Environmental Pollution 267 (2020) 115471

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

Machine learning reveals that prolonged exposure to air pollution is associated with SARS-CoV-2 mortality and infectivity in Italy[☆]

Roberto Cazzola Gatti [a, b, *], Alena Velichevskaya [a], Andrea Tateo [c], Nicola Amoroso [c, d, 1], Alfonso Monaco [e, 1]

[a] *Biological Institute, Tomsk State University, Russia*
[b] *Konrad Lorenz Institute for Evolution and Cognition Research, Austria*
[c] *Università degli Studi di Bari "A. Moro", Dipartimento Interateneo di Fisica, Bari, Italy*
[d] *Università degli Studi di Bari "A. Moro", Dipartimento di Farmacia - Scienze del Farmaco, Bari, Italy*
[e] *Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Bari, Bari, Italy*

ARTICLE INFO

Article history:
Received 15 June 2020
Received in revised form
28 July 2020
Accepted 19 August 2020
Available online 21 August 2020

Keywords:
SARS-CoV-2
Pollution
PM2.5
Italy
Mortality
Industry
Farms

ABSTRACT

Air pollution can increase the risk of respiratory diseases, enhancing the susceptibility to viral and bacterial infections. Some studies suggest that small air particles facilitate the spread of viruses and also of the new coronavirus, besides the direct person-to-person contagion. However, the effects of the exposure to particulate matter and other contaminants on SARS-CoV-2 has been poorly explored. Here we examined the possible reasons why the new coronavirus differently impacted on Italian regional and provincial populations. With the help of artificial intelligence, we studied the importance of air pollution for mortality and positivity rates of the SARS-CoV-2 outbreak in Italy. We discovered that among several environmental, health, and socio-economic factors, air pollution and fine particulate matter (PM2.5), as its main component, resulted as the most important predictors of SARS-CoV-2 effects. We also found that the emissions from industries, farms, and road traffic - in order of importance - might be responsible for more than 70% of the deaths associated with SARS-CoV-2 nationwide. Given the major contribution played by air pollution (much more important than other health and socio-economic factors, as we discovered), we projected that, with an increase of 5—10% in air pollution, similar future pathogens may inflate the epidemic toll of Italy by 21—32% additional cases, whose 19—28% more positives and 4—14% more deaths. Our findings, demonstrating that fine-particulate (PM2.5) pollutant level is the most important factor to predict SARS-CoV-2 effects that would worsen even with a slight decrease of air quality, highlight that the imperative of productivity before health and environmental protection is, indeed, a short-term/small-minded resolution.

1. Introduction

The year 2020 has started with Chinese authorities alerting WHO that several cases of unusual pneumonia had appeared in the area of Wuhan (Huang et al., 2020; Wang et al., 2020). In a few weeks, despite travel restrictions, Italy became the first Western country with the most serious outbreak of SARS-CoV-2 (Chinazzi et al., 2020). After six months and millions of people infected worldwide, Italy still ranks first among European Union's countries in the death toll (Dong et al., 2020). During the initial pandemic wave, Italy has immediately adopted social distancing measures but it appears clear that the casualties due to this new coronavirus spread have not affected each Italian region equally (Livingston and Bucher, 2020).

Although it is well-known the dynamics of epidemics can be shaped and explained by a combination of several health and socio-economic factors (Onder et al., 2020), the role played by environmental causes is still poorly explored. Nonetheless, there is increasing support on the link between severe viral respiratory disease and air pollution (Chauhan and Johnston, 2003; Ségala et al., 2008; Guan et al., 2016; Yao et al., 2019; Ogen, 2020). For instance, a previous study conducted in northern Italy showed the clinical severity of bronchiolitis in children living in highly polluted

[☆] This paper has been recommended for acceptance by Da Chen.
[*] Corresponding author
E-mail address: roberto.cazzola-gatti@kli.ac.at (R. Cazzola Gatti).
[1] Equal last author contribution.

https://doi.org/10.1016/j.envpol.2020.115471
0269-7491/© 2020 Elsevier Ltd. All rights reserved.

areas and that PM10 exposure is associated with increased hospitalizations for respiratory syncytial virus bronchiolitis among infants in Lombardy, Italy (Carugno et al., 2018). Accordingly, the geographical distribution of SARS-CoV-2 mortality (Fig. 1A) and air quality (Fig. 1B) in Italy unveils an alarming pattern. In fact, it seems no coincidence that the area with the worst air quality and most severely affected by the new coronavirus in Italy is the Po Valley. The death toll of regions such as Lombardy, Emilia-Romagna, Piedmont, and Veneto almost reaches 80% of the total national deaths (Fig. S1). Big cities in the Po Valley regions show an increased fatality rate between 150 and 250% (ISTAT, 2020). The heavy anthropic activity in this area, known as the "Industrial Triangle", deriving from a high density of factories, vehicular traffic, and intensive farming and agriculture (Ciccarelli and Fenoaltea, 2013), produces significant emissions of air pollutants. These cannot easily dissolve because air recycling is reduced by the specific topography (a plain surrounded by the Alps) and climatic conditions (high humidity and weak winds) that trap fine particulates (Bigi et al., 2012).

Several authors suggested that COVID lockdown measures reduced anthropogenic impacts in many areas of the world (Dutheil et al., 2020; Cazzolla Gatti, 2020; Muhammad et al., 2020; Gautam, 2020; Bherwani et al., 2020; Sharma et al., 2020). However, the rationale behind our study was motivated by some evidence that

