Proceedings

Analysis of changes in pollutant concentrations levels using a meteorological normalization technique based on a machine learning algorithm †

Roberta Valentina Gagliardi [1]*, Claudio Andenna [2]

[1] Istituto Superiore di Sanità, Viale Regina Elena 299, 00161, Rome, Italy; roberta.gagliardi@iss.it
[2] INAIL-DIT, Via del Torraccio di Torrenova 7, 00133, Rome, Italy; c.andenna@inail.it
* Correspondence: roberta.gagliardi@iss.it; Tel.: (39 06 49902878)
† Presented at the 4th International Electronic Conference on Atmospheric Sciences, online, 16-31 July 2021.

Abstract: In this study, a methodological procedure combining a technique of meteorological normalization, based on a random forest algorithm, with trend analysis and the change points detections in air quality time series is developed to analyze changes in pollutant concentrations levels. Data of air pollutants and meteorological parameters, collected over the period 2013-2019 in a rural area affected by anthropic sources of air pollutants, are used to test the procedure. The results appear to be promising in revealing, in a robust way, changes in pollutant levels not clearly observable in the original data.

Keywords: air pollution; machine learning; meteorological normalization, trend analysis, change-points detection.

Citation:.;; Title. *Proceedings* **2021**, *65*, x. https://doi.org/10.3390/xxxxx

Received: date
Accepted: date
Published: date

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

It is widely documented that air pollution is a leading cause of human morbidity and mortality globally [1]. According to the World Health Organization WHO [2], ambient air pollution accounts for an estimated 4.2 million premature deaths per year due to stroke, heart disease, lung cancer, acute and chronic respiratory diseases and 91% of the world's population live in places where air pollution levels exceed WHO Air Quality Guidelines limits [3]. In the European context, Italy presents several critical issues in terms of high-pollution areas [4], prompting the European Commission to call Italy to comply with the requirements of Directive 2008/50/EC on ambient air quality and cleaner air for Europe [5] with regard to particulate matter [6].

To design effective and well targeted strategies aimed at preventing or reducing health damages associated with the exposure to the atmospheric pollution, accurate information on the reals levels and on the trends of pollutants concentrations are required. To this purpose, the well known confounding effects of meteorology on the observed pollutants concentrations, occurring over multiple scales in time and space, must be considered [7], [8], [9]. Among the techniques accounting for changes over time in the air quality time series due to meteorology, referred as "meteorological normalization techniques", a new approach based on machine learning (ML) predictive algorithms has recently emerged [10], [11], which basically reduces air quality time series variability with statistical modelling. Once the confounding weather effects have been removed, further and more robust statistical evaluations can be carried out in the resulting normalised time series. For example, the trend patterns analysis, (i.e. concentration changes over a period of time [12]), and the detection of change points (i.e. unexpected, structural, changes in time series data properties, such as the mean or variance [13]), can be investigated in a more reliable way.

Aims of the work is to develop a methodological procedure to account for the confounding effects of weather variability in air quality time series concentrations and to more accurately explain the variability in the measured pollutant concentrations.

To this end, we developed a three-stage methodology. First, the effects of local weather in the air quality time series were removed using a technique of meteorological-normalization, based on a random forest (RF) ML algorithm. Secondly, trend analysis and change points detection were carried out to assess changes in the normalized signal. Finally, results obtained by the first two stages were jointly examined with the publicly available metadata to formulate some hypothesis on the potential link between the observed pollutant concentrations and the anthropic sources existing in the area. This procedure was applied on a data set comprising daily averaged data of air pollutants concentrations and meteorological parameters as well as temporal variables. Data were collected, over the 2013-2019 period, in a semi-rural area of Southern Italy interested by an anthropic source of air pollutants potentially influencing air quality. The obtained results appear to be promising in producing a reliable estimate of actual changes in the pollutant concentrations time series for use in air pollution exposure assessment studies.

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) (Figure 1); more details on the examined area can be found in [14]. The valley is characterized by the presence of the largest on-shore western European reservoir of crude oil and gas and of an oil pre-treatment plant (identified as Centro Olio Val d’Agri – hereafter COVA) in a populated area. The plant produces conveyed, diffuse and fugitive emissions of gases and particulate, which can affect the air quality and potentially pose a health risks for the population living in the area. Furthermore, the industrial processes taking place in the plant involve dangerous substances (toxics and flammables) for man and environment. An air quality control network, consisting of five monitoring stations, is operating in the area, managed by the Environmental Protection Agency of the Basilicata Region (ARPAB). 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.

