applied sciences

MDPI

Article

Predictive Analyses of Traffic Level in the City of Barcelona: From ARIMA to eXtreme Gradient Boosting

Eloi Garcia [1], Laura Calvet [2,*], Patricia Carracedo [3], Carles Serrat [1], Pau Miró [3] and Mohammad Peyman [4]

[1] Department of Mathematics, Barcelona School of Building Constrution, Universitat Politècnica de Catalunya-BarcelonaTECH, 08028 Barcelona, Spain
[2] Department of Telecommunications and Systems Engineering, Autonomous University of Barcelona, 08193 Bellaterra, Spain
[3] Department of Statistics and Operations Research, Universitat Politècnica de València, 03801 Alcoy, Spain
[4] Research Center on Production Management and Engineering, Universitat Politècnica de València, Plaza Ferrandiz-Salvador, 03801 Alcoy, Spain
* Correspondence: laura.calvet.linan@uab.cat

Abstract: This study delves into the intricate dynamics of urban mobility, a pivotal aspect for policymakers, businesses, and communities alike. By deciphering patterns of movement within a city, stakeholders can craft targeted interventions to mitigate traffic congestion peaks, optimizing both resource allocation and individual travel routes. Focused on Barcelona, Spain, this paper draws on data sourced from the city council's open data service. Through a blend of exploratory analysis, visualization techniques, and modeling methodologies—including time series analysis and the eXtreme Gradient Boosting (XGBoost) algorithm—the research endeavors to forecast traffic conditions. Additionally, a study of variable importance is carried out, and Shapley Additive Explanations are applied to enhance the interpretability of model outputs. Findings underscore the limitations of traditional forecasting methods in capturing the nuanced spatial and temporal dependencies present in traffic flows, particularly over medium- to long-term horizons. However, the XGBoost model demonstrates robust performance, with the area under ROC curves consistently exceeding 80%, indicating its efficacy in handling non-linear traffic data variables.

Keywords: traffic level; eXtreme Gradient Boosting; forecasting; mobility; open data

Citation: Garcia, E.; Calvet, L.; Carracedo, P.; Serrat, C.; Miró, P.; Peyman, M. Predictive Analyses of Traffic Level in the City of Barcelona: From ARIMA to eXtreme Gradient Boosting. *Appl. Sci.* **2024**, *14*, 4432. https://doi.org/10.3390/app14114432

Academic Editor: Roberto Carballedo

Received: 2 April 2024
Revised: 16 May 2024
Accepted: 21 May 2024
Published: 23 May 2024

1. Introduction

Mobility has evolved as a hallmark of modernity around the world. A key policy objective in recent decades has been for citizens to move faster and further more safely and more comfortably. In this process, urban mobility has played a fundamental role, which is strongly related to mass motorization. European drivers spend around 20 min on average behind the wheel daily on personal trips on working days and between 25 and 30 min on business trips [1]. Barcelona is the second most populous city in Spain and the seventh in the European Union (EU) [2]. According to [3], there were more than 6 million daily displacements in 2017; 35.3% corresponds to active mobility, 40.1% to public transport, and 24.6% to private transport (67.7% and 29.8% correspond to car and motorcycle, respectively).

Unfortunately, mobility produces negative effects such as congestion, pollution, accidents, deaths from traffic accidents, and greenhouse gas emissions [4,5]. In order to satisfy its new climate objectives, the EU has launched a series of political initiatives to reduce the negative effects of cars and, in turn, promote public transport [6].

In this context, accurate real-time prediction of traffic conditions using statistical and Artificial Intelligence (AI) methods is of vital importance for road users, private sectors, and governments. For example, identifying commuting trends may help to decide in which areas public transportation should be reinforced. Knowing mobility trends and urban

design may enable the selection of the optimal number and location of public charging stations for electric vehicles, which may increase the use of these vehicles. For companies, being able to estimate their social and environmental impacts, combined with intelligent methods, helps them to design distribution processes that minimize these impacts, which tend to be correlated to economical costs [7]. Thus, high quality data are potentially useful for modeling and addressing optimization problems related to urban mobility [8].

Regarding the availability of data, there is an increasing number of initiatives aiming to share data for the common good and for the benefit of anyone. For instance, the city of Chicago has an open data portal (https://data.cityofchicago.org/, accessed on 2 April 2024) which enables the creation of maps and graphs to gain insights about plenty of facts (salaries, violence, vaccinations, etc.). One of the advantages of these types of portals is that data are frequently updated (faster than the statistics typically offered by national statistics institutes). There are also initiatives proposed by private companies such as Carbon Footprint Ltd. (https://www.carbonfootprint.com/aboutus.html, accessed on 2 April 2024), which offers some online calculators to estimate the average carbon footprint of individuals and small businesses for free. Despite the emergence of these initiatives, ref. [5] states that there is a lack of sufficiently granular indicators and related data on urban mobility (such as modal split, environmental impacts, congestion, energy use). The reason is that cities are not required to collect and provide these data. However, the increasing adoption of Internet of Things technologies (e.g., intelligent traffic lights, road sensors, or sensors in trash and recycling containers) and awareness of the potential of open data portals suggest that the availability of data will continue to increase during the next years.

Traffic modelization and prediction constitute an arduous endeavor because of the high non-linearity and complexity of traffic flow. Recently, classic statistical models have been challenged by machine learning and deep learning methods in traffic prediction tasks [9]. This is due to the fact that traditional methods cannot make predictions in the medium–long term and fail to consider the spatial and temporal dependencies in the data [10].

In this context, the purpose of this article is to examine the traffic levels in Barcelona utilizing data from the Ajuntament de Barcelona's open data service, employing both classic and machine learning methods. The primary contributions include conducting an exploratory analysis of the traffic data with visualization techniques, developing predictive models utilizing both time series analysis and the eXtreme Gradient Boosting (XGBoost) algorithm, and examining the results along with a discussion. The specific flow of the research methodology employed in this article is depicted in Figure 1. To our knowledge, this study represents the first attempt to analyze open data on traffic levels utilizing visualization techniques, time series analysis, and the XGBoost algorithm.

Next, we outline the structure of the document. Section 2 reviews related work, focusing on traffic flow studies, prediction, and models in Barcelona. Section 3 details the methodology, encompassing visualization techniques, time series analysis, and the XGBoost algorithm. Section 4 outlines the application, detailing the dataset employed and presenting and discussing the results obtained through the various components of the methodology. Section 5 provides a discussion based on the results obtained in the previous sections. Finally, Section 6 draws pertinent conclusions and outlines potential lines of future research.

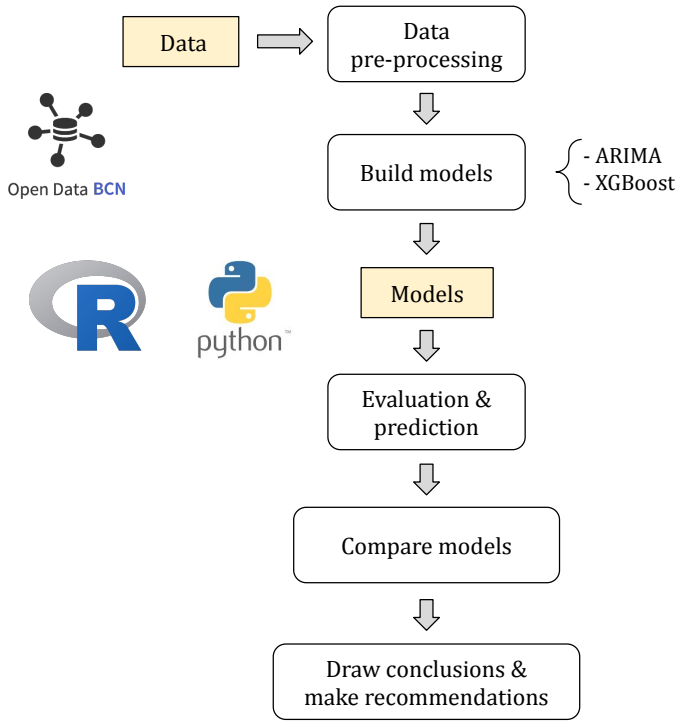


Figure 1. Diagram of research flow.

