Applying Machine Learning to Weather and Pollution Data Analysis for a Better Management of Local Areas: The Case of Napoli, Italy

Lelio Campanile[1], Pasquale Cantiello[2], Mauro Iacono[1], Roberta Lotito[1], Fiammetta Marulli[1] and Michele Mastroianni[1]

[1]*Dipartimento di Matematica e Fisica, Università degli Studi della Campania "L. Vanvitelli", viale Lincoln 5, Caserta, Italy*

[2]*Osservatorio Vesuviano, Istituto Nazionale di Geofisica e Vulcanologia, via Diocleziano 328, Napoli, Italy*

Keywords: Air Quality, Forecasting, Machine Learning, Regression, Data Analysis, Campania.

Abstract: Local pollution is a problem that affects urban areas and has effects on the quality of life and on health conditions. In order to not develop strict measures and to better manage territories, the national authorities have applied a vast range of predictive models. Actually, the application of machine learning has been studied in the last decades in various cases with various declination to simplify this problem. In this paper, we apply a regression-based analysis technique to a dataset containing official historical local pollution and weather data to look for criteria that allow forecasting critical conditions. The methods was applied to the case study of Napoli, Italy, where the local environmental protection agency manages a set of fixed monitoring stations where both chemical and meteorological data are recorded. The joining of the two raw dataset was overcome by the use of a maximum inclusion strategy as performing the joining action with "outer" mode. Among the four different regression models applied, namely the Linear Regression Model calculated with Ordinary Least Square (LN-OLS), the Ridge regression Model (Ridge), the Lasso Model (Lasso) and Supervised Nearest Neighbors Regression (KNN), the Ridge regression model was found to better perform with an R_2 (Coefficient of Determination) value equal to 0.77 and low value for both *MAE* (Mean Absolute Error) and *MSE* (Mean Squared Error), equal to 0.12 and 0.04 respectively.

1 INTRODUCTION

Since the beginning of the Industrial Revolution, one of the most affected environmental sector is the air: indeed, exhausted gases from human activities have changed the local atmosphere composition and this variation led to a relationship change between human and local area. It is an instilled thought that near industrial area or even in high urbanization area the air quality is poor, while, in order to 'breathe fresh air', it is necessary to go to a natural place like seafront or mountains. As matter of fact, the connection between the presence of some compounds, their concentration and the onset of specific disease is widely studied. The direct consequences is the different use of the territory with economic implication.

The natural atmospheric composition is well known: nitrogen accounts for 78%, oxygen 21% and argon 0.9%. Gases like carbon dioxide, nitrous oxides, methane, and ozone are trace compounds that account for about a tenth of one percent of the atmosphere. The presence and concentration of the trace gases can be characteristic for particular area (volcanic areas with the presence of hydrogen sulphide, rice paddies with methane presence): the variation of these compounds, both qualitatively and quantitatively, from their standard presence is called pollution. Actually, in consideration of the human activities and aerial transport/dispersion, it is nearly impossible to find a location on Hearth unpolluted. However, it is clear that the pollution from a city center is different compared both to an industrial area and to a local mountain village. In addition, even between two similar cities, the atmospheric pollution can be completely different due to meteorological asset, city architecture, regional location. Generally, we can say that the air quality depends on two classes of influence: the first regards the natural condition such as local weather, presence of water body, presence of emitting geological structures and so on. The second class regards the human-effect presence and so

Campanile, L., Cantiello, P., Iacono, M., Lotito, R., Marulli, F. and Mastroianni, M.
Applying Machine Learning to Weather and Pollution Data Analysis for a Better Management of Local Areas: The Case of Napoli, Italy.
DOI: 10.5220/0010540003540363
In *Proceedings of the 6th International Conference on Internet of Things, Big Data and Security (IoTBDS 2021)*, pages 354-363
ISBN: 978-989-758-504-3

it is related to all the human activities that emit in the atmosphere. Looking at a micro-scale area, new factors can be added: for example, the city architecture can influence the wind direction and speed resulting in different dispersion scenario or the implementation of a green area can decrease local contamination.

In order to restrain the continuous pollution and to try restoring some area to a better status, worldwide regulations have been issued: most of them imposed concentration limits both for outdoor air regarding pollutants like carbon monoxide, BTEX (Benzene - Toluene - Ethylbenzene - Xylene), nitrogen oxides, particular matter, and limit for specific emitting sources as industrial plants and human activities with the use of organic solvents. The overall strategy is to limit the emitting source where and when it is possible, and to check the air quality as result of the previous described factor. With the collected data, logged from stationary and mobile station, there is the possibility to assess the air quality by the use of some indicators based on the detected concentration of specific pollutant. A similar indicator is used in Europe, namely *European Air Quality Index* (EAQI), as established by 2008/50/CE directive. The hourly index is based on concentration values for up to five key pollutants and it reflects the potential impact of air quality on health, driven by the pollutant for which concentrations are poorest due to associated health impacts. The data are collected by stationary stations managed by local authorities.

In Italy, the transposition of the European directive took place with the enactment of law D.lgs. 155/2010 that has established a unified regulatory framework for the assessment and management of ambient air quality. Regions are assigned the responsibility to assess this quality, to classify the regional territory into zones and agglomerations, and to draw up plans and programs to maintain ambient air quality where it is good and to improve it in other cases. The national law imposes limits to outdoor pollutants concentration and, since the private transport sector has been identified as the major contributor to city pollution, in case of exceeding daily limits, the city administration restricts the access for private transport.