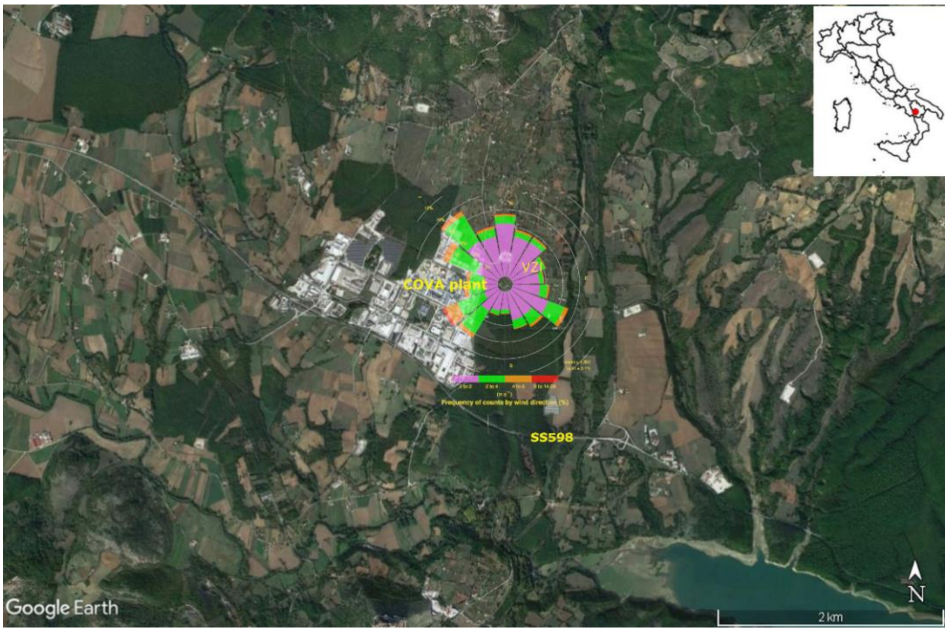


Figure 1**.** Map of the study area: the VZI monitoring site, the COVA plant, the national road SS598 and the wind rose based on the hourly data at the VZI station over the study period (2013-2019).

2.2. *Observational dataset*

Four gaseous pollutants, namely nitrogen oxides (NO_x), sulfur dioxide (SO_2), carbon oxide (CO) and hydrogen sulfide (H_2S) were selected for the analysis as proxies of anthropic sources existing in the area. For these pollutants, a strong evidence of respiratory and cardiovascular health effects is documented [15], [16]. Hourly data of NO_x, SO_2, CO, H_2S and of several meteorological variables (respectively: temperature (T), atmospheric pressure (P), relative humidity (RH), wind direction (wd) and wind speed (ws)), were downloaded from the official website of ARPAB [17] and combined to form the whole data set used. Overall, a data set consisting of more than 356000 observations covering the 2013-2019 period was set up. The time series of all predictors considered respected the required 75% proportion of valid data. Subsequently, the data were daily averaged and a set of other time-based variables was added to create the final data set for the RF models development. In particular, the day of the week, the Julian day (number of days since 1 January, 'Jday') and the date Unix of the observations (number of seconds since 1 January 1970, 'trend') were included in the model development as proxies for local traffic sources and to account for seasonal and long-term variability, respectively.

2.3. *Methodological approach*

The methodological approach to assess changes in pollutant concentrations levels, adopted in the present study, consists of the following main steps: i) for each pollutant a RF model was developed and, once its performances and interpretability have been analyzed to ensure its reliability, the meteorological normalization of the concentrations predicted by the RF model was carried out. ii) After that, the estimation of the main change-points time location in the normalized signal and the trend analysis were performed. iii) Combining the results of the previous stages with the available metadata, some hypothesis on the potential link between the normalised time series and specific events were formulated.

2.3. 1 Meteorological normalization procedure

The strategy for the meteorological normalization follows the work described in [18], as subsequently implemented in [19] and [20], and was based on two steps: first, a RF model was built and validated for each of the pollutants analyzed in the present study; second, the meteorological normalization of the predicted pollutants concentrations was carried out.