2. Related Work

2.1. Traffic Flow Prediction

Traffic flow prediction offers accurate projections of traffic volume within a specified area at future time intervals. The exploration of traffic forecasting proves valuable in alleviating congestion and promoting safer, more cost-effective travel [11]. The number of studies on traffic flow prediction has increased significantly during recent years due to a rising awareness of the environmental and social impacts of the increasing traffic congestion in most cities and the development of big data and deep learning methodologies. There are different approaches to predict traffic flow.

For instance, recently, ref. [12] reviewed and compared hybrid deep learning models. The authors recognized the escalating complexity of models, which now incorporate a growing array of finer-grained, multi-type data sourced from transportation systems. Most of the approaches reviewed use convolutional neural networks, recurrent neural networks, and long- and short-term memory units. Ref. [13] provided a quick review of machine learning algorithms for short-term traffic forecasting and introduced the challenges, followed by various ways for modeling temporal and/or geographical dependencies. Also, ref. [14] integrated a bootstrap methodology with the conventional parametric ARIMA model to form an ensemble approach. This ensemble, derived from random subsamples of data, aims to enhance prediction accuracy while maintaining adherence to theoretical principles.

The traffic flow data are classified as similar, volatile, and irregular parts. Based on the autoregressive integrated moving average and generalized autoregressive conditional heteroscedasticity (ARIMA-GARCH) model, ref. [15] created a methodology to forecast the similar and volatile portions. Also, the study compared the strengths and weaknesses of linear and non-linear hybrid methods.

Ref. [16] harnessed cutting-edge object detection and tracking technologies to pinpoint, categorize, monitor, and gauge vehicles on the road via video analysis. This comprehensive approach offers sturdy backing for urban traffic management strategies and future planning. Leveraging digital twin technology, they created a virtual replica of traffic patterns using camera data, forming the foundation for training various algorithmic models.

2.2. Traffic Models in the City of Barcelona

There has been several recent works on traffic models analyzing the specific case of Barcelona. For instance, ref. [17] compared the performance of different traffic assignment models through simulation. Such methodologies evaluate traffic levels within a track or road network based on its physical attributes, functional dynamics, and anticipated traffic volume. The case study focused on the province of Barcelona, and experiments compared five static traffic assignment approaches (all or nothing, stochastic assignment with a simulation-based method, the method of successive averages, incremental assignment, and user equilibrium using the Frank and Wolfe algorithm). Regression analysis was used to study the differences between estimated and observed flow. The best model was user equilibrium with an R^2 index of 0.93 for light vehicles and 0.89 for heavy vehicles.

Moreover, ref. [18] introduced two models, one for taxi stand services and the other for one-way carsharing, in the context of Barcelona's taxi demand. Each model determines optimal factors such as the number of cars and depots, depot capacity, system unitary costs, and level of service. Results indicate comparable operation between the two systems, although taxi services prove up to three times costlier due to driver hiring expenses. Ref. [19] developed a coupled macroscopic traffic and emission modeling system tailored to the Barcelona metropolitan area to estimate hourly road transport emissions at the road link level. Their analysis included an emission sensitivity assessment and investigation into typically high-uncertainty emission factors. The study found significant sensitivity to inputs such as vehicle fleet composition and meteorological impacts on diesel engines, with non-exhaust sources contributing substantially to total PM emissions. Discrepancies between the macroscopic and microscopic systems increased with congestion levels, particularly in NO_x emissions, reaching up to 65%. Ref. [20] examined methods for evaluating the impacts of pluvial flooding events on traffic flows under current and future climate change scenarios in Barcelona and Bristol. Using both meso-scale and micro-scale traffic models, they found that increased flood intensity led to greater disruptions in traffic flows, with climate change exacerbating these effects. Ref. [21] conducted a quantitative assessment of shared mobility service usage among residents in the Barcelona metropolitan region. Through 600 questionnaires responses from commuting travelers, they identified preferences based on factors such as age, regular commutes, and personal incomes. Findings revealed varied trends for intra-city and inter-city commutes, with younger demographics showing higher predicted use of ridesharing, carsharing, and ride-hailing services. Additionally, passengers tended to opt for the services that best suited their needs on each occasion rather than relying solely on one mode of transportation.

3. Methodology

To examine traffic levels in Barcelona utilizing data from the Ajuntament de Barcelona's open data service, this paper employs visualization techniques, time series analysis, and the XGBoost algorithm.

3.1. Visualization Techniques

Visualizing traffic data is crucial for effective city traffic management. It aids in comprehending the dynamics of moving entities and uncovering patterns related to traffic, social dynamics, spatial geography, and economic trends [22]. Some aspects of the potential application field of traffic visualization are identifying real-time traffic jams by monitoring traffic situations, discovering mobility patterns of vehicles and pedestrians, and improving route planning according to the traffic density. Commonly used visualization techniques include line charts, bar charts, heatmaps, histograms, and geospatial maps. Geospatial maps represent the most commonly employed method for visualizing traffic data, and 2D visualization is more commonly used in the literature than its 3D counterpart [23].

While 2D geospatial maps provide a good representation of the spatial dimensions of data, they fail to present the temporal dimension of data, which is especially important in traffic data visualizations. There are different approaches for representing spatial data with

temporal components. The most widely studied method in the literature is the space–time cube (STC) model [24]. In an STC model, the spatial components of the data are depicted along the x and y axes, while the temporal dimension is represented along the z axis. This model works well with small-to-medium-sized data; however, when the data are too large, the visualization becomes difficult.

Another common approach to represent space and time is to perform animation of a map to display geospatial changes in time. The visualization technique used in this article uses the HeatMapWithTime plugin of Folium library [25]. A heatmap is a powerful visualization method for storytelling, especially for geospatial data. This plugin allows creation of an interactive animation of heatmaps that allows end users to manipulate the visual representation according to their needs. Some of the potential interaction options are overview, zoom, pause, play, loop, and also play at different speeds for a period of time. The heatmap allows us to visualize the traffic density of different sections of the city throughout the day. Every dot in the heatmap represents the location of a vehicle; by analyzing the location of vehicles at different time intervals, potential mobility patterns of vehicles and the traffic density of the streets can be identified. The dataset employed in this article (described in Section 4.1) does not contain vehicle information but only street section data; however, with the provided information, we can manually assign a certain number of vehicles for every street section according to the traffic density of the sections, i.e., higher-traffic-density sections have a higher number of vehicles, and lower-density sections have a lower number of vehicles.

3.2. Time Series Analysis

The predictive models used are autoregressive integrated moving average (ARIMA) models [26]. The general case of ARIMA (p,d,q) can be written as follows:

$$Y_t = \alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \quad (1)$$

where the parameters α are the autoregresive part (ar), and θ are the moving average part (ma). The models are built by section because the behavior is more stable. They are implemented using free software R (version 4.03) [27] and the forecast package for R [28]. As these models have been widely utilized across various applications for an extended period, we direct interested readers to [29] for further elaboration.

