

Reducing Pollution Health Impact With Air Quality Prediction Assisted by Mobility Data

Juan Morales-García^{ID}, Emilio Ramos-Sorroche, Sara Balderas-Díaz^{ID}, Gabriel Guerrero-Contreras^{ID},
Andrés Muñoz^{ID}, Jose Santa^{ID}, and Fernando Terroso-Sáenz

Abstract—Countries all around the world recognise the impact of air quality on public health, advocating for city centre decarbonisation and pollutant monitoring via Internet of Things technologies. Using data collected from these systems, it is possible to generate models that predict pollution based on regular patterns where mobility data can enhance the accuracy and robustness of these advanced machine learning models. This paper follows this approach, utilising vehicle traffic data from image recognition, on-site vehicle detectors, and synthetic data to maximise prediction accuracy in various urban environments. The results reveal that this proposal improves prediction for traffic-related pollutants, such as SO_2 and $\text{PM}_{2.5}$, which are linked to severe respiratory diseases. These results also highlight the role of synthetic data in enhancing prediction performance under limited datasets.

Index Terms—Pollution, prediction, machine learning, health, graph neural networks, internet of things, road traffic.

I. INTRODUCTION

AIR quality has become of prime importance among the measures to improve living conditions in cities. Many municipalities are implementing policies to reduce pollution caused by near industry activity and road traffic. It is demonstrated that air quality has a direct impact on health, being related to serious respiratory diseases including asthma, lung cancer, stroke, and lower respiratory infections [1].

Received 15 July 2024; revised 12 October 2024; accepted 21 November 2024. This work was supported in part by MCIN/AEI/10.13039/501100011033/ and by ERDF A way to do Europe through the Projects AwESome PID2021-122215NB-C33, ALLEGRO PID2020-112827GB-I00, ONOFRE-3 PID2020-112675RBC41, ONOFRE 4 PID2023-148104OB-C42, and WILLIOT CNS2022-136150, and in part by the Seneca Foundation - Science and Technology Agency of the Region of Murcia through the Project MovIoT 22228/PDC/23. (Corresponding author: Andrés Muñoz.)

Juan Morales-García is with the Department of Computer Engineering, Catholic University of Murcia (UCAM), 30107 Murcia, Spain (e-mail: jmorales8@ucam.edu).

Emilio Ramos-Sorroche, Jose Santa, and Fernando Terroso-Sáenz are with the School of Telecommunication Engineering, Technical University of Cartagena (UPCT), 30202 Cartagena, Spain (e-mail: emilio.ramos@upct.es; jose.santa@upct.es; fernando.terroso@upct.es).

Sara Balderas-Díaz, Gabriel Guerrero-Contreras, and Andrés Muñoz are with the Department of Computer Engineering, University of Cádiz (UCA), 11003 Cádiz, Spain (e-mail: sara.balderas@uca.es; gabriel.guerrero@uca.es; andres.munoz@uca.es).

Digital Object Identifier 10.1109/JBHI.2024.3508466

At the same time, the deployment of sensors and networks has increased dramatically thanks to the implementation of Smart City concepts. As a matter of fact, monitoring air quality is one of the most considered applications in these scenarios [2]. Recent research addresses the exploitation of data collected in smart cities for the particular case of air quality monitoring, by using Machine Learning (ML) algorithms to predict the evolution of air contamination.

Classic regression models based on Support Vector Machines or Linear Regression, as well as classification algorithms like Decision Trees or Random Forests, are being superseded by deep learning approaches based on neural networks. Among these, Graph Neural Networks (GNN) have gained popularity for modelling relationships between different spatial areas or data sources. Recent studies have shown that GNNs offer significant advantages over other neural network architectures for modelling air pollution in cities [3]. Unlike CNNs, which are better suited for regular data (such as images), GNNs allow direct modelling of the data structure as graphs, capturing complex spatial and temporal interactions between entities. In that sense, sensors and monitoring stations in cities are not arranged regularly, and their interactions depend on the city's infrastructure, naturally forming a graph. This ability to leverage both node-level features and their interactions is key to improving predictive accuracy in urban pollution forecasting systems [4].

While some works focus on extending the typology of sensors used in this task, including particle matter and climate/gas sensors [5], [6], the complementary use of indirect measurements can significantly contribute to the pollution estimation process. This is the case for road traffic, whose contribution to pollution is undeniable. It is considered that more than 70% of pollution is due to combustion vehicles, despite the gradual increase in electric and hybrid models.

Although the introduction of mobility data into the ML equation has been initially presented in the literature for pollution forecasting [7], its overall impact depending on type of data source (e.g., traffic sensors or crowdsensing schemes), the pollutant object to prediction, the targeted prediction time horizon, and even the potential different evaluation metrics, have not been considered until now. Moreover, the limitations of IoT deployments regarding data quality, bearing in mind noisy features of atmospheric pollution sensors, and disposal of significant data records in recent installations have not been properly considered in current studies.

This paper proposes improving air quality predictions in cities by incorporating additional mobility data from vehicle identification using image recognition from cameras and traffic sensors. This mobility data is combined with information from particle matter, climate, and gas sensors. Real-time data from specific

cities is also enhanced with synthetic data to support GNNs when faced with limited datasets. This is the first study to examine the impact of mobility and the type of mobility data sources on air pollution prediction. Hence, the main contributions of the work are:

- 1) Integration of mobility data as an extra feature in GNN-based pollution prediction in urban areas.
- 2) Assessment of the impact of mobility data sources and type of city in the prediction, considering current road traffic restrictions in European cities.
- 3) Use of synthetic data for GNN training aimed at improving prediction accuracy.

The rest of the paper is organised as follows: Section II reviews the related literature; Section III details the overall methodology followed in the study, including the cities and data sources considered, data preprocessing, and the creation of the prediction model; Section IV presents and analyses the results; finally, Section V highlights the main conclusions of this work and suggests future research directions.

II. RELATED WORK

Recent advancements in air quality monitoring have leveraged ML, particularly GNNs, to enhance prediction accuracy and analyze urban pollution dynamics. In particular, a GNN was developed to predict human mobility patterns using open datasets, which included data on human movements and weather conditions [8]. The model accurately predicted human movement between urban areas, demonstrating the ability of GNNs to manage and analyse large-scale spatial data, which is crucial for applications like nationwide mobility tracking.

A system integrating mobility and pollution data from various transportation sources was introduced in [7], using ML and data fusion techniques to improve pollution prediction, optimize infrastructure, manage traffic, and inform public health policies to reduce pollution exposure. DL models were employed to predict urban mobility patterns through the integration of synthetic data, facilitating urban transportation planning and management through the efficient handling of large spatiotemporal datasets [9]. Furthermore, a system integrating mobility and pollution data from various transportation sources was introduced in [7], using ML and data fusion techniques to improve pollution prediction, optimize infrastructure, manage traffic, and inform public health policies to reduce pollution exposure. LSTM networks and MLPs were applied to predict emergency room visits for respiratory diseases triggered by PM_{2.5} pollution in Beijing [10], while gradient boosting and Gaussian process regression were used to forecast respiratory disease occurrences [11].

A study in Belgium used mobile phone data from 5 million users to refine NO₂ exposure estimates by incorporating individual mobility, improving health impact assessments [12]. Another study highlighted how accounting for both home and work locations when assessing PM_{2.5} exposure reduced misclassification compared to using only residential data [13].

A nonlinear seasonal grey model was proposed to predict PM_{2.5} concentrations in the Yangtze River Delta, incorporating a seasonal adjustment factor, the Weibull Bernoulli grey model, and a cultural algorithm to capture the seasonality and non-linearity of the data [14]. Similarly, an ML model using XGBoost and SHAP analysis was employed to assess the impact of vehicle electrification on urban air quality and human

health in Valencia, Spain, revealing significant reductions in NO₂ associated premature mortality. However, the study also emphasized the need for complementary strategies to further reduce traffic and control various pollution sources to safeguard human health [15].

In [16], a Pollution Weather Prediction System (PWP) was developed for real-time air quality monitoring and prediction using sensor data. The system, based on linear regression and artificial neural networks, effectively forecasted air quality indices to support public health decisions. Following the well-established link between pollution and health, in [17], a significant correlation between PM_{2.5} pollution and COVID-19 mortality in Western Europe was found, emphasizing the need to reduce PM pollution to mitigate respiratory diseases. A mobility-marked Hawkes model was also introduced in [18], using telecommunication data to improve COVID-19 spread predictions by incorporating mobility patterns, thereby enhancing public health decision-making.

