Short-term prediction of NO_2 concentration on ground-based weather sensors data in urban scale area with graph neural networks

Carlo Milesi[1], Francesco Pisanu[2], Carlino Casari[1]

[1] CRS4 (Center for Advanced Studies, Research and Development in Sardinia), Pula, Italy

[2] LIPN, Université Sorbonne Paris Nord, Villetaneuse, France

Correspondence to: (carlo.milesi@crs4.it)

Highlights

● Urban ground-level analysis based on NO_2 time series from 01/01/2018 to 01/04/2022

∉ Predictive models based on machine learning to forecast hourly NO_2 concentrations

∉ Graph Neural Networks combined with Long Short-Term Memory (GNN-LSTM) to obtain multidimensional predictions for the entire network

∉ Identification of elements that significantly alter the result.

Abstract

Air pollution and its associated pollutants have gained significant scientific interest recently, particularly due to their impact on human health. Urban areas, that are affected by polluting human activities, prioritize monitoring air pollution trends and diffusion for understanding and evaluating projections in the future. Nitrogen dioxide (NO_2) is among the most harmful pollutants to human health due to its reaction with respiratory tract moisture, resulting in damaging nitric acid in the lungs, throat, and nasal passages. This study introduces predictive models based on machine learning to forecast hourly NO_2 concentrations measured by a sensor network in an urban area. Specifically, several graphs are constructed based on different metrics across the sensor network. The models utilize graph neural networks combined with long short-term memory to obtain multidimensional predictions for the entire network. The proposed methods have been tested using recent data on NO_2 concentration measurements obtained from a sensor network located in the coastal metropolitan area of Cagliari, the largest city on the island of Sardinia.

The study highlights an improvement in results using a Graph Neural Network, and this improvement is even more pronounced when adding the correlation component to the multidimensional analysis. Furthermore, it allows you to identify any elements within the graph that significantly alter the result.

Keywords

Nitrogen Dioxide (NO_2), ground-based weather sensors, Neural Networks, GNN-LSTM

1 Introduction

Air pollution is a pressing environmental issue that poses significant risks to human health and the environment.(Neto et al., 2023) Among the various air pollutants, nitrogen dioxide (NO_2) is of particular concern due to its adverse impacts on respiratory health and its role as a precursor to the formation of other harmful air pollutants, such as ground-level ozone and particulate matter. (Landrigan et al., 2018; Rafaj et al., 2018)

With the advancement of technology and the availability of extensive air quality monitoring networks, researchers are increasingly turning to machine learning and deep learning approaches to improve the accuracy and efficiency of air quality analysis (Fan et al., 2017; Seng et al., 2021).

The convergence of machine learning and deep learning with traditional approaches in air quality analysis has enlarged the horizons for predictive modeling. (Lam et al., 2023; Rodeschini et al., 2023; Seyyedi et al., 2023; Xiao et al., 2022)

In recent years, numerous machine learning-based approaches have been developed for predicting air quality events (Méndez et al., 2023). Particularly, there has been a growing interest among researchers in employing Graph Neural Networks (GNNs) for tackling multidimensional problems related to urban traffic (Du et al., 2018; Jiang & Luo, 2021), parking availability (Zhang et al., 2019), and other similar scenarios (Rahmani et al., 2023; Xu et al., 2021).

From the point of view of air quality, there are some works mainly concerning the prediction of PM 2.5 (Chowdhury et al., 2023; Li et al., 2023; Qi et al., 2019). However, applying a Graph Neural Network combined with a Long Short-Term Memory (GNN-LSTM) model to the prediction of NO_2 is a little-explored topic.

In this context, we have developed a predictive model to predict hourly NO_2 concentrations in urban areas, which might be helpful for public health management and decision-making. This work aims to present such a predictive model based on machine learning, specifically utilizing a GNN-LSTM, to achieve multidimensional predictions for the entire sensor network in the metropolitan city of Cagliari.

ARPAS (Agenzia Regionale per la Protezione dell'Ambiente della Sardegna - Sardinia Environment Agency) provides a valuable collection of air quality data for the area of interest, with good spatial coverage and temporal resolution. By deploying these sensors in various locations, including industrial neighborhoods and areas near high-traffic zones, it becomes possible to complement the sparse regulatory network data and improve the overall accuracy of predictive models.

GNNs are widely used to reduce the effect of the irregularities in the samplings (Qasim et al., 2019). However, GNN-LSTM models suffer high irregularity in the sampling, which is common in real-world datasets, and are not commonly used in contexts where the lying graph model is unclear.

To further reduce the effect of irregularities, attention networks have been developed, (Brody et al., n.d.; Liu et al., 2021; Wang et al., 2019) that is GNN structures enriched by adding an attention layer that aims to find good weights on the edges to represent the data structure and offer significant advantages in terms of improved feature relevance, and enhanced information flow. However, their increased complexity and sensitivity to hyperparameters may pose challenges in terms of scalability, parameter tuning, and interpretability. Instead, we aim to use preprocessing methods that have less impact, on execution time, on the network's structure.

Moreover, we show that these methods generally outperform univariate LSTM models for our problem, which we use as a reference for state-of-the-art comparisons.