3.3. eXtreme Gradient Boosting

eXtreme Gradient Boosting [30] (XGBoost) is a scalable machine learning technique for tree gradient boosting. XGBoost employs a methodology similar to other gradient boosting techniques, constructing a mathematical framework to predict y_i from input $\boldsymbol{x}_i$

Specifically, XGBoost forms an ensemble $\mathcal{F} = \{f(\boldsymbol{x}) = w_{q(\boldsymbol{x})}\}(q : \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T)$ comprising K Classification and Regression Trees (CART). Each f_k corresponds to a distinct tree q with its own structure characterized by the number of leaves T and the set of leaf weights w.

Given this established framework, the predicted output is determined by a K additive function as follows:

$$\hat{y}_i = \phi(\boldsymbol{x}) = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i),\ f_k \in \mathcal{F} \quad (2)$$

In this specific instance, the data values $\boldsymbol{x}_i$ include factors such as the geographical location of the evaluated section, the time of data collection, and other data engineering variables discussed later. Rather than purely categorizing like conventional decision trees, this model incorporates continuous scores obtained from the weights w_i associated with each leaf across all defined trees q.

To enable the learning process of the K specified trees, a convex loss function acts as a regularized objective:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

Here, l represents a differentiable loss function measuring the disparity between the target $\hat{y}_i$ and the predicted value y_i, while Ω is defined as a regularization term penalizing potential overfitting of the model expressed as follows:

$$\Omega(f) = \lambda T + \frac{1}{2}\lambda ||w||^2 \tag{4}$$

Instead of adhering to optimization methods based in the Euclidean space, XGBoost undergoes training via an additive approach. In this method, for a given i-th instance at the t-th iteration, a distinct f_t is introduced to minimize the objective function, represented as follows:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\boldsymbol{x}_i)) + \Omega(f_t) \tag{5}$$

Here, $\hat{y}^{(t)}$ denotes the prediction at that specific instance and iteration. This process means that f_t is incrementally included based on its performance across optimization instances and iterations. The objective function, derived by expanding the loss function's second-order Taylor series with respect to $\hat{y}_i$ for optimization purposes, is expressed using its first- and second-order gradient statistics g_i and h_i, following

$$\begin{aligned} &\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} \left[l(y_i, \hat{y}^{(t-1)} + g_i f_t(\boldsymbol{x}_i) + \frac{1}{2} h_i f_t^2(\boldsymbol{x}_i) \right] + \Omega(f_t) \\ &\text{where } g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) \text{ and} \\ &h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}). \end{aligned} \tag{6}$$

$$\tilde{\mathcal{L}}^{(t)} \simeq \sum_{i=1}^{n} \left[g_i f_t(\boldsymbol{x}_i) + \frac{1}{2} h_i f_t^2(\boldsymbol{x}_i) \right] + \Omega(f_t) \tag{7}$$

Furthermore, expanding the objective function through its regularization term Ω for a leaf set I_j within a given structure $q(x_i)$ can be achieved as follows:

$$\tilde{\mathcal{L}}^{(t)} \simeq \sum_{i=1}^{n} \left[g_i f_t(\boldsymbol{x}_i) + \frac{1}{2} h_i f_t^2(\boldsymbol{x}_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{8}$$

This formulation enables the determination of the optimal leaf weight w_j^* for leaf j and assesses the fit of the overall structure $q(x)$ using the following:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{9}$$

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{10}$$

XGBoost employs a greedy algorithm, initially starting from a single leaf and subsequently adding branches at each iteration using the scoring function in Equation (10) to evaluate improvements in the general tree structure q. The potential gain is assessed within an instance $I = I_L \cup I_R$, where I_L and I_R denote the sets of nodes on each branch after splitting.

$$Gain = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \quad (11)$$

3.3.1. Evaluation Metrics

The classification system's evaluation is determined using the area under the ROC curve (AUC) with an Over to Rest (OrV) strategy, a widely recognized measure of performance for machine learning models. In the OrV strategy, any misclassification is considered erroneous regardless of which two classes are mistaken. The AUC method is defined as follows:

$$\text{AUC} = \sum_i \{(1 - \beta_i \times \Delta\alpha) + \frac{1}{2}[1 - \beta\Delta\alpha]\} \quad (12)$$

Here, α represents the probability of a false positive, and $1 - \beta$ denotes the probability of a true positive.

3.3.2. Shapley Additive Explanations

Shapley Additive Explanations (SHAP) are a tool rooted in game theory used to interpret machine learning models [31]. It assesses the contribution of each variable to the final output by iteratively introducing one variable into the model at a time and evaluating the expected value of the model's output function. This method allows SHAP to calculate the average contribution while considering the effects of all potential variable orderings.

The average contribution of variable x in model f can be calculated as follows:

$$g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \quad (13)$$

where $x' \in \{0, x_i\}^M$ represents the count of input variables, and $\phi_i \in \mathbb{R}$.

SHAP, and, in particular, Tree SHAP [32], provides insights into how variables impact tree-based machine learning models. It examines three fundamental properties: local accuracy ensures that the function linking x to x', denoted as $h_x(x')$, matches the set of variables x, ensuring the approximation of f corresponds to the output of f; missingness stipulates that absent values have no bearing on the model's output ($x'_i = 0 \rightarrow \phi_i = 0$); and consistency guarantees that if an input's contribution to the model increases or stays constant, the Shapley value follows suit, regardless of other inputs. Tree SHAP analyzes the model using an input dataset X of dimensions $N \times M$, generating a matrix of SHAP values for each variable in every tuple in X, ensuring consistent explanations for individual predictions while highlighting contributions with sign indicators.

3.4. Variable Analysis

Once the model is computed, the importance scores for each variable can be extracted. This presents a score for how valuable each variable selected is in absolute terms and averaged across all the trees that form the XGBoost model. We study the importance of each parameter following three main scoring systems:

- Gain: It denotes the relative contribution of each variable to the model, calculated by summing up the contribution of each variable for every tree generated. It shows the importance of a variable when generating a prediction;
- Cover: It indicates the relative importance of each variable based on the number of observations associated with it. This is determined by summing the second derivative of the loss function over all training data points falling into each node defined by the variable;
- Frequency: It represents the percentage indicating how often a specific variable appears in the trees of the model relative to the total number of trees.

The model is implemented using standard Python 3.10.8 distribution [33] with its correspondent libraries for XGBoost implementation.

4. Application

This section describes the dataset explored and presents the results obtained from applying the visualization techniques and models introduced in the preceding section.

4.1. Description of the Dataset

The dataset is called "Traffic state information by sections of the city of Barcelona" and is freely accessible from the dataset catalogue of the Open Data BCN service (opendata-ajuntament.barcelona.cat/en, accessed on 2 April 2024). According to its website, Open Data BCN constitutes "a movement driven by public administrations with the main objective of maximize available public resources, exposing the information generated or guarded by public bodies, allowing its access and use for the common good and for the benefit of anyone and any entity interested".

The dataset contains historical data collected since December 2017 and is updated monthly. It describes the traffic state for a set of 527 sections and has an update frequency of 5 min. The traffic states are 0 = no data, 1 = very fluid, 2 = fluid, 3 = dense, 4 = very dense, 5 = congestion, and 6 = cut off (closed). The traffic state is assessed using various sensors embedded beneath the asphalt, including those detecting magnetic field changes caused by passing metal masses (vehicles), infrared sensors, and cameras equipped with image processing capabilities. Data from each detection station are qualitatively interpreted, typically using a scale ranging from 1 to 5, based on predefined thresholds specific to each station.