In this scenario, the implementation of forecasting modeling systems have become increasingly important in order to understand the future impacts of the human activities and to manage local areas. Forecasting can both be applied to the new emitting sources in order to understand their relative impact on local air, and directly to outdoor air quality to understand its development. It is clear that in the first case the problem resolution is easier since all the factors that affect the final result are well known (characteristics of the emitting source, pollutant concentration, plant layout, etc.). In the second case, the factors that took part in the game are various and not always so well defined: indeed, the local impact detected by stationary stations is due to a series of events such as particular wind direction, local traffic, presence of a new apartment blocks and so on. Hence, in the situation of a micro-scale forecasting, the boundary between the influence classes for the air quality is very blurred. To help solve this problem, the use of machine learning techniques seems to be a promising practice.

In the last years, many scholars have studied the implementation of forecasting modeling with machine learning: the results may significantly vary, depending on the used dataset and the implementation made. The machine learning help is based on the assumption of a black box mechanism for the air quality: the forecast is essentially based on the 'training' on a specific dataset, which results in the extrapolation of a statistical set of rules that can be applied to the newly collected data. Globally, the shown trends indicate an improvement in the forecast on an extended area, such as a region, or at national level, with high level datasets. The forecast buffer time can also vary according to the used mechanism.

In this paper, an application of a forecasting modeling approach implemented by a machine learning based technique is presented for an Italian city where air quality is assessed by means of stationary stations controlled by local authorities according to D.lgs. 155/2010. The aim of this research is to understand if this application can lead to a good forecast on a focused area with a few analyzing stations and local weather stations in order to better manage the area before the limits imposed are exceeded. The main original contribution is the application of this kind of analysis on combined official pollution and weather data about Campania region: at the best of our knowledge, no such analysis is available in literature. In addition, at the moment as per practice the data collected from the station are firstly validated by a third part before they are used for forecasting purposes. Indeed, this quality check and assurance (QC/QA) is an essential phase and it is usually handmade by few technicians. For these reasons, it could easily be affected by errors and hence data loss. Consequently, for this research we only used raw data in order to check how they perform without any preliminary screening.

After this section, the paper is organized as follows: next section presents related work, and a brief background on this field is summarized; then the case study and the used dataset are described; subsequently, the methodology used to develop the forecasting model by means of machine learning; results

and discussion close the paper, together with future work and developments.

2 RELATED WORKS

Several type of forecasting methods have been discussed in literature, regardless of a specific context as in (Chatfield, 2000) or (Hyndman and Athanasopoulos, 2018). Forecasting methods can be in general divided into three main categories: those that only deal with expert opinion, those based on physical models (Brandt et al., 2001), and those that instead make predictions based on a statistical analysis of values in a series (Armstrong, 2001).

A discussion based on published reviews and recent analyses about challenges, applications and advances, main gaps and trends along with research needs for atmospheric composition and air quality modeling and forecasting can be found in (Baklanov and Zhang, 2020).

A model to predict emission concentrations of PM_{10}, SO_2, O_3 for a selected number of forward time steps is proposed in (Domańska and Wojtylak, 2014) and named e-APFM. It requires large historical data series for weather forecast, meteorological and pollution data enriched with information about the wind direction in sectors and meteorological stations.

Wind strength and direction is a key feature in the propagation of pollutants. In (Contreras and Ferri, 2016) several regression models have been compared to be able to predict the levels of four different pollutants (CO, NO_2, SO_2, O_3) in several locations of a city. Different techniques to incorporate wind strength and direction in the regression models have been studied and applied within an interpolation method.

A recent review on the intelligent modeling strategies in the air quality forecasting has been published (Liu et al., 2021) with the summarizing of representative and efficient predictive models along with some possible research directions of the air pollution forecasting and monitoring (Campanile et al., 2020).

A feature ranking method to forecast multiple air pollutants simultaneously over two cities is proposed in (Masmoudi et al., 2020). It is based on a combination of an ensemble method for Multi-Target Regression problems and the RandomForest permutation importance measure, so allowing to obtain good results even with a restricted subset of features.

A spatiotemporal air pollution analysis that involves large geographical areas and spans over a long time period can be surely classified as a big data problem. In (Tong, 2020) the state-of-the-art machine learning and deep learning methods are introduced for the generation of more accurate estimations of $PM_{2.5}$ concentration.

A framework for air pollution monitoring and forecasting that combines deep learning and domain-decomposition techniques to allow model deployment extending beyond the domain(s) on which it has been trained is presented in (Haehnel et al., 2020).

Neural networks have been applied in pollution forecasting: AirPoolTool, an online tool using neural networks, publishes +1, +2, and +3 days predictions of air pollutants updated twice a day (Kurt et al., 2008); a deep multi-output LSTM (DM-LSTM) neural network model incorporated with three deep learning algorithms is presented in (Zhou et al., 2019); a model using Artificial Neural Networks (ANN) to forecast pollutant concentration in a highly polluted region, trained using meteorological parameters equipped with real time correction is presented in (Agarwal et al., 2020); an approach for particulate matter ($PM_{2.5}$) prediction for Delhi with both Support Vector Machines and ANN is described in (Masood and Ahmad, 2020); a parameterised non-intrusive reduced order model (P-NIROM) based on proper orthogonal decomposition and machine learning methods has been developed to model reduction of pollutant transport equations in order to build a rapid response tool for regional air pollution modeling (Xiao et al., 2019).