To assess the effectiveness of the proposed GNN-LSTM model, this study compares the predictive results obtained from the model with predictions made on individual sensors. Subsequently, in the second phase, the model's multidimensional prediction is recalculated, incorporating correlation matrices and missing data percentage vectors between the sensors, which provides a way to evaluate the impact of these factors on model performance. By considering the correlation between sensors and addressing missing data issues, the model seeks to provide more robust and accurate predictions of NO_2 concentrations.

Modeling NO_2 concentrations in Cagliari presents challenges due to its geographical characteristics (Liang et al., 2023). Indeed, it is situated in a coastal location with a hilly hinterland and a large marshy area. These elements contribute to complex air circulation patterns and irregularities in the dispersion of pollutants. In addition, industrial activities, particularly the Sarroch industrial area which hosts a big oil refinery and is situated in the west of the city center, further complicate the modeling problem for the NO_2 concentrations (Kumar et al., 2017).

The presence of all these factors requires careful consideration of the complexity of the air prediction model both from the point of view of the final accuracy and the computational load at the cost of making the model hard for interpretation. Thus, by applying interpretable filters to different graph structures, our models aim to clarify certain patterns and allow us to identify a simpler network. Furthermore, this approach is more effective in the description of the NO_2 trend in the study area when compared to several state-of-the-art machine learning techniques.

2 Material and methods

This section is divided into two parts. The first part describes the study area from a topographical and urban perspective. The second part analyses the origin and geographical location of the data sources used and the method by which the data are collected.

2.1 Study region

The geographic area examined in this paper is the metropolitan area of Cagliari, located in southern Sardinia (Italy); the same area was analyzed in depth in the article (De Santis et al., 2023) by the same authors in which an analysis of the temporal variations of NO_2 in urban areas was carried out through satellite data and ground sensor data, the correlation of NO_2 with vehicular traffic dynamics during the COVID-19 pandemic was evaluated.

The metropolitan area of Cagliari lies between 8,820-9,486 longitude and 38,915-39,423 latitude (EPSG: 4326), has an area of about 1250 km^2 and has a population of about 420000 inhabitants, who reside in 17 different cities. Cagliari, the regional capital of Sardinia, represents the city with the highest density.

The climate in this region is mild during winter and hot and dry during summer. The topography is complex and includes a large coastal area, rural areas, and heavily populated areas. In the western part, precisely in the locality of Sarroch - Porto Foxi, is the SARAS refinery, one of the most essential petrochemical sites nationally and in the Mediterranean, which covers a vast area characterized by oil docks with berths for seventeen ships.

The port of Cagliari is an important hub for transshipment activities in the western Mediterranean and, in recent years, has seen a steady growth in tourist activity from a cruise perspective.

2.2 Data sources

We considered hourly measurements of NO_2 concentration in µg/m^3, provided by eight air quality monitoring stations made available by ARPAS, located in five cities within the metropolitan area of Cagliari: Cagliari, Quartu Sant'Elena, Monserrato, Assemini, Sarroch. Here (https://portal.sardegnasira.it/mappa-stazioni-misura, https://www.regione.sardegna.it/documenti/1_73_20171115100918.pdf, https://www.sardegnaambiente.it/documenti/21_393_20180802104517.pdf) one can see the details of all the control units. They are identified by the EoI (Exchange of Information, EU Member States Decision 97/101/EC) code and station type. In Table 1 we report the position and the station types corresponding to distinct locations. Figure 1 shows their position (red markers) on a base map that includes the study area.

In situ Multi-parameter control unit	EoI code	Coordinates (Lon; Lat)	Altitude (over sea level)	Inlet height	Station type
CENCA1	IT2056A	9.115; 39.2358	33 m	3 m	Traffic
CENQU1	IT2040A	9.1881; 39.2328	8 m	3 m	Background
CENMO1	IT1993A	9.1364; 39.2606	15 m	3 m	Background
CENAS6	IT1278A	8.9833; 39.2373	14 m	3 m	Background
CENAS8	IT1397A	8.9948; 39.2260	9 m	3 m	Industrial
CENAS9	IT2049A	9.0114; 39.2864	5 m	3 m	Background
CENSA2	IT1269A	9.0163; 39.0691	22 m	3 m	Industrial
CENSA3	IT1947A	9.0089; 39.0667	56 m	3 m	Background

Table 1: Air quality ground measurement characteristics: IDs, Geolocation (EPSG: 4326), Altitude, Inlet height, and Station type.

All the control units considered measure surface NO_2, the detection limit for NO_2 is up to $0.5\mu g/m^2$, a reference accuracy value for surface NO_2 measurements is within ± 15% (Kelly et al., 1990).

The data is made available under a CC BY-NC-ND license. ARPAS reserves the right to update the data within 60 days of initial publication.

The dataset used for analysis comprises 293,591 records from all 8 stations and covers the period from 01/01/2018 to 01/04/2022.

