Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 207 (2022) 573–582

Procedia Computer Science

www.elsevier.com/locate/procedia

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)

The Influence of Environmental Factors on the Spread of COVID-19 in Italy

Andrea Loreggia[a,*], Anna Passarelli[b], Maria Silvia Pini[b]

[a] *University of Brescia - Department of Information Engineering, Via Branze 38, 25121, Brescia, Italy*
[b] *University of Padova - Department of Information Engineering, Via Gradenigo 6/b, 35131, Padova, Italy*

Abstract

The aim of this work is to investigate possible relationships between air quality and the spread of the pandemic. We evaluate the performance of machine learning techniques in predicting new cases. Specifically, we describe a cross-correlation analysis on daily COVID-19 cases and environmental factors, such as temperature, relative humidity, and atmospheric pollutants. Our analysis confirms a significant association of some environmental parameters with the spread of the virus. This suggests that machine learning models trained using environmental parameters might provide accurate predictions about the number of infected cases. Our empirical evaluation shows that temperature and ozone are negatively correlated with confirmed cases (therefore, the higher the values of these parameters, the lower the number of infected cases), whereas atmospheric particulate matter and nitrogen dioxide are positively correlated. We developed and compared three different predictive models to test whether these technologies can be useful to estimate the evolution of the pandemic.

Keywords: Air Quality Effects; COVID-19 Pandemic; Machine Learning; Correlation Analysis

1. Introduction

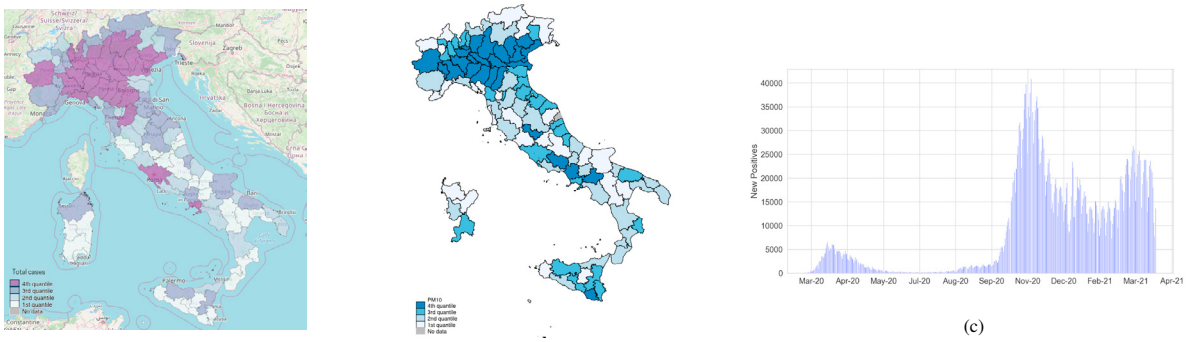
The new coronavirus SARS-CoV-2 is responsible for the respiratory disease named COVID-19. It was first identified on the 9^{th} of January 2020 by the Municipal Health Commission of Wuhan (China) which reported to the World Health Organization (WHO) a cluster of pneumonia cases of unknown origin in the city of Wuhan, in the Chinese province of Hubei. The spread of COVID-19 was then declared a global pandemic by WHO on the 11^{th} of March 2020 [**?**]. On the 31^{st} of March 2022, the World Health Organization reported the number of confirmed global cases

* Corresponding author:
E-mail address: andrea.loreggia@unibs.it

1877-0509
10.1016/j.procs.2022.09.112

of COVID-19 exceeds 480 million, with almost 6.1 million deaths. At that date in Italy, almost 15 million positive cases and more than 131 thousand deaths have been recorded [1].

The scale of the public health emergency caused by COVID-19 has no precedent in recent decades and it will surely have serious social and economic consequences. Indeed, the rapid spread of this global pandemic has immediately raised urgent issues, which need a coordinated study to slow down the evolution of the disease. In this context, Artificial Intelligence (AI) techniques can represent great support for government institutions and health organizations in order to provide information on the mechanisms which describe how the virus spread and, possibly, on the methodologies to be adopted to contrast it in the most effective way. If properly implemented and used, machine learning algorithms can help in analyzing data relating to some areas affected by the infection. For example, in [24] machine learning techniques are adopted for detecting infected patients with the minimum time penalty, while in Chen et al. [4] AI is employed for chest CT images for COVID-19 identification. By analyzing available historical data, machine learning models can be trained to predict possible developments of the pandemic as well as the impacts on the population. Understanding a complex system such as the spread of the pandemic is an important challenge in a country's sustainable development process. The concept of sustainable development has spread widely in recent decades and generally consists of a combination of three goals: the social goal, the economic goal, and the environmental goal. Policy makers play a key role in these scenarios in order to reach these goals, but they have to be properly informed in order to make the right decision. This process may be helped and improved by adopting the right technology such as artificial intelligence techniques. During the last months, the attention of many researchers has moved to this new challenge that has involved the entire planet. For instance, a group of researchers and engineers has created a global collaboration, called CORD-19, which collects thousands of scientific publications focused on the new coronavirus [30].