Pollution forecasting can be improved by using real-time data from sensors: a wireless sensor network that gathers pollutant concentrations has been used in Bengaluri city in South India (Belavadi et al., 2020), and IoT-based techniques with vehicles equipped with sensors embedded have been experimented (Shetty et al., 2020) that dynamically help the prediction of pollution level in the vehicle location based on the previous and current data.

From all the reviewed works it is clear that, in order to achieve a good pollution forecasting, it is mandatory to combine emission values detected by sensors with meteorological conditions, in particular wind strength and direction. The origin and the quality of acquired data is also a key factor for the success.

3 THE CASE STUDY

Regione Campania, the authority governing the homonym region located in southern Italy, according to D.Lgs. 155/2010 has implemented an air monitoring network by using mobile and stationary stations. After the last upgrade, the network configuration includes 36 fixed monitoring stations and 5 mobile lab-

oratories directly operated by the local environmental protection agency (ARPAC) plus 6 more fixed stations owned by third parties. The location of each station was defined in order to have a capillary disposition on the whole area and, at the same time, to cover all the sensitive receptors: hence, the stationary stations are usually located on the roofs of schools, hospitals and so on. Mobile laboratories are used accordingly to the necessities. Besides this network, there is another one, entirely devolved to analyse air quality near waste treatment plants: in this work this second network is not taken into consideration. The stations analyse the pollutant prescribed by D.Lgs. 155/2010 based on their locations: generally, the pollutants are nitrogen oxide, carbon monoxide, particular matter (P.M. both 10 and 2.5), ozone, benzene and sulphur dioxide. Data are collected hourly and then validated by applying national guidelines: after the validation process, the dataset is used for different purpouse by ARPAC while the raw data are available to the public. In addition, since 2018, some of the stations have been equipped with meteorological station to collect site-specific information. Collected data are basically used by the European Commission to create a specific air quality map (European Environmental Agency,).

Actually, data are also used to implement a forecasting system that is the base of official reports disclosure. ARPAC uses mathematical models to predict the spatial and temporal distribution of pollutants over a given area. In this context, the behaviour of the atmosphere in its layers in contact with the ground is decisive, where convective motions, thermal inversions and turbulence develop as a result of solar radiation and terrestrial re-irradiation. Meteorological monitoring is managed by CEntro MEtereologico e Climatologico (CEMEC) by means of the same stations, which implement specific sensors. The forecasting analysis is developed for a time window of three days and it has a low resolution, meaning that the resulting isopleths (curves showing the same pollutants concentration) cover a large area (Figure 1).

For this work, the first check was made on the open data regarding the concentration and the meteorological specifics. The open data related to pollutants concentration span since 2006 to 2018 (ARPAC,), while data about meteorological conditions are available since 2018 (Centro Meteorologico e Climatologico,). In addition, not all the stations are able to determinate the required pollutants: indeed, the sensors need to be checked frequently, hence they can be out of use for maintenance or, essentially, may be not designed to cover all kinds of analyses.

The spatiotemporal discrepancy is analysed in order to choose the boundary system. By matching the

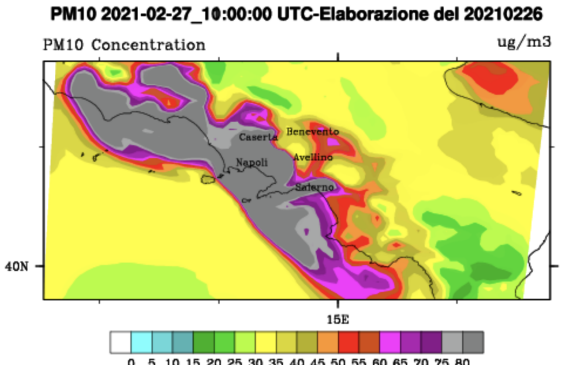

Figure 1: Screenshot of the hourly forecasting of PM_{10} made by ARPAC for February 26th, 2021. The area analysed was all the Campania region. The major cities are identified with the black dots. In order to give a distance set-up, it is highlighted that the distance Napoli - Caserta is 30 km circa. The prediction is calculated for three days.

two datasets, it is clear that the discrepancy is overcome only since 2018 and only for those stations that have recorded meteorological data. Figure 2 shows the location of all the stations that provide both meteorological and concentration data and in Table 1 stations coordinates are reported.

We decided to work only on the stations in the city of Napoli, so to restrict the area to be studies, aiming at getting a better resolution for that specific city. The chosen stations are reported in Table 2 with the specification of the pollutants analysed by each of them.

Figure 2: Satellite image of Napoli gulf (south-west), part of Salerno gulf (south) and Piana Campana (north). The red dots represent the locations of the stations with both meteorological and concentration acquisition systems.

4 METHODOLOGY

One of the major concerns in any experiment which tries to extract information from raw data, to obtain

Table 1: Stations locations.

Station Code	Name	Long	Lat
IT0898A	NA06	14.25	40.85
IT0934A	BN32	14.78	41.13
IT0936A	AV41	14.78	40.91
IT1491A	NA07	14.27	40.85
IT1493A	NA09	14.35	40.85
IT1504A	SA22	14.77	40.68

Table 2: Specific pollutants analysed by single station. T stands for true, hence the contaminant is analysed. F stands for false, so the contaminant is not analysed.

