MACHINE LEARNING METEOROLOGICAL NORMALIZATION MODELS FOR TREND ANALYSIS OF AIR QUALITY TIME SERIES

ROBERTA VALENTINA GAGLIARDI & CLAUDIO ANDENNA
Istituto Superiore di Sanità, Italy; INAIL-DIT, Italy

ABSTRACT

Air pollution is a major environmental cause of morbidity and mortality worldwide, representing a top public health objective, especially in areas interested by the presence of anthropic emissions sources. Correctly assessing how pollutant emissions influence the air quality is, therefore, crucial for the design and/or implementation of effective measures from the public health perspectives. The impact of local emission sources on air quality is strongly modulated by meteorological conditions, which can mask the real trends in the observed pollutant concentrations. However, the confounding effect of meteorology in air quality time series can be accounted for by techniques of meteorological normalisation. In this study, the performances of a meteorological normalisation technique based on machine learning (ML) algorithms were investigated. To these purposes, two ML models (gradient boosted regression (GBM) and random forest (RF)) were developed and subsequently used to calculate meteorologically normalised trends of nitrogen oxide (NO_x) concentrations time series. Both models were trained on daily averaged data of NO_x concentrations and meteorological parameters, as well as on temporal variables; data were acquired, over the 2013–2019 period, in a rural area affected by anthropic sources of air pollutants. Results obtained show that both models are able to explain more than 70% of the variance in the NO_x observed concentrations and that the meteorological normalization technique based on both algorithms represent a robust method to account for the confounding effect of meteorology in air quality time series. Moreover, the GBM/RF ML models allowed to analyse the dependence of the observed concentrations on each explanatory variables used in the models, shedding light on the role of local meteorological processes in the observed pollutant concentrations. This knowledge can help in defining air pollution control strategies that are increasingly effective in preventing and/or mitigating health damage associated with exposure to atmospheric pollution.

Keywords: air pollution, boosted regression trees, machine learning, meteorology, random forest, trend analysis.

1 INTRODUCTION

Air pollution is a major risk factor to human health. According to the World Health Organization (WHO), ambient air pollution causes more than 4 million premature deaths every year worldwide, and more than 90% of the population lives in areas exceeding the WHO guideline limits [1]. Furthermore, air pollution is one of the major factors contributing to climate change, especially in terms of global warming; at the same time, climate change can perturb the long-range transport, chemical processing and local meteorology that influence air pollution [2]. In the European context, Italy presents several criticalities in terms of high-polluted areas [3]; moreover, due to its geographical position at the centre of the Mediterranean area, it is also a 'hot spot' for climate change because of the intense photochemical activity, the crossing of air masses of different origin and the strong anthropogenic pressure [4].

To manage the air quality issues, the environmental or health decision makers need reliable estimates of pollutant concentration levels and related trends as input for decisions. The quantitative assessment of the real trend of pollutant concentrations is complicated by the variability of air pollution due to variations in local and synoptic meteorological conditions and seasonal effects, as well as the non-linear responses between emissions and concentrations

ISSN: 2398-2640 (paper format), ISSN: 2398-2659 (online), http://www.witpress.com/journals
DOI: 10.2495/EI-V4-N4-375-387

of air pollutants strongly affected by meteorology over multiple scales in time and space [5]. Therefore, to avoid that weather effects mask the actual trends in the observed pollutant concentrations, the confounding effect of meteorology in air quality time series must be accounted for. Once the weather effects have been removed, further statistical evaluations can be carried out in the resulting air quality time series, obtaining more robust estimation of pollutant trends or more reliable air quality predictions.

The process of accounting for changes in meteorology over time in an air quality time series, which is referred as 'meteorological normalization', can be carried out through several statistical techniques [6]. An emerging approach to meteorological normalization is based on machine-learning (ML) algorithms [7]. It mainly consists of a two-stage process reducing air quality time series variability with statistical modelling: first, an ML model, linking air quality and weather data, at a location of interest is used to predict pollutant concentrations as a function of meteorological parameters [8]. Second, if the model explains an adequate amount of variance in the predicted air quality variable, it can be used under a range of meteorological conditions, with the associate average referred to as meteorological–normalized time series [9].

Among the most popular ML algorithms, those based on decision trees methods, such as gradient boosting regression (GBM) [10] and random forest (RF) [11], are extensively used in the air quality field [12], [13]. Both these algorithms use a set of independent variables (explanatory or predictors variables) and an ensemble of decision trees to make predictions of a variable of interest (target/dependent variable). GBM/RF models are characterized by strong predictive performances and remarkable capability of insights on the relationships between variables. Their increasingly widespread use is due to their ability to model nonlinear relationships, to manage qualitative and quantitative variables, to remain robust despite missing data and outliers, to reduce overfitting and to require a limited number of user-defined parameters for model fitting/selection purposes. Furthermore, thanks to the interpretability of the GBM/RF models, it is possible to provide the functional relationships between each predictors and the dependent variable, improving model understanding and trustworthiness.

