Long Term Predictions of NO_2 Average Values via Deep Learning

Pierfrancesco Bellini, Stefano Bilotta, Daniele Cenni, Enrico Collini, Paolo Nesi(✉), Gianni Pantaleo, and Michela Paolucci

DISIT Lab, DINFO Dept, University of Florence, Florence, Italy
paolo.nesi@unifi.it
https://www.disit.org, https://www.snap4city.org

Abstract. Forecasting future values of air quality related metrics and specific pollutant concentration could be of pivotal importance in recent Smart City perspectives. A number of pollutants are dangerous for people's health and impact on environment and climate. In order to control and reduce the emissions, national and international organizations have defined guidelines and targeted limits to be respected currently, and to be progressively reduced along the year/months. On this regard, the European Union has set limits for the concentration of the yearly mean value of NO_2 which must not exceed 40 $\mu g/m^3$. To this end, in this paper, we propose a model and tool to compute long terms predictions, up to 180 days in advance, of the progressive mean value of NO_2 with a precision needed to enable decision makers to perform corrections. The solution proposed is based on machine learning approach taking into account measures of pollutant, traffic flow, weather and environmental variables coming from sensors on the field. A comparison of different techniques has been provided. The research activity has been developed in the context to TRAFAIR CEF project of EC which aimed to study the effect of traffic and of other human activities on NO and NO_2. The data and the solution have been developed by exploiting the Snap4City platform; the validation of the solution has been performed by using actual measured data from years 2014 to 2020 in the area of Florence, Italy. The results are accessible via a monitoring dashboard on Snap4City which reports real time values and predictions in real time.

1 Introduction

In the context of smart cities, tools for air quality monitoring are one of the main pillars. In recent years, the concentrations of air pollutants have reached critical values in the majority of industrialized cities over the world, and a large number of actions have been targeted by governs and international institutions to reduce them. Thanks to the development of Internet of Things (IoT) technologies, it has been possible to acquire useful data, for instance through air quality sensors, that can be used to develop real-time data analytics and predictive machine learning models. The search field of air quality predictions is experiencing an increasing interest due to its relevance. Despite this, the majority of related works are based on short-term predictions starting from hourly

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12956, pp. 595–610, 2021.
https://doi.org/10.1007/978-3-030-87010-2_44

values up to a few days from the prediction time. The majority of state-of-the-art tools to forecast future air pollutant concentrations are machine learning techniques that need consistent historical data representing the features that determine or influence the specific air pollutant, in order to be trained to predict future concentrations. In [15], it has been shown that this problem can be solved using multivariate data. Recently, Deep learning based methods have been proposed for air quality prediction, such as [9, 14]. The Long Short-Term Network (LSTM) model has been used in [16], in which several air quality pollutant factors have been predicted such as CO, NO_2, O3, PM10 in short term. In [13], the importance of taking into account meteorological and temporal features for the development of predictive models for air pollutants has been highlighted. Therefore, data analytic enabled a number of applications that range from the development of territorial heatmaps [4] to alerting systems when the NO_2 concentration reaches dangerous values. In this context, the TRAFAIR project (in which this research has been developed) has been focused on the short and mid-term prediction of NOX (NO + NO_2) on the basis of traffic flow and on other factors [6, 17].

Among the air pollutants, nitrogen dioxide (NO_2) can cause serious problems not only for people's health but also for the environment [19]. Exposure to NO_2 has been linked to increased mortality of a relative risk factor of 1.04 every 10 $\mu g/m^3$ in the annual NO_2 concentration [8]. The European Union has created a legislative program [7] in which the limits of air pollutants concentrations, to preserve people's health are specified. For the nitrogen dioxide, the maximum yearly mean concentration value is set to 40 $\mu g/m^3$. Other limits have been imposed on other pollutants, while the overcome of the mean value over the year 1seems to be the most effective in pushing cities towards the control and thus it forces them to the improvement of air quality. In fact, the reduction of the mean NO_2 leads to the corresponding reduction of other pollutants and of GHG (greenhouse gas) also provoked by traffic and heating in general (https://en.wikipedia.org/wiki/Greenhouse_gas).

This paper aims to present a system to assess and produce long terms predictions of the yearly mean NO_2 concentration. The yearly mean nitrogen dioxide concentration is typically assessed by the European Commission in the most critical points of the city, that are the major roads, and since it is a long term average, it is particularly complex to correct it by imposing last minute traffic restrictions. Being the metrics, a long term average over the year, it is hard to revert the trend by closing roads for a few days; even a drastic total closure per a number of days risks to create a marginal reduction on progressive mean. Our work differentiates from the above-mentioned papers in the field of air quality prediction because it does not focus on predicting the hourly or daily concentration of NO_2. We focused on long term prediction of the yearly mean value of the NO_2 concentration. In the presented work, in order to develop a monitoring, predictive tool and a dashboard to study the trend of the yearly mean value of NO_2 concentration, we have developed six predictive LSTM deep neural networks to forecast the future progressive mean values of nitrogen dioxide for 30, 60, 90, 120, 150 and 180 days ahead of the current day. The LSTM solution has been selected among a number of solutions compared in the paper. The work presented in this paper has been developed in the context of TRAFAIR CEF Project of EC, on the basis of the data collected in the city

of Florence. Indeed since 2014, the city of Florence has not respected the limit imposed by the EU for the mean value concentration of NO_2.

