UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

MASTER THESIS

A New Spatio-Temporal Neural Network Approach for Traffic Accident Forecasting

Author:
Rodrigo de Medrano López

Supervisor:
José Luis Aznarte Mellado, PhD

A thesis submitted in fulfillment of the requirements for the degree of MSc. Advanced Methods in Artificial Intelligence

in the

Departament of Artificial Intelligence

September 13, 2019

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

Abstract

ETS de Ingeniería Informática
Departament of Artificial Intelligence

MSc. Advanced Methods in Artificial Intelligence

A New Spatio-Temporal Neural Network Approach for Traffic Accident Forecasting

by Rodrigo de Medrano López

Traffic accidents forecasting represents a major priority for traffic governmental organisms around the world to ensure a decrease in life, property and economic losses. The increasing amounts of traffic accident data have been used to train machine learning predictors, although this is a challenging task due to the relative rareness of accidents, inter-dependencies of traffic accidents both in time and space and high dependency on human behavior. Recently, deep learning techniques have shown significant prediction improvements over traditional models, but some difficulties and open questions remain around their applicability, accuracy and ability to provide practical information. This paper proposes a new spatio-temporal deep learning framework based on a latent model for simultaneously predicting the number of traffic accidents in each neighborhood in Madrid, Spain, over varying training and prediction time horizons.

Acknowledgements

Cuando uno se embarca en la realización de un trabajo de este calibre, se puede tener por seguro que el camino no será fácil. Por eso, y sin desestimar el trabajo propio que haga cada uno, es imprescindible rodearse de personas que estén dispuestas a remar en la misma dirección que uno mismo.

Por ello, me gustaría agradecer especialmente la labor, apoyo y sobre todo confianza que el Dr. José Luis Aznarte ha depositado en mí desde el principio. Sin duda, este trabajo no habría sido posible sin su tutela.

Igualmente a mi familia, Irene y amigos, quienes al final son los que están tanto en las buenas como en las malas en el día a día. Siempre habéis sabido ser el apoyo que quizás no merecía, pero sí que necesitaba.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Through this first chapter, a general vision of the problem will be offered and its importance will be highlighted. Objectives and previous research on the topic are also discussed. At the end, there is a brief overview of the thesis.

1.1 Presentation

Nowadays, the urbanization trend around the globe has introduced new opportunities and issues in the cities. One of the most important aspects of the modern society is related to the use of motorized vehicles as a method of transport. Although very efficient in several ways [13], motor vehicles imply problems related to traffic and health care. For example, pollution and traffic accidents are some of the principal causes of death in cities all over the world [8, 28].

This is the reason why the scientific interest for traffic accidents has increased in the past decades, and proposing solutions is a crucial issue for the sake of improving transportation and public safety. Being capable of understanding and reducing accidents has become an important commit in many cities, as they not only cause significant life losses, but also property and economic ones [21].

In this work, an effort will be put to study the traffic accident phenomenon in the city of Madrid. This has been the subject of several lines of research in the past, although most previous studies on traffic accident prediction conducted by domain researchers simply applied classical prediction models on limited data without addressing many challenges properly, thus leading to unsatisfactory performances. For instance, the imbalanced severity classes, non-linear relationship between dependent and independent variables or spatial heterogeneity are usual problems to deal with in order to improve previous results in the field. In addition, traffic accidents show a potential problem when using quantitative methodologies for their prediction: there is a great dependence between accidents and human behaviour, being distractions or merely human actions cause of almost 60% of deadly traffic accidents in Spain [10] (and even a larger percentage for non-deadly accidents).

Although predicting the exact space-temporal position of accidents is out of the scope with actual techniques due to its complexity [17, 34], much progress might be done by characterizing important parts of the problem.

Trying to reduce the dimensionality of the space as much as possible, discovering relevant features or improving previous models are some examples of what can be done to provide insight in this particular problem.

In this context, this work presents the problem as a spatio-temporal series in which traffic intensity and meteorological variables play a central rol in predicting values for the traffic accidents series. For this purpose, we propose a new model, called XSTNN (from Exogenous Spatio-Temporal Neural Network) and consisting of a deep learning approach for traffic accident regression based on spatio-temporal data. The model, which is an extension of the one proposed by Ziat et al. [37] (Spatio-Temporal Neural Network, STNN) through the addition of external variables, is based on partitioning space into grid cells and taking advantage of the spatial relations existing in the series. A number of urban and environmental variables such as traffic intensity, rainfall, temperature and wind are collected and map-matched with each grid cell. Given the number of accidents as well as the other urban and environmental features at each location, we learn a model to forecast the number of accidents that will occur in each grid cell in future timesteps.

By presenting the number of traffic accidents as a spatio-temporal series and learning how to model it, it is possible (for example) to increase emergency service's response time, focus the efforts to avoid potential dangers, create real-time safe routes recommendation systems and, in short, reduce the losses that were discussed at the beginning. To the best of our knowledge, this is the first work that tackles the traffic accident forecast problem in the city of Madrid.

1.2 Motivation

The urban context is a perfect example of a complex system. Although it has been analyzed for years from different perspectives [2], there is still plenty of research and debate about every topic and decision related to them. In general, cities are the perfect stage for (at this moment) unpredictable events. This, together with the fact that every aspect has often a big impact on the everyday life, makes cities an interesting area of study.

As it was noted in the previous section, in the specific case of traffic accidents two potential problems are noteworthy: an extraordinary loss of resources and the worsening of traffic flow. As the complex system a city is, both issues might affect in the short and long term: for example, by the decrease of productivity (traffic jam) or the cost of medical care for severely injured people respectively. In any case, traffic accidents are a matter of capital importance nowadays.

Based on this last premise, previous studies have shown a strong dependency between time [34] or space [6] and collisions, but focused in one at a time. Researches pointing out the relation between time and space and the importance of using both are scarce, and they usually use classical approaches. Notwithstanding, few works have used more sophisticated tools based in Artificial Intelligence. Hence, this work propose the use of both modern and classical methodologies such as Recurrent Neural Networks

(RNN), Gradient Boost methods, linear and naive models to gain insight in the spatio-temporal study of the field.

Besides, a significant portion of the factors that contribute to collision-risk are straightforward. Most would expect that busy, in poor condition, high-speed roads endure a disproportionate number of automobile collisions. What is not straightforward, however, is how those variables influence the system. For this purpose, this works set the hypothesis that meteorological and traffic features are also relevant. To test these hypothesis, weather and traffic flow variables will be used.

1.3 Previous work

Although very much studied, traffic accidents have been treated mostly in a ”classical” context, by simply using statistical analysis in an attempt to understand better the phenomenon and the circumstances surrounding them. Examples that illustrate this situation can be found in [1, 14, 24]. There also are several works dealing with these methodologies and their typical issues, as for example [15, 17]. A long list of studies tackle the issue from the severity of the injuries perspective. Within this last group, [20, 19] are some examples. Although instructive, most of these previous research fail to be able to apply all this knowledge to predict future events.

In a closer line to our work, during the last decade a considerably number of Artificial Intelligence-based approaches have appeared, taking advantage of the large datasets which are available nowadays. We can cite [3, 11, 12, 35] as examples. As a first glance in the matter, these works provide new tools for solving the problem, but they lack relevant information in their analysis. In order to get more sophisticated and precise systems, last researches focus their efforts in new models as Variational Autoencoders and Deep Neural Networks for detecting and understanding better traffic accidents [36, 25, 32].

Until now, the references presented here were all lumped under the same hypothesis: ignoring the importance of the spatial dimension in the traffic accidents forecasting. However, a number of studies have pointed out how relevant this variable is in order to get appropriate results [31, 23, 6]. Since then, more and more researches focus their efforts in the spatio-temporal (and not just temporal) prediction problem. We can cite [30, 22, 33] as some of the most relevant works, some of them being classified under the label of deep learning. Specifically, some of these last references point at exogenous variables as helpful in the forecast process.

1.4 Objectives

The main objectives of this research are:

- To explore existing methodologies and propose new ones for forecasting traffic accident spatio-temporal series.

- To provide a practical case study in the city of Madrid.

In the course of this project, it is expected that, pursuing these goals, we may offer new information and perspectives in the use of spatio-temporal neural models for traffic accidents prediction.

1.5 Problem formulation

Given a spatial grid S, where each grid is represented as s_i, and a timestep t_j, we aim to learn a model to predict the number of accidents in each grid s_i during each time slot t_j. This mean that a spatio-temporal sample writes as $x(s_i; t_j) : j = 1, ..., T; i = 1, ..., S$.

More precisely, we propose that each grid s_i represents a neighborhood of Madrid as it is expected that each neighborhood presents different peculiarities that might be related to traffic accidents. Moreover, we use an hour as the length of our timestep t_j. Without loss of generality, other values could be chosen for s_i and t_j. We work with data from year 2018 for both the training and validation sets. Only in-city accidents are treated, as road accidents present different peculiarities. Chapter 2 reiterates and expands all these ideas.

1.6 Thesis overview

This work is subsequently organized as follows:

- Chapter 2 introduces data description along with the respective cleaning and pre-processing work in order to boost its predictive nature and improve the performance of the models that will be presented in Chapter 3.

- Chapter 3 attempts a review of existing models capable of dealing with traffic accidents prediction. The advantages and drawbacks of these models are pointed out, and an extension of a previous Artificial Neural Network (ANN) based model is presented, along with a discussion of its advantages with respect to existing methods.

- Chapter 4 explains the experiments undergone to test our proposal.

- In Chapter 5 the results acquired using the approaches of Chapter 3, data explained in Chapter 2 and settings exposed in Chapter 4 are shown.

- Finally, Chapter 6 concludes this work, points out the main contribution of this work, discusses ways in which the problems of our system can be addressed, and presents possible future work.

As *an addendum*, in appendix A we provide information about Madrid neighborhoods as they will represent our spatial grid.

Chapter 2

Data analysis and description

> Along this second chapter, a presentation and examination of the data is provided. In order to get a precise vision of it, data will be cleaned and several analysis will be performed.

As it was established on the first chapter, the problem we are trying to tackle will be treated as a spatio-temporal series problem. For this reason, the data used in the regression needs to adapt to both the spatial and temporal dimensions of the series. In this section, the data and its properties will be presented, and it will be shown that its nature is appropriate for the proposed theoretical framework.

At this point, it is necessary to distinguish two types of datasets that will have a different impact and nature on our study: the main dataset is the traffic accident one, as it will be the base of the regression problem. The remaining data (traffic and weather data) is expected to contribute to the performance of the different models, but it will be used as exogenous variables respect to the series.

As it may be seen, the accidents themselves form the time series. Meanwhile, the rest of the data intends to improve the performance of the different models used.

2.1 Data presentation

Now it is time to present all the data that has been used. As previously advanced, there are three sources or datasets. For each one of them, a summary of their variables and the granularity of their spatio-temporal information are shown.

- **Traffic accident data:** Provided by *Portal de datos abiertos del Ayuntamiento de Madrid* [1], it summarizes all the information related to car crashes in the city of Madrid. Specifically, for every accident it shows physical location (although not geographical), date (year, month and day), time (hour), sex and severity for each person involved and several meteorological conditions. The last two variables of this dataset were not taken in consideration, as they were not relevant or there were better sources

[1] https://datos.madrid.es/portal/site/egob/

for them (specifically Weather data later in this same section). For example, sex can be relevant when making statistics of the phenomena, but irrelevant when trying to predict new accidents.

Spatial information is presented as city addresses (street and number or intersection), while temporal information is limited to the hour in which the accident was reported.

- **Traffic data:** As before, provided by *Portal de datos abiertos del Ayuntamiento de Madrid*[1]. This dataset contains historical data of traffic measurement points in the city of Madrid. The measurements are taken every hour at each point, including traffic intensity in number of cars per hour and average speed in m/s. Some other traffic parameters, although unused in this project, are present in this set too.

  Spatial information is given with the coordinates (longitude and latitude) of measurement points, while temporal information is taken every 15 minutes.

- **Weather data:** Weather data was provided by the *Red Meteorológica Municipal*[2]. Weather observations consist of hourly temperature in Celsius degrees, solar radiation in W/m^2, wind speed measured in ms^{-1}, wind direction in degrees, daily rainfall in mmh^{-1}, pressure in mbar, degree of humidity in percentage and ultraviolet radiation in mWm^{-2} records.

