. The techniques employed to detect an anomaly in a sequence of sensor readings can be easily extrapolated to change detection, thus their common grounds have been applied in our model. The change detection mechanism is based in the definition of a threshold φ and a set $\mathcal{W}$ of warnings which has a maximum size of W_{max} warnings. Whenever the difference between the observed actual value ($o_t \in \mathcal{P}$) and the predicted one ($o_{t_d}^{\odot c}$) exceeds φ, a warning is raised and stored in $\mathcal{W}$. When W_{max} consecutive warnings have been raised, the sequence is interpreted as a change in the data. As this happens, the adaptation mechanism is triggered. The process is described in Algorithm 1.

The key aspect of this process leans on the definition of φ and W_{max}. Decreasing the threshold or the warning window size would result in more

Algorithm 1: Change detection mechanism

Input: Prediction $o^{\odot_c}$ array for the current day, individual observation $o_t \in \mathcal{P}$, array of previous warnings $\mathcal{W}$, maximum size W_{max}, threshold φ
Output: Change detected, updated $\overline{W}$

```
if abs(o_t − o_{t_d}^{⊙_c}) > φ then
    W = W ∪ o_t else
        W = ∅
    end
end
if |W| = W_max then
    return TRUE, W
    else
        return FALSE, W
    end
end
```

change alerts, and an increase of false positives, *i.e.* days that are correctly classified and the assigned centroid represents adequately the traffic of that day, but slight changes due to noise provoke a change detection and thus and adaptation (search for another traffic profile). On the other extreme, increasing the threshold and window size makes the detector less sensitive, and days that should be reassigned could be left out of the adaptation process. This becomes even more convoluted as the prediction baseline $o^{\odot_c}$ is smoothed, for it is an average of all of the elements in the cluster, but the readings o_t are noisy, making the discrepancy more likely to occur. Intuitively, a change should be detected and corrected as soon as possible, in order to make the rest of the day predictions accurate, thus sensitivity should be boosted. Nonetheless, some parts of the day are more sensitive *per se*. This is a result of the considerable differences between traffic variability during different parts of a day: for instance, between 2 a.m. and 5 a.m., traffic profiles tend to be barely variable, and a even a low threshold could trigger change alerts. However, that same threshold would be useless during rush hours, where great variations happen from day to day. If φ or W_{max} are too low, many false alarms might be risen every day during the first hours.

For the reasons exposed above, both φ and W_{max} should be well defined to maximize the detection of true alerts while minimizing the false positives. Either φ or W_{max} should be adaptive to mitigate the effects of the variability of traffic during the day. As their adaptation results in similar outputs (increase of any of them reduces sensitivity and their decrease enhances it), for this research one of them (φ) has been fixed, while W_{max} has been tailored for an optimal change detection performance. The value of φ is obtained for each 5 minute period of the day of a certain cluster and it is the standard deviation of all the measurements taken at that certain slot p within the days of that cluster, as per (4.3), resulting in 288

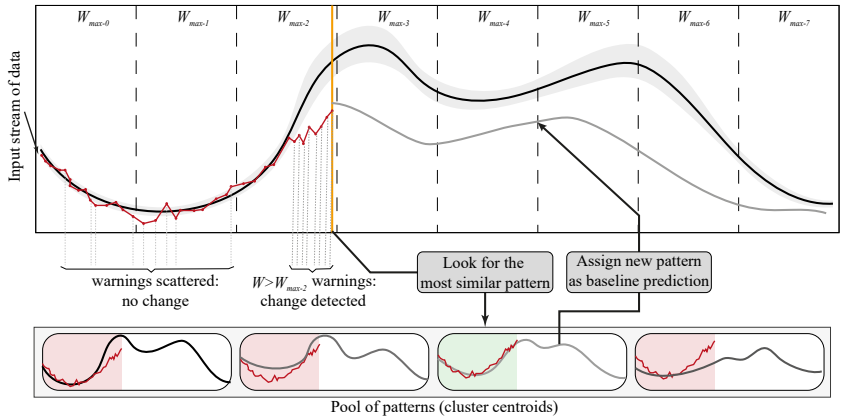

FIGURE 4.4: Mechanism for change detection and adaptation. When a set of warnings exceeds the maximum size of the warning window it raises a detection alert and finds the closest pattern.

thresholds φ_p^c per cluster. Such a standard deviation can be estimated as:

$$\varphi_p^c = \sqrt{\frac{1}{N-1} \sum_{\boldsymbol{o}^d \in \mathcal{H}_c} \left[o_p^c - \left(\frac{1}{N} \sum_{\boldsymbol{o}^d \in \mathcal{H}_c} o_p^d \right) \right]^2}, \tag{4.3}$$

where o_p^c are all observations of a certain 5 minute slot p of cluster c.

This definition allows for greater variations during peak hours, when the volume of vehicles can be very different even for the same type of days, and lower tolerances in periods with more flat traffic profiles. The adaptive window size W_{max} permits to compensate (if needed) for this sensitivity variation. This adaptive nature is furnished by means of 8 values of W_{max}, one for each 3-hours segment of the day. Thus, the number of warnings needed to raise an alert is different throughout the day, and can be forced to be less sensitive during some periods. Optimal sizes for each period are obtained after a grid search procedure is performed over a set of validation days from $\boldsymbol{\mathcal{H}}$ that aims at minimizing false positives and maximizing true alerts.

Adaptation is performed whenever Algorithm 1 returns a `TRUE` value. This process consists of seeking a better candidate for the current traffic profile. To this end, the sequence of available observations for the current day is compared to all the cluster centroids $\{o_p^{\odot c}\}_{c=1}^{c}$ in terms of Euclidean distance. The closest cluster is assigned as the prediction for the rest of the day. If the new cluster is found to be the same as the previous one, the day can be a member of the cluster, but it is possible that is closer to other members than to the centroid itself. Thus, the distance to each of the members is computed and a distance-weighted average is assigned as traffic profile for that particular day as a baseline prediction. This intracluster averaging procedure smooths slightly the assigned profile, as assigning the values of a particular member (the closest) would transfer

all the noise embedded in that member. However, this smoothing procedure is not enough, as more weighted days introduce more noise; for this reason, an additional moving average smoothing is performed in these cases. This process is repeated for each observation of the day in course, so several changes can occur during the day, and if a new assignation is erroneous it can be corrected later, although it is penalized in the prediction performance.

4.2.3.2 Clustering and eSNN Update

A key feature of the online detection and adaptation mechanism is the update of clustering and prediction models when new knowledge is found in incoming days. While the detection and adaptation can be processed completely online as new observations are received, the update of the classifier requires that a whole day is processed, for its samples are daily-based. Accordingly, between 23:55 of one day and 0:00 of the next day, the eSNN classifier is updated with new knowledge acquired during the last day. As defined in Section 4.2.2.3 the structure of eSNN allows for a quick update that can operate immediately after the last observation is received, and the classification for the new day can be obtained before the first observation of the next day. The following scenarios are considered:

1. No change detected, implying that the consecutive warnings condition has not been met for that particular day. Hypothetically, this could lead to interpreting that the day has been correctly classified, although it might have risen abundant non-consecutive warnings, suggesting a day with a noisy traffic profile. Anyhow, this result suggests that a fair amount of observations distributed along all the day have fallen within the area defined by the pattern thresholds. The classifier is updated by adding a new instance with the original class and by evolving the eSNN model incrementally.
2. Change detected, implying that one or more sequences of consecutive warnings has risen at least one alert. The alert is followed by an adaptation which requires a change of assigned pattern, which ultimately can result in a change of cluster belonging. Besides, more than one alert and cluster change can happen for each single day. This results in a final predicted set of values that can have been obtained from different clusters. As the time passes for each change detection, more observations are available in order to compare them to a sub-sequence of other patterns. This means that the later the detection is produced, the similarity found with other pattern is more significant, for it is based on more values. For the online detection and adaptation phase this situation is inevitable, and the early adaptations will work hypothetically worse than the later ones. Nonetheless, the classifier adaptation phase is performed in the end of the day, so all the real observations can be used to estimate the most proper cluster. When a change is detected, the whole set of actual observations is compared with the available cluster centroids using the same Euclidean distance that is used during the

clustering phase. The label of the closest cluster is used as the category for this day, and the eSNN classifier is updated accordingly.

An update mechanism is also devised for the clustering process. Once data of a complete day are available, the classification, detection and adaptation system can define a proper class for that day by using the criterion explained before. This class also identifies the cluster to which the day belongs. The clustering updating mechanism consists of aggregating the 288 actual observations to this selected cluster for which their membership is the most adequate. The aggregation encompasses adding the day as a new instance to the pool of selected cluster members, and recomputing the cluster centroid considering also the newly added member. Thus, an update of the cluster members and centroids is performed each day, at the same time that the eSNN is evolved. This allows not only the clusters and classifier to be updated, but also to grow and eventually promote single member clusters that derive from the noise cluster.

4.3 Results and Discussion

Methods described in previous Section have been tested with traffic profiles registered by sensors deployed in 6 different locations over the Madrid city network, all placed in urban areas with different traffic profiles. Four of them (A, B, C, D) are located in main roads while the other two (E, F) are installed in side residential streets with lower amounts of traffic. Particularly, location E is a street with no points of interest and almost no buildings, holding exceptionally low levels of traffic, with long periods of 0 vehicles passing by. For the sake of space, the location-based plots contained in this Section represent data of only of the sensor in location A, although the deeper analysis is presented and discussed for all of them. This sections presents in the first place the outcomes of the clustering process and the performance of the predictive model without any adaptation mechanism. In the second place, the results after applying our proposed adaptation method are analyzed.

4.3.1 Offline Prediction Analysis

The initial clustering stage is crucial for a proper operation of the whole proposed method. It provides the classes for the proxy model that will be used in the classification stage, but it also favors a deeper understanding of the traffic behavior in each site, and to examine visually the types of days. Figure 4.5 shows the clustering results for the traffic of loop A. Analogously, the optimization process that leads to these 9 clusters for location A has yielded different values for ϵ, the distance metric among the elements in the cluster, which leads to different groupings for the other placements (Table 4.1). The minimum number of elements required for each cluster is set to 2 for each of them, allowing for small clusters to form. The number of clusters is similar for most of them, although differences in the number of days in the noise cluster are more noticeable, and symptomatic of the

large differences in variability among the loops. Reducing the amount of noise implies making the distance metric more flexible to include those days. Thus, more days are clustered together and the number of clusters is reduced, being able to represent less particular circumstances and entailing greater errors in subsequent phases.

TABLE 4.1: Number of clusters and noise elements after performing DBSCAN clustering with optimized parameters at each location.

Location	A	B	C	D	E	F
ϵ	25	20	19	34	3	22
Number of clusters	9	8	8	7	8	7
Days in Noise Cluster	6	5	4	1	16	7

In the particular case of location A, most days in $\boldsymbol{\mathcal{H}}$ belong to cluster 0, which represents regular working days without Fridays. These have a different behavior in this particular location, and most of them are assembled in their own cluster. The same happens with Saturdays and Sundays. Beyond these 4 *basic* clusters, other special cases can be found with a very distinct traffic profile such as the last 4 days belonging to Easter, and the Christmas week. Days highlighted in black correspond to noise, *i.e.* days that do not fit in any of the existing clusters but are not able to conform a cluster by themselves, according to the density metric and minimum elements per cluster constraints. Days like New Year, with a highly active night and low traffic during day, and Christmas are in this cluster, but also the day of Epiphany, an important festivity in Spain, and the day previous to Easter. The same analysis in location E reveals that most of 16 noise days belong to Christmas and Easter periods. This anticipates bad results in subsequent stages, as only a few holidays or special days are characterized in the clusters, and the test set contains at least 10 of these days.

While a noise cluster exists (with all the noise days in it), a noise class can be assigned to some of the samples in the training set, and consequently, when classifying the test set, an instance could be classified as noise. Grouping all the noisy instances and giving them a class entity can bring classification problems, but more importantly, adaptation problems. The variability inside the noise cluster would be too high, and thus, the low amount of alarms based on the cluster standard deviations would lead those days to stay in the noise cluster. Hence, days reported as noise are regarded as clusters with one element. This allows them to be represented in the classification model and considered for the change and adaptation mechanism. Each noise day is expressed as one sample with different class in the training dataset, so their impact in the classification is expected to be reduced; however, when a similar cluster is sought in the adaptation phase, they compete as new cluster candidates in the same way than the other clusters. Moreover, as the clusters are updated every day, new daily patterns can be included in these single-member clusters, allowing them to grow and be more representative also in the classification phase.

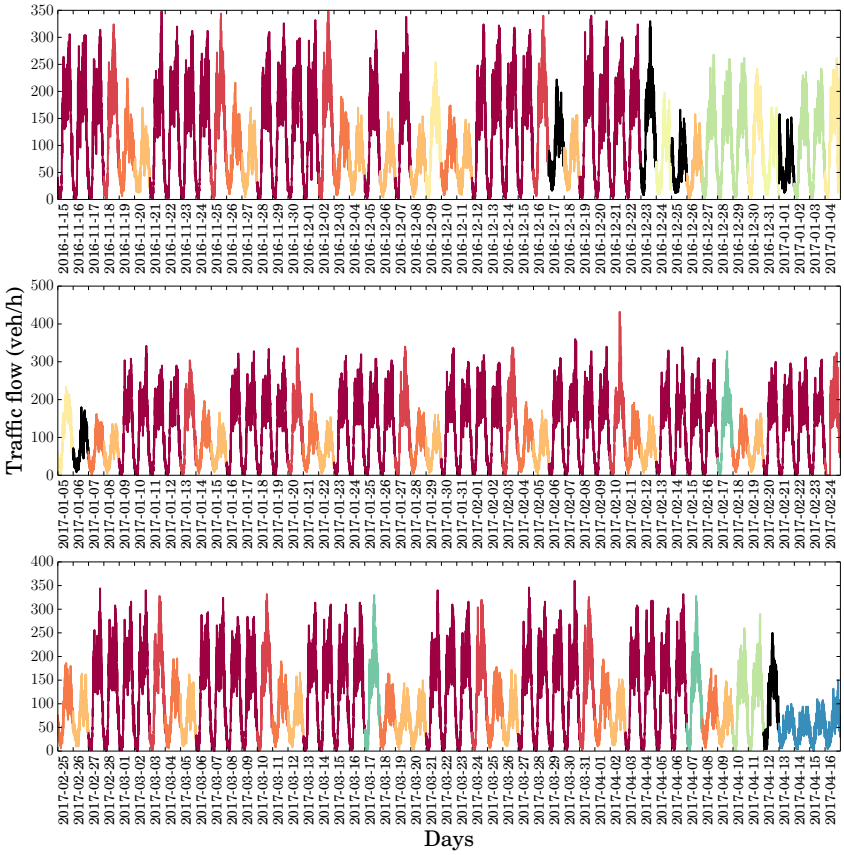

FIGURE 4.5: $|\mathcal{H}| = 153$ days of traffic in location A after clustering. Days are colored by their cluster membership. For visualizing purposes, observations are divided into three consecutive blocks of 51 days, shown in a row-wise manner.

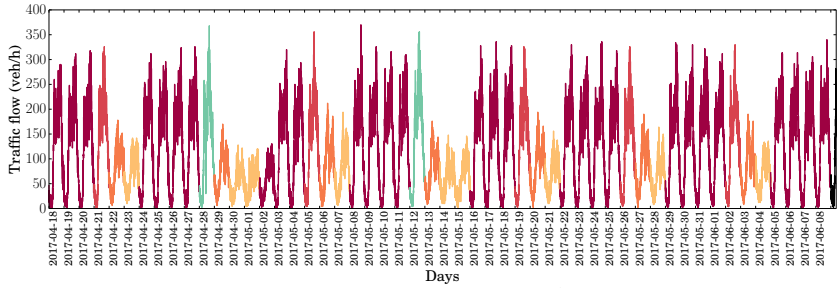

FIGURE 4.6: $|\mathcal{P}| = 52$ test days of traffic in location A after classification process.

In order to have a reference for comparison after change and adaptation mechanism is tested, an initial offline classification has been performed for all the 52 test days. The coefficient of determination R^2, which shows the likelihood of real values to fall within the predicted ones, and the Normalized Root Mean Squared Error (NRMSE) are presented for each day. R^2 metric is analogous to the one presented in Chapter 3 (Expression

3.8), although some of its notation has been adapted to the variables used in this Chapter, as follows:

$$R^2 \doteq 1.0 - \frac{\sum\limits_{\forall t_d} (o_{t_d} - \hat{o}_t)}{\sum\limits_{\forall t_d} (o_{t_d} - \bar{o}_t)}. \tag{4.4}$$

The same adaptation process is performed on RMSE metric, which provides the same insights that the one used in Chapter 3 (Expression 3.7). Besides the change in notation, in this case RMSE is normalized with respect to the average of vehicles passing each day, otherwise each RMSE measurement would have different meanings, depending on the day. Normalized RMSE (NRMSE) is defined as

$$\text{NRMSE} \doteq \frac{\sqrt{\frac{1}{N}\sum\limits_{\forall t_d} (o_{t_d} - \hat{o}_t)^2}}{\overline{o_{t_d}}}, \tag{4.5}$$

where N denotes the number of actual values of each day, $\overline{o_{t_d}}$ stands for the average of real observations for day d, and $\hat{o}_t$ is the predicted value for o_t.

Figure 4.6 shows the results for location A, and in Figure 4.7, the R^2 values for the rest of locations. These results are obtained training the model with half of the year, while the test days are part of the other half; some of the days in the test set can correspond to situations that have never been observed by the model, and thus have traffic profiles unknown to the model. Even considering this, forecasts obtained in A obtain a R^2 value superior to 0.9 for 33 of the 52 days, and as it can be observed in each individual graph, for these cases the actual observations are fit to the prediction lines (their assigned cluster centroids). This means that a basic approach that models patterns without adaptation is useful for more than half of the days, and the error produced in each day will be imputable to slight unpredictable variations.

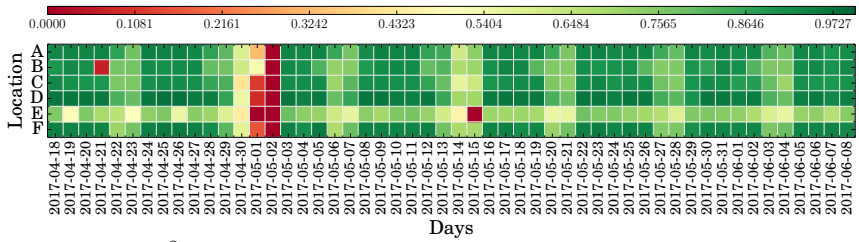

FIGURE 4.7: R^2 values for each test day of each location. Any value lower than 0 has been assigned value 0 in order to provide a representative color map.

As opposed to this initial good classification of what could be deemed *normal* days, there are some other days for which the prediction fails, in some cases they could have been assigned to a better cluster. This is the case, for instance, of May 2^{nd}, while others like June 3^{rd} and 4^{th} are apparently in the correct cluster there are greater variations that make the

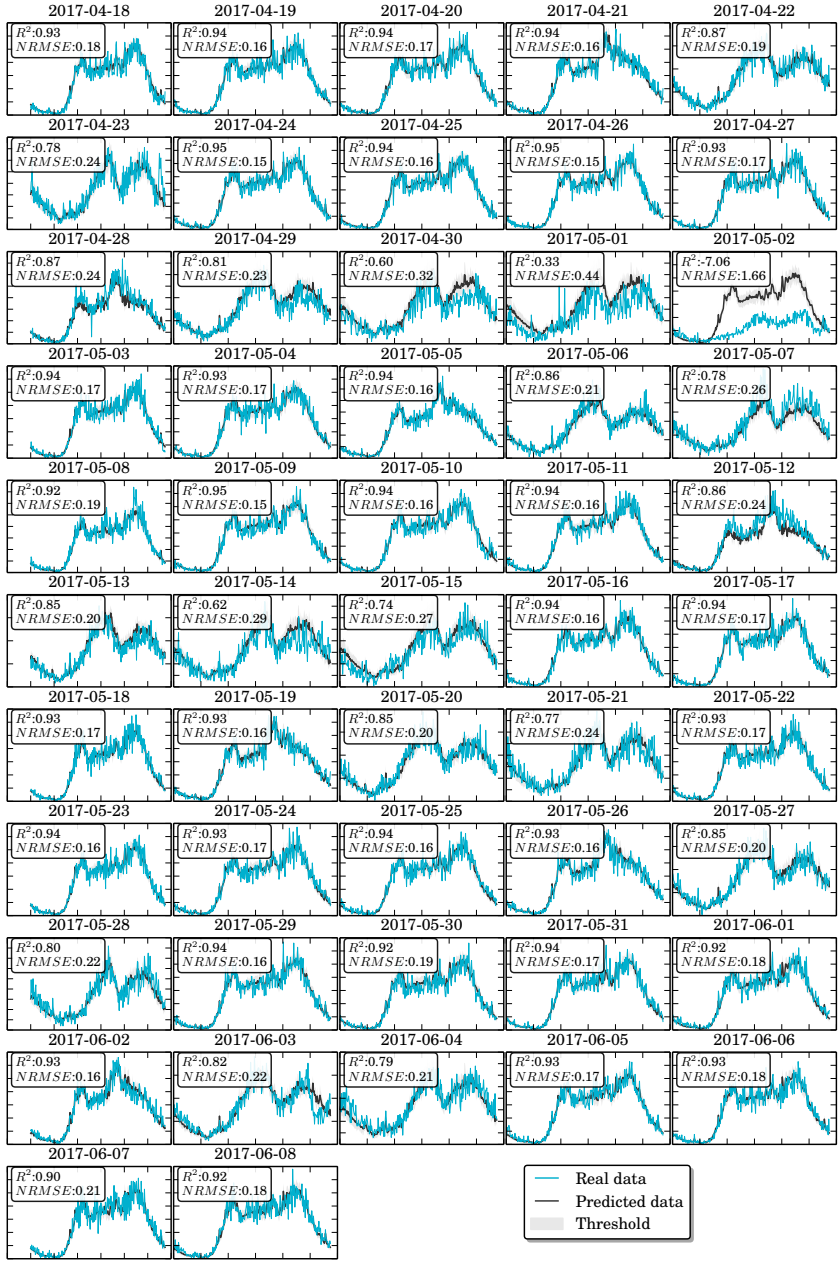

FIGURE 4.8: Test days, with real values, data predicted by eSNN model, and R^2 and NRMSE values.

predictions less useful. In the first case, a local holiday that has no similar traffic profile in the training set, and in the second, the celebrations of the winning of a sports championship affected a large area in the city center. By observing the rest of locations in Figure 4.7 an analogous conclusion

holds: the classification model is completely unable to properly assign patterns to the long weekend in the beginning of May, and the performance is also poor for other days in the different locations. For instance, in location B, Fridays are consistently predicted worse than in other locations. In general, predictions for weekends are less effective; although they are correctly classified, the information available for the cluster of Sundays is less accurate, there are less previous Sundays than weekdays to learn from, and it is likely that the activities that take place on weekends are very different in the test months (May, June), than in the train months (mainly winter). A model trained with longer periods could have yielded different clusters attending to the season or other circumstances. It is also observable that for location E, performances are lower, mainly due to the low dynamic range of its registered traffic data: for most days, volume readings oscillate between 0 and 25 vehicles with long series of real values equal to 0. Also, the mean traffic volume in this location is of about 11 vehicles and its standard deviation is close to 10. This produces highly noisy instances, reducing the possibilities of grouping days in meaningful clusters, which are more similar among themselves. Besides, in this scale, light changes imply broader errors, and a minor traffic increase of 5 more vehicles at some point, can represent a 20 or 25% relative increase of the expected measurement. Anyhow, any of the previously described shortcomings are useful to assess the performance of the adaptation mechanism presented in this work, whose results are discussed next.