Name	$C_6\,H_6$	CO	NO_2	O_3	SO_2
NA06	T	T	T	F	F
BN32	F	F	T	F	F
AV41	T	F	T	T	F
NA07	T	T	T	F	T
NA09	T	T	T	F	F
SA22	F	F	T	F	F

hidden information and make predictions by machine learning algorithms or other techniques, is the data collection phase.

Data used in this experiment have been collected by ARPAC fixed stations located in the area managed by Regione Campania; those data belong to two main types of information: air pollution data and weather data. Once acquired, those datasets should be merged to have an unique dataset that includes all the values. The resulting dataset is a good starting point for data analysis and for applying regression models to obtain predictions.

Besides the P.M., which is not revealed by these stations, another important parameter to consider and worth to predict is the *CO*: like the other compounds, the carbon monoxide is used to classify the regional area but, in contrast with the others, its average benchmarks are calculated every 8 hours, hence the precision of the forecasting should be superior than that of the forecasting of a pollutant with an average benchmark equal to 1 year.

4.1 Dataset Description

The air pollution dataset is downloaded from the official ARPAC website (ARPAC,). The downloaded dataset is in csv (comma separated values) format and contains the entire acquired data volume. It is an aggregate file that contains data from all sensors and all stations collected hourly for year 2018.

The format used for the measurements in the csv file is rather inconvenient for data analysis, as for each sensor and station a different row is present. We have transformed the dataset to obtain a file which has all sensors value for each hour on the same line in order to make the dataset more compact and manageable.

Parameters in the dataset are summarized in Table 3: the relevant difference in the number of measurements among the parameters depends on stations, as stations only perform some measurements, like described before.

Table 3: Parameters of air pollution dataset.

Parameter	Description	Unit	n.
$C_6\,H_6$	Benzene	mg/m^3	29,205
CO	Carbon monoxide	mg/m^3	23,199
NO_2	Nitrogen dioxide	mg/m^3	35,903
O_3	Ozone	mg/m^3	3,907
SO_2	Sulfur dioxide	mg/m^3	7,786

Weather data are downloaded from CEMEC website. Differently from the previous case, in this one a data file for each day must be downloaded. Each file contains the measurements of a sensor and station per line. Before applying the same transformation performed on the air pollution dataset, all files have been merged in order to have an unique and continuous dataset.

Parameters in this dataset are summarized in Table 4.

Table 4: Parameters of weather dataset.

Parameter	Description	Unit	n.
AlbeInf	Lower albedo	W/m^2	55,031
AlbeSup	Upper albedo	W/m^2	55,031
Rainfall	-	mm	139,601
RadSG	Day radiation	W/m^2	92,997
RadSN	Night radiation	W/m^2	92,993
Temperature	-	C	155,250
Humidity	-	%	141,235
Pressure	-	hPa	134,735
UVA	-	W/m^2	55,507
UVB	-	W/m^2	55,529
Wind direction	-	degrees	70,149
Wind speed	-	m/s	70,232

4.2 Data Fusion

Once we had the raw data of both datasets available in a manageable form, to maximize possible exploitation the weather and the air pollution databases were merged on same common information: the date and time value of the measurement and the identifier of the station which measured them. This part is crucial to preserve data correctness and to expand the useful information content of the datasets.

Performing this transformation required coping with the imbalance in the number of values available for the parameters in both datasets. To address this issue we decided to utilize a maximum inclusion strategy by performing the joining action with "outer" mode to preserve the largest quantity of data and avoid data loss.

If, on one hand, this procedure has allowed to benefit of all available data, on the other hand it has added a new issue: the resulting dataset has a large number of missing elements caused by the missing values for same parameters in the source datasets and amplified by the magnitude effect of Cartesian product performed in the merging process.

4.3 Data Analysis

The data analysis of the new comprehensive dataset has been executed mainly in Jupyter Notebook environment using the Python programming language and the Pandas library.

The first step was oriented to obtain an insight into the properties of each attribute of the dataset; the table in figure 6 summarizes descriptive statistics of data.

As regression analysis could suffer bad performances if there are highly correlated input features, it is crucial to investigate the correlation between input and output attributes. Consequently, in order to guide a proper features selection, the second step explores the relationship between couples of variables: the most common method for calculating this is Pearson's Correlation Coefficient. Figure 7 shows the correlation matrix in which it can been seen the correlation between all pairs of attributes. Figure 3 shows the correlation matrix in graphical form.

Finally, a graphical analysis has been performed to point out the characteristics of the attributes. Figure 4 and in Figure 5 provide a general outlook of the distribution of each attributes.

The data analysis reveals some interesting facts about the attributes. The AlbeInf, AlbeSup, RadSG, RadSN, UVA and UVB have averagely high correlation among them. It was widely expected, since they measure different aspect of the solar radiation.

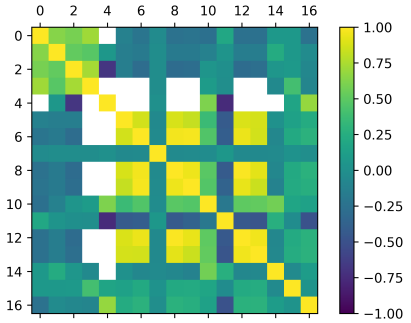

Figure 3: Correlation matrix plot.

In addition, the *CO* parameter has a moderate correlation with C_6H_6, NO_2 and SO_2, differently from all the other attributes with which there is no evidence of considerable correlation.

The last step in this phase consists in feature selection and data cleaning, therefore we drop some features from the original dataset and drop all the lines with missing data.