In the development of each RF model (theoretical insight can be found [21]) the pollutant included in the dataset represented the dependent variable (or target) while meteorological and time-dependent features represented the explanatory variables (or predictors). The 80% of the whole observed dataset randomly sampled (training dataset) was used to build up the prediction model. The remaining 20% (testing dataset) was used to test the prediction accuracy of the model. The best model for each pollutant was built on the training dataset using the best combination of the tuning parameters selected on the basis of the R^2 metric as evaluated on the testing dataset. The tuning parameters used in the work are the number of predictors randomly sampled to determine each split (mtry) and the minimum number of observations in a terminal node (min node size). The number of trees (n trees) was set at 1000. The RF model have an inherent procedure producing the relative importance of predictors that is, the measure of the impact of each feature on the accuracy of the model. Thus, the relative importance resulting from the developed models was analyzed to identify the most important predictors. The performances of the selected optimal RF model were fully assessed by comparing predicted and observed pollutants concentrations values using a set of statistical indicators [22] evaluated on the testing dataset (see Annex 1 for the relevant equations).

Once established that the RF model explained an adequate amount of variance in the predicted air quality variable, it was used to predict the pollutant concentrations resampling only the meteorological explanatory variables from the whole study period without replacement and randomly allocating them to a dependent variable observation. The advantage of this procedure is that the normalization process involves only the weather conditions but not the seasonal or weekly variations, so that the resulting normalized series is more closely related to emissions changes rather than changes due to meteorological effects. This procedure was repeated a number of times (300), then all the predictions were aggregated using the arithmetic mean to obtain the meteorological normalized concentration.

2.3. 3 Trend and structural change analysis

The goal of determine if there is a trend in the normalized concentrations over time was achieved using the Theil-Sen regression technique, which calculates the median slope of all possible slopes that may occur between the data points [23]. 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. As far as the trend analysis is concerned, the unadjusted trends we estimated are the product of both emission and meteorological changes, while the weather-adjusted trends remove the influence of weather changes on air quality. Consequently, the difference between the unadjusted and weather-adjusted trends reflects the impact of meteorological changes or weather penalties.

For a more in-depth analysis of the trend so achieved, an investigation about the structural changes in the normalised time series was carried out [24]. In the present study, we adopted the Wild Binary Segmentation (wbs) change point detection method [25] to detect the number and potential locations of change points with no prior assumptions.

2.3. 4 Metadata analysis

Finally, an attempt was made to acquire the available appropriate metadata allowing to properly interpreting the results. Data concerning plant operation, the timing of significant events related to the plant activities and the traffic flows in the Agri Valley were examined. The former were downloaded from official sources (i.e. the websites of the company that manages the plant) [26], while the traffic flows of heavy and light vehicles concerning the national road SS 598 were provided by the Azienda Nazionale Autonoma delle Strade (ANAS) [27].

All data loading, processing, analysis, statistical modelling and visualization were performed in the R version 4.1.0 (R Foundation for Statistical Computing, Vienna, Austria). It was mainly used the *Openair* package for air quality and trend analysis [28], the *rmweather* package [11] [18] for the meteorological normalisation, with the underlying *ranger* package [29] and *tuneRanger* package [30] for the development and tuning of the RF model and the *wbs* package [31] for change points analysis.

3. Results and discussion

3.1. Statistical analysis

The descriptive statistics per year and pollutant is reported in Table 1.

Table 1. Statistical summary of hourly data of NO_x, SO_2, CO, H_2S registered at the VZI monitoring station from January 2013 to December 2019. Mean concentration in bold and, in rounded brackets, the min and maximum values.

Year	NO_x µg/m^3	SO_2 µg/m^3	CO mg/m^3	H_2S µg/m^3
2013	**14.98** (0.00-118.29)	**5.63** (0.50-350.90)	**0.338** (0.00-1.10)	**2.18** (0.28-241.61)
2014	**20.34** (0.75-143.07)	**3.28** (0.00-195.20)	**0.370** (0.00-1.90)	**3.58** (0.69-43.85)
2015	**20.15** (0.00-186.07)	**7.00** (0.00-247.10)	**0.332** (0.00-1.30)	**2.86** (0.28-219.27)
2016	**16.84** (0.00-133.44)	**6.11** (0.03-175.80)	**0.424** (0.05-1.64)	**2.96** (0.30-272.35)
2017	**16.35** (2.02-117.05)	**6.08** (0.38-378.92)	**0.393** (0.00-2.11)	**3.08** (0.54-75.61)
2018	**13.03** (0.19-122.50)	**6.10** (0.09-281.03)	**0.381** (0.00-1.44)	**3.72** (0.08-62.56)
2019	**14.66** (0.26-105.57)	**3.60** (0.11-277.95)	**0.377** (0.00-2.23)	**3.01** (0.29-76.19)
All years	**16.63** (0.00-186.06)	**5.41** (0.00-378.92)	**0.374** (0.00-2.23)	**3.06** (0.08-272.35)