The paper is organized as follows. In Sect. 2, the description of data and selection of features are reported and discussed. Section 3 describes the production of the predictive models for mean progressive NO_2, and provides also the comparison with other machine learning techniques to demonstrate that the LSTM has been the better ranked. In Sect. 4, a description of the Real Time monitoring and prediction service set up on Snap4City are provided. Conclusions are drawn in Sect. 5.

2 Data Description and Feature Identification

Nitrogen dioxide, NO_2, is generated for the most part in the atmosphere for the oxidation of nitrogen monoxide (NO), which is produced by combustion processes, in particular by traffic of vehicles, heating houses, and industrial activities [1, 20]. However, other factors may influence NO_2 values. According to the introduction, this paper aimed to present a long term solution for predicting the yearly mean values of NO_2 concentration, in advance a much as possible with respect to the date in which the taxation to city may be produced. According to the European rules, the metrics is typically assessed at the end of the year, while the cities need to keep those metrics under control much in advance to take countermeasures on time. The progressive mean value in the specific year of study is calculated as the NO_2 mean concentration day after day, and dividing this by the number of passed days. The data for the problem as formulated, are structured as a time series. The city of Florence since 2014 have not respected the limit imposed by the EU for the mean value of NO_2, that has to be lower than 40 $\mu g/m^3$. The reference values have been estimated on the basis of the data acquired from sensor FI-GRAMSCI of regional agency of environment (Agenzia regionale per la protezione ambientale della Toscana, ARPAT) [1], recorded for year 2014 as yearly mean of 63.396 $\mu g/m^3$, for 2015: 65.173 $\mu g/m^3$, 2016 36.794 $\mu g/m^3$, for 2017: 63.396 $\mu g/m^3$, for 2018: 60.256 $\mu g/m^3$, for 2019: 56.111 $\mu g/m^3$, and for 2020 a value of 42.632 $\mu g/m^3$ (despite of the COVID-19). The data are accessible on the Snap4City infrastructure of DISIT lab https://www.snap4city.org [3, 5].

In order to build a long term predictive model, a number of features have been tested and relevant feature identified as described in the following. They refer to historical values of NO_2, traffic flow, weather conditions, heating conditions, etc. Therefore, a set of derived features have been computed according to the physical meaning of what we would like to predict, which is the progressive mean value of the measured variable in $\mu g/m^3$.

The most relevant predictor for the model is related to the traffic flow data. According to our analysis performed on ServiceMap (Snap4City) by DISIT lab, the position of the above mentioned ARPAT-Gramsci sensor for NO_2 has been analyzed. This sensor is positioned on "Viale Antonio Gramsci" in the city of Florence, and it detects the hourly mean value of NO_2 expressed in $\mu g/m^3$. The historical data cover the years starting from 2014. The time granularity is hourly. Using the same tool, the FI055ZTL00201 traffic sensor [18] has been considered. This sensor detects the number of vehicles that transit per hour in the same "Viale Antonio Gramsci" of Florence, in which the above mentioned

ARPAT-Gramsci sensor is also positioned. The prediction can only be performed on a single pollutant sensor since, that sensor is the sensor used by the EC to emit the taxation, and thus to take into account also other NO_2 sensors in other parts of the city has been demonstrated to be non relevant. As demonstrated in TRAFAIR, the pollutant are volatile and are moved by wind and thus a limit life 2on air. Figure 1 shows the position of the selected sensors.



Fig. 1. Positions of FI055ZTL00201 traffic sensor and ARPATGRAMSCI air quality sensor shown on ServiceMap tool of https://www.snap4city.org portal and service.

The historical data of traffic covers the years starting from 2014 with a granularity of 5 min. It should be noted that, the traffic flow sensors provide data in terms of traffic density and/or in terms of number of vehicles passed in the unit of time (vechicleFlow), which in our case was an hour. On the other hand, we need to have values which take into account the total amount of pollutant produced over the whole day because the NO_2 taken into account is a cumulative data. Therefore, we estimated the following features:

$$\text{numberOfVehicles}_i = \sum_{l=1}^{24} \text{vehicleFlow}_{i_l} \tag{1}$$

$$\text{numberOfVehiclesCumulated}_i = \sum_{k=1}^{i} \text{numberOfVehicles}_i \tag{2}$$

Another relevant aspect, which influences the traffic and thus the pollutant, is represented by the environmental conditions. To this end, meteorological data have been acquired through Snap4City and IlMeteo.it [12] for the city of Florence. The data provided by these resources comprehend various parameters: minimum, mean, and maximum temperature of the day expressed in °C; Dew point also expressed in °C; mean and maximum wind speed expressed in km/h; humidity of the day expressed in percentage; the air pressure of the day expressed in millibar (mb). These data covers the historical period of interest starting from 2014 with daily values. The environmental conditions influence the NO_2 production with the usage of the house heating, and propagation with wind. The former factor has been taken into account exploiting a formula derived

from TRAFAIR project [17]. Then, using the mean daily temperature, it is possible to determine, through a parametric formula, the domestic NOx produced in a day.