To explore the city of Barcelona's traffic situation information in March 2022, the software used was R version 4.0.3 (10 December 2020) [27], employing RStudio as integrated development environment [34]. The dataset contains five variables: ID section (where section refers to road segment in Barcelona), data (year, month, day, hour, minute, and second), current state, and expected state. There are 4,599,656 rows since the information of the 527 sections is updated approximately every 5 min. First, we transform the dataset to consider the current state every 5 min exactly. This process requires introducing new rows with a 0 (no data) both in current and expected states.

The distribution of the seven current traffic states is shown in Figure 2. Overall, congestion (0.7%) and 'no data' (42.9%) have the minimum and maximum proportions, respectively. The rest of the states, sorted by proportion, are fluid (29.1%), very fluid (20.1%), dense (4.7%), very dense (1.7%), and cut off (0.8%).

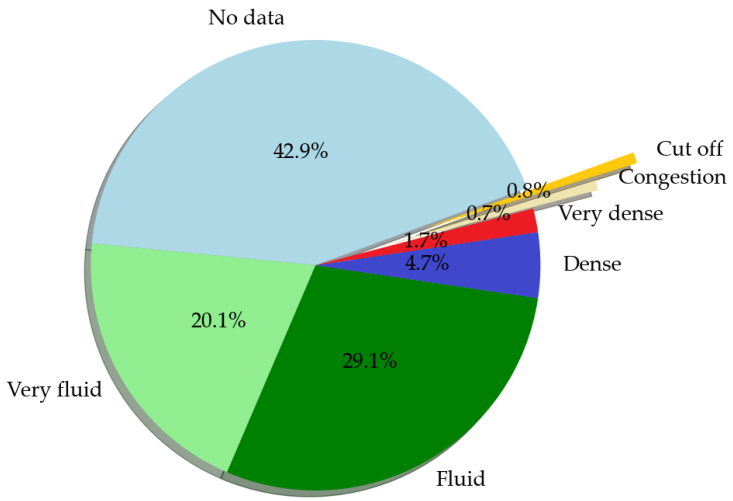


Figure 2. Distribution of traffic states. Data source: "Traffic state information by sections of the city of Barcelona" (Open Data BCN)—March 2022.

4.2. Visualization Techniques

Figure 3 represents the traffic state of the city of Barcelona for Friday 25 March 2022 at different time intervals. The lines represent the sections of the streets, with each color indicating the traffic condition of those sections: light blue for very fluid, dark blue for fluid, yellow for dense, orange for very dense, and red for congestion. As we can observe, most sections of the city have a traffic state of 1 to 2 (very fluid to fluid traffic), and rush hours (9:00 a.m.) have more sections with dense or higher traffic (3–5) than non-rush hours (3:00 p.m.).

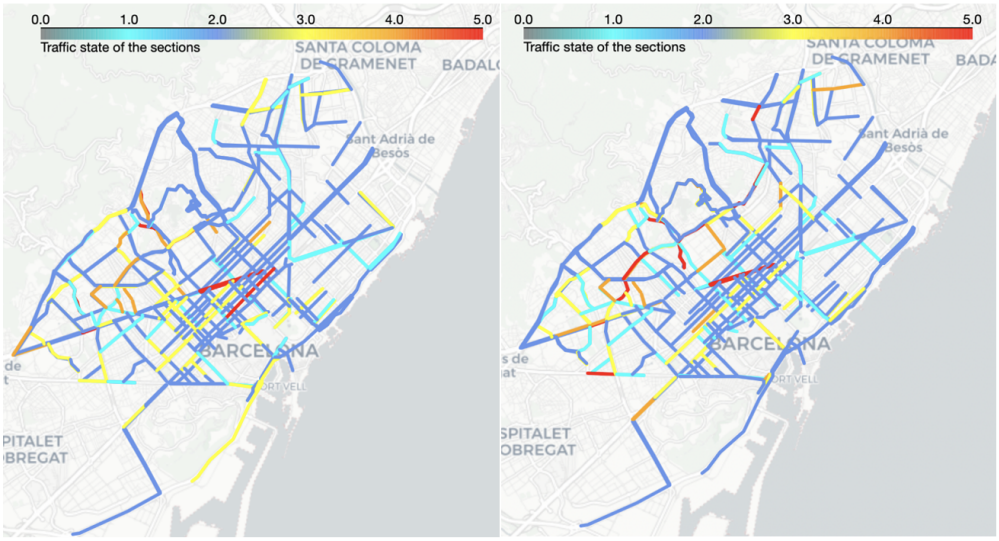


Figure 3. Geospatial traffic state map. Traffic state at 9:00 (**left**) and at 15:00 (**right**).

Figure 4 shows the traffic state for the same time intervals as Figure 3 but in heatmap form. The top two subfigures depict a broader area, whereas the bottom two subfigures concentrate on the central area. As we can observe, the dense area of the heatmap (that is, areas marked with an orange color) is located in sections with higher traffic or areas with a higher number of sections.

4.3. Time Series Analysis

The predictive models are applied to data from the city of Barcelona's traffic situation information in March 2022. Due to the large number of monitored sections, it is important to know whether their behavior is similar. For this purpose, 10 sections belonging to the same street are chosen. We would like to highlight that, even though the sections belong to the same street, there are times when the traffic conditions are not similar, even with changes depending on whether it is a weekend or a working day.

Table 1 shows the output of the time series models. The first two columns identify the section and the fitted model. The subsequent five columns display the coefficients, while the last two columns present the performance measures AIC and BIC. It is not necessary to differentiate the series in any case because the models are stable. Table 2 shows the Mean Error (ME) and the Root-Mean-Square Error (RMSE) of each model. The tests of independence (Box–Ljung test for residuals), homoscedasticity (Box–Ljung test for squared residuals), and normality (Shapiro–Wilk normality test) are applied to each of the models, and their p-values are shown. None of the models complies with the initial assumptions, which leads us to confirm that these models do not explain the behavior of the data very well. In conclusion, due to the high non-linearity and complexity of the traffic flow, classical methods cannot make good predictions; moreover, they do not fit well with the space–time structure of the traffic data. For this reason, in the next subsection, a machine learning method is applied.

Figure 4. Heatmap traffic state map. Traffic state at 9:00 (**left**) and at 15:00 (**right**).

Table 1. Output of time series by section.

Section	Model	ar1	ar2	ar3	ma1	ma2	AIC	BIC
1	ARIMA(2,0,2)	1.568	−0.655		−0.839	−0.315	934.270	961.940
2	ARIMA(3,0,1)	0.067	0.824	−0.211	0.965		105.650	133.320
3	ARIMA(2,0,1)	1.799	−0.872		−0.910		−39.040	−15.980
4	ARIMA(2,0,2)	1.668	−0.742		−0.639	−0.193	1126.190	1153.860
5	ARIMA(2,0,2)	1.506	−0.583		−0.622	−0.249	1469.640	1497.310
6	ARIMA(1,0,1)	0.638			0.259		1172.080	1190.520
7	ARIMA(2,0,1)	1.785	−0.850		−0.900		69.040	92.100
8	ARIMA(1,0,0)	0.589					1138.790	1152.620
9	ARIMA(2,0,0)	0.788	−0.227				1541.360	1559.800
10	ARIMA(2,0,1)	1.778	−0.838		−0.867		−33.030	−9.970

Table 2. Output of time series errors by section.