For regulated pollutants, time series analysis showed a general compliance with the limits set for by the existing national [32] and European legislation [5]. It is worth noting that, for the sole Agri Valley, a regional law [33] identifies limit values more stringent than those in force at national level for SO_2 and H_2S that are considered proxies of local hydrocarbon emissive processes. This law sets at 280 µg/m^3 and 100 µg/m^3 the hourly and daily limit values for the protection of human health for SO_2, and 32 µg/m^3 the daily limit for H_2S. The hourly limit value for SO_2 rarely exceeded these limits and each time in different years. As far as the climate is concerned, the cold and rainy winters as well as cool summers with frequent rainfall [34], typically registered in the area, define an area at sub-continental climate. Based on the analysis of wind data, the mean value of ws was 1.8 ms^{-1}, with the higher values generally measured during daytime. The wind rose, superimposed on the map in Figure 1, showed a prevailing wind direction from the SW to NW sector, over the period ranging between January 2013 and December 2019.

3.2. RF models development and performances

For each examined pollutant, the RF model, trained with the selection of the tuning parameters listed in Table 2, took the form shown by equation 1:

$$pollutant \sim rf(T, H, ws, wd, P, Jday, weekday, trend), \tag{1}$$

where *rf* is the function implementing the random forest algorithm in the R software environment.

Table 2. RF model tuning parameters for each of the selected pollutant.

pollutant	mtry	min nod size	n trees
NO_X	4	2	1000
SO_2	4	6	1000
CO	7	2	1000
H_2S	5	4	1000

The RF models performances were evaluated through the statistical indicators, whose resulting values were summarized in Table *3*.

Table 3. Statistical indicators of RF model 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.

pollutant	R^2	MBE [μg/m^3]	MAE [μg/m^3]	RMSE [μg/m^3]	IoA
NO_X	0.723	0.380	3.700	5.406	0.723
SO_2	0.458	0.177	1.519	3.201	0.726
CO	0.704	0.004	0.057	0.077	0.757
H_2S	0.683	0.069	0.366	0.700	0.806

The R^2 values show that the RF models can explain about the 70% of the total NO_x, CO and H_2S variability, while the model showed a moderate explanatory ability for SO_2 (R^2 values of 0.46).

The relative importance of the selected predictors for the examined pollutants are presented in Figure 2.

The overall contribute of the top four predictors explained over 85% of the variance for NO_2 and SO_2, and over 90% of the variance for CO and H_2S. For SO_2, CO and H_2S, the temporal variables, i.e. trend and Jday, were the most important predictors, indicating in the seasonality and long-term trend the strongest driving features.