$$NO_xDomestic_i = \left(K + A * Tmedia_i + B * Tmedia_i^2\right) * 1000 \tag{3}$$

where: $K = 2.22488$, $A = -0.14828$, $B = 0.00276$ computed and validated in TRAFAIR. Please note that they impact on pollutant only during the winter period and for about 1/10 of the NO_2 produced. So that, it is a corrective factor. The second aspects directly considering the wind as possible features. In order to forecast the progressive mean values of NO_2 it has been necessary to apply some pre-processing operations. Firstly, the conversion to the chosen time granularity, that is daily. The last operation consisted of deriving the cumulated progressive mean features starting from the NO_2, traffic values, and also for the domestic NOx and deriving some temporal features from the DateTime of the prediction day (Date, Year, Month, dayOfTheMonth, dayOfTheYear, dayOfTheWeek, weekEnd, festivity, workingDay, ferialDay). Considering the year j and the i-th day of this year as the day of the prediction, in the next lines the detailed formulas used to obtain the derived features are reported.

$$NO_2Cumulated_i = \sum_{k=1}^{i} NO_{2_k} \tag{4}$$

$$NO_2progressiveMean_i = \frac{NO_2Cumulated_i}{i} \tag{5}$$

$$NO_XDomesticCumulated_i = \sum_{k=1}^{i} NO_XDomestic_k \tag{6}$$

$$NO_XDomesticCumulated_i = \frac{NO_XDomestic_k}{i} \tag{7}$$

The initial dataset taken into account is reported in Table 1 with the details of the features.

The aim was to create a set of long term predictive models to be used in a real-time process that every day of the current year may generate the prediction on the basis of collected data, and thus of the calculated features. The computed predictions shall be also be displayed in a monitoring dashboard on Snap4City, which can be used for the city control monitoring and forecast.

2.1 Feature Analysis

As above mentioned, the research aimed to identify a prediction model for the progressive mean values of nitrogen dioxide, NO_2. After having performed several testing, a machine learning technique has been chosen discharging the more classical ARIMA/SARIMA, ARIMAX approaches which are unsuitable for long term predictions. The process of feature analysis has been an important operation to reduce the dimension of the input space in terms of features, with the aim of selecting the most relevant features which can lead to generalize the model, reducing the eventual overfitting, and simplifying the computational architecture in real time. Moreover, this allows a significant reduction of

Table 1. Initial Data-set taken into account with details

Metric	Details
Date	UTC format of the day of prediction YYYY-MM-DD
Year	Of the observation {2014, …, 2020}
Month	Of the observation {1, …12}
dayOfTheYear	Day number in the year {1, …365/366}
dayOfTheMonth	Day number in the month {1, …31}
dayOfTheWeek	Day of the week {1, …, 7}
weekend	Saturday or sunday 1, 0 otherwis
festivity	Festivity 1, 0 otherwise
workingDay	Not a saturday or sunday and it is not a festivity
ferialDay	1 if the day is not a sunday or a festivity
NO_2	The NO_2 hourly mean of the observation day in $\mu g/m^3$
Tmin	The min temperature of the day in °C
Tmean	The mean temperature of the day in °C
Tmax	The max temperature of the day in °C
dewpoint	The dew point temperature in °C
windMean	The mean value of the wind of the day in km/h
windMax	The max value of the wind of the day in km/h
Humidity	The humidity of the day in %
pressioneSLM	The air pressure in millibar (mb)
NOx	The NOx value of the day in kg
numberOfVehicles	The number of vehicles of the day
NO_2cumulated	The cumulated value of NO_2 up to the day
NO_2progressiveMean	The progressive mean value of NO_2 up to the day
numberOfVehiclesCumulated	The number of vehicles cumulated up to the day
NOxDomesticCumulated	The cumulated value of NOx up to the day
NOxDomesticProgressiveMean	The progressive mean value of NOx up to the day

the computational time which is performed every day. A first analysis was performed using Scatterplots and correlation matrices which did not shown useful information on the linear correlation of the target feature (the progressive mean value of the NO_2), and thus have not been reported in this paper. In order to select the most relevant features, Principal Component Analysis (PCA) has been applied. The PCA is used for multivariate problems, feature engineering and for machine learning [2]. The results of the PCA have been reported in Fig. 2. The trade-off for the explained variance has let us to selected the first 5 Principal Components, the figure reported the first 7.

The first component, which is largely the most relevant as shown in Table 2, includes the progressive mean features of NO_2 and NOxDomestic with also the min temperature, the dew point, and the cumulated number of vehicles. In the second component, the features cumulated are the most relevant, with the mean and max temperature and the humidity. The third component includes the wind features; the fourth the air pressure; and in the fifth the daily NO_2 with the number of vehicles.

According to the above reported analysis, we performed a number of tests with the aim of identifying a compromise between complexity (in terms of number of features) and precision. The identified compromise for the predictive features of the models has been to use: NO_2, Tmean, humidity, windMean, NOxDomestic, numberOfV ehicles, NO_2cumulated, NO_2progressiveMean, numberOfV ehiclesCumulated, with associated Month and dayOfT heY ear for their identification in the time series.

