Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

BREAKING THROUGH THE TRAFFIC CONGESTION: ASYNCHRONOUS TIME SERIES DATA INTEGRATION AND XGBOOST FOR ACCURATE TRAFFIC DENSITY PREDICTION

Eloi Garcia
Carles Serrat

Department of Mathematics - IEMAE - EPSEB
Universitat Politècnica de Catalunya-BarcelonaTECH
08028 Barcelona, SPAIN

Fatos Xhafa

Department of Computer Science
Universitat Politècnica de Catalunya-BarcelonaTECH
08034 Barcelona, SPAIN

1 ABSTRACT

The proliferation of data collection from smart cities has resulted in an exponential growth in the volume of measurements available for analysis. However, collecting all parameters concurrently at the same location is not feasible due to the complex nature of the real world. We present an innovative methodology that enriches asynchronous time series data from a variety of sources to facilitate data enrichment and city-wide behaviour simulation. A case study on OpenDataBCN attests to the efficacy of this approach via an XGBoost model, predicated on geographical coordinates and timestamp disparities. The consolidation of data from different sources improves the richness and granularity of information at disposal for analysis, thereby revealing previously hidden patterns and relationships, exhibiting new insights and underscoring the potential of this methodology for sustainable and efficient data enrichment processes as well as new possibilities for simulation based on smart city datasets.

2 INTRODUCTION

In recent years, the rapid urbanization and growth of cities have led to an increase in traffic congestion, posing significant challenges for urban planners and transportation authorities (Aluko 2019). The ability to accurately predict traffic density is crucial for effective traffic management and the development of sustainable urban mobility strategies. With the proliferation of open data platforms, such as OpenDataBCN (Ajuntament de Barcelona 2023) in Barcelona, there is an unprecedented opportunity to leverage vast amounts of heterogeneous data sources to develop accurate traffic prediction models. However, the integration of these diverse data sources, particularly those with asynchronous time series and varying geographical coordinates, remains a significant challenge. This paper aims to address this challenge by proposing a novel methodology for the asynchronous join of time series data based on geographical coordinates and timestamp differences, enabling the incorporation of previously unavailable information for improved traffic density prediction (Garcia et al. 2023).

In response to the exponential growth of geospatial data in urban areas, various approaches have emerged to address the correlated attribute patterns. This surge in data availability can be attributed to the widespread adoption of large-scale Internet of Things systems and the standardization of Cloud technologies (Xhafa

979-8-3503-6966-3/23/$31.00 ©2023 IEEE

et al. 2020). However, integrating the collected data poses challenges in terms of interconnection due to the complex nature of the real world.

In particular, first approaches like Khatri et al. (2006) or Donnay and Linke (2022) appear to address this topic, starting with the additional layers of complexity that the temporal and geographical dimensions add to data storage and manipulation, but always based on predefined pairs of time series or fixed timestamps. Moreover, approaches like Harada et al. (2020) go a step further by also exploring the correlations between the collected data and these geographical and temporal magnitudes. However, all these approaches fall short when generating data that makes possible the implementation of models for analysis and prediction, since they do not provide an integrated output that facilitates the analysis when this data is not gathered simultaneously and in the same geographical range.

The key contribution of this paper lies in the development of an innovative data integration technique that allows for the combination of asynchronous time series data from various sources, enhancing the richness and granularity of the information available for analysis. This advancement holds relevant implications for the simulation scene, where the fusion of diverse datasets can improve the accuracy and realism of the simulated models. Although no preliminary benchmarks exist to demonstrate the effectiveness of this methodology due to its novelty, a case study with illustrative purposes will be developed in Section 5. By utilizing the abundant open data provided by OpenDataBCN, this methodology exemplifies the potential of open data platforms in facilitating data-driven decision-making and fostering a deeper understanding of urban dynamics. Moreover, the integration of these diverse data sources enables the discovery of previously concealed patterns and relationships within these dynamics.

It is worth noticing that the relationship between traffic density and pollution is inherently complex, influenced by a multitude of factors. While the correlation between traffic density and emissions is not direct, the literature suggests the existence of some degree of association (Lipfert et al. 2006; Godec et al. 2021), therefore making this data collection a good fit for the development of an illustrative example. It is important to acknowledge the intricacies involved in this relationship and the various factors that contribute to the emission levels, that go beyond the scope of this case, that is developed around pollution data, although other geospatial and temporal datasets could be integrated with the same methodology. For example, public transportation availability, noise levels, among others, can be integrated and implemented using the same methods.

To evaluate the effectiveness of this data integration approach, we employ the end-to-end tree boosting system XGBoost (Chen and Guestrin 2016) model to analyze the resulting time series data for predicting traffic density in Barcelona. XGBoost is a powerful machine learning algorithm that has demonstrated exceptional performance in various prediction tasks, making it a suitable choice for this study. Through extensive experimentation and validation, we demonstrate the superiority of our proposed methodology in terms of prediction accuracy and the ability to uncover novel insights from the integrated data.

