AGILE: GIScience Series, 3, 6, 2022. https://doi.org/10.5194/agile-giss-3-6-2022
Proceedings of the 25th AGILE Conference on Geographic Information Science, 2022.
Editors: E. Parseliunas, A. Mansourian, P. Partsinevelos, and J. Suziedelyte-Visockiene.
This contribution underwent peer review based on a full paper submission.

Exploratory Analysis and Feature Selection for the Prediction of Nitrogen Dioxide

Ditsuhi Iskandaryan[1], Silvana Di Sabatino[2], Francisco Ramos[1], and Sergio Trilles[1]

[1]Institute of New Imaging Technologies (INIT), Universitat Jaume I, Av. Vicente Sos Baynat s/n, 12071 Castelló de la Plana, Spain

[2]Department of Physics and Astronomy, University of Bologna, Via Irnerio 46, 40127 Bologna, Italy

Correspondence: Ditsuhi Iskandaryan (iskandar@uji.es)

Abstract. Nitrogen dioxide is one of the most hazardous pollutants identified by the World Health Organisation. Predicting and reducing pollutants is becoming a very urgent task and many methods have been used to predict their concentration, such as physical or machine learning models. In addition to choosing the right model, it is also critical to choose the appropriate features. This work focuses on the spatiotemporal prediction of nitrogen dioxide concentration using Bidirectional Convolutional LSTM integrated with the exploration of nitrogen dioxide and associated features, as well as the implementation of feature selection methods. The Root Mean Square Error and the Mean Absolute Error were used to evaluate the proposed approach.

Keywords. nitrogen dioxide prediction, feature selection, mRMR, mutual information, machine learning

1 Introduction

Air pollution can have a severe effect on the environment and health. Reducing pollution is a challenge both locally and globally. Before applying certain measurements in order to reduce pollution, it is essential to monitor and identify dynamic changes. A more efficient way can be to predict the concentration in advance and find out the future trend before it can cause any negative consequences. Hazardous pollutants vary by location, which is the result of a combination of many factors such as topography, economic development, etc. (Zhao et al. (2020); Yang et al. (2020); Sun and Gu (2008)). According to the study by Sasha Khomenko et al. (Khomenko et al. (2021)) related to premature mortality due to air pollution in European cities, in which pollutant particles smaller than 2.5 micrometres in diameter ($PM_{2.5}$) and nitrogen dioxide (NO_2) were considered, Madrid, which is the case study of the current work, was found to be at the top of the ranking of European cities with the highest NO_2 mortality burden. Taking into consideration the importance of NO_2 for Madrid, it was selected as an air pollutant for predictive analysis.

Before performing the predictive analysis, a very important step is to conduct exploratory analyses, to identify existing relationships between NO_2 and other variables, and to select the best combination among the existing features by applying feature selection techniques. Having many features is not always good, as it causes many issues related to the curse of dimensionality (Altman and Krzywinski (2018); Verleysen and François (2005)), runtime execution, etc. Redundant and unnecessary data can lead to poor model performance. In addition to the above reasons, another reason for choosing the most optimal features is to prevent the lack of data; for example, if a certain feature is recorded and available for the city of Madrid, it may not be available for another case study. Thus, the ability to perform an analysis with a minimum number of features allows us to generalise the model, expand the application's spatial dimension, and reduce the execution time. Therefore, it is very important to select the minimum optimal features.

Many authors have implemented feature selection techniques in order to obtain better results. For example, Just et al. (2020) applied recursive feature selection based on least mean absolute SHAP values to predict $PM_{2.5}$; Shah and Mishra (2020) used correlation to predict $PM_{2.5}$; Xu and Ren (2019) used maximum relevance-minimum redundancy; Zheng et al. (2020) used recursive feature elimination with cross-validation for air quality health index prediction; Masmoudi et al. (2020) used Ensemble of Regressor Chains-guided Feature Ranking; and Liu and Chen (2020) applied three-stage feature selection, including Pearson's test, mutual information and binary grey wolf optimisation to predict the air quality index.

These aforementioned works confirm the advantage and importance of the implementation of feature selection methods, but they lacked the detailed exploratory analysis from the point of view of physical aspects, to which the current work is devoted, including the exploratory analysis and the importance of feature selection. The feature selection techniques implemented in this work are mutual information and maximum relevance-minimum redundancy techniques.

Regarding spatiotemporal prediction of air pollution, numerous studies have been conducted; in particular, with the help of machine learning and deep learning methods, more accurate prediction becomes achievable. One of these studies applied the STAR model based on the combination of CNN and LSTM to predict $PM_{2.5}$ and PM_{10} using the Seoul dataset (Bui et al. (2020)). Liu et al. (2022) implemented Support Vector Machine on the spatiotemporal features extracted using a geographic information system. Yan et al. (2021) proposed multi-time multi-site deep learning models (LSTM, CNN, CNN-LSTM) to forecast air quality. The model established in the scope of the current work is the Bidirectional Convolutional LSTM (BiConvLSTM). The advantage of BiConvLSTM was confirmed by several applications, including violence detection (Hanson et al. (2018)) and planetary gearbox fault diagnosis (Shi et al. (2022)). The architecture of BiConvLSTM allows predictions in the spatiotemporal dimension to be made with greater accuracy. However, compared to LSTM and ConvLSTM, BiConvLSTM takes longer to reach data convergence (Iskandaryan et al. (2022)). It is worth mentioning that in terms of spatial dimension, BiConvLSTM allows prediction not just for certain station locations, but also for areas where there are no air quality monitoring stations. The latter can be achieved using additional features and their corresponding locations.

The main objective of the current research is to use BiConvLSTM in combination with exploratory analysis and feature selection techniques to forecast the next 6 hours of NO_2 concentration in the spatiotemporal dimension. The following are the main questions that the ongoing work tries to answer: Which feature extraction technique is better: mutual information or minimum redundancy-maximum relevance? What is the best combination of the features causing the performance of the best model? Which wind direction transformation affects obtaining the best model performance?

The rest of the work has the following structure. Section 2 describes the datasets and software implemented. Section 3 introduces exploratory analysis by revealing the existing relationships between features. Section 4 illustrates the models and techniques used in this work. Section 5 presents the experiments and the results obtained, and section 6 summarises and reveals the main conclusions.