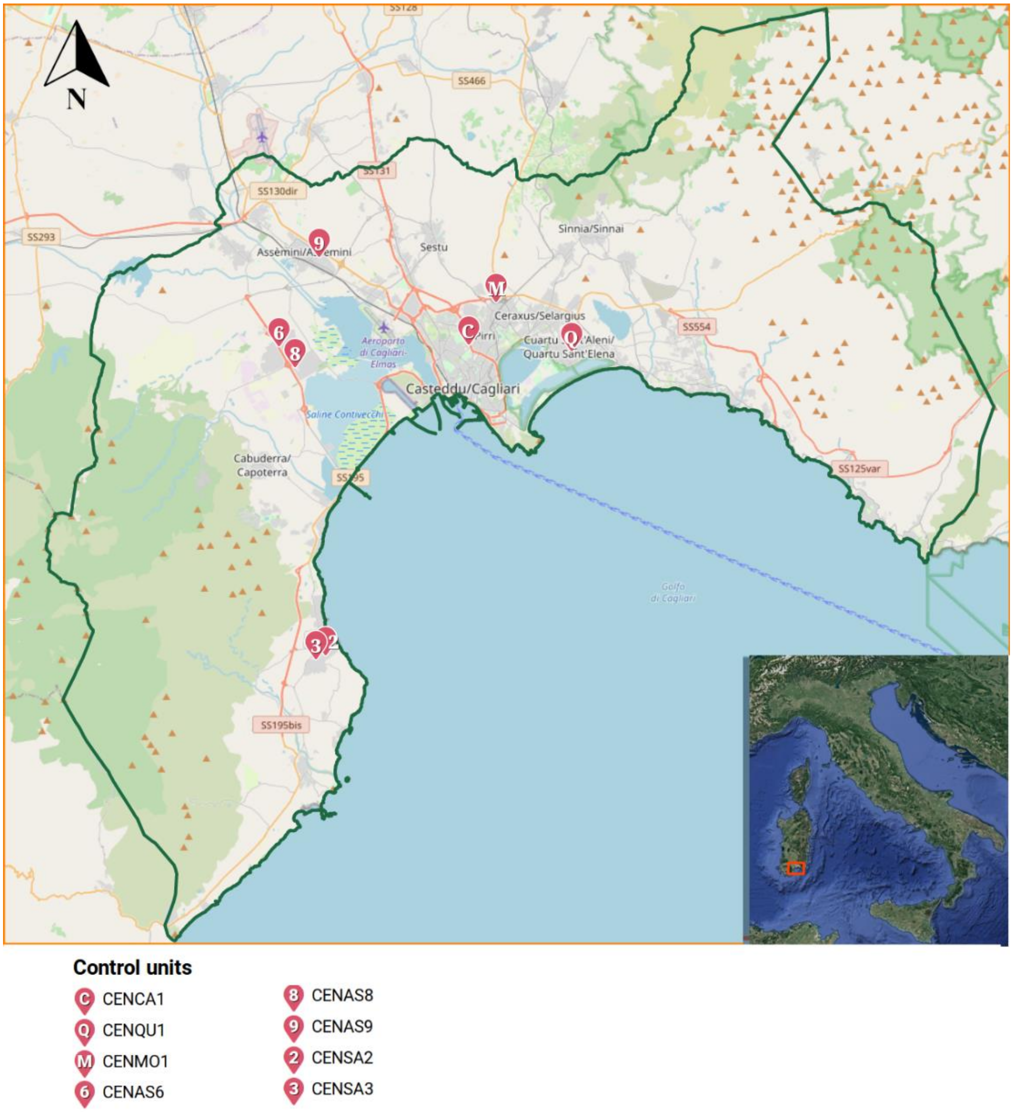

Figure 1: The study area with the control units indicated. Bounding box (EPSG:4326, format: [min lon, min lat, max lon, max lat]): [8.82020751573493, 38.9154851198457, 9.48642514779543, 39.4233721770646] (Background map © OpenStreetMap contributors https://www.openstreetmap.org/copyright/en).

2.3 Methodology

This section is divided into three parts. The first part describes the preprocessing techniques used to enhance the consistency and robustness of the dataset. The second part explains the approach and the steps taken to achieve the result. Finally, the third part outlines the implementation of algorithms used, along with their parameters, and the tools utilized.

2.3.1 Data manipulation

The control unit data may be incomplete due to network malfunctions or sensor issues. There are two scenarios: either the data is present but invalid and marked as ‘NULL’ or the data is absent; the percentage of missing data given by the sum of the two cases varies from a maximum of 9.07% for CENSA3 to a minimum of 6.68% for CENQU1. In Table 2 we see an analysis of the completeness of the dataset used.

We use a linear interpolation algorithm to complete the datasets with missing data. The input data, then are scaled through a standardization algorithm. We used the following standardization formula:

$$X_0 = \frac{x - \mu}{\sigma} \qquad (1)$$

where μ is the mean, and σ is the standard deviation. At the end of the process, the result is rescaled to the original values.