Therefore, based on the above, the integration of mobility data, advanced machine learning models, and detailed environmental data has significantly improved the accuracy and robustness of pollution prediction models. However, many studies lack the incorporation of vehicle traffic data or synthetic data required to maximize prediction accuracy in diverse urban environments. While advanced techniques and multiple data sources are used, a focus on traffic data or synthetic data for model training is often missing. Our study addresses these gaps by integrating detailed traffic data and synthetic datasets, resulting in significantly reduced prediction errors for pollutants such as PM₁₀, PM_{2.5}, CO, and SO₂. This combined use of real and synthetic data provides a more accurate and holistic framework for urban air quality monitoring, which is essential for assessing pollution exposure and its health impacts to inform effective public health interventions.

III. METHODOLOGY

In this section, we describe the target areas considered in our work, detailing the processes of collecting real data from different IoT deployments. This includes data preprocessing and the generation of synthetic data to augment the datasets used by artificial intelligence models, with the aim of enhancing prediction quality (Fig. 1). We also outline the methods used to forecast air pollution in each study area.

A. Study Area

This study has been performed in three different Spanish cities, namely Madrid (MAD), Bilbao (BIL) and Cartagena (CAR). These cities display significant geographical diversity. Madrid is located in the centre of the Iberian Peninsula, Bilbao is in the north of the peninsula, near the coast of the Bay of Biscay, and Cartagena is in the south-west of Spain. This geographical variation allows us to explore a wide range of climatic and geographical conditions that can influence the dispersion of pollutants.

The selected urban areas for our study have characteristics well-suited to our research objectives (see Fig. 2). Each area covers approximately 1.9 km² in Madrid, with 12 urban traffic cameras and 5 air-pollution sensors; 1.8 km² in Bilbao, with 17 traffic cameras and 2 air-pollution sensors; and 5 km² in Cartagena, with 3 traffic measurement points and 10 air-pollution sensors. These zones are situated in the downtown areas of their

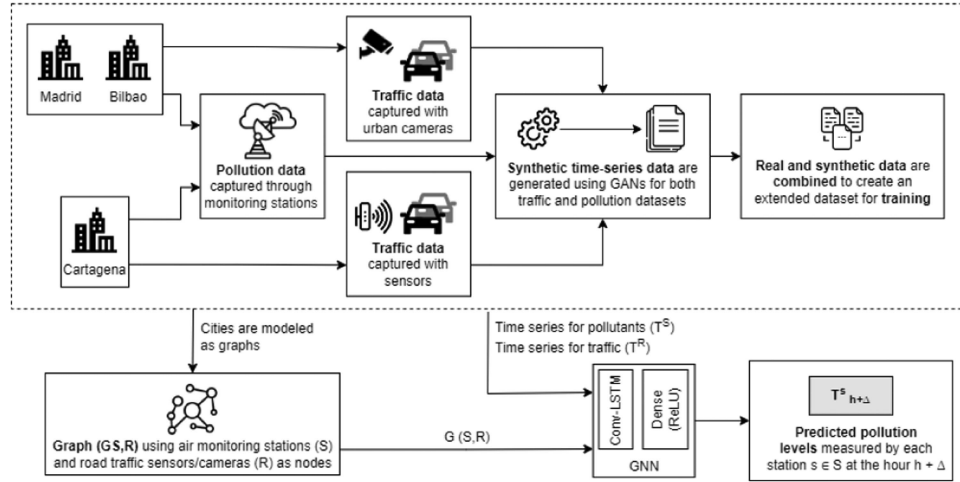


Fig. 1. Methodology flowchart outlining the integration of real and synthetic traffic and pollution data to predict urban air pollution levels.

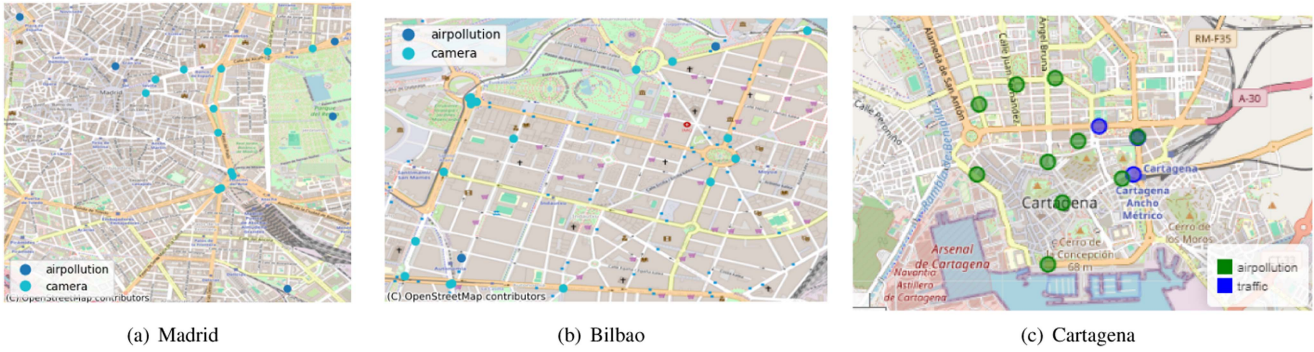


Fig. 2. Maps of the study area for each city.

respective cities, characterised by high population density and significant traffic volume.

B. Data Collection and Preprocessing

In this section we describe the different air pollution and traffic data feeds used for the present study.

1) *Air Pollution Data*: In order to collect the air pollution data from Madrid and Bilbao, we utilised their official open-data websites.¹ Concerning Cartagena, air pollutants are measured by a set of environmental stations maintained by the City Council. Through an agreement between Technical University of Cartagena (UPCT) and the local administration, we are able to obtain environmental data from 26 stations using a web service that allows queries filtered by date, although only 10 are used within the target area. Data used in this work belong to the period 13th September to 3rd October 2023. Data have been preprocessed for temporal resampling with hourly averages, and imputation is used to fill gaps using interpolation from closest measurements.

It is worth mentioning that the different air-pollution sensors measured different pollutant palettes, as Table I shows. This setting heterogeneity must be considered by the prediction model as we will see in Section III-D.

TABLE I

AIR POLLUTANTS MEASURED BY THE AIR QUALITY STATIONS IN EACH CITY

City	Sensor	SO ₂	CO	PM _{2.5}	PM ₁₀	O ₃
Madrid	m1	✓	✓			
	m2	✓	✓	✓	✓	✓
	m3			✓	✓	
	m4					✓
	m5		✓	✓		
Bilbao	b1	✓	✓	✓	✓	
	b2	✓	✓	✓	✓	✓
Cartagena	c1	✓	✓	✓	✓	✓
	c2	✓	✓	✓	✓	✓
	c3	✓	✓	✓	✓	✓
	c4	✓	✓	✓	✓	✓
	c5	✓	✓	✓	✓	✓
	c6	✓	✓	✓	✓	✓
	c7	✓	✓	✓	✓	✓
	c8	✓	✓	✓	✓	✓
	c9	✓	✓	✓	✓	✓
	c10	✓	✓	✓	✓	✓

2) *Road Traffic Data*: The road traffic data was collected using two distinct methods: image recognition from urban traffic cameras and vehicle detection via traffic sensors.

First, in the cities of Bilbao and Madrid, traffic cameras captured scenes every five minutes. The images were resized to

¹<https://www.ingurumena.ejgv.euskadi.eus>, <https://datos.madrid.es/>

a maximum width of 640 pixels, preserving their aspect ratios and maintaining full colour to ensure image quality. During data collection, occasional image freezes were noted from certain cameras. To avoid redundant analysis caused by these freezes, each image was assigned a unique hash code upon download, allowing for the identification and removal of duplicates before analysis.

The image analysis involved detecting and categorizing vehicle types (motorcycles, cars, buses, trucks) due to their varying impacts on traffic and pollution. YOLOv8x, a deep learning model for real-time object detection, was selected for its accuracy.