2 Data and Software Availability

This section introduces the datasets that have been used in the analyses. The datasets used in this work are NO_2 (μg/m3), meteorological data and traffic data from the period January-June 2019 (training set) and January-June 2020 (validation and testing set), and the location of the monitoring stations in the city of Madrid. The data were obtained from the Open Data portal of the City Council of Madrid[1]. There are 24 air quality control stations, 26 meteorological control stations and more than 4,000 traffic measurement points (shapefiles of measurement point locations are also provided for each month). The datasets include the following variables (Table 1 shows summary statistics of each type of data of the periods used in the analyses):

- Air Quality Data - NO_2 ($\mu g/m^3$).
- Meteorological Data - Ultraviolet radiation (Mw/m^2), Wind speed (m/s), Wind direction, Temperature (^oC), Relative humidity (%), Barometric pressure (mb), Solar irradiance (W/m^2), Precipitation (l/m^2).
- Traffic Data - Since the attributes of the traffic data can be specific to a certain area, the traffic attributes selected with their definition for the city of Madrid are shown below.
  - Intensity - Intensity of the measurement point in a period of 15 minutes (vehicles/hour).
  - Occupancy time - Measurement point occupancy time in a period of 15 minutes (%).
  - Load - Vehicle loading in a 15-minute period. This is a parameter that takes into account intensity, occupation and capacity of the road and establishes the degree of road use from 0 to 100.
  - Average speed - Average speed of the vehicles in a period of 15 minutes (km/h). Only for M30 intercity measuring points.

Although the traffic data is recorded every 15 minutes, since the NO_2 and meteorological data are at hourly rates, the traffic data was filtered and only hourly records were selected (for example, with entries at 13:00, 13:15, 13:30, 13:45 and 14:00, we simply selected the entries at 13:00 and 14:00 and the same logic was applied for the entire period).

Since the monitoring stations and measurement points are different for each dataset, the first task is to combine them spatially and temporally. Therefore, the grid was created in a given area, which was defined as a selected part of Madrid with a width and height of 1,000 metres within the following extent: Top -4,486,449.725263 metres; Bottom -4,466,449.725263 metres; Left -434,215.234430 metres;

[1]Portal de datos abiertos del Ayuntamiento de Madrid: https://bit.ly/3FFRiQM

Table 1. Summary statistics of the periods January-June 2019 and January-June 2020 for each data type.

Phenomena	Descriptors	January-June 2019	January-June 2020
Nitrogen dioxide	Mean (SD)	36.69 (30.85)	26.03 (25.35)
	Median [Min,Max]	27.0 [0.0, 328]	17.0 [0.0, 326]
UV	Mean (SD)	15.83 (30.27)	-
	Median [Min,Max]	1.0 [0.0, 199]	-
Wind speed	Mean (SD)	1.41 (1.11)	1.31 (1.05)
	Median [Min,Max]	1.14 [0.0, 8.75]	1.05 [0.0, 8.97]
Wind direction	Mean (SD)	167.80 (105.72)	140.82 (98.35)
	Median [Min,Max]	182.0 [0.0, 359]	135.0 [0.0, 359]
Temperature	Mean (SD)	13.38 (8.09)	13.63 (7.6)
	Median [Min,Max]	12.5 [-55.0, 47.3]	12.6 [-55.0, 44.6]
Humidity	Mean (SD)	48.73 (21.60)	60.76 (22.77)
	Median [Min,Max]	47.0 [-25, 100]	62.0 [-25, 100]
Pressure	Mean (SD)	943.3 (34.91)	940.62 (63.28)
	Median [Min,Max]	945.0 [0.0, 962.0]	945.0 [0.0, 1073.0]
Solar irradiance	Mean (SD)	220.73 (301.06)	191.95 (279.83)
	Median [Min,Max]	11.0 [0.0, 1103.0]	9.0 [0.0, 1113.0]
Precipitation	Mean (SD)	0.03 (0.41)	0.03 (0.27)
	Median [Min,Max]	0.0 [0.0, 30.4]	0.0 [0.0, 13.5]
Intensity	Count_non_zero	885863 (59.98%)	892197 (60.09%)
	Mean (SD)	245.69 (402.73)	161.45 (313.33)
	Median [Min,Max]	63.0 [0.0, 6348.0]	34.19 [0.0, 6588.0]
Occupancy time	Count_non_zero	845031 (57.21%)	822652 (55.41%)
	Mean (SD)	3.96 (6.36)	2.57 (4.9)
	Median [Min,Max]	0.95 [0.0, 100.0]	0.42 [0.0, 99.0]
Load	Count_non_zero	881500 (59.68%)	884950 (59.60%)
	Mean (SD)	11.65 (14.91)	7.85 (11.75)
	Median [Min,Max]	4.0 [0.0, 100.0]	2.2 [0.0, 100.0]
Average speed	Count_non_zero	233415 (15.8%)	223052 (15.0%)
	Mean (SD)	4.39 (13.28)	4.04 (12.96)
	Median [Min,Max]	0.0 [0.0, 96.5]	0.0 [-127.0, 127.0]

Right -451,215.234430 metres (Figure 1). There are 340 cells (20 by 17), which cover 56.27% of the total area of the city of Madrid. The logic behind selecting this area was to have a minimum extension to include all the air quality control stations so as to obtain higher accuracy. The value of each cell includes the values of NO_2, meteorological and traffic attributes recorded from assigned stations at a particular time. The value of the cells that do not contain any stations was set to zero, and in the case of several stations, the average value was set. The procedure described above was repeated for each hour of the selected period. The source code can be accessed at the GitHub repository[2].

The following software was used to process the data: *ArcGIS Pro*[3] with *ArcPy* package[4] for combining air quality, meteorological and traffic data both spatially and temporally (the data generated are available at the Zenodo repository[5]), and for creating maps; *WRPLOT VIEW* platform[6] for generating wind roses; *Openair R package*[7] for generating polar plots; *Plotly Python Graphing Library*[8] for generating graphs; and *Google Colab* cloud service[9] for running the models.

3 Exploratory Analysis

This step illustrates the results of observing features in the defined area. Examining features makes it possible to take a deep look at the data and, by correlating with NO_2, to understand which features are the most important for predicting NO_2, as well as selecting these features as a subset and using them as input for further predictive analyses. The exploratory analysis includes comparative analysis, which

[2]GitHub repository: https://github.com/Ditsuhi/ExploratoryAnalysis_FeatureSelection

[3]ArcGIS Pro: https://pro.arcgis.com/en/pro-app/latest/get-started/get-started.htm

[4]ArcPy package: https://pro.arcgis.com/en/pro-app/2.8/arcpy/get-started/what-is-arcpy-.htm

[5]Zenodo repository: https://doi.org/10.5281/zenodo.6076631

[6]WRPLOT VIEW: https://www.weblakes.com/software/freeware/wrplot-view/

[7]Openair R package: https://bookdown.org/david_carslaw/openair/polar-plots.html