The aim of this study is to explore the performances of both GBM/RF algorithms as a basis of a meteorological normalization technique by assessing the global accuracy metrics, as well as the interactions between the target and the explanatory variables selected for the models development. Moreover, trend analysis of the normalized air quality time series is also performed to quantitatively assess the changes in the ambient air pollution.

To these purposes, two GBM/RF models were built, validated and subsequently applied to calculate meteorologically normalized air quality time series. Both models were trained on daily averaged data of NO_x and meteorological parameters as well as on time variables; data were acquired, over the 2013–2019 period, in a semi-rural area affected by anthropic sources of air pollutants. A comparison between GBM and RF models was made on the basis of several statistical indicators [14]. Trend analysis was carried out on the normalized NO_x concentrations using the Theil–Sen regression technique [15]. Finally, the abilities of both models in ranking, visualizing and predicting the relationship between NO_x concentrations and its driving factors were analysed and graphically illustrated.

The paper is structured as follows: Section 2 describes the study area, the data used and the main steps adopted for the meteorological normalization procedure based on the GBM and RF models. A comparison between GBM and RF models' predictive performances, as well as the results obtained with the meteorological normalization procedure and the subsequent trend analysis, are presented in Section 3. It also includes a description of the outputs of the GBM and RF models. Finally, Section 4 summarizes the main findings of this work.

2 MATERIALS AND METHODS

2.1 Study area

The study area is the Agri Valley, located in the South-West part of the Basilicata Region (Southern Italy) (Fig. 1). Moreover, the site location is at the centre of the Mediterranean area, one of the most responsive regions to climate change.

Starting from the early 1990s, the largest on-shore western European reservoir of crude oil and gas in a populated area and an oil pre-treatment plant (identified as Centro Olio Val d'Agri – hereafter COVA) are operating in the valley [16]. The COVA plant determines emissions of gases and particulates, which can affect the air quality and potentially pose health risks for the population living in the area. Continuous concentration measurements of regulated pollutants and of several pollutants specifically related to oil/gas extraction activities are provided by an air quality control network, consisting of five monitoring stations. At the stations are also measured the following meteorological parameters: temperature (T), atmospheric pressure (P), relative humidity (RH), solar radiation (SR), wind direction (wd) and wind speed (ws). The Environmental Protection Agency of the Basilicata Region (ARPAB), managing the network, validates and makes public these data. More details about the methods and the instrumentation used for the measurements can be found elsewhere [17]. For the purpose of this work, data were obtained from the monitoring station closest to the COVA plant, named Viggiano (VZI, 40°18'50''N, 15°54'16''E, 603 m a.s.l.), categorized as an industrial station in a rural area. It is located at about 350 m from the industrial site and about 1000 m from a national road (SS598) characterized by a moderate volume of traffic produced by cars and heavy vehicles.

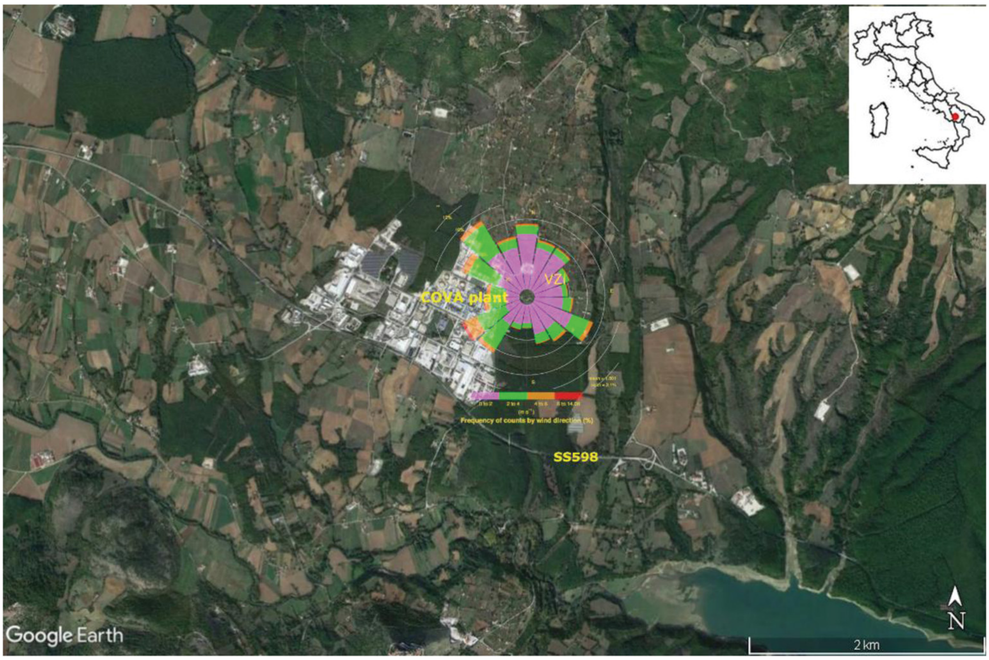


Figure 1: Map of the area: the VZI monitoring site, the COVA plant and the wind rose based on hourly data at the VZI station over the study period (2013–2019).