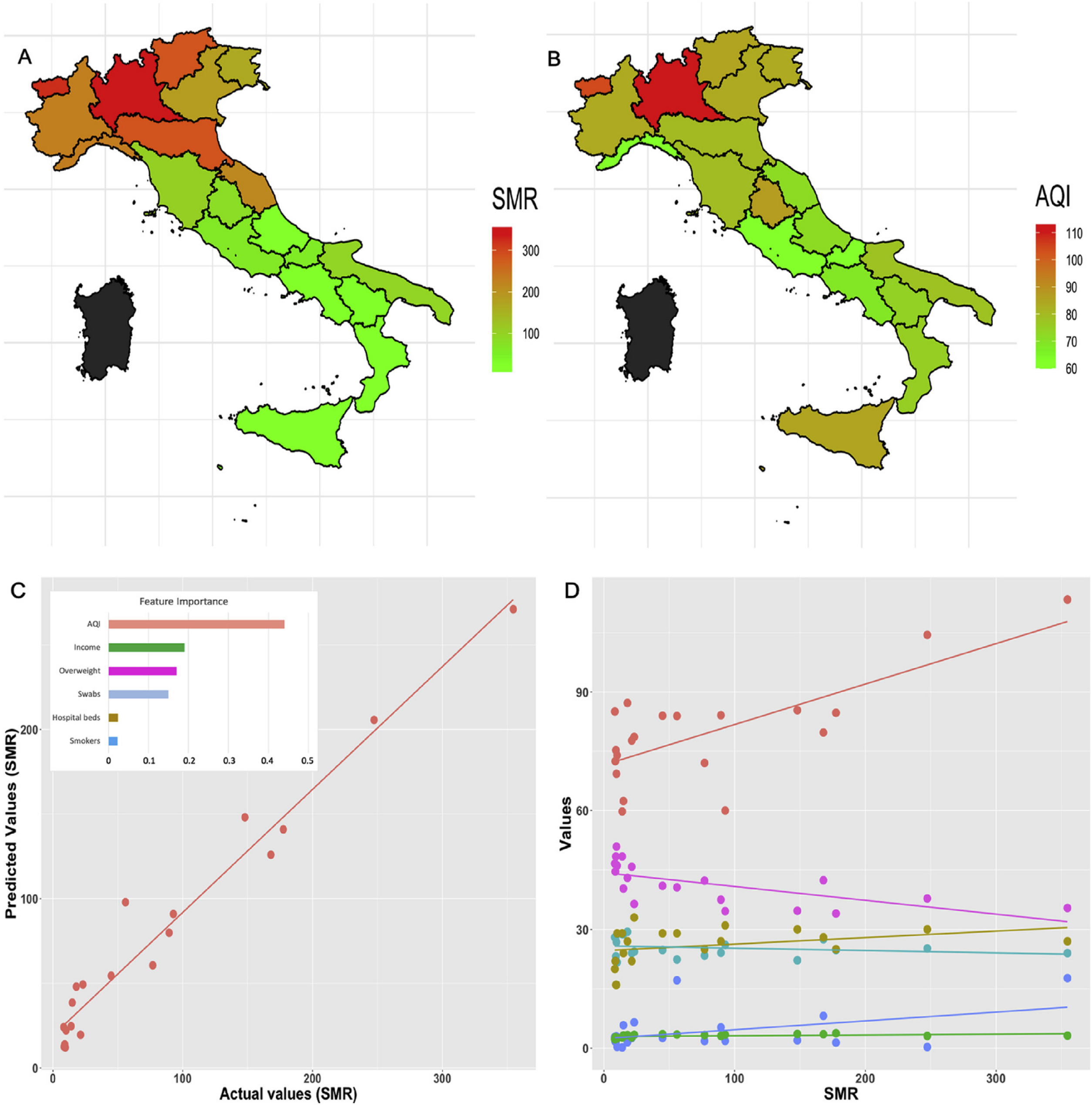


Fig. 1. The geographical distribution of air pollution (AQI) and the role of health, environmental, and socio-economic factors on SARS-CoV-2 mortality (SMR) in Italy. (A) SMR values of Northern Italian regions are much higher than in the Southern regions. (B) The situation is strikingly similar when considering pollution over the Italian Peninsula; for this representation, Air Quality Index (AQI) is used as a proxy of pollution (where high values mean higher pollution). AQI and SMR values are in percent. (C) The agreement between the SMR actual values and the Random Forests predictions ($R^2 = 0.95$), which follows a positive linear trend; in the smaller panel: the importance of the different factors with AQI ranking as the most important one. (D) Also when individually analysed (colours of points and lines as in the in part C's inset), AQI is the only factor showing a significant ($P < 0.05$, Bonferroni correction) and positive correlation with SMR ($R^2 = 0.54$). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

air pollution plays a role in spreading the viruses, emerged from previous studies on other epidemics (Ciencewicki and Jaspers, 2007; Wong et al., 2010; Nenna et al., 2017), which was also advanced for SARS-CoV-2 expanding virulence in specific areas of Italy (Conticini et al., 2020; Sciomer et al., 2020). Research conducted in Beijing showed that small-size particulate (PM2.5) in the air directly influences the transmission of influenza virus (Liang et al., 2014). In fact, fine particulate matter remains suspended longer in the air than heavier particles and its fine size facilitates virus suspension, their long-range diffusion, and the deep penetration into the bronchi of the lungs (Brugha and Grigg, 2014). In people chronically exposed to air pollution, this may induce chronic inflammation of the respiratory airways, which leads to excessive mucus production and decreased ciliary activity facilitating the development of severe respiratory diseases after viral infections (Xing et al., 2016). However, very few studies have evaluated the effects of prolonged exposure to air pollution on SARS-CoV-2019 susceptibility.

As far as we know, our study is the first one addressing this problem within a quantitative framework which allows the evaluation of the statistical association between pollution and SARS-CoV-2 effects.

Preliminary hypotheses advanced in the USA (Wu et al., 2020) and in Italy (Fattorini and Regoli, 2020), suggested that long-term exposure to air pollution and, particularly, fine particles may be linked to the COVID-19 death rate. Although a vivid debate has involved the scientific community, a systematic evaluation of the association between SARS-CoV-2 infectivity and air pollution factors is still lacking. Some authors highlighted that these preliminary studies had to be interpreted as more hypothesis-generating rather than confirmatory and called for future research to clarify the role of air pollution and the other factors, which may have mutually contributed to the diffusion of SARS-CoV-2 (Fattorini and Regoli, 2020). Moreover, these early studies linearly investigated the association between the pandemics and individually-taken factors.

Here we examined the possible reasons why the new coronavirus differently impacted on Italian regional and provincial populations with the help of artificial intelligence through a multivariate approach, trying to incorporate the effects of several factors and their interactions in a comprehensive non-linear model. Using standardized mortality and infectivity rates and including other important determinants such as age, lifestyle, and socio-economic factors, this works provides the first attempt to quantitative and more comprehensive modelling of SARS-CoV-2 pandemics.

With a machine learning algorithm, we were able to assess the association of pollution and environmental factors with the SARS-CoV-2 diffusion in Italy. In particular, we studied the importance of prolonged exposure to air pollution for mortality and positivity rates. We included in our analysis other factors potentially linked to the pandemics, such as the number of smokers, the level of obesity and overweight, the mean annual income per family, the number of public hospital beds, and the number of viral tests performed. Then, we determined the specific contribution of different components and sources of air pollution to the pandemics diffusion and severity. Finally, we projected the effects of a decrease in air quality over future pandemics.

2. Materials and methods

2.1. Data collection and standardization

Epidemiological data on positivity and mortality at the regional scale (20 Italian regions out of 21 total, excluding Sardinia for which air quality measurements were not available), and total cases at provincial scale (99 Italian provinces out of 107, excluding 5 from Sardinia and 3 - Fermo, Imperia, and Ragusa - from which air quality data were not available) were collected from the Italian Civil Protection's data repository (https://github.com/pcm-dpc/COVID-19/tree/master/dati-regionii) and (https://github.com/pcm-dpc/COVID-19/tree/master/dati-province).

In this study, we used the term "positivity" for the total number of people tested positive on SARS-CoV-2 swabs updated at June 06, 2020, in each Italian region; "mortality" for the number of people dead because of SARS-CoV-2 updated at June 06, 2020, in each Italian region; and "total cases" for the total number of SARS-CoV-2 cases, which is the sum of dead people, tested positive, and recovered, updated at June 06, 2020, in each Italian province.

We collected the last 5-year (2015–2019) Air Quality Index (AQI) from data provided by http://moniqa.dii.unipi.it/, which allowed us to synthetically represent the state of air quality at provincial and regional scales considering at the same time the data of several atmospheric pollutants. This index represents a unitless indicator of immediate reading. The calculation of the index is performed by dividing the measurement of the pollutant, by its reference limit, established by the Italian Legislative Decree 155/2010. Moreover, we collected the last 5-years (2015–2019) time series of the concentration of 7 principal air pollutants (namely PM2.5, PM10, NO_2, SO_2, CO, Benzene and O_3) from the measurements of the Regional Environmental Protection Agency (ARPA) of each province and region considered in this study for a total of 7720 singular entries.

We indicated as "Overweight" the percentages of subjects by Italian region in excess weight (overweight + obese subjects) calculated from 2015 to 2018 provided by https://www.epicentro.iss.it/passi/dati/sovrappeso. By overweight subjects, we mean people with body mass index (BMI) between 25 and 29.9. While subjects with BMI $\geq$30 are considered obese.