4.3.2 Online Processing Results

Once an offline clustering and classification iteration is completed as a comparison reference, the proposed change detection and adaptation mechanism, along with the clustering update system are activated to assess their performance. An averaged result per location is provided in Table 4.2 for both considered metrics.

TABLE 4.2: R^2 and NRMSE measurements obtained for each location, averaged for the 52 days in the sample, and Wilcoxon p-values for each pair of sample sets.

Location		***A***	***B***	***C***	***D***	***E***	***F***
R^2	No adaptation	0.722	0.782	0.756	0.781	0.582	0.729
	Adaptation	0.878	0.840	0.849	0.901	0.654	0.845
	Wilcoxon p-value	0.001	0.009	0.001	$8.8e-4$	0.005	$6.5e-6$
NRMSE	No adaptation	0.229	0.277	0.255	0.216	0.476	0.250
	Adaptation	0.195	0.257	0.233	0.194	0.454	0.240
	Wilcoxon p-value	0.001	0.009	0.001	$7.2e-4$	0.005	$3.8e-6$

Reduction of error is visible from any of the error perspectives: in general, in every location, the introduction of a change detection and adaptation mechanism has improved the non-adaptive pattern prediction counterpart. The statistical significance of these results has been tested via a Wilcoxon test, comparing the set of results obtained for each day without change detection and adaptation, and those with this mechanism active.

The null hypothesis on this test states that there is no statistically meaningful difference between the two measurements, Therefore, these results certify strong evidence against this hypothesis, and the improvements obtained by our proposed method can be declared as statistically significant. However, these averages conceal some relevant aspects of the overall improved outcomes; it has been shown in Figure 4.7 that a great fraction of test days is properly classified and predicted, with some days for which a bad classification produces a large prediction error. Hence, a better classification for these particular days might make a remarkable difference in averaged results. In order to examine this postulated hypothesis, we inspect particularized results in terms of coefficient of determination for each day and location in Figure 4.9.

It can be observed in this plot that although there are still days with low quality forecasts, and most predicted days remain similar, a general improvement has been achieved. Days for which pattern predictions were ineffective in the non-adaptive setting get better approximations after adaptations are made. For instance, May 1^{st} and 2^{nd} are clearly improved for all locations, and very poorly predicted days like April 21^{st} in location B or May 15^{th} in location E have been also adapted for the better. On the other hand, there are days with slightly worse predictions, due to diverse factors, that can be noted for location A in Figure 4.10, where a daily detail of the previous results is shown.

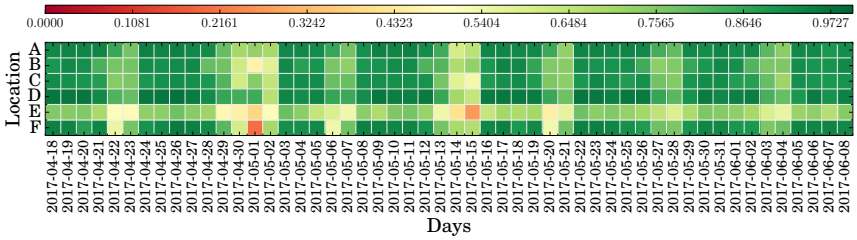

FIGURE 4.9: R^2 values for each test day of each location after adaptation. Any value lower than 0 has been assigned value 0 in order to provide a representative color map.

The change detection points are signalized with vertical lines, while black line represents originally assigned pattern, and the red line shows the final set of predictions that have been assigned to a day, after the initial assignment and subsequent changes. It is relevant to note that originally assigned patterns can be different that the ones presented in Figure 4.8, as an adaptation of the classifier and the cluster centroids is performed every day in this online version. This is noticeable for example in April 29^{th}, which was assigned offline to cluster 2, the cluster corresponding to Saturdays, and in the online version it is initially assigned to cluster 8, although then, early in the day is reassigned to cluster 2. Updates in the classifier have unchained this bad classification and the adaptation mechanism has corrected it, resulting in a day predicted with a little less accuracy that with the online version (NRMSE of 0.24 versus the previous 0.23); this situation is also found in May 19^{th} and 23^{rd}, with no prediction

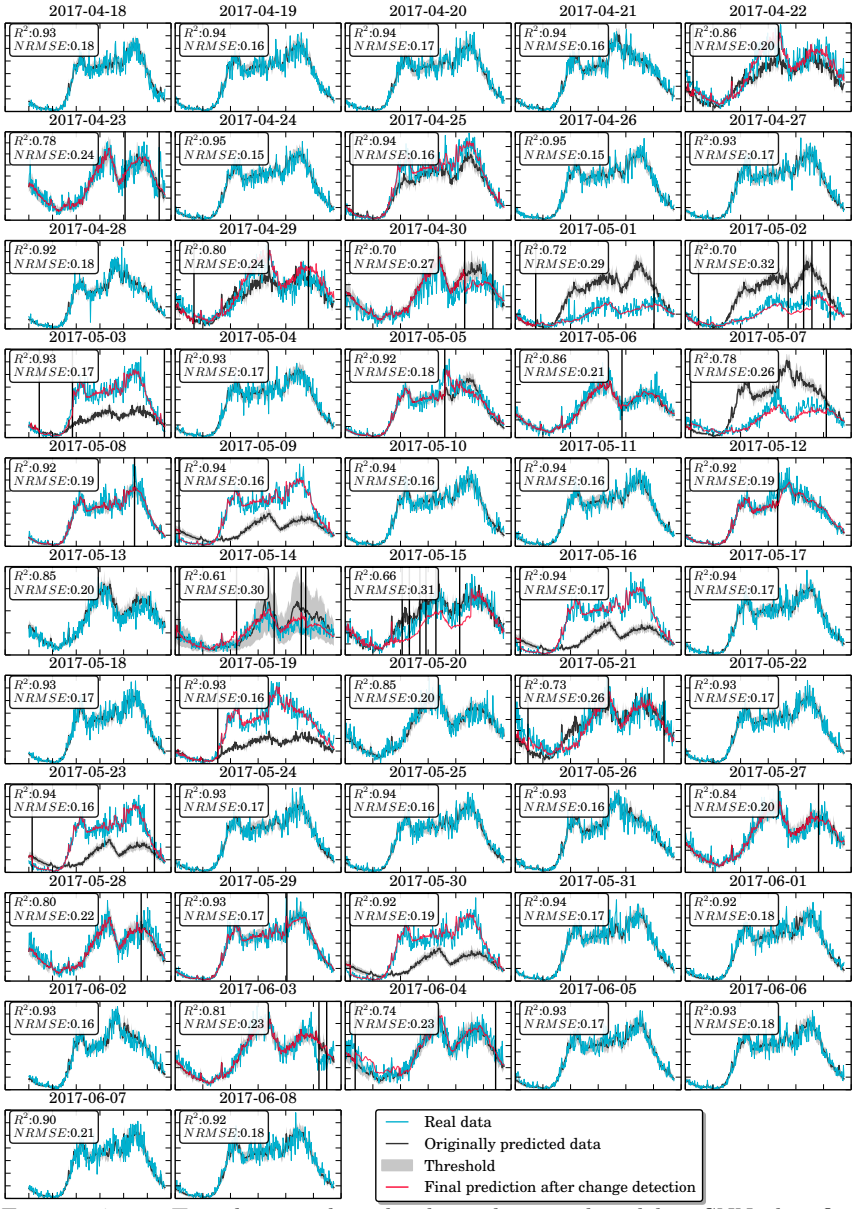

FIGURE 4.10: Test days, with real values, data predicted by eSNN classifier and new predictions after detecting changes and online updating classifier and clusters for location *A*. Vertical lines represent pattern change detection points.

accuracy loss, as a result of a good adaptation. A similar case is produced when the same pattern is originally assigned in both offline and online methods, but the online one looks for a better candidate within the same cluster for days like May 29^{th} or June 3^{rd} and 4^{th}. In all of them, the prediction performance is slightly reduced and keeping the original pattern would result in a better forecast. On its counterpart, the same scenario in

May 12^{th} results in a considerable improvement; this is more obvious for days like May 1^{st} and 2^{nd}, for which the original prediction was utterly useless, and the updated classifier and adaptation mechanism have jointly provided a more accurate forecast.

Besides the particular results for location A, Table 4.3 displays the number of days for which the prediction has been significantly improved or degraded (more than 5% gain or loss with respect to the original error) for each location, as well as the average accuracy gain or loss that those have entailed. In general, for all locations more than 60% of days have remained without significant change. For some locations there are more days with bad predictions that those daily predictions with good ones, for the reasons described above.

TABLE 4.3: Number of days and amount of NRMSE that have been improved and deteriorated after applying the change detection and adaptation mechanism and their average gain and loss amounts.

Location		***A***	***B***	***C***	***D***	***E***	***F***
Improvement > 5%	Days	8	10	7	8	10	5
	Avg. gain	0.207	0.113	0.202	0.212	0.168	0.271
Degradation > 5%	Days	4	5	9	11	8	14
	Avg. loss	0.024	0.032	0.024	0.043	0.054	0.053
No significant change	Days	40	37	36	33	34	33

This greater amount of days with worse forecasting results, however, does not involve a greater error, as for most of these days, the error gap is negligible. For instance, in location F, 14 days have obtained worse forecasting results, but they only account for an NRMSE increase of 0.053 points, while the improvements introduced in just 5 days have reduced the error on an average of 0.271 points. These averages are detailed in Figure 4.11, where the accuracy gain (error reduction) and loss (error increase) for each location are shown as violin plots. The error gaps for most days are very close to 0, whereas few days for each location are accountable for most the accuracy gain presented in Table 4.2.

All of the presented results and assessments of the model heretofore point at the same direction: an online adaptation mechanism leads to a model that can obtain better predictions even for days that have not been observed previously, but at the cost of a negligible penalty on prediction accuracy for some other days that would have been predicted better otherwise. With only 4 months of data to train, predictions are obtained for another 2 months, never observed by the model. Most test days (more than 70% for all locations and more than 80% for A and D) are predicted with great accuracy ($R^2 > 0.8$) in the long-term, without any adaptation. Location E is excluded from this analysis, since its noisy data have lead to poor results. However, it should be remarked that the characteristics of this location (a very low-valued traffic profile, no points of interest, almost no residences, and no connections with major arteries) make this case less conclusive for the study of the traffic of an urban area.

Most difficult days to predict in the used data are May 2^{nd} and 15^{th}, as they are local holidays exclusive to Madrid city and region, and they are

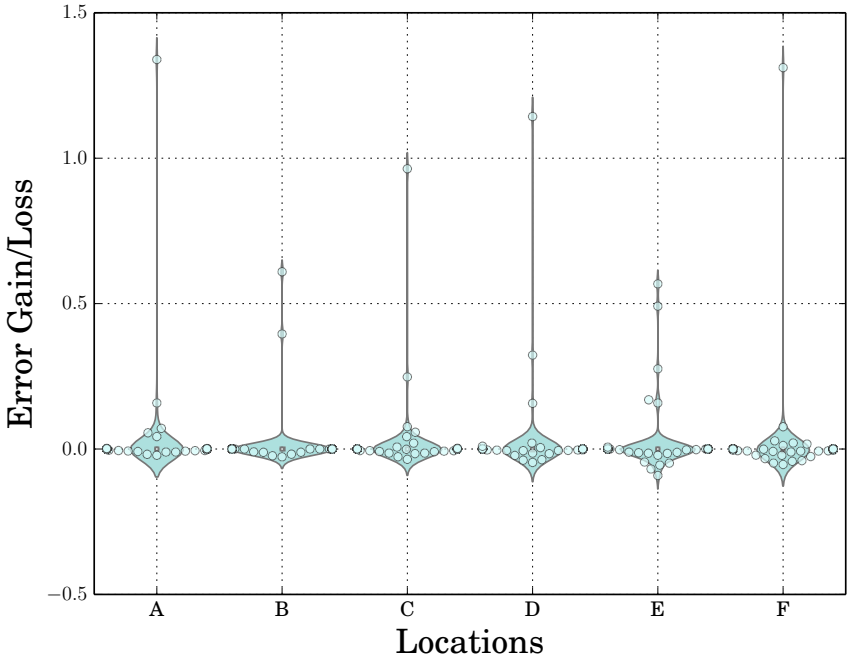

FIGURE 4.11: NRMSE error differences distribution for each location. Values above 0 represent days with less error, (and quantify their relative gain), while negative values represent days with more error and correspondingly their relative loss.

next to weekends and other holidays. Such circumstances have never been produced in the train dataset, nor have they been observed by the model. The clustering classification model fails to predict accurate patterns for these days, and they represent an example of how the outcomes are when unexpected conditions are playing. The adaptation mechanism enhances the accuracy of predictions and they can be used as a more approximate estimation of the real traffic. In all cases, even in the noisy-profiled location *E*, baseline predictions have benefited from the adaptation in global terms.

4.4 Conclusions

In this chapter a long-term urban traffic volume forecasting method with an adaptation mechanism is presented and tested with real data in different locations of the city of Madrid. The proposed method takes advantage of well-established traffic characterization models, which group periods of time (in this case days) with similar traffic, and use those patterns as predictions; such predictive models fail, however, when unexpected events occur, so an adaptive mechanism is designed to detect and adapt to those circumstances. The presented case study is intended to serve as a reference of the performance gains the long-term prediction and adaptation engine can achieve. In the absence of other sources of traffic related data, the

model has been built and assessed under the premise that pattern classification issues (to which the model adapts) would arise from calendar related conditions. Nonetheless, the model could and should be further enhanced upon the availability of more and more diverse data, such as events, traffic incidents, weather or planned road works.

In a real application case, it is likely that traffic data would be available throughout a whole year or more, and could enrich significantly the training data substrate for the predictive model. Almost all calendar situations could have been observed and trained previously, and hypothetically, better outcomes could be obtained in the offline phase. However, there are other considerations that could be added to the model to form more fine-grained clusters and classification, such as recurring events. In our case study there are two days clearly affected by a sports event celebration. The offline model provided a pattern that was mostly correct except for the night between the two days spanned by the event at hand. The proposed adaptation mechanism detected the anomaly, but was not able to provide another closer pattern. If in the clustering modeling phase this kind of event was contemplated, a cluster with these kinds of days would be available for the any of the forecasting phases, and the prediction could have been more accurate. Introducing events is an obvious next step in the development of this kind of urban traffic forecasting system. Sports events, demonstrations or parades can have a high impact even on spatially distant traffic, and they are foreseeable, so they can be used for modeling without relying on predictions. On the other hand, other factors like weather are proven to affect traffic notably, but if a model is built on past weather conditions, future weather conditions would be necessary to obtain forecasts with it; the quality of weather forecasts would have then a notorious impact on the model predictive performance. Nevertheless, with the proposed adaptive scheme, weather forecasts could be also valuable inputs, as they could be corrected during the adaptation phase.

Beyond including other features, the long-term operation of a system like the one described in this chapter would require other type of adaptations. The use of an eSNN allows the model to be constantly updated without retraining, and in combination with the clusters update, permits the model to operate indefinitely. However, in pursuance of a system that keeps learning, and adapts to long-term drift in data, the update of the whole clustering process is desirable. Unlike with the eSNN, introducing new knowledge to the clustering algorithm would imply running the whole process so as to forget the old data distribution, which could dramatically alter the cluster space, centroids and proxy dataset. The cluster space is stable to a certain extent due to traffic data seasonality, so the clustering should not be recomputed afresh on a daily basis, but rather once a year, once all types of days have occurred and they start to happen again. Thus, year after year, slight variations in traffic behavior would be captured by the representation of clusters, and the way in which the proxy dataset is built. If variations in traffic are a result of the alteration of the road profile (for instance, due to road works), the whole process described in Section 4.2.2.1 should be also performed again considering only observations taken

after the alteration. An automated high level adaptation mechanism that inspects the evolution of traffic whole days compared to known patterns could be implemented to detect these kind of variations.

Lastly, we consider that at the possibility of a random shift, the robustness of a model does not pass through the prediction of the random event, but through the ability to adapt to it. Lacking the proper data, we have tested our model behavior against the unexpected with unknown traffic profiles, which are what in essence would produce an unforeseen circumstance. Traffic incidents are stochastic events that produce severe traffic alterations, specially when they occur during certain time frames. Although they are rare in urban contexts, they tend to be recurrent in certain locations. The proposed forecasting system should be fit to deal with this kind of events; if it is possible to find traffic profiles of days with an incident and cluster them apart, the change and adaptation mechanism should be able to reassign the traffic profile to another one in which an incident has happened. Future developments of the presented model could potentially involve all the major aspects that affect traffic in order to have better and more automated forecasting systems.

Chapter 5

Environmental Insights from Road Traffic Forecasting Models

Once we have dealt with improved data-based models for traffic forecasting, the focus is now placed to the exploration of how this information can be exploited within a real application: the study of traffic-related pollution over a wide urban area. In this regard, this chapter concentrates again on the city of Madrid, the capital city of Spain, with 3.1 million inhabitants and a densely populated urban area (5225 inh/km^2) situated at an elevation of 667 meters over the sea level. As shown in Figure 5.1, the star-shaped design of the Spanish road network makes Madrid the central transport hub of the entire country. This fact, combined with the 4.2 million registered vehicles in the region, yields a heavy traffic supporting metropolis undergoing severe congestion issues through its road network. As a consequence of this, road traffic is widely acknowledged as the main source of air pollutants in Madrid [218]. In quantitative terms, NO_x and CO emissions are related to traffic in more than 80% in the city [219], 48% of PM_{10} mass was proven to be contributed by vehicle emissions [220], and 65% of tropospheric O_3 formation is on account of traffic-related precursors [221]. This close relationship between traffic and pollution comes along with severe health implications: indeed, worldwide epidemiological and toxicological studies have linked these traffic related pollutants to respiratory issues [222], [223], cardiovascular health effects [224] and lung cancer risk [225]. In 2013, the specialized cancer agency of the World Health Organization – the International Agency for Research on Cancer (IARC) – announced that outdoor air pollution has been officially classified as an carcinogenic agent for humans (Group 1) [226].

5.1 Related Work

Even though the number or vehicles has increased significantly over the last two decades [227], levels of NO, NO_2, CO and PM_{10} have featured a decreasing trend in Madrid [228] as a result of the pollution abatement policies promoted by the European Parliament (Directive 98/69/EC relating

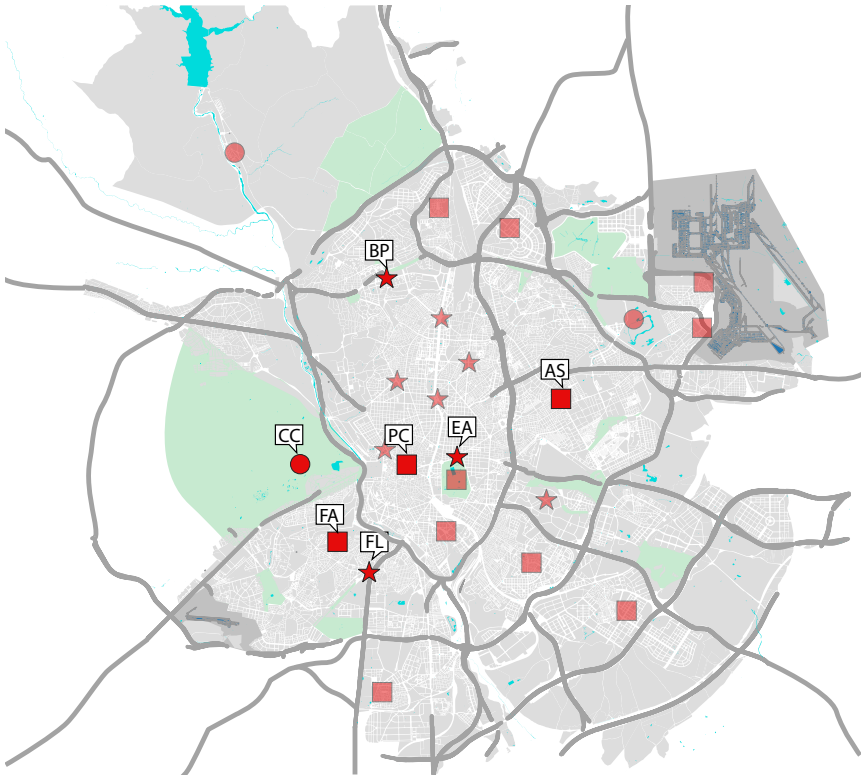

FIGURE 5.1: Radial distribution of the road network around Madrid, and location of the 24 urban air quality stations deployed over Madrid: urban background (□), roadside traffic (★) and suburban (○). Selected stations are tagged.

to measures to be taken against air pollution by emissions from motor vehicles[1]). The implementation of such regulatory laws and other subsequent sets of measures involves not only administrations, which are compelled to materialize control and management over traffic, home and industry pollutants, but also vehicle manufacturers, with more severe regulations for the exhaust emissions. Another relevant factor for this decreasing trend is the economic recession, which started in Spain in 2008 and has led to lower levels of fuel consumption [228]. On the other side, despite this NO_x reduction an upward trend is found in tropospheric O_3 concentration in the last decade [221]. O_3 is formed within a complex photochemical process that requires, among others, anthropogenic and natural sources of NO_x and volatile organic compounds, collectively referred to as ozone precursors, enhanced by favorable meteorological conditions (high temperatures and strong solar radiation).