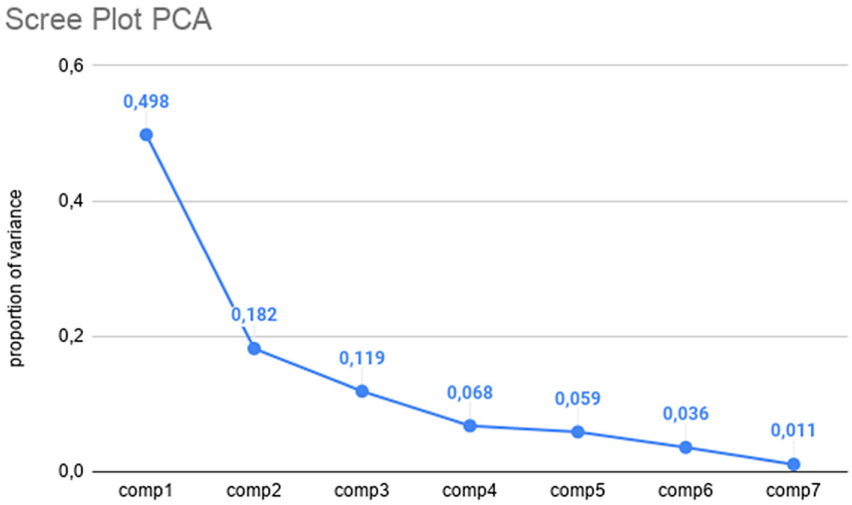


Fig. 2. Principal components with their corresponding explained variance.

Table 2. Principal components analysis (a part)

Parametro	comp1	comp2	comp3	comp4	comp5
NO_2	0.21492	0.03753	0.21523	0.12079	0.49583
NO_2cumulated	−0.29702	0.33402	−0.09504	−0.03905	0.03549
NO_2progressiveMean	0.31897	−0.25867	0.10213	0.05706	−0.04563
Tmin	−0.30745	−0.27795	0.06725	0.10825	0.08754
Tmean	−0.29595	−0.31203	0.16324	0.00393	0.13399
Tmax	−0.2687	−0.31676	0.24441	−0.06649	0.1375
dewPoint	−0.31326	−0.15871	0.17945	0.23102	0.04431
windMean.km.h	−0.00725	−0.28206	−0.6145	−0.14964	0.07701

(*continued*)

Table 2. (*continued*)

Parametro	comp1	comp2	comp3	comp4	comp5
windMax.km.h	−0.03454	−0.30142	−0.59137	−0.03938	0.09913
humidity	0.01218	0.43378	0.04680	−0.42877	−0.07932
pressioneSLM.mb	0.04822	−0.01663	0.18496	−0.91479	0.22794
numberOfVehicles	0.14502	0.16311	−0.12736	0.21015	0.78224
numberOfVehiclesCumulated	−0.29235	0.34455	−0.0991	−0.03161	0.03886
NOxDomestic	0.30408	0.27471	−0.06801	0.00842	−0.13415
NOxDomesticCumulated	−0.30356	0.30434	−0.11701	−0.04954	0.05715
NOxDomesticProgressiveMean	0.34165	−0.1894	0.07221	0.05133	−0.04398

3 Predictive Models

In order to create a reliable solution to compute predictions of the progressive mean value of NO_2 ahead, a number of models have been created. The questions to be solved have been: (1) it is possible to create a reliable predictive model for the progressive mean value of NO_2?, (2) how much in advance the model can provide acceptable predictions? Please note that, the literature provides only short terms predictions of NO_2 which are not useful to perform corrections in time, since the progressive mean is hard to be corrected as above explained. Therefore, we targeted the study to perform a set of predictive models aiming at computing reliable long terms prediction, for example 30, 60, 90, 120, 150, and 180 days in advance with respect to the current day. For this reason we have developed 6 specific LSTM models instead of one multi-output, which does not produced good enough results. The dataset available covered the years 2014, 2015, 2016, 2017, 2018, 2019, and 2020. However, the year 2020 reported a significantly different trend for the progressive mean values of the nitrogen dioxide (due to the restrictions for the Covid-19 pandemic), as it is shown in Fig. 3. For this reason, the data used for training have been those of the years 2014–2017. In addition, once the model has been obtained, it has been validated against the values of year 2018, and tested for precision assessment against data of 4year 2019, as reported in the following. We have also reported the results for year 2020 for completeness, but due to the COVID-19 they cannot be taken into account as a good example and validation of the model, as it can be observed in Fig. 3.

The LSTM model has been adopted [11] with its update [10]. Before selecting the LSTM we have tested a number of techniques as reported in Sect. 3.2. The architecture of the predictive models has been set to 3 layers with optimized hyperparameters for every temporal target through a Randomized Search CV:

- The first layer is made by 64 or 32 LSTM units
- The intermediate layer is a Dense layer with 64, 32 or 16 units with also a dropout rate of 0.1, 0.2, 0.5.
- The final layer has only one neuron to predict the selected time target.