In the case of the city of Cartagena, traffic data was collected using 12 Sky Light Traffic Sensors from Nabla Quadro, installed along three streets (see Fig. 2(c)). These sensors employ a combination of ambient light variation, infrared detection, and magnetic field measurements to detect vehicles without requiring additional infrastructure. The data collected by the sensors includes the number of vehicles per lane and their direction of travel. This information is regularly transmitted to a central server in our laboratory via a Long-Range Wide Area Network (LoRaWAN). To ensure consistency and reliability in the analysis, data from all 12 sensors was aggregated at each of the three deployment sites. Finally, as for the case of air pollution data, all values have been reformatted, removing irrelevant information and adapting them to a common format for GNN training.

Using distinct data collection technologies for different cities allows us to validate the consistency of the results within each specific context. In Bilbao and Madrid, cameras provide detailed visual data for vehicle classification, while in Cartagena, sensors deliver precise traffic flow measurements. In particular, the two distinct traffic data collection methods validate the consistency of the results, ensuring robustness regardless of the capture technique. This comparison reduces the risk of technology-specific bias and confirms the reliability of the findings.

As a result of this data collection process, two types of time series are generated for each city: 1) a multivariate time series comprising the pollutants' measurements per hour, $\mathcal{T}^S$; and 2) another multivariate sequence comprising the vehicles counting per hour, $\mathcal{T}^R$.

C. Generation of Synthetic Data Using Time Series GANs

A Generative Adversarial Network (GAN) consists of two neural networks that are trained jointly and competitively: the generator and the discriminator. The generator's goal is to produce synthetic data from a random noise input, attempting to make these data as similar as possible to real data. Meanwhile, the discriminator receives both real and synthetic data and learns to distinguish between them. During training, both models improve their abilities adversarially: the generator tries to create more realistic data to deceive the discriminator, while the discriminator becomes more proficient at accurately identifying generated data. As the process progresses, the generator learns to produce data that are nearly indistinguishable from real data, thus achieving its goal of realistic data synthesis. In the case of time series, specific adaptations are required due to the sequential nature of the data. Since there are correlations across time, the generator must learn to maintain temporal consistency and the dependencies between values. To address this, specific architectures such as recurrent neural networks (RNNs) can be

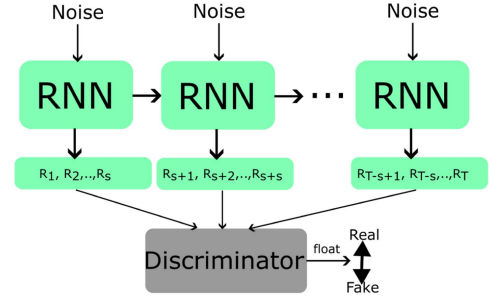


Fig. 3. DoppelGANger architecture used for the synthetic data generation.

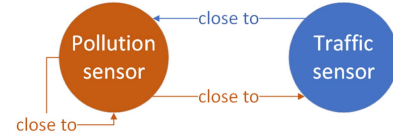


Fig. 4. Schema of the heterogeneous graphs used as input for the GNN models.

used in both components (generator and discriminator), enabling the effective modeling of temporal dynamics.

For the generation of synthetic of data, this work employed *DoppelGANger*, a GAN architecture tailored for sequential data proposed in [19]. As depicted in Fig. 3, the GAN architecture employed utilises RNNs to generate synthetic time series data. The generative aspect of *DoppelGANger* utilises a layer of LSTM cells with 100 units, employing a batch generation strategy. This approach enables the model to produce S consecutive records of synthetic time series data per pass, departing from the conventional method of generating a single record at a time, as exemplified by $R_1, R_2, \dots, R_S$ in Fig. 3.

This strategy facilitates a more precise capture of temporal correlations in extended series while minimising the number of steps required for data generation. Additionally, the GAN incorporates a normalisation mechanism for each input time series to alleviate the model-collapse issue commonly observed in many GAN models. The discriminator, comprising a MultiLayer Perceptron (MLP) with up to five layers, each containing 200 neurons and followed by a ReLU activation function, employs the *Wasserstein* loss function to evaluate differences between the real and the synthetic data.

Let us denote these sequences of synthetic air pollution and traffic data as $\mathcal{T}_{synth}^S$ and $\mathcal{T}_{synth}^R$, respectively.

D. Definition of the Heterogeneous Graph

In order to define the heterogeneous graph encoding using the aforementioned sensors deployments, it must include two different entities: the air monitoring stations ($\mathcal{S}$) and the road-traffic sensors ($\mathcal{R}$). Let us denote this graph as $\mathcal{G}^{\mathcal{S}, \mathcal{R}}$. Fig. 4 shows the schema of this graph. The graph incorporates several *close-to* edges among the two entities to model their spatial relationships, where each edge is assigned a weight based on the geographical distance between each pair of entities involved. This network of close-to relations enables the predictor to encode air pollution patterns that occur not only in nearby locations but also in distant ones, which might emerge in certain scenarios [20].

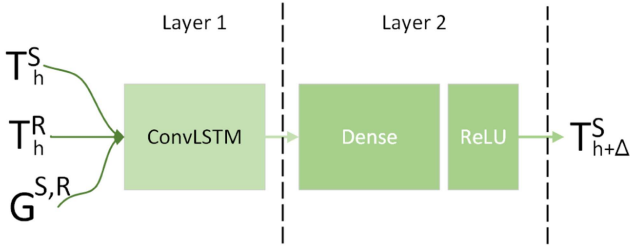


Fig. 5. Layered structure of the proposed GNN.

Lastly, it is worth mentioning that this design allows to include new data-sources and entities of interest in the graph. For example, if we wanted to enrich the model with meteorological data then we could add a new node type, *station*, defining the different weather stations installed across the target city and link them to the existing sensor nodes with *close-to* edges.

E. Air Pollution Prediction

The air pollution forecasting task is defined as the following regression problem:

Given the hour h , the input graph $\mathcal{G}^{S,R}$, and the sets $\mathcal{T}_h^S$ and $\mathcal{T}_h^R$, comprising the time series of all the air monitoring stations $\mathcal{S}$ and road-traffic sensors $\mathcal{R}$, respectively, we aim at finding a mapping function $\mathcal{F}$

$$\mathcal{F}(\mathcal{G}^{S,R}, \mathcal{T}_h^S, \mathcal{T}_h^R) \rightarrow \mathcal{T}_{h+\Delta}^S \quad (1)$$

where $\mathcal{T}_{h+\Delta}^S$ are the predicted pollution levels of the particles measured by each station $s \in \mathcal{S}$ at the hour $h + \Delta$, where $\Delta (\geq 1)$ is the time horizon of the prediction.

F. Graph Neural Network Design

Finally, given the heterogeneous graph and the regression problem expressed in (1), we developed the Graph Neural Network architecture depicted in Fig. 5.

As shown in this figure, the proposed architecture consists of two distinct layers. More specifically, the first layer is responsible for capturing the temporal evolution of the time series generated by the air-pollution stations and the road-traffic sensors. To achieve this, we apply an Integrated Graph Convolutional Long Short Term Memory (Conv-LSTM) model [21].

This model generates embeddings that represent the latent features of air-pollution and road-traffic sensors based on their temporal evolution. These embeddings are then passed through a dense layer with the Rectified Linear Unit (ReLU) activation function. This dense layer enables the model to learn relationships among pollutants and stations that are not explicitly encoded in the heterogeneous graph.

IV. RESULTS

In this section, we present the main results of the evaluation performed in the three target settings using the data and methods stated in the previous section.

A. Evaluation Metrics

For a precise evaluation of the models, three different metrics are employed:

TABLE II
MODEL PARAMETERS FOR THE EXPERIMENTS

Type	Parameter	Value
Training	Loss	Mean Squared Error (MSE)
	Learning factor	0,001
	Batch size	12
	Optimizer	Adam
	Early Stop	patience: 5, minDelta: 0.001
Conv-LSTM	Num. of epochs	900
	Output size	128
Dense	Num. cells	2
	Input size	128
	Output size	Num. of target particles

- **Mean Absolute Error (MAE):** This metric denotes the mean deviation between the predicted and actual values at each forecast point, thereby indicating the average magnitude of the prediction errors. It is defined by the formula:

$$\text{MAE}(y, \hat{y}) = \frac{\sum_{i=0}^{N-1} |y_i - \hat{y}_i|}{N}$$

- **Mean Squared Error (MSE):** This metric measures the average of the squares of the errors or deviations, thereby acting as an indicator of prediction accuracy. It is defined by the formula: $\text{MSE}(y, \hat{y}) = \frac{\sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2}{N}$

- **Mean Absolute Percentage Error (MAPE):** This metric expresses the magnitude of the absolute error in percentage terms, offering a relative measure of prediction accuracy. It is defined by the formula: $\text{MAPE}(y, \hat{y}) = \frac{100\%}{N} \sum_{i=0}^{N-1} \frac{|y_i - \hat{y}_i|}{y_i}$

B. Hyperparameter Tuning

In order to train the GNN, we used a 80-10-10 split of the time series for training, testing and evaluation, respectively. Concerning the hyperparameters used in the evaluation of the proposal, Table II comprises the most relevant ones related to the two layers of the GNN (Section III-F).