Although vehicle emissions, industry and heating produce most of the atmospheric pollutants, the climatological characteristics of the region play an important role in how the pollutants are dispersed or conserved. Precipitation can help dissipating heavier pollutants, while wind can help

[1] http://eur-lex.europa.eu/legal-content/EN/NOT/?uri=CELEX:31998L0069

dispersing lighter ones [229]–[232]. The literature also evinces that the lack of wind and precipitation combined with high-pressure atmospheric conditions curb pollutant dispersion, and high UV radiation levels unchain the forenamed O_3 effects. A study made in Oslo (Norway) in 2004 [233] analyzed how distinct meteorological conditions impact on different pollutants, and elucidated that the number of vehicles is the most important factor in a city under such conditions. In a city like Madrid, with stable atmospheric conditions, the pollution should be strongly dominated by the prevailing meteorological conditions, setting traffic aside in a less relevant role, as pollutants accumulate until they are washed away by meteorological agents. Thence, two cities with similar anthropogenic emission levels may have acutely different pollution levels if they have antithetic meteorologic features. The dry and stable climate of Madrid, with less than 60 days of precipitation in 2015, results in a highly meteorology-dependent pollution. Furthermore, topography and urban street disposition and building types play also an important role in pollution concentration and dispersion. The term *street canyon* referred originally to narrow streets flanked by buildings. This definition has been updated and characteristics like the height of buildings of each side, the length of the street, the number of crossing streets and the number of openings in the walls configure different types of street canyons [234]. Street canyons may produce diverse effects in pollutant concentrations depending on the direction and speed of wind, creating vortexes of pollution when wind is perpendicular to the street and in-flow channels when wind runs parallel to the street.

In relation with the meteorological influence, 2015 has been the warmest year ever recorded in a global scale [235]. This fact, along with the incipient overcoming of the economic crisis, has implied high pollution issues over the city of Madrid during late 2015. Evidences abound: levels of NO_2 exceeded the 200 $\mu g/m^3$ limit up to 95 times during 2015 in El Pilar district. Considering all air quality stations in the city, the NO_2 level limit has been exceeded an average of 23 times during the year. Likewise, levels of tropospheric O_3 exceeded the 120 $\mu g/m^3$ limit an average of 10 times a year with a top of 68 excesses at one of the monitoring stations. On the contrary, PM_{10} and CO levels remained below the recommended limits [236]. This series of data evidences has motivated authorities to undertake traffic containment measures such as speed and parking limitations or public transport reinforcement [237], to the point of foreseeing stringent traffic restrictions in the inner city should the previous measures not lower down pollution to admissible levels. However, such countermeasures are not new, as similar action plans have been put to practice for years in other cities [238], [239]. Effectiveness of the implementation of these policies is not strongly evidenced [240]; although in some cities they do have an impact in pollutant levels [241], in other locations their relevance is milder. On one hand, research efforts have been invested on explaining the behavior of traffic emissions [242], [243] in order to understand how traffic pollutants are produced and how this knowledge should be exploited so as to diminish them. A report on this subject [244] showed that the factors influencing traffic emissions can range from drivers' aggressiveness

to the number of stops they make if the traffic is congested. The first kind of factors are out of reach for traffic management, but the latter can be tuned so as to reduce emissions. However, this tuning might cause a negative impact in the level of service of the road network, and therefore should be implemented with caution.

Pollution models can help traffic managers to take decisions efficiently, by selecting the most adequate traffic management strategy [245]. In literature, meteorological data are the main input for the models [246], [247], while some researchers use only traffic data [248], and a slighter proportion of researchers build their models with both traffic and meteorological data as inputs [249], [250]. This chapter will examine the relevance of road traffic variables and meteorological conditions in order to understand and predict the levels of pollutant agents in different kinds of locations of the city of Madrid, using historic traffic, pollution and meteorological data of 2015 as inputs. To this end, an ML methodology will be followed. In the previous literature a great part of prediction models designed for this purpose hinge on neural networks [251]. Variations in the neural network model and improvements in the pre-processing of input data are introduced by [252] to enhance its predictive capabilities. Other ML techniques such as decision trees [253] or SVM [254] have also been used to predict pollution from meteorological data [255], [256]. Linear regression has also been used to model PM_{10} concentrations [250]. This chapter joins these previous works from a new perspective: not only it explores the performance when predicting pollution using combinations of meteorological and traffic inputs and an ensemble supervised learning model, but also analyzes a quantitative measure of the importance of each variable as estimated during the training process of the model itself. Furthermore, the selected locations of air quality stations and traffic loops utilized in this study are characterized by different configurations in regards to their surrounding topography and urban street disposition. As discussed and concluded from the performed data analysis the impact of meteorological conditions do prevail on the pollution levels of this city, which might ultimately outgain any traffic-based countermeasure promoted by relevant authorities and stakeholders.

5.2 Materials and Methods

In this chapter data are provided by Open Portal of the Madrid City Council, not only for traffic, as in the previous ones, but also for pollution levels. The Meteorological State Agency of Spain (AEMET) provided the weather conditions data. Data correspond to the year 2015, from January to November, as December traffic data are undisclosed at the time of the development of this study. Meteorological conditions operate in large areas and traffic loops are available all over the city; therefore, the selection of the input data has been made by defining a set of targeted air quality stations and their closest traffic loops with enough diversity to represent different urban topographic characteristics and neighborhood types (e.g. downtown, residential).

5.2.1 Pollution Data

Madrid Air Quality System maintains 24 stations in the metropolitan area, which provide a variety of pollutant readings. Equipment to measure NO_x levels is available at all stations. There are two so-called "super-stations" which provide, besides NO_x, SO_2, CO, PM_{10}, $PM_{2.5}$, O_3, heavy metals and benzopyrenes readings, whereas for the rest the set of available measurements vary. According to the European and Spanish legislation there are three types of air quality stations: urban background (□), representative of urban population exposure to pollutants, in which the pollution levels should be distributed among different sources [257]; roadside traffic (★), representing mainly emissions originated in close roads; and suburban (◯), in the outskirts of the city, which record the highest levels of ozone. Figure 5.1 depicts the location and type of the air quality stations deployed in Madrid.

In this chapter, the focus is set on those pollutants most closely related to traffic [218], [258]. Several of those pollutant agents, such as sulfur dioxide (SO_2), are originated mainly in industrial processes. In contrast, other pollutants like nitrogen oxides (NO_x) are directly related to road traffic, and particulate matters ($PM_{2.5}$ and PM_{10}) are also appreciably contributed by vehicle emissions [218], [259]. Ozone (O_3) levels depend mainly on meteorological conditions, but changes in NO_x emissions strongly influence O_3 trends [260]. Therefore the O_3 concentration embodies a good indicator of traffic-related pollution, specially in locations where cloudy weather conditions are infrequent as it occurs in Madrid. Consequently, air quality stations have been selected for the study considering their capabilities to measure NO, NO_2, O_3, $PM_{2.5}$ and PM_{10}, their location over the city and their proximity to traffic measuring points. Regardless the latter condition, we have also included a suburban station (*Casa de Campo*) in the study in order to reinforce the analysis with a location relatively far from any local traffic source. The selected stations are depicted in Figure 5.1 and described as follows:

- *Escuelas Aguirre* (RS-EA): Roadside "super-station" with equipment capable of measuring CO, NO, NO_2, SO_2, O_3, benzene, hydrocarbons, $PM_{2.5}$ and PM_{10}. This station is located in the junction of three main roads in the center of Madrid with four or more lanes. Although it supports large amounts of traffic, it is placed next to *El Retiro*, a 350 acre park, which may mitigate the effects of surrounding traffic. The station is placed in an open area, flanked by the park and six-story buildings, not forming a typical street canyon.

- *Barrio del Pilar* (RS-BP): Roadside station located in a residential area in the north of the city. The station is placed in a park, next to a junction of two four-lane streets. The station is surrounded by a wide open area, but closer streets form street canyons flanked by thirteen-story buildings. The M30 ring highway, one of the roads with heavier traffic in Madrid, is located 230 meters north. Interestingly for the purpose of this study, measurements recorded by this station violated

the NO_2 limit levels 95 times in 2015. RS-BP provides readings of CO, NO, NO_2 and O_3.

- *Plaza de Fernández Ladreda* (RS-FL): Roadside station in a residential area, supplying measurements of CO, NO, NO_2, and O_3. It is located in a junction of 4 important streets and a highway, but surrounded by trees of a small park nearby. In front of its location there is a greater park (*Parque Emperatriz María de Austria*).

- *Plaza del Carmen* (UB-PC): Urban background station located less than 2 km west of the RS-EA station and close to *Gran Via*, an important artery with heavy bus traffic. Nonetheless, the station is placed in a pedestrian public square, surrounded by buildings that isolate it from the direct impact of *Gran Via* traffic. CO, NO, NO_2, and O_3 levels are provided by this station.

- *Arturo Soria* (UB-AS): Urban background station in a location similar to that of RS-BP: a residential working-class district with a main road crossing 50 meters south. However, this station is not so close to highways, and the surrounding buildings are lower (i.e. three-story), hence avoiding street canyons with their implications in pollution. It also has trees in both sidewalks and in the median strip. The station provides CO, NO, NO_2, and O_3 measurements.

- *Farolillo* (UB-FA): Urban background station in a residential working class district with no important roads in 1 km around. The station is placed in a small public square surrounded by short buildings that conform typical street canyons with low traffic impact. Besides CO, NO, NO_2, SO_2 and O_3, this station provides PM_{10} readings, which will be helpful in order to assess the impact of close traffic on this pollutant.

- *Casa de Campo* (SU-CC): Suburban "super-station" in a large park in western Madrid. The closest roads are located at 1.6 km, which limits the impact of local traffic emissions in the measurements re-corded in this site. Thus, models built for this station will be compared to those obtained for stations with closer local traffic sources.

It should be noted that the above stations supply the most complete sets of pollutants, with NO, NO_2, and O_3 covered in all of them, and PM_{10} in three of them. The rest of stations provide only part of these pollutant agents, with several of them missing, not being useful for comparison purposes. Readings of the selected pollutants are published by the stations in $\mu g/m^3$ with a hourly resolution.

5.2.2 Traffic Data

As in previous chapters, the traffic data have been obtained from the Madrid City council open data portal, detailed in Appendix A. In this chapter, aggregated readings of the traffic flow in intervals of 15 minutes have been used. To this end, a distinct subset of the ATRs deployed over

the city is chosen for each air quality station based on their proximity to each other. Although pollutant agents can be dispersed in greater areas and affected by diverse factors, this study is focused on the direct impact of close traffic; consequently, a 100-meter radius has been set around each station[2] so as to discriminate the subset of ATRs that characterizes the traffic in the surroundings of the station at hand. Differences in traffic volume are expected to have a slighter impact in pollution than the meteorological conditions, which is close to be the same for all areas.

5.2.3 Meteorological Data

The present study requires meteorological data with both good quality and high temporal resolution. Consequently, although it is possible to access daily resumes of many of the meteorological observatories hosted by AEMET[3], we decided to use the meteorological observations provided by the Aerodrome Meteorological Office of the Adolfo Suárez Madrid-Barajas International Airport. Specifically, we have used the METAR and SPECI reports. METAR is a coded report normally generated every half an hour throughout the worldwide network of operative airports. On the other hand, SPECI stands for the code of an aerodrome special meteorological report. SPECI briefings are generated when there is significant deterioration or improvement in airport meteorological conditions (METAR and SPECI reports[4]). Both reports are freely available on the Internet and provide interesting meteorological data. For this study, we have selected the following data:

- Precipitation, extracted from the group that informs about present weather phenomena observed at or near the aerodrome.
- Temperature and dew point, obtained from the group that informs about both variables (measured in degrees Celsius).
- Wind intensity, obtained from the wind group. For this study, we have converted this variable from knots to kilometers/hour.
- Cloud type and cover, obtained from the group used to report sky condition for atmospheric layers aloft. This group informs only about some cloud types (mainly Cumulus Congestus and Cumulonimbus), and gives interesting information regarding cloud cover, which is encoded into five categories: sky clear, few, scattered, broken and overcast. Not having UV radiation data makes these variables relevant for the regression of ozone levels.

Finally, regarding the quality of the meteorological data used in this work, it is interesting to note that all the meteorological observing systems that the Aerodrome Meteorological Office Staff utilize for preparing METAR and SPECI reports (e.g. thermometers, anemometers, and

[2]Except for UB-FA, which is 520 meters from the closest ATR.

[3]`http://www.aemet.es/en/datos_abiertos/catalogo`

[4]ICAO Annex 3 to the Convention Meteorological Service for International Air Navigation

ceilometers, among others) are managed under a Quality Management System certified to ISO 9001:2008. Besides the aforementioned fine-grained temporal data, the open data bank of the Madrid City Council provides monthly aggregated temperature in Celsius and precipitation in millimeters since 1988, as well as monthly aggregated maximum wind speeds in kilometers/hour and UV radiation levels in Joules/m2 since 2012, all captured from the Madrid-Retiro meteorological observatory (latitude: 40.41 N, longitude: 3.68 W, altitude: 667 m). These data are not used in the regression models, as they do not have enough resolution; instead, they provide the means to compare the climate in Madrid during 2015 to the averaged sequence of previous years.

5.2.4 Regression Model and Feature Importances

As anticipated in the introduction of this chapter we will resort to one of the most utilized ensemble methods for supervised learning problems RF, which have gained momentum in the last decade by virtue of their ability to handle multidimensional classification and regression problem with excellent accuracy and small chances to overfit [183]. In their naive definition RF consists of an ensemble of weak tree learners, each trained on a sampled subset of the available data, from where the predicted output is taken by aggregating and averaging the individual predictions of all such compounding trees. This particular construction method, which blends together the concepts of bagging and random feature selection, have been demonstrated to improve performance over other ML algorithms and linear regression models [261]. As a byproduct of this training procedure RF also provide an embedded method for quantifying the predictive importance that each of the predictor variables in the dataset possesses with regards to the target variable to be predicted. Specifically, the value of the feature importance reflects the mean decrease in accuracy when testing the model with out-of-bag observations [261], either for classification or regression problems. Specifically, the importance value of the j-th feature after training the RF model is computed by first permuting the values of this feature among the training data and next computing the average out-of-bag score difference between the original, unaltered dataset and that obtained after permuting the variable. Scores are normalized by the standard deviation of such differences in performance. This provides a numerical estimation $I_j \in [0, 1]$ that denotes the importance of such a variable during the training process of the model, i.e. $I_j \to 1$ if the k-th feature is predictively relevant for the variable to be predicted, and will approach 0 otherwise.

Since this study deals with regression results discussed in what follows will be interpreted by jointly analyzing cross-validated predictive performance scores – using the so-called coefficient of determination R^2 and the Mean Fractional Bias (MFB) metric – and variable importances obtained for each dataset. The R^2 score measures how fit the model is to predict future data instances, with its best score being 1.0 for error-free prediction, and its worst value equal to -1 corresponding to a model that performs

worse than a constant prediction equal to the expected value of the target variable. Mathematically speaking the R^2 score computed over N predicted samples $\widehat{\mathbf{y}} = \{\widehat{y}_n\}_{n=1}^N$ corresponding to the true instances $\mathbf{y} = \{y_n\}_{n=1}^N$ will be given as in previous chapters by

$$R^2(\mathbf{y}, \widehat{\mathbf{y}}) \doteq 1 - \frac{\sum_{n=1}^N (y_n - \widehat{y}_n)^2}{\sum_{n=1}^N (y_n - \bar{y})^2}, \tag{5.1}$$

where $\bar{y} \doteq \sum_{n=1}^N y_n/N$. The MFB score provides a symmetrical measure bounded in the range [-2,+2] that builds upon the concept of bias, which measures the tendency of a model to over- or under-predict [262]. A performance region is usually defined in the [-0.5,+0.5] interval that indicates a good performance of the model [263]. The best value of this score (MFB = 0) means that there is no bias between the predicted and the observed value. This score is computed over the predicted and observed values of the target variable as

$$\text{MFB} \doteq \frac{2}{N} \sum_{n=1}^{N} \frac{\widehat{y}_n - y_n}{\widehat{y}_n + y_n}. \tag{5.2}$$

An statistically meaningful estimation of the first-order statistics (mean, standard deviation) of the above scores will be computed using K-fold shuffled cross-validation over the datasets in question.

5.2.5 Preprocessing of the Datasets

The data sources chosen in this chapter deliver different temporal resolution data: hourly for pollution and meteorological data, and 15-minutely for traffic data. The first preprocessing step involves bringing resolution uniformity to the data, thence, 15-minute traffic slots were aggregated into one-hour slots, decreasing the noise produced by outliers, and maintaining the characteristics of original distribution. Each instance of a dataset is created with these 1-hour uniform data. Instances contain the hour of the day, the traffic flow value of each loop in the dataset for that hour, meteorological parameters during the same period and pollutant levels read by the monitoring station at hand. Besides, three additional columns accounting for the type of day, public holidays and month were added to the dataset, leaving the instances as shown in Figure 5.2.

A dataset is created for each monitoring station, each then split into four sub-datasets which are combinations of the three types of available input variables: all features (labeled with ① in the remainder of the chapter); only traffic and meteorology features (marked with ②); only traffic and temporal features (correspondingly, ③); and finally, only meteorology and temporal features (④). This split into subsets aims at assessing the impact of the lack of each of the variables in the prediction of 4 pollutants: CO, NO, NO_2 and O_3. This process results in 16 datasets per location. Also, 3 additional datasets are created to build models for predicting PM_{10}. Datasets are initially created with 8760 instances (one per hour through all the year). There are, though, instances with incomplete

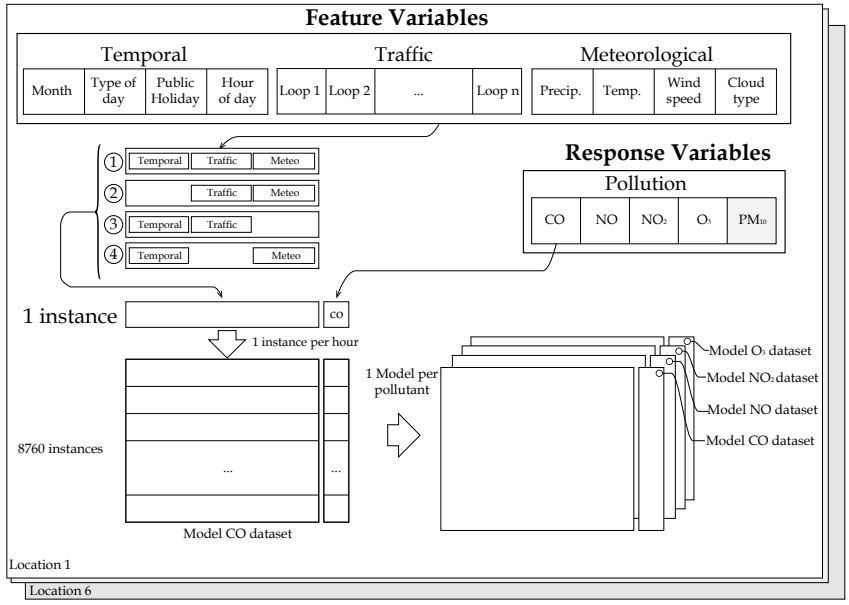

FIGURE 5.2: Example of dataset instances. The last four target variables correspond to pollutant readings, and are used separately when the regression models are built.

data, such as all the ones corresponding to the month of December 2015 or the first part of January 2015, for which there are no traffic data or lines with faulty meteorological readings. Such incomplete instances are deleted from the finally processed datasets.

5.3 Results and Discussion

As emphasized in the introduction to this chapter, the experiments and results next discussed are oriented towards quantitatively assessing the influence of meteorological conditions and traffic variables on the local pollution levels in different parts of Madrid in 2015. For this purpose this section gravitates on the analysis of the interactions among such variables, which is done by both visual inspection and supervised learning.

5.3.1 Traffic Characteristics in Selected Zones

Besides the already highlighted criteria for selecting among station locations, the disparate traffic levels recorded thereat constitute another reason why they were chosen. These traffic behavioral differences are observable in Figure 5.3, which result after aggregating the data captured by the ATRs of each considered zone. Columns in each subplot represent the daily averaged traffic flow, i.e. traffic is averaged first over the available ATR readings for each hour, then the average over 24 hours delivers the

value represented in this plot. SU-CC location is not considered in these graphs, as this location has no local traffic.