Control Unit	Tot Hour	Saved hour	Hour missed	Total hour missed	Pct. Missed (%)
CENCA1	37224	36646	2235	2813	7.55
CENQU1	37224	36716	1980	2488	6.68
CENMO1	37224	36683	2005	2546	6.83
CENAS6	37224	36726	2105	2603	6.99
CENAS8	37224	36731	2840	3333	8.95
CENAS9	37224	36628	2638	3234	8.68

CENSA2	37224	36646	1919	2497	6.70
CENSA3	37224	36815	2968	3377	9.07

Table 2: The table shows for each control unit the completeness of the data in the period between 2018-01-01 00:00 - 2022-04-01 00:00

Tot hour: indicates the maximum number of hours of operation in the span of hours examined

Saved hour indicates the number of hours in which the monitoring unit returned a value

Missed hour: hours where the control unit returned a value that was not useful

Total hour missed: the sum of the previous value plus the hours where the control unit did not return a value

Pct. missed: percentage of hours where the control unit did not return value (Hour real missed) compared to the total (tot hour)

2.3.2 The modeling approaches

The objective of the study is to predict the concentration of NO_2 of each station at a given time by considering a graph *G=(V, E)* where *V* is the set of nodes and *E* is the set of edges. We encode the graph's structure with the associated adjacency matrix, that is a 0,1-matrix of size $|V|\times|V|$ and whose *(i,j)* entry equals 1 if and only if $i=j$ or the edge v_iv_j belongs to *E*. The graph considered in the study is an undirected unweighted graph, and the learning task is a regression of nodes, where the goal is to forecast future values starting from historical time series data, considering both temporal dependencies and graph structure (J. L. Gross and J. Yellen, 2003).

To create the adjacency matrix, we measured the distance between each control unit using EPSG 3003, which is the most accurate geographic reference system for the region under consideration. The greatest distance recorded was the one between CENAS9 and CENSA3, which amounted to 24.3 km. The control units closest to each other were CENSA2 and CENSA3, with a distance of 700 m.

Control Unit	CENCA1	CENQU1	CENMO1	CENAS6	CENAS8	CENAS9	CENSA2	CENSA3
CENCA1	0.0	6.3	3.1	11.5	10.6	10.5	20.7	21.2

CENQU1	6.3	0.0	5.4	17.8	16.8	16.4	23.7	24.3
CENMO1	3.1	5.4	0.0	13.5	12.8	11.1	23.7	24.2
CENAS6	11.5	17.8	13.5	0.0	1.6	5.9	18.9	19.1
CENAS8	10.6	16.8	12.8	1.6	0.0	6.8	17.5	17.7
CENAS9	10.5	16.4	11.1	5.9	6.8	0.0	24.1	24.3
CENSA2	20.7	23.7	23.7	18.9	17.5	24.1	0.0	0.7
CENSA3	21.2	24.3	24.2	19.1	17.7	24.3	0.7	0.0

table 3: The distances matrix shows the mutual distance between each control unit, where the distance is expressed in km

We consider three methods to outperform the univariate model. One uses the distance-based adjacency matrix, another adds a filter on the edges through the correlation matrix, and a third one adds a filter on the edges setting a threshold for the missing value percentages depending on each node.

To construct the correlation matrix (Table 4), we calculated the Pearson index for each pair of nodes. In the case of the correlation matrix, we considered five scenarios, in each of which we excluded edges between pairs of nodes by setting a threshold for the absolute value of the correlation index to be greater than or equal to 0.3, 0.4, 0.5, 0.6, and 0.7.

Control Unit	CENCA1	CENQU1	CENMO1	CENAS6	CENAS8	CENAS9	CENSA2	CENSA3
CENCA1	1.00	0.71	0.72	0.40	0.38	0.66	0.46	0.52
CENQU1	0.71	1.00	0.78	0.29	0.44	0.68	0.54	0.60
CENMO1	0.72	0.78	1.00	0.35	0.46	0.68	0.49	0.55
CENAS6	0.40	0.29	0.35	1.00	0.52	0.31	0.31	0.33
CENAS8	0.38	0.44	0.46	0.52	1.00	0.45	0.44	0.43

CENAS9	0.66	0.68	0.68	0.31	0.45	1.00	0.46	0.53
CENSA2	0.46	0.54	0.49	0.31	0.44	0.46	1.00	0.71
CENSA3	0.52	0.60	0.55	0.33	0.43	0.53	0.71	1.00

Table 4: The correlation matrix shows Pearson coefficient between each control unit.

Regarding the missing value percentage-based filter, once determined the percentage of missing data for each node, we chose three scenarios by adding this value of -0.5, 0, and +0.5. This choice sets a unique threshold depending on the missing value percentage of each station. That is, we filter the neighbours of each node of the graph by using the following formula:

$$X_i + \gamma \leq X, \qquad (2)$$

where X_i is the percentage of the missing data of the node to evaluate, γ is an arbitrary constant, and X is the node considered. By definition, this filter might destroy the symmetry of the adjacency matrix. Indeed, while if a node is adjacent to a second one, the latter might be not adjacent to the first. Thus, the corresponding matrix is not an adjacency matrix of a graph. More properly, the corresponding structure is a multilayer graph, and each row of the matrix encodes the local graph structure depending on the corresponding node.

We define several distance-based adjacency matrices, by increasing a radius ε centred to each node from 0.1 km to 25.1 km with a step of at least 1 km (for instance, the graphs corresponding to radii from 7.1 to 10.1 are equal to each other, and so are the associated models, hence we kept only the first one). For example, by setting ε at of 0.1 km the corresponding adjacency matrix is the identity matrix, while for ε set at 25.1 km, we produce a complete graph.

$$\varepsilon = [0.1, 1.1, \ldots\ldots\ldots\ldots\ldots.., 24.1, 25.1] \quad (3)$$

To conclude, we used a single-step GNN-LSTM (Hochreiter & Schmidhuber, 1997) model to forecast NO_2 levels for each control unit, predicting a single future value based on past data. This will be done for each scenario and each ε.