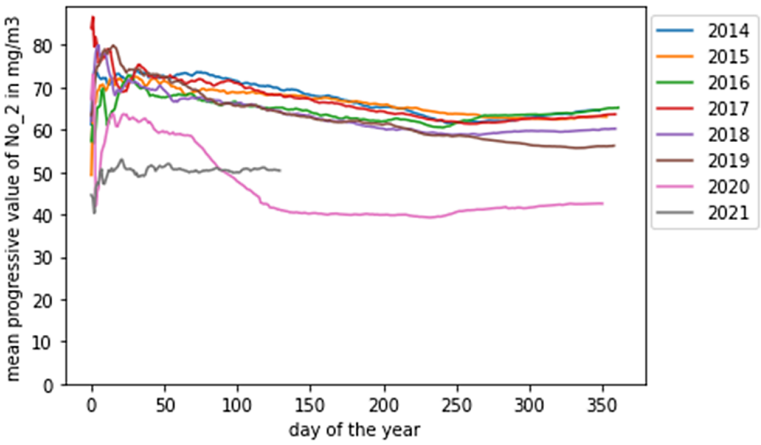


Fig. 3. Progressive mean trends for NO_2 in the years 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021 according to the ARPAT sensor adopted by the EC as a reference. Please note that NO_2 scale starts from 38 $\mu g/m^3$ on Y axis.

For each model, the Adam Optimizer has been chosen among learning rates 0.05, 0.005, 0.0005, or 0.00005. The considered loss has been assessed by using the Mean Squared Error (MSE). The batch size changed between 64 and 32; meanwhile the number of epochs has been set to a maximum value of 1000, while the training strategy used the Early Stopping method for determining the optimum epoch number which minimize the Medium Average Error (MAE) of the validation set, allowing also the restoring of the weights of the best model.

The input data for to the model have been organized through a multiple sliding window that contains the data of 20 days preceding the i-th day of prediction in the

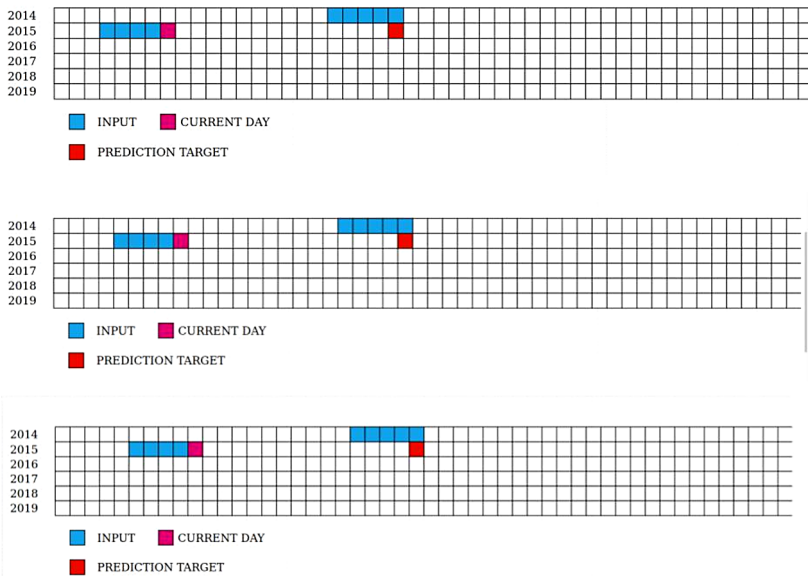


Fig. 4. Structure of the sliding window in the day (time instants): i, i + 1, i + 2.

considered year, and 20 days before the target day of the previous year. The structure of the multiple sliding windows approach is shown in Fig. 4.

The input features to the models are: Month, dayOfTheYear, NO_2, Tmean, humidity windMean,NOxDomestic, numberOfVehicles, NO_2Cumulated, NO_2progressiveMean, numberOfVehiclesCumulated.

Let's deepen into the structure of the model that predicts the progressive mean value of NO_2 30 days ahead. The hyperparameters for the 6 predictive models developed are reported in Table 3, which reports the 3 top model configurations resulted from the Randomized Cross Validation and in bold are highlighted the best ones.

Table 3. Hyperparameters for the models developed

negMSE	Model	LSTMunits	intermediateUnits	dropoutRate	learningRate	batchSize
−0.0017	30	64	64	0.1	0.00005	32
−0.0062	30	32	16	0.1	0.0005	64
−0.0077	30	64	16	0.5	0.05	32
−0.0031	60	32	64	0.1	0.0005	64
−0.0047	60	64	64	0.5	0.005	32
−0.0069	60	64	16	0.25	0.005	32
−0.0038	90	64	64	0.1	0.0005	32
−0.0049	90	64	64	0.1	0.0005	64
−0.0053	90	64	16	0.25	0.00005	32
−0.0059	120	64	64	0.25	0.005	64
−0.0092	120	32	16	0.1	0.005	64
−0.0116	120	32	32	0.5	0.0005	32
−0.0061	150	64	64	0.5	0.05	64
−0.0066	150	32	32	0.1	0.05	32
−0.0071	150	32	16	0.1	0.005	64
−0.0069	180	32	64	0.1	0.0005	64
−0.0124	180	64	64	0.25	0.005	64
−0.0158	180	32	64	0.5	0.05	64

The negative Mean Squared Error is used when minimizing the test score metrics of the combinations. The architectural structure of the LSTM neural network is visible in Fig. 5.