C. Exploratory Data Analysis

The following section presents an exploratory data analysis (EDA), along with a brief discussion regarding the limitations of synthetic data and its use in deep learning models.

Table III presents a statistical analysis comparing real and synthetic data. This analysis includes the total number of records, the mean of the variable, its standard deviation, the minimum and maximum values, as well as the quartiles. Additionally, the table shows the difference between the values of the real and synthetic time series for each of these statistics and for each of the studied variables. As can be observed, the difference between the statistics of the real and synthetic data is minimal, indicating that both time series are very similar in terms of their statistical properties.

To visually verify the similarity between the synthetic data generated and the real data, Fig. 6 presents the measurements of both types of data over a one-week period for all variables studied in Bilbao. It can be observed that the trends are similar across all data. Regarding pollutants, it is worth highlighting the case of the O3 and SO2 variables, which form an almost periodic pattern of peaks and valleys; the match is nearly perfect, except for a slight variation in one of the valleys at the beginning. As for traffic, the Car and Bus variables stand out, as they, like the

TABLE III
COMPARATIVE ANALYSIS OF REAL AND SYNTHETIC DATA

		count	mean	std	min	25%	50%	75%	max
CO	Real	350.00	0.33	0.09	0.14	0.27	0.33	0.39	0.67
	Synthetic	87,600.00	0.33	0.08	0.15	0.27	0.32	0.38	0.66
	Difference	87,250.00	-0.00	-0.01	0.01	0.00	-0.01	-0.01	-0.01
O3	Real	350.00	35.06	25.90	1.00	10.00	29.25	59.38	96.00
	Synthetic	87,600.00	35.65	25.73	1.10	10.79	32.06	58.47	95.77
	Difference	87,250.00	0.59	-0.17	0.10	0.79	2.81	-0.91	-0.23
PM2.5	Real	350.00	9.76	4.45	1.33	6.33	9.00	12.67	23.67
	Synthetic	87,600.00	9.44	3.92	1.56	6.51	8.88	11.84	23.63
	Difference	87,250.00	-0.32	-0.53	0.23	0.18	-0.12	-0.83	-0.04
PM10	Real	350.00	17.32	7.31	3.67	12.00	16.33	21.67	45.33
	Synthetic	87,600.00	16.98	6.56	3.88	12.06	15.99	20.87	42.41
	Difference	87,250.00	-0.34	-0.75	0.21	0.06	-0.34	-0.80	-2.92
SO2	Real	350.00	7.23	2.81	4.33	5.50	6.33	8.00	28.00
	Synthetic	87,600.00	7.30	2.71	4.34	5.55	6.52	8.14	27.87
	Difference	87,250.00	0.07	-0.10	0.01	0.05	0.19	0.14	-0.13
Car	Real	350.00	1,605.74	677.93	638.00	988.50	1,544.50	2,027.75	3,367.00
	Synthetic	85,690.00	1,622.76	677.47	638.00	996.00	1,565.00	2,047.00	3,367.00
	Difference	85,340.00	17.02	-0.46	0.00	7.50	20.50	19.25	0.00
Motorcycle	Real	350.00	4.65	3.97	0.00	1.00	4.00	7.00	22.00
	Synthetic	87,600.00	4.71	3.99	0.00	1.00	4.00	7.00	22.00
	Difference	87,250.00	0.05	0.03	0.00	0.00	0.00	0.00	0.00
Bus	Real	350.00	128.06	91.31	1.00	34.75	119.50	202.75	332.00
	Synthetic	86,526.00	132.78	91.41	1.00	39.00	125.00	209.00	332.00
	Difference	86,176.00	4.71	0.10	0.00	4.25	5.50	6.25	0.00
Truck	Real	350.00	177.86	130.67	15.00	75.00	113.00	274.00	572.00
	Synthetic	86,316.00	188.20	133.20	15.00	79.00	131.00	297.00	572.00
	Difference	85,966.00	10.33	2.53	0.00	4.00	18.00	23.00	0.00

For each studied variable (shown in the first column), the following statistics are provided: count, mean, standard deviation (std), minimum value (min), first quartile (25%), second quartile or median (50%), third quartile (75%), and maximum value (max). These statistics are reported for both real and synthetic data, along with the differences between them.

pollutants, exhibit very similar behaviors and trends between the real and synthetic data.

Although synthetic data are beneficial for overcoming data scarcity, they present certain limitations that can significantly affect the performance of machine learning models. These limitations arise from the differences between synthetic data and real-world data, potentially leading to biases, inaccuracies, and reliability issues. For instance, one of the most important limitations is that synthetic data may not accurately reproduce the statistical properties of real-world data, which could result in models performing poorly when applied to real-world tasks. Another limitation stems from the noise and irregularities typically present in real data, which are often absent in synthetic data. This lack of noise could prevent trained models from handling noisy data, thus reducing their ability to generalize effectively. Additionally, potential knowledge gaps within a domain should be noted, as synthetic data are based on real data for their generation. This means that new situations not previously observed cannot be generated, introducing a certain bias into the synthetic data. Lastly, it is important to highlight the computational costs of generating synthetic data, as they often require expensive hardware to run generative models and a significant amount of computing time.

Based on the previously presented statistical and graphical analysis, it can be observed that the synthetic data generated suggest that they may be suitable for use in training models.

D. Performance Analysis

In this section, we describe in detail the performance results obtained by the GNN architecture in the three cities.

1) *Usage of Traffic Data as Input*: Table IV shows the obtained metrics for Madrid. In order to evaluate the suitability

of using road-traffic data as a exogenous variable of the GNN, the table shows the results when the GNN takes the form of a univariate predictor that only considers air pollution $\mathcal{T}_h^S$ as input, and then the results when the GNN is fed with both air pollution $\mathcal{T}_h^S$ and road-traffic $\mathcal{T}_h^R$ time series (see Section III-E).

From this table, we can see that the usage of traffic data as input of the GNN significantly reduced the prediction error for most of the pollutants when $\Delta = 24$ h. For example, the predictor reduced its MSE from 90.000 to 77.016 for PM_{10} and its MAE from 3,605 to 3,376 for PM_{25} . Given $\Delta = 12$ h, the prediction improvement is observed for pollutants CO, O_3 and SO_2 but not for PM_{10} and PM_{25} .

Table IV also shows the metrics obtained in Bilbao. In this case, the largest accuracy improvements when we injected traffic data occurred when the time horizon Δ was set to 12 hours, achieving an error decrease in all the the pollutants and metrics. As a matter of fact, the MAE of O_3 and $PM_{2.5}$ was reduced from 22.524 to 17.983 and from 6.058 to 5.752, respectively, when we included $\mathcal{T}^{R,BIL}$ as input of the GNN. Such an improvement was also observed when $\Delta = 24$ h for particles SO_2 and $PM_{2.5}$.

Table IV also presents the metrics obtained in Cartagena. In this case, the greatest improvement when including $\mathcal{T}^{R,CTG}$ occurs for the case where $\Delta = 12$ h, in which all the error metrics are significantly reduced, except for the CO case. For the case of $\Delta = 24$ h, the error metrics are less conclusive when including $\mathcal{T}^{R,CTG}$, improving in some cases, such as SO_2 , and worsening in others, such as PM_{10} particles.