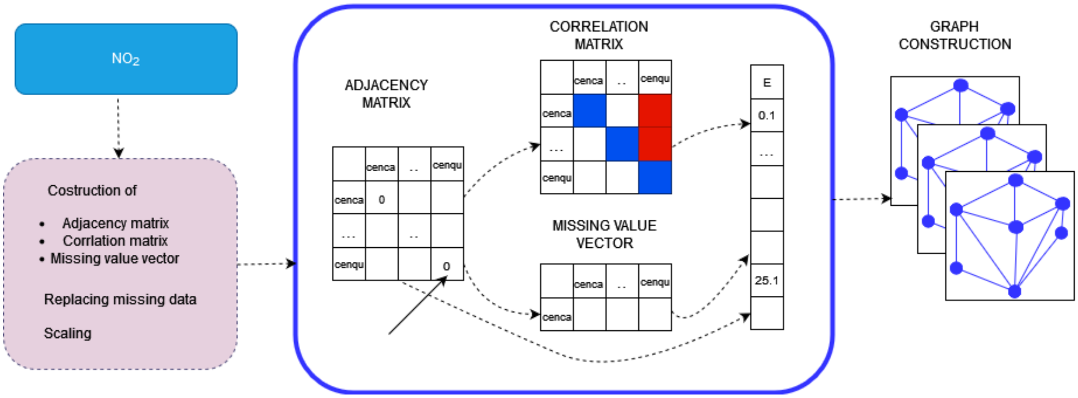

Figure 2: Diagram describing the process of data preprocessing and graph generation.

2.3.3 Model implementation and execution

The machine learning model used in this work was built and managed using Keras a framework based on TensorFlow, (https://www.tensorflow.org) a Python-based library with comprehensive tools.

We implemented a GNN-LSTM with 300 neurons in the first and second hidden layers, and 8 neurons in the output layer, a dropout with a probability of 0.2 is used in all fully connected layers. The input shape is 1-time step considering a time window of 25 values and a batch size of 48. We used the Rectified Linear Unit (ReLU) activation function. Furthermore, we utilized the Adam optimizer because it has shown good generality in deep learning models and faster convergence ability. The model is trained for a maximum of 200 epochs with an Early Stopping callback fixed at 5, which aims to prevent overfitting of the model (Inadagbo et al., 2023). To split the dataset, we allocated 50% of the data for training, 20% for validation, and 30% for testing. We trained each model ten times and chose the best one according to the validation result.

The models are trained by optimizing the Mean Square Error (MSE), which quantifies the average of squared errors between predicted and true hourly NO_2 values. The following is the corresponding mathematical formula:

$$MSE = \frac{1}{N}\sum_{i=0}^{n}(Y_i - Y_i')^2, \qquad (4)$$

where N is the number of samples, Y_i is the observed value, Y_i' is the corresponding predicted value. The MSE metric assigns more weight to larger errors, making it the preferred choice for our application (Casella & Berger, 2002). This characteristic is in line with our specific context, where we aim to identify substantial increases in NO_2 levels and, as such, we prioritize the modeling of true values that significantly deviate from zero. The same metric was used to test the model.

All simulations were performed using the CRS4 computing center, taking advantage of a Dual Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz with 32 cores and 768 GB RAM, equipped with three 40GB NVIDIA A100 Tensor Core video cards.

3 Results and discussion

In this section, we present the results and analysis of the trained models. Table 5 displays the results rounded to the nearest hundredth of all the models we studied. The column labels "Corr" and "Miss" indicate that the correlation index filter and the missing value percentage filter have been applied to the models. Each column corresponds to a unique scenario for some fixed threshold and the rows correspond to a fixed value of ε. The elements in the cells correspond to the MSE associated with the best-trained model compared to the ten validated ones. In the last three rows, the cells contain as indicators the mean, median, and standard deviation of each scenario. For certain scenarios, the corresponding filter implies that the model is invariant to some values of ε. Therefore, we refrain from replicating the model to avoid creating imbalances in the indicators.

For every scenario with ε set at 0.1, all the proposed models are equivalent to the univariate one. Indeed, the distance of each couple of nodes is greater than 100m. So, no filter should be applied to the adjacency matrix. However, while the application of a correlation filter or a nonpositive missing value percentage threshold does not alter the model (formula 2), the introduction of a missing value percentage threshold augmented of +0.5

implies that each node does not have an adequate level of accuracy for its representation. Consequently, the corresponding filtered result is a null matrix, which does not encode the graph's structure. Therefore, we assume that, for ε set at 0.1, this model is also equivalent to the univariate one.

The integration of a GNN-LSTM into our problem leads to a reduction of the MSE across all nodes in most of the cases. As delineated in Table 5, the GNN-LSTM architecture consistently produces better predictions in fourteen out of eighteen models. The same consistency is not generally attained through the other methods here studied. Nonetheless, the best outcome in each column performed better than the univariate case.