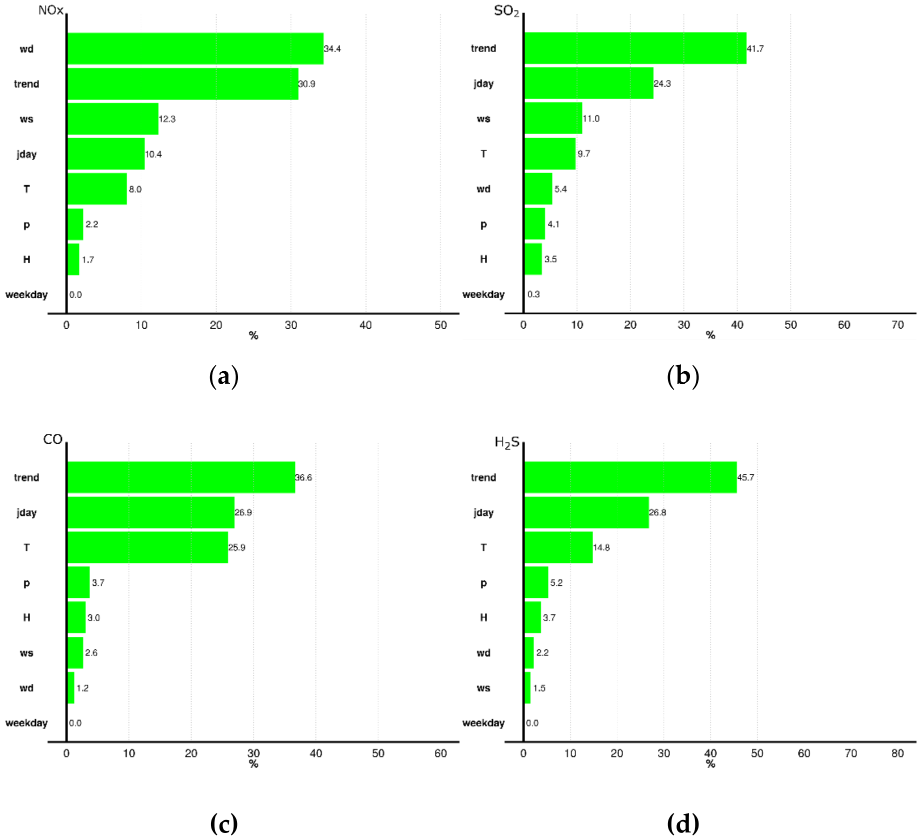
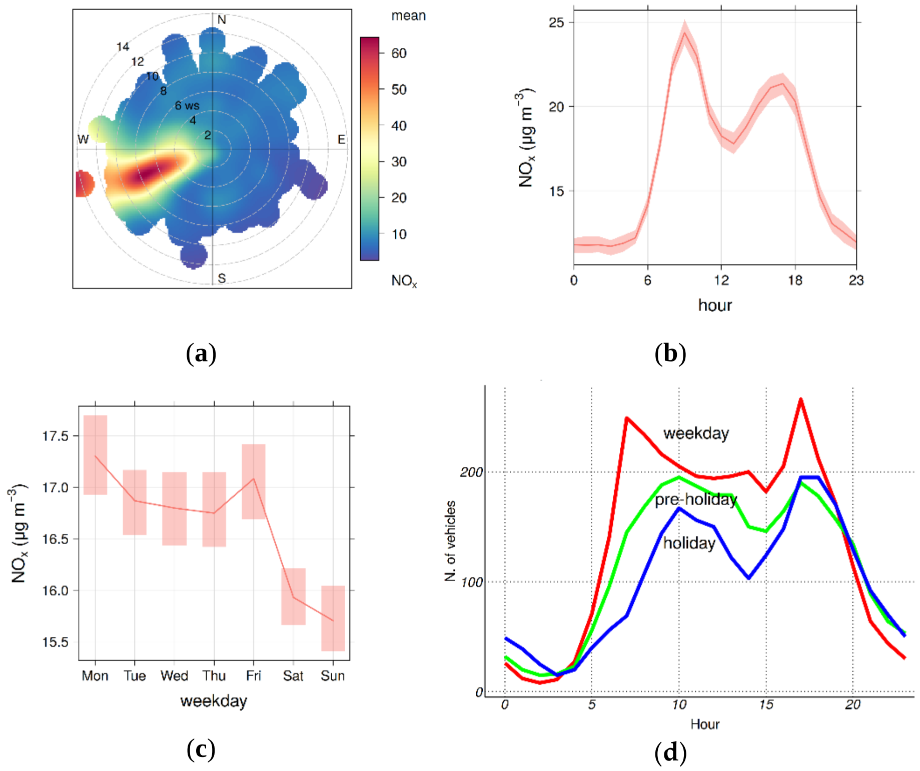


Figure 2. Relative importance of predictors for each of the selected pollutants.

The most important contribute to NO_x variability, instead, was due to the wind direction, closely followed by trend, and to a lesser extent by ws and Jday. It is worth looking more closely to the dependence of NO_x from wd. The bivariate polar plot (Figure 3a) confirmed the strong directionality of NO_x concentrations associated to winds from WSW, that is in the direction of both several of the COVA plant conveyed emissive sources and the SS598 national road. The hypothesis of a traffic contribution to NO_x was supported by the analysis of the daily and weekly NO_x pattern (Figure 3 b and c). The former tends to be significantly bimodal (higher concentrations in the early morning and late afternoon coinciding with the commuting hours). The latter shows a clear decrease of NO_x concentrations on Saturday and Sunday when traffic is usually lower. Both these patterns were also confirmed by the analysis of the metadata concerning the traffic flows of cars and heavy vehicles for the national road SS598 provided by ANAS for the year 2019 (Figure 3d).



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

(**c**) (**d**)

Figure 3. Polar plot (a), daily (b) and weekly (c) profiles of hourly NO_x concentrations. Also shown on the plots b) and c) is the 95 % confidence interval in the mean. d) Average hourly trend of traffic flows of the national road SS598 in 2019.