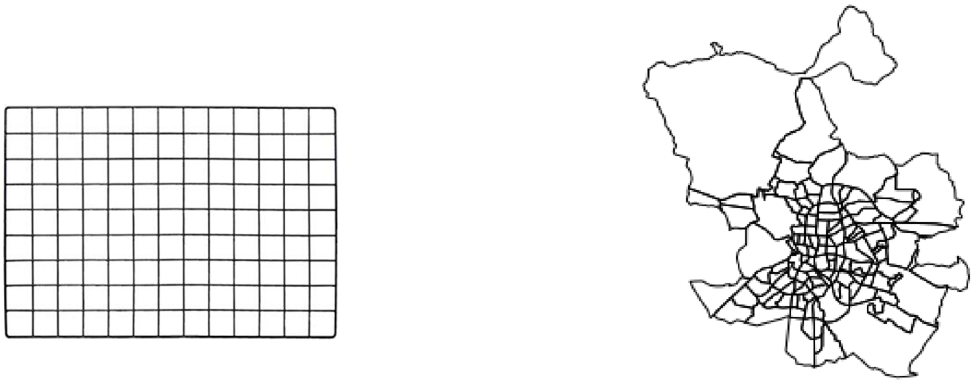
  Weather information is taken along six different stations. It is reported hourly.

Given the nature of the project, spatio-temporal granularity is of special relevance since it limits the scope of the prediction. For example, any model which uses the traffic accident dataset will present automatically a systematic error due to poor precision in spatio-temporal data collection.

In addition, there is one more dataset which is necessary in this project. As it was pointed out in Section 1.3, the choice of spatial zones is of great importance. Not only the election of the spatial mesh is important, but it should be clear that a model with similar spatial relations to the real ones will be more willing to correctly capture the dynamics of the series. Usually, and for simplicity, the spatial grid is formed by uniform squares [22, 33] (figure 2.1a) or by road segments and intersections [6, 30]. For a more realistic model, it has been decided to use Madrid neighborhoods as spatial zones (figure 2.1b). Although this decision might introduce extra difficulties in the work, the unique properties demographically and economically speaking of neighborhoods are expected to be beneficial. This extra dataset (geographical format) can be found in *Portal de datos abiertos del Ayuntamiento de Madrid*[1] again. Spatial relations chosen for this work will be described in Section 4.4.

Lastly, and as it was pointed out previously, some variables from the original datasets were not relevant. Nevertheless, all the features used throughout this work have been indicated.

[2] http://www.mambiente.madrid.es



(A) A classical squared grid for spatial zones.

(B) Our proposed mesh grid for spatial zones based on Madrid neighborhoods.

FIGURE 2.1: Two mesh grid choices for a spatial problem in the city of Madrid.

2.2 Data cleaning

During this section, we present the actual cleaning work that was necessary to work through the data. It is worth noting that only the year 2018 will be modeled.

For every particular dataset, we have the following aspects.

- **Traffic accident data:** Firstly, Google Maps Api [3] was used for geocoding the adresses provided in the dataset. Specifically, coordinates in longitude and latitude so all the data presents the same format.

  Secondly, a neighborhood is assigned to each crash.

  Thirdly, each accident was repeated for every person involved. In this work, it is only important the number of accidents without further information about the event. For this reason, this repetition was eliminated.

  No missing values were reported.

- **Traffic data:** In regards to traffic intensity, it is worth highligthing that is the only set that does not present its information hourly, but every 15 minutes. In order to have a final homogeneous dataset, average over every entire hour is calculated. Note that typical deviation of traffic intensity over and hour represents less than 10% of the real values on average.

  In addition, the average of the traffic intensity is taken for each neighborhood as if every measurement point was a different sample from the same phenomenon for every zone. Once more, the standard deviation that results from this decision is less than 5% respect to the mean,

[3] https://cloud.google.com/maps-platform/

showing that there is a predisposition to have similar traffic conditions for each neighborhood.

Missing values represent a small percentage of the dataset ($\sim 0.8\%$), being replaced by the average of traffic intensity from all adjacent neighborhoods at the same time.

- **Weather data:** While the real data was taken in six subestations in the city of Madrid, our own data consists of average hourly variables from those six substations. Although this decision could be seen as a loss of information, this approximation is enough for a first insight. Also, assigning different meteorological variables for each accident depending on its location supposes an extra difficulty when using a spatial mesh (the six subestations) different from the one used in this work (neighborhoods of Madrid).

  Missing values were a small percentage ($< 0.1\%$), so they were replaced by the average of the previous and the next value for each feature.

There are some other general aspects that are worth to mention. For example, only accident in urban roads are considered, as other types of roads present different properties that may affect the series. Some neighborhoods (three in particular) do not present traffic or accidents data so they were removed, giving a final number of 131 neighborhoods. Finally, for generating the dataset that will feed the different models all the features are condensed as elements of a matrix. In this matrix, every row intends to be a temporal step while every column represents a different neighborhood. A scheme of this final dataset is presented in table 2.1.

	Palacio	Embajadores	...	Corralejos
01/01/2018, 00:00	Number of accidents, traffic conditions, weather			
01/01/2018, 01:00				
...				
31/12/2018, 23:00				

TABLE 2.1: Scheme of the final-cleaned dataset used in this project.

In summary, at this point we have information related to the number of traffic accidents (including 0), traffic condition and weather for every neighborhood every hour. All this together forms the basis for the spatio-temporal regression problem that this work tries to cover.

2.3 Data analysis

Although stochastic by nature, traffic accidents show some properties that make them suitable for a spatio-temporal series approach. Through this section, a time and spatial analysis of the series is carried out, pointing out its

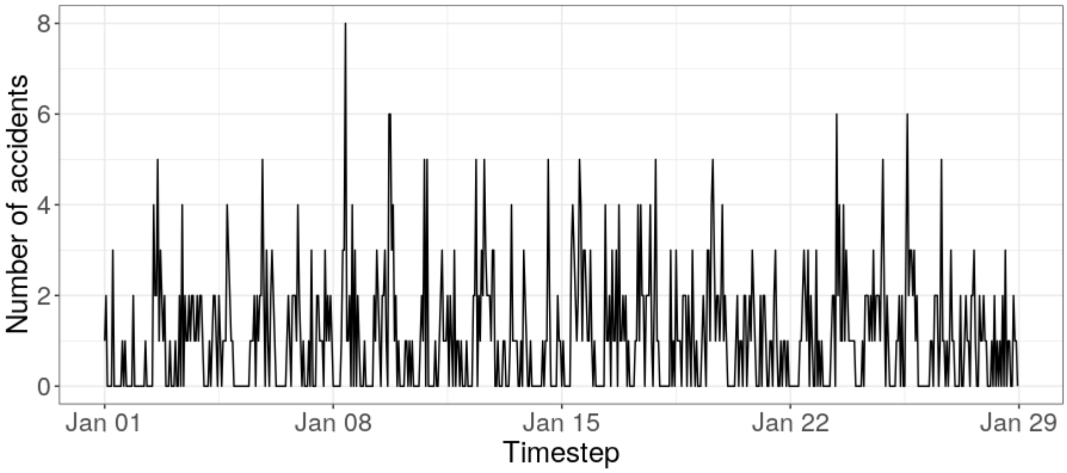


FIGURE 2.2: Total number of accidents for each timestep in Madrid.

main characteristics. Then, we go deeper in understanding the relations between the main series (car crashes) and the exogenous variables (traffic and weather).

In all cases, analysis have been done for the same data extract that will be used in the experiments. To be more precise and as a reminder, 131 neighborhoods of Madrid and year 2018.

2.3.1 Time series study

First, some basics statistics are presented before (table 2.2) and after normalizing (table 2.3) from the entire series for the traffic accidents dataset (no spatial grid considered). Usually, normalize is considered a good practice for the sake of facilitating and balancing the calculations in the different models. For this work, each series was rescaled between 0 and 1. Remark again that the time series is studied in Madrid, without any spatial differentiation.

Statistics without normalizing
Min. : 0
1st Qu. : 0
Median : 1
Mean : 1.074
3rd Qu. : 2
Max. : 11

TABLE 2.2: Dataset statistics before reescaling

Statistics having normalized
Min. : 0
1st Qu. : 0
Median : 0.091
Mean : 0.098
3rd Qu. : 0.182
Max. : 1

TABLE 2.3: Dataset statistics after reescaling

From this tables it is easy to see how infrequent accidents are. In this context, and from the frequentist probability point of view, the odds of an accident taking place anytime in an hour and at any neighborhood is about

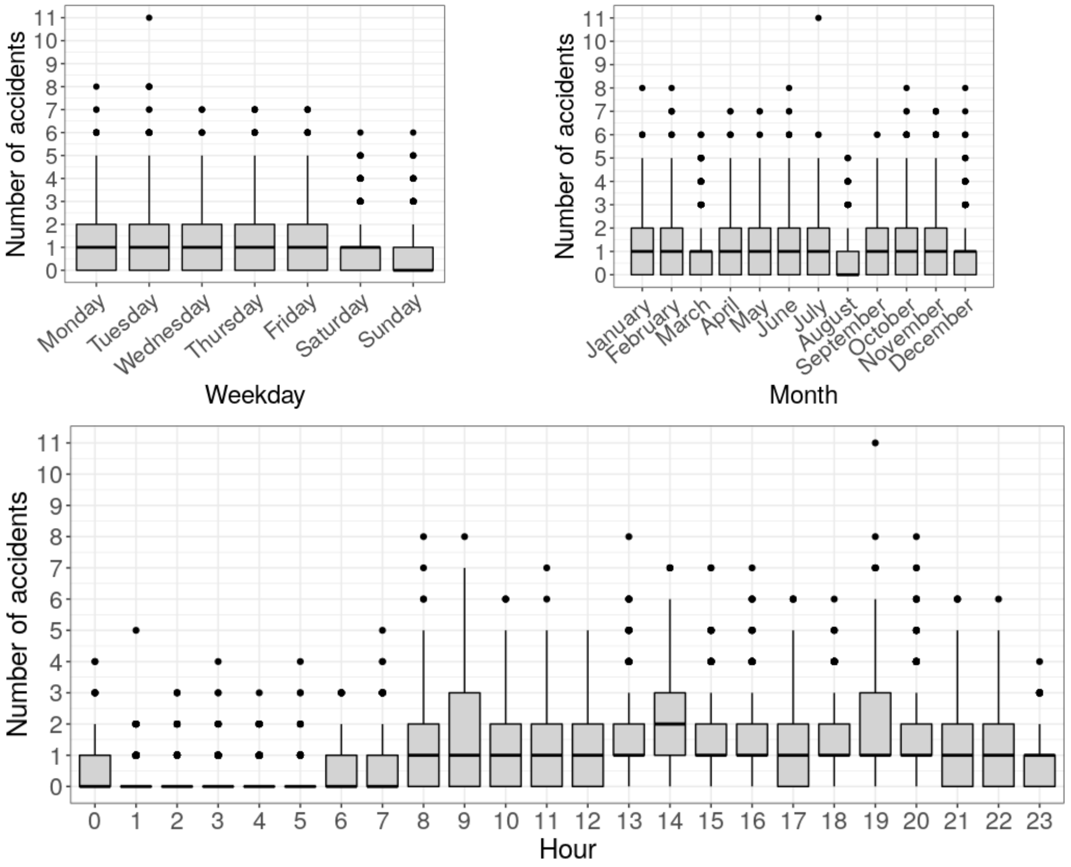


FIGURE 2.3: Periodicities of the traffic accidents series. (a) Number of accidents depending on day of the week. Weekends present less number of accidents. (b) Number of accidents for each month. August seems to be safer. (c) Number of accidents depending on hour of the day. In this case we have the most clear difference.

0.8%. For this reason (among other peculiarities introduced in Chapter 1), traffic accidents series are considered specially difficult to forecast.

For a deeper insight in the matter, figure 2.2 offers the total number of accidents per timestep (hour) during a four week period(in order to facilitate its comprehension and clarity). In this plot, some tendencies can be observed, although it is not clear how the series is distributed over time.

As we have just seen, our series is not a classical time series. Nevertheless, by making a deeper analysis of the data it is possible to find some characteristic periodicities that reveal a hidden time dependence in traffic accidents. Fig. 2.3 shows a clear pattern that depends on several time dimensions.

These figures establish a relation between time and accidents, contributing to reinforce the idea of treating our data as a time series. Moreover, they let us understand better the phenomenon: the number of accidents is directly related to the schedules of daily life in Madrid. They reach their highest levels during the week, coinciding with daily commuting. On the contrary, August, nights and weekends usually carry less displacements, with the respective decrease in the likelihood of traffic accidents happening.

Now that the time dependency of the series has been established, some usual time series analysis is provided. For this analysis, a 24 timestep frequency has been considered, meaning that the series is expected to be repeated every 24 hours:

- **Trend:** Although there are no proven "automatic" techniques to identify trend components in the time series data, as long as the trend is monotonous (consistently increasing, decreasing) or stable that part of data analysis is typically not very difficult. In our specific case, the series is stable.
- **Seasonality:** Seasonal dependency (seasonality) is another general component of the time series pattern. Seasonal patterns of time series can be examined via correlograms. The correlogram (autocorrelogram) displays graphically and numerically the autocorrelation function (ACF), that is, serial correlation coefficients (and their standard errors) for consecutive lags in a specified range of lags. Concretely, a stationary time series will have the autocorrelation fall to zero fairly quickly but for a non-stationary series it drops gradually. Figure 2.4 illustrate this idea, showing a clear fall to zero as a proof of a some-stationary series for our problem.