Some model results exhibited anomalies (see, for instance, ε set at 2.1. in the GNN column). We supposed that it was due to the presence of sensor CENAS9. Indeed, this control unit adversely impacted the graph structure, as the MSE for its predictions exceeded the values observed for other control units by more than fivefold on several occasions. Extensive analyses were conducted on the signals recorded, yet no aberrant patterns or anomalies were discernible (see supplementary materials). Consequently, we decided to remove CENAS9 to verify the validity of our assumption.

Table 6 illustrates the outcomes after the remotion of CENAS9 from the network, demonstrating the elimination of all outlier values. An improvement in the average performance across all scenarios is also highlighted.

When considering all stations, the MSE of the best GNN-LSTM model, whose corresponding adjacency matrix is distance-based, equals 23.96, while applying a correlation-based or a missing value-based filter, the best MSEs are 23.23 and 24.35, respectively. Comparatively, the univariate approach yielded an MSE value of 24.99, while all the other scenarios outperformed the univariate model, achieving 23.23 as the best performance. Moreover, the medians of seven scenarios out of nine are lower than the univariate one.

Once removed CENAS9, the univariate approach yielded an MSE value of 21.5, while all the other methods and thresholds outperformed the univariate model, achieving 20.39 as the best performance in the case of a correlation index filter whose threshold is fixed to 0.4. Moreover, both mean and median, for each scenario, are smaller than the best value obtained from the univariate model. Finally, also the standard deviation of each scenario is relatively small.

ε	**Univariate**	**GNN**	**Corr 0.3**	**Corr 0.4**	**Corr 0.5**	**Corr 0.6**	**Corr 0.7**	**Miss -0.5**	**Miss 0**	**Miss +0.5**
0.1	**24,99**	24,99	24,99	24,99	24,99	24,99	24,99	24,99	24,99	24,99
1.1		26,62	26,62	26,62	26,62	26,62	26,62	24,76	24,76	24,97
2.1		47,22	47,22	47,22	47,22	24,82	24,82	27,6	27,6	25,82
3.1		25,79	25,79	25,79	25,79	26,23	26,23	25,3	25,3	25,65
6.1		24,57	24,57	35,86	35,86	24,76	**24,76**	37,87	24,44	25,84
7.1		24,81	24,81	25,64	25,06	25,41	25,41	25,1	24,71	25,11
11.1		24,65	24,65	25,33	24,9	24,93	26,75	24,55	24,7	24,68
12.1		24,46	24,46	24,69	—	—	—	**24,47**	24,74	25,51
13.1		24,67	24,67	26,86	—	—	—	25,86	25,45	25,14
14.1		24,18	24,18	—	—	—	—	24,99	26,12	24,73
17.1		24,57	24,57	24,54	24,26	24,85	—	24,48	26,99	24,83
18.1		**23,96**	24,02	24,33	—	—	—	25,32	24,91	24,89
19.1		24,48	23,85	—	—	—	—	24,94	24,86	27,36
21,1		24,13	24,12	25,98	—	—	—	24,9	**24,35**	24,5
22.1		24,9	23,77	24,26	24,73	—	—	24,91	25,03	26,55
24,1		24,87	23,77	**23,71**	24,71	—	—	25,28	51,03	**24,4**
25.1		24,04	**23,23**	23,82	**24,15**	**24,18**	—	24,77	24,67	24,83
mean	24,99	26,05	25,84	27,31	28,03	25,20	26,65	25,89	26,74	25,28
median	24,99	24,65	24,57	25,33	24,99	24,93	25,41	24,99	24,91	24,99
std	0	5,49	5,56	6,23	7,18	0,76	0,86	3,17	6,32	0,77

Tab.5: The results pertain to the MSE of the test for the best-trained model selected from among ten ones. The best model for each scenario is highlighted in bold. Grey values indicate that the corresponding model is equivalent to the preceding one in the same row. The cells with "—" indicate models equivalent to the previous case of the same scenario. All values are reported rounding to the nearest hundredth.

ε	**Univariate**	**GNN**	**Corr 0.3**	**Corr 0.4**	**Corr 0.5**	**Corr 0.6**	**Corr 0.7**	**Miss –0.5**	**Miss 0**	**Miss +0.5**
0,1	**21,5**	21,5	21,5	21,5	21,5	21,5	21,5	21,5	21,5	21,5
1,1		21,78	21,78	21,78	21,78	21,78	21,78	21,35	21,35	21,46
2,1		21,4	21,4	21,4	21,4	—	—	21,51	21,51	21,52
3,1		21,36	21,36	21,36	21,36	21,3	21,3	21,85	23,18	21,63
6,1		21,3	21,3	21,3	21,3	21,56	21,56	21,45	21,3	—
7,1		21,55	21,55	21,55	21,55	21,19	**21,19**	21,62	21,4	21,17
11,1		20,91	20,91	21,55	—	—	—	21,15	21,1	21,46
12,1		21,06	21,06	21,2	—	—	—	21,43	21,37	21,09
13,1		21,8	21,8	21,57	—	—	—	21,23	21,53	21,59
14,1		21,14	21,14	—	—	—	—	21,08	20,97	—
17,1		21,26	21,26	20,98	—	—	—	21,03	21,4	21,38
18,1		20,68	21,13	20,7	—	—	—	21,08	21,25	21,36
19,1		20,73	20,97	—	—	—	—	21,13	20,96	21,24
21,1		21,06	21,13	20,86	—	—	—	**20,94**	21,2	21,31
22,1		**20,51**	21,03	21,01	21,13	—	—	21,04	20,97	21,63
24,1		20,95	**20,55**	**20,39**	21	—	—	21,24	**20,78**	—
25,1		20,72	20,58	20,55	**20,8**	**21,17**	—	21,01	21,09	**20,82**
mean	21,5	21,16	21,2	21,18	21,31	21,42	21,47	21,27	21,34	21,37
median	21,5	21,14	21,14	21,3	21,36	21,4	21,5	21,23	21,3	21,42
std	0	0,38	0,35	0,41	0,30	0,24	0,23	0,25	0,52	0,23