[8]Plotly Python Graphing Library: https://plotly.com/python/

[9]Google Colab: https://colab.research.google.com/notebooks/intro.ipynb

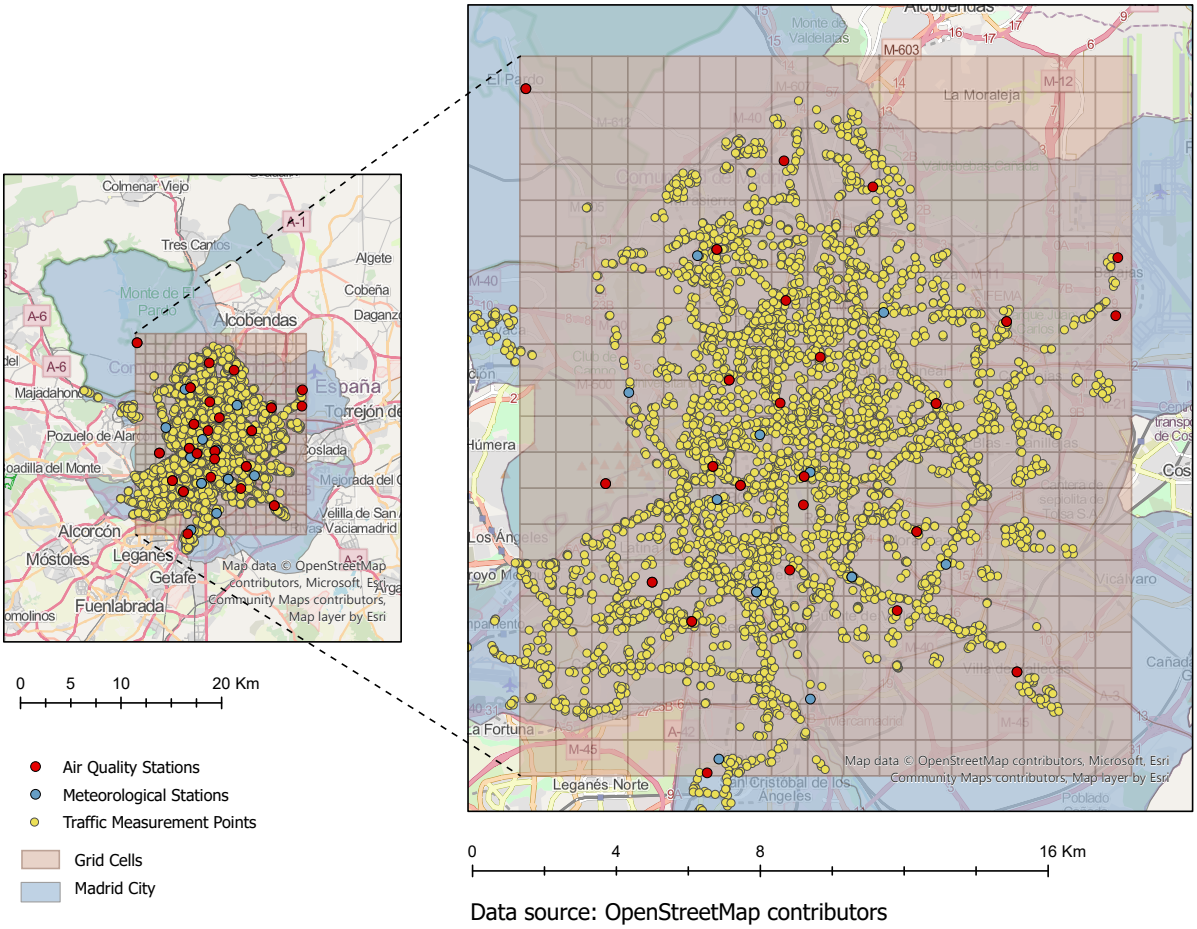

Figure 1. Air quality stations, meteorological stations, traffic measurement points (January 2019) and grid cells segments on the defined area of the city of Madrid.

was performed for January and June 2019, to explore the variables' behaviour during two different seasonal periods.

First of all, a wind rose was created for each station, and it turned out that out of 26 meteorological stations, only ten stations provide data on wind speed and direction. Then, based on the wind roses obtained, a map was generated showing the dominant wind directions at each station, marked with different colours (Figure 2).

The output showed that in January we had the following dominant directions with the station's id, respectively - North: 173, 217; Northeast: 214, 96; East: 72; Southwest: 138; South: 42; West: 242, 47, 5 - and in June: South: 42, 214; Southwest: 72, 138, 173, 217, 242; West: 5, 47, 96. Wind speed was classified based on the Beaufort scale (Delmar-Morgan (1959); Huler (2007)).

After generating a wind rose for each station, it was found that higher wind speed does not always coincide with the dominant direction. For example, Figure 3 shows that at the station with id=96 during January, which had calm winds 3.63%, the predominant direction is northeast, but a higher wind speed was found in the westerly direction.

To reveal a relationship between concentration and wind speed, the time series of those variables were plotted to see how they change over time. For example, Figure 4 shows a time series of NO_2 and wind speed during January for the station with id=5. It can be observed that these two variables are inversely proportional; particularly, increasing wind speed assumes lower concentration due to increased dilution through advection and increased mechanical turbulence. This finding can also be seen in the scatter plot, which was plotted by taking the y-axis for NO_2 and the x-axis for wind speed (Figure 5). The trendline is based on locally weighted scatter plot smoothing (Cleveland (1979)). Also, in Figure 6, which shows the time series of NO_2 and wind speed of a typical day during January (a typical day is a day on which hourly data is the average