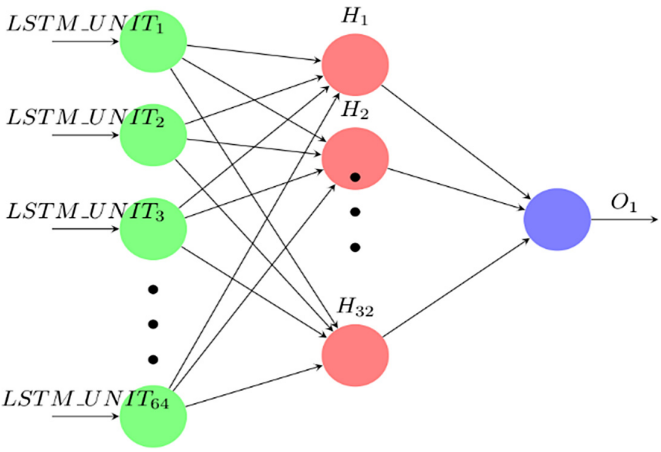


Fig. 5. Example of LSTM structure for the prediction model.

3.1 Experimental Results and Validation

The quantitative metrics used to evaluate the predictive models have been the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), the Mean Absolute Percentage Error (MAPE), and the R-Squared (R2) which is the coefficient of determination. The results obtained by using the test set of year 2019 are reported in Table 4, and the outputs are visible in Fig. 6.

Table 4. Assessment of the predictive models with respect to the actual values of the 2019

Metric	model30	model60	model90	model120	model150	model180
MAE	1.21	1.31	1.52	2.04	2.31	2.37
RMSE	2.16	2.61	4.18	6.77	7.83	7.93
MAPE	1.99	2.20	2.65	3.57	4.07	4.18
R2	0.91	0.83	0.80	0.54	0.45	0.14

According to the results, larger errors in predictions are obtained by the models which try to provide longer terms predictions. They start from a MAE of 1.21 $\mu g/m^3$ for the prediction of 30 days ahead up to 2.37 $\mu g/m^3$ for the model of 180 days ahead. In percentage, these results correspond to the 1.99% for the 30 days predictions and to the 4.18% for 180 days predictions. Please note that the precision in prediction is acceptable even in the worst case, since the error is very low with respect of the 40 $\mu g/m^3$ of the reference value of the EC. We can state that obtaining a prediction 180 days in advance allows decision makers to put in place the needed measures to correct the NO_2 trend of the city.

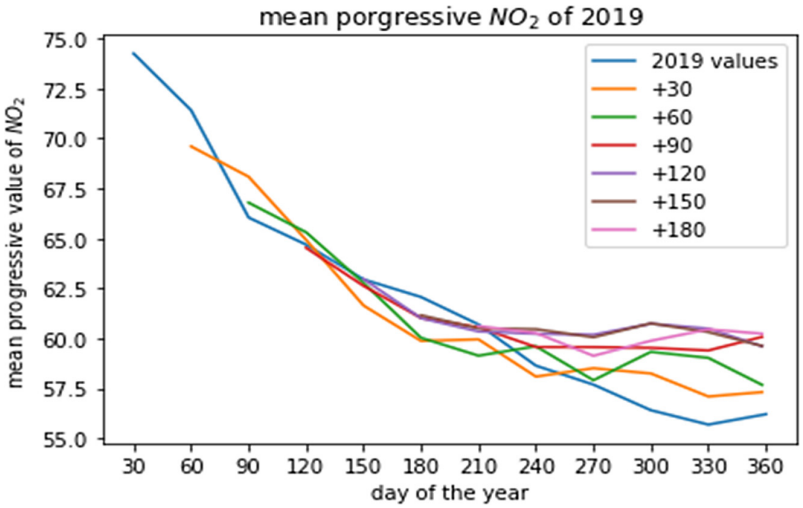


Fig. 6. Trend of the 6 predictive models with respect to the actual progressive mean NO_2 values of year 2019. Please note that NO_2 scale starts from 55 $\mu g/m^3$ on Y axis.

The results of the models for the year 2020 are visible in Fig. 7. Year 2020 has been a particular one due to the COVID-19 pandemic and the lockdowns that changed the volume of traffic vehicles, the amount of heating, and thus also the NO_2 concentration has been substantially changed, as it can be observed by its temporal trend with respect to the typical trends of the previous years. For this reason, the model puts in production

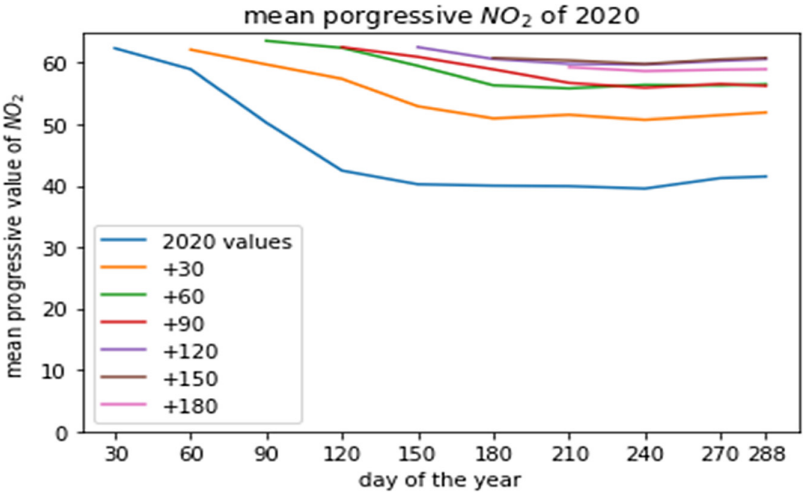


Fig. 7. Trends of the 6 predictive models with respect to the actual progressive mean NO_2 values of year 2020.

to generate the dashboard data for the year 2021 uses sliding windows with learning model based on 2019 and past data, instead of on those of the previous year.