All in all, we can see two different patterns across the three cities with respect to the benefit of using traffic data in the prediction pipeline. On the one hand, the forecasting of SO_2 and $PM_{2.5}$ values were improved in the three use cases for almost all the time horizons under consideration. The rationale of this multi-dimensional improvement is that road traffic is one

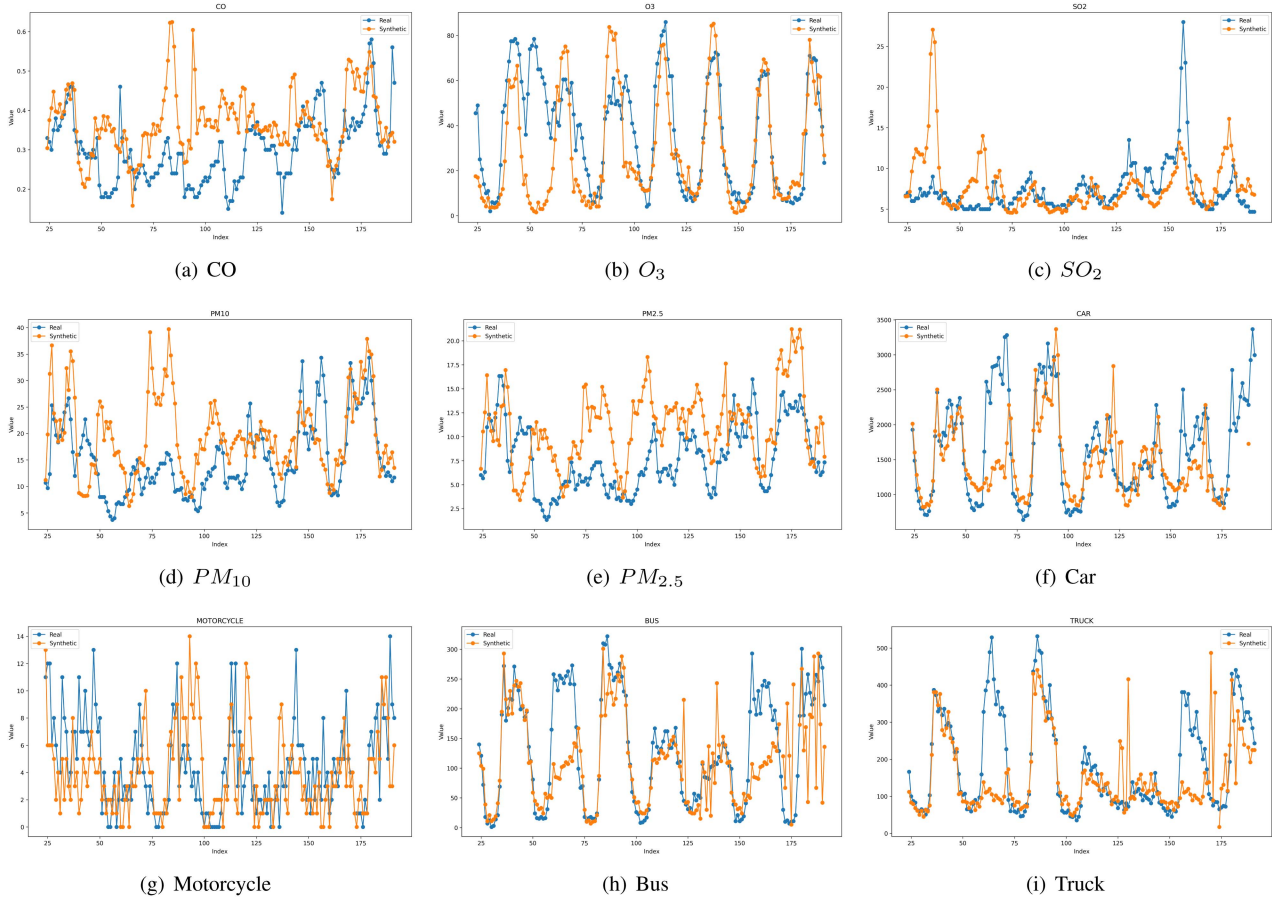


Fig. 6. Time series plots for each studied variable in Bilbao, including both real (blue) and synthetic (orange) data, over a one-week time window.

TABLE IV
EVALUATION METRICS FOR MADRID, BILBAO, AND CARTAGENA WITH REAL TIME SERIES

Δ (h)	Pollutant	Traffic	Madrid			Bilbao			Cartagena		
			MAE	MSE	MAPE	MAE	MSE	MAPE	MAE	MSE	MAPE
12	CO	No	0.187	0.071	73.653	0.135	0.029	27.542	51.964	5.347	49.380
		Yes	0.171	0.069	60.016	0.122	0.023	25.105	52.624	5.486	50.570
	O3	No	21.033	617.888	215.885	22.524	581.190	176.263	12.006	609.074	40.369
		Yes	20.190	573.907	207.973	17.983	362.640	139.997	11.729	590.235	40.184
	PM10	No	11.639	345.247	29.364	11.508	217.083	36.122	22.651	700.855	212.881
		Yes	11.998	356.882	30.451	11.119	200.390	35.021	22.500	698.634	212.379
	PM25	No	4.884	43.007	34.658	6.058	52.283	36.090	9.088	115.539	218.939
		Yes	5.025	45.100	35.434	5.752	47.104	33.988	8.980	113.556	217.147
	SO2	No	0.570	0.428	50.796	2.407	14.507	18.844	5.995	159.763	94.885
		Yes	0.554	0.462	48.176	2.192	12.057	17.762	5.739	145.351	89.871
	CO	No	0.112	0.021	47.496	0.094	0.012	19.966	26.789	1.504	21.897
		Yes	0.110	0.019	48.016	0.134	0.023	29.841	27.008	1.515	20.853
24	O3	No	7.948	123.471	80.506	11.045	213.604	23.408	11.197	638.143	31.065
		Yes	7.840	111.061	73.788	11.220	214.336	24.401	10.969	626.450	31.069
	PM10	No	6.509	90.000	20.597	4.730	44.138	19.679	19.661	506.990	199.276
		Yes	5.979	77.016	18.355	4.796	44.158	20.021	20.021	524.264	201.789
	PM25	No	3.605	17.472	97.627	7.248	73.386	39.389	5.557	43.190	153.083
		Yes	3.376	15.376	90.490	7.230	73.327	39.190	5.540	43.196	153.517
	SO2	No	0.615	0.525	56.138	0.980	1.477	16.848	2.924	30.106	32.584
		Yes	0.628	0.570	56.801	0.952	1.414	16.464	2.597	22.635	30.168

The traffic column indicates whether the traffic flow has been used as input of the GNN. The best metric value for each (pollutant, Δ) tuple is depicted in bold.

of the most important sources of these two pollutants in urban regions [22]. On the other hand, regarding the Δ values for which the usage of traffic data provides a general improvement across particles varies depending on the city, whilst $\Delta = 12$ h is the best configuration in Cartagena and Bilbao, $\Delta = 24$ h is the best configuration in Madrid. This might be due to the nature of the

urban regions in each city. In Madrid, the sensors are installed in a *Low Emission Zone* (LEZ), which significantly restricts access and circulation for most polluting vehicles through a complex access control system.² However, neither Bilbao nor Cartagena

²<https://tinyurl.com/4h8zzz64>

TABLE V
EVALUATION METRICS FOR MADRID, BILBAO, AND CARTAGENA WITH REAL AND SYNTHETIC TIME SERIES, $\langle \mathcal{T}_{synth}^S \cup \mathcal{T}^S, \mathcal{T}_{synth}^R \cup \mathcal{T}^R \rangle$