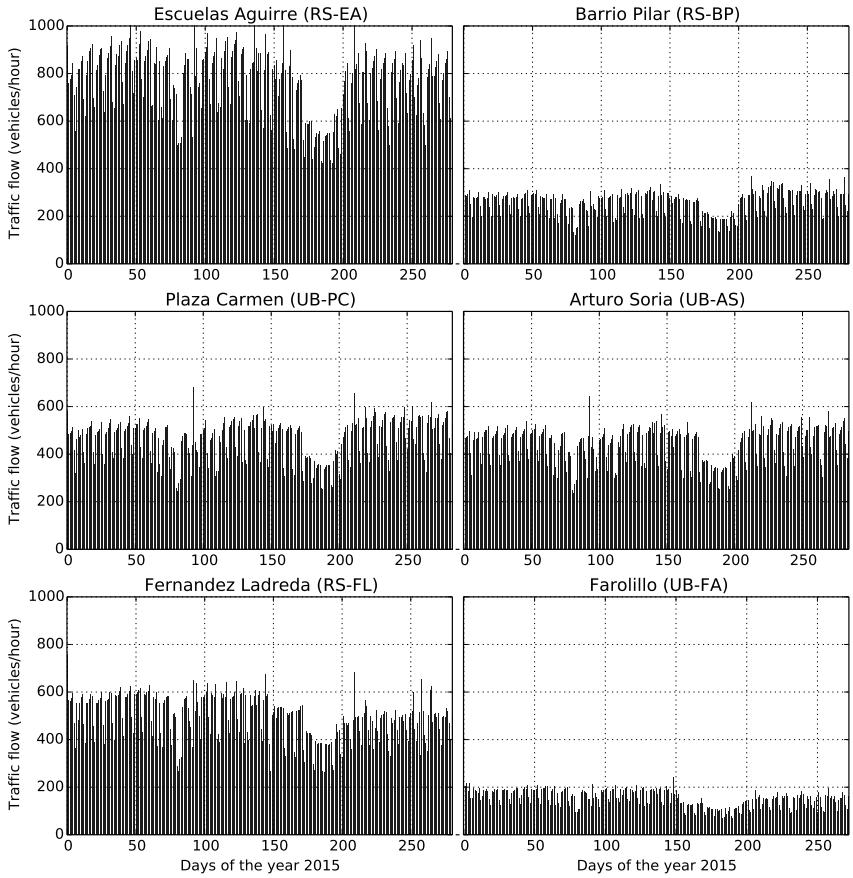

FIGURE 5.3: Comparison among six zones of the day-average traffic flow along the year. The horizontal axis represents each one of the days in the sample, and is shorter than 365 days because no traffic data were available for the first 14 days of January and the whole December 2015, and days with incomplete data were removed.

As expected, Figure 5.3 shows that the traffic flow in the most central zone (namely, RS-EA) is the highest one, whereas UB-FA supports the lowest traffic flow levels. It can be also observed that the area around RS-BP bears less traffic than those of UB-PC and UB-AS, being the first a roadside location and the latter urban background locations. RS-BP and UB-AS share characteristics: working class districts, close to main roads and more than 5 km far from the city center. Nevertheless, UB-AS almost doubles the average traffic in the nearby ATRs, and is considered a urban background measuring site, while RS-BP is contemplated as a kerbside location. When it comes to pollution, an opposite trend is noted: according to [236], RS-BP has exceeded the limit levels of NO_2 up to 95 times during 2015. UB-AS, holding more traffic, only exceeded the

level 18 times (below the 20-times alert level). Although both locations share similar demographic features, the higher presence of trees on both sides and the center of the road, and the lower amount of crossings, which induce to stop and start the engines of the vehicles, are the main elements present in UB-AS that explain the opposite trends in traffic and pollution [264]. As aforementioned, UB-FA location is in a working class district with no important roads around, and this situation is observable in the traffic readings. Traffic levels are similar in UB-PC and UB-AS; both are urban background locations, but the first is a small public square in the center, with ATRs in a one-way road (the rest of surrounding roads are pedestrian streets), while the latter is in a residential area and the ATRs are placed along two- or four-lane roads. Similar traffic flows in different type of roads imply higher occupation of the road in the smaller ones.

Aside from the discrepancies in the 6 zones, there are concurrences relative to the temporal dimension of the readings. In all 6 locations it is possible to observe a variably acute decrease in the beginning of the second half of the year. This decrease corresponds to August and late July, and it is more abrupt in center zones (RS-EA, UB-PC, where the working population is stable until the end of July), than in residential ones (RS-BP, UB-FA). There is also a noticeable decrease in the end of the first third of the year, corresponding to Easter holidays. This drop presents mostly the same duration in any area, as these holidays have always the same duration and the date is predefined. Additionally, a pattern is observable in each week: traffic levels increase gradually on weekdays, and drop in weekends.

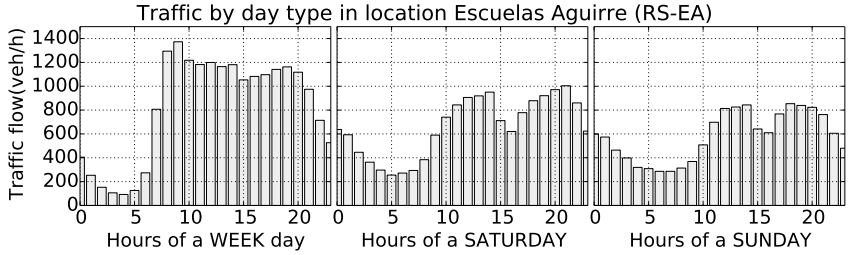

FIGURE 5.4: Average traffic per hour and day of the week in the RS-EA location. In this location, Sundays are similar to Saturdays, and both present more traffic by night and less by day than weekdays.

A closer look at this pattern is provided in Figure 5.4 for the traffic captured in the RS-EA location. This traffic characterization is usual in the literature related to traffic modeling [265], [266] and temporal features were found to have a decisive relevance in long-term traffic forecasting [34]. Based on this rationale, temporal features are subsequently incorporated to the dataset in order to improve the performance of the regression techniques, as previously explained in Section 5.2.5.

5.3.2 Climate in Madrid during 2015

Climate in Madrid is typically dry, with maximum temperatures rounding the 35 °C during summertime, and negative values in January and February. Figure 5.5 displays the daily average, maximum and minimum temperatures recorded in the utilized aerodrome observatory during 2015. As described in Section 5.2.3, METAR and SPECI reports provide qualitative information about the precipitation episodes, but not quantitative. For this reason, and for visualization purposes, the number of hours in which precipitation has taken place each day are also included in the plot. Remarkably, only on 6 days of the entire year it rained for more than 8 hours, all of them aligned with temperature declines.

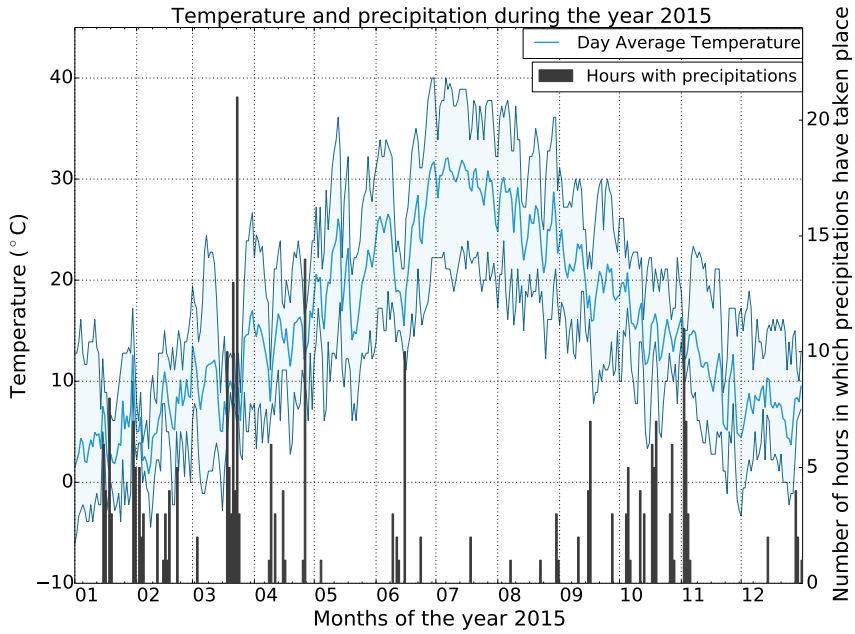

FIGURE 5.5: Temperature and precipitation in Madrid during 2015. Temperature is shown as a line with a shaded region from its minimum to its maximum value. Precipitations are shown as the number of hours at which there were precipitations in each day.

The monthly average historic data provided by the Madrid City Council show that temperatures are in line with the typical temperatures of this region through the year (Figure 5.6A), with some values above the average in July, November and December. Late autumn and winter have been specially dry, though, with 29.1 mm and 4.2 mm precipitated in November and December, as opposed to 51.4 mm and 40.3 mm averages for the same months (Figure 5.6B).

This lack of precipitations hinders dispersion and reduction of some pollutants such as PM_{10} and $PM_{2.5}$ [229]–[232], and are related to the pollution peaks recorded in the last part of the year in Madrid, which

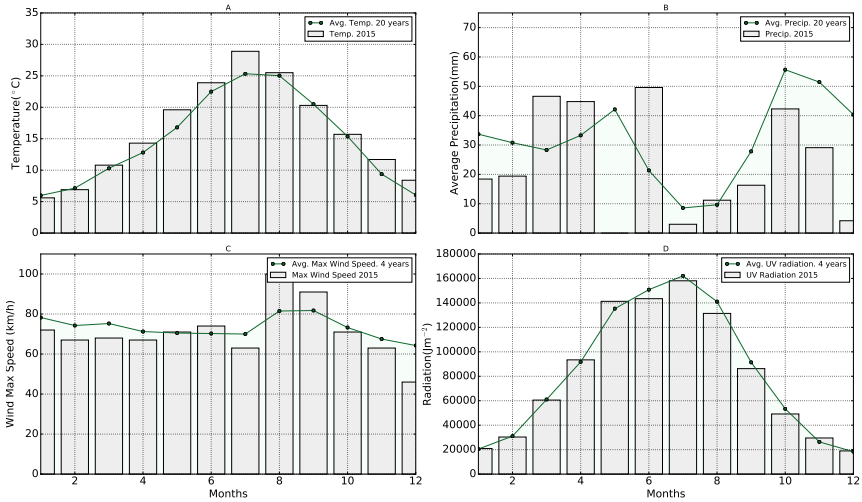

FIGURE 5.6: A) Average month temperatures in 2015 compared with average month temperatures of the period 1995-2015. B) Monthly precipitations in 2015 compared with average month precipitations of the period 1995-2015. C) Monthly wind max speed in 2015 compared with average month max speed of the period 2012-2015. D) Monthly total UV radiation in 2015 compared with average month UV radiation of the period 2012-2015.

ultimately lead to traffic restrictions. The other two relevant factors analyzed in this study, wind speed (able to disperse pollutant particles, or bring them from somewhere else) and UV radiation (instigator of chemical reactions that transform some pollutants into others), have maintained values close to the historic records, which are constrained to 4 years, in the available public data (Figures 5.6C and 5.6D). The maximum wind speed attained in December is the most different recorded data (46 km/h vs average 64 km/h). This along with the previously exposed data buttress the relevance of the change in typical winter meteorological conditions in Madrid that may be behind the utmost pollution levels recorded in the last months of 2015 in this region of Spain.

5.3.3 Pollution Characteristics in Selected Zones

Air quality monitoring stations have been selected considering the pollutant agents they are able to measure, their distance to direct sources of traffic pollution, and the type of station, defined by their location. As described in Section 5.2.1, the selected 6 stations are urban background and roadside, and they are placed in locations with dissimilar characteristics; different levels of pollution and traffic are expected. Figure 5.7 shows the distribution of CO, NO, NO_2, and O_3 through the year.

General seasonal trends are visible in the figure. O_3 reaches its maximum in summer months (35-120 $\mu g/m^3$) when meteorological conditions

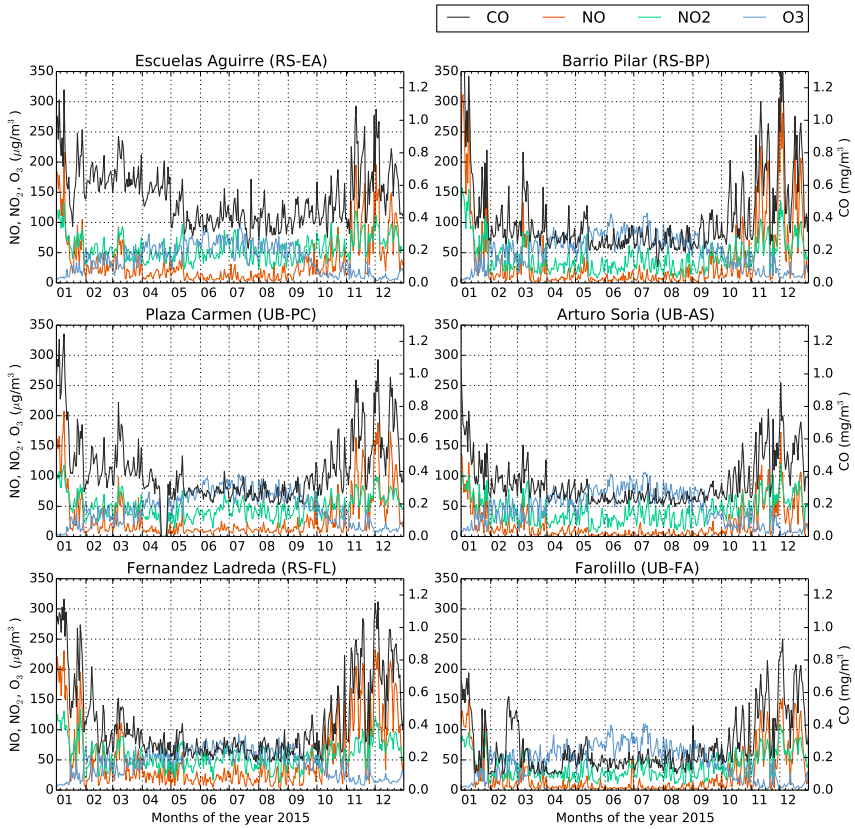

FIGURE 5.7: Daily averaged pollution levels for CO, NO, NO_2 and O_3 through 2015 for the 6 zones considered in this chapter.

facilitate its formation, whereas minimums (5-25 μg/m^3) are found in winter months (less solar radiation, shorter days), coherently with other studies carried out in southern Europe [260]. NO_2 levels remain stable through the year, closer to the 50 μg/m^3 line in roadside traffic stations and to 40 μg/m^3 in the urban background stations. Peaks attained by the CO and NO pollutants coincide when heating systems are active and ozone plunges.

PM_{10} levels are only available in 3 stations, and represented in Figure 5.8. RS-EA and UB-FA locations are the most dissimilar in the entire sample, as detailed in Section 5.2.1, and they are 5.2 km away. Yet, their PM_{10} measurements are resembling. Top and bottom peaks are produced in almost the same parts of the year, and the values follow roughly coincidental lines. Aside from natural sources and Saharan dust being significant contributors of PM_{10}, traffic is also a proven relevant source of this pollutant [220]. Furthermore the SU-CC station, far from traffic influence, presents lower yet similarly shaped PM_{10} levels through the year that constitute a background concentration over which traffic in the rest of locations is contributed. In light of the PM_{10} levels recorded in the rest of locations, traffic related share of this pollutant concentration can be concluded to be

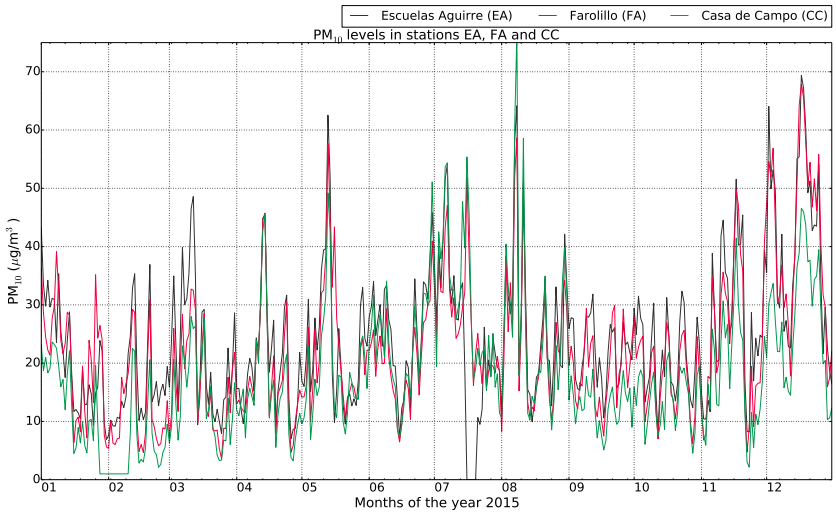

FIGURE 5.8: Daily averaged pollution levels for PM_{10} through 2015 comparing the two most antithetic traffic locations, and the control location SU-CC.

scarcely contributed by local sources, and represents mainly a background traffic contribution, which is in line with the conclusions drawn in [228]. On the basis of this evidence, and not having PM_{10} measurements for all the selected areas, this chapter will focus on the analysis of CO, NO, NO_2, and O_3.

5.3.4 Relations among Pollution, Traffic and Meteorological Conditions in Selected Zones

Air pollution in big cities is produced in a significant level by road vehicle emissions, with the modifying influence of meteorological agents. When cross-matching the previously presented data it is possible to discern similar effects for the city of Madrid. In order to assess the impact of meteorological conditions and local traffic in local pollution levels, the concentration of CO, NO, NO_2 and O_3 data of each site were analyzed with different temporal scales and overlaid with traffic levels. Annual results for the RS-EA location are shown in Figure 5.9. The plot depicts pollutant levels running seasonally, with increased O_3 during the summer months and the consequent increment of NO_2 and decrease of NO. Winter months undergo peaks of NO, coinciding with less ozone presence and heating systems being active.

Although these trends can be a priori expected, there is no apparent relation with traffic. When traffic plummets in August, NO and CO levels are maintained. The NO concentration even peaks over 30 $\mu g/m^3$ on August 26^{th}, the 9 daylight hours with overcast cloud coverage, might be behind the low levels of O_3 (under 50 $\mu g/m^3$). In Figure 5.5 it is possible to observe a week in the late March when it rained for several days, and temperatures decreased in 6-7 °C relative to the previous trend.

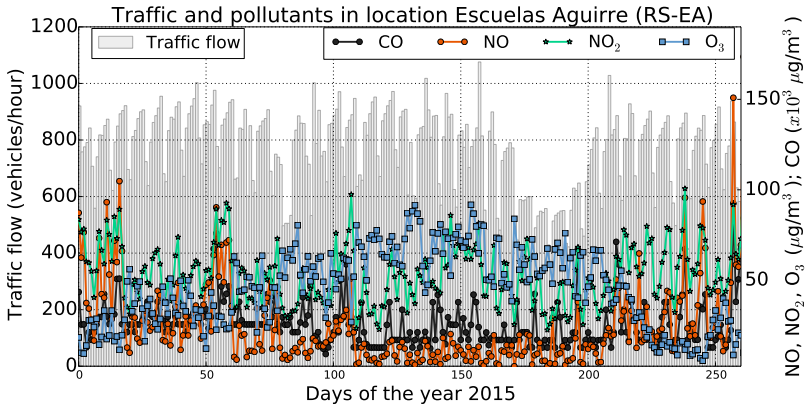

FIGURE 5.9: Traffic and pollution levels through 2015 in RS-EA.

This corresponds in Figure 5.9 to the NO peak and O_3 valley after day 50. Traffic was about the same as in the previous week, but pollution increased. When examining the same data for UB-FA (the most different location), a similar decoupling is found as shown in Figure 5.10. Traffic is flatter through the year, and although pollution levels are lower, they follow a similar trend.

Anyhow, presented results provide a comparative insight on the predictability of pollution based on traffic and meteorological conditions. In the first place, the worst results are obtained when no temporal information is provided to the model (Table 5.1, figures under the ② label). Regardless the particular location or the predicted pollutant, the best R^2 scores are achieved for every model in any other combination of features, which is a revealing indicator of the seasonality of the data along time. Scores labeled under ③ in Table 5.1 correspond to the results obtained without meteorological data, and even if slightly better than those of ②, they are still far from the best obtained. Required temporal variables combined with traffic levels seem to perform poorly without meteorological information. Among these results, locations with greater traffic flow levels achieve the best scores in traffic-emitted pollutants (NO and CO): RS-EA and RS-FL are close to important junctions, while UB-PC and UB-FA are far or blocked from main arteries. O_3 is not linked to traffic as directly as CO and NO, hence its scores teeter among locations not connected with each traffic density; this effect is transferred to NO_2 scores. The results obtained for the SU-CC location behave in a similar way for CO, NO, NO_2 and O_3: models with the three types of inputs and fed with temporal and meteorology information obtain the best performances, with barely significant differences among them. This represents a first evidence of a low relevance of local sources of traffic emissions, as models from all kind of locations – with acute differences in local traffic – perform in a very similar fashion. A posterior analysis will delve into this statement.

This detachment is not found if the time scale is varied. In Figure 5.11 traffic levels are plotted by day divided by working and non-working

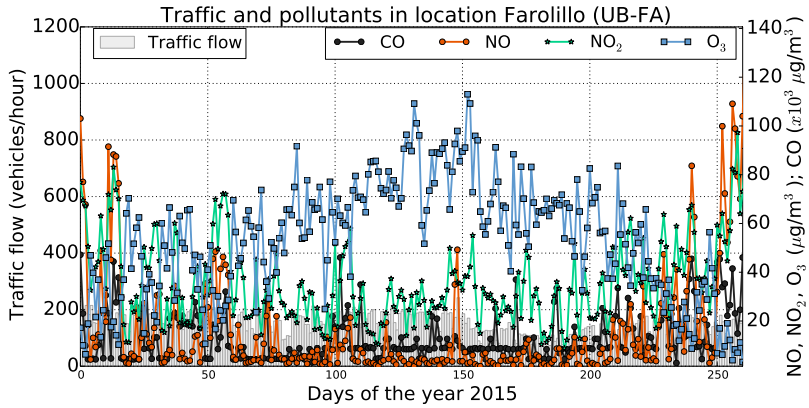

FIGURE 5.10: Traffic and pollution levels through 2015 in UB-FA.

days, summer (April to October) and winter (November to March) months. Working and non-working days separation affects the traffic levels, and season separation impacts on both traffic and pollution levels. By inspecting these figures further, a closer relation between traffic and pollutants is discovered: when traffic starts in the first hours of the morning, specially in working days, the formation of NO and CO is triggered. After 10 AM, and particularly in summer, the ozone formation increases, combining itself with NO and reducing its levels. The decay of both starts at the same time as the night (less traffic producing NO and less sunlight inducing O_3). Lighter traffic in weekends provokes higher levels of O_3, which is known as the *weekend effect* [260]. Differences between these two scopes have been already been noted by [230], [257], [259], [260] for diverse pollutants and over different cities.