We indicated as "Smokers" the percentage of smokers in each Italian region calculated from 2015 to 2018 extracted from https://www.epicentro.iss.it/passi/dati/fumo. Smokers are the subjects who have smoked 100 or more cigarettes in their life and who declared that they continue to smoke at the time of the survey.

We indicated as "Hospital beds" the available beds in public hospitals, per regional population, in 2017. Data were collected from (Italian Ministry of Health, 2019).

The most recent available data on the main sources of air pollution were collected from I.Stat, the warehouse of statistics currently produced by the Italian National Institute of Statistics (https://www.istat.it/en/). As variables, we selected:

- as "Farm", the average regional number of animals per breeding farm type from 2015 to 2019: caprine, bovine, equine, swine, ovine, and bufalin;
- as "Industry", the regional number of factories at the regional level selected as potential contributors to air pollution in 2017: extraction of minerals and from mines, water and garbage management, energy supply, transport and storage, and construction;
- as "Agriculture", the regional surface covered in hectares by all crops in 2019;
- as "Firewood", the regional consumption of biomass (in tonnes) per heating purposes (firewood and pellets) in 2017;
- as "Cars", the regional number of circulating vehicles in 2018.

Data on the total regional number of "Incinerators" in 2017 were collected from (ISPRA, Rapporto Rifiuti Urbani, 2017) and data on the regional number of "Airports" were collected from (ENAC, Aeroporti Certificati, 2019).

Data for "Traffic" were collected from (ACI, Annuario statistico,

2019) as the average regional quantity of fuel (gasoline and regular in tonnes) sold annually from 2005 to 2018.

All data were standardized to the total regional population size.

2.2. Data analysis

All the processing and statistical analyses were performed in R version 3.6.1 (R Core Team, 2020).

Following a procedure widely used in the literature, to carry out a statistical comparison between the different resident populations, neutralizing the effects deriving from their different age structures and population size, the data on positivity at the regional scale (for which age classes were available) were standardized with the direct method (Curtin and Klein, 1995), while the data on mortality at the regional scale (for which age classes were not available) were standardized with the indirect method (Wilcox and Russell, 1986). The standardization procedure leads to removing the effect of any age differences between two populations, keeping the real differences in disease frequency. Both direct and indirect standardizations involve the calculation of numbers of expected events (i.e. deaths and the number of positive people in our study), which are compared to the number of observed events. Standardized ratios were calculated as the ratio between the observed number of deaths/infected in the study population (regional or provincial) and the number of deaths/infected would be expected, based on the age- and sex-specific rates in the standard population (Italy) and the population size of the study population by the same age/sex groups.

In the direct method of standardization, age-adjusted rates are derived by applying the category-specific mortality rates of each population to a single standard population. This produced age-standardized positivity ratios (SPR) that Italian regions would have if they had the same age distribution as the standard population (Italy).

The indirect method is based on the ratio between the deaths observed in a territory and those expected in the same. The expected deaths were calculated by applying the corresponding specific mortality ratios of the population assumed as standard (the national one in this analysis) to the average annual population by age and sex classes of each territorial unit (regions, in this analysis). The Standardized Mortality Ratio (SMR), therefore, expresses the relationship between the deaths observed in a specific territory and the expected deaths if in the same territory there was the annual mortality, specific for age groups, of the population used as standard.

We calculated the mean value of the 5-year (2015–2019) Air Quality Index (AQI) at the provincial and regional levels by averaging the indexes recorded in each trimester of the 5 years. Moreover, we calculated the mean value of the 5-year time series of the 7 main pollutants at the provincial and regional scale by averaging the indexes recorded in each year (about 20 measurements, one per trimester, in each of the 99 provinces).

We used a Machine Learning algorithm based on Artificial Intelligence as a regressor to predict SMR and SPR and to identify which variables were important for the prediction model. Random Forest (RF) (Breiman, 2001) is a bagging technique that exploits the strength of tree ensemble classifiers. The main modification with standard tree classifiers is that each tree is grown using a substantially different set of features, as a consequence the algorithm prevents each tree from being too correlated to other trees. This ensures that the learning is not strongly dependent on any single feature. Our choice is motivated by several considerations: (i) it is robust as it does not require any particular tuning, in fact, it depends on just two different parameters, which are the number of trees *n* and the number of features *m* sampled to grow each leaf within a tree; (ii) it estimates variable importance and therefore provides a straightforward interpretation for the model; (iii) thanks to its out-of-bag estimation Random Forest generate an unbiased estimate of the generalization error. Furthermore, the randomization procedure at the base of random forest significantly reduces the problem of overfitting (Breiman, 1996).

For our analysis, we adopted a default configuration for Random Forest, with $n = 500$ trees and $m = f/3$ with *f* being the number of features. As previously mentioned, one of the main advantages of Random Forest is the possibility to internally assess the importance of each feature for the model accuracy. We evaluated the feature importance using the mean decrease impurity. During the training phase, it is possible to estimate how much each selected feature decreases the impurity of a tree. In Random Forest, the impurity decrease from each feature is averaged on all trees to estimate the importance ranking of variables. The impurity is measured by the residual sum of squares.

Considering the *N* trees of our model the prediction for each element x_i of sample X is given by the average of predictions of all *N* trees:

$$\widehat{y}(x_i) = \frac{1}{N}\sum_{j=1}^{N}\Theta_j(x_i) \tag{1}$$

where Θ_j indicates a single tree grown by randomly selecting m variables.

The mean-squared generalization error for each predictor tree is:

$$E(Y - \Theta(X))^2 \tag{2}$$

with Y that represents the expected numerical outcome. For accurate Random Forest regression is required a low correlation between residuals of differing regressor trees and a minimization of the prediction error function for the individual trees, defined in (2) (Segal, 2004).

In the present work, we assessed the problem of predicting the rates of mortality and positivity in 20 Italian regions based on heterogeneous characterizations. In particular, we built three Random Forest regression models for each indicator (SMR and SPR): the first in which we used the 6 environmental, health, and socio-economic variables as input features, the second was implemented with information on sources of air pollution, finally, in the third application we modelled our regressor through concentrations of 7 main pollutants. We verified the goodness of our models employing the *root-mean-square error* (*RMSE*) and the coefficient of determination (R^2). Random Forest was also employed to project the increase in the SARS-CoV-like estimated mortality (SMR) and positivity (SPR) with a 5–10% higher AQI (i.e. air pollution worsening).

A linear regression analysis was performed and a Pearson's correlation coefficient was calculated to study the relationship between the total number of recorded cases at the provincial level (adjusted by the provincial population size) and the provincial AQI. The trend line's equation was then employed to project the increase in the SARS-CoV-like estimated cases with a 5–10% higher AQI (i.e. worsening of air pollution). We also compared the positivity rate of each province with the value estimated with a simple linear model where this rate was expressed as a function of the AQI. The difference between the actual and the expected value is the plotted residuals.

As a secondary analysis to assess the robustness of our results gathered with Random Forest, we applied the Canonical Correlation Analysis (CCA) (Hotelling, 1992) to research the best correlation between two following sets of data: the first composed by the 6 environmental, health, and socio-economic variables (Overweight, Smokers, Hospital beds, Income, AQI and Swabs) and the second with the SPR and SMR. Our purpose was to evaluate which of the 6 variables provided the greatest contribution (weight) in the correlation. Given two independent datasets X and Y, the CCA identifies the pairs of linear combinations, one of each dataset, most closely related to each other. In other words, CCA detects the best correlation that can be obtained between two independent datasets.