2.2 Data preparedness

The meteorological normalization procedure was applied at the NO_x concentrations time series, which are key air pollutants also playing an important role in tropospheric chemistry as precursors of tropospheric ozone and secondary aerosols [18]. Being predominantly emitted during fuel combustion, such as by vehicle engines, industry processes and domestic heating, NO_x can be considered as an indicator of the anthropic emissive sources existing in the examined area, which mainly consist in the conveyed emissions produced by the COVA plant and in the local traffic sources. Therefore, to form the whole data set used for the GBM/RF models development, hourly data of NO_x concentrations together with the meteorological variables P, RH, T, ws and wd were downloaded from the official website of ARPAB [19] and combined together. Overall, a data set consisting of more than 59000 observations covering the 2013–2019 period was set up. The time series of all predictors considered respected the required 75% proportion of valid data. The daily average of data was used as input to the model; this time resolution balances the need to preserve the pattern of data at a temporal scale consistent with the examined phenomena and the need to reduce the noisy data and the computational resource demand. Subsequently, a range of other variables was added to the ML models development. These are the day of the week (weekday), the Julian day, i.e., number of days of 1 January (Jday) and the date Unix (trend), i.e., the number of seconds since 1 January 1970; they can be interpreted as proxy for local traffic sources or to account for seasonal and long-term variability, respectively. The day of the week was a categorical variable, while all others were numeric; moreover, all variables were used within their response scale. All data loading, processing, statistical analysis and modelling were accomplished in the R software environment (version 4.1.0; Foundation for Statistical Computing, Vienna, Austria) and its packages.

3 METHODOLOGICAL APPROACH

The methodological approach adopted in the present study consists of the following main steps. First, GBM/RF models were developed, and their performances were estimated and compared. Second, the meteorological normalization was carried out on the predicted NO_x concentrations by each model, and the relevant trend analysis was subsequently performed. Finally, the interpretability of the models was evaluated to ensure their plausibility and reliability.

3.1 GBM/RF models development

Theoretical insights of both GBM and RF models' development are beyond the scope of the present paper and can be found in [10] and [11], respectively. Here, it intends to recall only those concepts that are necessary for the understanding of what will be discussed later. Both GBM and RF are ensemble models that have been developed to optimize predictive performance by training multiple 'weak learners' and merging their results to build a 'strong learner'. GBM is a step-wise, additive-type model that sequentially fits new tree-based models. Each fitted model at every step attempts to compensate for the shortcomings of the previous fitted models. The final model aggregates the results from each step and a strong learner is achieved. RF generates a large number of individual models in a parallel way. In the training procedure, each tree is built based on a random subset of the original data (with replacement). In addition, a randomly selected subset of predictors is chosen for each built tree, and the RF predictions are the averaged output of all aggregations. To build both GBM/RF models, the whole observed dataset was randomly partitioned into a training dataset (80% of the observations) and a testing dataset (20% of the observations) used for model performance evaluation. The best model for each of the two ML techniques was obtained by tuning the relative hyper-parameters. For the GBM model, this process was carried out using the *gbm* R package [20] on a grid of possible values, while the *tuneRanger* R package [21] was used for RF model tuning. In both cases, the best combination of hyper-parameters, based on R^2 metric, was chosen to build the optimal models on the training dataset. For GBM, the best combination was: learning rate = 0.005, tree complexity = 5, bag fraction = 0.5; number of trees = 5,450. For RF: number of variables sampled to determine each split = 4, minimum number of terminal nodes = 2; number of trees = 1000.

Prediction performances of both models were evaluated by comparing predicted and observed NO_x concentration values using a range of statistical indicators. The coefficient of determination (R^2), the index of agreement (IoA), the mean bias error (MBE), the mean absolute error (MAE) and the root mean square error (RMSE) are used in this work. The relevant equations are provided in Appendix A. High accuracy (R^2 and IoA close to 1) and minimal errors (MBE, MAE and RMSE close to 0) are the desired performances for an optimal prediction model.