Δ (h)	pollutant	Traffic	Madrid			Bilbao			Cartagena		
			MAE	MSE	MAPE	MAE	MSE	MAPE	MAE	MSE	MAPE
12	CO	No	0.133	0.065	63.908	0.099	0.018	37.940	44.778	5,284	83.805
		Yes	0.131	0.064	63.934	0.102	0.018	38.081	46.104	5,439	78.666
	O ₃	No	29.273	1,201.805	66.732	23.867	787.826	92.751	17.026	1,093	2,227
		Yes	29.384	1,206.623	66.862	23.645	766.287	91.474	16.509	1,071	2,328
	PM ₁₀	No	8.262	171.062	54.177	5.276	49.287	41.638	17.919	656.093	86.520
		Yes	8.304	170.975	54.220	5.238	48.694	41.441	20.070	791.319	95.619
	PM ₂₅	No	4.256	38.913	62.110	2.851	17.751	34.477	7.348	121.175	94.480
		Yes	4.205	37.274	60.980	2.815	17.608	34.432	8.222	141.140	106.245
	SO ₂	No	0.785	1.383	89.562	2.447	35.505	46.511	7.620	276.484	208.644
		Yes	0.750	1.189	83.989	2.395	34.954	46.118	7.260	256.184	163.342
24	CO	No	0.099	0.040	52.801	0.098	0.017	37.867	35.532	3,867	35.660
		Yes	0.101	0.039	52.064	0.099	0.018	37.929	33.501	3,625	33.943
	O ₃	No	13.707	293.541	33.027	20.676	652.684	83.776	8.899	366.510	1,095.757
		Yes	13.282	279.773	32.271	20.674	653.144	83.791	8.141	347.465	940.491
	PM ₁₀	No	8.289	140.065	51.869	4.732	41.491	36.990	17.913	671.148	79.080
		Yes	7.888	117.490	48.005	4.771	42.142	37.256	17.882	667.614	80.409
	PM ₂₅	No	4.087	30.384	56.547	3.062	19.730	37.361	6.807	92.870	84.468
		Yes	4.001	27.493	53.884	3.059	19.491	37.122	6.844	93.184	85.516
	SO ₂	No	0.676	0.891	75.085	1.836	15.704	41.872	3.860	81.920	137.655
		Yes	0.692	0.963	77.745	1.801	15.292	41.336	2.979	46.388	87.893

The traffic column indicates whether the traffic flow has been used as input of the GNN. The best metric value for each $\langle \text{pollutant}, \Delta \rangle$ tuple is depicted in bold.

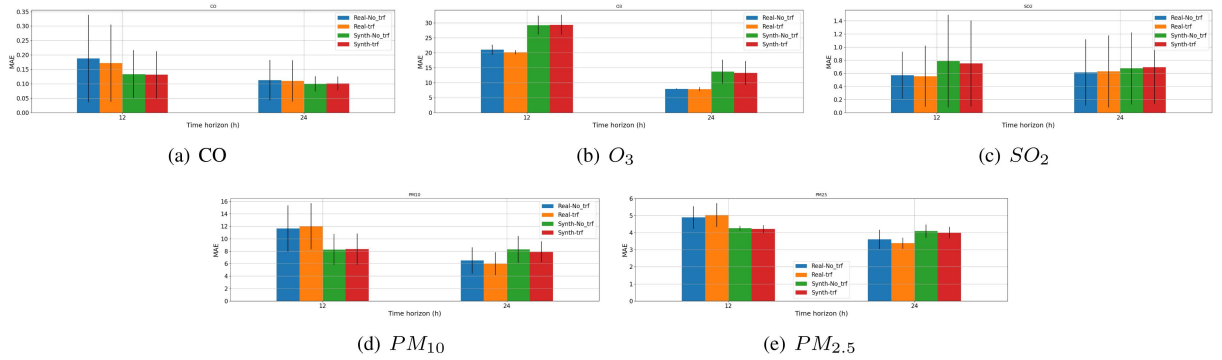


Fig. 7. MAE for Madrid considering the real vs synthetic data and the usage or not of traffic data as primary input.

do have such restricted LEZs, hence combustion vehicles can circulate more freely than in Madrid. This makes traffic a more relevant source of pollutants in these two cities than in Madrid. As a side effect, this allows the $\mathcal{T}^R$ of these cities to be more directly correlated to the air pollution and, thus, provide better short-term predictions.

2) *Usage of Synthetic Data for Model Training:* In order to test the suitability of integrating the synthetic datasets generated for each city, we concatenated $\mathcal{T}_{synth}^S$ and $\mathcal{T}^S$ series to compose a single time series. Due to the 80-10-10 split of the dataset (see Section IV-B), the GNN is mainly trained with the synthetic part of the dataset and evaluated with the real data section.

Table V shows the metrics obtained with such time series composition. We can see that for the two patterns observed when only real data still remains (Section IV-D1), the usage of traffic data improves the prediction accuracy for the SO_2 and the $PM_{2.5}$ values for almost all the Δ and metric values. Note that the largest improvement occurs when $\Delta = 24$ h in Madrid and when $\Delta = 12$ h in Bilbao and Cartagena. This might reveal that the usage of synthetic data did not meaningfully varied the impact of using traffic data on the global performance of the predictive solution.

In order to evaluate whether the usage of synthetic data actually reduce or not the prediction error of the GNN, Fig. 7 shows the MAE obtained in Madrid for the target pollutants. In this city, it was observed that the usage of synthetic data reduced the prediction error for CO across both target Δ , regardless of whether traffic data was used as input. As a matter of fact, the blue and green bars in Fig. 7(a) shows that the usage of synthetic data when we use the univariate version of the GNN (feed with $\mathcal{T}^S$) reduces its MAE from 0.187 to 0.133. Likewise, error reduction is observed when we use the GNN with traffic data from 0.112 to 0.099, according to the orange and red bars in Fig. 7(a). Larger reduction is observed for PM_{10} in Fig. 7(d), when $\Delta = 12$ h, computing 35.67% improvement (from 12.861 to 8.274), with no traffic data, and 35.52% improvement (from 12.997 to 8.381) when traffic data is used in the GNN. Lastly, Fig. 7(e) shows that the MAE of $PM_{2.5}$ also decreases when $\Delta = 12$ h, improving 16.32% (from 5.025 to 4.205) when traffic data is used (orange and red bars), and 12.86% (from 4.884 to 4.256) when no traffic data is injected to the GNN (blue and green bars in the plot).

Fig. 8 shows the effect of synthetic data in Bilbao. Similarly to Madrid, we observe a meaningful reduction in the prediction

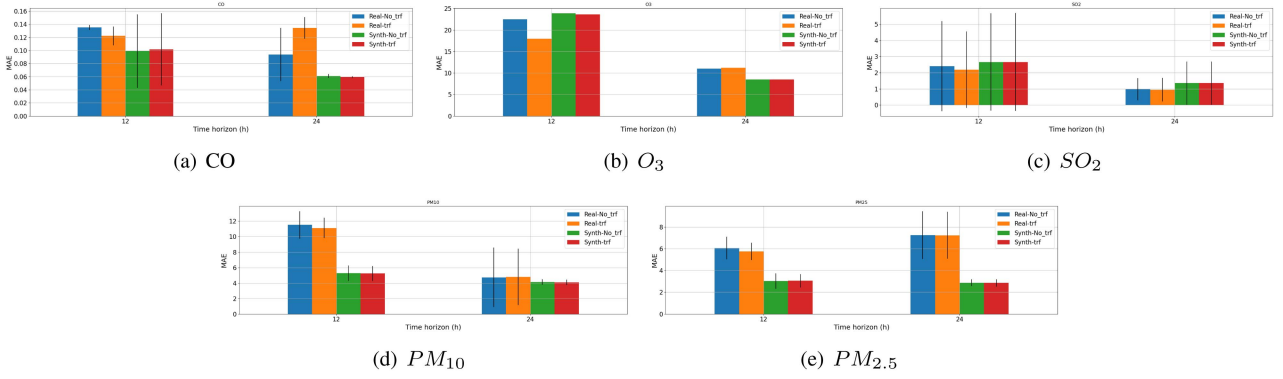


Fig. 8. MAE for Bilbao considering the real vs synthetic data and the usage or not of traffic data as primary input.

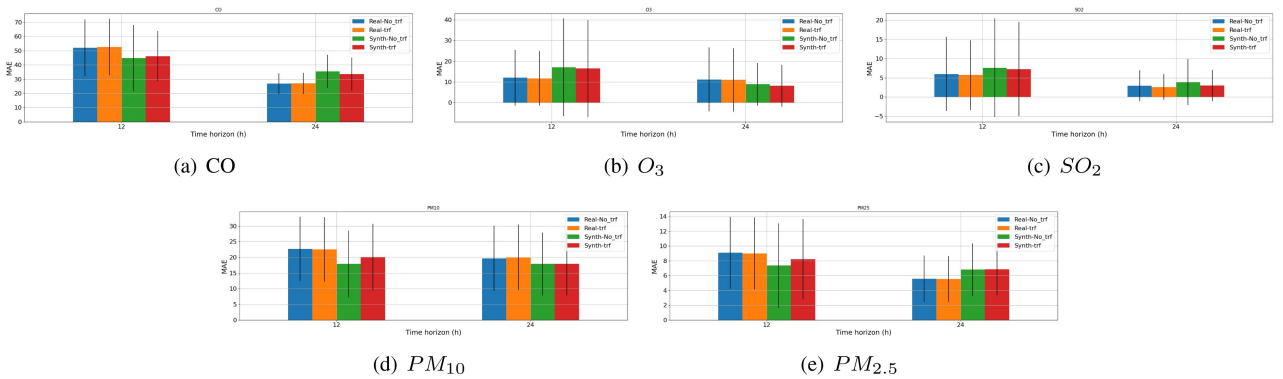


Fig. 9. MAE for Cartagena considering the real vs synthetic data and the usage or not of traffic data as primary input.

error for CO, PM_{10} and $PM_{2.5}$ for the two target Δ . For instance, the PM_{10} MAE is reduced 54.16% (from 11.508 to 5.276) when $\Delta = 12$ h and using traffic data, according to the orange and red bars in Fig. 8(d). Moreover, the MAE of $PM_{2.5}$ is reduced using traffic data 45.31% (from 5.752 to 3.046) when $\Delta = 24$ h, according to the orange and red bars in Fig. 8(e).