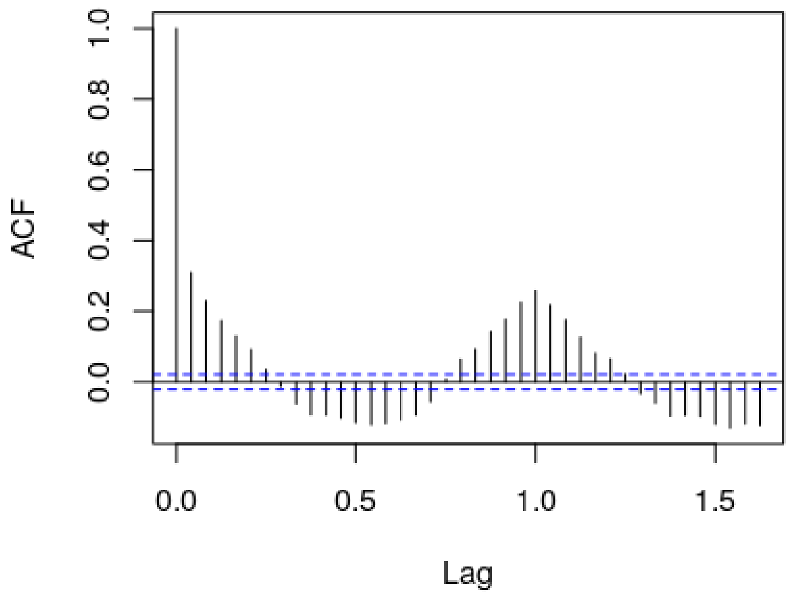


FIGURE 2.4: Autocorrelation of the time series associated to traffic accidents in Madrid.

It is recalled again that, until now, the data aggregate for the entire city of Madrid has been used. If each neighborhood is treated as a different series, as will be done in this work, each of these will have a different temporal behavior. However, this initial analysis lets us understand better the phenomenon.

2.3.2 Spatial series study

As we did in the previous section, spatial dependency can be seen by plotting the total number of accidents for each spatial zone. For a more clear insight, at first districts are used instead of neighborhoods (as there are only 21 of them). These districts are the next level in territorial division policy in Madrid, gathering neighborhoods with similar characteristics. Figure 2.5 shows this spatial dependency.

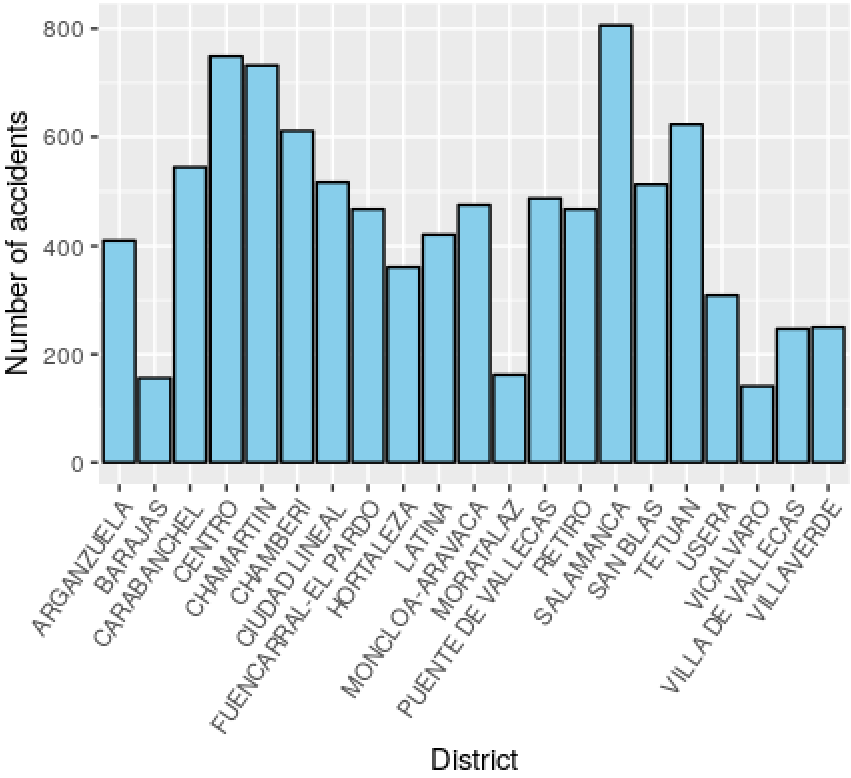


FIGURE 2.5: Total number of accidents by district of Madrid.

From this last figure it should be clear that different districts (and, in consequence, neighborhoods) present different peculiarities that might be related to traffic accidents. The conclusion from this analysis is double: not only exist a clear dependency, but the election of the neighborhoods as spatial zones that was made before shows to be relevant. Moreover, the fact of using a known grid is expected to guarantee a better understanding and extrapolation of the results.

For a more concise analysis, figure 2.6 illustrates the same idea discussed before but for our own mesh grid: Madrid neighborhoods. Usually, the highest traffic accident region lies in the major commercial and business areas.

2.3.3 Relations between datasets

Lastly, we defend the election of the external data that will feed the models: traffic and weather data. Several studies conclude that there is a clear relation between those variables and traffic accidents [6, 16].

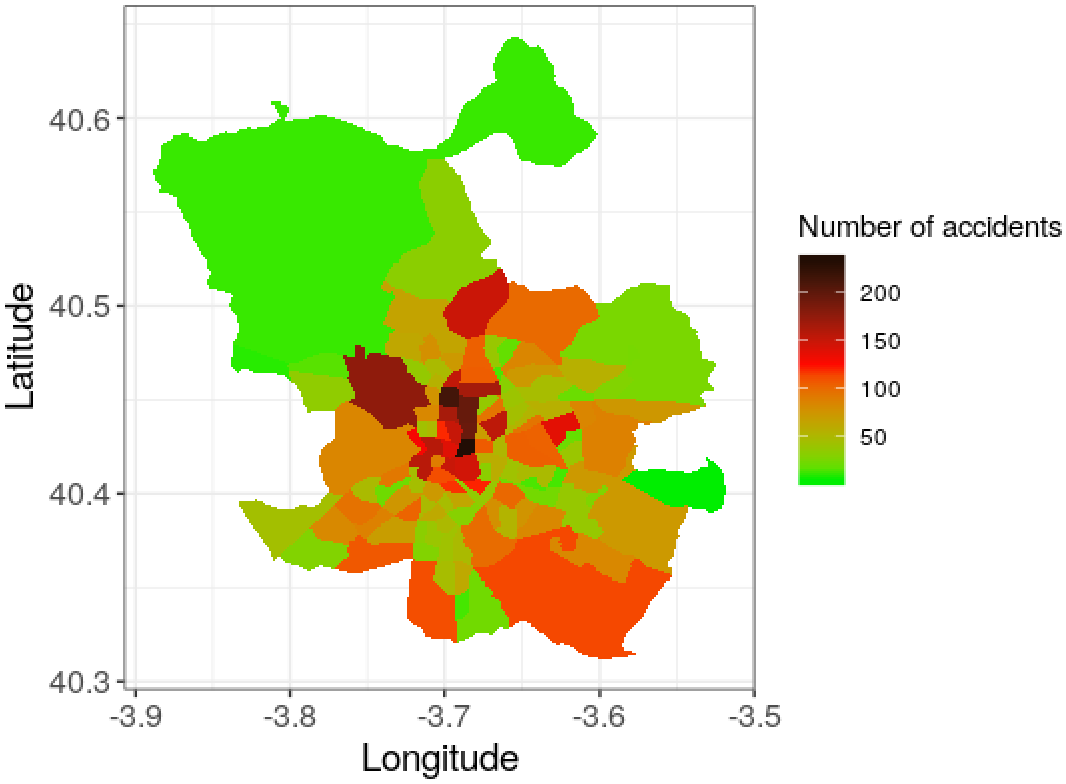


FIGURE 2.6: Total number of accidents by neighborhoods of Madrid.

In figure 2.7, the Pearson correlation coefficient is represented for each exogenous variable respect to traffic accidents. This magnitude lets us have a better understanding of the linear association between datasets.

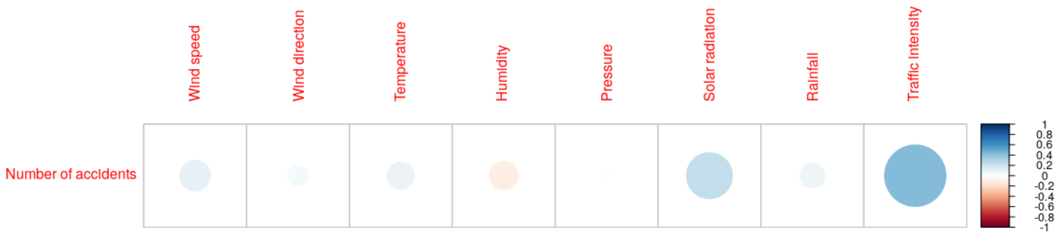
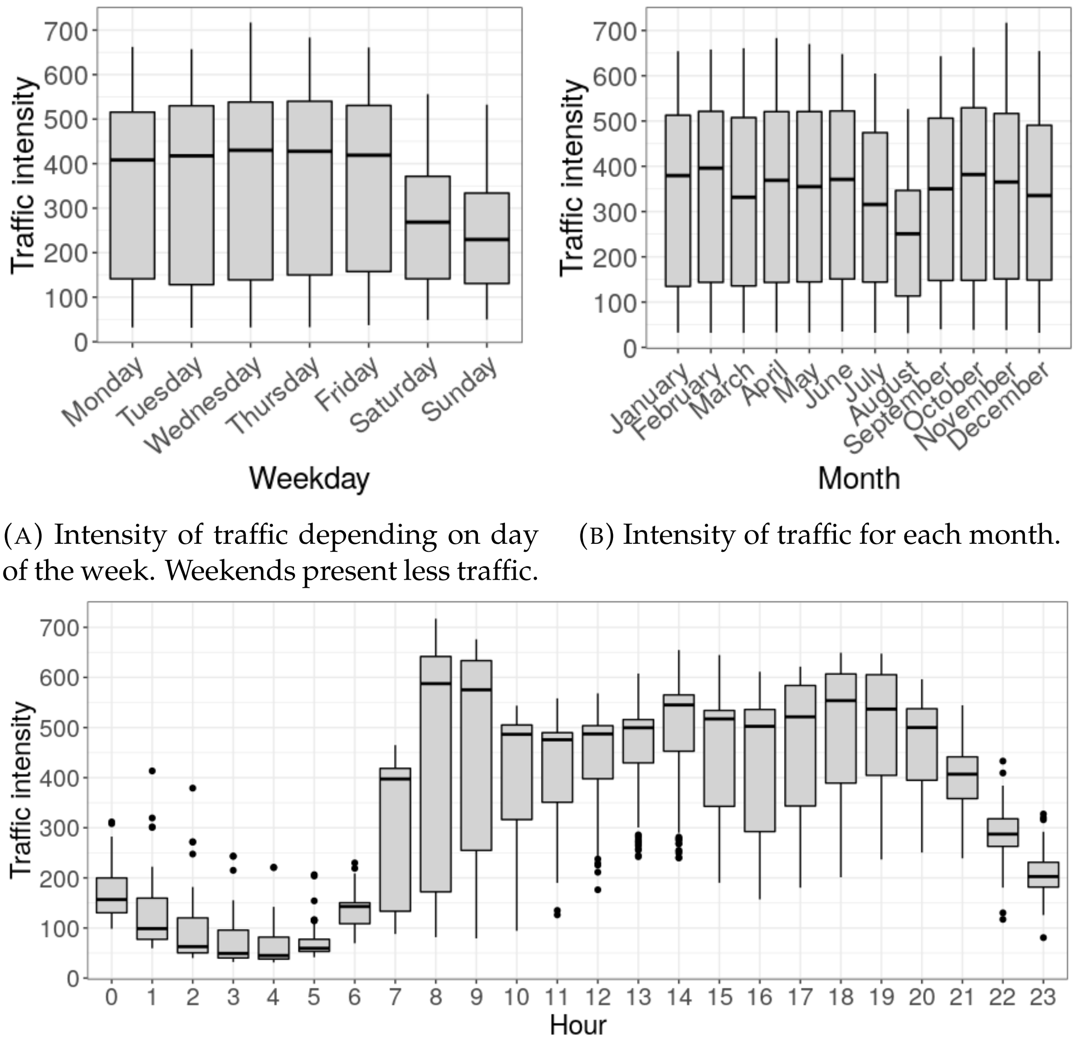


FIGURE 2.7: Correlation diagram of traffic accidents respect to exogenous variables.