In conclusion, this paper presents a novel approach to the asynchronous join of time series data based on geographical coordinates and timestamp differences, enabling the effective utilization of open data platforms such as OpenDataBCN for improved traffic density prediction and metropolitan behaviour simulation. By harnessing the richness of open data and advanced machine learning techniques, we contribute to the ongoing efforts to make urban data more understandable and profitable for the general public, ultimately promoting more sustainable and efficient urban mobility solutions through the development of new simulations with richer and more precise data.

3 DATA SEMANTIC ENRICHMENT MODEL AND IMPLEMENTATION

3.1 Model

In this section, we provide a detailed description of the model used to associate an initial set of values with the closest available measurements based on geographical proximity and minimum difference in timestamps while considering the availability of additional data (Garcia et al. 2023). This model is designed to provide an accurate representation of the predefined conditions in different sections of a city, based on the closest measurements available. In the following section, we describe the model in detail, including the inputs, outputs, and the development of the data enrichment process.

3.1.1 Input

Let C be the set of sections of the city, and let T be the set of timestamps. For each section $c \in C$, let D_c be a dataset of initial values for c at different timestamps $t \in T$. Let S be a set of geographic locations associated with each dataset D_c. For each timestamp $t \in T$, let M_t be a set of measurements taken at different locations $s \in S_t$ at time t. Each measurement $m \in M_t$ has a geographic location s_m and a timestamp t_m, and may be associated with additional data D'_m that is available only at s_m.

3.1.2 Output

For each section $c \in C$ and timestamp $t \in T$, find the closest measurement $m_{c,t}$ in M_t to the initial values dataset D_c based on both geographical proximity and minimum difference on timestamps. If additional data D'_m is available for $m_{c,t}$, associate it with the corresponding initial dataset value in D_c.

3.1.3 Semantic Data Model

For each section $c \in C$ and timestamp $t \in T$, we can compute the closest measurement $m_{c,t}$ in M_t as follows:

STEP 1. Compute the geographic distances $d(s, D_c)$ between each location $s \in S$ associated with D_c and each location $s' \in S_t$ in M_t:

$$d(s, D_c) = \min_{s' \in S_t} \{\text{distance}(s, s')\},$$

where $\text{distance}(s, s')$ is the geographical distance between s and s'.

STEP 2. Find the measurement $m_{c,t}$ that minimizes the sum of the geographic distance and the absolute time difference with D_c:

$$m_{c,t} = \arg\min_{m \in M_t} \{d(s_m, D_c) + |\min_{m' \in M_t} \{|t_m - t_{m'}|\} - t|\}.$$

In other words, we find the measurement m in M_t that has the minimum sum of geographic distance with D_c and the absolute time difference with the closest measurement in M_t to t. The closest measurement in M_t to t is obtained by computing the minimum time difference between all pairs of measurements in M_t.

STEP 3. If additional data D'_m is available for $m_{c,t}$, associate it with the corresponding initial set value in D_c:

$$\text{initial set}(c,t) = \begin{cases} (\text{initial set}(c,t), \text{value}(m_{c,t})) & \text{if } D'_m \nexists \text{ for } m_{c,t}, \\ (\text{initial set}(c,t), \text{value}(m_{c,t}), D'_m) & \text{otherwise.} \end{cases}$$

In other words, if additional data is not available for $m_{c,t}$, we simply associate the initial dataset value from the closest measurement in M_t to D_c. Otherwise, we associate both the initial set value and the additional data D'_m.

3.2 Computational Complexity

We denote n as the number of rows in the *initial_dataset* input variable and m as the number of rows in the *extra_dataset* input variable. Since the initial definition of the algorithm contains nested loops that search for the closest measuring station and timestamp for each point in the initial dataset, the computational complexity is defined as $O(n^3)$. This means that the running time increases significantly as the size of the dataset grows.

The proposed model has a for loop that iterates over every point in the initial dataset, which has a size proportional to n. For each point, the function must calculate the distance to every point in the extra dataset, which has a size proportional to m. The calculation of normalized distances between the initial point and every point in the extra dataset has a time complexity of $O(m)$, as it must loop over every point in the extra dataset. Similarly, the normalization of the timestamps has a time complexity of $O(m)$ as it also must loop over every point in the extra dataset. After normalization, the function must combine the two measures into a single distance value for each point in the extra dataset. This operation has a time complexity of $O(m)$ as it involves element-wise addition of two arrays of length m. Finally, the function must find the point in the extra dataset that has the minimum combined distance from the current initial point. This operation has a time complexity of $O(m)$ as it must loop over every point in the extra dataset to find the minimum distance.

Therefore, the total time complexity of the proposed function is $O(nm^2)$. This is because for every point in the initial dataset, the function must perform $O(m^2)$ calculations to find the point in the extra dataset with the minimum distance. Any other computational improvement had to be discarded since the complexity of the possible geographical references could not be supported by the definition of the implementation.