Assuming X and Y, two independent datasets, composed by *N* cases with *n* and *m* features respectively, the CCA technique allows to find, for each case, *k* pairs of canonical variables, being *k* the minimum between *n* and *m*. Each pair has as a first element a linear combination of the features of the dataset X and as a second element a combination of the features of the dataset Y. The *k* pairs are ordered in such a way that the first pair has a greater correlation than the second, the second has a greater correlation than the third, and so on. In addition to the scores, the CCA algorithm returns the canonical factors (the coefficients to calculate each canonical variable) and the canonical correlations, the *k* correlations relating to each pair of canonical variables.

Since one of the two dataset in our study contains only two variables (SPR and SMR), CCA returns only two pairs of canonical variable; with the first to be preferred correlation as it yields a stronger correlation between the linear combination of SPR and SMR and the linear combination of the six environmental variables. In addition, to avoid the problems of overfitting related to the small number of available cases in the model (20 Italian regions), we only considered all possible pairs by combining the six variables two by two. We applied the CCA to each of these pairs to estimate the best correlation existing with SMR and SPR dataset. This analysis confirms what has already been achieved by applying the CCA to the entire dataset. In fact Fig. 2 shows that only the couples with AQI have a correlation coefficient above 70% respect to SPR and SMR combination.

3. Results

3.1. Relations between SARS-CoV-2 and health, ecological and socio-economic factors

First, we analysed the relation between the last 5-year exposure to air pollution (hereafter Air Quality Index, AQI; see Appendix A for Supplementary Methods), the number of smokers, the level of obesity and overweight, the mean annual income per family, the number of public hospital beds, and the number of SARS-CoV-2 tests performed with the rates of mortality (SMR; see Methods) and positivity (SPR) in 20 Italian regions. SMR and SPR are standardized rates that represent the percentage of the increase or decrease in mortality of a study cohort (regional populations, in this study) compared to the general population (Italy, in this study). We adjusted all rates to the national norm to remove the differences in size and age distribution of the regional populations, which may represent confounding factors (Naing, 2000). Then, we modelled how SMR and SPR vary with these factors with a popular supervised learning algorithm from Artificial Intelligence (A.I.), which is Random Forests (see Appendix A for Supplementary Methods). Our model can explain the SMR ($R^2 = 0.95$ and RMSE = 28.9; Fig. 1C) and SPR ($R^2 = 0.93$ and RMSE = 20.3; Fig. S2) values with high accuracy. AQI ranked first in importance among six different factors and is the only one showing a significant and positive correlation with SMR ($R^2 = 0.54$; Fig. 1D). This Machine Learning evidence was also confirmed by a Canonical Correlation Analysis (CCA), which showed that SMR and SPR have always higher correlations with all the pairwise combinations of the six variables that include AQI (Fig. 2).

3.2. Importance of the different sources and components of air pollution

Because the prolonged exposure to air pollution resulted as the most important factor to explain the differential mortality and positivity to the SARS-Cov-2 in Italian regions, we investigated deeper to determine the relative contribution of main sources of air pollution. We evaluated the importance for SMR and SPR of 8 principal sectors related to local air pollution such as industry, farms, agriculture, vehicular traffic, cars, household firewood, airports, and incinerators, which are confirmed sources of air pollution and emit Particulate Matter (PM), Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2), Carbon Monoxide (CO), Ozone (O_3) precursors, Benzene, etc. (Holman, 1999). The Random Forest analysis showed that Industry (28%), followed by Farm (22%) and Road traffic (19%) were the most relevant air pollution sectors linked to an increase in mortality rates of the 20 Italian regions (Fig. 3A). Moreover, road traffic resulted in the most important variable related to SARS-CoV-2 positivity (Fig. 3C).

Our machine learning analysis also revealed that the exposure to high levels of PM2.5 (particulate matter that has a diameter smaller than 2.5 μm) is the most important contaminant to explain the high mortality and positivity due to the new coronavirus propagation in the most polluted regions of Italy (Fig. 3B and D) among other AQI components (namely PM10, O_3, SO_2, $NO_{2,}$ CO, and Benzene), which also contributed to higher mortality levels in some regions such as Aosta Valley and Trentino Alto-Adige (Fig. 4).

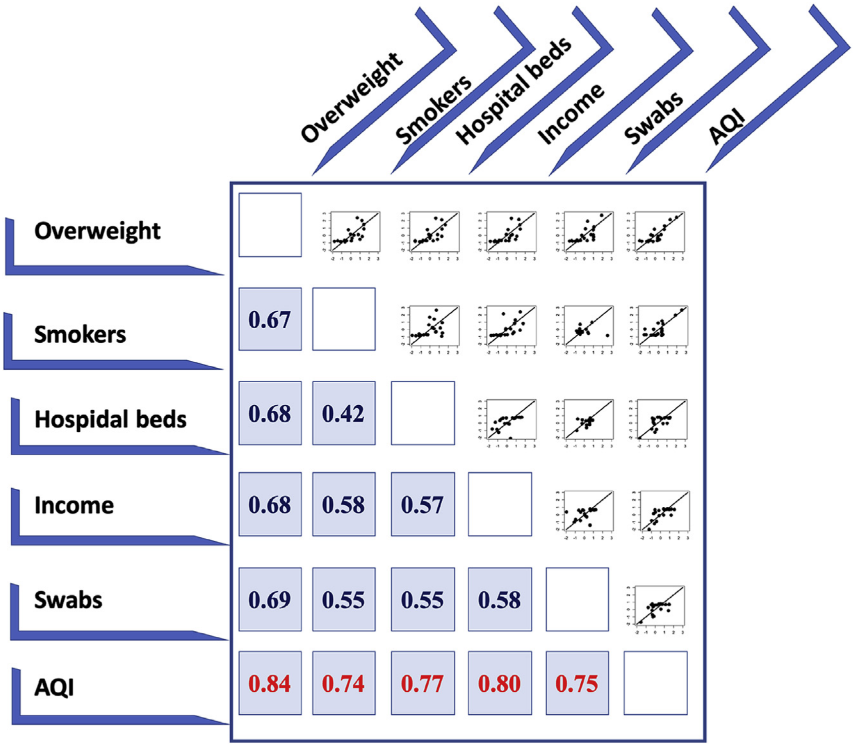
3.3. Effects of air pollution at a local scale and on future epidemics

Because the impact of the AQI on SARS-CoV-2 infection showed a strong geographical pattern at a regional level, we zoomed our analysis in on a smaller provincial scale to better understand the effect of the prolonged exposure to air pollution (Fig. S3). From a regression between the total recorded cases in 99 provinces (adjusted by their population) and the AQI in each of them (Fig. 5A), we confirmed a positive linear trend (slope = 0.52, r = 0.44). Interesting additional elements, however, came out. Six provinces showed to be evident outliers (Fig. 5B): 5 of them (Cremona, Lodi, Piacenza, Bergamo, and Brescia) with an excess of cases than those predicted by the AQI in our model and one province (Siracusa) with a lack of cases than those expected by its highest level of exposure to air pollution.

Given the major contribution played by air pollution (much more important than other health and socio-economic factors, as we discovered from the regional analysis), we estimated the expected effects of a decrease in air quality on new potential SARS-CoV-like epidemics. Alarmingly, we estimated that with an overall increase between 5% and 10% of air pollution at the national scale, there would be 21–32% additional cases, whose 19–28% more positives and 4–14% more deaths, to sum to the future epidemic tolls of Italy.