3.3. Meteorological normalized air pollutants time series

Daily concentrations of the observed and normalized data for NO_x, SO_2, CO and H_2S are shown in Figure 3. Also shown in the figure is a blue solid line representing the line joining the wbs change-points. As result of the meteorological normalization process, clear differences can be seen between the observed and normalized concentrations with the latter being a much smoother data series. Trend in the normalized pollutants concentrations was less volatile and noisy compared to the observed values and showed the extent to which changes in emissions influence the pollution level measured at the examined site. Moreover, number and location of change points identified by the wbs methods appears to detect the main structural changes in the normalized time series. Linking these structural changes with specific events through the available metadata should allow formulating hypotheses about what originated them. It is worth dwelling on two specific events corresponding to the periods represented by the grey areas in the Figure 4. By means of the available metadata at [26], it is known that the first corresponds to a plant shutdown, from April to August 2016, for a judicial investigations.

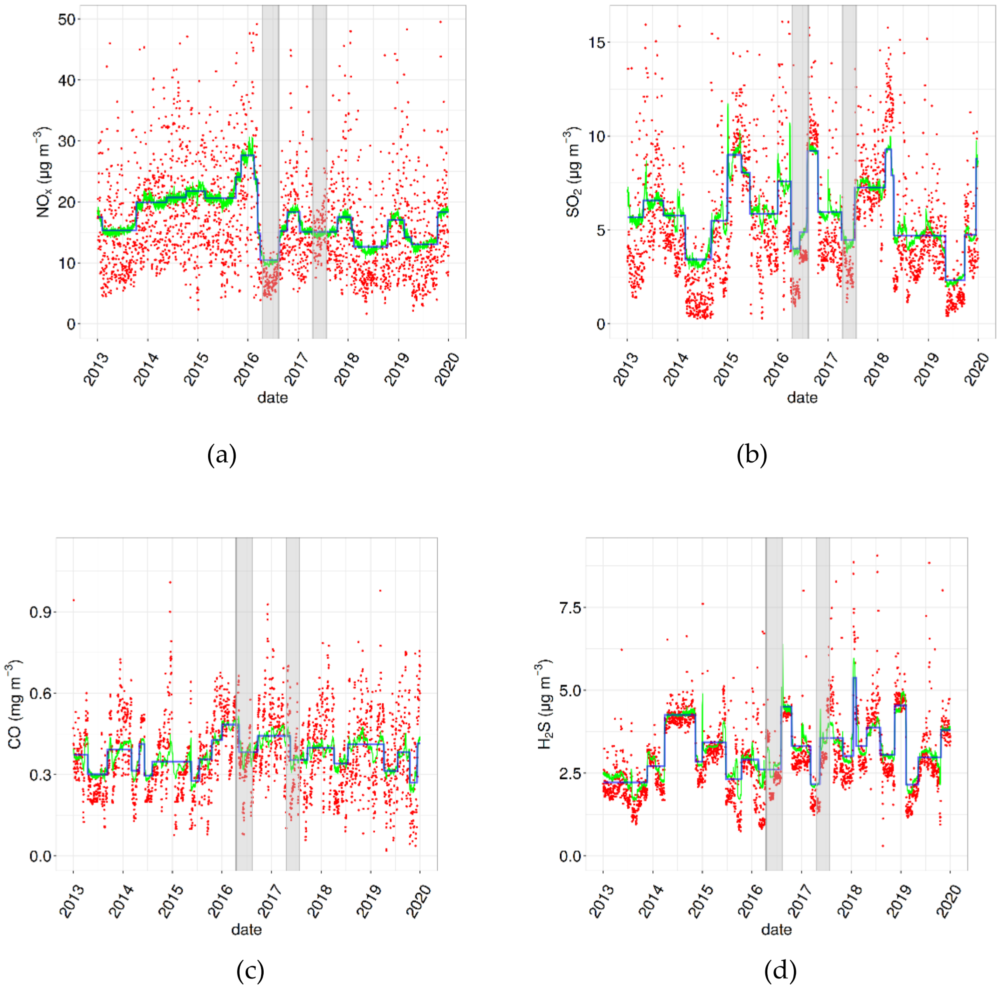


Figure 4. Daily averages of observed (red dots) and meteorologically normalized concentrations (green lines) of (a) NO_x, (b) SO_2 , (c) CO and (d) H_2S. The blue solid line represents the line joining the wbs change-points, while the grey areas show the periods of COVA plant shutdowns.

The second consists in another plant shutdown, from April to July 2017, due to a major accident, caused by the release of hydrocarbons from a storage unit. As far as the SO_2, and CO signals are concerned, a decrease in concentrations corresponding to these periods can be observed in Figure 4. With respect to the NO_x pollutant, a strong correspondence was found between the normalized concentrations trend and the event occurred at the COVA plant in 2016. The lack of correspondence with the event registered in 2017 may be due to other sources contributing to the observed NO_x level. H_2S, instead, seems to be less affected by these closures period, as expected, since this pollutant is representative of the fugitive emissions from oil tanks and piping of the COVA plant.