Section	Model	ME	RMSE	Independence Test (*p*-Value)	Homoscedasticity Test (*p*-Value)	Normality Test (*p*-Value)
1	ARIMA(2,0,2)	5.27×10^{-5}	0.449	**0.7556**	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
2	ARIMA(3,0,1)	0.0008	0.257	3.5×10^{-11}	**0.078**	$<2.2 \times 10^{-16}$
3	ARIMA(2,0,1)	−0.0005	0.234	1.594×10^{-11}	4.331×10^{-6}	$<2.2 \times 10^{-16}$
4	ARIMA(2,0,2)	−0.0001	0.511	0.00015	5.707×10^{-6}	$<2.2 \times 10^{-16}$
5	ARIMA(2,0,2)	0.0007	0.644	**0.2001**	3.201×10^{-10}	$<2.2 \times 10^{-16}$
6	ARIMA(1,0,1)	0.0009	0.529	5.668×10^{-9}	2.065×10^{-14}	$<2.2 \times 10^{-16}$
7	ARIMA(2,0,1)	−0.0001	0.251	5.951×10^{-8}	**0.093**	$<2.2 \times 10^{-16}$
8	ARIMA(1,0,0)	−0.0002	0.518	**0.1294**	1.208×10^{-8}	$<2.2 \times 10^{-16}$
9	ARIMA(2,0,0)	0.0003	0.678	1.056×10^{-5}	1.762×10^{-8}	$<2.2 \times 10^{-16}$
10	ARIMA(2,0,1)	3.99×10^{-5}	0.235	4.441×10^{-16}	1.478×10^{-7}	$<2.2 \times 10^{-16}$

4.4. eXtreme Gradient Boosting

For streamlined parameter importance analysis, we classify the diverse densities derived from the original dataset into separate classification categories: 0 = very fluid, 1 = fluid, 2 = dense, 3 = very dense, 4 = congestion. The dataset, organized according to Table 3, contains information spanning the entirety of 2019. We partition the dataset into distinct sets for both training and testing stages during the evaluation process. Specifically, 70% of the dataset is earmarked for training purposes, while the remaining 30% is preserved for parameter evaluation. Data spanning January to March 2022 are processed for final analysis and predictions as presented in the model performance section.

Table 3. Description of the dataset for the XGBoost model.

Variable	Description	Type	Range
Status	Current traffic status of the section.	number	0 to 4
FromNorth	Starting Latitude (North).	number	2099 to 2222
FromWest	Starting Longitude (West).	number	41,338 to 41,450
ToNorth	Ending Latitude (North).	number	2100 to 2223
ToWest	Ending Longitude (West).	number	41,338 to 41,449
DailyHour	Measurement hour.	number	0 to 23
DailyMinute	Measurement minute.	number	0 to 60
Weekday	Weekday of the measurement.	number	1 to 7
DayMonth	Measurement day of the month.	number	1 to 31
Holiday	Boolean value representing the existence of a holiday on that day.	Boolean	False or True
Number of records: 24,533,298			

To optimize XGBoost's performance, parameter tuning is conducted using a cross-validation technique known as grid search with five folds. This involves exploring various parameter configurations within sensible ranges, as detailed in Table 4.

Table 4. Parameter matrix for XGBoost.

Parameter	Selected value	Options
max_depth	10	5, 7, 9, 10, 11
eta	0.3	0.1, 0.2, 0.3, 0.4
gamma	1	0.5, 1, 1.5, 2, 5
subsample	1	0.6, 0.8, 1
objective	multi:softmax	-
num_class	5	-

The Area Under the Curve (AUC) depicted in Figure 5 illustrates the consistently strong performance of the model, with AUC values exceeding 80% across all ROC curves. Notably, the model exhibits particularly strong performance in extreme cases, with a true positive rate exceeding 85%. However, there is a slight decline in performance observed in intermediate classes.

On average, this model achieves an accuracy rate of 74.48% for the 2019 dataset; therefore, we can affirm that this model represents with accuracy the importance of the parameters with a relevant correlation between the true labels and the predicted ones.

Computing time can also be counted in model performance once it is trained, like in Table 5. For reference, average times for predictive values for a particular day, week, or month's worth of data are presented considering that this model's training and execution was carried out on an 11th Gen Intel(R) i5-1135G7 that runs at 2.40 GHz with 16 GB of RAM and Windows 10 Pro 22H2—64 bits as the main operative system. It is concluded that this technology is perfectly adaptable for a dynamic interpretation of the data if this is to be implemented as a prediction tool by Barcelona council.

For this model, Figure 6a shows the parameters ranked by frequency of appearance, where FromNorth, DailyHour, and DayMonth represent the majority of the relevance in order to classify the data rows. FromWest also shows a relatively superior relevance. These results are consistent with our initial hypotheses since the geographical location, the time, and the day of the month influence, and, therefore, help to explain, the behavior of current traffic status. Variables like DailyMinute and Weekday appear to be labeled as less determinant in the model. Moreover, checking not only the frequency of appearance but its cover (the relative number of observations linked in prediction with the variable) in Figure 6b and its gain (the relative contribution of each variable on each tree for every prediction) in Figure 6c, it can be seen how variables such as ToWest and ToNorth, even though their percentage of appearance is not outstanding compared to others in terms of variable importance, have gain and cover which show how this is also defining the behavior of the model and therefore have a relevant impact on the outcome.

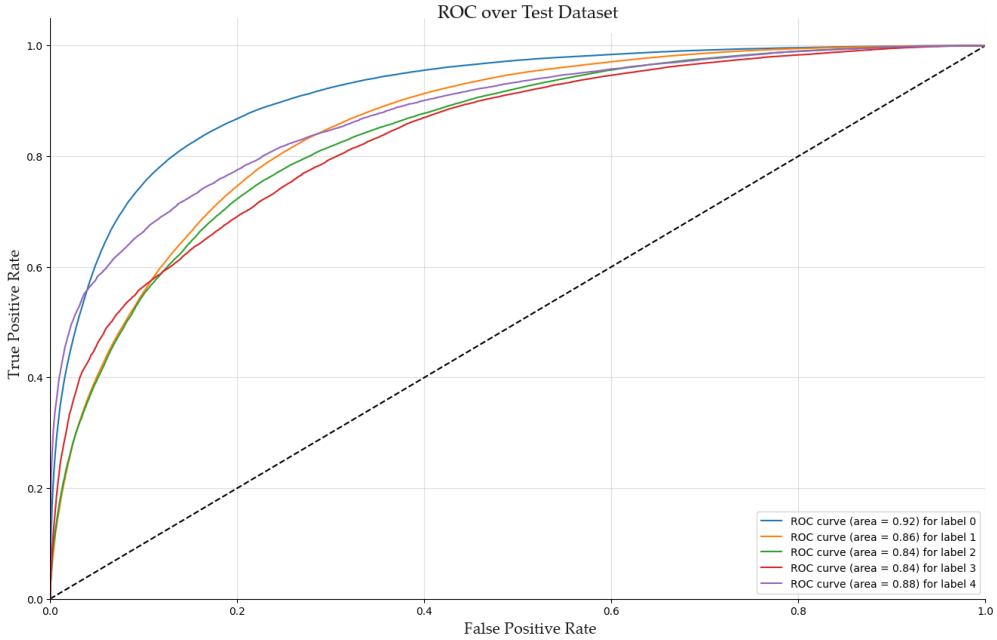
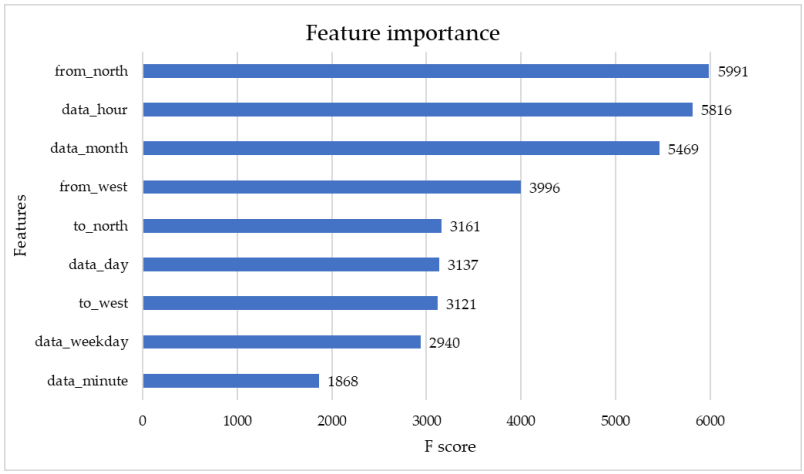