Although a linear relation is not entirely representative of how two variables depend on each other (this relation could be not linear), diagram 2.7 lets us confirm that the chosen exogenous variables are relevant for the target feature.

Solar radiation has shown to be relevant in previous studies [30], due to its relation to the day-night cycle and the loss of visibility by sunlight. In the concrete case of traffic data, plotting traffic intensity against same time intervals that figure 2.3 reveal a similar pattern. Figure 2.8 illustrate this idea.



(A) Intensity of traffic depending on day of the week. Weekends present less traffic.

(B) Intensity of traffic for each month.

(C) Number of accidents depending on hour of the day. In this case we have the most clear difference.

FIGURE 2.8: Periodicities of the traffic series.

This last figure reflect how similar are both series. It should not be a surprise: the more intensity of traffic (number of vehicles per hour), the more number of traffic accidents we would expect to happen.

Chapter 3

Models for spatio-temporal series regression

This chapter reviews the models used along this thesis. After a quick look at notation, Section 3.2 presents a deep neural network approach for spatio-temporal series. Section 3.3 describes a new model based on the previous one created for this project. Finally, Section 3.4 explains the rest of the proposed models that will form the baseline.

3.1 Notation

Let us first introduce the notation that will be used throughout this chapter. We denote n as the number of series, T their length and m the dimensionality of them. In our specific domain, there will be as many series (n) as spatial zones. Moreover, $m = 1$ as every series will be composed of only one dimension: traffic accidents.

If we call X as the values of all the series between instants 1 and T, then X is a tensor in $\mathbb{R}^{T\times n\times m}$. At last, $X_t \in \mathbb{R}^{n\times m}$ is a tensor that denotes the values of all the series at time t.

3.2 The STNN model

Proposed by Edouard Delasalles, Ali Ziat, Ludovic Denoyer and Patrick Gallinari [37], the STNN model is a deep neural network approach capable of learning temporal and spatial dependencies through a structured latent space. Our model preserves this nature but it is an improvement from the point of view of its usability, allowing us to make use of external (or exogenous) variables. Concretely, the model learns these spatio-temporal dependencies through a structured latent dynamical representation, while a decoder predicts the observations from the latent space.

3.2.1 The main idea

If we do not consider spatial relations, the problem is equivalent but simpler. This allows us to present the model in an easier way without loss of generality.

Let Z_t be the latent representation, or latent factors, of the series at time t. The model has two principal components: the dynamic function (denoted as g), and the decoder function (called d). The first one is in charge of controlling the dynamics of the system, calculating the next latent state based on the previous one: $Z_{t+1} = g(Z_t)$. The second one is a decoder which maps latent factors Z_t onto a prediction of the actual series values at time t: $\tilde{X}_t = d(Z_t)$, $\tilde{X}_t$ being the prediction computed at time t.

As it should be clear, the parameters of both functions (g and d) are learned so that the essence of the series is captured. Unlike usual neural networks, the latent representation Z_t is treated as a parameter too, distinguishing this model and making it more flexible than usual recurrent neural networks.

Having been presented, the next step is defining the learning problem. As it was established before, two mapping functions (g and d) together with the latent factors Z_t are learned from data. Consequently, the loss function that is proposed gathers all these elements. Let $\mathcal{L}(g, d, Z)$ be this objective function:

$$\mathcal{L}(d, g, Z) = \frac{1}{T}\sum_{t}\Delta(d(Z_t), X_t) + \lambda\frac{1}{T}\sum_{t=1}^{T-1}||Z_{t+1} - g(Z_t)||^2 \tag{3.1}$$

In this expression, the first term tries to measure how well the decoder works, while the second term captures the ability of the model to capture the dynamics of the series via the latent space. The hyperparameter λ needs to be fixed for every problem, and contribute to balance the importance of this second term.

As usual in neural networks, the problem to tackle is minimizing the loss. In mathematical terms:

$$d^*, g^*, Z^* = \arg\min_{d,g,Z}\mathcal{L}(d, g, Z) \tag{3.2}$$

Inference is done by calculating new latent factors via the g function as much steps as necessary. Formally, if the learned vector is Z_T, the latent space at time $T + \tau$ will have this form:

$$\tilde{Z}_\tau = g \circ g \circ \ldots \circ g(Z_T) = g^{(\tau)}(Z_T) \tag{3.3}$$

In summary, the network learns the dynamic of the series via a latent representation (function g), how to translate from this latent space to our series (function d) and the latent structure itself (Z_t). Hence, it is capable of predicting new steps of the series by applying the dynamic function as many times as required.

Lastly, the learning problem can be solved with Stochastic Gradient Descent (SDG) algorithms, sampling a pair (Z_t, Z_{t+1}) and updating the three set of parameters described before according to the gradient of (3.1).

3.2.2 Application to spatio-temporal series

Until now, we have just explained the model without any spatial component or the form of g and d. The idea behind the spatial component is to consider each zone as a different series with its own latent representation at each time step. For a latent space dimension of N, Z_t is a $n \times N$ tensor such that $Z_{t,i} \in \mathbb{R}^N$ is the latent factor of series i at time t. Thus, we have the following relations:

$$d : \mathbb{R}^{n \times N} \rightarrow \mathbb{R}^{n \times m} \tag{3.4a}$$

$$g : \mathbb{R}^{n \times N} \rightarrow \mathbb{R}^{n \times N} \tag{3.4b}$$

Not only each spatial zone has a series, but spatial information is integrated in the dynamic component of the model through a matrix $W \in \mathbb{R}_+^{n \times n}$ that shares information between all the zones. Although this matrix will be provided in the relevant parts of this work, the actual model is also capable of learning or refining it by defining W elements as actual learnable parameters.

The latent representation of each series at time $t+1$ depends on the previous state of all the series (included itself). Hence, we can separate the calculation of a new state by two different sources: intra-dependency in the first term of the right-hand side of (3.5) and inter-dependency in the second term. The first one aims to get the dynamic of each series as an individual entity, whereas the second one is devised to exploit spatial relations between all series. This way, the model considers a different temporal series in each spatial zone while keeping information about the spatial relation between all of them. Formally, the dynamic model $g(Z_t)$ is designed as follows:

$$Z_{t+1} = h(Z_t \Theta^{(0)} + W Z_t \Theta^{(1)}) \tag{3.5}$$

In this last equation, h is a nonlinear function ($h = tanh$ in this project) and Θ denotes a parametrized function $\Theta \in \mathbb{R}^{N \times N}$. In this case, Θ will be a linear function or a multilayer perceptron (MLPs), although could be any parametrized function - see Section 4.3. Figure 3.1 shows a diagram of the model.

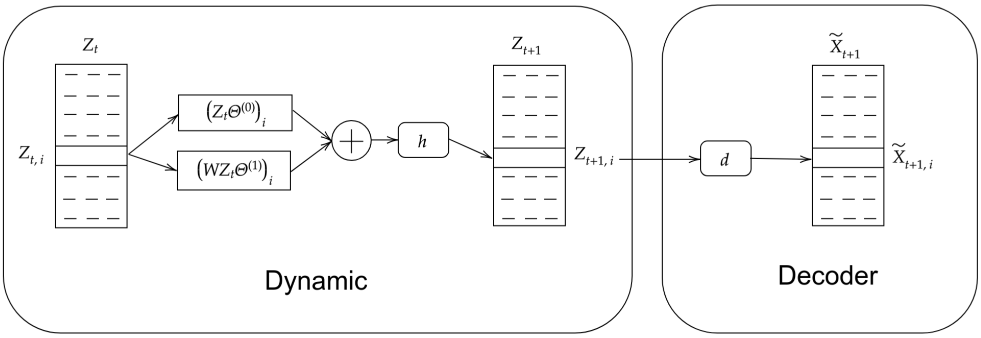


FIGURE 3.1: Architecture of the STNN model as described in Section 3.2.2.

At the end, the optimization problem can be written as:

$$
\begin{aligned}
d^*, Z^*, \Theta^{(0)*}, \Theta^{(1)*} = \underset{d,Z,\Theta^{(0)},\Theta^{(1)}}{\arg\min} \; & \frac{1}{T}\sum_t \Delta(d(Z_t), X_t) \\
& + \lambda \frac{1}{T}\sum_{t=1}^{T-1} ||Z_{t+1} - h(Z_t\Theta^{(0)} + WZ_t\Theta^{(1)})||^2
\end{aligned} \tag{3.6}
$$

This is a modified extract of the real and complete model. For a deeper and broader lecture about the STNN model, we reference again [37].

3.3 The XSTNN model

The main limitation of the STNN is that it is not able to take into account the exogenous variables which might be related to the process being modelled and which could enrich the internal representation and, thus, improve the predictions. In this section, a new model called XSTNN that aims to resolve this problem is proposed. Based on the STNN, we introduce exogenous variables as extra information that might be beneficial for the performance of the spatio-temporal regression. Thus, the new model is expected to retain all the benefits of the STNN but improving it by providing extra knowledge to the system.

Overall, the model is the same as the STNN. Both the optimization problem and the training (loss function, learning algorithm, inference, etc) are applicable to the XSTNN model.

Now, let us consider a set of exogenous variables, Λ. This variables are temporal series, so they can be treated on the same way we did previously, meaning that Λ_t denotes the slice of Λ at time t. As in Section 3.2, we will denote Z as the latent space, X as the value of the own series and $\tilde{X}$ being the prediction of the STNN. There are several ways in which Λ could be introduced. We remark three options:

- Construct a new neural network that includes $\tilde{X}$ and Λ as inputs, generating the value of the final series as output. This output would be compared with the real value of the series.

- After using equation (3.5) but before decoding the latent space, change its value using Λ. This might be done with a linear mapping or a MLP.

- Change equation (3.5) so that the latent space is modified directly by Λ.

Because of the nature of the model, we believe that the last option is the most appropriate one. By introducing Λ in the estimation of Z_t, the model learns the dynamics taking into account external information too. As the premise of this work is to assume that exogenous variables might change the dynamic of the series, learning to mold the system in function of both meets our requirements the best.

Once the main idea has been explained, it is necessary to answer some other questions. Specifically, there are a few alternatives for reconstruct equation (3.5) in the way is intended. Moreover, a discussion about what time step to use with Λ is desirable: when computing Z_t, both Λ_t and Λ_{t+1} might be beneficial. The first one represents the idea of a previous state having an effect on the next one, whereas the second option symbolizes the conception of an actual state modifying the series.

Let us now introduce some possibilities. First, if exogenous data does not present spatial dependency, it can be more efficient to avoid the use of spatial relations for Λ. This version writes:

$$Z_{t+1} = h(Z_t\Theta^{(0)} + WZ_t\Theta^{(1)} + \Lambda_t\Theta^{(2)}) \tag{3.7}$$

On the contrary, when exogenous variables may exhibit spatial dependency, the same treatment that Z has will be provided to Λ. This notion is captured as follows:

$$Z_{t+1} = h(Z_t\Theta^{(0)} + WZ_t\Theta^{(1)} + \Lambda_t\Theta^{(2)} + W\Lambda_t\Theta^{(3)}) \tag{3.8}$$

A diagram that represents this last option is presented in figure 3.2.

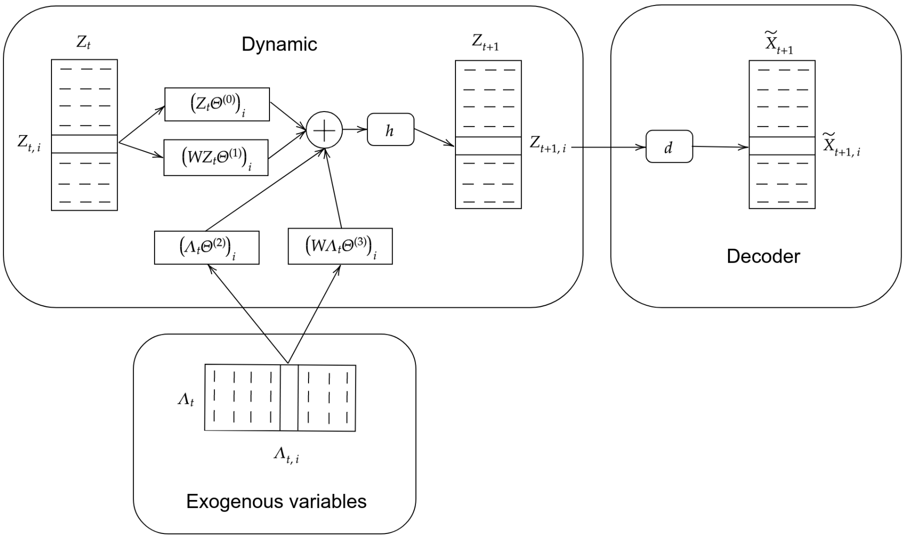


FIGURE 3.2: Architecture of the XSTNN model as described in Section 3.3.