In terms of space complexity, the function creates a new dataset to store the output, which has a size proportional to n. Additionally, the function creates a spatial index for the extra dataset, which has a size proportional to m. With this, we can state that the total space complexity of the function is $O(n+m)$.

4 CASE STUDY: FROM TIME SERIES DATA TO SEMANTICALLY ENRICHED DATA

The growth of urban areas has brought about a significant increase in the number of vehicles on the roads, resulting in an unprecedented level of traffic congestion. As a result, it has become increasingly important for city planners and transportation authorities to gain insights into traffic patterns and trends, in order to optimize traffic flow and reduce congestion. In this case study, we explore the use of an algorithm for associating traffic density values with measurements and additional data, as a means of moving from time series data to semantically enriched data. By applying this algorithm to a real-world dataset, we aim to demonstrate the potential of this approach in improving the accuracy and reliability of traffic analysis, and supporting the development of more effective traffic management strategies.

4.1 Preprocessing, Feature Engineering, and Standardization

A preprocessing method is defined with the ultimate goal of standardization and interoperability, with the adoption of time and space filters, and it involves several key steps (see processing workflow in Figure 1).

In the first step, the user selects a dataset of interest from the Open Data Barcelona portal, specifying a specific time and geographic range. Subsequently, built-in methods request the data through an API, returning the raw dataset in its original form. It is important to emphasize that each chosen dataset requires a unique preprocessing and cleaning method due to the differences in data sources and structures.

To ensure the necessary treatment for each dataset, we have developed a modular approach that involves the removal of invalid and redundant data, identification and elimination of duplicate entries, and standardization of geographical and temporal properties.

Finally, this method employs built-in predefinitions of different sections of the city to extract the selected period and section of data chosen by the user. This facilitates targeted analysis of specific neighborhoods, streets, and other relevant city subdivisions. In summary, the proposed preprocessing method plays a crucial role in enabling the analysis of Open Data Barcelona datasets by providing clean and standardized data that is readily available for further exploration and analysis.

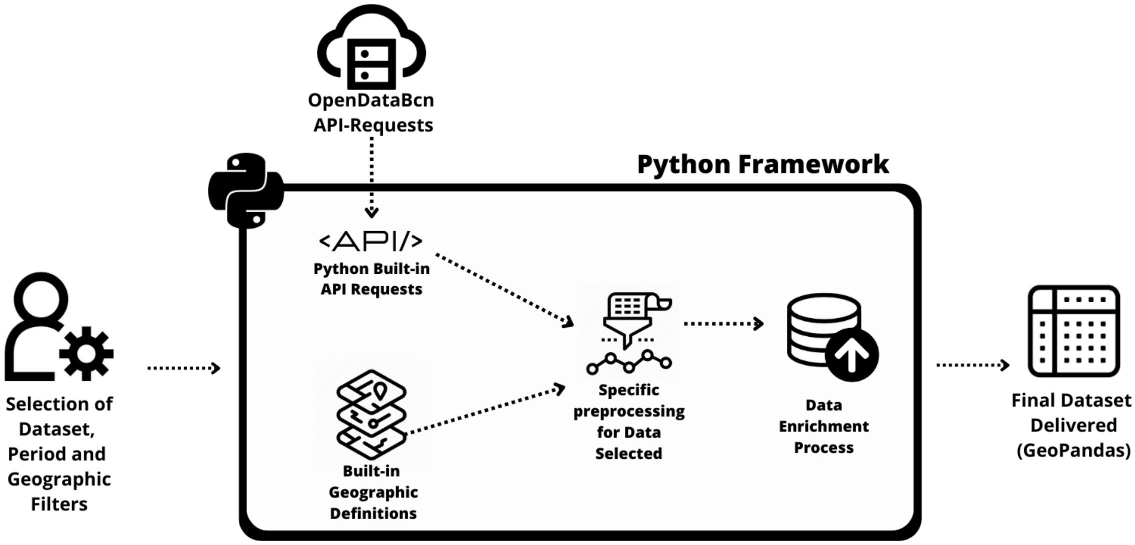


Figure 1: Preprocessing flow diagram designed for the use case.

4.2 Open Data Barcelona Data Selection and Geographic Sections Definition

We present the datasets included in the case study for the asynchronous join process. The table includes information on the number of different geographical locations contained in each dataset, their geographic definition, and the type of information that can be retrieved from them. Additionally, the estimated update rate and periodicity in which the data is updated according to what Open Data Barcelona states in each dataset metadata.

- Traffic Density: Traffic Density values and Initial Locations in a categorical internal system (from no traffic to jam). Updated every 15 minutes (when there is data available) and collected from 534 sections of streets defined in the city map of Barcelona.
- Pollutants: Eight pollution sensors coordinates (points) spread out on Barcelona city that detect different pollutants every hour (when there is data available). For this case, carbon monoxide measured in mg/m^3, nitrogen monoxide measured in $\mu g/m^3$ and black carbon concentration measured in $\mu g/m^3$.