These variables are used to build the datasets defined in Section 5.2.5.RF regression models are built for each of the 112 datasets (Figure 5.2) and evaluated by applying shuffled cross-validation with $K = 10$ folds. For instance, the first dataset is created with temporal, traffic and meteorological variables as features, and CO measurement as the target variable in the RS-EA location. One data sample is created for every hour of the year, resulting in 8760 samples that are later cleansed by removing those with missing attributes. For the cross-validation, the dataset is split into 4 sub-datasets, each containing a shuffled random fourth portion of the original instances. The model is then trained with 3 of the 4 sub-datasets and tested with the remaining one, rendering performance metrics that are stored for subsequent processing. This process is repeated 10 times for each model, with different compositions of each sub-dataset, and the overall performance is averaged among the results of the 10 executions. Coefficients of determination are extracted to observe how the response variables are fitted by the model, whereas the MFB of each model is obtained to evaluate the performance of the model.

Table 5.1 shows the averaged R^2 and MFB values of the 112 models. These performance values were achieved after several iterations in which

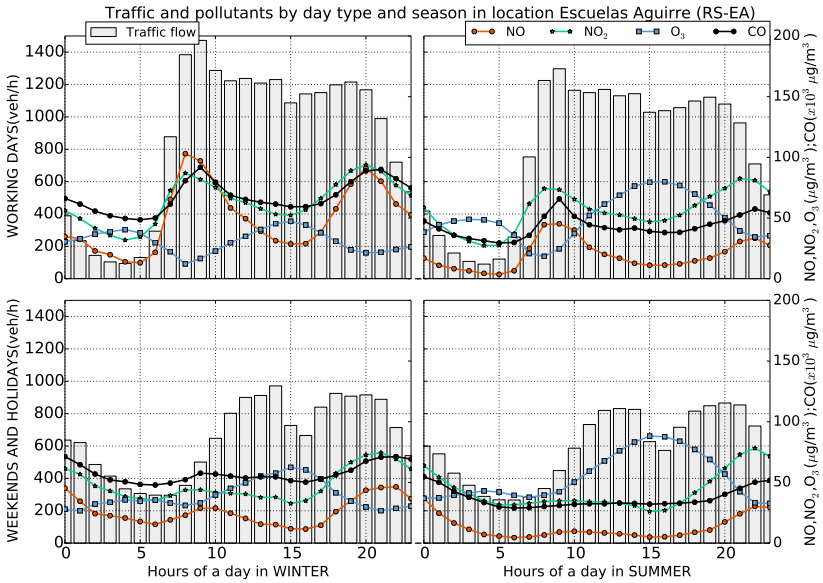

FIGURE 5.11: Hourly average pollutant and traffic readings by day type and season in RS-EA.

the RF model was refined, setting its parameters by way of a grid search procedure. Even after the parameters of the model were tuned for a better performance of the model, the obtained R^2 scores are in general low (under 0.7). The MFB scores were in general under 0.3, which reveal a fair performance of the models except for the NO concentrations. However, if predicting pollutant levels were the main purpose of these models the input features to each model would need to be predicted as well. For instance, to predict NO at a certain moment in the future, traffic and meteorological conditions at that moment would be also required; being a future instant would make necessary to predict the traffic and meteorological conditions at that point. Relying in predicted values as inputs to the predictive model would probably lead to worse results.

PM_{10} is also a pollutant influenced by traffic, but as seen in Section 5.3.3, it is affected mainly by general contributions, not local. The performance scores reported for the models in the SU-CC location bolster this idea: a much lower direct traffic influence helps this pollutant to be more predictable. Local traffic emissions in RS-EA and UB-FA may introduce small-scale variations that make it more difficult to predict. The relevance of traffic sources in the total PM_{10} concentration levels in Madrid could be addressed in depth in a future research. Other obtained results buttress further this observation: Table 5.1 (① and ④) contain the best R^2 scores, being very similar. Although ① comprehends most of the best R^2 scores obtained, and ④ all of best MFB scores (in bold type), the differences for both metrics between ① and ④ are in the 10^{-2} or even 10^{-3} order of magnitude. Once again, the higher improvement when incorporating traffic is produced in RS-EA.

TABLE 5.1: R^2/ MFB scores of 112 (+12) models.

	RS-EA	RS-BP	RS-FL	UB-PC	UB-AS	UB-FA	SU-CC
			① **Temporal + Meteorology + Traffic**				
CO	**0.46**/0.11	**0.38**/0.11	**0.57**/0.09	0.47/0.14	**0.44**/0.13	0.14/0.21	0.29/0.14
NO	**0.35**/0.33	**0.32**/0.47	**0.43**/0.31	**0.39**/0.41	**0.31**/0.30	**0.39**/0.42	**0.31**/0.35
NO_2	**0.53**/0.07	0.48/0.14	0.49/0.07	**0.51**/0.13	**0.51**/0.08	0.52/0.12	**0.53**/0.21
O_3	**0.70**/0.04	0.72/0.07	0.71/0.05	**0.76**/0.05	**0.72**/0.06	**0.71**/0.11	0.69/0.04
PM_{10}	0.38/0.15	-	-	-	-	0.37/0.17	**0.48**/0.06
			② **Traffic + Meteorology**				
CO	0.32/0.14	0.28/0.13	0.38/0.11	0.31/0.17	0.20/0.16	0.30/0.22	0.15/0.14
NO	0.21/0.37	0.22/0.53	0.28/0.34	0.20/0.45	0.10/0.37	0.28/0.51	0.16/0.5
NO_2	0.35/0.10	0.34/0.18	0.33/0.09	0.29/0.16	0.26/0.11	0.37/0.16	0.38/0.24
O_3	0.53/0.05	0.57/0.09	0.55/0.06	0.59/0.06	0.51/0.08	0.58/0.13	0.58/0.05
PM_{10}	0.21/0.22	-	-	-	-	0.21/0.23	0.39/0.10
			③ **Temporal + Traffic**				
CO	0.32/0.11	0.28/0.11	0.41/0.09	0.22/0.12	0.26/0.14	0.24/0.21	0.06/0.14
NO	0.21/0.32	0.19/0.44	0.27/0.29	0.16/0.42	0.17/0.31	0.19/0.43	0.16/0.37
NO_2	0.38/0.08	0.33/0.16	0.31/0.08	0.31/0.13	0.29/0.09	0.34/0.14	0.28/0.24
O_3	0.58/0.04	0.60/0.07	0.57/0.05	0.65/0.05	0.57/0.06	0.57/0.11	0.53/0.05
PM_{10}	0.22/0.22	-	-	-	-	0.25/0.21	0.25/0.15
			④ **Temporal + Meteorology**				
CO	0.43/**0.09**	0.37/**0.09**	0.56/**0.09**	**0.47**/**0.13**	0.39/**0.12**	**0.42**/**0.19**	**0.31**/**0.13**
NO	0.31/**0.31**	0.32/**0.42**	0.43/**0.28**	0.40/**0.39**	0.26/**0.27**	0.37/**0.34**	0.31/**0.28**
NO_2	0.52/**0.06**	**0.49**/**0.13**	**0.49**/**0.07**	0.51/**0.12**	0.49/**0.07**	**0.54**/**0.10**	0.51/**0.18**
O_3	0.68/**0.04**	**0.72**/**0.07**	**0.71**/**0.05**	0.75/**0.05**	0.68/**0.06**	0.69/**0.09**	**0.69**/**0.04**
PM_{10}	**0.39**/**0.18**	-	-	-	-	**0.37**/**0.19**	0.47/**0.07**

TABLE 5.2: Wilcoxon p-values comparing pairs of R^2 result sets.

	RS-EA	RS-BP	RS-FL	UB-PC	UB-AS	UB-FA	SU-CC
CO	0.05	0.57	0.05	0.50	0.01	0.16	0.23
NO	0.24	0.79	0.24	0.24	0.03	0.01	0.71
NO_2	0.38	0.95	0.87	0.24	0.02	0.87	0.51
O_3	0.07	0.20	0.20	0.01	0.01	0.16	0.62

A non-parametric Wilcoxon hypothesis test has been performed for the R^2 scores in order to confirm this conjecture and to shed light on the statistical significance of such performance gaps. Table 5.2 shows the Wilcoxon test results (p-values); in general, statistically significant discrepancies (p-values close to 0) are few. This means that for most cases, there is no evidence that the medians of the score sets being compared differ significantly from each other. UB-AS location presents, though, the opposite outcome: R^2 values have the larger differences between traffic and no-traffic datasets, and Wilcoxon p value rejects the hypothesis that the difference is due to chance. Figures 5.7 and 5.3 bolster this idea: UB-AS and UB-PC share very similar traffic levels, but NO and CO concentrations read in UB-AS are lower, with NO ranking from 1 to 15 $\mu g/m^3$ in UB-AS and from 10 to 30 $\mu g/m^3$ intervals in UB-PC. These differences might be caused by a variety of factors: natural ventilation of the area, presence of vegetation or others. Regardless of the reason, UB-AS – a

background pollution measuring station – is more affected by local traffic than the other stations. A deeper analysis, left out of the scope of this study, should determine its causes, possibly by resorting to topographical urban models. In those locations where O_3 concentrations are best estimated with traffic data (RS-EA, UB-PC, UB-AS, UB-FA), the Wilcoxon test shows a relevant difference not due to chance. Road traffic emissions are linked to O_3 as they modify its concentrations when combined.

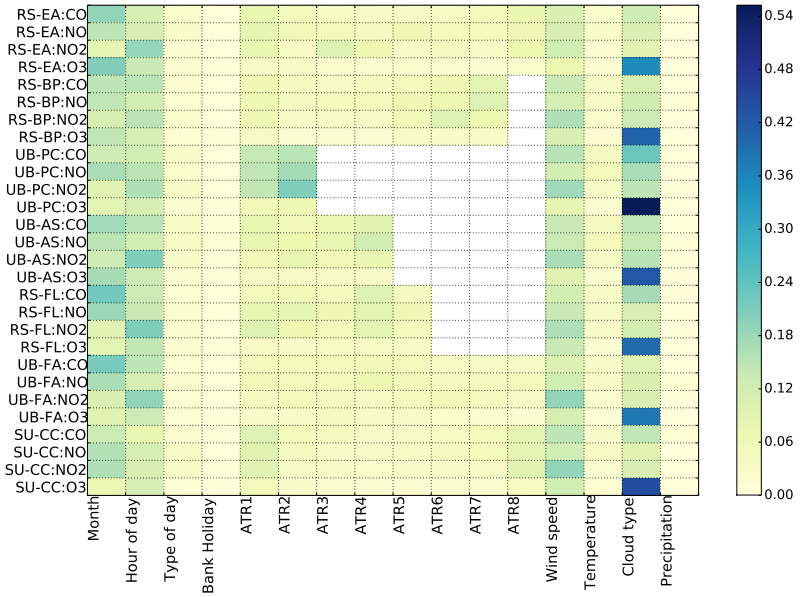

FIGURE 5.12: Feature importance of each variable for each dataset. Blank cells are part of datasets for which less ATR readings have been used due to the distance to the air monitoring station criteria.

The previous set of analyses is completed by Figure 5.12, where the feature importances of the different predictors – as provided by the RF regressor – is plotted as a heat map in order to analyze in detail the coupling of features and response variable. A noteworthy outcome is the low relevance of precipitation feature, which is in fact the less relevant for all datasets. This could be attributed to the lack of precipitations in 2015 in Madrid. Precipitations are indeed an important pollution-modifying factor, but when held in so infrequently it does not make a good general predictive feature. Also, the use of a discrete scale of precipitation levels, in the dearth of real millimeter readings, could reduce the relevance of this feature. On the other hand, cloud types are provided to the model in a similar discrete scale, and they are a generally relevant feature, and particularly the most relevant feature to predict O_3. Despite temporal features have been found to be the most determining ones, two of them – public holidays and day types – are scarcely relevant. These variables were useful in [34] to predict traffic, but in the long cycles of pollutants they have no effective value. On the other hand, months seem to be the most relevant (darker shade) for predicting CO and NO in almost every

location, whereas hour of day influences specially in NO_2. CO and NO are produced by different sources and maintained in the air in cycles that depend on meteorological and season factors, and NO_2 levels are highly driven by the hour of day, as its production is linked to O_3 and the latter exhibit day and night cycles. Wind speed is a good NO_2 predictor, with importance values around 15% and peaking in 20% at the UB-FA location (the one with lower buildings, and in a flatter area more exposed to wind effects). The importance of ATR readings is in general low, which not only validates the R^2 scores and the analysis, but also finds that in all cases traffic levels relevance for predicting CO, NO or NO_2 is doubled for predicting O_3.

Although in Table 5.1 O_3 predictions always render better performance scores, the analysis of the feature importances shows that this performance is in any case linked to other variables, specially the cloud type. Other conspicuous outcome is the relatively high importance of traffic features in the UB-PC dataset; this is related to the distribution of the feature importance, which sums 1 for each dataset. Distributing the importance among less features adds relative weight to them. For the same reason, cloud type feature predicting O_3 is more relevant as the dataset is smaller (0.35-0.4 for RS-EA and UB-FA, and 0.55 for UB-PC). Nonetheless, aggregated ATR measurements importance for each dataset gives an average 0.44 importance for all the ATRs in RS-EA and 0.27 for UB-PC, which confirm that traffic is more relevant in RS-EA than in UB-PC, as results in Table 5.1 clearly show. The aggregated importance is useful for comparison purposes, but not for feature importance analysis, because of the way RF sub-samples the dataset. In each tree a random subset of features is used to build the model, and only a small portion of them will have all ATR reading features concurring in the same subset.

The fact that incorporating traffic levels to the predictive model could worsen the scores was unexpected, and makes local traffic readings become noisy in some cases. This does not mean that traffic is irrelevant for pollution – it is indeed one of the main contributors to pollution in Madrid [218] –, but rather that its effects on particular sites are limited and stringently linked to localities. Other research contributions such as [257], [259], [267] have addressed similar subjects and found significant disparities among urban background and roadside stations, being the latter highly influenced by vehicular emissions. In Madrid only measuring sites with heavier traffic and/or higher vegetation density have been shown to be influenced by traffic slightly over the direct effects of meteorology and seasonality.

5.4 Conclusions

Many are the sources of air pollutant agents: industry, agriculture and livestock farming, road and air traffic, forest fires, natural sources like volcanoes or particles drawn by wind, among others. A wide variety of elements modify the concentration of different pollutants, mainly meteorological agents, but also topography, tree and shrub presence, building

distribution or water streams like rivers. This chapter has examined the effects of local road traffic, meteorological conditions and temporal variables on air pollution in Madrid. Data collected from 6 air monitoring stations, 33 ATRs and data from a meteorological observatory were used to build supervised learning models and analyze the relationships among these variables. Results have shown that pollutant agent levels in the 6 evaluated locations were weakly linked to local vehicular emissions. An additional suburban station was also analyzed to account for the influence of regional background PM values on the local measurements.

The outcomes provided by RF regression and the analysis of the importance of the features during the training process of the model suggest that seasonal features are the most relevant when predicting CO, NO and NO_2, with daily seasonality affecting both residential and downtown areas. PM_{10} concentrations have been studied for two of the locations, unveiling a slight impact of local traffic in these particles. Within the Madrid urban area, CO, NO, NO_2 and O_3 concentrations are strongly influenced by meteorological factors, specifically wind speed and cloud type, with less temperature influence. Precipitations, a usually influential actor in pollution alteration, have been proven to have a minor effect in pollutant concentrations over Madrid during 2015. The fact that this year has been specially dry, and the measurement scale used might be the most likely contributors to the poor predictive performance of this variable. Meteorological agents like precipitation in millimeters, wind direction, humidity or pressure have been left out of the study in the lack of proper public sources of data. Using actual precipitation or UV radiation readings instead of a proxy scale could improve estimator results.

The meager impact of traffic emissions on pollution levels is a remarkable outcome of the analysis that should be observed cautiously. Vehicular exhaust chemicals have been proven to be the major contributors to Madrid pollution levels. However, other factors such as its flat topography, the absence of zones with concentration of high buildings and its dry, atmospherically stable conditions contribute to a global background pollution that affects in similar ways to different areas. As global pollution increases daily with contributions from all city traffic, industry, heating systems and others, the local contributions decrease relatively to the volume of general accumulated concentrations of pollutants. Thus, only locations supporting heavy traffic produce a contribution of traffic-related pollutant chemicals largely enough to impact on local concentrations of pollutants under study.

The overall results of this study suggest that countermeasures to reduce pollution based on restricting traffic for short periods of time could have a modest impact if meteorological conditions to mitigate the current accumulated pollution do not occur. The particular climatic characteristics of Madrid and the increasing road traffic lead to the belief that long-term measures such as permanent low-emissions zones, campaigns for promoting the use of public transportation or policies favoring the widespread adoption of electric vehicles could help containing city-wide pollution issues in a more effective manner.

Chapter 6

Concluding Remarks

Traffic forecasting is one of the key elements to consider in the growing area of Intelligent Transportation Systems, as it can be ascertained from the huge interest it attracts within the research community. Thanks to the upsurge of data availability and of the techniques that allow to handle them, important advances have been made in recent years, although some of traffic prediction dimensions, like the forecasting horizon remain unaffected. This Thesis has aimed at shedding light on the long-term forecasts, as well as on other relevant aspects such as the imputation of missing data and the application of ML techniques and traffic data to other related fields. In particular, the contributions and findings of this dissertation can be categorized in four main blocks:

- **State of the Art of the Field**

After analyzing more than 100 related studies published during the last 2 years, the state of the art on traffic forecasting was updated. Some aspects of traffic are common to all previous reviews on the subject, and analyzing these commonalities portrays what researchers have found fundamental about this discipline. A set of common issues and challenges is found in all previous reviews, regarding prediction scope and context, most suitable model selection, metrics for different models comparison, and hybridization of models to improve performance. These aspects remain pertinent, 30 years after the first survey that exposed them, and regardless the shift to data-driven modeling. Besides, new issues and challenges arise in the light of this now prevailing way of modeling. Collecting data from new sources involves fusing different types, and managing data aggregation and resolution are some of the most recent challenges proposed by literature reviews. Additional challenges have been introduced in this chapter: increasing the prediction horizon, incorporating exogenous factors to models, and adding data ageing mechanisms that allow models to adapt their learning to changing circumstances. The latter two are linked to data-driven modeling, and help achieving the first. Further prediction horizons are specially useful for traffic management, and having days or weeks forecast can change traffic managing measures from being reactive to being proactive. Their performance has always been considered poor, compared to short-term predictions, but these are slightly useful if there is no time for reaction. This performance can be improved when the models

are data driven, by incorporating new sources of information that complete the traffic configuration scheme, and with adaptive learning methods. Another issue rises here that was anticipated since the first reviews on the field: the impact that road users having the information of future traffic status may imprint on the short-term traffic patterns themselves.

- **Data Engineering**

Understanding and preprocessing data should be a mandatory first step of any data-driven analysis or forecasting endeavor, in what is commonly known as data engineering. Many of the features can be correlated among them, or new features can be obtained from combinations of the known ones. The number of instances can be increased if there are too few, or reduced if there are repeated or non valid ones. Besides, and specially when the data are obtained from real-world sources, they can carry noise, include erroneous values or be incomplete. This latter case is recurrent in traffic forecasting, where researchers have to deal with sequences of data that contain abundant gaps. Although this is a crucial part of the process, in most of research a preprocessing stage is not reported and it is kept out of the results and posterior analysis. When preprocessing is explicitly performed, there is no common ground on how to deal with gaps, or the implications of using one technique or other. In Chapter 3 the most prevailing imputing strategies have been analyzed, and new techniques have been presented and assessed, intended to deal with long series of missing data, which are more usual than literature on imputation suggests. Beyond the importance of introducing methods to handle this long-gaps issue, a relevant contribution has been made regarding the way in which imputation methods are evaluated, normally comparing imputed values to real ones, and not considering the actual usage of imputed data for the forecasting. From this perspective, the analysis has revealed that for some types of missing data, the adopted kind of imputation method is scarcely relevant.

- **Long-term Traffic Forecasting Modeling**

Classic forecasting techniques like time series modeling or in general any method that relies solely on past observations are not enough to obtain long-term predictions. However, there are divergent approaches that allow for a long-term traffic characterization, based on finding patterns that can be used as forecasts. A pattern clustering and classification scheme has been proven to comprise a useful tool to obtain long-term predictions. The whole long-term forecasting scheme proposed in Chapter 4 automatically finds the characteristic days (clustering) and uses them as predictions for future days that fit in the patterns (classification). The initial design of this scheme is crucial, and when all its parameters are optimized and the features used are carefully selected, it is able to provide accurate predictions months away for the latest day that the training has observed, as a favorable by-product of the traffic seasonality. This operation mode yields accurate results for most days, and with enough past data, it can be tuned to perform even better. However, unforeseen circumstances can happen

and evolve to a very different traffic profile than the expected one, ultimately misleading the model. For these hindmost cases, an adaptation mechanism can provide the means to obtain acceptable levels of forecasting accuracy. This systems operates online, and seeks prediction failures, providing alternate predictions for those cases. Such predictive design allows, when properly trained, to work autonomously generating long-term forecasts and correcting them online when necessary. These corrections yield average prediction gains, resulting in a more robust forecasting system with an improved overall performance.

- **Applying Traffic Forecasting**

The domain of application of traffic forecasting is obvious, as it is an end in itself; predicting the future levels of traffic is useful *per se*, for traffic managers and road users. However, there are other interests underlying the need for traffic management: the real demand is to alleviate congestion and to reduce its impact in noise and pollution. In fact, road traffic is one of the main contributors to air pollution in cities, reason for which municipal governments around the world develop environmental policies and establish restrictive circulation measures to curb the impact of vehicle emissions. Predictions of traffic levels are of clear interest for the implementation of these measures, allowing authorities to supply adapted restrictions. However, the correlations between traffic in a city location and the pollutant levels registered there can be less linear that intuition conveys. Building predictive models that incorporate traffic, pollutant levels, and weather variables to analyze pollution in a big city produced an unexpected outcome: in a local context, and given a certain meteorologic circumstances, the impact of road traffic in pollution is limited. Cities with very stable dry weather can accumulate high levels of background pollution that are slightly altered by local sources of exhaust emissions, making the pollution highly related to weather factors and less related to the amount of vehicles passing by. This situation should not be disregarded by administrations in those cities, where short-term measures could have a significant economic and social impact, but no impact on pollution levels.