4. Discussion

The results of our artificial intelligence model showed the existence of a strong association between prolonged exposure to air pollution and SARS-Cov-2 mortality and positivity. In fact, as a measure of this association, we considered two different metrics such as R^2 and RMSE, and a corollary analysis involving partial correlations. According to both metrics, it is possible to model mortality and positivity with great accuracy using pollution data.



	Overweight	Smokers	Hospital beds	Income	Swabs	AQI
Overweight						
Smokers	0.67					
Hospidal beds	0.68	0.42				
Income	0.68	0.58	0.57			
Swabs	0.69	0.55	0.55	0.58		
AQI	0.84	0.74	0.77	0.80	0.75	

Fig. 2. Results of the CCA applied among all pairs of the 6 considered variables and SMR and SPR indexes. For each pair of variables, the below-diagonal cells show the correlation value for the first pair of canonical components, and the relative scatterplot is symmetrically plotted in the upper part. The highest correlation coefficients in red, Pearson's r > 0.7, are obtained for all pairs that include AQI. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

As far as we know, this is the first study to investigate and model this association with a multivariate model, while so far simple single variable approaches have been preferred (Wu et al., 2020; Fattorini and Regoli, 2020). Considering the feature importance, we were able to assess the contribution of each considered factor and disentangle the role of those factors showing minor importance. Surprisingly, none of the health variables such as the number of swabs, smokers, and overweight people show any significant ($P > 0.05$, Bonferroni correction) positive trend with mortality and positivity rates. The same consideration holds for socio-economic factors such as family income and healthcare. We think this finding is particularly intriguing because it may also highlight the fact that, in Italy, access to the health-care system does not depend on the socio-economic conditions of patients. Moreover, it can be considered an indirect validation of our hypotheses about the primary role played by pollution factors because this shows that air quality affects all citizens despite their wage or lifestyle.

We then studied the relative contribution of the main sources of air pollution. We showed that industry and farms were the most important, among 8 principal sectors, sources of air pollution for SARS-CoV-2 mortality and that vehicular traffic had a major impact on positivity. Industry, particularly from metallurgic, gas and oil, plastic, construction, and chemical factories may severely pollute the air of the surrounding areas emitting, through smokestacks, substances such as Particulate Matter (PM), Nitrogen dioxide (NO_2), Carbon Monoxide (CO), Ozone (O_3) precursors, etc. (Wang et al., 2016; Zhang et al., 2020). Farms, particularly intensive indoor ones, are considered important emitters of PM precursors created by secondary reactions in the atmosphere with sulphur dioxide (SO_2), nitrogen oxides (NO_x), volatile organic compounds (VOCs) and ammonia (NH_3) (Zhao et al., 2017). Intensive and extensive agriculture and farms, in fact, generate fumes from nitrogen-rich fertilizers and animal waste, which mainly contain ammonia and combine in the air with combustion emissions to form solid small particles. Road transport is a major source of air pollution and widely confirmed to harm human health and the environment (Samet, 2007). Vehicles emit a range of pollutants including NOx, particulate matter, and O_3 precursors. The use of biomass such as firewood and pellets for domestic heating is recognized as a heavy threat for air quality because its burning emits black smoke from smokestacks full of PM, sulphur oxides (SO_x), NO_2, dioxins, and furans (Boman et al., 2003). The incineration of waste, besides CO, SO_x, NO_x, etc., may release in the atmosphere heavy metals, Polyvinyl Chlorides (PVC), dioxins, and furans (Ranzi et al., 2011). Finally, airports are among the largest sources of air pollution and contribute to increasing the level of CO, O_3, $NO_{x,}$ and PM in the air surrounding runways (Schlenker and Walker, 2016).

Considering that more than three-quarter of the deaths associated with SARS-CoV-2 in the first six months of 2020 in Italy was reported in northern regions, our result that the number of factories plays a major role in the SARS-CoV-2 deaths connected to air pollution is not surprising. North-western Italy, which comprises the first Italian industrial triangle (also called To-Mi-Ge) corresponding to the summits of Turin, Milan, and Genoa, is the area in which large-scale industrialization of the Italian economy took place between the end of the 19th century and the beginning of the 20th century (Ciccarelli and Fenoaltea, 2013). Moreover, a report published in 2019 by the European Environment Agency (European Environment Agency, Air quality in Europe, 2018) showed that the Po Valley, i.e. the area between the Alpine chain, the Northern Apennines and the Adriatic Sea, is the most affected region by the concentration of air pollutants of the whole Europe.

Among the 7 main components of air pollution, we discovered that fine particles (PM2.5) played a major role in this pandemic. This small particulate is primarily produced by the combustion of fuel in engines in vehicles, the rubbing of brakes and tires, domestic

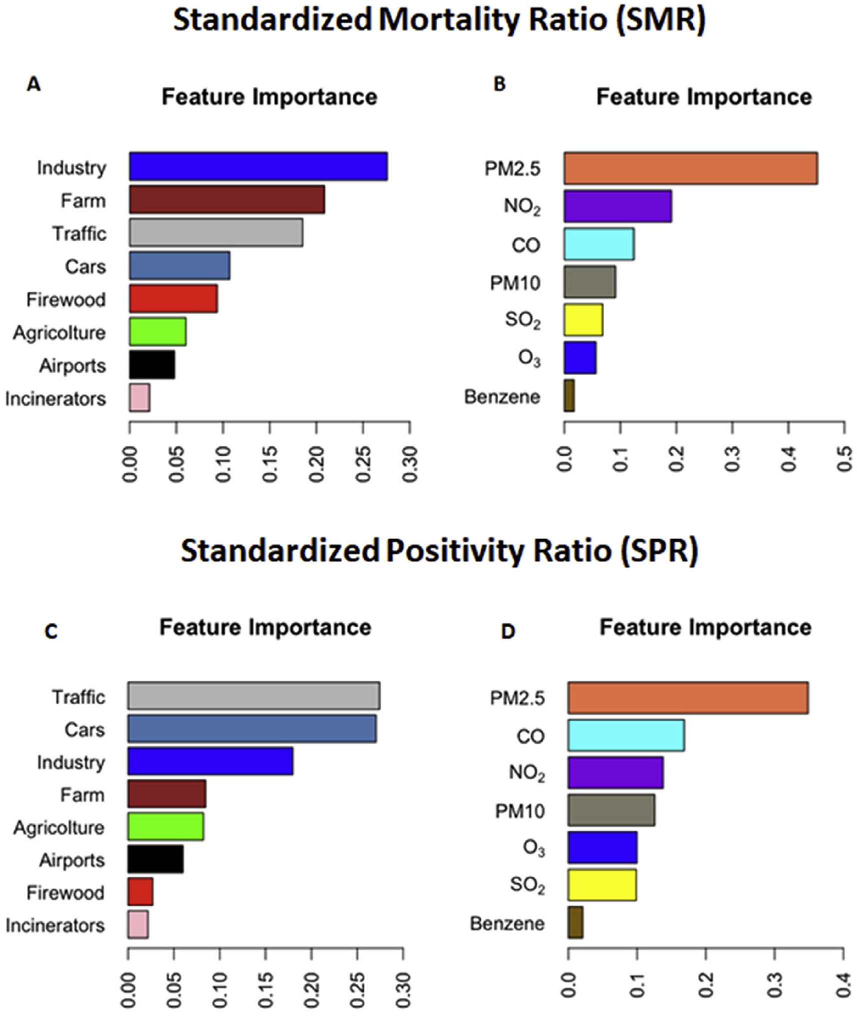


Fig. 3. Feature importance of air pollution components and its main sources for SARS-CoV-2 mortality and positivity. The feature importance produced by Random Forest model for the relation of SMR with (A) major sources, which shows that Industry, Farm and Road traffic are - respectively - the most relevant sources of air pollution connected to an increase in the mortality rates; and with (B) contaminants, which shows that PM2.5 is the most relevant component of air pollution connected to an increase in the mortality rate. The feature importance produced by Random Forest model for the relation of SPR, at the regional scale, with (C) major sources, which shows that Road traffic and Cars, Industry, and Farm are the most relevant sources of air pollution and with (D) contaminants, which shows that PM2.5, followed by CO and NO_2 are the most important components of air pollution, connected to an increase in the positivity rates.