4.3 Computational Results

In this section, we present the computational results obtained from using the methodologies and implementations discussed in this study, to merge the selected data into a homogeneous time series for the different street sections in Barcelona, keeping the original objective of providing a comprehensive view of the traffic data in Barcelona and presenting a unified time series data.

Combining data from different sources provides a broader perspective of the city's traffic and helps in identifying trends and patterns that may not have been apparent in individual datasets, as shown in Figure 2. The resulting time series provides a valuable resource for analyzing and understanding the traffic patterns in Barcelona. For example, correlations between certain pollution levels and areas of the city that are more prone to traffic congestion can now be evaluated, along with tracking changes in traffic patterns over time based on the availability of other methods of public transportation.

One significant advantage of the proposed methodology is that it enables the application of more advanced analysis methods like windowed cross-correlation (Boker et al. 2002), a time series analysis

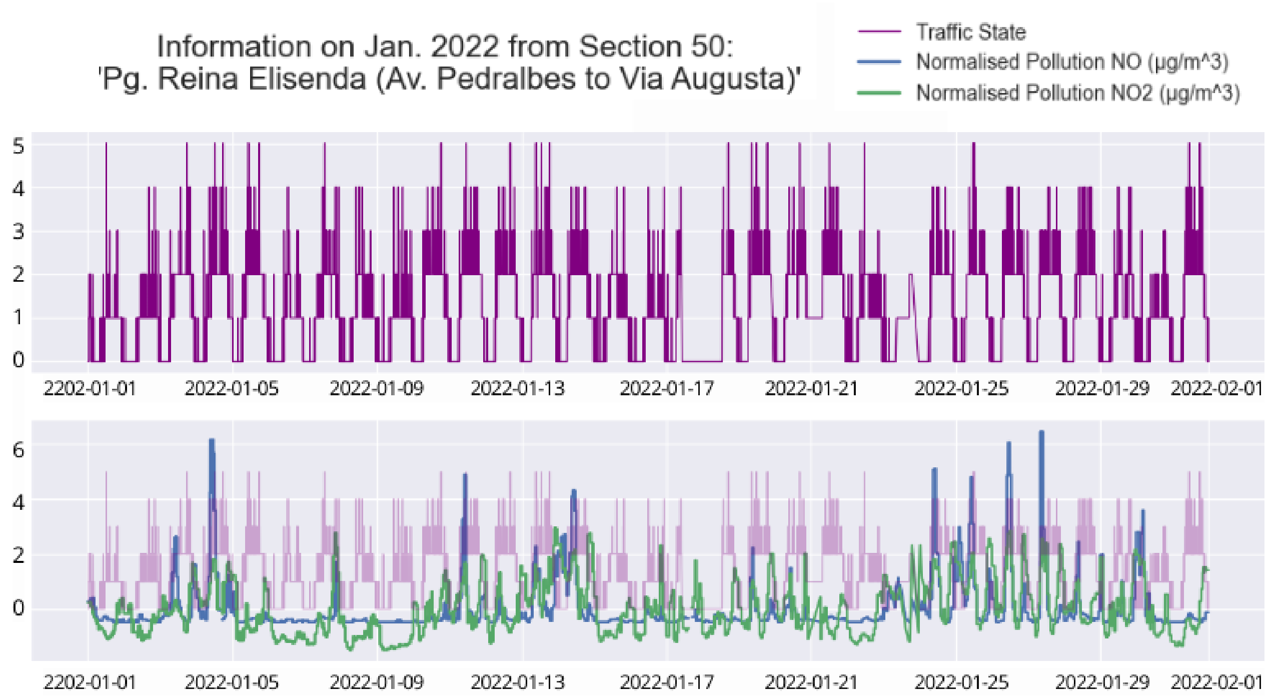


Figure 2: Comparison of two time series: original traffic density (top), generated normalised NO/NO2 pollution (bottom).

technique used to identify correlations and time lags between different time series data points. With this kind of tools now available, users can identify time lags between two time series that came from different sources.

4.4 The Novelty of the Methodology: Lack of Comparative Benchmark

The methodology proposed in this study for moving from time series data to semantically enriched data is a novel approach specifically designed for analyzing temporal patterns in urban areas. Due to its novelty, there are currently no established comparative benchmarks for evaluating its performance. While the absence of a comparative benchmark poses a challenge, the computational results obtained from real-world datasets can demonstrate the methodology's potential and effectiveness, showing the marginal contribution of the new variables not available before, and improvements on overall performance.

As this methodology gains traction, future research and adoption may lead to the emergence of comparative benchmarks. These benchmarks would facilitate a more comprehensive evaluation and further advancements in semantically enriched data analysis for traffic management.