3.2 Meteorological normalization procedure and trend analysis

The procedure adopted for the meteorological normalization has been proposed in [22], which modified that originally proposed in [7] and consists in normalizing the NO_x concentrations with the GBM/RF model, resampling the meteorological explanatory variables from the whole study period. In this way, the normalization process preserves the emission changes in the normalized concentrations. For both models, this process was repeated 300 times after all the predictions were aggregated using the arithmetic mean to obtain the meteorological normalized concentration. The benefit of this approach is that the trend calculated in this way will more closely relate to emission changes rather than changes due to meteorological effects. The meteorological normalization of GBM model was conducted using the *deweather* R package [23] modified to use the optimized hyper-parameters values, with the underlying *gbm* R package; for the RF model, the meteorological normalization was carried out using the *rmweather* R package [7], with the underlying *ranger* R package.

Once normalized, the NO_x concentrations time series were object of further statistical analysis. The Theil–Sen regression technique was used to calculate the direction of a trend in the normalized concentrations over time. The Theil–Sen method assesses the median slope of all possible slopes that may occur between the data points. It is regarded as more suitable than the linear-regression method, as it gives more accurate confidence intervals with non-normal distributed data and it is not affected as much by outliers. All the regression parameters, among which is the p-value for the slope, are estimated through bootstrap resampling. In our calculations, the trends were based on monthly averages, and they were adjusted for seasonal variations, as these can have a significant effect on monthly data.

3.3 GBM/RF models interpretability

Among the undoubted advantages of the decision tree ML models are several tools allowing to interpret the GBM/RF models, enhancing their understanding and trustworthiness. These tools are the relative importance of predictors and the partial dependence plots between variables. For RF models, the variable importance measures the impact of each feature on the accuracy of the model [11]. For GBM models, the variable importance is determined by a variable's average relative influence across all trees generated by the GBM algorithm [10]. The partial plots illustrate the effect of each explanatory variables on the dependent variable after accounting for the average effects of all other variables; in this way, the plausibility and reliability of the model can be verified.

4 RESULTS AND DISCUSSIONS

4.1 Statistical analysis

The descriptive statistics of each variable used in the model are summarized in Table 1.

Time series analysis of air pollution metrics for the regulated pollutants measured at the VZI station showed a general compliance with the limits set by the existing national [24] and European legislation [25], [26]. As far as the climate is concerned, the cold and rainy winters, as well as cool summers with frequent rainfall [27], typically registered in the area, define an area at sub-continental climate. During the study period, the mean temperature was 13.78 °C, the mean relative humidity was 71.1%, while pressure was rather static. The mean value of ws was 1.8 ms^{-1} with the higher values generally measured during daytime. As shown by the wind rose in Fig. 1, the prevailing wind direction was from the SW–NW sector.

4..2 Models comparison

The GBM/RF models took the form shown by equation 1:

$$NO_x = gbm / rf\left(T, H, ws, wd, P, jday, weekday, trend\right), \tag{1}$$

where *gbm/rf* were the functions implementing the GBM and RF technique in the R software environment. The resulting predictive performances and behaviour of both models were compared through the statistical indicators shown in Table 2. The results suggest that both models were highly predictive based on the range of R^2 values from 0.73 (RF) to 0.76 (GBM). The R^2 values suggest that GBM/RF models can explain more than 70% of the total NO$_x$ variability. Both models showed very similar performances both in terms of minimal errors based on the average values of MBE, MAE and RMSE and in terms of prediction accuracy.

Table 1: Statistical summary of hourly data of NO$_x$ and meteorological parameters registered at the VZI monitoring station from January 2013 to December 2019. m.u. = measurement unit.

Parameter	m.u.	Min	Max	Mean	Median
NO$_x$	µg/m^3	0.00	186.06	16.63	11.17
RH	%	5.68	100.00	71.31	74.50
ws	ms^{-1}	0.00	14.08	1.80	1.38
T	°C	-10.90	41.69	13.78	13.10
P	hPa	872.00	971.00	945.20	945.30

Table 2: Statistical indicators of the GBM and RF models performances for the testing data set. Legend: R^2 = coefficient of determination, MBE = mean bias error, MAE = mean absolute error, RMSE = root men square error and IoA = index of agreement.

NO_x	$\boldsymbol{R^2}$	**MBE [μg/m³]**	**MAE [μg/m³]**	**RMSE [μg/m³]**	**IoA**
GBM	0.76	–0.45	3.72	5.52	0.76
RF	0.73	–0.09	3.53	5.39	0.76

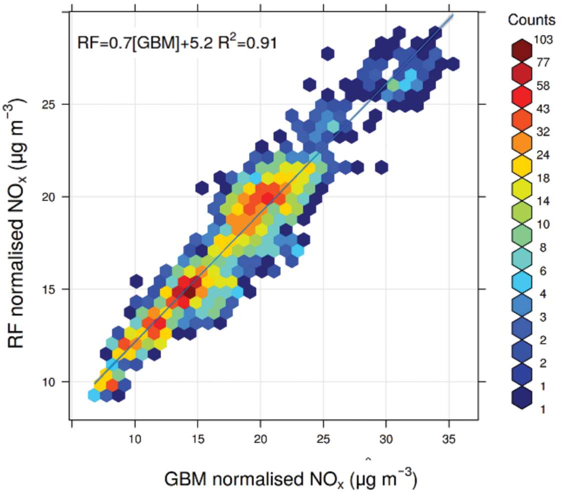


Figure 2: Scatter plot of the GBM *vs*. RF normalized NO_x concentrations.