3.2 Model Comparison

The sliding window approach proposed for past data has been tested with also other machine learning techniques.

The first comparison has been with a deep neural network, DNN, with the same number of layers as the LSTM architecture described in the previous section. Also, the hyperparameters tuning has been the same as the one described in Sect. 3.1 with the same values but of course, the units of the layers are not LSTM units but the ones used for the DNN.

Our approach has been tested with also ensemble learning techniques. In particular, the techniques chosen have been Random Forest (RF) and Extreme Gradient Boosting machines (XGBoost).

Regarding the implementation of the ensemble learning 6techniques the number of trees parameter for the RF has been set to 300, with min sample split set to 2, min number of samples allowed for a leaf equal to 1, without limits on the maximum number of features considered to split a node and the number of leafs and with the construction of bootstrapped datasets for creating the trees. The XGBoost regressor uses the least-squares loss function with learning rate optimized with values 0.1, 0.01, and 0.001 with max depth equal to 3 and min sample spilt, min sample leaf, max number of features equal to the ones chosen for the RF.

The results of these techniques vs LSTM have been compared in term of MAPE, Mean Absolute Percentage Error results for the prediction targets of 30, 60, 90, 120, 150, and 180 days. The results are reported in Table 5.

Table 5. MAPE for the prediction targets of 30, 60, 90, 120, 150 and 180 days resulted for the machine learning techniques LSTM, DNN, RF, and XGBoost

Target day	LSTM	DNN	XGBoost	RF
30	2.16	4.87	5,26	5,26
60	2.61	6.67	6,52	6,56
90	4.18	7.00	7,64	7,76
120	6.77	6.86	8,81	8,93
150	7.83	8.99	9,35	9,40
180	7.93	9.25	9,90	10

The results proved that the proposed LSTM approach outperforms the other techniques presented in terms of MAPE for every prediction target. For the 30 days prediction, the LSTM performs an error of about 2% compared to the 5% of the other techniques and performs better for the other targets up to the 180 days target with a MAPE of 7.93% where the others recorded MAPEs greater than 9%.

4 Dashboard for Real Time Monitoring and Prediction

As presented in previous section, substantially, we provided positive answers to the above reported two questions, regarding feasibility of the predictive model and the capability of providing predictions in advance enough to be exploitable by decision makers. On the other hand, a daily tool to visualize the results generated by the predictive models for the progressive mean NO_2 has been developed as a monitoring dashboard. It has been realized by exploiting the facilities of the DISIT Lab with its Snap4City Dashboard Builder of Snap4City https://www.snap4city.org [5]. The dashboards can easily exploit data collected in real time by the platform, as well as real-time rendering of results generated by Node-RED processes, which are called IOT Apps. The IOT App process, Node-Red/node.JS, exploiting Snap4City MicroServices, is executed daily and uses a Python script to generate the inputs for the predictive models and make the predictions for every temporal target. The dashboard developed is reported in Fig. 8 and it is made of three areas:

- The first contains the trends of the predictions made from the current day of the years for the temporal targets of 30, 60, 90, 120, 150 and 180 days ahead. These are reported through a bar series plot with a color scheme that darkens as the time target increases.
- The second contains a temporal multi series plot that shows the whole set of prevision models with respect to the trend of the actual values. In the plot, the horizontal green line is the EU limit value of 40 $\mu g/m^3$.
- The third on the right shows a heatmap of the NO_2 in Florence with the possibility to show the position of air quality and traffic sensors. It is also possible to monitor the trends of the traffic, NO_2, and of many other pollutants of the last weeks and months and years from the current day.

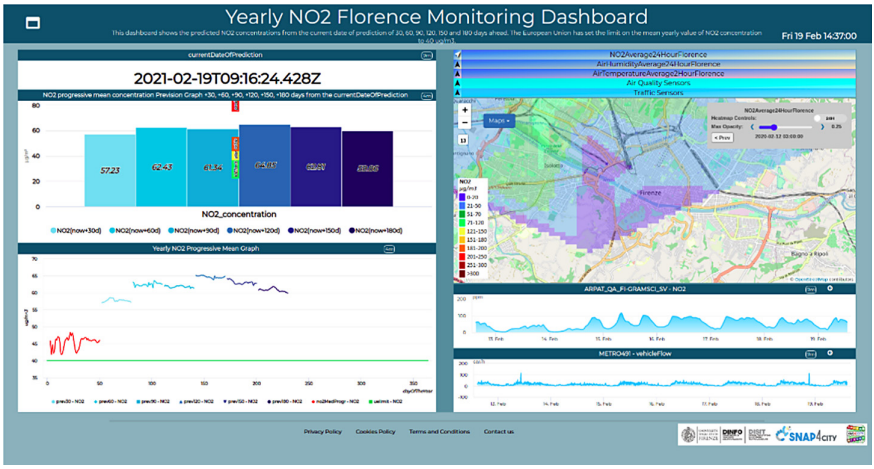


Fig. 8. Monitoring dashboard reporting real time value, prediction and actual cumulative values of NO_2 of the sensors considered for the official metrics.