3.3.1 Limitations

Before explaining the rest of the models, we would like to point out the two principal limitations of our model:

- Using an specific matrix W for a concrete problem means that, for different circumstances (for example, a different spatial grid), a retraining is needed.

- Both the dynamic and the decoder functions are stationary, meaning that they do not change over time. In [37] a method to tackle this problem is proposed.

3.4 Other models

For the sake of testing the new model, not only the base STNN but other methods are proposed for comparison. Through this section a few guidelines of these new methodologies are given. All of them are well known in the field of Data Science and statistics, consequently they only will be introduced and contextualized in our problem.

3.4.1 Mean

A simple-naive model which forecast new values of the series using the mean of past values from the same series. In other words, it uses the values of the T' training timesteps for computing the mean which will be used as the value of any new prediction. Formally:

$$\bar{X}_{T'} = \frac{1}{T'} \sum_{t} X_t \tag{3.9}$$

where $\bar{X}_{T'}$ denotes the mean computed for T' timesteps.

3.4.2 Persistence

The second simple-naive model which will be used in this project. In this case, the last value for each series (each neighborhood) is used for making the prediction, assigning it for all forecasted timesteps (T'). It writes as:

$$\tilde{X}_{T'} = X_T \tag{3.10}$$

where T is the last timestep from which we know the series value and $\tilde{X}_{T'}$ denotes the prediction as previously.

3.4.3 Linear regression

The core idea behind linear regression is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Typically, is described as follow:

$$y = X\beta \tag{3.11}$$

where y, X and β are matrix denoting response variables, regressors and regression coefficients respectively.

Compared with the rest of the models, a change of notation has been introduced. Concretely, y is the value of the series, while X represents all the variables explained in Chapter 2 (time, space, traffic and meteorology).

Although simple, its capability of forecasting relies in the fact that a linear relation between y and X is expected.

3.4.4 XGBoost

Tree-Based Models create a single (e.g. for CART) or many (e.g. Random Forest or XGBoost) decision trees which create conditional ”splits” in the data to arrive at their predictions. For the models which create many trees, the model averages the predictions of all the trees to create its final prediction. These models are very robust and perform quite well in several applications [18].

Particularly, XGBoost was proposed by Tianqi Chen [5]. It is based on the concept of Gradient Boosting (GB): while SGD optimize the parameters of some fixed architecture, GB does not assume any fixed architecture learning both, the function that best approximate the data and its parameters. Usually, for regression problems RMSE and regularization are used as loss functions. When the model to fit is bounded to tree models, they are called Gradient Boosted Trees. Specifically, XGBoost is one of the fastest implementations of gradient boosted trees.

It does this by tackling one of the major inefficiencies of gradient boosted trees: considering the potential loss for all possible splits to create a new branch (especially if you consider the case where there are thousands of features, and therefore thousands of possible splits). XGBoost works over this inefficiency by looking at the distribution of features across all data points in a leaf and using this information to reduce the search space of possible feature splits.

However, this kind of algorithms tend to overfit, so a correct adjustment of their parameters through a parameter-tuning process is necessary.

Chapter 4

Experimental settings

In this chapter we describe the decisions taken and experiments conducted in order to evaluate the performance of the several approaches introduced in Chapter 3.

4.1 Introduction, purpose and organization

In tune with the nature of this project, the proposed experiments for this thesis tries to evaluate the actual performance of several methodologies for spatio-temporal regression in the special case of traffic accidents. Therefore, the different models presented in Chapter 3 will be trained and evaluated in the dataset advanced in Chapter 2. All these methods will be tested in a spatio-temporal context, with the same conditions in terms of space zones and timesteps. In addition, all aspects concerning to the experiments must be as similar as possible among all approaches. In case any model needs a different treatment for any detail, it will be indicated.

In practice, we proceed as follows:

- Firstly, an architecture is chosen for every model when necessary. For that end, an hyper-parametrization (neural models) and parameter tuning (XGBoost) is executed. The validation scheme used along this procedure is explained in Section 4.2.
- Secondly, the training is conducted. The STNN and XSTNN models use spatial dependencies by nature. XGBoost can make use of neighborhoods as a variable, and naive models needs to be executed in every spatial zone as a different series. Again, the final result is evaluated using the validation scheme presented in Section 4.2.

To evaluate the accuracy and precision of the prediction, we selected Mean Absolute Error (MAE) and Bias as our metrics. In a spatio-temporal context [29], they are defined as:

$$MAE = \frac{1}{TS}\sum_{j=1}^{T}\sum_{i=1}^{S} \mid x_{s_i;t_j} - \tilde{x}_{s_i;t_j} \mid \tag{4.1}$$

$$Bias = \frac{1}{TS}\sum_{j=1}^{T}\sum_{i=1}^{S}(x_{s_i;t_j} - \tilde{x}_{s_i;t_j}) \tag{4.2}$$

where, as it was defined in Section 1.5, $x(s_i; t_j) : j = 1, ..., T; i = 1, ..., S$ is a spatio-temporal sample from the real series, $\tilde{x}(s_i; t_j)$ makes reference to the predicted series, S is the total number of spatial grids and T the total number of timesteps.

We set up the neural networks experiments on *Google Colab* [1]. For the other three models, a external machine proportionated by *Departamento de Inteligencia Artificial, UNED* [2] was used. The STNN and the XSTNN [3] were built upon PyTorch. The Mean, Persistence, linear regression and XGboost models are built upon R, the last one made use of the package *xgboost*[5].

4.2 Validation

In time series, some typical validation methods are not recommended. Concretely, when dealing with this type of data, traditional schemes like cross-validation (k-fold) or holdout should not be used for two reasons:

- **Temporal dependencies:** With time series data, particular care must be taken in splitting the data in order to prevent data leakage. For example, it could happen that the test set is a predecessor of the train set. Additionally, our proposed model has as central axis of its functioning the learning of the temporal dynamics of the series, lacking sense to test a set of data prior to the trained one.

- **Arbitrary choice of Test Set:** The choice of the test set is fairly arbitrary, and that choice may mean that the test set error is a poor estimation of error on an independent test set.

To validate the different proposed methodologies, a time series cross-validation scheme called rolling origin is used [26]. Rolling origin is an evaluation technique according to which the forecasting origin is updated successively and the forecasts are produced from each origin. This technique allows obtaining several forecast errors for time series, which gives a better understanding of how the models perform. There are different options of how this can be done, showing figure 4.1 the one we have chosen for this work.

Notice that in this figure 4.1 and in our proposal for the validation scheme, several models are trained for each method presented in Chapter 3. These models have an increasing in-sample size. However, and as it was pointed out before, it is not the only way. Another option is the constant in-sample validation, meaning that the train set has always the same size and its origin changes at the same time as the test set does.

Furthermore, if the train set origin is always the same (as we did), this could be considered as a rolling origin with a constant holdout sample size. Otherwise, it is called non-constant holdout sample size.

[1] https://colab.research.google.com/

[2] http://www.ia.uned.es/

[3] Code available at https://github.com/rdemedrano/xstnn

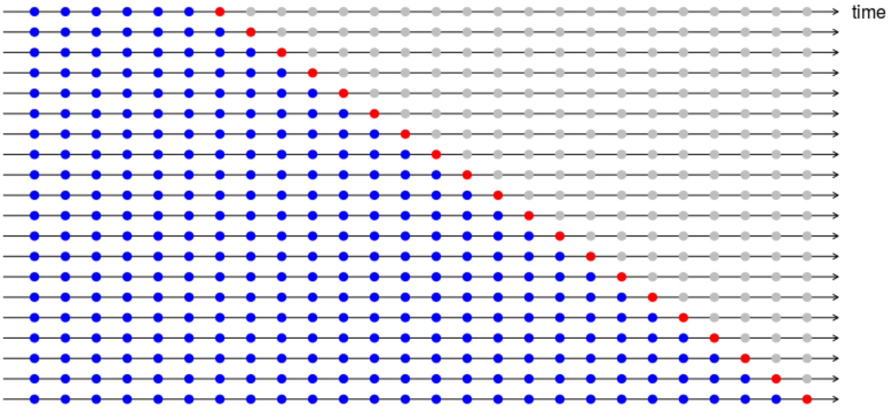


FIGURE 4.1: An example of rolling origin cross-validation for time series. Blue dots represent the train set, whereas red dots show the test set. Figure from [7].

In general, this validation scheme is considered robust and an almost unbiased estimation of the true error [7]. Nevertheless, its computational cost and the need to train several models are its disadvantages.

4.2.1 Setup for the experiments

Let us now describe how the previous procedure is applied in our own experiments. Consider the following steps:

1. The traffic accidents dataset is splitted in 10 successive sets, that is to say, starting all sets from 1 of january of 2018 at 00:00, each of those ten sets end at a different date between 14 of february at 23:00 and 31 of december at 18:00. To consider: all datasets are equally spaced and a minimun of 45 days has been set for training.

2. As a test set, we consider predictions within a + 5 horizon. For a train set of T timesteps, this means that the evaluation of the quality of the model will be made over $T + 1$ to $T + 5$ timesteps.

3. Finally, the 10 split sets are trained and validate over $T + 1$ to $T + 5$ timesteps. The final error is the average of all validations. The datasets have been chosen with the purpose that different hours and week days are tested for a more complete and extensible validation.

Again, this procedure is equivalent for all models presented in Chapter 3.

At this point, it should be clear that the total error for the entire space throughout the complete test time interval is the average of each error over $T + 1$ to $T + 5$ timesteps. Solely comment that a different error is calculated for every neighborhood, being the total error for a timestep T the average of all spatial grids at that same time T.

4.3 Hyper-parametrization and parameter tuning

In order to achieve the best possible results on our model, and also on our baselines, we grid-searched hyper-parameters on each model. Hence, each hyper-parameter is selected by the cross-validation approach presented before. First, we will present both the possibilities and chosen values. After that, we will discuss some of the elections. We detail below the values tested for each hyper-parameter in table 4.1:

STNN	Learning rate	0.0001, 0.003, 0.01, 0.1, 1
	λ	0.001, 0.01, 0.1, 1, 10
	n_z	1, 2, 3, 5, 10
	g(Z)	Linear, MLP(2,2), MLP(5,5), MLP(10,10), MLP(20,20)
	Minibatch size	128, 256, 512, 1024
	Dropout	0, 0.25, 0.35, 0.5, 0.8
XSTNN	Learning rate	0.0001, 0.003, 0.01, 0.1, 1
	λ	0.001, 0.01, 0.1, 1, 10
	n_z	1, 2, 3, 5, 10
	g(Z)	Linear, MLP(2,2), MLP(5,5), MLP(10,10), MLP(20,20)
	Minibatch size	128, 256, 512, 1024
	Dropout	0, 0.25, 0.35, 0.5, 0.8
XGBoost	Number of rounds	40, 60, 80, 100, 120
	Max. depth	1, 5, 10, 15, 20
	η	0.0001, 0.001, 0.01, 0.1, 1
	γ	0, 1, 2, 3, 4
	Min. child weight	0, 0.5, 1, 1.5, 2
	Subsample	0, 0.2, 0.5, 0.7, 1

TABLE 4.1: Values tested for each hyper-parameter. n_z is the dimension of the latent space. The remaining variables were presented in Chapter 3 or are commonly used parameters.

After the validation process, the values that have been chosen as the best alternative are presented in table 4.2.

Let us point out that, although MAE has been the principal metric, computational cost has been determinant too. It is the case of, for example, minibatch size and $g(Z)$: despite the fact that several values exhibited similar or slightly better performance, the computational effort introduced by them makes these hyper-parameters worse candidates.

Specially intriguing is the fact of a linear function being capable of modeling the dynamics of the system $g(Z)$ with such a good performance. It is expected that, with enough computational resources, a deeper MLP could overperformed the linear function. Additionally, new functions that represent the dynamics could be proposed as was mentioned in Chapter 3.

As it can be seen from table 4.2, the λ parameter is smaller in the case of the STNN. This means that this model will be more permissive with the dynamic's part of the loss in (3.1).

STNN	Learning rate	0.01
	λ	0.01
	n_z	2
	$g(Z)$	Linear
	Minibatch size	512
	Dropout	0.25
XSTNN	Learning rate	0.01
	λ	0.1
	n_z	2
	$g(Z)$	Linear
	Minibatch size	512
	Dropout	0.35
XGBoost	Number of rounds	80
	Max. depth	15
	η	0.1
	γ	1
	Min. child weight	1
	Subsample	0.7

TABLE 4.2: Values chosen for each hyper-parameter.