5 EXTREME GRADIENT BOOSTING MODEL COMPARISON

Extreme Gradient Boosting (Chen and Guestrin 2016) (XGBoost) is a scalable machine learning method for tree gradient boosting, suitable for predicting traffic density categorizations and providing the predictive capacbilities of each value upon after-training analysis. In addition, XGBoost can handle missing values and has the ability to identify and prioritize the most informative features, which is useful for feature selection (Alsahaf et al. 2022).

XGBoost generates an ensemble $\mathscr{F} = \{f(\boldsymbol{x}) = w_{q(\boldsymbol{x})}\}(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T)$ of K-Classification and Regression Trees (CART), predicting output as $\hat{y}_i = \phi(\boldsymbol{x}) = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i),\ f_k \in \mathscr{F}$.

The learning process involves a regularized objective convex loss function:

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

where $\Omega(f) = \lambda T + \frac{1}{2}\lambda ||w||^2$ penalizes over-fitting.

XGBoost is trained using an additive method, minimizing the objective function like

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

adding f_t greedily based on optimization performance.

The objective function can be simplified using second-order Taylor expansion to

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

The function can be expanded again through its regularization term Ω for instance set I of leaf j for given structure $q(x_i)$ as $I_j = \{i | q(x_i) = j\}$ to

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

The optimal leaf weight w_j^* and goodness of fit of overall structure $q(x)$ can be found as

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

and

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

XGBoost starts from a single leaf and on every iteration a new branch is added using a greedy algorithm, measuring possible gain as

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

5.1 Evaluation Metrics

5.1.1 Area Under the ROC Curve (AUC)

The evaluation of this classification system is chosen to be the area under the ROC curve (AUC) with an Over to Rest (OrV) strategy, since it is a very well-known measure of performance for machine learning models. The Over the Rest strategy implies that any missclassification is considered a wrong classification no matter which two classes are being mistaken. With this, we define our AUC method as

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

where α is the probability of a false positive and $1 - \beta$ the probability of a true positive.

5.1.2 SHapley Additive Explanations (SHAP)

SHapley Additive Explanations (SHAP) is a machine learning model interpretation tool based on game theory (Lundberg and Lee 2017). With SHAP, the attribution of every variable to the final output is measured by taking one variable into the model at a time and measuring the expected value of the function of the model's output. With this changes produced by the addition of each variable, SHAP computes the average contribution by measuring the impact of every possible variable orderings.

The average contribution of the variable x in the model f can be computed as:

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

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

A single understandable solution is generated with SHAP since three main properties are present:

- Local accuracy: when the function that relates x to x' defined as $h_x(x')$ equals the set of variables x, and therefore the approximation of f equals to the output of f.
- Missigness: when the values that are missing have no impact on the output of the model, $x'_i = 0 \rightarrow \phi_i = 0$.
- Consistency: if a model changes so that some input's contribution increases or stays the same, the Shapley value also has to increase or stay the same regardless of the other inputs.

SHAP, and in particular Tree SHAP (Lundberg et al. 2020), can provide explanations for the impact of variables in tree-based machine learning models such as the one used. Tree SHAP performs an exploration of the model with an input dataset X of size $N \times M$ and produces a matrix of equal size with the SHAP values for every variable on every tuple in X. With this values, consistency of explanation in guaranteed for individual predictions, along with an identification of the contribution with a sign indicator.

5.2 Experimental Results

In this section, the results of the prediction model, its performance metrics and a variable analysis are described and analysed.

5.2.1 Classes Identified on the Data

For the purpose of applying this method for parameter importance analysis, we consider the different densities obtained from the original dataset as different classification options following: 0 = very fluid, 1 = fluid, 2 = dense, 3 = very dense, 4 = congestion.

5.2.2 Final Dataset

The dataset contains information structured following Table 1 with data from the entire year 2019. We split the dataset in two different sets for the training and test steps of the training and evaluation process. With this, we assign a 70% of the dataset to the training step and we keep the remaining 30% for evaluation of the parameters taken. Finally, data from January to March of 2022 is processed for the final analysis and predictions shown at the model performance section. The last three parameters, marked with $*$ on Table 1, are added on another iteration of the model with the same parameters to compare the impact on the accuracy on the new model.

For the XGBoost, the parameters were tuned with a cross validation tool defined as a grid search with 5 folds, containing different possible configurations for the following parameters within reasonable ranges, as we can see on Table 2.

Table 1: Dataset description for the XGBoost model.