The dashboard is accessible through any browser using the following link: https://www.snap4city.org/dashboardSmartCity/view/index.php?idasboard=MzA2OQ.

5 Conclusions and Future Developments

A number of pollutants are very critical for people's health, environment and climate. In order to reduce the emissions, national and international organizations have defined guidelines and targeted limits to be respected. On this regard, the European Union has set limits to the concentration for the yearly mean value of NO_2 which must not exceed 40 $\mu g/m^3$. In this paper, we described a model and tool to compute long terms predictions for 30, 60, 90, 120, 150, and 180 days ahead from the current day of estimation. The access to reliable long term predictions may allow decision 7makers to perform corrections. These models has produced results on a test set composed with the data of 2019, starting from a MAE of 1.21 $\mu g/m^3$ for the 30 days ahead prediction up to a 2.37 $\mu g/m^3$ for the 180 days prediction, that in percentage corresponds to 1.99% and 4.18%. The solution proposed is based on LSTM approach of deep learning taking into account measures of pollutant, traffic and environmental variables coming from sensors on the field. The LSTM solution has been demonstrated to be better ranked with respect to DNN, RF and XGBoost. The research activity has been developed in the context to TRAFAIR CEF project in which aimed to study the effect of traffic and other human activities on the NO and NO_2. The data and the solution have been developed exploiting the Snap4City platform, and the validation of the solution has been performed by using actual data from 2014 to 2020 in the area of Florence, Italy, from a large number of features and not only historical data. The results are accessible via a monitoring dashboard on Snap4City which report real time values and the predictions in real time. The approach presented in this paper can be further applied to the other air pollutants like PM2.5, PM10, CO, for which the EU has set yearly limits on their concentrations, and of course, this can be applied to other Smart Cities scenarios whenever the available data cover a sufficient historical range.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable.

The following bibliography provides a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4], as well as a URL [5].

References

1. Agenzia regionale per la protezione ambientale della Toscana. Biossido di azoto (2021). http://www.arpat.toscana.it/temi-ambientali/aria/monitoraggio/inquinanti-monitorati/biossido-di-azoto
2. Azid, A., Juahir, H., Latif, M.T., Zain, S., Osman, R.: Feed-forward artificial neural network model for air pollutant index prediction in the southern region of peninsular Malaysia. J. Environ. Prot. **4**(1), 1–10 (2013)
3. Badii, C., Bellini, P., Difino, A., Nesi, P.: Smart city IoT platform respecting GDPR privacy and security aspects. IEEE Access **8**, 23601–23623 (2020)

4. Badii, C., et al.: High density real-time air quality derived services from IoT networks. Sensors **20**(18), 5435 (2020)
5. Bellini, P., Cenni, D., Marazzini, M., Mitolo, N., Nesi, P., Paolucci, M.: Smart city control room dashboards: big data infrastructure, from data to decision support. J. Vis. Lang. Comput. **4**, 75–82 (2018)
6. Bilotta, S., Nesi, P.: Traffic flow reconstruction by solving indeterminacy on traffic distribution at junctions. Future Gener. Comput. Syst. **114**, 649–660 (2021)
7. European Commission: European union commission air quality standards (2020). https://ec.europa.eu/environment/air/quality/standards.htm
8. Faustini, A., Rapp, R., Forastiere, F.: Nitrogen dioxide and mortality: review and meta-analysis of long-term studies. Eur. Respir. J. **44**(3), 744–753 (2014)
9. Freeman, B.S., Taylor, G., Gharabaghi, B., The, J.: Forecasting air quality time series using deep learning. J. Air Waste Manage. Assoc. **68**(8), 866–886 (2018)
10. Gers, F., Schmidhuber, J., Cummins, F.: Learning to forget: continual prediction with LSTM. Neural Comput. **12**, 2451–2471 (2000)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
12. IlMeteo.it: Dati meteorologici forniti da ilmeteo.it (2021). https://www.ilmeteo.it/meteo/firenze
13. Iskandaryan, D., Ramos, F., Trilles, S.: Air quality prediction in smart cities using machine learning technologies based on sensor data: a review. Appl. Sci. **10**(7), 2401 (2020)
14. Xing, J., et al.: Deep learning for prediction of the air quality response to emission changes. Environ. Sci. Technol. **54**, 8589–8600 (2020)
15. Masih, A.: Machine learning algorithms in air quality modeling. Glob. J. Environ. Sci. Manage. **5**(4), 515–534 (2019)
16. Navares, R., Aznarte, J.L.: Predicting air quality with deep learning LSTM: towards comprehensive models. Ecol. Inform. **55**, 101019 (2020)
17. Po, L., et al.: TRAFAIR: understanding traffic flow to improve air quality. In: 2019 IEEE International Smart Cities Conference (ISC2), pp. 36–43. IEEE (2019)
18. Regione Toscana: Regione toscana - osservatorio dei trasporti (2021). http://www501.regione.toscana.it/osservatoriotrasporti/
19. Sheng, Q., Zhu, Z.: Effects of nitrogen dioxide on biochemical responses in 41 garden plants. Plants **8**(2), 45 (2019)
20. Wikipedia Contributors: Nitrogen dioxide wikipedia the free encyclopedia (2020)