6.1 List of Publications

As a result of the research conducted while pursuing this doctoral degree, several contributions were published in journals and conferences of the traffic forecasting area, which are listed below:

- **Journal publications**
  - I. Laña, J. Del Ser, A. Padró, M. Vélez, and C. Casanova-Mateo, "The role of local urban traffic and meteorological conditions in air pollution: A data-based case study in Madrid, Spain", Atmospheric Environment, vol. 145, pp. 424–438, 2016.
    JCR 3.629 49/229 **Q1** *Environmental Sciences*

  - I. Laña, J. Del Ser, M. Vélez, and E. I. Vlahogianni, "Road traffic forecasting: Recent advances and new challenges", IEEE Intelligent Transportation Systems Magazine, vol.10, pp. 93–109, 2018.
    JCR 3.019 65/260 **Q1** *Engineering, Electric & Electronical*
  - I. Laña, I. Olabarrieta, J. Del Ser, and M. Vélez, "On the imputation of missing data for road traffic forecasting: New insights and novel techniques", Transportation Research Part C: Emerging Technologies, vol. 90, pp. 18–33, 2018.
    JCR 3.968 6/35 **Q1** *Transportation Science and Technology*
  - I. Laña, J. L. Lobo, E. Capecci, J. Del Ser, and N. Kasabov, "Adaptive long-term traffic forecasting with evolving spiking neural networks", under second review round in Transportation Research Part C: Emerging Technologies, 2018.
    JCR 3.968 6/35 **Q1** *Transportation Science and Technology*

- **Conference publications**
  - I. Laña, J. Del Ser, and I. Olabarrieta, "Understanding daily mobility patterns in urban road networks using traffic flow analytics", in IEEE Network Operations and Management Symposium (NOMS), 2016.
  - I. Laña, J. Del Ser, M. Vélez, and I. Oregi, "Joint feature selection and parameter tuning for short-term traffic flow forecasting based on heuristically optimized multi-layer neural networks", in International Conference on Harmony Search Algorithm, pp. 91–100, 2017.
  - I. Laña, J. Del Ser, and M. Vélez, "A novel fireworks algorithm with wind inertia dynamics and its application to traffic forecasting", in IEEE Congress on Evolutionary Computation (CEC), pp. 706–713, 2017.
  - I. Laña, E. Capecci, J. Del Ser, J.L. Lobo, and N. Kasabov, "Road traffic forecasting using NeuCube and dynamic evolving spiking neural networks", accepted for its presentation in the 12th International Symposium on Intelligent Distributed Computing (IDC), 2018.

6.1.1 Other Publications

Besides, the author has also collaborated in the research yielding the following publications:

- **Journal publications**
  - A.I. Torre-Bastida, J. Del Ser, M.N. Bilbao, M. Illardia, S. Campos-Cordobes, and Ibai Laña, "Big Data for transportation and mobility: recent Advances, trends and challenges", accepted for its publication in IET Intelligent Transport Systems, in press, 2018.

JCR 1.387 21/35 **Q3** *Transportation Science and Technology*

- J.L. Lobo, I. Laña, J. Del Ser, M.N. Bilbao, and N. Kasabov, "Evolving spiking neural networks for online learning over drifting data streams", Neural Networks, Volume 108, pp. 1-19, 2018.

  JCR 7.197 7/132 **Q1** *Computer Science, Artificial Intelligence*

- E. Capecci, J.L. Lobo, I. Laña, J.I. Espinosa-Ramos, and N. Kasabov, "Modelling gene interaction networks from time-series gene expression data using Evolving Spiking Neural Networks", accepted for its publication in Evolving Systems, 2018.

- **Conference publications**

  - J.L. Lobo, J. Del Ser, M.N. Bilbao, I. Laña, and S. Salcedo-Sanz, "A probabilistic sample matchmaking strategy for imbalanced data streams with concept drift" in International Symposium on Intelligent and Distributed Computing, pp. 237-246, 2016.
  - I. Olabarrieta, A.I. Torre-Bastida, S. Campos-Cordobes, I. Laña, and J. Del Ser, "A heuristically optimized complex event processing engine for Big Data stream analytics", in International Conference on Harmony Search Algorithm, pp. 101-111, 2017.
  - J. Del Ser, A.I. Torre-Bastida, I. Laña, M. N. Bilbao, C. Perfecto, "Nature-inspired heuristics for the multiple-vehicle selective pickup and delivery problem under maximum profit and incentive fairness criteria", in IEEE Congress on Evolutionary Computation (CEC), pp. 480–487, 2017.
  - E. Osaba, J. Del Ser, A.J. Nebro, I. Laña, M.N. Bilbao, and J. Sanchez-Medina, "Multi-objective optimization of bike routes for last-mile package delivery with drop-offs", accepted for its presentation in the 21st IEEE International Conference on Intelligent Transportation Systems, 2018.
  - J.L. Lobo, J. Del Ser, I. Laña, M. N. Bilbao, and N. Kasabov, "Drift detection over non-stationary data streams using evolving spiking neural networks", accepted for its presentation in the 12th International Symposium on Intelligent Distributed Computing (IDC), 2018.
  - D. Nandini, E. Capecci, I. Laña, L. Koefoed, G. Kishore Shahi, and N. Kasabov, "Modelling and analysis of temporal gene expression data using spiking neural networks", accepted for its presentation at the 25th International Conference on Neural Information Processing (ICONIP), 2018.

- **Book chapters**

  - S.Campos-Cordobés, J. Del Ser, I Laña, I. Olabarrieta, J. Sánchez-Cubillo, J. Sánchez-Medina, and A.I. Torre-Bastida, "Big Data in road transport and mobility research"", Intelligent Vehicles, pp. 175-205, Elsevier, 2018.

6.2 Future Research Lines

There is a voluminous body of literature on traffic forecasting and the abundance of very similar works might suggest that the field has been exhausted. The outlook, however, is promising, and many aspects are yet to be explored. Some of them have been partly object of study during the development of this Thesis, but researching them have originated new questions and reopened long-standing issues. The most noteworthy issues are listed below as future lines of research:

- External variables: any of the methods presented in this Thesis could have been improved upon the availability of other sources of data. Inputs like weather, events or incidents are essential to the estimation of traffic, but their public availability is restricted. Sometimes a good source of traffic data is found in a city, but there is no weather data available for that city or *vice versa*. In other cases, the source is available but the granularity of data is insufficient, or they are not geolocated. For instance, there is a complete and timestamped open data source of traffic incidents in the city of Madrid, but there is no spatial information, so it is not possible to link it to traffic data traces. This kind of inputs could enrich considerably predictive models, and especially in the long-term ones, they allow for pattern sets that consider more diverse situations and thus perform better. An interesting further development of the long-term prediction paradigm presented in this Thesis would involve considering training periods longer than a year, and including external variables to the model.

- Concept drift and adaptation mechanism: when circumstances not contemplated by the long-term scheme happen, the change detection and adaptation system is able not only to provide more accurate predictions, but also to slightly modify itself and the pattern identification system so in the future it performs better in similar circumstances. Clusters are updated with new days that are similar to their pattern, but in the presence of these new days, the whole clustering configuration could change, should the clustering process be performed again. This adaptation has not been considered, but if made in the proper moment it would confer the system more robustness, and it could help dealing with long-term concept drift. Besides, this opens the door for implementing life-long learning strategies that keep the models acquiring new knowledge.

- Self-affecting predictions: an issue awaits future researchers who deal with highly effective predictions, specially the long-term ones: when they become massively adopted by drivers to take routing decisions, those will be able to modify the traffic and affect the predictions themselves, rendering them useless. At some point of the adoption of prediction technology, anticipating a congestion situation in a road could lead to many drivers taking an alternative route and generating the congestion in other road. Hybrid schemes encompassing predictive models and simulation tools could help addressing this issue.

- Multi-location forecasts or network-level predictions, which would help in part to deal with the previous point: this is in general a commonly accepted challenge in traffic forecasting field, as most research is focused on obtaining results for one or more nodes of the road that operate independently. It is possible to find works in recent years pointing in that direction, but it is still hard to find systems that can produce forecasts for a whole network at a time, possibly exploiting intra-detector spatial correlations. They could become essential for intelligent signaling and also for intelligent routing in autonomous driving.

- Data fusion and Big Data: the era of connected vehicles and sensed roads is near, and profuse data are becoming more available to exploit. Implementation of traffic forecasting tools in Big Data architectures also allows for real time predictions based on several types of input, taken from different open and not open sources in an effective way. Nevertheless, a significant work is required in this field; learning methods ought to be parallelized to be capable of mining chunks of intercorrelated data without jeopardizing the quality of the predicted variable. Likewise, the interaction of traffic information with non-structured datasets is also challenging due to the essential differences that may eventual characterize them in terms of ageing, drift, graph-like nature and spatial connectedness.

Appendix A

Open Road Traffic Data

All experiments conducted during this Thesis have been built upon real traffic data, collected from a public source maintained by the City Council of Madrid (Spain), which has currently more than 7800 vehicle detection sensors deployed through its road network. These consist of 200 optical detectors, 1400 urban freeway sensors, and around 6200 sensors placed under traffic lights in urban roads intersections. These detecting devices are arranged in around 4100 measuring stations, which commonly embody more than one sensor, depending on the lanes under measurement. Urban freeway sensors are integrated in 300 measuring stations placed in the main beltway of the city, the so-called M-30 highway, and they are able to characterize vehicles and measure speed. The rest of them count vehicles passing through all the lanes covered by the measuring station and aggregate this tally to provide the following data:

- **Flow**: Vehicles per hour that pass through the measuring station, extrapolating linearly to one hour the computed vehicles during the measuring time frame.
- **Occupancy**: Percentage of occupancy of the control point, obtained from the number of vehicles and the maximum capacity of that particular segment of the road.
- **Load**: Parameter obtained as a function of flow and occupancy, as well as the infrastructure features, ranging from 0 (no load) to 100 (fully loaded road).
- **Level of Service**: Qualitative measure obtained from the previous ones that allows for an intelligible interpretation of the other metrics.

Data collected by these measuring stations are published in a live feed every minute in Madrid Open Data portal[1], and posted historically in the form of 15-minute aggregated periods. Using the fine-grained data requires the implementation of a tool to capture them and to run it during a period as long as the desired extension of data. This was performed for the experiments conducted in Chapter 4. In the rest of experiments, historical 15-minute data were used.

[1] `http://datos.madrid.es`

Bibliography

[1] B. Van Arem, H. R. Kirby, M. J. M. Van Der Vlist, and J. C. Whittaker, "Recent advances and applications in the field of short-term traffic forecasting", *International Journal of Forecasting*, vol. 13, no. 1, pp. 1–12, 1997.

[2] R. J. Weiland and L. B. Purser, "Intelligent transportation systems", *Transportation in the new millennium*, 2000.

[3] M. S. Ahmed and A. R. Cook, *Analysis of freeway traffic time-series data by using Box-Jenkins techniques*, 722. 1979.

[4] M. Levin and Y. Tsao, "On forecasting freeway occupancies and volumes", *Transportation Research Record*, no. 773, pp. 47–49, 1980.

[5] C. K. Moorthy and B. G. Ratcliffe, "Short term traffic forecasting using time series methods", *Transportation Planning and Technology*, vol. 12, no. 1, pp. 45–56, 1988.

[6] N. L. Nihan and K. O. Holmesland, "Use of the Box and Jenkins time series technique in traffic forecasting", *Transportation*, vol. 9, no. 2, pp. 125–143, 1980.

[7] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory", *Transportation Research Part B: Methodological*, vol. 18, no. 1, pp. 1–11, 1984.

[8] E. I. Vlahogianni, J. C. Golias, and M. G. Karlaftis, "Short−term traffic forecasting: Overview of objectives and methods", *Transport Reviews*, vol. 24, no. 5, pp. 533–557, 2004.

[9] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights", *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 3, pp. 387–399, 2011.

[10] F. Su, H. Dong, L. Jia, Y. Qin, and Z. Tian, "Long-term forecasting oriented to urban expressway traffic situation", *Advances in Mechanical Engineering*, vol. 8, no. 1, pp. 1–16, 2016.

[11] C. Lamboley, J. C. Santucci, and M. Danech-Pajouh, "24 or 48 Hour Advance Traffic Forecast in Urban and Periurban Environment: The example of Paris", in *4th World Congress on Intelligent Transport Systems, Mobility for Everyone*, 1997.

[12] C. van Hinsbergen, J. van Lint, and F. Sanders, "Short term traffic prediction models", in *14th World Congress on Intelligent Transportation Systems (ITS)*, 2007.

[13] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where we're going", *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 3–19, 2014.

[14] P. Næss and A. Strand, "Traffic forecasting at 'strategic', 'tactical' and 'operational' level", *Journal of Critical Realism*, vol. 51, no. 2, pp. 41–48, 2015.

[15] C. Kai, K. Zhang, and Y. Hamamatsu, "Traffic Information Real-Time Monitoring Based on A Short-Long Term Algorithm", in *IEEE International Conference on Systems, Man, and Cybernetics*, 2006, pp. 651–656.

[16] H. J. van Lint and C. P. van Hinsbergen, "Short-term traffic and travel time prediction models", *Artificial Intelligence Applications to Critical Transportation Issues*, vol. 22, pp. 22–41, 2012.

[17] A. Bhaskar, E. Chung, and A. G. Dumont, "Fusing loop detector and probe vehicle data to estimate travel time statistics on signalized urban networks", *Computer-Aided Civil and Infrastructure Engineering*, vol. 26, no. 6, pp. 433–450, 2011.

[18] E. Bolshinsky and R. Friedman, "Traffic flow forecast survey", Computer Science Department, Technion, Tech. Rep., 2012.

[19] L. Dimitriou, T. Tsekeris, and A. Stathopoulos, "Adaptive hybrid fuzzy rule-based system approach for modeling and predicting urban traffic flow", *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 5, pp. 554–573, 2008.

[20] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights", *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 3, pp. 387–399, 2011.

[21] G. Marfia and M. Roccetti, "Vehicular congestion detection and short-term forecasting: A new model with results", *IEEE Transactions on Vehicular Technology*, vol. 60, no. 7, pp. 2936–2948, 2011.

[22] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach", *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 3, pp. 211–234, 2005.

[23] H. Lin, R. Zito, and M. Taylor, "A review of travel-time prediction in transport and logistics", in *Eastern Asia Society for Transportation Studies*, vol. 5, 2005, pp. 1433–1448.

[24] S. Cohen and Z. Christoforou, "Travel time estimation between loop detectors and FCD: A compatibility study on the Lille Network, France", *Transportation Research Procedia*, vol. 10, pp. 245–255, 2015.

[25] X. Fei, C. C. Lu, and K. Liu, "A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction", *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 1306–1318, 2011.

[26] C. J. Lu, "An adaptive system for predicting freeway travel times", *International Journal of Information Technology & Decision Making*, vol. 11, no. 04, pp. 727–747, 2012.

[27] F. Zheng and H. Van Zuylen, "Urban link travel time estimation based on sparse probe vehicle data", *Transportation Research Part C: Emerging Technologies*, vol. 31, pp. 145–157, 2013.

[28] L. Moreira-Matias, J. Mendes-Moreira, J. de Sousa, and J. Gama, "Improving mass transit operations by using AVL-based systems: a survey", *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 1636–1653, 2015.

[29] E. Castillo, Z. Grande, A. Calviño, W. Y. Szeto, and H. K. Lo, "A state-of-the-art review of the sensor location, flow observability, estimation, and prediction problems in traffic networks", *Journal of Sensors*, vol. 2015, 2015.

[30] U. Mori, A. Mendiburu, M. Álvarez, and J. A. Lozano, "A review of travel time estimation and forecasting for Advanced Traveller Information Systems", *Transportmetrica A: Transport Science*, vol. 11, no. 2, pp. 119–157, 2015.

[31] S. Oh, Y. Byon, K. Jang, and H. Yeo, "Short-term travel-time prediction on highway: A review of the data-driven approach", *Transport Reviews*, vol. 35, no. 1, pp. 4–32, 2015.

[32] A. Stathopoulos and M. G. Karlaftis, "A multivariate state space approach for urban traffic flow modeling and prediction", *Transportation Research Part C: Emerging Technologies*, vol. 11, no. 2, pp. 121–135, 2003.

[33] R. Chrobok, O. Kaumann, J. Wahle, and M. Schreckenberg, "Different methods of traffic forecast based on real data", *European Journal of Operational Research*, vol. 155, no. 3, pp. 558–568, 2004.

[34] I. Laña, J. Del Ser, and I. Olabarrieta, "Understanding daily mobility patterns in urban road networks using traffic flow analytics", in *IEEE Network Operations and Management Symposium (NOMS)*, 2016.

[35] N. E. Faouzi, H. Leung, and A. Kurian, "Data fusion in intelligent transportation systems: Progress and challenges – A survey", *Information Fusion*, vol. 12, no. 1, pp. 4–10, 2011.

[36] N. E. Faouzi and L. A. Klein, "Data Fusion for ITS: Techniques and research needs", *Transportation Research Procedia*, vol. 15, pp. 495–512, 2016.

[37] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Y. Wang, "Traffic flow prediction with Big Data: A deep learning approach", *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2015.

[38] S. Hoogendoorn and P. Bovy, "State-of-the-art of vehicular traffic flow modelling", *Journal of Systems and Control Engineering*, vol. 4, no. 215, pp. 283–303, 2001.

[39] Y. Kamarianakis and P. Prastacos, "Forecasting traffic flow conditions in an urban network: Comparison of multivariate and univariate approaches", *Transportation Research Record*, vol. 1857, no. 1, pp. 74–84, 2003.

[40] C. Zhang, S. Sun, and G. Yu, "A bayesian network approach to time series forecasting of short-term traffic flows", in *7th International IEEE Conference on Intelligent Transportation Systems (ITS)*, 2004, pp. 216–221.

[41] H. Yin, S. Wong, J. Xu, and C. Wong, "Urban traffic flow prediction using a fuzzy-neural approach", *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 2, pp. 85–98, 2002.

[42] M. Annunziato and S. Pizzuti, "A smart-adaptive-system based on evolutionary computation and neural networks for the on-line short-term urban traffic prediction", in *European Symposium on Intelligent Technologies, Hybrid Systems and their Implementation on Smart Adaptive Systems (EUNITE)*, 2004.

[43] H. Dia, "An object-oriented neural network approach to short-term traffic forecasting", *European Journal of Operational Research*, vol. 131, no. 2, pp. 253–261, 2001.

[44] W. C. Hong, Y. Dong, F. Zheng, and C. Y. Lai, "Forecasting urban traffic flow by SVR with continuous ACO", *Applied Mathematical Modelling*, vol. 35, no. 3, pp. 1282–1291, 2011.

[45] E. I. Vlahogianni, "Optimization of traffic forecasting: Intelligent surrogate modeling", *Transportation Research Part C: Emerging Technologies*, vol. 55, pp. 14–23, 2015.

[46] B. L. Smith, B. M. Williams, and R. K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting", *Transportation Research Part C: Emerging Technologies*, vol. 10, pp. 303–321, 2002.

[47] D. I. Tselentis, E. I. Vlahogianni, and M. G. Karlaftis, "Improving short-term traffic forecasts: To combine models or not to combine?", *IET Intelligent Transport Systems*, vol. 9, no. 2, pp. 193–201, 2014.

[48] P. Cai, Y. Wang, G. Lu, P. Chen, C. Ding, and J. Sun, "A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting", *Transportation Research Part C: Emerging Technologies*, vol. 62, pp. 21–34, 2016.

[49] J. Zhong and S. Ling, "Key factors of K-nearest neighbors non-parametric regression in short-time traffic flow forecasting", in *21st International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2014, pp. 9–12.

[50] P. Dell'acqua, F. Bellotti, R. Berta, and A. De Gloria, "Time-aware multivariate nearest neighbor regression methods for traffic flow prediction", *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3393–3402, 2015.

[51] D. Xia, B. Wang, H. Li, Y. Li, and Z. Zhang, "A distributed spatial-temporal weighted model on MapReduce for short-term traffic flow forecasting", *Neurocomputing*, vol. 179, pp. 246–26, 2016.

[52] A. Abadi, T. Rajabioun, and P. A. Ioannou, "Networks with limited traffic data", *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 653–662, 2015.

[53] J. Xu, D. Deng, U. Demiryurek, C. Shahabi, and M. Van Der Schaar, "Mining the situation: spatiotemporal traffic prediction with Big Data", *IEEE Journal on Selected Topics in Signal Processing*, vol. 9, no. 4, pp. 702–715, 2015.

[54] J. Wang, I. Tsapakis, and C. Zhong, "A space-time delay neural network model for travel time prediction", *Engineering Applications of Artificial Intelligence*, vol. 52, pp. 145–160, 2016.

[55] Y. Wu, F. Chen, C. Lu, and S. Yang, "Urban traffic flow prediction using a Spatio-Temporal Random Effects model", *Journal of Intelligent Transportation Systems. Technology, Planning, and Operations*, vol. 2450, pp. 1–12, 2015.

[56] A. Ladino, A. Kibangou, H. Fourati, and C. C. de Wit, "Travel time forecasting from clustered time series via optimal fusion strategy", in *European Control Conference (ECC)*, 2016.

[57] D. J. Dailey, "Travel-time estimation using cross-correlation techniques", *Transportation Research Part B: Methodological*, vol. 27, no. 2, pp. 97–107, 1993.

[58] C. J. Lan and G. A. Davis, "Real-time estimation of turning movement proportions from partial counts on urban networks", *Transportation Research Part C: Emerging Technologies*, vol. 7, no. 5, pp. 305–327, 1999.

[59] B. S. Westgate, D. B. Woodard, D. S. Matteson, and S. G. Henderson, "Large-network travel time distribution estimation for ambulances", *European Journal of Operational Research*, vol. 252, no. 1, pp. 322–333, 2016.