Regarding Cartagena, Fig. 9 shows the impact when synthetic data is used. In this scenario, the results are somewhat more diverse. There is a reduction in error when using synthetic data at $\Delta = 12$ h for CO, PM_{10} and $PM_{2.5}$, both for the case of either using traffic data or not. The greatest reduction is for the case of PM_{10} , where a reduction is observed in Fig. 9(d) from 22,651 to 17,919 without traffic data. It is also observed that the reduction is more evident when traffic data are not used. For the $\Delta = 24$ h case, only a slight reduction in MAE is observed for O_3 and PM_{10} pollutants.

E. Comparison With Baseline Models

To ensure that a GNN is the most suitable neural network model for the type of data analyzed, a prediction has been made with the same data, including traffic and synthetic data, using a CNLSTM and MLP model, obtaining the mean errors (between all pollutants on each city) that can be seen in Table VI. Comparing these results with those obtained with the GNN used for this study, it can be seen that in most cases the results improve greatly in the case of the GNN.

TABLE VI

COMPARISON BETWEEN AVERAGE MSE OF POLLUTANTS FOR EACH CITY AND THE DIFFERENT NEURAL NETWORK MODELS, RESULTS CORRESPOND TO PREDICTIONS WITH SYNTHETIC + REAL DATA WITH TRAFFIC

Δ (h)	City	Model	MSE (mean)	MSE (σ)
12	Bilbao	CNNLSTM	417.550	729.033
		MLP	2481.287	5106.817
		GNN	173.512	331.875
	Madrid	CNNLSTM	612.695	1323.555
		MLP	1078.069	2279.092
		GNN	283.225	520.931
	Cartagena	CNNLSTM	10762.526	23976.051
		MLP	7402.680	16354.823
		GNN	1539.7286	2212.898
24	Bilbao	CNNLSTM	309.269	485.976
		MLP	1893.051	3899.074
		GNN	146.017	146.017
	Madrid	CNNLSTM	440.798	907.189
		MLP	1265.820	2708.317
		GNN	85.151	118.932
	Cartagena	CNNLSTM	7977.345	17740.541
		MLP	6813.230	15036.133
		GNN	955.930	1512.356

The bold values indicate the lowest MSE per city and time horizon.

F. Lessons Learnt

Given these results, we can draw some interesting remarks. First, it is worth highlighting the improvement that the usage of synthetic data has provided in the anticipation of certain pollutants, with a reduction above 50% of the error in certain

cases. This opens an interesting research line to use data generation tools to speed up the deployment of accurate air-pollution predictors and avoid the cold-start problem that might occur in urban settings where IoT networks have just been installed.

We have also observed that the improvement provided by traffic or synthetic time series occurs at different time horizons depending on the city. As we have previously stated, this might be due to the traffic restrictions imposed on combustion vehicles in the three cities, which impacts on the number of polluting vehicles in the target areas.

Moreover, the usage of traffic data as exogenous variable and synthetic data improved the prediction accuracy of the GNN models for two particular pollutants, SO_2 and $PM_{2.5}$. Attending to results obtained when synthetic data are used, improvements are evident in CO , PM_{10} and $PM_{2.5}$ pollutants. These particles are strongly related to road traffic patterns, as indicated in [23]. Several studies found that exposure to $PM_{2.5}$ can cause inflammation in the respiratory tract and damage the epithelial barrier, facilitating respiratory infections and exacerbating allergic conditions [24], [25]. Furthermore, prolonged exposure to these pollutants is associated with an increase in hospitalisations for respiratory diseases, especially in children and the elderly [26], [27].

An extensive analysis regarding long-term exposure to traffic-related air pollution, mainly citing our highlighted pollutants in prediction accuracy CO , PM_{10} and $PM_{2.5}$, indicates that health outcomes extend to [23]: birth outcomes, including low weight at term or considering gestational age; diseases in children, including asthma and lower respiratory infections; diseases in adults, receiving the most relevant impact, including ischemic heart diseases, lung cancer, asthma, and even diabetes.

Improving pollution forecasting can help reduce respiratory infections and health risks by enabling proactive measures. Predictions can assist in managing traffic flows and imposing temporary restrictions on industries to prevent dangerous pollution levels. Urban planners and policymakers can use these forecasts to adjust traffic signals, redirect vehicles from pollution hot-spots, or impose vehicle capacity limits in city centres [28], [29]. Additionally, data can trigger low-emission zones, restrict high-polluting vehicles, and optimise public transportation schedules. Long-term insights can guide urban planning, such as road layouts or green spaces, to reduce pollution exposure before development begins [30].

The complexity of Graph Neural Networks (GNNs) offers both benefits and challenges, especially for real-world applications like urban pollution prediction. A key trade-off is the higher computational cost compared to simpler models, as GNNs involve complex operations like graph convolutions and message passing, which are resource-intensive with large urban graphs. Additionally, issues with data availability and quality, such as missing or noisy data, can negatively affect GNN performance since they rely on node connectivity. To address these challenges, strategies like using efficient graph operators (e.g., GraphSAGE) [31], mini-batch training, data imputation, and leveraging external data sources can improve robustness. Periodic retraining or incremental learning can also help maintain GNN performance as urban conditions change.

V. CONCLUSION AND FUTURE WORK

Air pollution has become a critical public health issue, especially in urban areas. Despite advances in IoT systems to monitor

pollution from advanced sensors, and ML-powered schemes to perform predictions, there is still a need for integrating new sources of data to create better informed models, including traffic information.

This study has demonstrated the effectiveness of integrating mobility data with IoT-based air quality monitoring infrastructures to enhance the accuracy of pollution prediction in urban environments. By utilising vehicle traffic data obtained through image recognition and traffic sensors, alongside synthetic data generation to extend limited datasets in some scenarios, our approach significantly improved the prediction accuracy for key pollutants such as SO_2 and $PM_{2.5}$. These pollutants are closely linked to road traffic patterns and have substantial impacts on public health, particularly in causing respiratory infections. The inclusion of Graph Neural Networks (GNNs) allowed for sophisticated modelling of spatial and temporal relationships between air quality and mobility data, highlighting the potential for advanced machine learning techniques to address complex environmental challenges.

The diverse data collection methods, including traffic cameras in Bilbao and Madrid and sensors in Cartagena, create a comprehensive dataset that offers valuable insights into urban traffic patterns, adaptable to similar cities. While cameras may be affected by weather or obstructions, they still capture key traffic trends consistently. In Cartagena, the sensors provide reliable vehicle counts and directional data, although lacking vehicle classification. These varied data sources enhance the model's robustness across different urban environments.

Some limitations have been identified in this work: first, synthetic data significantly improved prediction accuracy, but there is a risk of overfitting, limiting the model's generalisation to new scenarios; second, the variability in prediction improvements between different cities, likely due to differences in traffic patterns and the sensor infrastructure, suggesting that the model may require to carefully select the areas under study, which would reduce the scalability of the proposal; and third, real-world data quality pose a challenge for maintaining the model's robustness, as GNNs rely on the accuracy of connected data points for optimal performance.