Lastly, both models show a good performance with $n_z = 2$, allowing us to reckon that there is no need of a specially high multi-dimensional latent space for traffic accidents.

It is important to point out that both neural model uses same optimizer parameters. Concretely, an early-stopping approach using Adam optimizer with the settings: $\beta_1 = 0.0$, $\beta_2 = 0.999$, $\epsilon = 10^{-9}$ and $w_d = 10^{-6}$ for both methodologies.

4.4 Spatial relations

To close this chapter, let us discuss the election for spatial relations of the model. Explained in Chapter 3, equations (3.5) and (3.8) precise spatial information summarised in a matrix W. As this matrix will content all the information related to spatial zones and their relation, it is important to construct it in a way that let us gain (or not lose) as much knowledge as possible.

For this reason, it has been decided to use the inverse of spatial distance as the main metric. Thus, all zones are in some way related but in a bigger degree the closer they are. The precise matrix is illustrated in figure 4.2.

Notice matrix W is normalized by row in order to avoid vanishing or exploding problems. If not, it would be possible a highly predominance of certain zones over the rest, creating decompensations in the predictions.

Other options for this matrix could be a simple adjacency matrix (1 if two spatial zones are colliding, 0 in any other case) or a representation of a graph structure.

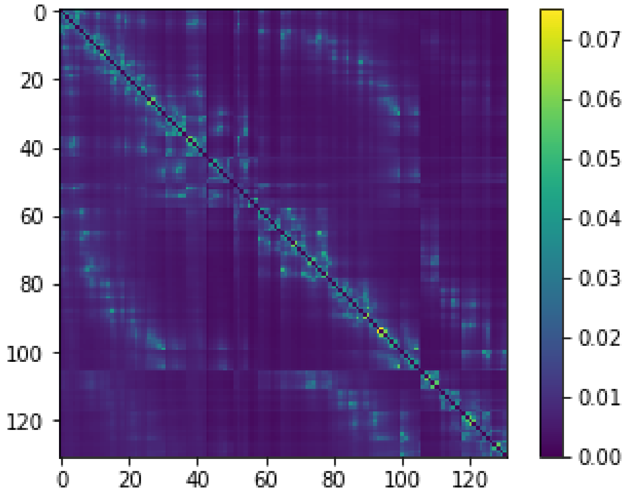


FIGURE 4.2: Spatial relations used during the experiments. Representation of matrix W.

Chapter 5

Results and discussion

The following chapter presents results and discussion on the different experiments and settings exposed in Chapter 4.

Before explaining the results, we will establish what questions we wish to answer. They are stated as follow: (1) Are the results of the proposed model better when compared with benchmark methods, including classical predictive models, tree-based models and STNN? (2) Is our proposed model capable of managing different spatial regions or timesteps? (3) Do the forecasting results make sense? Does our model provide more insights on the problem? (4) Are the predicted accident locations correlated with the ground truth spatially?

Through this questions, we expect to evaluate if the XSTNN model supposes a step forward in the prediction of traffic accidents.

5.1 General results

In order to identify in a quantitative way the performance of the different models and baselines, table 5.1 provides the average prediction error for $T+1$ to $T+5$. From this first insight it should be clear that both STNN and XSTNN outperform the other models. As Mean model, Persistence model and XGboost were trained taking into account the existence of a spatial grid but without establishing relations between them, these results confirm that making use of prior spatial information is beneficial for the regression problem. Beyond that, the XSTNN presents a better performance than its base model, the STNN.

Model	MAE	Bias
XSTNN	$\mathbf{0.0041 \pm 0.0006}$	-0.0006 ± 0.0004
STNN	0.0045 ± 0.0006	-0.0004 ± 0.0006
XGBoost	0.0052 ± 0.0006	0.0004 ± 0.0006
Linear regression	0.0050 ± 0.0006	0.0002 ± 0.0007
Mean	0.0052 ± 0.0007	0.0003 ± 0.0007
Persistence	0.0055 ± 0.0008	0.0006 ± 0.0007

TABLE 5.1: Performance for $T+1$ to $T+5$ traffic accident regression.

For a more detailed vision, Fig. 5.1 shows the distribution of the metrics and the average error by timestep. From this figure, same conclusions can be extracted as before: the XSTNN model presents a better general behaviour compared to the rest of the models. Again, the fact of introducing spatial knowledge to the problem stands as an appropriated approach for this particular series, and our results reinforce the idea that introducing exogenous variables is favorable for the regression problem. However, it is worth noting that there is not a clear relation between errors and timestep. Although an increment on the error by timestep in the prediction is usually expected (cumulative error), the randomness of traffic accidents do not let us extract clear conclusions from this aspect.

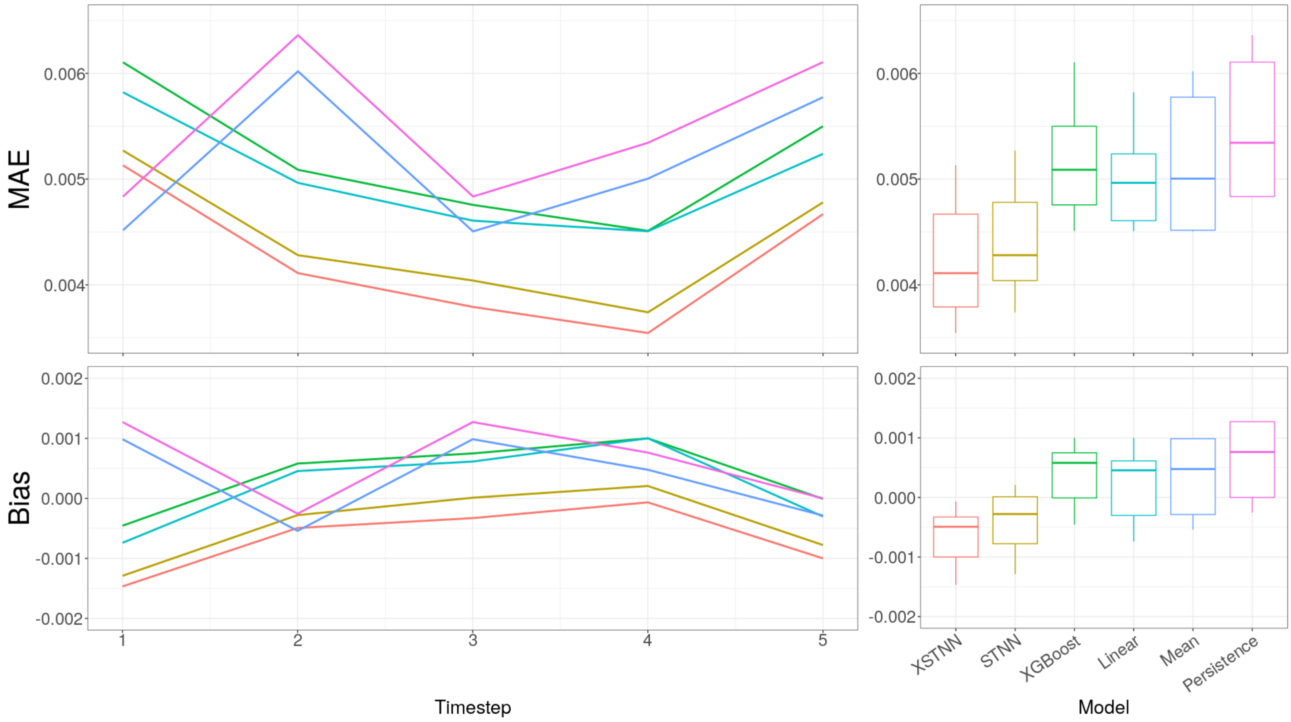


FIGURE 5.1: Forecasting performance (MAE and bias) of the different models by timestep together with the calculated distributions.

5.2 Reasoning in an spatio-temporal dimension

Beyond the quantitative analysis, now we show some accomplishments from our proposed model respect to the STNN. For that purpose, we will take a deeper look into a concrete example, without loss of generality.

Let us introduce the following situation: we forecast the accident regression series from 17 p.m. to 21 p.m. on a Wednesday. From Fig. 2.3 we know this situation corresponds to a high risk circumstance for traffic accidents to happen. In this context, Fig. 5.2 illustrate a comparison of our two principal models with a levelplot (time in x axis, neighborhoods in y axis and coloured by traffic accidents). The relation between neighborhoods and their correspondent number can be found in Appendix A. Let us expose several ideas:

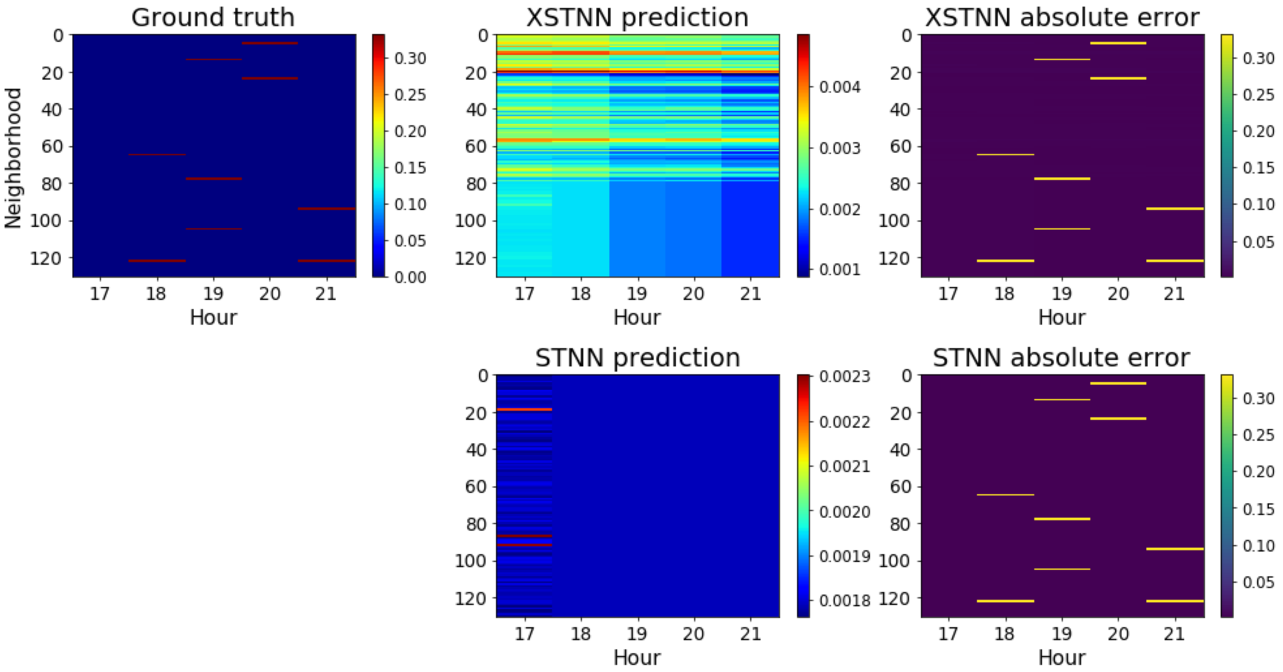


FIGURE 5.2: A practical example of the operation of both networks, XSTNN and STNN, for a same situation. From 17 p.m. to 21 p.m. on a Wednesday.

First of all, and unfortunately, the regression problem is far from being solved. A comparison of colorbars from both, STNN and XSTNN predictions, with the ground truth corroborates this statement. As Chen et. al. has documented, after some analysis of traffic accident data, it is difficult to predict whether traffic accidents will happen or not directly, because complex factors can affect traffic accidents, and some factors, such as the distraction of drivers, cannot be observed and collected in advance [4]. Nevertheless, our XSTNN model has proved to be a new step in the right direction, outperforming the rest of baselines models (table 5.1).

Secondly, the next natural question that rises is about the reason of this improvement. Again, Fig. 5.2 sheds light on this matter. Whereas the STNN quickly truncates its values close to 0 for every neighborhood and timestep, the XSTNN takes some risks and it is able to differentiate between time intervals and spatial zones. As the most likely situation is having no accidents for each hour and neighborhood, both networks have values approaching to 0 as outputs.

Certainly, taking more risks does not ensure a better performance in the regression problem. It is necessary that the model manages to elucidate which time intervals and neighborhoods are more important for the problem that we have in hand as a function of past events. In this concrete case, the model has learned to prioritize neighborhoods from 1 to 80, as they report a vast majority of the total number of traffic accidents in the city of Madrid. Besides, the XSTNN reveals a negative trend over the hours as we would expect.