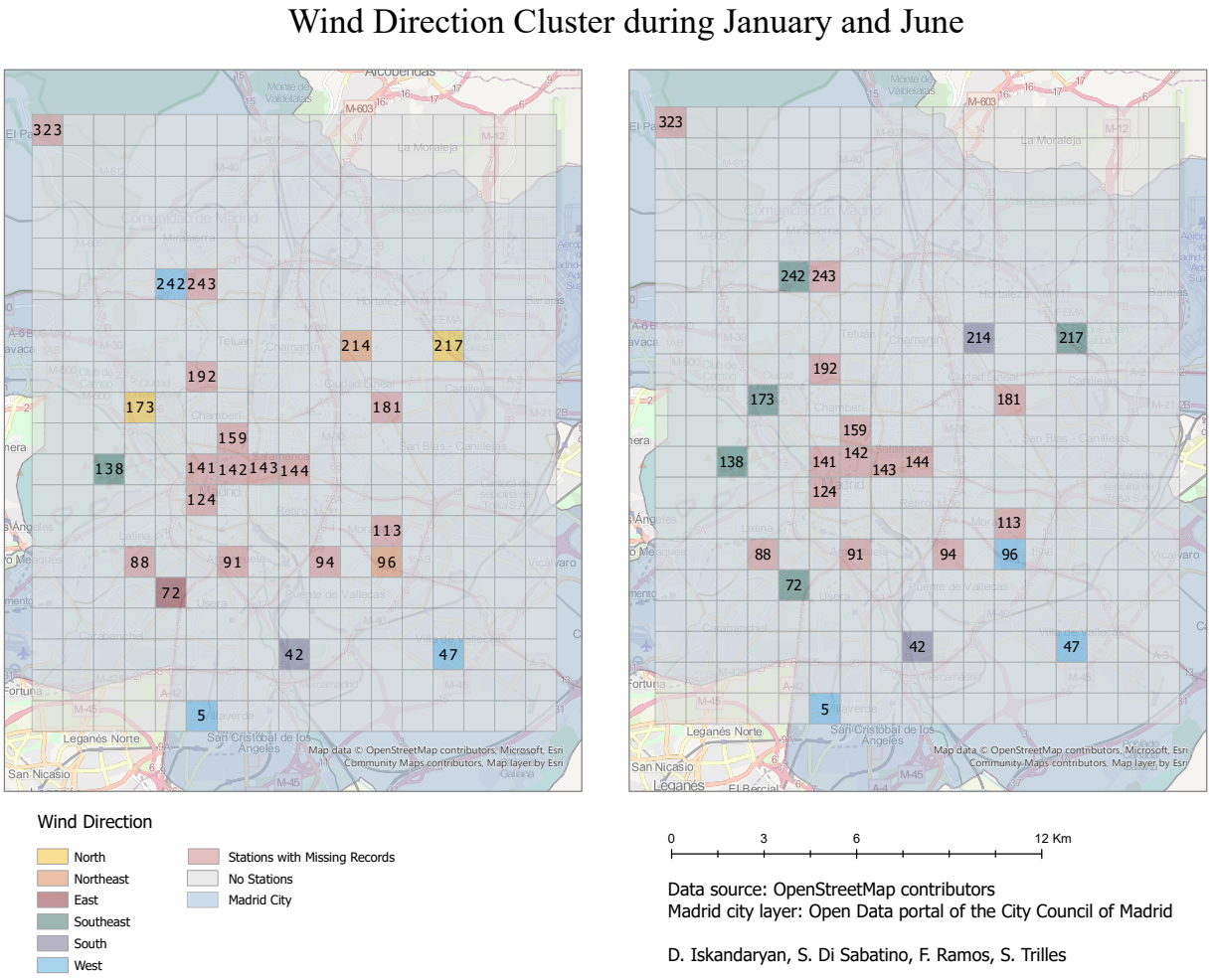

Figure 2. Wind direction cluster during January (left) and June (right) 2019 in the city of Madrid.

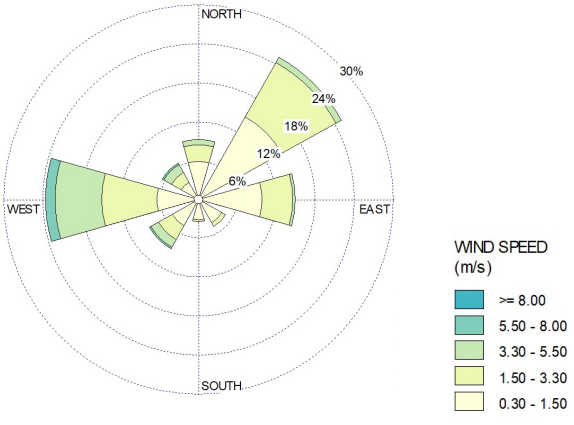

Figure 3. Wind Rose at the station with id=96 during January 2019 in the city of Madrid.

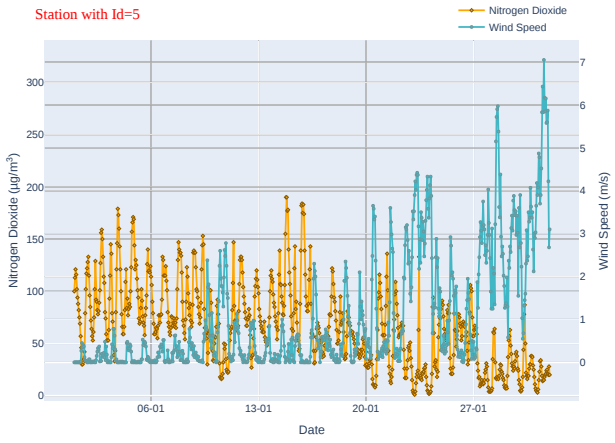

Figure 4. Time series of NO_2 and wind speed during January 2019 at the station with id=5 in the city of Madrid.

of all the records for a given hour over a specified period). Again, it can be seen that the concentration decreases with increasing wind speed.

Another analysis was performed to generate polar plots using the openair R package to show the relationship of concentration, wind speed and wind direction. Figure 7 and

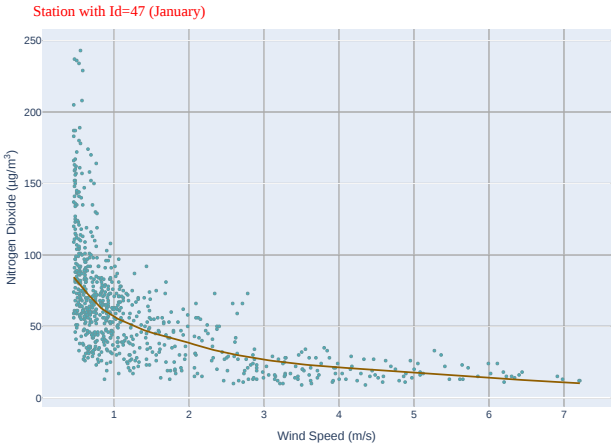

Figure 5. Scatter plot of NO_2 and wind speed during January 2019 at the station with id=47 in the city of Madrid.

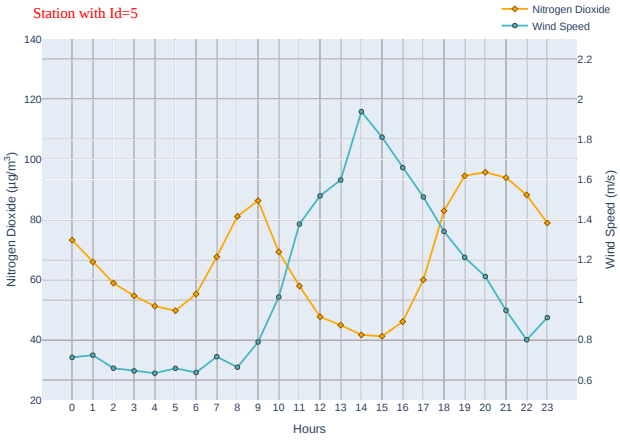

Figure 6. Time series of NO_2 and wind speed of a Typical Day during January 2019 at the station with id=5 in the city of Madrid.