Variable	Description	Type	Range
Section ID	Unique identifier of the Barcelona Road sections.	number	1 to 534
Status	Current traffic status of the section.	number	0 to 4
FromNorth	Geographic coordinate system Latitude where the section starts (North).	number	2,099 to 2,222
FromWest	Geographic coordinate system Longitude where the section starts (West).	number	41,338 to 41,450
ToNorth	Geographic coordinate system Latitude where the section ends (North).	number	2,100 to 2,223
ToWest	Geographic coordinate system Longitude where the section ends (West).	number	41,338 to 41,449
DailyHour	Hour of the day when the status was measured.	number	0 to 23
DailyMinute	Minute of the hour when the status was measured.	number	0 to 60
Weekday	Day of the week when the status was measured.	number	1 to 7
DayMonth	Day of the month when the status was measured.	number	1 to 31
Holiday	Boolean value representing a holiday on the corresponding day.	Boolean	False or True
*pollution_6	Carbon monoxide density level measured in mg/m^3.	number	0.057 to 0.5
*pollution_7	Nitrogen monoxide density level measured in $\mu g/m^3$.	number	1 to 41
*pollution_22	Black carbon density level measured in $\mu g/m^3$.	number	125.43 to 3183
Number of records: 24,533,298			

Table 2: Parameter matrix for XGBoost.

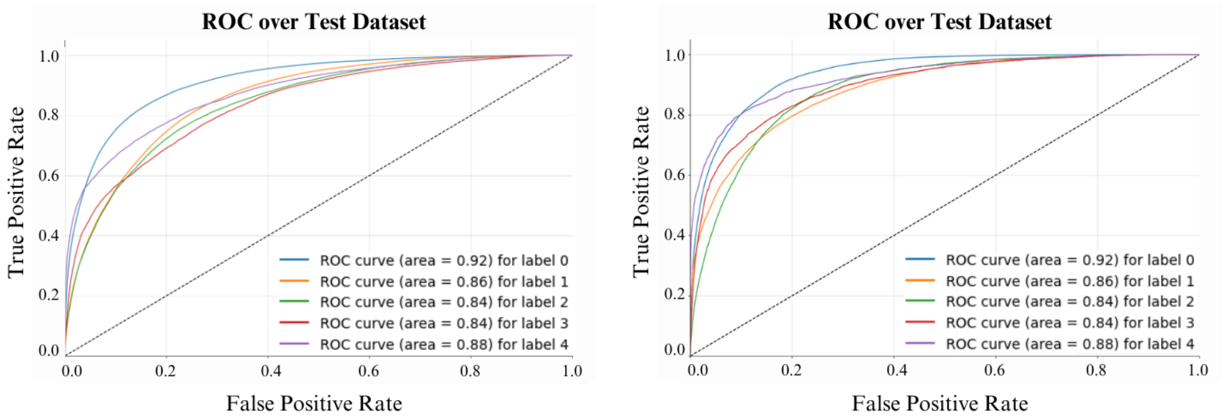
Parameter	Final value	Cross validation candidates
max_depth	10	5, 7, 9, 10, 11
eta	0.3	0.1, 0.2, 0.3, 0.4
gamma	1	0.5, 1, 1.5, 2, 5
subsample	1	0.6, 0.8, 1
objective	multi:softmax	-
num_class	5	-

5.2.3 Model Performance

The prediction accuracy of the initial model, specifically when considering the middle classes of traffic density, has been identified as a significant weakness. These middle classes represent approximately 10% of the total dependent variable values and have shown lower accuracy in predictions, as evidenced by the ROC plot.

However, with the inclusion of the new variables, the performance of the model has significantly improved. The Area Under the Curve (AUC) in Figure 3 demonstrates the overall positive performance of the model. Notably, the AUC values for each ROC curve are consistently above 80%, even without the addition of the pollution data.

It is important to highlight that the initial model performs better, on average, for extreme cases, with a true positive rate exceeding 85%. However, there is a drop in performance for the intermediate classes. This drop is mitigated in the model that incorporates the pollution data, indicating an increase in accuracy specially where the original variable composition failed. The introduction of the new variables, such as pollutant measurements, has contributed to improving the prediction accuracy of the model, particularly for the middle classes of traffic density. The ROC plot demonstrates that the model's performance is highly positive, even without the inclusion of pollution data.



(a) AUC for the initial XGBoost model.

(b) AUC for the XGBoost model with pollution data.

Figure 3: Effects on the AUC after the incorporation of the pollution data.

On average, this model presents an accuracy rank of 71.56% for the data of 2019 without the addition of the pollution data. This accuracy increases to 79.65% with the pollution data added as parameters in the model.

5.2.4 Variable Analysis and Shapley Additive Explanations

The importance on average of every variable for every class can be assessed by using Shapley Additive Explanations, a method from cooperative game theory that increase the interpretability of the final output values. Figure 4 shows the model's impact of the different combinations of variables for traffic density class 1, with similar results being found in the remaining classes.