As shown in Fig. 2, the NO_x-normalized trends from both models agree very well with each other, R^2 =0.91, confirming the substantial similarity in the predictive ability.

4.3 Meteorological normalization and trend analysis

The observed daily concentrations of NO_x were compared with the normalized concentrations predicted with both GBM and RF models (Fig. 3). The meteorological normalized signal highlights the trends of the NO_x concentrations with respect to the observed data. It is worth noting a relevant decrease around February–March 2016 that is consistent with the reduced emissive activity of the COVA plant due to a general plant shutdown for judicial investigations to which the plant was subjected, approximately from end of March to early August 2016, and to the consequent lower traffic regime around the plant.

Trend analysis of the normalized NO_x concentrations, performed with the Theil–Sen method, shows a statistically significant decreasing pattern: –0.66 [–1.17, –0.34] μg m^3 $year^1$ in GBM and –0.62 [–0.99, –0.37] μg m^3 $year^1$ in RF model, respectively, where the square bracket represents the 95% confidence intervals. This is in line with the general decreasing trend of nitrogen oxides registered over the whole national territory. The observed data give about the same trend (–0.66 [–1.13, –0.27] μg m^3 $year^1$) probably supporting the hypothesis

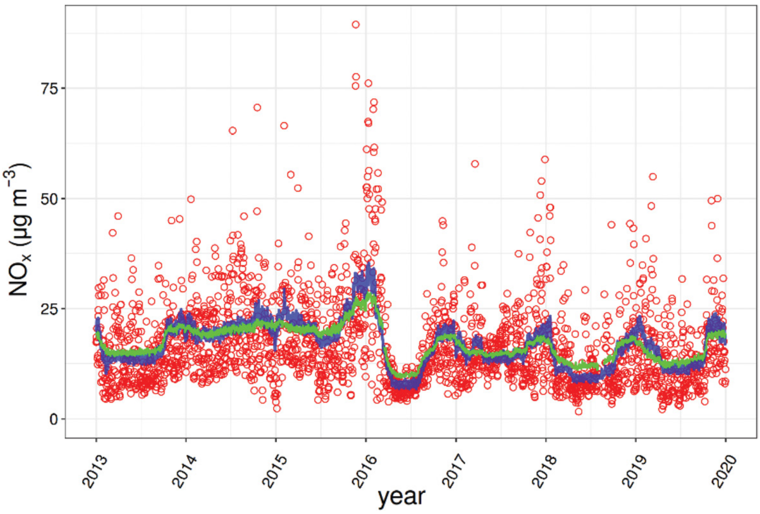


Figure 3: GBM and RF normalized NO_x concentrations (blue and green lines respectively). Red dots represent the daily averages of the observed NO_x concentrations.

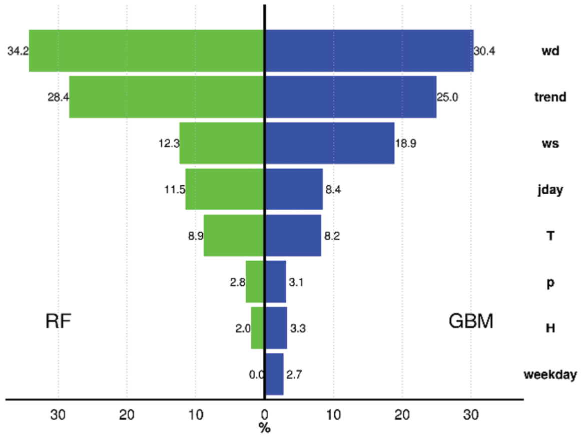


Figure 4: Relative importance of the explanatory variables in the GBM and RF models.

that the change in local sources emissions could overcome the effects of local meteorology in the observed NO_x variability.

4.4 Models interpretability

The relative importance of predictors and the partial dependence plots can be used to shed light on the role of single predictors on the NO_x concentrations variability. In Fig. 4, it is shown that the relative importance of the predictors is normalized to 100% and in a descending order.