Figure 8 show polar plots at the station with id=47 during January and during June. It can be seen that in the central part with a lower wind speed, the concentration is higher, and at the edges with a higher wind speed, the concentration is lower. In the polar plots obtained using the average NO_2, it is noticeable that the concentration is lower in June than in January, which can be explained by domestic heating.

An additional analysis was carried out to determine the relationship between non-dimensional concentration and non-dimensional wind speed. The formula to calculate the non-dimensional concentration (Eq. (1)) and non-dimensional wind speed (Eq. (2)) are illustrated below (Stull (2015)).

$$C\_ADIM = C * U * L * H / EMISSIONS \quad (1)$$

$$U\_ADIM = U / U\_arp \quad (2)$$

where,

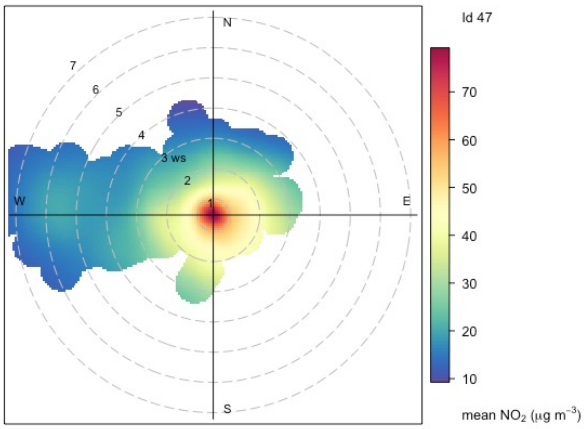

Figure 7. Polar plot of wind speed, wind direction and mean concentration of NO_2 during January 2019 at the station with id=47 in the city of Madrid.

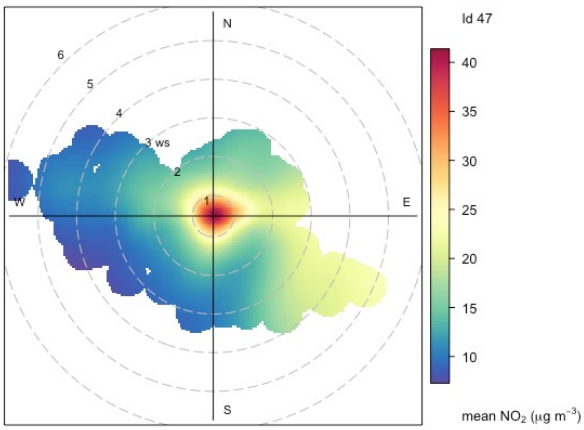

Figure 8. Polar plot of wind speed, wind direction and mean concentration of NO_2 during June 2019 at the station with id=47 in the city of Madrid.

- C_ADIM – Non-dimensional concentration,
- C – Concentration (μg/m3),
- U – Wind speed (m/s),
- L – Road length (km) in a certain cell (it was calculated using ArcGIS Pro software),
- H – Planetary boundary layer height (m) in the Adolfo Suárez Madrid–Barajas Airport, which was generated by the ERA5 model (European Centre for Medium-Range Weather Forecasts[10]),
- EMISSIONS – Nitrogen Oxides (NOx),
- U_ADIM – Non-dimensional wind speed,

[10]ECMWF: https://www.ecmwf.int/en/about

- U_arp – Wind speed [10m] in the Adolfo Suárez Madrid–Barajas Airport (m/s), which was obtained from the ERA5 model.

Figure 9 and Figure 10 show the scatter plots of the non-dimensional concentration and non-dimensional wind speed. In those plots the relationship is not clear.

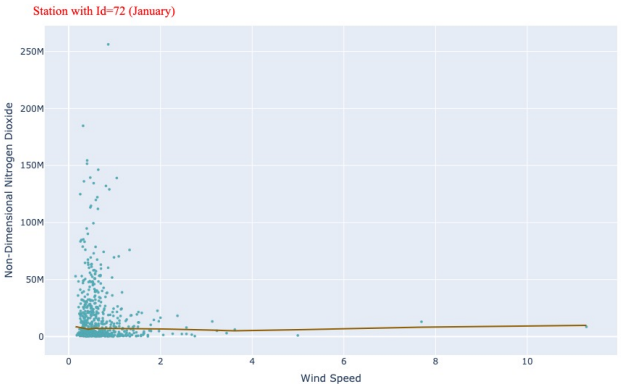

Figure 9. Scatter plot of the non-dimensional NO_2 concentration and non-dimensional wind speed during January 2019 at the station with id=72 in the city of Madrid.

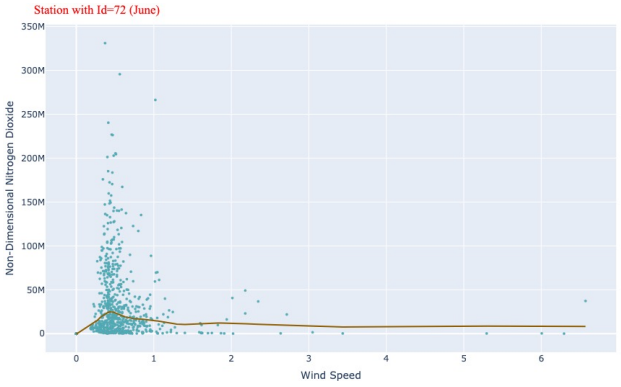

Figure 10. Scatter plot of the non-dimensional NO_2 concentration and non-dimensional wind speed during June 2019 at the station with id=72 in the city of Madrid.

Overall, observing the plots related to concentration and wind speed, it was found that the concentration in the station with id=72 is higher and wind speed is lower for January and June; the concentration in the station with id=138 in January is the lowest; concentration and wind speed are more correlated during winter than in summer.

The above mentioned analyses were performed between NO_2 and other variables, although it was challenging to reveal a certain correlation from the plots. All plots generated during the exploratory analyses are available in the GitHub repository[11].

For future selection, several points must be considered. First, the following variables can be excluded for future predictive analysis: UV and precipitation. Regarding UV, it was observed that in January it was recorded only in three stations that have NO_2 records (station: id=47, id=38, id=217), and June has no UV records; moreover, there were no records for the period from January to June 2020. Regarding precipitation, it was found that around 99% of data were 0. Another feature that may be excluded is average traffic speed. This is because the average traffic speed is available only for the M30 road, which is 15.8% of the case study. (Figure 11 shows average traffic speed for a period of one week). However, it will be included in further analyses in order to track the results after implementing feature selection methods.