As XSTNN learns better to distinguish between time ranges and spatial zones, it is possible to find other situations in which, again, this model offers more information and assimilates the system's dynamics in a better way. For example, and to corroborate that the XSTNN behaves better in a variety of situations, Fig. 5.3 gives evidence of a totally different state on a Sunday from 6 a.m. to 10 a.m. In this context, we will expect a higher risk at last late hours and at past 9 a.m., the XSTNN correspondingly adapting its output to this situation. On the contrary, the STNN is not capable of learning the corresponding dynamic. Unlike previously (Fig. 5.2), this time the XSTNN takes less risks and its output is closer to 0 as we would expect less accidents on a Sunday morning that a Wednesday on the evening as before.

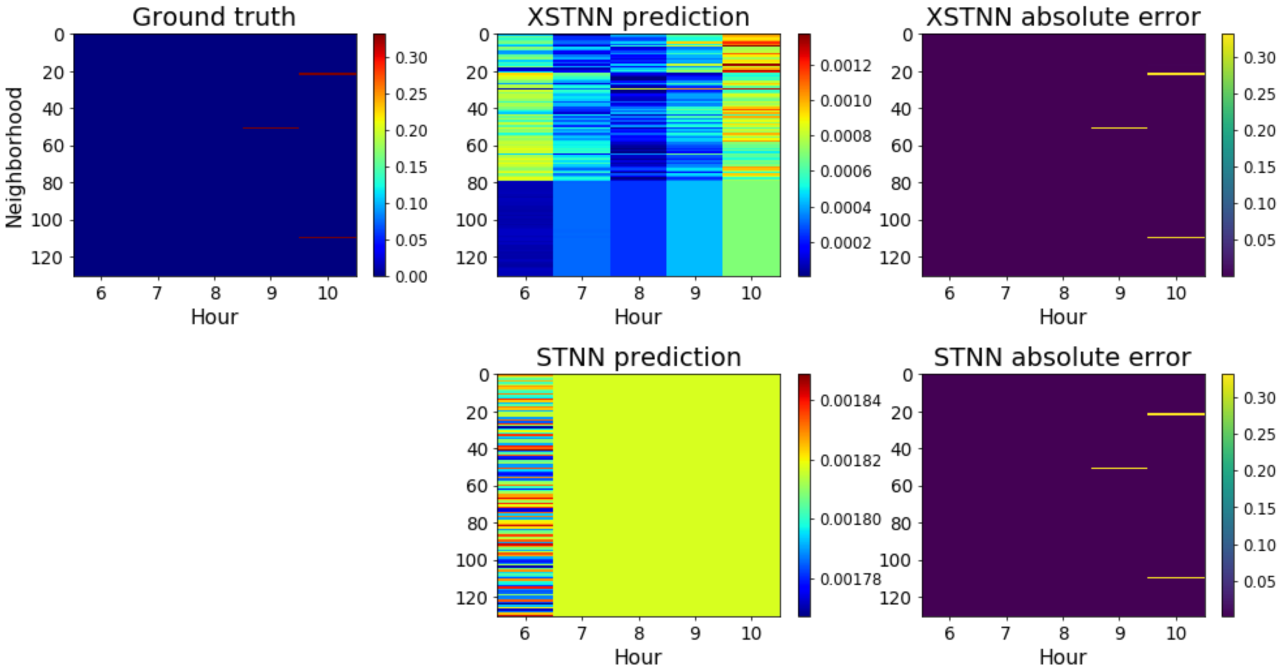


FIGURE 5.3: A practical example of the operation of both networks, XSTNN and STNN, for a same situation. From 6 a.m. to 10 a.m. on a Sunday.

5.3 Spatial dependency

Through the previous discussion in Section 5.2, we have pointed out how the XSTNN infers properties based on the time condition and the concrete spacial zone. For this last case, Fig. 5.4 offers an analysis of spatial risk for each neighborhood. Both series, the real and the predicted ones were reescaled for a direct comparison between them. This way, it is clear that the XSTNN is capable of reasoning in both dimension, temporal and spatial.

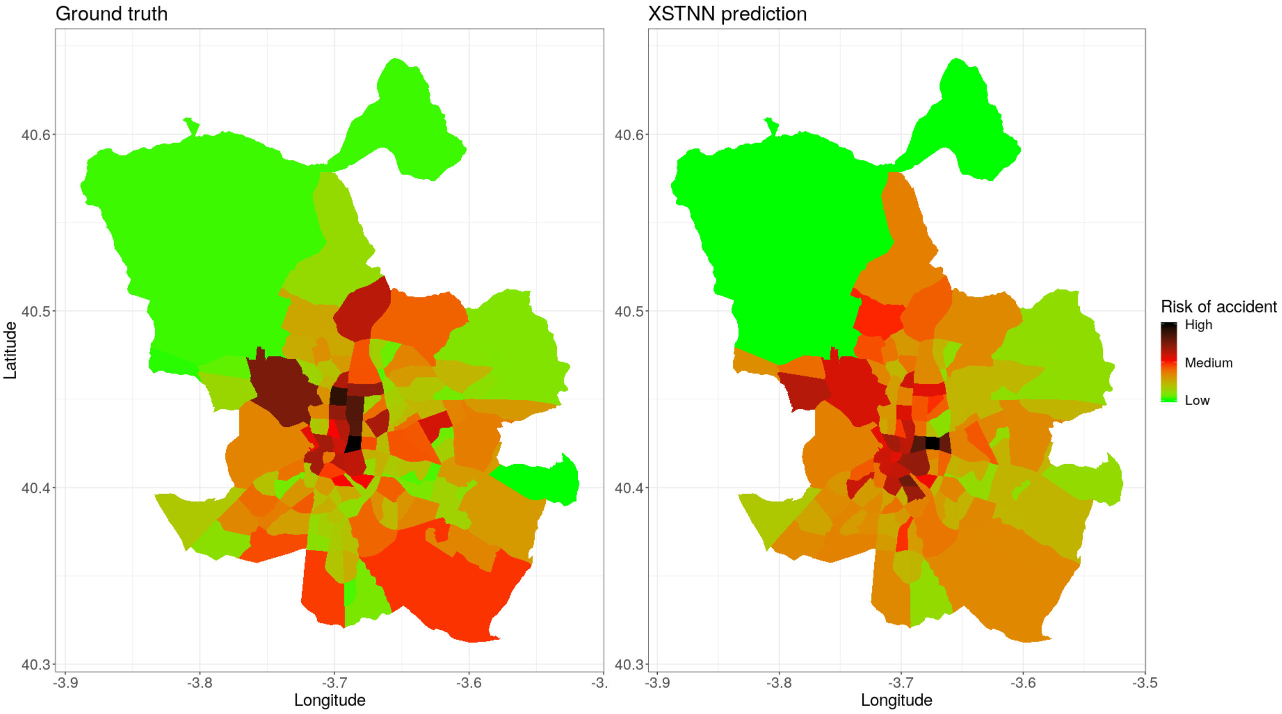


FIGURE 5.4: Spacial risk in the same scale for the ground truth (left) and the XSTNN (right).

In summary, the XSTNN reports a better understanding and learning of the dynamic of the system, being more flexible and creative in its prediction. These features translate into a better performance than their direct rivals and let us answer the questions we established at the beginning of this chapter in a positive way.

5.4 Feature importance

With respect to feature importance, the XGBoost method lets us get insight from this matter in a simple way. By plotting the contribution from every variable to the final tree, we can see these results. Concretely, Fig. 5.5 shows this idea.

Although this last figure reveals an expected dependency on spatial location and traffic intensity, the XGBoost method is relying primarily on features that are closely associated with the traffic condition, resting importance from the rest of variables. While decisions like those might oversimplify the model, Fig. 5.5 supports our initial hypothesis of the importance that external variables and a good spatial distribution might have in the modeling of traffic accidents.

Nevertheless, it is important to point out that feature importance not necessarily manifest a direct relation between the predictor and the prediction. In our case, while temperature or solar radiation might be seen as important contributors, it is more likely that day-night cycle, season, time of the day or the fact that the sun could be in driver's eyes are the real causes/explanations.

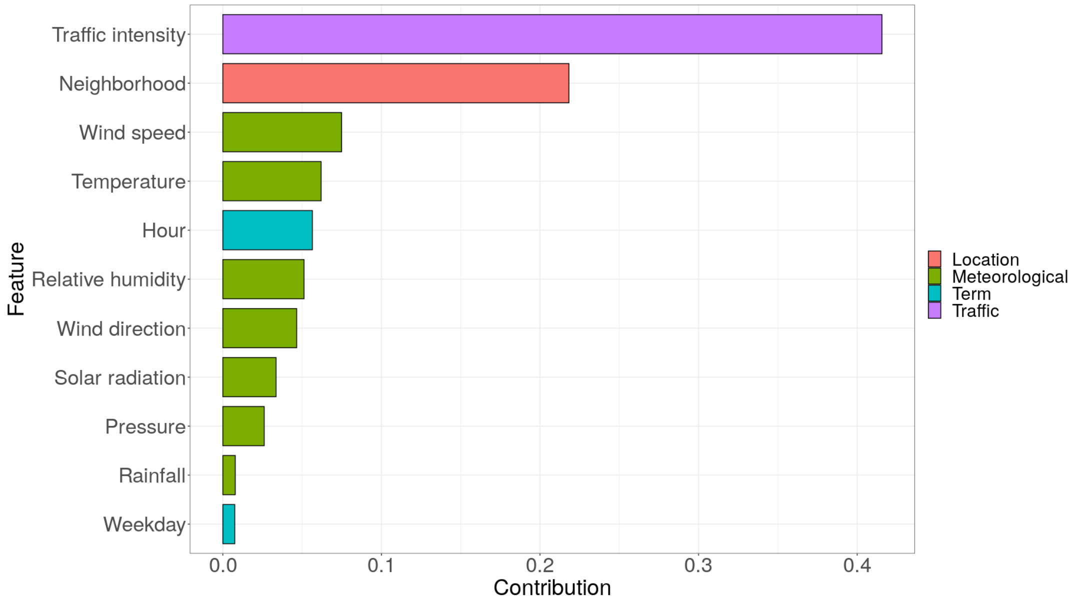


FIGURE 5.5: Feature contribution by type of data.

5.5 Reproducibility

As it was pointed out in Section 4.1, the code for our model is available at `https://github.com/rdemedrano/xstnn`. An example of its use can be found there too. Also, Chapter 4 explains in detail all the experimental process to obtain our results.

Chapter 6

Conclusions, contributions, future research and ethical aspects

Through this work, a new approach for spatio-temporal series forecasting called XSTNN has been proposed. The problem of traffic accidents prediction was tackled by this new neural network model, showing a better performance than the rest of baselines models. Although traffic accidents regression is challenging due to several difficulties, the XSTNN has proved to stand out for its capability of providing a deeper insight in the problem series. Thus, this thesis demonstrate that spatio-temporal neural networks are a promising field for traffic accident prediction in the future.

We would like to remark our two principal objectives, established in Section 1.4:

- To explore the existing methodologies and propose new ones for forecasting traffic accident spatio-temporal series.
- To provide a practical case study in the city of Madrid.

In general, these objectives can be considered fulfill. After applying two new neural models that had never been used in traffic accidents, one of which was proposed for this concrete work, we have now a deeper insight of this phenomenon in the city of Madrid adapting the network to the rhythms of life and particularities of this city.

As future lines of research that are direct extensions of this project, there are several paths that could be followed both for new spatio-temporal models and traffic accidents prediction:

- The XSTNN model might be extended by introducing more temporal terms from exogenous series for updating the latent space, as this could be beneficial. This way, more complicated time-relations could be explored.
- In the same way, several spatial relations might be used at once, improving the learning of the spatial dimension.
- Given that exogenous variables have shown to be helpful, future work in this model can be extended to incorporate other features that are not necessarily series, like economic or demographic variables.

Nevertheless, the study of traffic accidents might be expanded in much more fronts. As it was established in Section 1.1, human factors (as distractions or consumption of alcoholic substances) represent a major number of accident's causes, meaning that a complete forecasting system might need to include them in its dynamic. Although driver monitoring systems like [9, 27] are not new (and even some of them are commercially available), most of them show an error rate too high to be fully useful in real life. Moreover, this kind of systems are usually focused in several variables (frequently physicals, as open-eyes or hands on the steering wheel) but lack information on other relevant aspects, such as the possibility of having used illegal substances or the increase in driver fatigue.

As it has just been explained, driver monitoring systems try to avoid future accidents, but a continuous monitoring might be beneficial for the sake of forecasting the odds of traffic accidents happening in function of human behaviour.