(a) (b) (c)

Fig. 1: (1a) The choropleth map shows the distribution of COVID-19 total cases in the Italian districts (as of 3rd October 2020). The data are divided into quartiles. (1b) The choropleth map shows the average PM_{10} concentrations of 2018 in the Italian districts. The data are divided into quartiles [1]. 1c Number of diagnosed cases of COVID-19 in Italy by test/diagnosis date (from 24th of February 2020 to 8th of April 2021).

Identifying the main factors that contribute to the spread of the SARS-CoV-2 virus is certainly one of the current public health goals. However, the complexity of the phenomenon and the limited availability of information make this study particularly difficult. The possible relationship between fine particulate matter (PM) and SARS-CoV-2 immediately aroused particular interest, especially if one compares the distribution of infections and pollutants.

In [1], authors show that areas with high concentration of pollutants (i.e., dark blue districts in Figure 1b) mostly coincide with areas with high number of positive cases of COVID-19 (i.e., Figure 1a). Unfortunately, the time window used in the two figures is different, mostly due to a lack of information for Italian districts in 2019 and 2020 about the pollutants. However, this observation motivates the interest in investigating the possible correlations between air quality and daily confirmed cases of COVID-19. In addition, the graph of the new daily cases in Italy (Figure 1c) shows an important decrease in infections during the summer months. This leads us to think that there may be a relationship between the environmental temperature and the spread of the pandemic. In this work, we focus on the

[1] https://covid19.who.int/ - Last visited March 31, 2022

study of possible relationships among the number of new daily infected cases and the air quality in some Italian districts.

Contribution. In this work, we performed a cross-correlation analysis that highlights possible relationships among the number of daily cases and several factors related to the air quality. We exploit these correlations by developing and comparing three different supervised learning models. We trained these models to predict the number of new cases of COVID-19, showing that the number of infected cases can be computed in advance with good accuracy. These tools can be used to enrich the set of information available to governments and institutions for helping in making decisions in order to protect the population and stem the pandemic. Indeed, accurate predictive models might help modeling possible scenarios, helping government institutions to better manage the pandemic.

2. Related Works

The rapid spread of the COVID-19 pandemic has attracted the attention of numerous scholars and researchers from many different disciplines. The aim is twofold: on one side, scholars want to understand the modalities of transmission of the SARS-CoV-2 virus and its mechanisms of interaction with the host. On the other side, they want to investigate all possible contributing causes that may have played a key role in the number of infections and in the mortality rate of the disease. Currently, evidence indicates that the SARS-CoV-2 virus spreads mainly from person to person through the inhalation of respiratory droplets, which are normally released when an infected person speaks, coughs or sneezes [15]. However, it is hypothesized that the virus may be aerosolized during certain activities or procedures and may remain active for prolonged periods [28]. On the 30^{th} of January 2020, the Istituto Superiore di Sanità (ISS) confirmed the first two cases of SARS-CoV-2 infection in Italy: two tourists from Wuhan landed in Milan and then hospitalized in Rome. The first autochthonous positive case was confirmed on the 21^{st} of February 2020 and was a patient hospitalized in serious condition in Lodi. Always on the 21^{st} of February 2020 the first death of COVID-19 in the country was reported; he was a man from Vo' (Padova). From the 23^{rd} of February 2020, 11 municipalities in northern Italy (in Lombardy and Veneto) were quarantined and from the 10^{th} of March 2020 the lockdown was extended to the entire country, until the 3^{rd} of June 2020. Of course, the taken measures may have influenced the progress of the pandemic. The study by Lavezzo et al. [16], for example, shows that containment measures have helped to decrease the transmission of SARS-CoV-2 in the municipality of Vo'.