Overall, from all these observations and analyses, it can be summarised that the most correlated feature with NO_2 is wind speed, and the features that definitely have to be excluded are UV and precipitation. Further analysis will be performed based on features without the aforementioned excluded features.

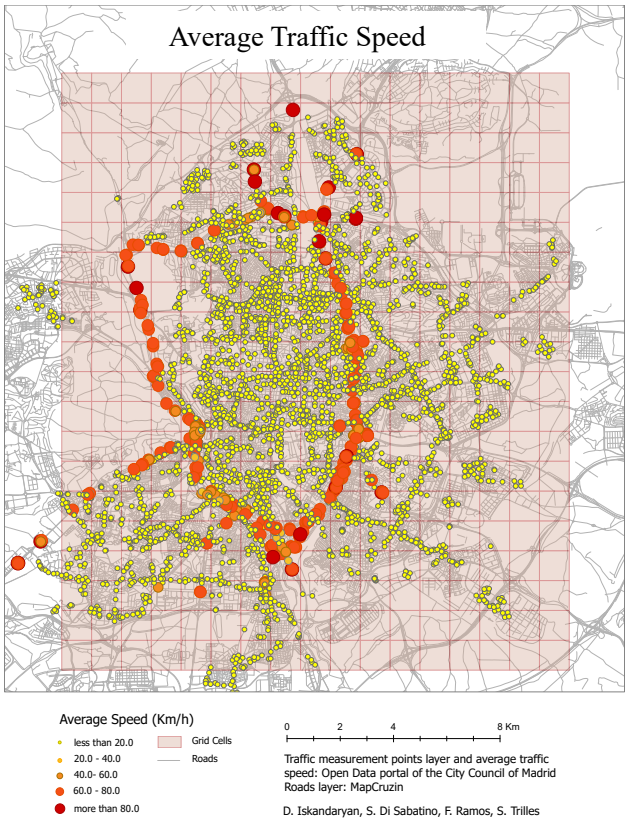

Figure 11. Average traffic speed for the period 1-7 January 2019 in the city of Madrid.

4 Methodology and Evaluation Metrics

This section describes the methods used for feature selection and predictive analysis, and the metrics used to evaluate model performance.

[11] GitHub repository: https://github.com/Ditsuhi/ExploratoryAnalysis_FeatureSelection

4.1 Feature Selection Techniques

Feature selection should be implemented to select the best combination of features, which will prompt the model to generalise data efficiently. In this work, the following feature selection techniques were used: Mutual Information (MI) and Maximum Relevance-Minimum Redundancy (mRMR) (Peng et al. (2005); Zhao et al. (2019)).

Mutual information: This technique calculates the mutuality between additional features and the target feature (NO_2). The formula to calculate mutual information is presented below (Eq. (3)).

$$MI(x;y) = \iint P(x_i,y) log \frac{P(x_i,y)}{P(x_i)P(y)} dx_i\, dy \\ = H(x) - H(x|y) \tag{3}$$

where $P(x_i,y)$ is the joint probability distribution of two variables, $P(x_i)$ and $P(y)$ are marginal distributions, $H(x)$ is the entropy for x, and $H(x|y)$ is the conditional entropy.

Maximum Relevance-Minimum Redundancy: mRMR selects the features that are the most relevant to the target also considering minimum redundancy concerning the features that have already been selected. The equation of the mRMR is the following (Eq. (4)).

$$score_i(f) = \frac{F(f,target)}{\sum_{s \in features\ selected\ until\ i-1} |corr(f,s)|/(i-1)} \tag{4}$$

where i is the i-th iteration, f is the feature that is being evaluated, F is the F-statistic and corr is the Pearson correlation.

4.2 Bidirectional Convolutional LSTM

Bidirectional Convolutional LSTM (BiConvLSTM) was chosen as the machine learning method for predicting NO_2 and comparing results obtained with different feature selection techniques. BiConvLSTM is an advanced version of ConvLSTM, which is able to preserve spatiotemporal information. It combines the LSTM unit (responsible for temporal information) and the convolutional layer (responsible for spatial information), and adding a bidirectional factor captures more information in the time dimension. Below is the mathematical expression of BiConvLSTM (Eq. (5)) (Song et al. (2018)).

$$Y_t = \tanh(W_y^{Hf} * H_t^f + W_y^{Hb} * H_{t-1}^b) \tag{5}$$

where H^f is the hidden state from the forward ConvLSTM unit, H^b is the hidden state from the backward ConvLSTM unit, and Y_t is the final output. The ConvLSTM can be formulated with the following equations (Eq. (6)) (Shi et al. (2015); Song et al. (2018)):

$$\begin{aligned} i_t &= \sigma(W_i^X * X_t + W_i^H * H_{t-1}) \\ f_t &= \sigma(W_f^X * X_t + W_{f^H} * H_{t-1}) \\ o_t &= \sigma(W_o^X * X_t + W_o^H * H_{t-1}) \\ C_t &= f_t \otimes C_{t-1} + i_t \otimes tanh(W_c^X * X_t \\ &\quad + W_c^H * H_{t-1}) \\ H_t &= o_t \otimes tanh(C_t) \end{aligned} \tag{6}$$

where i_t is the input gate, f_t is the forget gate, and o_t is the output gate (these gates control the flow of information through the cell), W is the weight matrix in the forward ConvLSTM cell, X_t is the current input data, h_{t-1} is the previous hidden output, C_t is the cell state, "$*$" represents the convolution operation and "$\otimes$" represents the Hadamard product.

4.3 Evaluation Metrics

To evaluate the model performance, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) were chosen as evaluation metrics. RMSE measures the geometric difference between estimated and actual values and it is very sensitive to large errors (Eq. (7)), and MAE measures the average magnitude of the errors (Eq. (8)).

$$RMSE = \left(\frac{1}{n} \sum_{i=1}^{n} (E_i - A_i)^2 \right)^{1/2} \tag{7}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |E_i - A_i| \tag{8}$$

where n is the number of instances, and E_i and A_i are the estimated and actual values. The lower the value, the better the prediction.

5 Experiments and Results

This section describes the experiments and the results obtained. It should be mentioned that this work is the extension of the following work (Iskandaryan et al. (2022)). Therefore, the detailed description of data generation, data preprocessing steps and model development can be found in the aforementioned work. The main focus of the current work is feature selection and data transformation. The workflow is illustrated in Figure 12.