According to the obtained results, both models indicate in wind direction the most important contribution to NO_x variability, closely followed by trend and, to a lesser extent, by ws and Jday. The overall contribution of these top four predictors explain about 86.4% for RF model and about 82.7% for GBM model of the variance in NO_x. The partial dependence plots for the top four predictors identified by the GBM and RF models are shown in Fig. 5.

The partial dependencies of NO_x from each predictor are consistent between the two models. As evident, NO_x concentrations were strongly affected by wd. The highest concentrations of the NO_x are associated with winds blowing from SW to NW, i.e., in the direction of both several of the COVA plant conveyed emissive sources and the SS598 national road. The traffic contribution to the observed concentrations was supported by the analysis of the daily and weekly NO_x pattern. The former (Fig. 6a) tends to be significantly bimodal (higher concentrations in the early morning and late afternoon coinciding with the commuting hours). The latter (Fig. 6b) shows a clear decrease of NO_x concentrations on Saturday and Sunday when traffic is usually lower. The trend represented the second most relevant variable and clearly confirmed the decrease in NO_x concentrations since the beginning of 2016 due to the reasons above discussed. NO_x concentrations slightly decrease or remain constant when ws is lower than 2.5 ms^{-1}. For higher values of ws, NO_x concentrations grow until ws reaches values of 5/6 ms^{-1}, then they remain almost constant. As explained in [28], increases in NO_x concentrations with ws could be indicative of a buoyant plume from a source such

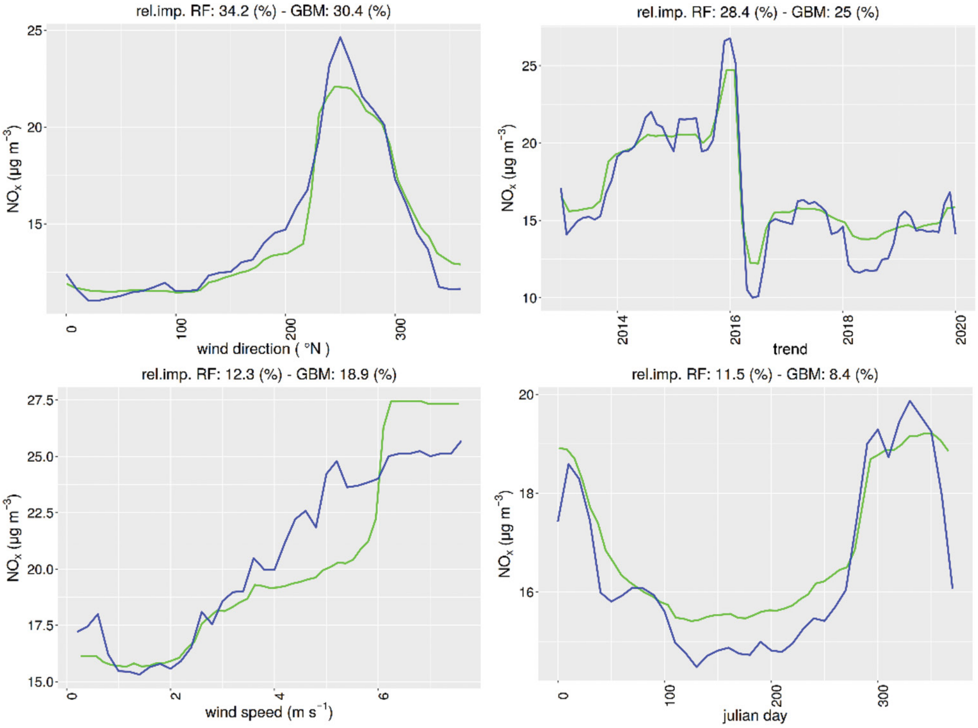
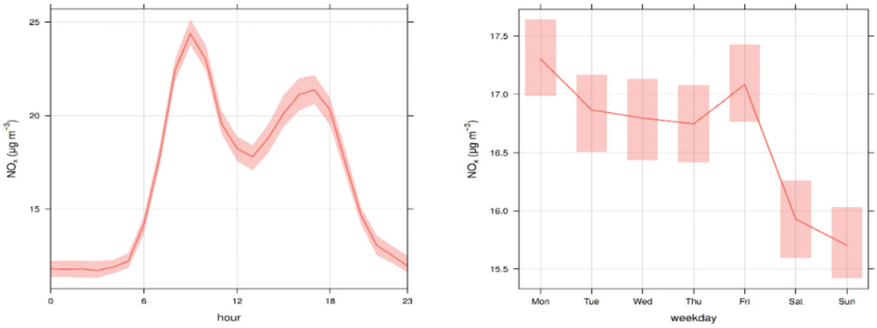


Figure 5: Partial dependence plots showing the variation in hourly NO_x concentrations as a function of the top four explanatory variables used in the GBM (blue line) and RF (green lines) models.