Tab 6: The results pertain to the MSE of the test for the best-trained model selected from among ten ones excluding CENAS9. The best model for each scenario is highlighted in bold. Grey values indicate that the corresponding model is equivalent to the preceding one in the same row. The cells with "—" indicate models equivalent to the previous case of the same scenario. All values are reported rounding to the nearest hundredth.

4 Conclusions

The study's findings highlight the improvements achieved through the utilization of various GNN-LSTM-based models for NO_2 concentration prediction. These advancements are more pronounced when filtering the adjacency matrix with a correlation index and a missing value percentage threshold. Moreover, the correlation filter method provides valuable insights into the specific nodes of the graph that significantly influence the overall outcome, enhancing our understanding of the complex dynamics of air pollution in urban environments.

Incorporating the considered filters into the GNN-LSTM models introduces uncertainty in the performance of the model, when the control unit CENAS9 is considered, and highlights the dependency among the time series of each node into the multidimensional analysis. Indeed, when the station CENAS9 is considered, several GNN-LSTM models were negatively affected (see Table 5). Furthermore, when CENAS9 was removed from the network, the overall network homogeneity improved, as the predictions for each node wee more precise (see Table 6). In both cases, a rule of thumb emerged, indicating that better results are generally concentrated around an absolute value of 0.5 for the correlation index threshold, while the distance threshold should be greater than 20 kilometers.

To summarize, our study highlights the efficacy of GNN-LSTM-based models in NO_2 concentration prediction and the additional benefits of incorporating the correlation-based and the missing value percentage-based filters into the analysis. The presence or absence of certain nodes in the graph influences the model's performance, and a simple backward analysis allows the recognition of these nodes. This research contributes several insights into air pollution dynamics and serves as a foundation for further investigations in this field.

Acknowledgments

This study was developed in the framework of the Artificial Intelligence project, funded by the Sardinian Regional Authorities ex art 9 L.R. 20/2015.

Author Contribution

C.M.: Conceptualization, Methodology, Software, Data Curation, Formal analysis, Writing - Original draft;

F.P.: Conceptualization, Methodology, Software, Data Curation, Formal analysis, Writing - Original draft;

C.C.: Supervision, Writing - Reviewing and Editing, Project administration, Funding acquisition.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Brody, S., Alon, U., & Yahav, E. (n.d.). *HOW ATTENTIVE ARE GRAPH ATTENTION NETWORKS?* Retrieved October 20, 2023, from https://github.com/tech-srl/how_attentive_are_

Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2nd Edition). Duxbury Press, Pacific Grove.

Chowdhury, S., Roychowdhury, S., & Chaudhuri, I. (2023). *Effect of air pollution on the growth of diabetic population*.

De Santis, D., Amici, S., Milesi, C., Muroni, D., Romanino, A., Casari, C., Cannas, V., & Del Frate, F. (2023). Tracking air quality trends and vehicle traffic dynamics at urban scale using satellite and ground data before and after the COVID-19 outbreak. *Science of The Total Environment*, *899*, 165464. https://doi.org/10.1016/j.scitotenv.2023.165464

Du, S., Li, T., Gong, X., & Horng, S.-J. (2018). *A Hybrid Method for Traffic Flow Forecasting Using Multimodal Deep Learning*.

Fan, J., Li, Q., Hou, J., Feng, X., Karimian, H., & Lin, S. (2017). A Spatiotemporal Prediction Framework for Air Pollution Based on Deep RNN. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *IV-4/W2*, 15–22. https://doi.org/10.5194/isprs-annals-IV-4-W2-15-2017

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Inadagbo, K., Arig, B., Alici, N., & Isik, M. (2023). *Exploiting FPGA Capabilities for Accelerated Biomedical Computing*.