As final result of data transformation, cleaning and features selection, we have obtained a dataset with the following features: C_6H_6, NO_2, SO_2, Rainfall, Temperature, Humidity, Pressure, Wind direction, Wind speed, and finally *CO* as target variable.

4.4 Evaluation Metrics

To validate the ability of the regression model to make good predictions, the dataset has been divided into a training and a test part, 70% and 30% respectively. The Mean Square Error (*MSE*), the Mean Absolute Error (*MAE*), and the Coefficient of determination (R^2) have been calculated to evaluate the performance of prediction.

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

$$\mathrm{MSE}(y,\hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2$$

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

MAE and *MSE* are risk metrics corresponding to the expected value of the error and the quadratic error, while R^2 (≤ 1) represents the proportion of variance of y and provides a general indication of goodness of fit of the model.

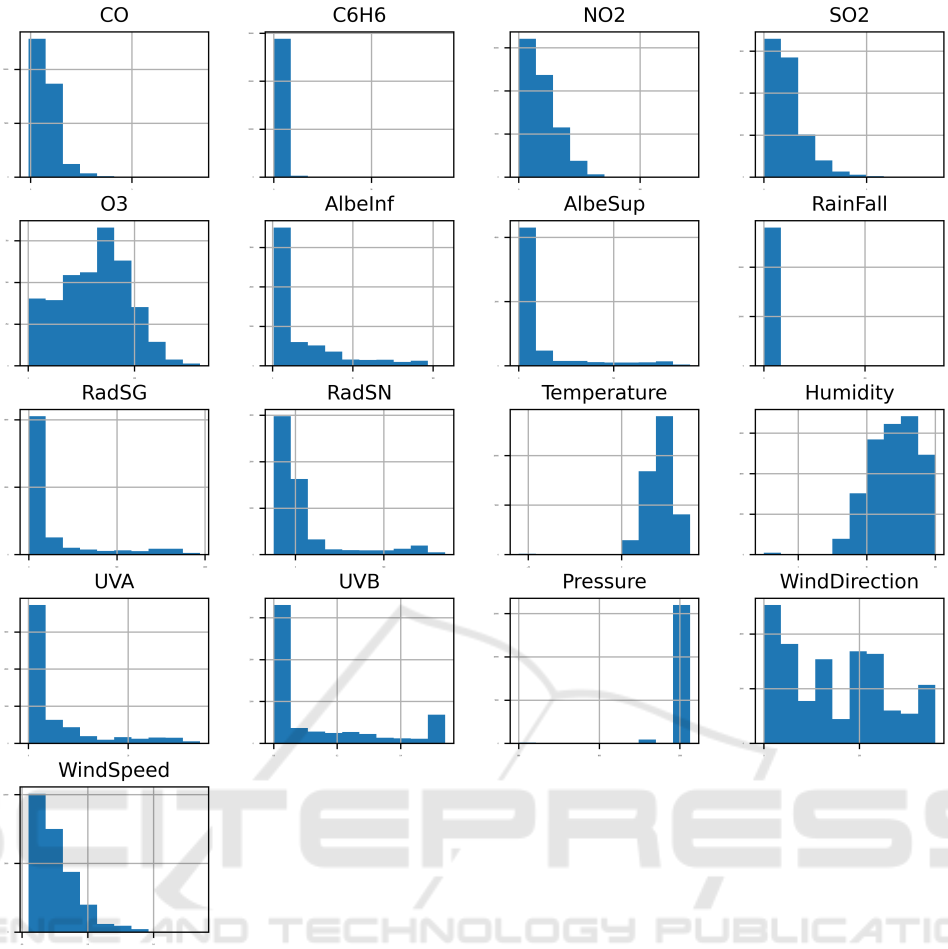

Figure 4: Outlook of distribution attributes.

4.5 Regression Models

In order to perform the analysis, we applied four different regression models, namely the Linear Regression Model calculated with Ordinary Least Square (LN-OLS), the Ridge regression model (Ridge), the Lasso model (Lasso) and Supervised Nearest Neighbors Regression (KNN).

We use the notation $x \in \mathbb{R}^m$ to describe the input data, with m input features, y for the target variable (CO). The Linear Model tries to approximate the predicted value $\hat{y}$ using a linear combination of the input features:

$$\hat{y}(w,x) = w_0 + w_1 x_1 + \ldots + w_p x_p$$

.

We also use the notation X to describe the matrix of input features and $w = (w_1, \ldots, w_p)$ for the vector of coefficients. Mathematically, the solution of the following problem provides us with the values of the coefficients w of the linear model, using the aforementioned methods:

$$OLS : \min_w \|Xw - y\|_2^2$$

$$Ridge : \min_w \|Xw - y\|_2^2 + \alpha\|w\|_2^2$$

$$Lasso : \min_w \frac{1}{2n_{\text{samples}}} \|Xw - y\|_2^2 + \alpha\|w\|_1$$

.

The KNN was selected because it is a non linear algorithm which uses a different approach, on a different basis with respect to the other three chosen ones: consequently, it is not possible to define an analogous, yet consistent, formal expression.