(**a**) (**b**)

Figure 6: Daily (a) and weekly (b) profiles of hourly NO$_x$ concentrations. Also shown on the plots is the 95 % confidence interval in the mean.

as a chimneystack, where the plume is brought down to ground level when the wind speed increases. This hypothesis is consistent with the presence of conveyed emissive sources in the proximity of the monitoring site. The pattern of NO$_x$ as a function of Jday reflects the seasonal variations in NO$_x$ and shows the expected trend with generally lower concentrations in summer and higher ones in winter: the NO$_x$ concentrations are elevated at lower ambient temperature (T <10°C) and during colder months (Oct–Feb), as these conditions limit the dispersion of emissions and there is a higher demand for domestic heating.

Overall, all the information derived from the analysis of the partial plots confirm the role of the oil plant and the vehicular traffic on the NO$_x$ variability at the study site.

5 CONCLUSIONS

In this work, GBM/RF algorithms were explored as a basis of a meteorological normalization technique. Both models showed high predictive ability in terms of accuracy and minimal errors being able to explain more than 70% of the variance in the NO$_x$ observed concentrations. The ML-based meteorological normalization technique allowed to estimate the real trend of NO$_x$ concentrations and to highlight the role of both the industrial activity and the vehicular traffic on the observed NO$_x$ concentration levels at the monitoring site. Although further refinements may be made in the future, for example, in terms of additional predictors or temporal analysis, the results produced are plausible and coherent with what is expected on the basis of the scientific literature. However, because these methods are all data driven, caution is required when generalizing the results obtained to different conditions and/or sites. Ultimately, the ML-based meteorological normalization techniques are found to represent a robust method to account for the confounding effect of meteorology in air quality time series, thus providing those reliable estimates of pollutant concentration levels and variability that represent a crucial input for the decisions concerning the environmental and public health protection.

ACKNOWLEDGEMENTS

The authors are grateful to the Environmental Protection Agency of Basilicata Region for providing the data used in this work.

REFERENCES

[1] World Health Organization, *Air Pollution*, WHO, available at https://www.who.int/health-topics/air-pollution#tab=tab_1 (accessed 19 May 2021).

[2] Fiore, A.M., Naik, V. & Leibensperger, E.M., Air quality and climate connections. *Journal of the Air & Waste Management Association,* **65(6)**, pp. 645–685, 2015.

[3] Donateo, A., Villani, M., Lo Feudo, T., Chianese, E., Recent Advances of Air Pollution Studies in Italy. *Atmosphere*, vol. **11**, https://doi.org/10.3390/atmos11101054, 2020.

[4] F. Giorgi, Climate change hot spot, *Geophys Res Lett*., vol. **33**, 10.1029/2006GL025734, 2006.

[5] Kinney, P.L., Climate change, air quality, and human health. *American Journal of Preventive Medicine,* vol. **35(5)**, pp. 459–467, 2008.

[6] Thompson, M.R.J., Cox, L., Guttorp, P. & Sampson, P., A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric Environment,* **35**, pp. 617–630, 2001.

[7] Grange, S., Carslaw, D., Lewis, A., Boleti, E. & Hueglin, C., Random forest meteorological normalisation models for Swiss PM10 trend analysis. *Atmospheric Chemistry and Physics,* **18**, pp. 6223–6239, 2018.

[8] Grange, S. & Carslaw, D.,Using meteorological normalisation to detect interventions in air quality time series. *Science of the Total Environment,* **653**, pp. 578–588, 2019.

[9] Petetin, H., Bowdalo, D., Soret, A., Guervara, M., Jorba, O., Serradell, K., Perez Garcia-Pardo, C., Meteorology-normalized impact of COVID-19 lokdown upon NO2 pollution in Spain. *Atmos. Chem. Phys*., vol. **20**, pp. 11119–11141, 2020.

[10] Friedman, J., Stochastic gradient boosting. *Computational Statistics & Data Analysis,* **38(4)**, pp. 367–378, 2002.

[11] Breiman, L., Random forests. *Machine Learning,* **45**, pp. 5–32, 2001.

[12] Gagliardi, R.V., Andenna, C. A Machine Learning Approach to Investigate the Surface Ozone Behavior. *Atmosphere*, vol. **11**, https://doi.org/10.3390/atmos11111173, 2020.

[13] Brokamp, C., Jandarov, R., Hossain, M. & Ryan, P., Predicting daily urban fine particulate matter concentrations using a random forest model. *Environmental Science & Technology,* **52**, pp. 4173–4179, 2018.

[14] Sayegh, A., Munir, S. & Habeedullah, T., Comparing the performance of statistical models for predicting PM10. *Aerosol and Air Quality Research,* **14**, pp. 653–665, 2014.