The results seems to confirm the goodness of this approach in identifying an atmospheric response in the observed data after an unplanned event or a change in emission sources. However, more stringent evidences are desirable to confirm this hypothesis, due to extreme complexity of the overall effects of the start/stop plant procedures on air quality.

Finally, Table 4 summarizes the results of the Theil-Sen regression analysis. For NO_x, a statistically significant trend for normalized and observed data were found, while less statistically significant normalized trends were found for H_2S and CO ($p<0.05$) and SO_2 ($p<0.1$).

Table 4. Theil-Sen slope and 95 % confidence intervals of the observed and normalized pollutants concentrations. The symbols shown next to the square bracket relate to how statistically significant the trend estimate is: $p < 0.001 = * * *$, $p < 0.01 = **$, $p < 0.05 = *$ and $p < 0.1 = +$.

pollutant		Theil-Sen slope (μg m^{-3} year^{-1})	95% confidence interval
NO_x	observed	-0.66	[-1.13, -0.27]***
	normalized	-0.65	[-1.07, -0.39]***
SO_2	observed	-0.03	[-0.32, 0.26]
	normalized	-0.19	[-0.39, 0.02]+
CO	observed	0.01	[0.00, 0.02] *
	normalized	0.01	[0.00, 0.01] *
H_2S	observed	0.12	[0.02, 0.20] *
	normalized	0.11	[0.04, 0.17] *

The comparison between the observed and normalized slopes of each pollutant show a generally scarce influence of the weather conditions to the trend of the pollutants. This result appears to be more stringent in the case of NO_x due the high statistical significance of the Theil-Sen analysis. This is consistent with the information deduced from the results illustrated above, which indicate in the local NO_x sources, mainly the COVA plant and the traffic, the main drivers of NO_x variability.

4. Conclusions

Ambient air pollution remains a great challenge for sustainable development and public health safeguard. Meteorological influences upon air quality trend analysis can complicate the evaluations of air pollution control efforts. The joined interpretation of the observed data of air pollutants, of the simulations produced by the RF models used to remove the effect of meteorology, and the subsequent statistical analysis, adopted in the present study, represents an effective tool to assess and quantify changes in air pollution. In particular, the technique of the meteorological normalization allows discriminating the contribution of meteorology from those of source's emissions, while the wbs method seems to be promising in correctly following main changes in the normalized pollutants concentrations. Since the RF models are data driven, caution is required when generalizing the results obtained to different conditions and/or sites. Moreover, a deeper knowledge of the study area characterized by an complex orography, a more comprehensive collection of the available metadata as well as a wider awareness of all natural or anthropic events affecting local air quality, can be obtained only through a close collaboration with the local environmental and health authorities who are the most informed on the criticalities of the examined territory.

Overall, our results show that the adopted procedure can improve qualitative trend assessment of observed air pollutants data and help in revealing shifts in pollutants levels that can not be clearly seen in the original data, so providing crucial information for the implementation of effective strategies to prevent the health impact of air pollution.

Author Contributions: All individuals listed as authors have equally contributed to the present work.

Funding: This research received no external funding.

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

Conflicts of Interest: The authors declare no conflict of interest.

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	$RMSE = \sqrt{\left(\frac{\sum_{i=1}^{N}(M_i - O_i)^2}{N}\right)}$
Coefficient of Determination	$R^2 = \left(\{\sum_{i=1}^{N}(M_i - \bar{M})(O_i - \bar{O})\} \Big/ \{\sum_{i=1}^{N}(M_i - \bar{M})^2(O_i - \bar{O})^2\}^{\frac{1}{2}} \right)^2$
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 $\sum_{i=1}^{N}\lvert M_i - O_i \rvert \leq c\sum_{i=1}^{N}\lvert O_i - \bar{O} \rvert$ $IoA = \frac{c\sum_{i=1}^{N}\lvert O_i - \bar{O} \rvert}{\sum_{i=1}^{N}\lvert M_i - O_i \rvert} - 1$, 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	

References

[1] Khomenko, S.; Cirach, M.; Pereira-Barboza, E.; Mueller, N.; Barrera-Gómez, J.; Rojas-Rueda, D.; de Hoogh, K.; Hoek, G.; Nieuwenhuijsen, M. Premature mortality due to air pollution in European cities: a health impact assessment. *Lancet*. Published online January 19, **2021** https://doi.org/10.1016/S2542-5196(20)30272-2.