The review of the literature, conducted on the variables that influence seasonal viruses affecting the respiratory system, underlines that there is a multitude of factors that can influence the behaviour of viruses, including humidity, pollution, human behaviour, physiological and demographic characteristics, human mobility, as well as climate change [26]. Each of these factors is important as it affects virus survival, virulence, and transmissibility between individuals. From several studies, that have examined the mechanisms underlying the seasonal nature of respiratory viral infections, it has been deduced that, in general, the two main factors contributing to the spread of virus infections are identifiable in changes in environmental parameters and in human behaviours. A recent investigation [19, 17], that has analysed the mechanisms of action of viruses, shows how the combination of favourable winter levels of humidity, temperature, and solar radiation can compromise our antiviral defense mechanisms, resulting in greater susceptibility of the host to respiratory viruses. Furthermore, various studies [6, 31, 13] report evidence in favour of an association between exposure to air pollutants and the increased risk of respiratory viral infections, although the potential cellular and molecular mechanisms underlying the increased susceptibility are still largely unknown.

The scientific literature that has investigated the possible relationships between environmental factors and SARS-CoV-2 is very large and there are also conflicting opinions. In [23] Setti et al. show how the PM_{10} limit exceedances may be compatible with a role of particulate matter as virus carrier. This hypothesis is also supported by the discovery of the presence of SARS-CoV-2 RNA on atmospheric particulate matter [22]. This is also supported by latest analysis (see for instance[7]). Furthermore, the results of a survey on 120 Chinese cities [32] reveal significantly positive associations between daily measurements of atmospheric particulate matter and nitrogen dioxide and COVID-19 confirmed cases, while sulphur dioxide is negatively associated.

Regarding ozone, many studies (e.g., [20, 14]) claim that it is particularly lethal against viruses due to its high oxidizing property. However, there are no studies confirming the role of ozone in the specific inactivation of SARS-CoV-2. Anyway, it was effective in killing the SARS-CoV virus of the 2003 epidemic [33] and therefore it could also be lethal against SARS-CoV-2 as both viruses come from the same group and have similar structures. This

hypothesis, however, would not seem to agree with the results of a recent research [32] that found a significantly positive association between ozone concentrations and daily COVID-19 confirmed cases.

Finally, the scientific literature argues that high temperature and relative humidity affect the environmental resistance of SARS-CoV-2, reducing its spread. Two different studies [18, 5] show that virus viability decayed more rapidly at higher temperatures, indicating that viral infectivity can be altered with the increase of temperature. Furthermore, a recent research [27] claims the existence of a negative correlation between the average temperature by country and the number of SARS-CoV-2 infections. Regarding relative humidity, another study [29] supports the existence of robust negative associations between humidity and transmission of COVID-19.

3. Background

In this work, we performed an analysis of the collected data, studying the correlation between environmental features and the target variable (i.e., the number of new daily infected cases).

Correlation Analysis. A correlation analysis [10] is a statistical study that evaluates the strength and the sign of a relationship between two variables. There exist many different indexes that can be employed to describe such relationships. The Pearson's correlation coefficient [2] r_p is usually adopted as correlation index. Given two vectors of values $\mathbf{X}$ and $\mathbf{Y}$, the Pearson correlation index can be calculated as $r_p = \frac{\sum_{i=0}^{n-1}(x_i-\hat{x})(y_i-\hat{y})}{\|\mathbf{X}-\hat{x}\|_2 \cdot \|\mathbf{Y}-\hat{y}\|_2}$, where n is the number of samples, x_i and y_i are the samples, $\|\cdot\|_2$ is the l_2-norm, and $\hat{x}$, and $\hat{y}$ correspond to the average values of $\mathbf{X}$ and $\mathbf{Y}$ respectively.

This index assumes that: (i) both variables are normally distributed; (ii) relationship between each of the two variables is linear; (iii) variables have continuous values; (iv) data are equally distributed about the regression line (also known as homoscedasticity). Another function is the so-called Spearman correlation coefficient [21]. This is a simple and efficient way to analyze the similarity of the shape of two time series, it is computed as $r = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$, where d_i is the difference between the two ranks of each observation and n is the number of observations. Spearman coefficient operates on raw data, it is based on the ranks of the data, it is insensitive to outliers and can operate with ordinal values. Due to the characteristics of the data, in this work, we adopted the Spearman correlation coefficient.

The value of the coefficient is in $[-1, 1]$ and it describes the relationship between the variables. Specifically, a high value, close to $+1$, indicates a positive correlation, while a low value, close to -1, indicates a negative correlation. A positive correlation exists when the increase in the value of one variable makes also increase the value of the other variable. On the other hand, if a negative correlation exists, then as the value of one variable increases, the value of the other variable decreases. Furthermore, when the index value is close or equal to 0 there is a poor or no correlation between the two variables, which means that increasing or decreasing one variable does not affect the value of the other variable. In particular, we will use the following terminology based on the absolute value of r, we say that there is no correlation or very weak correlation when $r < 0.3$; weak correlation when $0.3 \leq r < 0.5$; moderate when $0.5 \leq r < 0.7$; and we will say that the correlation is strong when $r \geq 0.7$.