Future work will focus on expanding the scope of this research to include a broader range of pollutants and urban environments. This will involve collaborations with local administrations and environmental agencies in several Spanish cities to access more comprehensive data sources during the next 12 months. We also plan to refer to the Spanish Meteorological Agency to integrate additional meteorological data and to Spanish Statistical Agency for including socio-economic factors. Further refinement of synthetic data generation techniques will be pursued in a second stage to ensure robustness and generalisation across different cities and conditions. Additionally, the development of real-time prediction models and decision-support systems for urban planners and policymakers will be a key goal, with pilot implementations scheduled for the next 18-24 months to optimise traffic management and reduce pollution exposure.

CODE AVAILABILITY

The source code of this project is available at https://github.com/fterroso/airpollution_gnn_uca_upct

REFERENCES

- [1] T. Sigsgaard and B. Hoffmann, "Assessing the health burden from air pollution," *Science*, vol. 384, no. 6691, pp. 33–34, 2024. [Online]. Available: <https://www.science.org/doi/abs/10.1126/science.abo3801>
- [2] S. Karkuzhali and S. Senthilkumar, *Role of Artificial Intelligence and IoT in Environmental Monitoring—A Survey*. Singapore: Springer Nature Singapore, 2022, pp. 141–155, doi: [10.1007/978-981-19-2083-7_8](https://doi.org/10.1007/978-981-19-2083-7_8).
- [3] Y.-J. Lu and C.-T. Li, "Forecasting urban sensory values through learning attention-adjusted graph spatio-temporal networks," *ACM Trans. Spatial Algorithms Syst.*, vol. 10, no. 1, p. 22, Mar. 2024, Art. no. 4, doi: [10.1145/3635140](https://doi.org/10.1145/3635140).
- [4] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.
- [5] G. Guerrero-Ulloa, A. Andrango-Catota, M. Abad-Alay, M. J. Hornos, and C. Rodríguez-Domínguez, "Development and assessment of an indoor air quality control IoT-based system," *Electronics*, vol. 12, no. 3, 2023, Art. no. 608.
- [6] L. Milivojević, S. M. Kurilić, Z. Božilović, S. Koprivica, and O. Krčadinac, "Study of particular air quality and meteorological parameters at a construction site," *Atmosphere*, vol. 14, no. 8, 2023, Art. no. 1267. [Online]. Available: <https://www.mdpi.com/2073-4433/14/8/1267>
- [7] I. M. Al Jawarneh, L. Foschini, and P. Bellavista, "Efficient integration of heterogeneous mobility-pollution Big Data for joint analytics at scale with QoS guarantees," *Future Internet*, vol. 15, no. 8, 2023, Art. no. 263.
- [8] F. Terroso-Sáenz and A. Muñoz, "Nation-wide human mobility prediction based on graph neural networks," *Appl. Intell.*, vol. 52, pp. 4144–4160, 2022.
- [9] Y. Wang, F. Currim, and S. Ram, "Deep learning of spatiotemporal patterns for urban mobility prediction using Big Data," *Inf. Syst. Res.*, vol. 33, no. 2, pp. 579–598, 2022.
- [10] J. Lu, P. Bu, X. Xia, N. Lu, L. Yao, and H. Jiang, "Feasibility of machine learning methods for predicting hospital emergency room visits for respiratory diseases," *Environ. Sci. Pollut. Res.*, vol. 28, pp. 29701–29709, 2021.
- [11] Y. Ku, S. B. Kwon, J.-H. Yoon, S.-K. Mun, and M. Chang, "Machine learning models for predicting the occurrence of respiratory diseases using climatic and air-pollution factors," *Clin. Exp. Otorhinolaryngol.*, vol. 15, no. 2, 2022, Art. no. 168.
- [12] B. Dewulf et al., "Dynamic assessment of exposure to air pollution using mobile phone data," *Int. J. Health Geographics*, vol. 15, pp. 1–14, 2016.
- [13] M. Nyhan, I. Kloog, R. Britter, C. Ratti, and P. Koutrakis, "Quantifying population exposure to air pollution using individual mobility patterns inferred from mobile phone data," *J. Exposure Sci. Environ. Epidemiol.*, vol. 29, no. 2, pp. 238–247, 2019.
- [14] W. Zhou, X. Wu, S. Ding, X. Ji, and W. Pan, "Predictions and mitigation strategies of PM_{2.5} concentration in the Yangtze River Delta of China based on a novel nonlinear seasonal grey model," *Environ. Pollut.*, vol. 276, 2021, Art. no. 116614.
- [15] V. Calatayud, J. J. Diéguez, E. Agathokleous, and P. Sicard, "Machine learning model to predict vehicle electrification impacts on urban air quality and related human health effects," *Environ. Res.*, vol. 228, 2023, Art. no. 115835.
- [16] S. Pandya et al., "Pollution weather prediction system: Smart outdoor pollution monitoring and prediction for healthy breathing and living," *Sensors*, vol. 20, no. 18, 2020, Art. no. 5448.
- [17] J.-B. Renard, J. Surcin, I. Annesi-Maesano, G. Delaunay, E. Poincelet, and G. Dixsaut, "Relation between PM_{2.5} pollution and COVID-19 mortality in western Europe for the 2020–2022 period," *Sci. Total Environ.*, vol. 848, 2022, Art. no. 157579.
- [18] A. Schwabe, J. Persson, and S. Feuerriegel, "Predicting COVID-19 spread from large-scale mobility data," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2021, pp. 3531–3539.
- [19] Z. Lin, A. Jain, C. Wang, G. Fanti, and V. Sekar, "Using GANs for sharing networked time series data: Challenges, initial promise, and open questions," in *Proc. ACM Internet Meas. Conf.*, 2020, pp. 464–483, doi: [10.1145/3419394.3423643](https://doi.org/10.1145/3419394.3423643).
- [20] R. Ignaccolo, S. Ghigo, and S. Bande, "Functional zoning for air quality," *Environ. Ecological Statist.*, vol. 20, pp. 109–127, 2013.
- [21] Y. Seo, M. Defferrard, P. Vandergheynst, and X. Bresson, "Structured sequence modeling with graph convolutional recurrent networks," in *Proc. Neural Inf. Process., 25th Int. Conf.*, 2018, pp. 362–373.
- [22] M. M. M. F. Jion et al., "A critical review and prospect of NO₂ and SO₂ pollution over asia: Hotspots, trends, and sources," *Sci. Total Environ.*, vol. 876, 2023, Art. no. 162851.
- [23] H. Boogaard et al., "Long-term exposure to traffic-related air pollution and selected health outcomes: A systematic review and meta-analysis," *Environ. Int.*, vol. 164, 2022, Art. no. 107262. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S016041202200188X>
- [24] J. Liu et al., "Particulate matter disrupts airway epithelial barrier via oxidative stress to promote pseudomonas aeruginosa infection," *J. Thoracic Dis.*, vol. 11, pp. 2617–2627, 2019.
- [25] Y. Xue, J. Chu, Y. Li, and X. Kong, "The influence of air pollution on respiratory microbiome: A link to respiratory disease," *Toxicol. Lett.*, vol. 334, pp. 14–20, 2020.
- [26] S. Adar, P. Filigrana, N. Clements, and J. Peel, "Ambient coarse particulate matter and human health: A systematic review and meta-analysis," *Curr. Environ. Health Rep.*, vol. 1, no. 3, pp. 258–274, 2014.
- [27] Y. Tian et al., "Short-term effects of ambient fine particulate matter pollution on hospital visits for chronic obstructive pulmonary disease in Beijing, China," *Environ. Health*, vol. 17, no. 1, pp. 1–8, 2018.
- [28] S. Dikshit, A. Atiq, M. Shahid, V. Dwivedi, and A. Thusu, "The use of artificial intelligence to optimize the routing of vehicles and reduce traffic congestion in urban areas," *EAI Endorsed Trans. Energy Web*, vol. 10, Dec. 2023. [Online]. Available: <https://publications.eai.eu/index.php/ew/article/view/4613>
- [29] J. L. Texcalac-Sangrador et al., "Speed limits and their effect on air pollution in Mexico city: A quasi-experimental study," *Sci. Total Environ.*, vol. 924, 2024, Art. no. 171506. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0048969724016474>
- [30] B. Augusto, D. Lopes, S. Rafael, M. Coelho, and J. Ferreira, "Assessing the impact of different urban morphology scenarios on air pollutant emissions distribution," *Sci. Total Environ.*, vol. 950, 2024, Art. no. 175341. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0048969724054913>
- [31] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, Eds., Curran Associates, Inc., 2017, vol. 30.