J. L. Gross and J. Yellen. (2003). *Handbook of Graph Theory.* (CRC Press, Ed.).

Jiang, W., & Luo, J. (2021). *Graph Neural Network for Traffic Forecasting: A Survey*. https://doi.org/10.1016/j.eswa.2022.117921

Kelly, T. J., Spicer, C. W., & Ward, G. F. (1990). An assessment of the luminol chemiluminescence technique for measurement of NO2 in ambient air. *Atmospheric Environment. Part A. General Topics*, *24*(9), 2397–2403. https://doi.org/10.1016/0960-1686(90)90332-H

Kumar, A., Patil, R. S., Dikshit, A. K., & Kumar, R. (2017). Application of WRF Model for Air Quality Modelling and AERMOD - A Survey. *Aerosol and Air Quality Research*, *17*(7), 1925–1937. https://doi.org/10.4209/aaqr.2016.06.0265

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., & Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*. https://doi.org/10.1126/science.adi2336

Landrigan, P. J., Fuller, R., Acosta, N. J. R., Adeyi, O., Arnold, R., Basu, N. (Nil), Baldé, A. B., Bertollini, R., Bose-O'Reilly, S., Boufford, J. I., Breysse, P. N., Chiles, T., Mahidol, C., Coll-Seck, A. M., Cropper, M. L., Fobil, J., Fuster, V., Greenstone, M., Haines, A., … Zhong, M. (2018). The Lancet Commission on pollution and health. *The Lancet*, *391*(10119), 462–512. https://doi.org/10.1016/S0140-6736(17)32345-0

Li, J., Crooks, J., Murdock, J., de Souza, P., Hohsfield, K., Obermann, B., & Stockman, T. (2023). A nested machine learning approach to short-term PM2.5 prediction in metropolitan areas using PM2.5 data from different sensor networks. *Science of The Total Environment*, *873*, 162336. https://doi.org/10.1016/j.scitotenv.2023.162336

Liang, M., Chao, Y., Tu, Y., & Xu, T. (2023). Vehicle Pollutant Dispersion in the Urban Atmospheric Environment: A Review of Mechanism, Modeling, and Application. *Atmosphere*, *14*(2), 279. https://doi.org/10.3390/atmos14020279

Liu, X., Tan, H., Chen, Q., & Lin, G. (2021). RAGAT: Relation Aware Graph Attention Network for Knowledge Graph Completion. *IEEE Access*, *9*, 20840–20849. https://doi.org/10.1109/ACCESS.2021.3055529

Méndez, M., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to forecast air quality: a survey. *Artificial Intelligence Review*, *56*(9), 10031–10066. https://doi.org/10.1007/s10462-023-10424-4

Neto, A. B., Ferraro, A. A., & Vieira, S. E. (2023). Acute and subchronic exposure to urban atmospheric pollutants aggravate acute respiratory failure in infants. *Scientific Reports*, *13*(1), 16888. https://doi.org/10.1038/s41598-023-43670-1

Qasim, S. R., Kieseler, J., Iiyama, Y., & Pierini, M. (2019). Learning representations of irregular particle-detector geometry with distance-weighted graph networks. *The European Physical Journal C*, *79*(7), 608. https://doi.org/10.1140/epjc/s10052-019-7113-9

Qi, Y., Li, Q., Karimian, H., & Liu, D. (2019). A hybrid model for spatiotemporal forecasting of PM2.5 based on graph convolutional neural network and long short-term memory. *Science of The Total Environment*, *664*, 1–10. https://doi.org/10.1016/j.scitotenv.2019.01.333

Rafaj, P., Kiesewetter, G., Gül, T., Schöpp, W., Cofala, J., Klimont, Z., Purohit, P., Heyes, C., Amann, M., Borken-Kleefeld, J., & Cozzi, L. (2018). Outlook for clean air in the context of sustainable development goals. *Global Environmental Change*, *53*, 1–11. https://doi.org/10.1016/j.gloenvcha.2018.08.008

Rahmani, S., Baghbani, A., Bouguila, N., & Patterson, Z. (2023). Graph Neural Networks for Intelligent Transportation Systems: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, *24*(8), 8846–8885. https://doi.org/10.1109/TITS.2023.3257759

Rodeschini, J., Fassò, A., Moro, A. F., & Finazzi, F. (2023). *Scenario analysis of livestock-related PM2.5 pollution based on heteroskedastic geostatistical modelling*.

Seng, D., Zhang, Q., Zhang, X., Chen, G., & Chen, X. (2021). Spatiotemporal prediction of air quality based on LSTM neural network. *Alexandria Engineering Journal*, *60*(2), 2021–2032. https://doi.org/10.1016/j.aej.2020.12.009

Seyyedi, A., Bohlouli, M., & Oskoee, S. N. (2023). Machine Learning and Physics: A Survey of Integrated Models. *ACM Computing Surveys*. https://doi.org/10.1145/3611383

Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., & Yu, P. S. (2019). Heterogeneous Graph Attention Network. *The World Wide Web Conference*, 2022–2032. https://doi.org/10.1145/3308558.3313562

Xiao, X., Jin, Z., Wang, S., Xu, J., Peng, Z., Wang, R., Shao, W., & Hui, Y. (2022). A dual-path dynamic directed graph convolutional network for air quality prediction. *Science of The Total Environment*, *827*, 154298. https://doi.org/10.1016/j.scitotenv.2022.154298

Xu, Z., Kang, Y., Cao, Y., & Li, Z. (2021). Spatiotemporal Graph Convolution Multifusion Network for Urban Vehicle Emission Prediction. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(8), 3342–3354. https://doi.org/10.1109/TNNLS.2020.3008702

Zhang, W., Liu, H., Liu, Y., Zhou, J., & Xiong, H. (2019). *Semi-Supervised Hierarchical Recurrent Graph Neural Network for City-Wide Parking Availability Prediction*.