In this stage, the BiConvLSTM will be applied to the selected subsets obtained using the feature selection methods, including mutual information and maximum

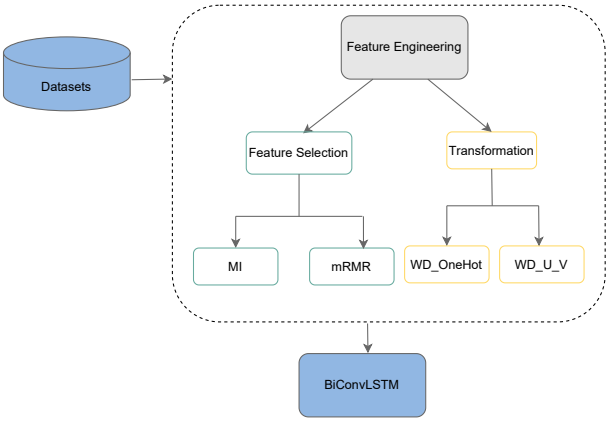

Figure 12. The workflow of the analysis focusing on the two components of feature engineering, including feature selection and transformation.

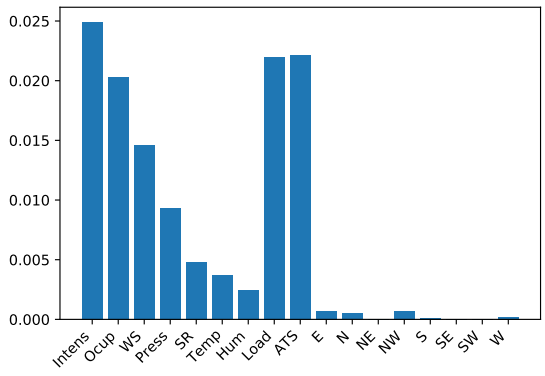

Figure 13. Feature selection using the mutual information technique (Wind direction with One Hot Encoder).

relevance-minimum redundancy. It should also be pointed out that based on the fact that wind direction is a cyclical feature, extra preprocessing steps must be implemented in order to transform wind direction data. Based on the transformation mechanism, the experiments were carried out with the following scenarios:

First scenario: Wind direction was converted to the following categories: north, east, south, west, southwest, northeast, southeast, northwest, and later it was included in the analysis by implementing One Hot Encoder[12].

Second scenario: Wind direction was converted to u and v components using the following equations (Eq. (9))[13].

$$u = ws * \cos(\theta)$$
$$v = ws * \sin(\theta) \tag{9}$$

where ws is the wind speed, θ is the wind direction using mathematical direction (mathematical wind direction = 270-meteorological wind direction).

Feature selection techniques were implemented for each scenario. Figure 13 and Figure 14 show the results of both scenarios based on the mutual information technique. The features selected were those with a score higher than 0.005. In Figure 13 it can be observed that among 17 features the following 6 were selected: intensity, occupancy time, wind speed, pressure, load and average traffic speed. In Figure 14, of 11 features the following 8 were selected: intensity, occupancy time, wind speed, pressure, load, average traffic speed, u component and v component.

After extracting the relative features using mutual information, the next step is to run BiConvLSTM. Table 2 shows the results. First of all, it can be seen that on including all the features the results of the first scenario outperformed the second scenario. However, the results of mutual information do not maintain the same trend. In particular, for the first scenario, mutual information deteriorated the results, but in the second scenario, mutual information improved the overall results. Considering also that after the implementation of mutual information in the second scenario two components, u and v, were chosen, and in the first scenario no category of wind direction was included, it can be concluded that wind direction is one of the important features for predicting NO_2. An additional finding is that, with all features included, the conversion of wind direction into categories and the subsequent implementation of One Hot Encoder outperformed the conversion to u and v components.

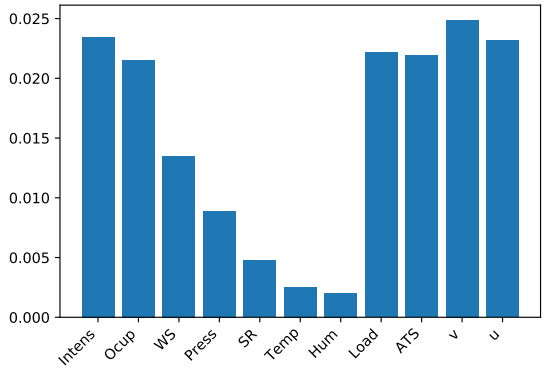

Figure 14. Feature selection using the mutual information technique (Wind direction with u and v components).

Table 2. RMSE and MAE of Scenarios I and II using BiConvLSTM (units in $\mu g/m^3$).

	All Features		Selected Features (MI)	
	RMSE	**MAE**	**RMSE**	**MAE**
Scenario I	18.99	12.89	26.92	20.00
Scenario II	24.87	16.49	22.32	16.89

[12]One Hot Encoder: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

[13]Wind: u and v Components: http://colaweb.gmu.edu/dev/clim301/lectures/wind/wind-uv

Regarding mRMR, the results are illustrated in Table 3 (first scenario) and Table 4 (second scenario). It can be seen that the results are significantly reduced. In the case of the first scenario, the best combination of the features is obtained when K=7 (RMSE–3.44, MAE–2.87). The selected features are: load, northwest direction, pressure, wind speed, average traffic speed, occupancy time and north direction. In the case of the second scenario, the best result was obtained when K=5 (RMSE–4.20, MAE–3.65). The selected features are load, pressure, wind speed, average traffic speed and occupancy time.

Table 3. RMSE and MAE of extracted features based on mRMR (K is the number of features) using BiConvLSTM (scenario I).

	RMSE (μg/m^3)	**MAE (μg/m^3)**
K=3	6.81	5.97
K=4	5.61	5.18
K=5	3.55	3.07
K=6	4.90	4.37
K=7	**3.44**	**2.87**
K=8	19.91	15.51

Table 4. RMSE and MAE of extracted features based on mRMR (K is the number of features) using BiConvLSTM (scenario II).

	RMSE (μg/m^3)	**MAE (μg/m^3)**
K=3	5.60	4.84
K=4	5.26	4.69
K=5	**4.20**	**3.65**
K=6	23.51	14.05
K=7	33.48	21.29
K=8	31.80	21.77

Following the outcome, it can be concluded that mRMR outperformed mutual information since this latter selects the most relevant features, while mRMR selects the relevant features with minimal redundancy. In addition, it is important to see which features were chosen and what these results are related to. In both cases, after implementing mRMR, the load was selected. Taking into account the definition of load (load is a combination of intensity, time of use and road capacity) and given the importance of traffic data for NO_2 production, the choice of this feature is obvious. The other features that yield better results are pressure, wind speed, average traffic speed and occupancy time. The last two features, as already mentioned, are chosen because of the importance of traffic data for NO_2 production. Regarding wind speed, as mentioned in the exploratory analysis, there is a strong correlation between wind speed and NO_2. Regarding the wind direction transformation, the u and v components were not included in the selected subsets after applying mRMR, although the northwest and north directions were included. The best subsets of the first scenario outperformed the second scenario, improving RMSE by 18.1% and MAE by 21.37%. Therefore, also in the case of implementing mRMR, it can be seen that wind direction conversion to categories surpassed the u and v conversion.