heating, and industrial emissions, including power plants. Nonetheless, the fact that we found that farms ranked as the second main source of air pollution-related to increased SARS-CoV-2 mortality is strong evidence that, together with industry, intensive farming can worsen the pandemic death toll. Major sources of NH_3 include agricultural activities (i.e. animal husbandry and fertilizers) and the regional inventory estimates of Lombardy (ARPA Lombardy, 2017) and Emilia-Romagna (ARPA Emilia Romagna, 2018) attribute ~95% of ammonia emissions to agricultural activities out of the annual total. Ammonia reacts with nitric and sulfuric acids leading to the formation of ammonium nitrate and ammonium sulphate, respectively, the two inorganic salts most present in the small particles. Therefore, intensive farming heavily contributes to the formation of secondary particulate matter.

We now have stronger evidence that high exposure to small particles is related to increased SARS-CoV-2 mortality. However, other air contaminants must not be underestimated. Our finding that in some regions with exceeding SARS-CoV-2 mortality there was prolonged exposure to high levels of other AQI components supports the idea of a synergistic effect of contaminants. Air pollution is composed of several pollutants and their complex interactions may exert even stronger pressure on human health and have bigger impacts on the environment than expected by considering the effects of contaminants in isolation. The fact that road traffic ranks only third in terms of importance as a source of air pollution shows that it is a contributing, but not the most relevant, compartment affecting the mortality to viral-induced respiratory infections, although it emerged as the most important variable related to SARS-CoV-2 positivity, which could mean that this source produces other air pollutants that, affecting the respiratory system, increased the susceptibility to the new coronavirus.

We recognize that at a smaller scale, besides air quality, additional and strictly local factors may have played a role in increasing the number of total SARS-CoV-2 cases in some provinces of Lombardy, which would explain the outliers we found in our regression at the provincial level. Although correlation does not imply causation, our findings would recommend significant interventions to reduce the air pollution deriving from industrial and farming production activities, home heating and transports to hopefully weaken the impact of future epidemics.

In contrast, according to our model, a province in Sicily could become a new outbreak location if measures to limit this new

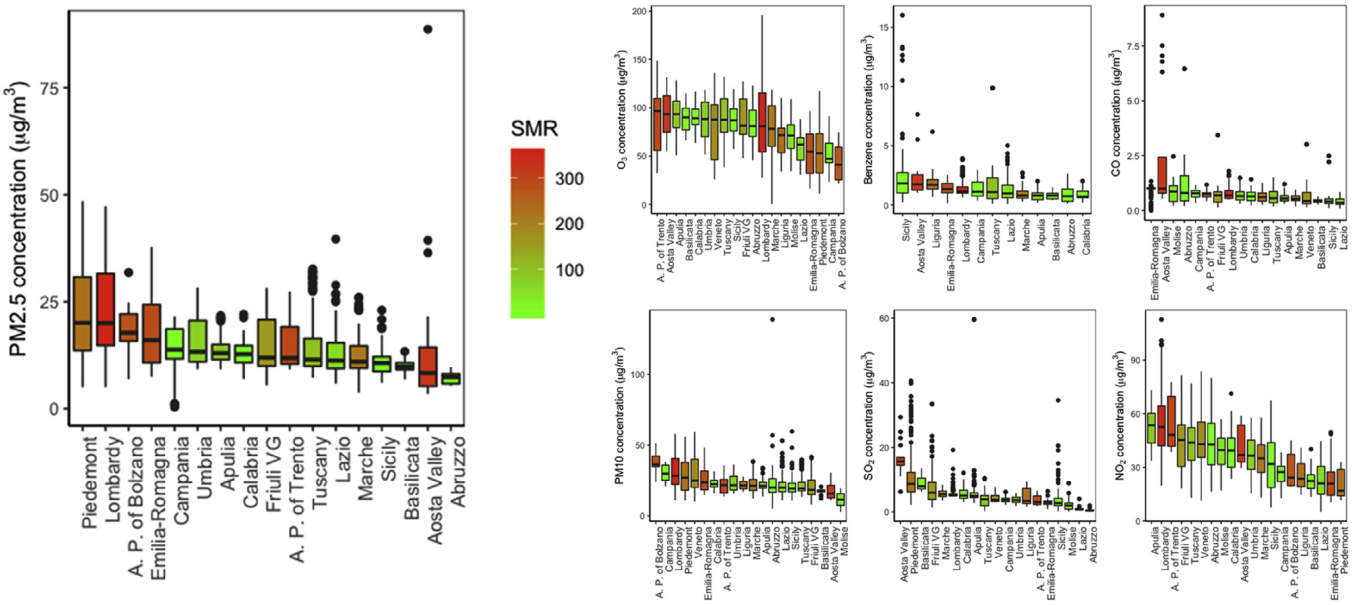


Fig. 4. The influence of prolonged exposure to all AQI components to SARS-CoV-2 mortality in Italian regions. Boxplots of seven air contaminants (namely PM2.5, PM10, O_3, Benzene, SO_2, NO_2, and CO) measured in the last-5 years (2015–2019) by the Regional Agencies for Environmental Protection ARPA that contribute to the evaluation of air quality and show their relation with SARS-CoV-2 mortality (the SMR depicted in a coloured scale of intensity in each box). Boxes are ranked from the highest to the lowest median value (horizontal line in the box) per each contaminant (in μg/m^3) and represent the first and fourth quartiles. Whiskers represent the maximum and minimum values excluding outliers (dots). The SMR colour scale is the same as in Fig. 1A. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

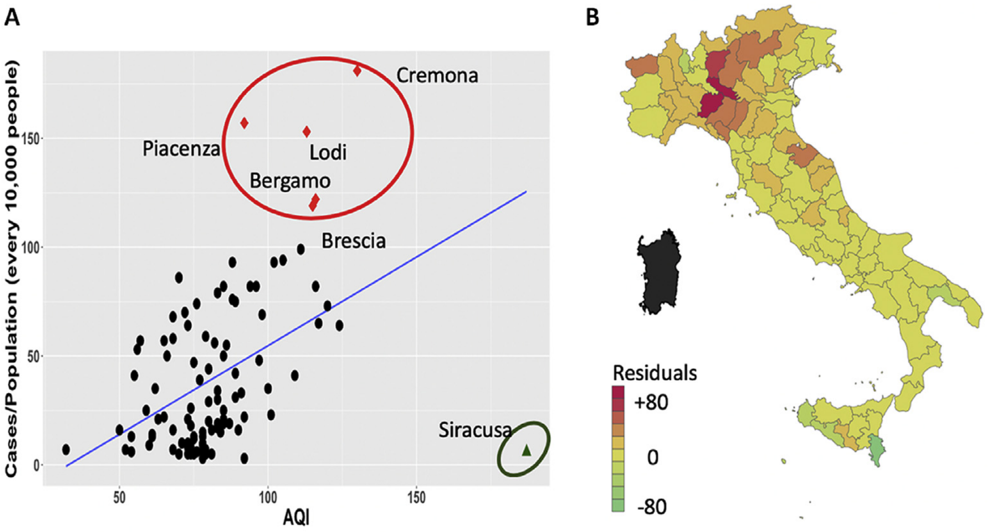


Fig. 5. The relation between SARS-CoV-2 total cases and Air Quality Index (AQI) at the provincial scale. (A) Most of the Italian provinces show a positive linear correlation (slope = 0.52, r = 0.44) between the Air Quality Index (AQI) and the total number of cases (adjusted for the provincial population size). However, some provinces are evident outliers. Those in the red set (above the trend line) show an excess of cases compared to what predicted by the linear correlation with the AQI. The only outlier in the green set (below the trend line) shows a lack of cases compared to what predicted by the linear correlation with the AQI. (B) This is also shown in the map of the analysis of residuals (which are the difference between the observed and predicted cases in each province). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

coronavirus and pathogens with similar respiratory effects are not well managed.