5 RESULTS AND DISCUSSION

To implement the various regression algorithms, we have used Python and the scikit-learn programming

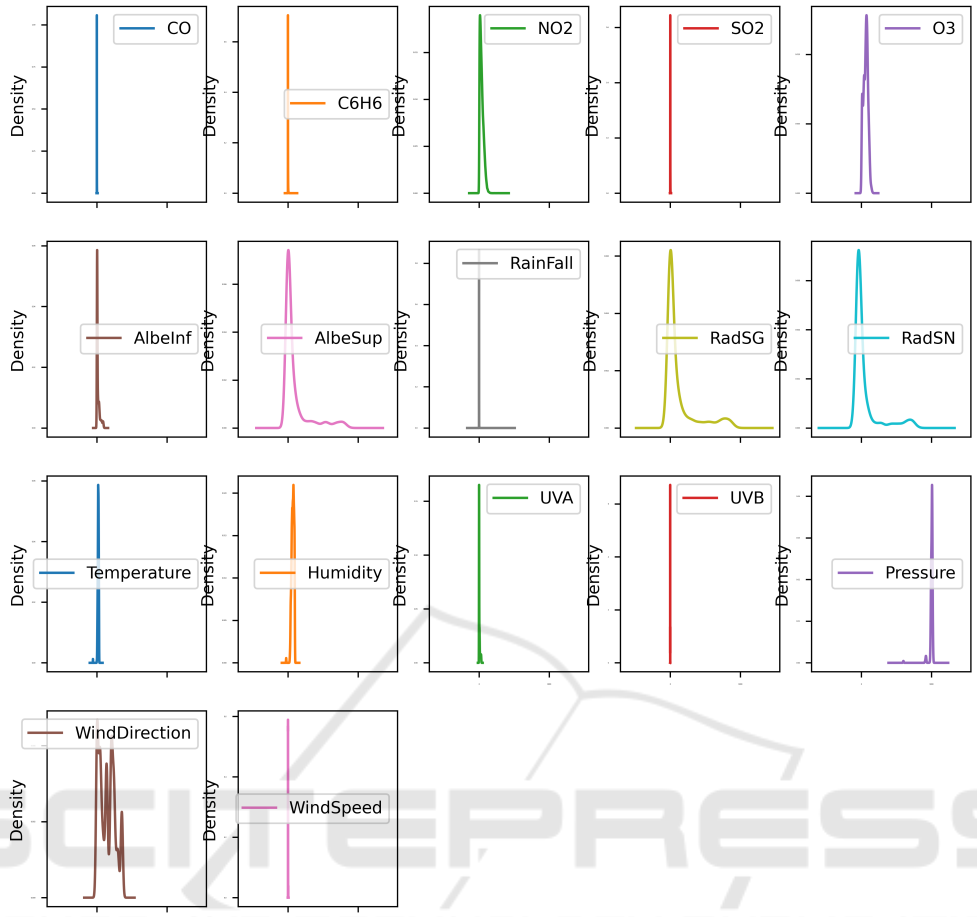

Figure 5: Outlook of attributes density.

	CO	C6H6	NO2	SO2	O3	AlbeInf	AlbeSup	RainFall	RadSG	RadSN	Temperature	Humidity	UVA	UVB	Pressure	WindDirection	WindSpeed
count	23199.000000	29205.000000	35903.000000	7786.000000	3907.000000	5690.000000	5690.000000	28056.000000	5690.000000	5690.000000	28056.000000	27818.000000	5738.000000	5738.000000	16582.000000	25274.000000	25274.000000
mean	0.807344	1.134922	37.718810	1.167062	62.880760	18.226991	100.785431	0.181401	125.293791	50.675707	20.270797	69.633865	5.024807	0.133651	1000.700965	149.807470	1.294758
std	0.560325	2.013477	26.305528	0.909716	31.585689	22.213847	194.870322	3.184699	229.271382	197.745682	7.709763	17.563240	8.021986	0.188412	33.376437	111.069902	0.872063
min	-0.096211	0.000000	0.000000	0.000000	0.417592	0.750000	0.000000	0.000000	-0.860000	-122.640000	-55.000000	-25.000000	0.060000	0.000000	600.000000	0.000000	0.250000
25%	0.449927	0.139128	16.924670	0.577022	38.433172	4.540000	0.000000	0.000000	0.000000	-47.740000	16.260000	57.222500	0.130000	0.000000	997.192500	45.000000	0.660000
50%	0.691268	0.488900	31.756836	0.954040	66.120540	5.700000	0.905000	0.000000	1.670000	-28.220000	20.850000	70.960000	0.495000	0.010000	1006.855000	141.000000	1.100000
75%	0.994990	1.290915	53.205110	1.525592	85.271128	26.037500	87.110000	0.000000	115.975000	29.817500	25.290000	83.450000	6.630000	0.230000	1010.400000	236.000000	1.720000
max	8.356315	87.676735	282.449340	8.348122	161.597300	107.470000	902.830000	339.200000	971.530000	845.860000	36.740000	100.000000	34.380000	0.540000	1025.680000	359.000000	6.770000

Figure 6: Statistical outlook of attributes of the dataset.

library. We use a cross-validation approach, in order to estimate the overall performance of the chosen machine learning algorithms with less variance than in the case of a single train-test split. The procedure starts by dividing the dataset in k parts, with $k = 10$ in our tests; then the algorithm is trained on $k-1$ parts. The result is a more reliable estimation of the performance of the machine learning algorithm.

We have performed tests using variables $X = (C_6H_6, NO_2, SO_2$, Rainfall, Temperature, Humidity, Pressure, Wind direction, Wind speed) as features and variable $y = CO$ as target. We have repeated tests for each of the above considered machine learning algorithms, obtaining comparable measurements of performance. The results are summarised in Table 5.

Table 5: Results.