[60] A. Bezuglov and G. Comert, "Short-term freeway traffic parameter prediction: Application of grey system theory models", *Expert Systems with Applications*, vol. 62, pp. 284–292, 2016.

[61] A. Salamanis, D. D. Kehagias, C. K. Filelis-Papadopoulos, D. Tzovaras, and G. A. Gravvanis, "Managing spatial graph dependencies in large volumes of traffic data for travel-time prediction", *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 6, pp. 1678–1687, 2016.

[62] Y. Byeong-Seo, S.-P. Kang, and C.-H. Park, "Travel time estimation using mobile data", in *Proceedings of Eastern Asia Society for Transportation Studies*, Citeseer, vol. 5, 2005, pp. 1533–1547.

[63] H. Yang, T. Dillon, and Y. P. Chen, "Evaluation of recent computational approaches in short-term traffic forecasting", in *4th IFIP International Conference on Artificial Intelligence*, vol. 465, 2015, pp. 108–116.

[64] S. Oh, Y. Kim, and J. Hong, "Urban traffic flow prediction system using a multifactor pattern recognition model", *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2744–2755, 2015.

[65] A. Wibisono, W. Jatmiko, H. A. Wisesa, B. Hardjono, and P. Mursanto, "Traffic Big Data prediction and visualization using Fast Incremental Model Trees-Drift Detection (FIMT-DD)", *Knowledge-Based Systems*, vol. 93, pp. 33–46, 2016.

[66] F. Jin and S. Sun, "Neural network multitask learning for traffic flow forecasting", in *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1897–1901.

[67] K. Jha, N. Sinha, S. Arkatkar, and A. K. Sarkar, "A comparative study on application of time series analysis for traffic forecasting in India : prospects and limitations", *Current Science*, vol. 110, no. 3, pp. 373–385, 2016.

[68] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data", *Transportation Research Part C: Emerging Technologies*, vol. 54, pp. 187–197, 2015.

[69] L. Li, X. Su, and Y. Zhang, "Trend modeling for traffic time series analysis : An integrated study", *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 1–10, 2015.

[70] Z. Hou and X. Li, "Repeatability and similarity of freeway traffic flow and long-term prediction under Big Data", *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 6, pp. 1786–1796, 2016.

[71] X. He, X. Gao, Y. Zhang, Z. H. Zhou, Z. Y. Liu, B. Fu, F. Hu, and Z. Zhang, "Research of traffic flow forecasting based on the information fusion of BP network sequence", in *International Conference on Intelligent Science and Big Data Engineering*, vol. 9243, 2015, pp. 548–558.

[72] H. Pan, J. Liu, S. Zhou, and Z. Niu, "A block regression model for short-term mobile traffic forecasting", in *IEEE/CIC ICCC 2015 Symposium on Next Generation Networking*, 2015, pp. 1–5.

[73] L. Li, X. Su, Y. Wang, Y. Lin, Z. Li, and Y. Li, "Robust causal dependence mining in big data network and its application to traffic flow predictions", *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 292–307, 2015.

[74] J. Wan, J. Liu, Z. Shao, A. V. Vasilakos, M. Imran, and K. Zhou, "Mobile crowd sensing for traffic prediction in internet of vehicles", *Sensors*, vol. 16, no. 88, pp. 1–15, 2016.

[75] F. Moretti, S. Pizzuti, S. Panzieri, and M. Annunziato, "Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling", *Neurocomputing*, vol. 167, pp. 3–7, 2015.

[76] A. Csikós, Z. J. Viharos, K. B. Kis, T. Tettamanti, and I. Varga, "Traffic speed prediction method for urban networks - an ANN approach", in *Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 2015, pp. 102–108.

[77] K. Kumar, M. Parida, and V. K. Katiyar, "Short term traffic flow prediction in heterogeneous condition using artificial neural network", *Transport*, vol. 30, no. 4, pp. 397–405, 2014.

[78] S. V. Kumar and L. Vanajakshi, "Short-term traffic flow prediction using seasonal ARIMA model with limited input data", *European Transportation Research Review*, vol. 7, no. 21, pp. 1–9, 2015.

[79] X. Niu, Y. Zhu, Q. Cao, X. Zhang, W. Xie, and K. Zheng, "An online-traffic-prediction based route finding mechanism for smart city", *International Journal of Distributed Sensor Networks*, vol. 2015, pp. 1–16, 2015.

[80] M. Meng, S. Chun-fu, W. Yiik-diew, W. Bo-bin, and L. I. Hui-xuan, "A two-stage short-term traffic flow prediction method based on AVL and AKNN techniques", *Journal of Central South University*, vol. 22, pp. 779–786, 2015.

[81] M. Hui, L. Bai, Y. Li, and Q. Wu, "Highway traffic flow nonlinear character analysis and prediction", *Mathematical Problems in Engineering*, vol. 2015, pp. 1–8, 2015.

[82] J. Abdi and B. Moshiri, "Application of temporal difference learning rules in short-term traffic flow prediction", *Expert Systems*, vol. 32, no. 1, pp. 49–64, 2015.

[83] M. Huang, "Intersection traffic flow forecasting based on ν -GSVR with a new hybrid evolutionary algorithm", *Neurocomputing*, vol. 147, pp. 343–349, 2015.

[84] C. Dong, Z. Xiong, C. Shao, and H. Zhang, "A spatial temporal based state space approach for freeway network traffic flow modelling and prediction", *Transportmetrica A: Transport Science*, vol. 11, no. 7, pp. 547–560, 2015.

[85] Z. Zhu, B. Peng, C. Xiong, and L. Zhang, "Short-term traffic flow prediction with linear conditional Gaussian Bayesian network", *Journal of Advanced Transportation*, pp. 1–13, 2016.

[86] J. Zhao and S. Sun, "High-order gaussian process dynamical models for traffic flow prediction", *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 2014–2019, 2016.

[87] K. Balsys and D. Eidukas, "Traffic flow detection and forecasting", *Electronics and Electrical Engineering*, vol. 5, no. 5, pp. 5–8, 2010.

[88] L. Si-yan, L. De-wei, X. Yu-geng, and T. Qi-feng, "A short-term traffic flow forecasting method and its applications", *Journal of Shanghai Jiaotong University (Science)*, vol. 20, no. 2, pp. 156–163, 2015.

[89] J. Mendes-moreira, A. M. Jorge, J. Freire de Sousa, and C. Soares, "Improving the accuracy of long-term travel time prediction using heterogeneous ensembles", *Neurocomputing*, vol. 150, pp. 428–439, 2015.

[90] J. Tang, Y. Zou, J. Ash, S. Zhang, F. Liu, and Y. Wang, "Travel time estimation using freeway point detector data based on evolving fuzzy neural inference system", *PloS one*, vol. 11, no. 2, pp. 1–24, 2016.

[91] Y. Wu, H. Tan, P. Jin, B. Shen, and B. Ran, "Short-term traffic flow prediction based on multilinear analysis and k-nearest neighbor regression", in *15th International Conference on Transportation Professionals (CICTP-2015)*, 2015, pp. 557–569.

[92] J. Rupnik, J. Davies, B. Fortuna, A. Duke, and S. S. Clarke, "Travel time prediction on highways", in *IEEE International Conference on Computer and Information Technology*, 2015, pp. 1435–1442.

[93] Y. Hou, P. Edara, and C. Sun, "Traffic flow forecasting for urban work zones", *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 1761–1770, 2015.

[94] Y. Chun-xia, F. Rui, and F. Yi-qin, "Prediction of short-term traffic flow based on similarity", *Journal of Highway and Transportation Research and Development*, vol. 10, no. 1, pp. 92–97, 2016.

[95] Y. Zhang and Y. Zhang, "A comparative study of three multivariate short-term freeway traffic flow forecasting methods with missing data", *Journal of Intelligent Transportation Systems*, vol. 20, no. 3, pp. 205–218, 2016.

[96] X. Li and W. Gao, "Prediction of traffic flow combination model based on data mining", *International Journal of Database Theory and Application*, vol. 8, no. 6, pp. 303–312, 2015.

[97] H. Lu, Z. Sun, W. Qu, and L. Wang, "Real-time corrected traffic correlation model for traffic flow forecasting", *Mathematical Problems in Engineering*, vol. 2015, pp. 1–11, 2015.

[98] W. Ma and R. Wang, "Traffic flow forecasting research based on bayesian normalized Elman neural network", in *IEEE Signal Processing and Signal Processing Education Workshop (SP/SPE)*, 2015, pp. 426–430.

[99] R. T. Das, K. K. Ang, and C. Quek, "ieRSPOP: A novel incremental rough set-based pseudo outer-product with ensemble learning", *Applied Soft Computing*, vol. 46, pp. 170–186, 2016.

[100] J. Melorose, R. Perroy, and S. Careas, "Analysis and prediction of spatiotemporal impact of traffic incidents for better mobility and safety in transportation systems", Tech. Rep., 2015.

[101] T. Ma, Z. Zhou, and B. Abdulhai, "Nonlinear multivariate time-space threshold vector error correction model for short term traffic state prediction", *Transportation Research Part B: Methodological*, vol. 76, pp. 27–47, 2015.

[102] L. Moreira-Matias and F. Alesiani, "Drift3Flow: Freeway incident prediction using real-time learning", in *IEEE 18th International Conference on Intelligent Transportation Systems*, 2015, pp. 566–571.

[103] P. Lopez-Garcia, E. Onieva, E. Osaba, A. D. Masegosa, and A. Perallos, "A hybrid method for short-term traffic congestion forecasting using genetic algorithms and cross entropy", *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 557–569, 2016.

[104] G. Fusco, C. Colombaroni, L. Comelli, and N. Isaenko, "Short-term traffic predictions on large urban traffic networks : applications of network-based machine learning models and dynamic traffic assignment models", in *2015 Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 2015, pp. 3–5.

[105] N. N. Yusof, A. Mohamed, and S. Abdul-Rahman, "Soft computing in data science", *Communications in Computer and Information Science*, vol. 545, pp. 43–53, 2015.

[106] Y. Cong, J. Wang, and X. Li, "Traffic flow forecasting by a least squares support vector machine with a fruit fly optimization algorithm", *Procedia Engineering*, vol. 137, pp. 59–68, 2016.

[107] S. Jeon and B. Hong, "Monte Carlo simulation-based traffic speed forecasting using historical big data", *Future Generation Computer Systems*, 2015.

[108] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction", *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 308–324, 2014.

[109] H. Yang and X. Hu, "Wavelet neural network with improved genetic algorithm for traffic flow time series prediction", *Optik*, vol. 127, pp. 8103–8110, 2016.

[110] I. Laña, J. Del Ser, M. Vélez, and I. Oregi, "Joint feature selection and parameter tuning for short-term traffic flow forecasting based on heuristically optimized multi-layer neural networks", in *International Conference on Harmony Search Algorithm*, 2017, pp. 91–100.

[111] Z. Ma, G. Luo, and D. Huang, "Short term traffic flow prediction based on on-line sequential extreme learning machine", in *8th International Conference on Advanced Computational Intelligence (ICACI)*, 2016, pp. 143–149.

[112] G. Zhu, K. Song, and P. Zhang, "A travel time prediction method for urban road traffic sensors data", in *International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI)*, 2015, pp. 29–32.

[113] L. Moreira-Matias, O. Cats, J. Gama, J. Mendes-Moreira, and J. F. de Sousa, "An online learning approach to eliminate bus bunching in real-time", *Applied Soft Computing*, vol. 47, pp. 460–482, 2016.

[114] Y. Xu, H. Chen, Q. Kong, X. Zhai, and Y. Liu, "Urban traffic flow prediction: A spatio-temporal variable selection-based approach", *Journal of Advanced Transportation*, vol. 50, pp. 489–506, 2016.

[115] I. Laña, J. Del Ser, A. Padró, M. Vélez, and C. Casanova-Mateo, "The role of local urban traffic and meteorological conditions in air pollution: A data-based case study in Madrid, Spain", *Atmospheric Environment*, vol. 145, pp. 424–438, 2016.

[116] H. Park and A. Haghani, "Real-time prediction of secondary incident occurrences using vehicle probe data", *Transportation Research Part C: Emerging Technologies*, 2014.

[117] F. C. Pereira, F. Rodrigues, and M. Ben-akiva, "Using data from the web to predict public transport arrivals under special events scenarios", *Journal of Intelligent Transportation System*, vol. 1, no. April 2015, pp. 37–41, 2015.

[118] S. C. Calvert, J. W. C. Van Lint, and S. P. Hoogendoorn, "A hybrid travel time prediction framework for planned motorway roadworks", in *IEEE Conference on Intelligent Transportation Systems (ITS)*, Madeira Island, 2010, pp. 1770–1776.

[119] S. Yousif, "Motorway roadworks: effects on traffic operations", *Highways and Transportation*, pp. 20–22, 2002.

[120] Z. Zhou, X. Yang, and P. Zheng, "A method for delay estimation using traffic monitoring data", in *International Conference on Automatic Control and Artificial Intelligence (ACAI)*, 2012, pp. 108–111.

[121] I. Žliobaitė, M. Pechenizkiy, and J. Gama, "An overview of concept drift applications", in *Big Data Analysis: New Algorithms for a New Society*, Springer, 2016, pp. 91–114.

[122] G. Souto and T. Liebig, *On event detection from spatial time series for urban traffic applications*, 9580. 2016, pp. 217–230.

[123] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Statistical methods for detecting nonlinearity and non-stationarity in univariate short-term time-series of traffic volume", *Transportation Research Part C: Emerging Technologies*, vol. 14, no. 5, pp. 351–367, 2006.

[124] Y. Kamarianakis, H. O. Gao, and P. Prastacos, "Characterizing regimes in daily cycles of urban traffic using smooth-transition regressions", *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 5, pp. 821–840, 2010.

[125] E. I. Vlahogianni and M. G. Karlaftis, "Temporal aggregation in traffic data: Implications for statistical characteristics and model choice", *Transportation Letters*, vol. 3, no. 1, pp. 37–49, 2011.

[126] J. Tang, Y. Wang, H. Wang, S. Zhang, and F. Liu, "Dynamic analysis of traffic time series at different temporal scales: A complex networks approach", *Physica A: Statistical Mechanics and its Applications*, vol. 405, pp. 303–315, 2014.

[127] R. Elwell and R. Polikar, "Incremental learning of concept drift in nonstationary environments.", *IEEE Transactions on Neural Networks*, vol. 22, no. 10, pp. 1517–31, 2011.

[128] E. I. Vlahogianni, *Computational intelligence and optimization for transportation Big Data: Challenges and opportunities*. Springer International Publishing, 2015, pp. 107–128.

[129] S. Gábor and I. Csabai, "The analogies of highway and computer network traffic", *Physica A: Statistical Mechanics and its Applications*, vol. 307, no. 3, pp. 516–526, 2002.

[130] S. Clement, N. Vogiatzis, M. Daniel, S. Wilson, and R. Clegg, "Road networks of the future: Can data networks help?", in *28th Australasian Transport Research Forum*.

[131] J. Baber, J. Kolodko, T. Noel, M. Parent, and L. Vlacic, "Cooperative autonomous driving: Intelligent vehicles sharing city roads", *IEEE Robotics & Automation Magazine*, vol. 12, no. 1, pp. 44–49, 2005.

[132] H. Feng and Y. Shu, "Study on network traffic prediction techniques", in *International Conference on Wireless Communications, Networking and Mobile Computing*, vol. 2, 2005, pp. 995–998.

[133] A. Sang and S. Li, "A predictability analysis of network traffic", *Computer Networks*, vol. 39, no. 4, pp. 329–345, 2002.

[134] Y. Chen, B. Yang, and Q. Meng, "Small-time scale network traffic prediction based on flexible neural tree", *Applied Soft Computing Journal*, vol. 12, no. 1, pp. 274–279, 2012.

[135] S. Salcedo-Sanz, D. Manjarres, Á. Pastor-Sánchez, J. Del Ser, J. A. Portilla-Figueras, and S. Gil-Lopez, "One-way urban traffic reconfiguration using a multi-objective harmony search approach", *Expert Systems with Applications*, vol. 40, no. 9, pp. 3341–3350, 2013.

[136] M. A. Abdel-Aty, R. Kitamura, and P. P. Jovanis, "Using stated preference data for studying the effect of advanced traffic information on drivers' route choice", *Transportation Research Part C: Emerging Technologies*, vol. 5, no. 1, pp. 39–50, 1997.

[137] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Spatio temporal short-term urban traffic volume forecasting using genetically optimized modular networks", *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, no. 5, pp. 317–325, 2007.

[138] F. Schimbinschi, X. V. Nguyen, J. Bailey, C. Leckie, H. Vu, and R. Kotagiri, "Traffic forecasting in complex urban networks: Leveraging big data and machine learning", in *IEEE International Conference on Big Data*, 2015, pp. 1019–1024.

[139] J. Van Lint, S. Hoogendoorn, and H. J. van Zuylen, "Accurate freeway travel time prediction with state-space neural networks under missing data", *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 5, pp. 347–369, 2005.

[140] M. Zhong, P. Lingras, and S. Sharma, "Estimation of missing traffic counts using factor, genetic, neural, and regression techniques", *Transportation Research Part C: Emerging Technologies*, vol. 12, no. 2, pp. 139–166, 2004.

[141] H. Chen, S. Grant-Muller, L. Mussone, and F. Montgomery, "A study of hybrid neural network approaches and the effects of missing data on traffic forecasting", *Neural Computing & Applications*, vol. 10, no. 3, pp. 277–286, 2001.

[142] S. Sun, G. Yu, and C. Zhang, "Short-term traffic flow forecasting using sampling markov chain method with incomplete data", in *IEEE Intelligent Vehicles Symposium*, 2004, pp. 437–441.

[143] L. Li, Y. Li, and Z. Li, "Efficient missing data imputing for traffic flow by considering temporal and spatial dependence", *Transportation Research Part C: emerging technologies*, vol. 34, pp. 108–120, 2013.

[144] J. L. Schafer, *Analysis of incomplete multivariate data*. CRC press, 1997.

[145] A. M. Moffat, D. Papale, M. Reichstein, D. Y. Hollinger, A. D. Richardson, A. G. Barr, C. Beckstein, B. H. Braswell, G. Churkina, A. R. Desai, *et al.*, "Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes", *Agricultural and Forest Meteorology*, vol. 147, no. 3, pp. 209–232, 2007.

[146] D. Kondrashov and M. Ghil, "Spatio-temporal filling of missing points in geophysical data sets", *Nonlinear Processes in Geophysics*, vol. 13, no. 2, pp. 151–159, 2006.

[147] F. M. Shrive, H. Stuart, H. Quan, and W. A. Ghali, "Dealing with missing data in a multi-question depression scale: A comparison of imputation methods", *BMC medical research methodology*, vol. 6, no. 1, p. 57, 2006.

[148] K. L. Sainani, "Dealing with missing data", *PM&R*, vol. 7, no. 9, pp. 990–994, 2015.

[149] F. Arteaga and A. Ferrer, "Dealing with missing data in MSPC: several methods, different interpretations, some examples", *Journal of Chemometrics*, vol. 16, no. 8-10, pp. 408–418, 2002.

[150] J. Sterne, I. White, J. Carlin, M. Spratt, P. Royston, M. Kenward, A. Wood, and J. Carpenter, "Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls", *British Medical Journal*, vol. 339, pp. 157–160, 2009.

[151] B. Smith, W. Scherer, and J. Conklin, "Exploring imputation techniques for missing data in transportation management systems", *Transportation Research Record: Journal of the Transportation Research Board*, no. 1836, pp. 132–142, 2003.

[152] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[153] K. Henrickson, Y. Zou, and Y. Wang, "Flexible and robust method for missing loop detector data imputation", *Transportation Research Record: Journal of the Transportation Research Board*, no. 2527, pp. 29–36, 2015.

[154] L. Qu, L. Li, Y. Zhang, and J. Hu, "PPCA-based missing data imputation for traffic flow volume: A systematical approach", *IEEE Transactions on intelligent transportation systems*, vol. 10, no. 3, pp. 512–522, 2009.

[155] C. Chen, Y. Wang, L. Li, J. Hu, and Z. Zhang, "The retrieval of intra-day trend and its influence on traffic prediction", *Transportation Research Part C: Emerging Technologies*, vol. 22, pp. 103–118, 2012.

[156] J.-M. Chiou, Y.-C. Zhang, W.-H. Chen, and C.-W. Chang, "A functional data approach to missing value imputation and outlier detection for traffic flow data", *Transportmetrica B: Transport Dynamics*, vol. 2, no. 2, pp. 106–129, 2014.

[157] M. T. Asif, N. Mitrovic, J. Dauwels, and P. Jaillet, "Matrix and tensor based methods for missing data estimation in large traffic networks", *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 1816–1825, 2016.

[158] B. Ran, H. Tan, Y. Wu, and P. J. Jin, "Tensor based missing traffic data completion with spatial–temporal correlation", *Physica A: Statistical Mechanics and its Applications*, vol. 446, pp. 54–63, 2016.

[159] Y. Li, Z. Li, and L. Li, "Missing traffic data: Comparison of imputation methods", *IET Intelligent Transport Systems*, vol. 8, no. 1, pp. 51–57, 2014.

[160] L. Li, X. Su, Y. Zhang, J. Hu, and Z. Li, "Traffic prediction, data compression, abnormal data detection and missing data imputation: An integrated study based on the decomposition of traffic time series", in *IEEE 17th International Conference on Intelligent Transportation Systems (ITS)*, 2014, pp. 282–289.

[161] Y. Bie, X. Wang, and T. Z. Qiu, "Online method to impute missing loop detector data for urban freeway traffic control", *Transportation Research Record: Journal of the Transportation Research Board*, no. 2593, pp. 37–46, 2016.

[162] M. Zhong, S. Sharma, and P. Lingras, "Matching patterns for updating missing values of traffic counts", *Transportation Planning and Technology*, vol. 29, no. 2, pp. 141–156, 2006.

[163] H. Tan, Y. Wu, B. Cheng, W. Wang, and B. Ran, "Robust missing traffic flow imputation considering nonnegativity and road capacity", *Mathematical Problems in Engineering*, vol. 2014, 2014.