Regarding the overall results, it should be noted that 2020 was a year with certain peculiarities, namely the coronavirus (COVID-19) pandemic and its consequences, including traffic restrictions and self-isolation. It would be ideal to choose a period other than 2020 in order to avoid the impact of COVID-19 on the analyses.

6 Conclusions

NO_2 prediction is a critical task. Numerous factors influence the formation of NO_2. Among all the factors, it is very important to choose the best minimum factors that will help to predict the concentration faster and more accurately. There are many methods for feature selection. The results showed that combining machine learning methods with domain knowledge can produce better results.

This work has focused on the application of mutual information and maximum relevance-minimum redundancy, obtaining the most relevant features related to NO_2, and comparing the results of both methods. Another direction was the preprocessing of wind direction data. Two conversion methods have been implemented: converting the wind direction into u and v components or into categorical data. The results show that the conversion of the wind direction in One Hot Encoder is superior to the conversion to the u and v components. Regarding feature selection methods, it was found that the implementation of mRMR yields better results compared to mutual information, given the fact that mRMR, in addition to selecting relevant features, tries to select the next relevant feature that has a minimum correlation with already selected features.

Acknowledgements. Ditsuhi Iskandaryan has been funded by the predoctoral programme PINV2018 - Universitat Jaume I (PREDOC/2018/61) and by the Pla de promoció de la investigació a l'UJI (E-2020-14). Sergio Trilles has been funded by the Juan de la Cierva - Incorporación postdoctoral programme of the Ministerio de Ciencia e Innovación - Spanish government (IJC2018-035017-I).

References

Altman, N. and Krzywinski, M.: The curse (s) of dimensionality, Nat Methods, 15, 399–400, 2018.

Bui, T.-C., Kim, J., Kang, T., Lee, D., Choi, J., Yang, I., Jung, K., and Cha, S. K.: Star: Spatio-temporal prediction of air quality using a multimodal approach, in: Proceedings of SAI Intelligent Systems Conference, pp. 389–406, Springer, 2020.

Cleveland, W. S.: Robust locally weighted regression and smoothing scatterplots, Journal of the American statistical association, 74, 829–836, 1979.

Delmar-Morgan, E.: The Beaufort Scale, The Journal of Navigation, 12, 100–102, 1959.

Hanson, A., Pnvr, K., Krishnagopal, S., and Davis, L.: Bidirectional convolutional lstm for the detection of violence in videos, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0–0, 2018.

Huler, S.: Defining the wind: the Beaufort scale and how a 19th-century admiral turned science into poetry, Crown, 2007.

Iskandaryan, D., Ramos, F., and Trilles, S.: Bidirectional convolutional LSTM for the prediction of nitrogen dioxide in the city of Madrid, Under Review: PLOS One, 2022.

Just, A. C., Arfer, K. B., Rush, J., Dorman, M., Shtein, A., Lyapustin, A., and Kloog, I.: Advancing methodologies for applying machine learning and evaluating spatiotemporal models of fine particulate matter (PM2. 5) using satellite data over large regions, Atmospheric Environment, 239, 117 649, 2020.

Khomenko, S., Cirach, M., Pereira-Barboza, E., Mueller, N., Barrera-Gómez, J., Rojas-Rueda, D., de Hoogh, K., Hoek, G., and Nieuwenhuijsen, M.: Premature mortality due to air pollution in European cities: A health impact assessment, The Lancet Planetary Health, 2021.

Liu, C.-C., Lin, T.-C., Yuan, K.-Y., and Chiueh, P.-T.: Spatio-temporal prediction and factor identification of urban air quality using support vector machine, Urban Climate, 41, 101 055, 2022.

Liu, H. and Chen, C.: Spatial air quality index prediction model based on decomposition, adaptive boosting, and three-stage feature selection: A case study in China, Journal of Cleaner Production, p. 121777, 2020.

Masmoudi, S., Elghazel, H., Taieb, D., Yazar, O., and Kallel, A.: A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection, Science of The Total Environment, 715, 136 991, 2020.

Peng, H., Long, F., and Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on pattern analysis and machine intelligence, 27, 1226–1238, 2005.

Shah, J. and Mishra, B.: Analytical Equations based Prediction Approach for PM2. 5 using Artificial Neural Network, arXiv preprint arXiv:2002.11416, 2020.

Shi, J., Peng, D., Peng, Z., Zhang, Z., Goebel, K., and Wu, D.: Planetary gearbox fault diagnosis using bidirectional-convolutional LSTM networks, Mechanical Systems and Signal Processing, 162, 107 996, 2022.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting, arXiv preprint arXiv:1506.04214, 2015.

Song, H., Wang, W., Zhao, S., Shen, J., and Lam, K.-M.: Pyramid dilated deeper convlstm for video salient object detection, in: Proceedings of the European conference on computer vision (ECCV), pp. 715–731, 2018.

Stull, R. B.: Practical meteorology: an algebra-based survey of atmospheric science, 2015.

Sun, R. and Gu, D.: Air pollution, economic development of communities, and health status among the elderly in urban China, American journal of epidemiology, 168, 1311–1318, 2008.

Verleysen, M. and François, D.: The curse of dimensionality in data mining and time series prediction, in: International work-conference on artificial neural networks, pp. 758–770, Springer, 2005.

Xu, X. and Ren, W.: Prediction of Air Pollution Concentration Based on mRMR and Echo State Network, Applied Sciences, 9, 1811, 2019.

Yan, R., Liao, J., Yang, J., Sun, W., Nong, M., and Li, F.: Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering, Expert Systems with Applications, 169, 114 513, 2021.

Yang, Q., Yuan, Q., Yue, L., and Li, T.: Investigation of the spatially varying relationships of PM2. 5 with meteorology, topography, and emissions over China in 2015 by using modified geographically weighted regression, Environmental Pollution, 262, 114 257, 2020.

Zhao, S., Yin, D., Yu, Y., Kang, S., Qin, D., and Dong, L.: PM2. 5 and O3 pollution during 2015–2019 over 367 Chinese cities: spatiotemporal variations, meteorological and topographical impacts, Environmental Pollution, 264, 114 694, 2020.

Zhao, Z., Anand, R., and Wang, M.: Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform, in: 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 442–452, IEEE, 2019.

Zheng, H., Cheng, Y., and Li, H.: Investigation of model ensemble for fine-grained air quality prediction, China Communications, 17, 207–223, 2020.