Model	R2	MAE	MSE
LR-OLS	0.73	0.12	0.04
Ridge	0.77	0.12	0.04
Lasso	0.60	0.13	0.05
KNN	0.68	0.12	0.04

This analysis intentionally uses only regression models: the prediction of CO from weather measures and other pollution-related values was not a straight-

	CO	C6H6	NO2	SO2	O3	AlbeInf	AlbeSup	RainFall	RadSG	RadSN	Temperature	Humidity	UVA	UVB	Pressure	WindDirection	WindSpeed
CO	1.000000	0.637214	0.591649	0.712027	NaN	-0.124526	-0.225244	-0.036840	-0.242321	-0.232810	-0.263667	0.196412	-0.251396	-0.253543	0.067847	0.062895	-0.244324
C6H6	0.637214	1.000000	0.521443	0.483636	0.060684	-0.144642	-0.184261	-0.020596	-0.175535	-0.181585	-0.118505	0.037448	-0.153498	-0.150508	0.228586	0.069769	-0.035123
NO2	0.591649	0.521443	1.000000	0.728671	-0.682951	-0.141511	-0.282882	-0.019462	-0.296857	-0.279879	0.041148	-0.002524	-0.293796	-0.273353	0.049208	0.074000	-0.074381
SO2	0.712027	0.483636	0.728671	1.000000	NaN	NaN	NaN	-0.022015	NaN	NaN	0.098162	-0.011997	NaN	NaN	-0.052765	0.404499	-0.091736
O3	NaN	0.060684	-0.682951	NaN	1.000000	NaN	NaN	0.006093	NaN	NaN	0.613961	-0.752549	NaN	NaN	NaN	0.164085	0.684651
AlbeInf	-0.124526	-0.144642	-0.141511	NaN	NaN	1.000000	0.866066	-0.052916	0.880740	0.820827	0.360121	-0.427446	0.889102	0.871220	0.014538	0.142863	0.226704
AlbeSup	-0.225244	-0.184261	-0.282882	NaN	NaN	0.866066	1.000000	-0.043336	0.961530	0.971608	0.457324	-0.405263	0.941738	0.867138	-0.076591	0.214191	0.284951
RainFall	-0.036840	-0.020596	-0.019462	-0.022015	0.006093	-0.052916	-0.043336	1.000000	-0.046671	-0.037875	-0.025351	0.059239	-0.047272	-0.047280	-0.041545	0.020386	-0.008200
RadSG	-0.242321	-0.175535	-0.296857	NaN	NaN	0.880740	0.961530	-0.046671	1.000000	0.951207	0.492132	-0.428779	0.984748	0.920293	-0.072163	0.215532	0.278628
RadSN	-0.232810	-0.181585	-0.279879	NaN	NaN	0.820827	0.971608	-0.037875	0.951207	1.000000	0.466819	-0.360236	0.930301	0.838612	-0.110924	0.219572	0.257333
Temperature	-0.263667	-0.118505	0.041148	0.098162	0.613961	0.360121	0.457324	-0.025351	0.492132	0.466819	1.000000	-0.201257	0.505579	0.520928	0.573868	0.174992	0.147824
Humidity	0.196412	0.037448	-0.002524	-0.011997	-0.752549	-0.427446	-0.405263	0.059239	-0.428779	-0.360236	-0.201257	1.000000	-0.443047	-0.470730	0.261683	-0.023654	-0.499300
UVA	-0.251396	-0.153498	-0.293796	NaN	NaN	0.889102	0.941738	-0.047272	0.984748	0.930301	0.505579	-0.443047	1.000000	0.954241	-0.077056	0.220877	0.277843
UVB	-0.253543	-0.150508	-0.273353	NaN	NaN	0.871220	0.867138	-0.047280	0.920293	0.838612	0.520928	-0.470730	0.954241	1.000000	-0.066457	0.237285	0.285999
Pressure	0.067847	0.228586	0.049208	-0.052765	NaN	0.014538	-0.076591	-0.041545	-0.072163	-0.110924	0.573868	0.261683	-0.077056	-0.066457	1.000000	-0.043642	0.104105
WindDirection	0.062895	0.069769	0.074000	0.404499	0.164085	0.142863	0.214191	0.020386	0.215532	0.219572	0.174992	-0.023654	0.220877	0.237285	-0.043642	1.000000	0.025844
WindSpeed	-0.244324	-0.035123	-0.074381	-0.091736	0.684651	0.226704	0.284951	-0.008200	0.278628	0.257333	0.147824	-0.499300	0.277843	0.285999	0.104105	0.025844	1.000000

Figure 7: Pearson correlation between attributes of dataset.

forward task for the regression machine learning algorithm, but with a R^2 value of 0.77 for the Ridge algorithm and a low value for both MAE and MSE, it becomes suitable for prediction.

6 CONCLUSIONS AND FUTURE WORKS

In this paper we studied the applicability of a simple machine learning technique on real open data managed by third parties. Specifically, the analysed case study is the Campania region (southern Italy), where ARPAC, the local environmental protection agency, developed and managed an air monitoring network by using mobile and stationary stations. The same network overlays with another network regarding meteorological measurements.