6.1 Ethical aspects

Lastly, we would like to provide some insight in the ethical and social repercussions that traffic accidents research (as this project) could have and that should be covered in the near future:

- First, there is a close relation between accidents and criminalization. It is very common to prevent accidents by criminalizing certain ways of acting, as speed limits or the use of cell phones while driving. While these measures might be beneficial, it is important to understand and analyse with caution the results obtained in order to avoid mistakes by criminalizing conducts that might not be part of the problem. For example, Section 5.4 shows that temperature might be related with accidents, but this relation is probably due to the day-night cycle, making no sense to take action depending on whether there is more temperature or less.

- By improving accidents prediction, an emerging view that a major role can and should be played by institutions is reasonable, implying that it is unclear how much responsability each part should take (drivers and prediction models owners). Simplifying, it could be argued that if some institution find a particular future situation risky and at the end an accident happens, this institution should have taken steps for the accident not to happen.

- Finally, and closely related to the use of driver monitoring systems, the knowledge of someone's situation (use of cell phone, drunkenness...) from a gadget could be seen as a violation of their privacy. However, it could be recognised that the degree of risk associated with driving may imply that the expectation of privacy on the road is not reasonable, as it would be in our homes.

Appendix A

Madrid neighborhoods

In the following table, a correspondence between name and numeration for Madrid neighborhoods is given:

Number	Name
1	Palacio
2	Embajadores
3	Cortes
4	Justicia
5	Universidad
6	Sol
7	Imperial
8	Acacias
9	Chopera
10	Legazpi
11	Delicias
12	Palos de Moguer
13	Atocha
14	Pacífico
15	Adelfas
16	Estrella
17	Ibiza
18	Jerónimos
19	Niño Jesús
20	Recoletos
21	Goya
22	Fuente del Berro
23	Guindalera
24	Lista
25	Castellana
26	El Viso
27	Prosperidad
28	Ciudad Jardín
29	Hispanoamérica
30	Nueva España
31	Castilla
32	Bellas Vistas
33	Cuatro Caminos

34	Castillejos
35	Almenara
36	Valdeacederas
37	Berruguete
38	Gaztambide
39	Arapiles
40	Trafalgar
41	Almagro
42	Rios Rosas
43	Vallehermoso
44	El Pardo
45	Fuentelareina
46	Peñagrande
47	Pilar
48	La Paz
49	Valverde
50	Mirasierra
51	El Goloso
52	Casa de Campo
53	Argüelles
54	Ciudad Universitaria
55	Valdezarza
56	Valdemarín
57	El Plantío
58	Aravaca
59	Cármenes
60	Puerta del Angel
61	Lucero
62	Aluche
63	Campamento
64	Cuatro Vientos
65	Águilas
66	Comillas
67	Opañel
68	San Isidro
69	Vista Alegre
70	Puerta Bonita
71	Buenavista
72	Abrantes
73	Orcasitas
74	Orcasur
75	San Fermín
76	Almendrales
77	Moscardó
78	Zofío
79	Pradolongo
80	Entrevías

81	San Diego
82	Palomeras Bajas
83	Palomeras Sureste
84	Portazgo
85	Numancia
86	Pavones
87	Horcajo
88	Marroquina
89	Media Legua
90	Fontarrón
91	Vinateros
92	Ventas
93	Pueblo Nuevo
94	Quintana
95	Concepción
96	San Pascual
97	San Juan Bautista
98	Colina
99	Atalaya
100	Costillares
101	Palomas
102	Piovera
103	Canillas
104	Pinar del Rey
105	Apostol Santiago
106	Valdefuentes
107	San Ándres
108	San Cristobal
109	Butarque
110	Los Rosales
111	Los Ángeles
112	Casco Histórico de Vallecas
113	Santa Eugenia
114	Ensanche de Vallecas
115	Casco Histórico de Vicálvaro
116	Valdebernardo
117	Valderribas
118	Cañaveral
119	Simancas
120	Hellín
121	Amposta
122	Arcos
123	Rosas
124	Rejas
125	Canillejas
126	Salvador
127	Alameda de Osuna

128	Aeropuerto
129	Casco Histórico de Barajas
130	Timón
131	Corralejos

Bibliography

[1] Mohamed A. Abdel-Aty and A. Essam Radwan. "Modeling traffic accident occurrence and involvement". In: *Accident Analysis & Prevention* 32.5 (Sept. 1, 2000), pp. 633–642. ISSN: 0001-4575. DOI: 10.1016/S0001-4575(99)00094-9.

[2] Michael Batty. "Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies". In: *CASA Working Papers (131). Centre for Advanced Spatial Analysis (UCL), London, UK.* (June 2012).

[3] Chen Chen. "Analysis and Forecast of Traffic Accident Big Data". In: *ITM Web of Conferences* 12 (2017), p. 04029. ISSN: 2271-2097. DOI: 10.1051/itmconf/20171204029.

[4] Quanjun Chen et al. "Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference". In: *AAAI*. 2016.

[5] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (2016), pp. 785–794. DOI: 10.1145/2939672.2939785. arXiv: 1603.02754.

[6] Provost D. Duer G. Ni S. *Automobile Collision Prediction in Louisville, KY*.

[7] R. J. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. URL: https://0texts.com/fpp2/.

[8] Frank J. Kelly and Julia C. Fussell. "Air pollution and public health: emerging hazards and improved understanding of risk". In: *Environmental Geochemistry and Health* 37.4 (2015), pp. 631–649.

[9] Neslihan Kose et al. "Real-Time Driver State Monitoring Using a CNN Based Spatio-Temporal Approach". In: *arXiv:1907.08009 [cs, eess]* (July 18, 2019). arXiv: 1907.08009. URL: http://arxiv.org/abs/1907.08009.

[10] *Las distracciones causan uno de cada tres accidentes mortales*. URL: http://www.dgt.es/es/prensa/notas-de-prensa/2018/20180917_campana_distracciones.shtml.

[11] Xiugang Li et al. "Predicting motor vehicle crashes using Support Vector Machine models". In: *Accident Analysis & Prevention* 40.4 (July 1, 2008), pp. 1611–1618. ISSN: 0001-4575. DOI: 10.1016/j.aap.2008.04.010.

[12] Lei Lin, Qian Wang, and Adel W. Sadek. "A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction". In: *Transportation Research Part C: Emerging Technologies*. Engineering and Applied Sciences Optimization (OPT-i) - Professor Matthew G. Karlaftis Memorial Issue 55 (June 1, 2015), pp. 444–459. ISSN: 0968-090X. DOI: 10.1016/j.trc.2015.03.015.

[13] T. A. Litman. "Transportation Cost and Benefit Analysis: Techniques, Estimates and Implications". In: (2009).

[14] Dominique Lord. "Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter". In: *Accident Analysis & Prevention* 38.4 (2006), pp. 751 –766. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2006.02.001.

[15] Dominique Lord and Fred Mannering. "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives". In: *Transportation Research Part A: Policy and Practice* 44.5 (June 1, 2010), pp. 291–305. ISSN: 0965-8564. DOI: 10.1016/j.tra.2010.02.001.

[16] Fanny Malin, Ilkka Norros, and Satu Innamaa. "Accident risk of road and weather conditions on different road types". In: *Accident Analysis & Prevention* 122 (Jan. 1, 2019), pp. 181–188. ISSN: 0001-4575. DOI: 10.1016/j.aap.2018.10.014.

[17] Fred L. Mannering and Chandra R. Bhat. "Analytic methods in accident research: Methodological frontier and future directions". In: *Analytic Methods in Accident Research* 1 (Jan. 1, 2014), pp. 1–22. ISSN: 2213-6657. DOI: 10.1016/j.amar.2013.09.001.

[18] Didrik Nielsen. "Tree Boosting With XGBoost". In: *Master thesis* (2016), p. 110.

[19] Juan de Oña, Griselda López, and Joaquín Abellán. "Extracting decision rules from police accident reports through decision trees". In: *Accident Analysis & Prevention* 50 (Jan. 1, 2013), pp. 1151–1160. ISSN: 0001-4575. DOI: 10.1016/j.aap.2012.09.006.

[20] Juan de Oña, Randa Oqab Mujalli, and Francisco J. Calvo. "Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks". In: *Accident Analysis & Prevention* 43.1 (Jan. 1, 2011), pp. 402–411. ISSN: 0001-4575. DOI: 10.1016/j.aap.2010.09.010.

[21] M. Peden et al. "World Report on Road Traffic Injury Prevention". In: (2004).

[22] Honglei Ren et al. "A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction". In: *arXiv:1710.09543 [cs]*. Oct. 26, 2017. arXiv: 1710.09543.

[23] Kyoung-Ah Rhee et al. "Spatial regression analysis of traffic crashes in Seoul". In: *Accident Analysis & Prevention* 91 (June 1, 2016), pp. 190–199. ISSN: 0001-4575. DOI: 10.1016/j.aap.2016.02.023.

[24] Arash M. Roshandeh, Bismark R. D. K. Agbelie, and Yongdoo Lee. "Statistical modeling of total crash frequency at highway intersections". In: *Journal of Traffic and Transportation Engineering (English Edition)* 3.2 (Apr. 1, 2016), pp. 166–171. ISSN: 2095-7564. DOI: 10.1016/j.jtte.2016.03.003.

[25] D. Singh and C. K. Mohan. "Deep Spatio-Temporal Representation for Detection of Road Accidents Using Stacked Autoencoder". In: *IEEE Transactions on Intelligent Transportation Systems* 20.3 (Mar. 2019), pp. 879–887. ISSN: 1524-9050. DOI: 10.1109/TITS.2018.2835308.

[26] Leonard J. Tashman. "Out-of-sample tests of forecasting accuracy: an analysis and review". In: *International Journal of Forecasting*. The M3-Competition 16.4 (Oct. 1, 2000), pp. 437–450. ISSN: 0169-2070. DOI: 10.1016/S0169-2070(00)00065-0.

[27] F. Vicente et al. "Driver Gaze Tracking and Eyes Off the Road Detection System". In: *IEEE Transactions on Intelligent Transportation Systems* 16.4 (Aug. 2015), pp. 2014–2027. ISSN: 1524-9050. DOI: 10.1109/TITS.2015.2396031.

[28] *WHO | Data*. WHO. URL: http://www.who.int/violence_injury_prevention/road_safety_status/2015/GSRRS2015_data/en/.

[29] Christopher K. Wikle, Andrew Zammit-Mangion, and Noel Cressie. *Spatio-Temporal Statistics with R*. 1st ed. Boca Raton, Florida : CRC Press, [2019]: Chapman and Hall/CRC, Feb. 18, 2019. ISBN: 978-1-351-76972-3. DOI: 10.1201/9781351769723. URL: https://www.taylorfrancis.com/books/9780429649783.

[30] Daniel Wilson. *Using Machine Learning to Predict Car Accident Risk*. Medium. May 3, 2018.

[31] Pengpeng Xu and Helai Huang. "Modeling crash spatial heterogeneity: Random parameter versus geographically weighting". In: *Accident Analysis & Prevention* 75 (Feb. 1, 2015), pp. 16–25. ISSN: 0001-4575. DOI: 10.1016/j.aap.2014.10.020.

[32] Y. Yu, M. Xu, and J. Gu. "Vision-based traffic accident detection using sparse spatio-temporal features and weighted extreme learning machine". In: *IET Intelligent Transport Systems* 13.9 (2019), pp. 1417–1428. ISSN: 1751-956X. DOI: 10.1049/iet-its.2018.5409.

[33] Zhuoning Yuan, Xun Zhou, and Tianbao Yang. "Hetero-ConvLSTM: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatio-Temporal Data". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*. the 24th ACM SIGKDD International Conference. London, United Kingdom: ACM Press, 2018, pp. 984–992. ISBN: 978-1-4503-5552-0. DOI: 10.1145/3219819.3219922.

[34] Guangnan Zhang, Kelvin K. W. Yau, and Guanghan Chen. "Risk factors associated with traffic violations and accident severity in China". In: *Accident Analysis & Prevention* 59 (Oct. 1, 2013), pp. 18–25. ISSN: 0001-4575. DOI: 10.1016/j.aap.2013.05.004.

[35] Zhenhua Zhang et al. "A deep learning approach for detecting traffic accidents from social media data". In: *Transportation Research Part C: Emerging Technologies* 86 (Jan. 1, 2018), pp. 580–596. ISSN: 0968-090X. DOI: 10.1016/j.trc.2017.11.027.

[36] M. Zheng et al. "Traffic Accident's Severity Prediction: A Deep-Learning Approach-Based CNN Network". In: *IEEE Access* 7 (2019), pp. 39897–39910. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2903319.

[37] Ali Ziat et al. "Spatio-Temporal Neural Networks for Space-Time Series Forecasting and Relations Discovery". In: *2017 IEEE International Conference on Data Mining (ICDM)* (Nov. 2017), pp. 705–714. DOI: 10.1109/ICDM.2017.80. arXiv: 1804.08562.