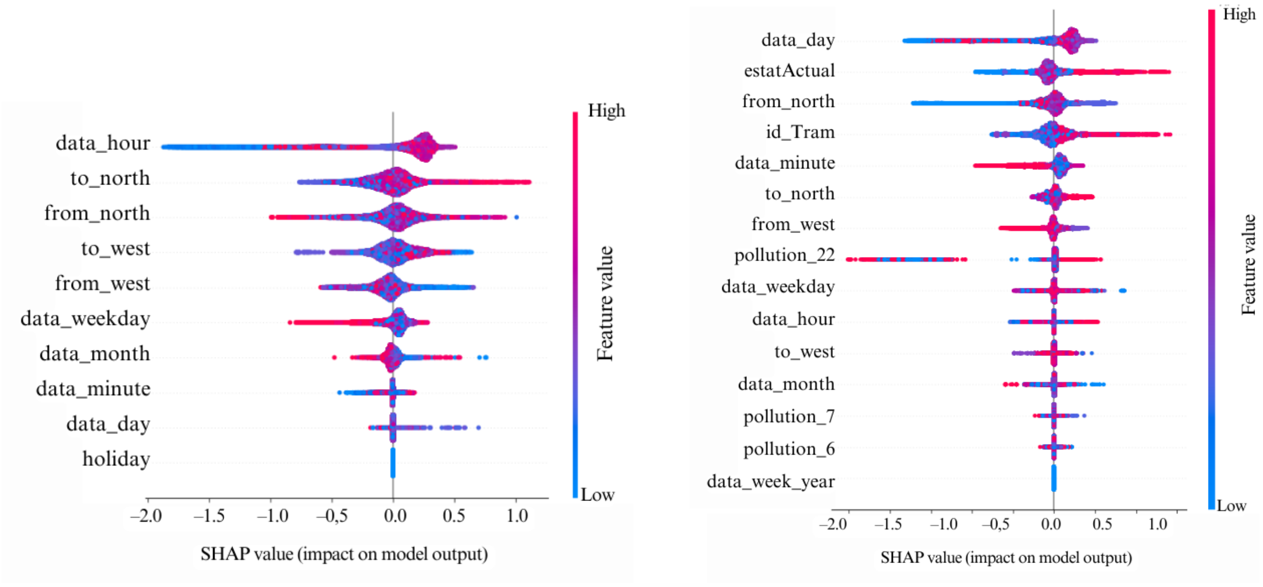
Interestingly, the analysis revealed that the new pollution index variables were not among the top three most important variables. However, we observed that they still had a positive effect on the model's classification for each traffic density. This suggests that the new pollution index variables provided valuable information for predicting traffic density, despite not being among the top-ranked variables.

The positive effect of the new pollution index variables on the model's prediction of traffic density may be due to their ability to capture the impact of pollution on traffic (European Environment Agency 2022). Pollution levels and traffic density are related, as high levels of pollution can be directly linked with traffic density, and they can reduce visibility and increase the likelihood of accidents. Furthermore, pollution can affect the health and well-being of individuals living in the area, which can impact their travel patterns and behaviours.

By including the new pollution index variables in the XGBoost model, we were able to capture this important relationship between pollution and traffic density, which contributed to the model's overall predictive power. Our findings highlight the importance of considering a broad range of variables when developing predictive models for traffic density, as even variables that may not be among the most important can still provide valuable insights and improve the accuracy of the model's predictions.

6 CONCLUSIONS

This study presents a novel approach to address this problem by integrating asynchronous time series data from diverse sources to develop a more accurate traffic density prediction. The study used an XGBoost model based on geographical coordinates and timestamp differences to demonstrate the effectiveness of this approach on OpenDataBCN. The integration of data from multiple sources improves the richness and granularity of information available for analysis, revealing previously hidden information that could not be applied without the methodology applied. The XGBoost model demonstrates superior prediction



(a) Impact on model without pollution data: Class 1. (b) Impact on model with pollution data: Class 1.

Figure 4: Impact on model output on predicted class 1.

accuracy and indicates the potential of this approach for sustainable and efficient urban mobility solutions. Furthermore, this approach can be useful for generating and analyzing city traffic data through simulation tools. By utilizing open data platforms and advanced machine learning techniques, this novel approach offers a new paradigm for addressing the challenges of urban mobility in a data-driven manner.The XGBoost model developed in this study demonstrates that the addition of pollution parameters to traffic density data using a data enrichment model for asynchronous data improves the precision of the model. This conclusion is supported by the results of the ROC plot, which shows that the model performs better when pollution parameters are included and a SHAP exploration that shows the relative importance that these variables have on the overall model performance.

ACKNOWLEDGMENTS

This work was partially funded by the Spanish Ministry of Science (PID2019-111100RB-C21 /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), and it was supported by the Departament de Recerca i Universitats de la Generalitat de Catalunya (Spain) (2021 SGR 01421 (GRBIO)).

REFERENCES

Ajuntament de Barcelona 2023. “Open Data BCN”. https://opendata-ajuntament.barcelona.cat/en/open-data-bcn. accessed 13th February 2023.

Alsahaf, A., N. Petkov, V. Shenoy, and G. Azzopardi. 2022. “A Framework for Feature Selection Through Boosting”. *Expert Systems with Applications* 187:115895.