As far as we know, this is the first study addressing this problem within a quantitative framework which allows the evaluation of the statistical association between pollution and SARS-CoV-2 effects. Nonetheless, we identified some limitations of our study to improve future research on the same topic. First of all we recognize that at a smaller scale, besides air quality, additional and strictly local factors may have played a role in increasing the number of total SARS-CoV-2 cases in some provinces of Lombardy and North Italy; accordingly, the correlation between pollution and epidemic effects could reasonably include the contribution of other co-factors. This study is far from being an exhaustive and comprehensive examination of all possible factors which may contribute to aggravated viral infections in the human lower respiratory tract; further studies could address this aspect. Lockdown measures and travel restrictions preventing the virus diffusion could have played a relevant role and partially account for regional differences in epidemic mortality and severity. Another important aspect to consider is time exposure, in this study we consider a pollutant exposure up to 5 years, but it could be interesting to investigate if and to which extent longer time exposure has a relevant effect on epidemic severity.

Nonetheless, the use of artificial intelligence to understand the importance of multiple factors represents an advancement of classical statistical analysis, which through a machine learning process sensitively increased the accuracy of our model.

5. Conclusions

Our results showed a strong correlation between SARS-CoV-2 mortality and positivity and the prolonged exposure to air pollution. We designed and evaluated a multivariate model to investigate how different factors would affect pandemics diffusion and severity. In particular, we observed that in our model air quality plays the most relevant role. Interestingly, other factors inherent to socio-economic conditions and lifestyles showed much less importance. The model is accurate and paves the way for the prediction of future outcomes: a further worsening of air quality might lead to even more dramatic consequences in future pandemics. We also showed stronger evidence that a prolonged exposure particularly to small particles is strongly related to an enhanced SARS-CoV-2 mortality. Together with this evidence, it might also be that the pollutants stagnation, resulting from a combination of specific climatic conditions, local anthropogenic emissions and regional topography, may promote a longer permanence of the of viral particles in the air, thus favouring an indirect diffusion in addition to the direct one from individual to individual. Further investigations can shed more light on these dynamics and the role of small particles in the diffusion of pathogens. However, we already have quite clear proof that prolonged exposure to air pollution in Italy, mainly from highly industrialized and intensively farmed areas and congested roads that produce fine particulate matter, enhances the risks associated with epidemics. This must be taken into account in country-based environmental policies on a global scale because it is highly probable that the relation between air pollution and SARS-CoV-like mortality we discovered in Italy is a more common pattern than thought before and may worsen the effects of pathogens' spread in other parts of the world.

Overall, our finding that air quality is the most important factor connected with mortality and positivity of SARS-CoV-like pathogens, which would worsen even with a slight increase of pollutants, makes clear that the imperative of productivity before health and environmental protection is, indeed, a short-term/small-minded resolution.

Funding

The authors received no specific funding for this work.

Data and materials availability

All processed data is available in the main text or the supplementary materials. Original data are freely available from the references mentioned in the main text or the supplementary materials. Code, and materials used in the analysis will be made available to any researcher for purposes of reproducing or extending the analysis.

CRediT authorship contribution statement

Roberto Cazzolla Gatti: Conceptualization, Resources, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Alena Velichevskaya:** Resources, Validation, Writing - review & editing. **Andrea Tateo:** Formal analysis, Investigation, Writing - review & editing. **Nicola Amoroso:** Conceptualization, Resources, Validation, Software, Formal analysis, Investigation, Writing - review & editing. **Alfonso Monaco:** Conceptualization, Resources, Validation, Formal analysis, Investigation, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Authors would like to thank IT resources made available by ReCaS, a project funded by the MIUR (Italian Ministry for Education, Universityand Re-search) in the "PON Ricerca e Competitività 2007–2013-AzioneI-Interventi di rafforzamento strutturale" PONa3_00052, Avviso254/Ric, University of Bari.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envpol.2020.115471.

References

ACI, Annuario statistico, 2019. http://www.aci.it/laci/studi-e-ricerche/dati-e-statistiche/annuario-statistico/annuario-statistico-2019.html.

ARPA Emilia Romagna, 2018. https://www.arpae.it/dettaglio_documento.asp?id=7361&amp;idlivello=1693. (Accessed 6 June 2020).

ARPA Lombardy, 2017. http://www.inemar.eu/xwiki/bin/view/Inemar/HomeLombardia. (Accessed 6 June 2020).

Bherwani, H., Nair, M., Musugu, K., Gautam, S., Gupta, A., Kapley, A., Kumar, R., 2020. Valuation of air pollution externalities: comparative assessment of economic damage and emission reduction under COVID-19 lockdown. Air Qual. Atmos. Health 13, 683–694.

Bigi, A., Ghermandi, G., Harrison, R.M., 2012. Analysis of the air pollution climate at a background site in the Po valley. J. Environ. Monit. 14 (2), 552–563. https://doi.org/10.1039/C1EM10728C.

Boman, B.C., Forsberg, A.B., Järvholm, B.G., 2003. Adverse health effects from ambient air pollution in relation to residential wood combustion in modern society. Scand. J. Work. Environ. Health 29, 251–260.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Brugha, R., Grigg, J., 2014. Urban air pollution and respiratory infections. Paediatr. Respir. Rev. 15 (2), 194–199. https://doi.org/10.1016/j.prrv.2014.03.001.

Carugno, M., Dentali, F., Mathieu, G., Fontanella, A., Mariani, J., Bordini, L., et al., 2018. PM10 exposure is associated with increased hospitalizations for respiratory syncytial virus bronchiolitis among infants in Lombardy, Italy. Environ. Res. 166, 452–457.

Cazzolla Gatti, R., 2020. Coronavirus outbreak is a symptom of Gaia's sickness. Ecol. Model. 426, 109075.

Chauhan, A.J., Johnston, S.L., 2003. Air pollution and infection in respiratory illness. Br. Med. Bull. 68 (1), 95–112. https://doi.org/10.1093/bmb/ldg022.

Chinazzi, M., et al., 2020. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science 368 (6489), 395–400. https://doi.org/10.1126/science.aba9757.

Ciccarelli, C., Fenoaltea, S., 2013. Through the magnifying glass: provincial aspects of industrial growth in post-Unification Italy 1. Econ. Hist. Rev. 66 (1), 57–85. https://doi.org/10.1111/j.1468-0289.2011.00643.x.

Ciencewicki, J., Jaspers, I., 2007. Air pollution and respiratory viral infection. Inhal. Toxicol. 19 (14), 1135–1146.