The first step was the analysis of available data and the creation of a merged dataset so to overcome the spatiotemporal discrepancy. Indeed, even if the same station can perform both the chemical analysis and meteorological recording, the two networks work on different layers; hence, data are recorded in two different datasets with different characteristics. The challenge was overcome with the use of identifiers that linked the same station between the dataset used. It is important to highlight that, other from "outer" mode used to preserve the largest quantity of data and avoid data loss, the data were not screened for their coherence. This decision was made in order to understand how coarse data perform with the machine learning technique. According to the correlation analysis performed on CO trends, this parameter emerged to have a moderate correlation with C_6H_6, NO_2 and SO_2, differently from all the other attributes with which there is no evidence of considerable correlation. This correction is explained by the characteristics of the compounds: all of them are related to vehicular traffic emissions. Next, four different regression models, namely the Linear Regression Model calculated with Ordinary Least Square (LN-OLS), the Ridge regression model (Ridge), the Lasso model (Lasso) and Supervised Nearest Neighbors Regression (KNN) were applied and evaluated by statistical means. Finally, the Ridge regression model was found to be the choice that fits the best among them with an R^2 value equal to 0.77 and low value for both MAE and MSE, equal to 0.12 and 0.04 respectively.

In this work, using regression algorithms we only scratched the surface in data prediction. The search for a better prediction should consider the use of deep learning algorithms: as a matter of fact, this methodology can analyse large datasets with considerable performances. Nevertheless, it is clear that performing an headline validation of the starting data can help both the techniques, as, at the moment, a human intervention in the first phases is still needed to address the right directions in data exploration.

ACKNOWLEDGEMENTS

This work has been partially funded by the internal competitive funding program "VALERE: VAnviteLli pEr la RicErca" of Università degli Studi della Campania "Luigi Vanvitelli" and by project "Attrazione e Mobilità dei Ricercatori" Italian PON Programme (PON_AIM 2018 num. AIM1878214-2).

REFERENCES

Agarwal, S., Sharma, S., Suresh, R., Rahman, M. H., Vranckx, S., Maiheu, B., Blyth, L., Janssen, S., Gargava, P., Shukla, V., et al. (2020). Air quality forecasting using artificial neural networks with real time dynamic error correction in highly polluted regions. *Science of The Total Environment*, 735:139454.

Armstrong, J. S. (2001). *Extrapolation for Time-Series and Cross-Sectional Data*, pages 217–243. Springer US, Boston, MA.

ARPAC. Organizzazione della rete di monitoraggio della Qualità dell'Aria. https://dati.arpacampania.it/dataset/rete-di-monitoraggio-della-qualita-dell-aria. Accessed: 2021-02-28.

Baklanov, A. and Zhang, Y. (2020). Advances in air quality modeling and forecasting. *Global Transitions*, 2:261–270.

Belavadi, S. V., Rajagopal, S., Ranjani, R., and Mohan, R. (2020). Air quality forecasting using lstm rnn and wireless sensor networks. *Procedia Computer Science*, 170:241–248.

Brandt, J., Christensen, J. H., Frohn, L. M., Palmgren, F., Berkowicz, R., and Zlatev, Z. (2001). Operational air pollution forecasts from european to local scale. *Atmospheric Environment*, 35:S91–S98.

Campanile, L., Iacono, M., Lotito, R., and Mastroianni, M. (2020). A wsn energy-aware approach for air pollution monitoring in waste treatment facility site: A case study for landfill monitoring odour. In *IoTBDS*, pages 526–532.

Centro Meteorologico e Climatologico. http://cemec.arpa.campania.it/meteoambientecampania/php/misure_suolo.php. Accessed: 2021-02-28.

Chatfield, C. (2000). *Time-series forecasting*. CRC press.

Contreras, L. and Ferri, C. (2016). Wind-sensitive interpolation of urban air pollution forecasts. *Procedia Computer Science*, 80:313–323.

Domańska, D. and Wojtylak, M. (2014). Explorative forecasting of air pollution. *Atmospheric Environment*, 92:19–30.

European Environmental Agency. European Air Quality Index. https://www.eea.europa.eu/themes/air/air-quality-index. Accessed: 2021-02-28.

Haehnel, P., Mareček, J., Monteil, J., and O'Donncha, F. (2020). Using deep learning to extend the range of air pollution monitoring and forecasting. *Journal of Computational Physics*, 408:109278.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Kurt, A., Gulbagci, B., Karaca, F., and Alagha, O. (2008). An online air pollution forecasting system using neural networks. *Environment international*, 34(5):592–598.

Liu, H., Yan, G., Duan, Z., and Chen, C. (2021). Intelligent modeling strategies for forecasting air quality time series: A review. *Applied Soft Computing*, page 106957.

Masmoudi, S., Elghazel, H., Taieb, D., Yazar, O., and Kallel, A. (2020). A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection. *Science of the Total Environment*, 715:136991.

Masood, A. and Ahmad, K. (2020). A model for particulate matter (pm2. 5) prediction for delhi based on machine learning approaches. *Procedia Computer Science*, 167:2101–2110.

Shetty, C., Sowmya, B., Seema, S., and Srinivasa, K. (2020). Air pollution control model using machine learning and iot techniques. In *Advances in Computers*, volume 117, pages 187–218. Elsevier.

Tong, W. (2020). Machine learning for spatiotemporal big data in air pollution. In *Spatiotemporal Analysis of Air Pollution and Its Application in Public Health*, pages 107–134. Elsevier.

Xiao, D., Fang, F., Zheng, J., Pain, C., and Navon, I. (2019). Machine learning-based rapid response tools for regional air pollution modelling. *Atmospheric environment*, 199:463–473.

Zhou, Y., Chang, F.-J., Chang, L.-C., Kao, I.-F., and Wang, Y.-S. (2019). Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. *Journal of cleaner production*, 209:134–145.