Aluko, O. 2019. “A Review of Urbanisation and Transport Challenges in Developing Countries”. *International Journal of Innovation Education and Research* 7:315–323.

Boker, S. M., J. L. Rotondo, M. Xu, and K. King. 2002. “Windowed Cross-Correlation and Peak Picking for the Analysis of Variability in the Association Between Behavioral Time Series”. *Psychological Methods* 7:338–355.

Chen, T., and C. Guestrin. 2016. “XGBoost”. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 6th–10th August, Halifax, NS, Canada, 785–794.

Donnay, K., and A. M. Linke. 2022. *Geomerge: Geospatial Data Integration*. Vienna, Austria: R Foundation for Statistical Computing.

European Environment Agency 2022. "Emissions from Road Traffic and Domestic Heating Behind Breaches of EU Air Quality Standards Across Europe". https://www.eea.europa.eu/highlights/emissions-from-road-traffic-and.

Garcia, E., M. Peyman, C. Serrat, and F. Xhafa. 2023. "Join Operation for Semantic Data Enrichment of Asynchronous Time Series Data". *Axioms* 12(4).

Godec, R., I. Jakovljević, S. Davila, K. Šega, I. Bešlić, J. Rinkovec, and G. Pehnec. 2021. "Air Pollution Levels Near Crossroads With Different Traffic Density and the Estimation of Health Risk". *Environmental Geochemistry and Health* 43(10):3935–3952.

Harada, K., Y. Sasaki, and M. Onizuka. 2020. "MISCELA: Discovering Simultaneous and Time-Delayed Correlated Attribute Patterns". *Distributed and Parallel Databases* 39(3):637–664.

Khatri, V., S. Ram, and R.T. Snodgrass. 2006. "On Augmenting Database Design-support Environments to Capture the Geo-spatio-temporal Data Semantics". *Information Systems* 31(2):98–133.

Lipfert, F., R. Wyzga, J. Baty, and J. Miller. 2006. "Traffic Density as a Surrogate Measure of Environmental Exposures in Studies of Air Pollution Health Effects: Long-Term Mortality in a Cohort of US Veterans". *Atmospheric Environment* 40(1):154–169.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. 2020. "From Local Explanations to Global Understanding with Explainable AI for Trees". *Nature Machine Intelligence* 2(1):56–67.

Lundberg, S.M., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions". In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 4th–9th December, Long Beach, CA, 4768—4777.

Xhafa, F., B. Kilic, and P. Krause. 2020. "Evaluation of IoT Stream Processing at Edge Computing Layer for Semantic Data Enrichment". *Future Generation Computer Systems* 105:730–736.

AUTHOR BIOGRAPHIES

ELOI GARCIA is a Research Assistant at the Universitat Politècnica de Catalunya-BarcelonaTECH (UPC) in Spain. He holds a MSc degree in Data Science Methodology from the Barcelona School of Economics (BSE). Garcia's research interests include data engineering, process mining, and geographical predictive models. In addition to his academic work, Garcia has worked as a data scientist for various companies, where he has applied his skills in data engineering and process mining to solve real-world business problems. Garcia's email address is eloi.garcia.climent@upc.edu and his website is https://futur.upc.edu/EloiGarciaCliment.

CARLES SERRAT is an Associate Professor at the Department of Mathematics at the Universitat Politècnica de Catalunya-BarcelonaTECH, at the Barcelona School of Building Construction (Catalonia, Spain). His areas of research include, but are not limited to, methodological and applied statistics as well as methaheuristics to fields like public health, construction, civil engineering, economy, logistics, and transport. Specifically he focuses on approaches based on survival analysis techniques, longitudinal data analysis, and missing data analysis. Professor Serrat has been granted for visiting scholarships at Harvard University and Hasselt University and visiting researcher stays at Open University of Catalonia, Trinity College Dublin, Universidad Nacional de Colombia, and Universidad de La Sabana. His email address is carles.serrat@upc.edu, and his website is https://futur.upc.edu/CarlesSerratPie.

FATOS XHAFA, PhD in Computer Science, is Full Professor at the Technical University of Catalonia (UPC), Barcelona, Spain. He has held various tenured and visiting professorship positions. He was a Visiting Professor at the University of Surrey, UK (2019/2020), Visiting Professor at the Birkbeck College, University of London, UK (2009/2010) and a Research Associate at Drexel University, Philadelphia, USA (2004/2005). Prof. Xhafa is a member of IEEE Communications Society, IEEE Systems, Man & Cybernetics Society and Founder Member of Emerging Technical Subcommittee of Internet of Things. His research interests include IoT and Cloud-to-thing continuum computing, massive data processing and collective intelligence and optimization, among others. He can be reached at fatos@cs.upc.edu. Please visit www.cs.upc.edu/~fatos/ and at http://dblp.unitrier.de/pers/hd/x/Xhafa:Fatos.