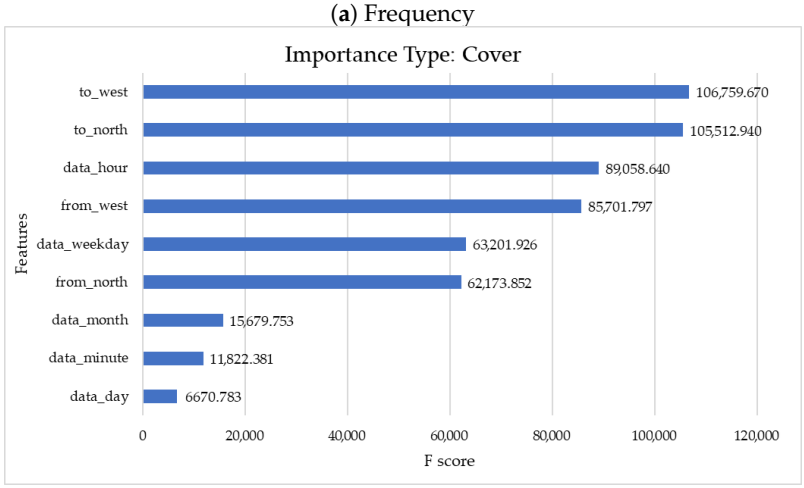
Figure 5. Area Under the Curve for the XGBoost model.

Table 5. Computing times.

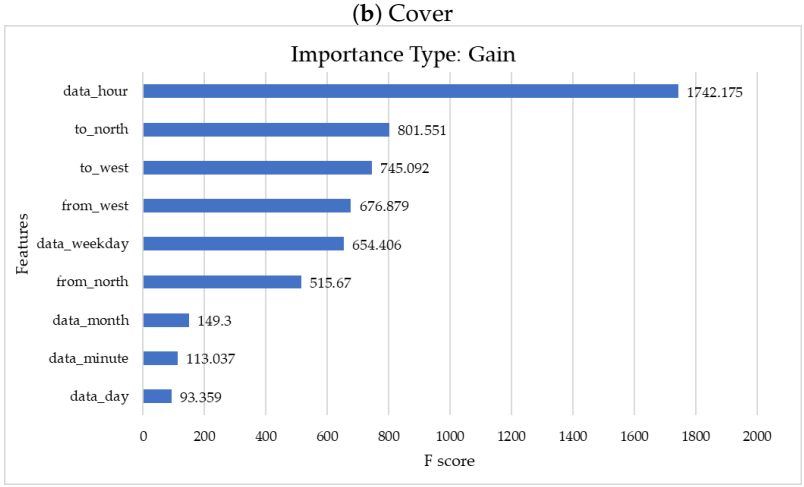
Time Lapse	Average Times after 50 Runs (s)
One day	0.037577
One week	0.229319
One month	1.363848



(**a**) Frequency



(**b**) Cover



(**c**) Gain

Figure 6. Variable analysis of the XGBoost model.

As Figure 7 implies, classes 0, 1, and 2 obtain their explainability from DayMonth in the majority of cases. The rest of the classes find their explanation impact from coordinates like ToWest and FromWest.

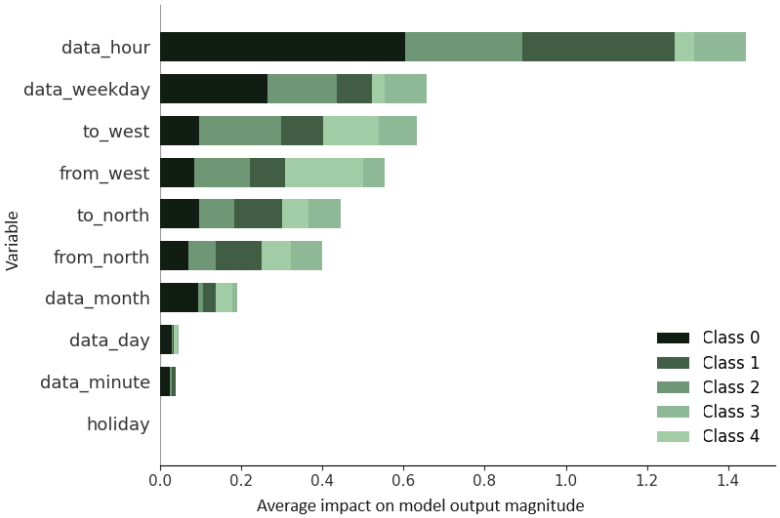
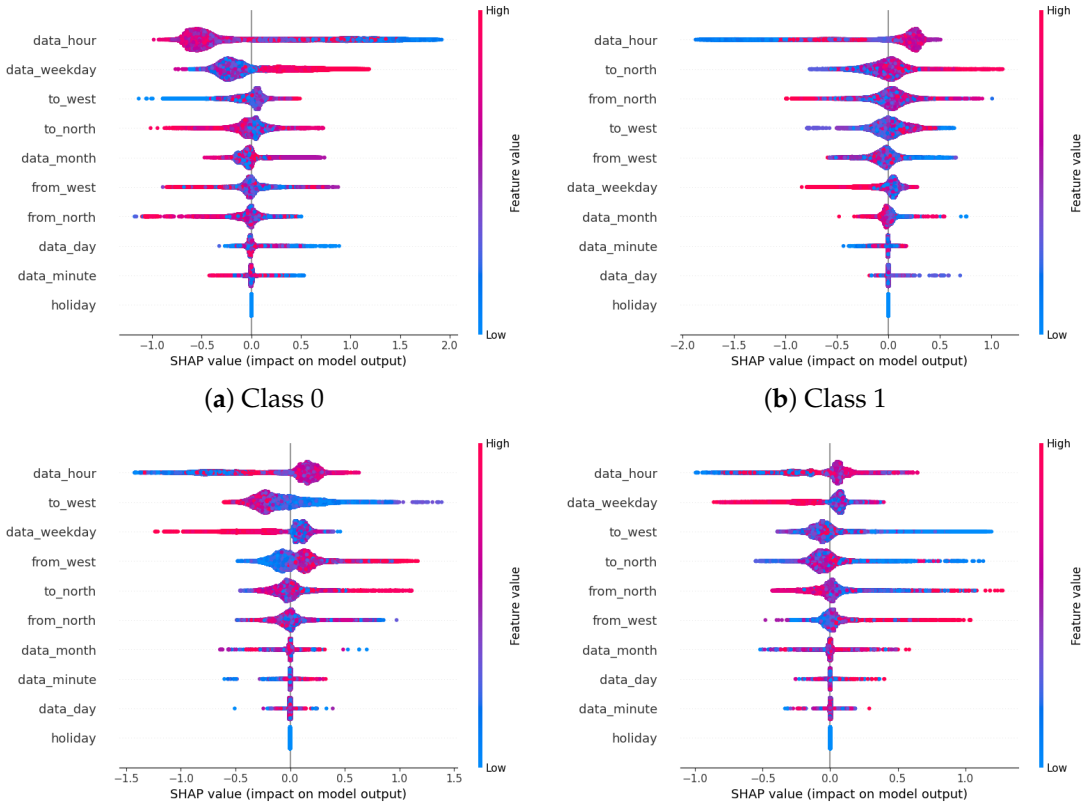


Figure 7. SHAP variable importance plot by classes.