[164] J. Haworth and T. Cheng, "Non-parametric regression for space–time forecasting under missing data", *Computers, Environment and Urban Systems*, vol. 36, no. 6, pp. 538–550, 2012.

[165] M. Zhong, S. Sharma, and P. Lingras, "Genetically designed models for accurate imputation of missing traffic counts", *Transportation Research Record: Journal of the Transportation Research Board*, no. 1879, pp. 71–79, 2004.

[166] J. Tang, G. Zhang, Y. Wang, H. Wang, and F. Liu, "A hybrid approach to integrate fuzzy c-means based imputation method with genetic algorithm for missing traffic volume data estimation", *Transportation Research Part C: Emerging Technologies*, vol. 51, pp. 29–40, 2015.

[167] W. C. Ku, G. R. Jagadeesh, A. Prakash, and T. Srikanthan, "A clustering-based approach for data-driven imputation of missing traffic data", in *IEEE Forum onIntegrated and Sustainable Transportation Systems (FISTS)*, 2016, pp. 1–6.

[168] Y. Duan, Y. Lv, Y.-L. Liu, and F.-Y. Wang, "An efficient realization of deep learning for traffic data imputation", *Transportation Research Part C: emerging technologies*, vol. 72, pp. 168–181, 2016.

[169] M. E. Whitlock and C. M. Queen, "Modelling a traffic network with missing data", *Journal of Forecasting*, vol. 19, no. 7, pp. 561–574, 2000.

[170] S. Sun, C. Zhang, and G. Yu, "A bayesian network approach to traffic flow forecasting", *IEEE Transactions on intelligent transportation systems*, vol. 7, no. 1, pp. 124–132, 2006.

[171] S. Sun and C. Zhang, "The selective random subspace predictor for traffic flow forecasting", *IEEE Transactions on intelligent transportation systems*, vol. 8, no. 2, pp. 367–373, 2007.

[172] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks", in *IEEE International Joint Conference on Neural Networks*, vol. 2, 2004, pp. 985–990.

[173] M. Treiber and D. Helbing, "Reconstructing the spatio-temporal traffic dynamics from stationary detector data", *Cooperative Transportation Dynamics*, vol. 1, no. 3, pp. 3–1, 2002.

[174] A. Abadi, T. Rajabioun, and P. A. Ioannou, "Traffic flow prediction for road transportation networks with limited traffic data", *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 653–662, 2015.

[175] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[176] E. Falge, D. Baldocchi, R. Olson, P. Anthoni, M. Aubinet, C. Bernhofer, G. Burba, R. Ceulemans, R. Clement, H. Dolman, *et al.*, "Gap filling strategies for long term energy flux data sets", *Agricultural and Forest Meteorology*, vol. 107, no. 1, pp. 71–77, 2001.

[177] S. Sun, C. Zhang, G. Yu, N. Lu, and F. Xiao, "Bayesian network methods for traffic flow forecasting with incomplete data", in *European Conference on Machine Learning (ECML)*, Springer, 2004, pp. 419–428.

[178] C. Chen, J. Kwon, J. Rice, A. Skabardonis, and P. Varaiya, "Detecting errors and imputing missing data for single-loop surveillance systems", *Transportation Research Record: Journal of the Transportation Research Board*, no. 1855, pp. 160–167, 2003.

[179] I. Laña, J. Del Ser, and M. Vélez, "A novel fireworks algorithm with wind inertia dynamics and its application to traffic forecasting", in *IEEE Congress on Evolutionary Computation (CEC)*, 2017, pp. 706–713.

[180] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise", in *KDD*, vol. 96, 1996, pp. 226–231.

[181] B. J. Frey and D. Dueck, "Clustering by passing messages between data points", *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[182] T. K. Ho, "Random decision forest", in *3rd International Conference on Document Analysis and Recognition*, 1995, pp. 278–282.

[183] L. Breiman, "Random forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[184] F. B. Hildebrand, *Introduction to numerical analysis*. Courier Corporation, 1987.

[185] F. Crawford, D. P. Watling, and R. D. Connors, "A statistical method for estimating predictable differences between daily traffic flow profiles", *Transportation Research Part B: Methodological*, vol. 95, pp. 196–213, 2017.

[186] I. M. Wagner-Muns, I. G. Guardiola, V. Samaranayke, and W. I. Kayani, "A functional data analysis approach to traffic volume forecasting", *IEEE Transactions on Intelligent Transportation Systems*, 2017.

[187] A. M. Andrew, "Spiking neuron models: Single neurons, populations, plasticity", *Kybernetes*, vol. 32, no. 7/8, 2003.

[188] F. Ponulak and A. Kasinski, "Introduction to spiking neural networks: Information processing, learning and applications", *Acta Neurobiologiae Experimentalis*, vol. 71, no. 4, pp. 409–433, 2011.

[189] S. G. Wysoski, L. Benuskova, and N. Kasabov, "Adaptive learning procedure for a network of spiking neurons and visual pattern recognition", in *International Conference on Advanced Concepts for Intelligent Vision Systems*, Springer, 2006, pp. 1133–1142.

[190] S. Kulkarni, S. P. Simon, and K. Sundareswaran, "A spiking neural network (SNN) forecast engine for short-term electrical load forecasting", *Applied Soft Computing*, vol. 13, no. 8, pp. 3628–3635, 2013.

[191] V. Sharma and D. Srinivasan, "A spiking neural network based on temporal encoding for electricity price time series forecasting in deregulated markets", in *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2010, pp. 1–8.

[192] D. Reid, A. J. Hussain, and H. Tawfik, "Spiking neural networks for financial data prediction", in *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–10.

[193] L. Yang and T. Zhongjian, "Prediction of grain yield based on spiking neural networks model", in *IEEE 3rd International Conference on Communication Software and Networks (ICCSN)*, 2011, pp. 171–174.

[194] M. Schmuker, T. Pfeil, and M. Nawrot, "A neuromorphic network for generic multivariate data classification", *National Academy of Sciences*, vol. 111, no. 6, pp. 2081–2086, 2014.

[195] E. Tu, N. Kasabov, and J. Yang, "Mapping temporal variables into the neucube for improved pattern recognition, predictive modeling, and understanding of stream data", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 6, pp. 1305–1317, 2017.

[196] N. Kasabov, *Evolving connectionist systems: the knowledge engineering approach*. Springer Science & Business Media, 2007.

[197] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation", *ACM computing surveys (CSUR)*, vol. 46, no. 4, p. 44, 2014.

[198] Z. Zhou, N. Chawla, Y. Jin, and G. Williams, "Big Data opportunities and challenges: Discussions from data analytics perspectives", *IEEE Computational Intelligence Magazine*, vol. 9, no. 4, pp. 62–74, 2014.

[199] G. I. Webb, R. Hyde, H. Cao, H. Nguyen, and F. Petitjean, "Characterizing concept drift", *Data Mining and Knowledge Discovery*, vol. 30, no. 4, pp. 964–994, 2016.

[200] I. Khamassi, M. Sayed-Mouchaweh, M. Hammami, and K. Ghédira, "Discussion and review on evolving data streams and concept drift adapting", *Evolving Systems*, pp. 1–23, 2016.

[201] I. Žliobaitė, M. Pechenizkiy, and J. Gama, "An overview of concept drift applications", in *Big Data Analysis: New Algorithms for a New Society*, Springer, 2016, pp. 91–114.

[202] I. Laña, I. I. Olabarrieta, J. Del Ser, and M. Vélez, "On the imputation of missing data for road traffic forecasting: New insights and novel techniques", *Transportation Research Part C: Emerging Technologies*, vol. 90, pp. 18–33, 2018.

[203] S. Li, Z. Shen, and F.-Y. Wang, "A weighted pattern recognition algorithm for short-term traffic flow forecasting", in *IEEE 9th International Conference on Networking, Sensing and Control (ICNSC)*, 2012, pp. 1–6.

[204] Z. Liu and S. Sharma, "Statistical investigations of statutory holiday effects on traffic volumes", *Transportation Research Record: Journal of the Transportation Research Board*, no. 1945, pp. 40–48, 2006.

[205] S. Schliebs, M. Defoin-Platel, and N. Kasabov, "Integrated feature and parameter optimization for an evolving spiking neural network", in *International Conference on Neural Information Processing*, Springer, 2008, pp. 1229–1236.

[206] S. G. Wysoski, L. Benuskova, and N. Kasabov, "Evolving spiking neural networks for audiovisual information processing", *Neural Networks*, vol. 23, no. 7, pp. 819–835, 2010.

[207] N. Kasabov, K. Dhoble, N. Nuntalid, and G. Indiveri, "Dynamic evolving spiking neural networks for on-line spatio-and spectro-temporal pattern recognition", *Neural Networks*, vol. 41, pp. 188–201, 2013.

[208] S. M. Bohte, J. N. Kok, and H. La Poutre, "Error-backpropagation in temporally encoded networks of spiking neurons", *Neurocomputing*, vol. 48, no. 1-4, pp. 17–37, 2002.

[209] C. Enroth-Cugell and J. G. Robson, "The contrast sensitivity of retinal ganglion cells of the cat", *The Journal of Physiology*, vol. 187, no. 3, pp. 517–552, 1966.

[210] S. Nirenberg and P. E. Latham, "Population coding in the retina", *Current Opinion in Neurobiology*, vol. 8, no. 4, pp. 488 –493, 1998, ISSN: 0959-4388.

[211] M. J. McMahon, O. S. Packer, and D. M. Dacey, "The classical receptive field surround of primate parasol ganglion cells is mediated primarily by a non-gabaergic pathway", *Journal of Neuroscience*, vol. 24, no. 15, pp. 3736–3745, 2004.

[212] S. M. Bohte, H. La Poutré, and J. N. Kok, "Unsupervised clustering with spiking neurons by sparse temporal coding and multilayer RBF networks", *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 426–435, 2002.

[213] S. Schliebs, M. Defoin-Platel, S. Worner, and N. Kasabov, "Integrated feature and parameter optimization for an evolving spiking neural network: Exploring heterogeneous probabilistic models", *Neural Networks*, vol. 22, no. 5, pp. 623–632, 2009.

[214] Q. Yu, H. Tang, J. Hu, and K. C. Tan, "Rapid feedforward computation by temporal encoding and learning with spiking neurons", in *Neuromorphic Cognitive Systems*, Springer, 2017, pp. 19–41.

[215] S. Thorpe and J. Gautrais, "Rank order coding", in *Computational Neuroscience: Trends in Research, 1998*, J. M. Bower, Ed. Boston, MA: Springer US, 1998, pp. 113–118.

[216] S. Schliebs and N. Kasabov, "Evolving spiking neural networks: A survey", *Evolving Systems*, vol. 4, no. 2, pp. 87–98, 2013.

[217] C. Alippi, *Intelligence for embedded systems*. Springer, 2014.

[218] "Informe de calidad y evaluación Ambiental", Ministerio de Agricultura, Alimentación y Medio Ambiente, Madrid, Tech. Rep., 2012, pp. 4–2.4–35.

[219] A. Monzón and M. J. Guerrero, "Valuation of social and health effects of transport-related air pollution in Madrid (Spain)", *Science of the Total Environment*, vol. 334-335, pp. 427–434, 2004.

[220] P. Salvador, B. Artíñano, D. G. Alonso, X. Querol, and A. Alastuey, "Identification and characterisation of sources of PM10 in Madrid (Spain) by statistical methods", *Atmospheric Environment*, vol. 38, no. 3, pp. 435–447, 2004.

[221] V. Valverde, M. T. Pay, and J. M. Baldasano, "Ozone attributed to Madrid and Barcelona on-road transport emissions: Characterization of plume dynamics over the Iberian Peninsula", *Science of the Total Environment*, vol. 543, pp. 670–682, 2016.

[222] M. Brauer, G. Hoek, P. Van Vliet, K. Meliefste, P. H. Fischer, A. Wijga, L. P. Koopman, H. J. Neijens, J. Gerritsen, M. Kerkhof, J. Heinrich, T. Bellander, and B. Brunekreef, "Air pollution from traffic and the development of respiratory infections and asthmatic and allergic symptoms in children", *American Journal of Respiratory and Critical Care Medicine*, vol. 166, no. 8, pp. 1092–1098, 2002.

[223] M. Zuurbier, G. Hoek, M. Oldenwening, K. Meliefste, P. van den Hazel, and B. Brunekreef, "Respiratory effects of commuters' exposure to air pollution in traffic", *Epidemiology*, vol. 22, no. 2, pp. 219–227, 2011.

[224] G. Hoek, R. M. Krishnan, R. Beelen, A. Peters, B. Ostro, B. Brunekreef, and J. D. Kaufman, "Long-term air pollution exposure and cardiorespiratory mortality: a review.", *Environmental Health: A Global Access Science Source*, vol. 12, no. 1, p. 43, 2013.

[225] O. Raaschou-Nielsen, Z. J. Andersen, M. Hvidberg, S. S. Jensen, M. Ketzel, M. Sørensen, S. Loft, K. Overvad, and A. Tjønneland, "Lung cancer incidence and long-term exposure to air pollution from traffic", *Environmental Health Perspectives*, vol. 119, no. 6, pp. 860–865, 2011.

[226] International Agency for Research on Cancer, *Air pollution and cancer*, 161. 2013.

[227] Dirección General de Tráfico, "Anuario Estadístico General", Tech. Rep., 2012.

[228] P. Salvador, B. Artíñano, M. M. Viana, A. Alastuey, and X. Querol, "Multicriteria approach to interpret the variability of the levels of particulate matter and gaseous pollutants in the madrid metropolitan area, during the 1999-2012 period", *Atmospheric Environment*, vol. 109, pp. 205–216, 2015.

[229] P. A. Kassomenos, H. A. Flocas, S. Lykoudis, and A. Skouloudis, "Spatial and temporal characteristics of the relationship between air quality status and mesoscale circulation over an urban Mediterranean basin", *Science of the Total Environment*, vol. 217, no. 1-2, pp. 37–57, 1998.

[230] P. A. Kassomenos, S. Vardoulakis, A. Chaloulakou, A. K. Paschalidou, G. Grivas, R. Borge, and J. Lumbreras, "Study of PM10 and PM2.5 levels in three European cities: Analysis of intra and inter urban variations", *Atmospheric Environment*, vol. 87, pp. 153–163, 2014.

[231] J. Kukkonen, M. Pohjola, R. S. Sokhi, L. Luhana, N. Kitwiroon, L. Fragkou, M. Rantamäki, E. Berge, V. Ødegaard, L. H. Slørdal, B. Denby, and S. Finardi, "Analysis and evaluation of selected local-scale PM10 air pollution episodes in four European cities: Helsinki, London, Milan and Oslo", in *Atmospheric Environment*, vol. 39, 2005, pp. 2759–2773.

[232] S. Vardoulakis and P. Kassomenos, "Sources and factors affecting PM10 levels in two European cities: Implications for local air quality management", *Atmospheric Environment*, vol. 42, no. 17, pp. 3949–3963, 2008.

[233] M. Aldrin and I. H. Haff, "Generalised additive modelling of air pollution, traffic volume and meteorology", *Atmospheric Environment*, vol. 39, no. 11, pp. 2145–2155, 2005.

[234] S. Vardoulakis, B. E. A. Fisher, K. Pericleous, and N. Gonzalez-Flesca, "Modelling air quality in street canyons: A review", *Atmospheric Environment*, vol. 37, no. 2, pp. 155–182, 2003.

[235] J. Hansen, M. Sato, R. Ruedy, G. A. Schmidt, and K. Lo, "Global Temperature in 2014 and 2015", Tech. Rep., 2015.

[236] "Resumen de la Calidad del Aire 2015", Ayuntamiento de Madrid, Madrid, Tech. Rep., 2016.

[237] F. Serrato, "Una ciudad en vilo por la polución", *El País*, 2015.

[238] F. Ferreira, P. Gomes, H. Tente, A. C. Carvalho, P. Pereira, and J. Monjardino, "Air quality improvements following implementation of Lisbon's Low Emission Zone", *Atmospheric Environment*, vol. 122, pp. 373–381, 2015.

[239] C. Holman, R. Harrison, and X. Querol, "Review of the efficacy of low emission zones to improve urban air quality in European cities", *Atmospheric Environment*, vol. 111, pp. 161–169, 2015.

[240] G. Titos, H. Lyamani, L. Drinovec, F. J. Olmo, G. Močnik, and L. Alados-Arboledas, "Evaluation of the impact of transportation changes on air quality", *Atmospheric Environment*, vol. 114, pp. 19–31, 2015.

[241] J. M. Baldasano, M. Gonçalves, A. Soret, and P. Jiménez-Guerrero, "Air pollution impacts of speed limitation measures in large cities: The need for improving traffic data in a metropolitan area", *Atmospheric Environment*, vol. 44, no. 25, pp. 2997–3006, 2010.

[242] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu, "Inferring gas consumption and pollution emission of vehicles throughout a city", in *20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 1027–1036.

[243] J. L. Reyna, M. V. Chester, S. Ahn, and A. M. Fraser, "Improving the accuracy of vehicle emissions profiles for urban transportation greenhouse gas and air pollution inventories", *Environmental Science and Technology*, vol. 49, no. 1, pp. 369–376, 2015.

[244] H. Frey, N. Rouphail, A. Unal, and J. Colyar, "Emissions Reduction Through Better Traffic Management: An Empirical Evaluation Based Upon On-Road Measurements", Tech. Rep., 2001.

[245] B. Barratt, R. Atkinson, H. Ross Anderson, S. Beevers, F. Kelly, I. Mudway, and P. Wilkinson, "Investigation into the use of the CUSUM technique in identifying changes in mean air pollution levels following introduction of a traffic management scheme", *Atmospheric Environment*, vol. 41, no. 8, pp. 1784–1791, 2007.

[246] B. Ando, S. Baglio, S. Graziani, and N. Pitrone, "Models for air quality management and assessment", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 3, pp. 358–363, 2000.

[247] P. Mlakar and M. Boinar, "Perceptron neural network - based model predicts air pollution", in *Intelligent Information Systems*, 1997, pp. 345–349.

[248] R. H. Keeler, "A machine learning model of Manhattan air pollution at high spatial resolution", PhD thesis, 2014, pp. 1–33.

[249] G. Ibarra-Berastegi, J. Saenz, A. Ezcurra, A. Elias, and A. Barona, "Using neural networks for short-term prediction of air pollution levels", pp. 498–502, 2009.

[250] I. González-Aparicio, J. Hidalgo, A. Baklanov, A. Padró, and O. Santa-Coloma, "An hourly PM10 diagnosis model for the Bilbao metropolitan area using a linear regression methodology", *Environmental Science and Pollution Research*, vol. 20, no. 7, pp. 4469–4483, 2013.

[251] J. Hopfield, "Artificial neural networks", *IEEE Circuits and Devices Magazine*, vol. 4, no. 5, pp. 3–10, 1988.

[252] K. S. Lei and F. Wan, "Pre-processing for missing data: A hybrid approach to air pollution prediction in Macau", in *IEEE International Conference on Automation and Logistics (ICAL)*, 2010, pp. 418–422.

[253] J. R. Quinlan, "Induction of Decision Trees", *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[254] V. Vapnik, "Support vector machine", *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[255] X. Xi, Z. Wei, R. Xiaoguang, W. Yijie, B. Xinxin, Y. Wenjun, and D. Jin, "A comprehensive evaluation of air pollution prediction improvement by a machine learning method", in *IEEE International Conference on Service Operations And Logistics, And Informatics (SOLI)*, 2015, pp. 176–181.

[256] E. Sahafizadeh and E. Ahmadi, "Prediction of air pollution of Boushehr city using data mining", in *Second International Conference on Environmental and Computer Science, 2009. ICECS '09*, 2009, pp. 33 –36.

[257] J. Lau, W. T. Hung, and C. S. Cheung, "Interpretation of air quality in relation to monitoring station's surroundings", *Atmospheric Environment*, vol. 43, no. 4, pp. 769–777, 2009.

[258] H. Mayer, "Air pollution in cities", *Atmospheric Environment*, vol. 33, no. 24-25, pp. 4029–4037, 1999.

[259] S. K. Pandey, K. H. Kim, S. Y. Chung, S. J. Cho, M. Y. Kim, and Z. H. Shon, "Long-term study of NOx behavior at urban roadside and background locations in Seoul, Korea", *Atmospheric Environment*, vol. 42, no. 4, pp. 607–622, 2008.

[260] M. Escudero, A. Lozano, J. Hierro, J. del Valle, and E. Mantilla, "Urban influence on increasing ozone concentrations in a characteristic Mediterranean agglomeration", *Atmospheric Environment*, vol. 99, pp. 322–332, 2014.

[261] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures", *Computational Statistics and Data Analysis*, vol. 52, no. 4, pp. 2249–2260, 2008.

[262] J. W. Boylan and A. G. Russell, "PM and light extinction model performance metrics, goals and criteria for three-dimensional air quality models", *Atmospheric Environment*, vol. 40, pp. 4946–4959, 2006.

[263] H. Zhang, J. Hu, S.-H. Chen, and C. Wiedinmyer, "Evaluation of a seven-year air quality simulation using the Weather Research and Forecasting (WRF)/Community Multiscale Air Quality (CMAQ) models in the eastern United States", *Science of the Total Environment*, vol. 473–474, pp. 275–285, 2013.

[264] J. Tu, Z. G. Xia, H. Wang, and W. Li, "Temporal variations in surface ozone and its precursors and meteorological effects at an urban site in China", *Atmospheric Research*, vol. 85, no. 3-4, pp. 310–337, 2007.

[265] H. Zou, Y. Yue, Q. Li, and Y. Shi, "A spatial analysis approach for describing spatial pattern of urban traffic state", in *13th International IEEE Conference on Intelligent Transportation Systems*, 2010, pp. 557–562.

[266] W. Weijermars and E. van Berkum, "Analyzing highway flow patterns using cluster analysis", in *IEEE Intelligent Transportation Systems Conference*, 2005, pp. 831–836.

[267] G. Raducan and S. Stefan, "Characterization of traffic-generated pollutants in Bucharest", *Atmosfera*, vol. 22, no. 1, pp. 99–110, 2009.