[15] Nunifu, T. & Fu, L., *Methods and Procedures for Trend Analysis of Air Quality Data*. Government of Alberta, Ministry of Environment and Parks: Edmonton, 2019.

[16] ENI, Il centro Olio Val d'agri, https://www.eni.com/eni-basilicata/chi-siamo/centro-olio-val-d-agri.page. Accessed on: 10 Mar. 2021.

[17] ARPAB, Inquinanti monitorati, http://www.arpab.it/aria/inquinanti.asp. Accessed on: 30 Mar. 2020.

[18] Seinfeld, J. & Pandis, S., *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, John Wiley & Sons: New York, 2006.

[19] ARPAB, Gli Open Data - qualità dell'aria www.arpab.it/opendata/q_aria_serie.asp. Accessed on: 10 Jan. 2020.

[20] Greenwell, B., Boehmke, B., Cunningham, J., Developers, G., gbm: geeneralized boosted regression models. *r package version 2.1.5*, 2019. https://cran.r-project.org/web/packages/gbm/index.html. Accessed on: 10 Jan. 2021.

[21] Probst, P., Wright, M., Boulestei, A., *Hyperparameters and Tuning Strategies for Random Forest*. https://arxiv.org/pdf/1804.03515.pdf, Accessed on: 20 December 2018.

[22] Shi, Z., Song, C., Liu, B., Lu, G., Xu, J., Vu, T., Elliot, R., Li, W., Bloss, W. & Harrison, R., Abrupt but smaller than expected changes in surface air quality attributable to COVID-19 lockdowns. *Science Advances,* **7**, pp. 1–10, 2021.

[23] Carslaw, D., Deweather, http://github.com/davidcarslaw/deweather. Accessed on: 10 March 2019.

[24] Legislative Decree 155/10. Attuazione della direttiva 2008/50/CE relativa alla qualità dell'aria ambiente e per un'aria più pulita in Europa. *Gazzetta Ufficiale n. 216 del 15.09.2010 - Suppl. Ordinario n. 217*, 2010.

[25] Directive 2008/50/EC on ambient air quality and cleaner air for Europe, Offcial Journal of the European Union, L 152/1, pp. 1–44, 11.6.2008.

[26] Arpab, Rapporto Ambientale Anuale, http://www.arpab.it/public/Rapporto-Ambientale-anno-2019.pdf. Accessed on: 10 Jun 2020.

[27] Prefettura - Ufficio Territoriale del Governo di Potenza, PEE Centro Olio Val d?agri di Viggiano, http://www.prefettura.it/potenza/contenuti/Pee_centro_olio_val_d_agri_di_viggiano_edizione_2013-64403.htm. Accessed on: 30 Mar. 2021.

[28] Carslaw, D.C., Beevers, S.D., Ropkins, K. & Bell, M.C., Detecting and quantifying aircraft and other on-airport contributions to ambient nitrogen oxides in the vicinity of a large international airport. *Atmospheric Environment,* **40**, pp. 5424–5434, 2006.

Appendix A

Statistic name	Equation
Mean bias error	$MBE = \frac{1}{N}\sum_{i=1}^{N} M_i - O_i$
Mean absolute error	$MAE = \frac{1}{N}\sum_{i=1}^{N} \lvert M_i - O_i \rvert$
Root mean squared error	$RMSE = \sqrt{\left(\frac{\sum_{i=1}^{N} \left(M_i - O_i \right)^2}{N} \right)}$
Coefficient of determination	$R^2 = \left(\left\{ \sum_{i=1}^{N} \left(M_i - \bar{M} \right)\left(O_i - \bar{O} \right) \right\} \Big/ \left\{ \sum_{i=1}^{N} \left(M_i - \bar{M} \right)^2 \left(O_i - \bar{O} \right)^2 \right\}^{\frac{1}{2}} \right)^2$
Index of agreement	$IoA = 1 - \frac{\sum_{i=1}^{N} \lvert M_i - O_i \rvert}{c\sum_{i=1}^{N} \lvert O_i - \bar{O} \rvert}$, when $\sum_{i=1}^{N} \lvert M_i - O_i \rvert \le c\sum_{i=1}^{N} \lvert O_i - \bar{O} \rvert$ $IoA = \frac{c\sum_{i=1}^{N} \lvert O_i - \bar{O} \rvert}{\sum_{i=1}^{N} \lvert M_i - O_i \rvert} - 1$, when $\sum_{i=1}^{N} \lvert M_i - O_i \rvert > c\sum_{i=1}^{N} \lvert O_i - \bar{O} \rvert$ with c = 2

where:

N = total number of hourly measurements; M_i = ith predicted value; O_i = ith observed value;

$\bar{M}$ = mean of the predicted values; $\bar{O}$ = mean of the observed values.