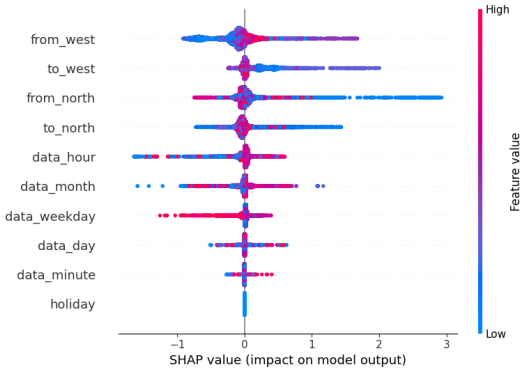
Lastly, Figure 8 illustrates the model's effect on various combinations of variables for each class. For class 0, the hour and day of the week have the greatest impact on the model. For class 1, the hour and North coordinates are the most important. For class 2, the hour and West coordinates have the greatest impact. For class 3, the hour and day of the week coordinates are the most important. In class 4, all four coordinates have the most impact on the model.



(**a**) Class 0 (**b**) Class 1

(**c**) Class 2 (**d**) Class 3

Figure 8. *Cont.*



(**e**) Class 4

Figure 8. Impact on model output by predicted class.

5. Discussion

Due to the increasing importance of traffic prediction for urban transportation management, numerous traffic flow prediction models have been studied over recent decades. This paper utilizes visualization techniques, such as heatmaps, to illustrate the traffic density across various city sections throughout the day. Analyzing vehicle locations at different time intervals reveals mobility patterns and street traffic density. An interactive animation plugin is identified, enabling end users to customize the visual representation with options including overview, zoom, pause, play, loop, and variable playback speeds.

Subsequently, traffic prediction is approached as a time series analysis problem, with ARIMA models being employed to forecast traffic flow across ten sections of the same street. Time series analysis, particularly ARIMA models, has historically been popular for traffic prediction due to ease of implementation and relatively high accuracy [35]. Nevertheless, the intricate non-linear spatio-temporal dependencies, coupled with external factors such as weekends, holidays, and weather, present challenges that surpass the capabilities of ARIMA. Our results indicate that none of the ARIMA models built satisfies the assumptions of independence, homoscedasticity, and normality. Thus, classical methods struggle to make accurate predictions and are not well suited to the spatio-temporal structure of traffic data.

To tackle this challenge, an XGBoost model is employed for traffic prediction. XGBoost requires less prior knowledge of traffic patterns, handles non-linear variables effectively, and consistently achieves robust performance with the area under ROC curves exceeding 80%. Its use facilitates analysis of variable importance and enhances interpretability through Shapley Additive Explanations.

In summary, while visualization techniques aid real-time analysis of traffic density and interactive animations for clearer pattern recognition, the XGBoost model provides insights into variable importance, interpretable outputs, and accurate predictions.

By accurately predicting traffic patterns, decision-makers can make more informed about transportation infrastructure, such as the location and number of roads, highways, and public transportation systems. This can help to optimize the use of limited resources and improve the efficiency of the transportation network [36]. In terms of reducing congestion and emissions, accurate traffic prediction can help decision-makers to implement strategies to reduce congestion and improve air quality. For instance, they could implement demand-management strategies like variable tolls or congestion pricing. These measures aim to incentivize drivers to opt for alternative modes of transportation or travel during off-peak hours. Also, traffic prediction can help to identify areas where accidents are more likely to occur, allowing them to implement strategies to reduce the likelihood of accidents and improve road/street safety. Hence, by foreseeing and outlining the variables that contribute to the forecast, this paper's contribution can assist a decision-maker in making quick and effective decisions in order to enhance traffic management and mitigate traffic congestion [37]. In this study, the significant variables that contribute to the forecast are the

geographical location, especially when the section starts in the North direction, the time, and the day of the month. However, some limitations need to be highlighted. In order to increase the usefulness of the data and analysis, data must be accessible, interconnected, and quantifiable to facilitate decision-making and drive innovation. In the case of the city council, this is not fully achieved, as the availability of information and the interconnection between datasets are severely restricted by the format and structure in which they are delivered. This hinders a broader perspective on some possible research questions and impedes the ability to unlock the full potential of the data. No clear descriptions were given for some of the data sources, a large number of missing data were detected, and the quantification of terms like traffic density was not available. Possible data connections with other datasets on the same portal were also not available.

6. Conclusions and Further Research

Studying the mobility patterns in smart cities may provide useful insights to support decision-making related to urban mobility. Optimizing processes such as the placement of electric vehicle charging stations or designing efficient routes may minimize economic costs while contributing to reducing the environmental impacts and increasing social welfare. In this context, this paper explored mobility patterns in the city of Barcelona (Spain), relying on the open data service of the Ajuntament de Barcelona. An exploratory analysis, visualization techniques, and predictive models have been presented. In addition, a discussion regarding the potential of these tools for policymakers has been provided.

The primary limitation of this research lies in its reliance on data spanning only one month. While the selection of this period facilitated building models and yielded intriguing insights, examining a more extended period (spanning years) would enable exploration of trends, seasonal fluctuations, variability, and the influence of COVID-19 and lockdown policies on traffic patterns, among other factors. This work opens up several lines of future research. First, the use of more variables could lead to more robust, powerful, and interpretable models. Interesting variables to consider are related to weather (rainwater level, temperature, etc.), traffic accidents, and events that attract a lot of people such as football games and international conferences. In addition, comparing open data portals of smart cities across the world (regarding datasets related to urban mobility and characteristics such as update frequency, granularity, and documentation) and studying the use that different agents have for them is another promising research line. Further lines in this topic could involve comparing the results of XGBoost to those of other machine learning algorithms and examining ways to improve its interpretability. Research on the potential benefits of combining XGBoost with the use of more advanced spatio-temporal models, in order to capture the spatio-temporal dependence, is another promising option for implementing better predictive models.

Author Contributions: Conceptualization, L.C., P.C., C.S. and P.M.; methodology, E.G., L.C., P.C., C.S. and P.M.; software, E.G. and M.P.; validation, E.G. and M.P.; investigation, L.C., P.C., C.S. and P.M.; data curation, E.G. and M.P.; writing—original draft preparation, E.G., L.C., P.C. and C.S.; writing—review and editing, L.C. and P.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially funded by the Spanish Ministry of Science (PID2019-111100RB-C21-C22/AEI/10.13039/501100011033), as well as by the Barcelona City Council and Fundació "la Caixa" under the framework of the Barcelona Science Plan 2020–2023 (grant 21S09355-001). The authors appreciate the support received from the research group GRBIO under the grant 2021 SGR 01421 from the Departament de Recerca i Universitats de la Generalitat de Catalunya (Spain).