Conticini, E., Frediani, B., Caro, D., 2020. Can atmospheric pollution be considered a co-factor in extremely high level of SARS-CoV-2 lethality in Northern Italy? Environ. Pollut. 114465 https://doi.org/10.1016/j.envpol.2020.114465.

Curtin, L.R., Klein, R.J., 1995. Direct Standardization (Age-adjusted Death Rates), vol. 6. US Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD.

Dong, E., Du, H., Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. 20 (5), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1.

Dutheil, F., Baker, J.S., Navel, V., 2020. COVID-19 as a factor influencing air pollution?

Environ. Pollut. 263, 114466.

ENAC, Aeroporti Certificati, 2019. https://www.enac.gov.it/aeroporti/infrastrutture-aeroportuali/aeroporti-in-italia/certificazione-degli-aeroporti/aeroporti.

European Environment Agency, Air quality in Europe, 2018. https://www.eea.europa.eu/publications/air-quality-in-europe-2018.

Fattorini, D., Regoli, F., 2020. Role of the chronic air pollution levels in the Covid-19 outbreak risk in Italy. Environ. Pollut. 114732.

Gautam, S., 2020. COVID – 19: air pollution remains low as people stay at home. Air Qual. Atmos. Health. https://doi.org/10.1007/s11869-020-00842-6.

Guan, W.J., Zheng, X.Y., Chung, K.F., Zhong, N.S., 2016. Impact of air pollution on the burden of chronic respiratory diseases in China: time for urgent action. Lancet 388 (10054), 1939–1951. https://doi.org/10.1016/S0140-6736(16)31597-5.

Holman, C., 1999. Sources of air pollution. In: Air Pollution and Health. Academic Press, Cambridge.

Hotelling, H., 1992. “Relations between Two Sets of Variates” in Breakthroughs in Statistics. Springer, New York.

Huang, C., et al., 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395 (10223), 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5.

ISPRA, 2017. Rapporto Rifiuti Urbani - Edizione 2017. http://www.isprambiente.gov.it/it/archivio/eventi/2017/ottobre/rapporto-rifiuti-urbani-edizione-2017.

ISTAT, 2020. https://www.istat.it/it/archivio/240401. (Accessed 6 June 2020).

Italian Ministry of Health, 2019. Annuario Statistico del Servizio Sanitario Nazionale 2017. http://www.salute.gov.it/imgs/C_17_pubblicazioni_2879_allegato.pdf.

Liang, Y., et al., 2014. PM2.5 in Beijing temporal pattern and its association with influenza. Environ. Health 13 (1), 102. https://doi.org/10.1186/1476-069X-13-102.

Livingston, E., Bucher, K., 2020. Coronavirus disease 2019 (COVID-19) in Italy. Jama 323 (14). https://doi.org/10.1001/jama.2020.4344, 1335-1335.

Muhammad, S., Long, X., Salman, M., 2020. Covid – 19 pandemic and environmental pollution: a blessing in disguise? Sci. Total Environ. 728, 138820.

Naing, N.N., 2000. Easy way to learn standardization: direct and indirect methods. MJMS 7 (1), 10–15.

Nenna, R., et al., 2017. Respiratory syncytial virus bronchiolitis, weather conditions and air pollution in an Italian urban area: an observational study. Environ. Res. 158, 188–193. https://doi.org/10.1016/j.envres.2017.06.014.

Ogen, Y., 2020. Assessing Nitrogen Dioxide (NO2) Levels as a Contributing Factor to the Coronavirus (COVID-19) Fatality Rate. Science of The Total Environment, 138605. https://doi.org/10.1016/j.scitotenv.2020.138605.

Onder, G., Rezza, G., Brusaferro, S., 2020. Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. Jama 4683. https://doi.org/10.1001/jama.2020 ([published in Jama First Release; not yet published in print]).

R Core Team, R., 2020. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/.

Ranzi, A., et al., 2011. Mortality and morbidity among people living close to incinerators: a cohort study based on dispersion modeling for exposure assessment. J. Environ. Health 10 (1), 22. https://doi.org/10.1186/1476-069X-10-22.

Samet, J.M., 2007. Traffic, air pollution, and health. Inhal. Toxicol. 19 (12), 1021–1027. https://doi.org/10.1080/08958370701533541.

Schlenker, W., Walker, W.R., 2016. Airports, air pollution, and contemporaneous health. Rev. Econ. Stud. 83 (2), 768–809. https://doi.org/10.1093/restud/rdv043.

Sciomer, S., Moscucci, F., Magrì, D., Badagliacca, R., Piccirillo, G., Agostoni, P., 2020. SARS-CoV-2 spread in Northern Italy: what about the pollution role? Environ. Monit. Assess. 192, 325.

Segal, M.R., 2004. Machine Learning Benchmarks and Random Forest Regression. UCSF: Center for Bioinformatics and Molecular Biostatistics.

Ségala, C., Poizeau, D., Mesbah, M., Willems, S., Maidenberg, M., 2008. Winter air pollution and infant bronchiolitis in Paris. Environ. Res. 106 (1), 96–100. https://doi.org/10.1016/j.envres.2007.05.003.

Sharma, S., Zhang, M., Gao, J., Zhang, H., Kota, S.H., 2020. Effect of restricted emissions during COVID-19 on air quality in India. Sci. Total Environ. 728, 138878.

Wang, K., et al., 2016. A comprehensive emission inventory of multiple air pollutants from iron and steel industry in China: temporal trends and spatial variation characteristics. Sci. Total Environ. 559, 7–14. https://doi.org/10.1016/j.scitotenv.2016.03.125.

Wang, C., Horby, P.W., Hayden, F.G., Gao, G.F., 2020. A novel coronavirus outbreak of global health concern. Lancet 395 (10223), 470–473.

Wilcox, A.J., Russell, I.T., 1986. Birthweight and perinatal mortality: III. Towards a new method of analysis. Int. J. Epidemiol. 15 (2), 188–196. https://doi.org/10.1093/ije/15.2.188.

Wong, C.M., et al., 2010. Part 4. Interaction between air pollution and respiratory viruses: time-series study of daily mortality and hospital admissions in Hong Kong. Res. Rep. 154, 283–362.

Wu, X., Nethery, R.C., Sabath, B.M., Braun, D., Dominici, F., 2020. Exposure to Air Pollution and COVID-19 Mortality in the United States. medRxiv. https://doi.org/10.1101/2020.04.05.20054502.

Xing, Y.F., Xu, Y.H., Shi, M.H.Y., Lian, X., 2016. The impact of PM2.5 on the human respiratory system. J. Thorac. Dis. 8 (1), E69–E74. https://doi.org/10.3978/j.issn.2072-1439.2016.01.19.

Yao, L., Zhan, B., Xian, A., Sun, W., Li, Q., Chen, J., 2019. Contribution of transregional transport to particle pollution and health effects in Shanghai during 2013–2017. Sci. Total Environ. 677, 564–570. https://doi.org/10.1016/j.scitotenv.2019.03.488.

Zhang, F., et al., 2020. Significant impact of coal combustion on VOCs emissions in winter in a North China rural site. Sci. Total Environ. 720, 137617. https://doi.org/10.1016/j.scitotenv.2020.137617.

Zhao, Z.Q., et al., 2017. Mitigating ammonia emission from agriculture reduces PM2.5 pollution in the Hai River Basin in China. Sci. Total Environ. 609, 1152–1160. https://doi.org/10.1016/j.scitotenv.2017.07.240.