[2] https://www.who.int/health-topics/air-pollution *(Accessed on 10 Jan* ***2021****).*

[3] World health Organization. Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide. **2005**.

[4] Donateo, A.; Villani, M.; Lo Feudo, T.; Chianese, E. Recent Adavences of Air Pollution Studies in Italy. *Atmosphere.* **2020**.

[5] Directive 2008/50/EC on ambient air quality and cleaner air for Europe, Official Journal of the European Union L 152/1, **2008**.

[6] https://ec.europa.eu/commission/presscorner/detail/IT/INF_20_1687.

[7] Elminir, H.,K. Dependence of urban air pollutants on meteorology. *Science of the Total Environment.* **2005,** *350*, 225–237.

[8] Jones, A. M.; Harrison, R. M.; Baker, J. The wind speed dependence of the concentrations of airborne particulate and NOx. *Atmospheric Environment.* **2010**, *44*, 1682-1690.

[9] Kinney, P. Climate Change, Air Quality, and Human Health. *American Journal of Preventive Medicine.* **2008**, *35*, 459-467.

[10] 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.* **2020**, *20*, 11119-11141.

[11] Grange, S.; Carslaw, D., Lewis, A.; Boleti, E.; Hueglin, C. Random forest meteorological normalisation models for Swiss PM10 trend analysis. *Atmos. Chem. phys.* **2018**, *18*, 6223-6239.

[12] Guerreiro, C.,B.,B.; Foltescu, V.; de Leeuw, F. Air quality status and trens in europe. *Atmospheric Environment.* **2014**, *98*, 376-384.

[13] Xiong, L.; Guo, S. Trend test and change-point detection for the Yichang hydrological station annual discharge series of the Yangtze River at the. *Hydrological Sciences–Journal–des Sciences Hydrologiques.* **2004**, 99-111.

[14] Gagliardi, R., V.; Andenna, C. A Machine Learning Approach to Investigate the Surface Ozone Behaviour. *Atmosphere.* **2020**, *11.*

[15] https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/health-impacts. (*Accessed on 10 Jan **2021***).

[16] Mousa. H., A., L. Short-term effects of subchronic low-level hydrogen sulfide exposure on oil field workers. *Environ Health Prev Med.* **2015**, *20*, 12-17.

[17] www.arpab.it/opendata/q_aria_serie.asp. (*Accessed on 10 Jan. **2021***).

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

[19] Vu, T., V.; Shi, Z.; Cheng,J.; Zhang, Q.; He, K.; Wang, S. Harrison, R., M. Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique. *Atmos. Chem. Phys.* **2019**, *19*, 11303–11314.

[20] 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.* **2021**, *7*, 1-10.

[21] Breiman, L. Random Forests. *Machine Learning.* **2001**, *45*, 5-32.

[22] Sayegh, A., S.; Munir, S.; Habeebullah, T., M. Comparing the Performance of Statistical Models for Predicting PM10. *Aerosol and Air Quality Research.* **2014**, *14*, p. 653–665.

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

[24] Aminikhanghahi, S.; Cook, D., J. A Survey of Methods for Time Series Change Point Detection,» *Knowl Inf Syst.* **2017**, *51*, 339-367.

[25] Fryzlewicz, P. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics.* **2014**. *42*, 2243-2281.

[26] https://www.eni.com/eni-basilicata/news/2021-elenco-news.page.

[27] https://www.stradeanas.it/it/strade.

[28] Carslaw, D. Openair-An R package for air quality data analysis. *Environ Modell Softw.* **2012**, 52-61.

[29] Wright , M.; Ziegler, A. ranger: a fast implementation of random forests for high diemnsional data in <C++ and R. *J. Stat. Software.* **2017**, 77, pp. 1-17.

[30] Probst, P.; Wright, M; Boulestei, A. Hyperparameters and Tuning Strategies for Random Forest, https://arxiv.org/pdf/1804.03515.pdf **2019**.

[31] Baranowski, R.; Fryzlewicz, P. wbs: wild binary segmentation for multiple change-point detection. R package version 1.1. **2014**.

[32] D. Lgs. 155/10, Attuazione della Direttiva 2008/50/CE relativa alla qualità dell'aria ambiente e per un'aria più pulita in Europa., Gazzetta Ufficiale **2010**.

[33] DGR della Regione Basilicata del 6 agosto **2013**, n. 983.

[34] http://www.prefettura.it/potenza/contenuti/Pee_centro_olio_val_d_agri_di_viggiano_edizione_2013-64403.htm. (*Accessed on 30 Mar. **2021***).