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in Open Data BCN service at opendata-ajuntament.barcelona.cat/en (accessed on 2 April 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pasaoglu, G.; Fiorello, D.; Martino, A.; Scarcella, G.; Alemanno, A.; Zubaryeva, A.; Thiel, C. *Driving and Parking Patterns of European Car Drivers—A Mobility Survey*; European Commission Joint Research Centre: Luxembourg, 2012.
2. Eurostat. Statistics on European Cities. 2017. Available online: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Statistics_on_European_cities (accessed on 2 April 2024).
3. Ajuntament de Barcelona. Dades Bàsiques de Mobilitat Barcelona. 2017, 2018. Available online: http://hdl.handle.net/11703/111727 (accessed on 2 April 2024).
4. European Environment Agency. Transport. 2019. Available online: https://www.eea.europa.eu/themes/transport/intro (accessed on 9 December 2022).
5. European Commission. New EU Urban Mobility Framework. Roadmap. 2021. Available online: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12916-Sustainable-transport-new-urban-mobility-framework_en (accessed on 2 April 2024).
6. European Council. Conclusions on 2030 Climate and Energy Policy Framework, SN 79/14. 2014. Available online: https://www.buildup.eu/en/practices/publications/european-council-23-and-24-october-2014-conclusions-2030-climate-and-energy (accessed on 2 April 2024).
7. Reyes-Rubiano, L.; Calvet, L.; Juan, A.A.; Faulin, J.; Bové, L. A biased-randomized variable neighborhood search for sustainable multi-depot vehicle routing problems. *J. Heuristics* **2020**, *26*, 401–422. [CrossRef]
8. Faulin, J.; Grasman, S.; Juan, A.; Hirsch, P. *Sustainable Transportation and Smart Logistics: Decision-Making Models and Solutions*; Elsevier: Amsterdam, The Netherlands, 2018.
9. Khan, A.B.F.; Ivan, P. Integrating Machine Learning and Deep Learning in Smart Cities for Enhanced Traffic Congestion Management: An Empirical Review. *J. Urban Dev. Manag.* **2023**, *2*, 211–221. [CrossRef]
10. Yu, B.; Yin, H.; Zhu, Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, Stockholm, Sweden, 13–19 July 2018; pp. 3634–3640.
11. Kashyap, A.A.; Raviraj, S.; Devarakonda, A.; Nayak K, S.R.; KV, S.; Bhat, S.J. Traffic flow prediction models—A review of deep learning techniques. *Cogent Eng.* **2022**, *9*, 2010510. [CrossRef]
12. Shi, Y.; Feng, H.; Geng, X.; Tang, X.; Wang, Y. A Survey of Hybrid Deep Learning Methods for Traffic Flow Prediction. In Proceedings of the 3rd International Conference on Advances in Image Processing, Chengdu, China, 8–10 November 2019; pp. 133–138.
13. Li, Y.; Shahabi, C. A brief overview of machine learning methods for short-term traffic forecasting and future directions. *Sigspatial Spec.* **2018**, *10*, 3–9. [CrossRef]
14. Shahriari, S.; Ghasri, M.; Sisson, S.; Rashidi, T. Ensemble of ARIMA: Combining parametric and bootstrapping technique for traffic flow prediction. *Transp. Transp. Sci.* **2020**, *16*, 1552–1573. [CrossRef]
15. Yao, R.; Zhang, W.; Zhang, L. Hybrid methods for short-term traffic flow prediction based on ARIMA-GARCH model and wavelet neural network. *J. Transp. Eng. Part Syst.* **2020**, *146*, 04020086. [CrossRef]
16. Kan, H.; Li, C.; Wang, Z. Enhancing Urban Traffic Management through YOLOv5 and DeepSORT Algorithms within Digital Twin Frameworks. *Mechatronics Intell. Transp. Syst.* **2024**, *3*, 39–54. [CrossRef]
17. Rojo, M. Evaluation of traffic assignment models through simulation. *Sustainability* **2020**, *12*, 5536. [CrossRef]
18. Cuevas, V.; Estrada, M.; Salanova, J.M. Management of on-demand transport services in urban contexts. Barcelona case study. *Transp. Res. Procedia* **2016**, *13*, 155–165. [CrossRef]
19. Rodriguez-Rey, D.; Guevara, M.; Linares, M.P.; Casanovas, J.; Salmerón, J.; Soret, A.; Jorba, O.; Tena, C.; García-Pando, C.P. A coupled macroscopic traffic and pollutant emission modelling system for Barcelona. *Transp. Res. Part Transp. Environ.* **2021**, *92*, 102725. [CrossRef]
20. Evans, B.; Chen, A.S.; Djordjević, S.; Webber, J.; Gómez, A.G.; Stevens, J. Investigating the effects of pluvial flooding and climate change on traffic flows in Barcelona and Bristol. *Sustainability* **2020**, *12*, 2330. [CrossRef]
21. Gilibert Junyent, M.; Ribas Vila, I.; Rodríguez Donaire, S. Analysis of mobility patterns and intended use of shared mobility services in the Barcelona region. In Proceedings of the ETC Conference Papers 2017; 2019; pp. 1–20.
22. Chen, W.; Guo, F.; Wang, F.Y. A survey of traffic data visualization. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 2970–2984. [CrossRef]
23. Clarinval, A.; Dumas, B. Intra-city traffic data visualization: A systematic literature review. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 6298–6315. [CrossRef]
24. Kraak, M.J. The space-time cube revisited from a geovisualization perspective. In Proceedings of the 21st International Cartographic Conference, Durban, South Africa, 10–16 August 2003; pp. 1988–1996.
25. Story, R. Folium. 2020. Version: 0.11.0. Available online: https://python-visualization.github.io/folium/ (accessed on 2 April 2024).
26. Douc, R.; Moulines, E.; Stoffer, D. *Nonlinear Time Series: Theory, Methods and Applications with R Examples*; CRC Press: Boca Raton, FL, USA, 2014.
27. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2023.

28. Hyndman, R.J.; Khandakar, Y. Automatic time series forecasting: The forecast package for R. *J. Stat. Softw.* **2008**, *27*, 1–22. [CrossRef]
29. Shumway, R.H.; Stoffer, D.S. Time Series Regression and ARIMA Models. In *Time Series Analysis and Its Applications*; Springer New York: New York, NY, USA, 2000; pp. 89–212. [CrossRef]
30. Chen, T.; Guestrin, C. XGBoost. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
31. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: New York, NY, USA, 2017; pp. 4768–4777.
32. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
33. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, USA, 2009.
34. RStudio Team. *RStudio: Integrated Development Environment for R*; RStudio, Inc.: Boston, MA, USA, 2019.
35. Ahmed, M.S.; Cook, A.R. *Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques*; Transportation Research Record: Thousand Oaks, CA, USA, 1979.
36. Zhu, L.; Yu, F.R.; Wang, Y.; Ning, B.; Tang, T. Big data analytics in intelligent transportation systems: A survey. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 383–398. [CrossRef]
37. Habtemichael, F.G.; Cetin, M. Short-term traffic flow rate forecasting based on identifying similar traffic patterns. *Transp. Res. Part C Emerg. Technol.* **2016**, *66*, 61–78. [CrossRef]