It is important to remember that the correlation analysis does not provide any indication of a cause-effect relationship between the variables. To establish a true causal condition, the variables should be completely isolated from any other possible confounding variable. If a correlation is found between air quality and SARS-CoV-2 infection, this would constitute just one more proof to be able to subsequently support any scientific demonstration.

Machine Learning Techniques. Machine learning is a branch of AI that studies and develops learning algorithms able to model intrinsic characteristics or relationships in the data. Usually, machine learning algorithms have a bottom-up approach, which means that they infer information from a collection of data called a dataset, which describes the studied scenario. Thus, a dataset is an M × N matrix in which each column corresponds to a variable (also called "feature") that describes a specific characteristic of the domain, and each row corresponds to a sample $c_i = (x_i; y_i)$ where $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,N})$ and y_i represents the "label" of the sample (i.e., the value of a variable that we want to predict), which is not always known a priori. The problem addressed in this paper is a supervised learning problem. The term "supervised" refers to the fact that in the set of samples the labels $y_1, \ldots, y_M$ are already known. In this approach, we assume that there exists an ideal function $f : X \to Y$ such that $f(x_i) = y_i$, where X is the space of all possible samples and Y is the set of all possible outputs. Supervised learning tries to find a function $\bar{f} : X \to Y$ that

approximates f as closely as possible, finding the same labels as f for most of the samples. In this work we adopt Random Forest [3], XGBoost [12, 11], and Neural Network [25].

4. Empirical Study

In this section, we describe the datasets and the results of the correlation analysis performed on them. We describe the predictive models and their performances on the different datasets.

4.1. Data Collection

The data used in this work comes from two different sources: one contains the daily details of the pandemic, and the other contains the environmental information of different districts or geographical areas around the world.

We focus on the pandemic situation in Italy. Italian data about the pandemic has been made available in a GitHub repository under a CC-BY-4.0 license from the Italian Civil Protection Department (ICP) [2]. In this repository, the number of total cases is available at the level of each Italian district. For security and privacy reasons, other information (i.e., the number of infected cases in a specific city or the number of deaths per district) is stored and protected in a platform of the Integrated Surveillance and thus accessible only by authorized people [3]. Therefore, only the daily number of total cases of COVID-19 in different areas is used to describe the progress of the pandemic in Italy.

Table 1: List of considered variables in the datasets.

Variable	*Description*	*Measure*	*Source*
date	Date		
humidity_median	Daily median of the relative humidity	percentage	ARPA
no2_median	Daily median concentration of NO_2 (nitrogen dioxide)	$\mu g/m^3$	ARPA
o3_median	Daily median concentration of O_3 (ozone)	$\mu g/m^3$	ARPA
pm10_median	Daily median concentration of PM_{10}	$\mu g/m^3$	ARPA
pm2.5_median	Daily median concentration of $PM_{2.5}$	$\mu g/m^3$	ARPA
so2_median	Daily median concentration of SO_2 (sulfur dioxide)	$\mu g/m^3$	ARPA
temp_median	Daily median temperature	Celsius	ARPA
total_cases	Cumulative number of COVID-19 cases		ICP
new_cases	Number of new daily cases of COVID-19		ICP

The time window of this study goes from the 1^{st} of January 2020 to the 8^{th} of April 2021 (date of our last measurement). We collected environmental data from the Air Quality Open Data Platform (AQODP) [4]. This platform was created by the World Air Quality Index project team and contains meteorological and air quality information of major cities around the world, unfortunately not all the data published in this website is validated. But it is worth noting that data of Italian districts contained in AQODP is provided by the ARPA (Agenzia Regionale per la Protezione Ambientale), which is an official and trusted source for this kind of data. The AQODP platform publishes information about 12 districts of Italy. Due to the fact that many features of 4 out of the 12 districts have missing values, we decided to focus exclusively on the eight districts with the most complete set of data. Specifically, we used data about Bologna, Brescia, Milan, Modena, Naples, Parma, Prato, and Rome. For each of these districts, we derived a dataset merging data from the two aforementioned sources: each row in a dataset describes information about environmental factors and the number of newly infected cases for a specific date. All the adopted variables are summarized in Table 1.

[2] https://github.com/pcm-dpc/COVID-19 - Last visited on March 30 2022
[3] https://www.epicentro.iss.it/en/coronavirus/sars-cov-2-integrated-surveillance-data - Last visited on March 30, 2022
[4] https://aqicn.org/data-platform/covid19/ - Last visited on March 30, 2022

4.2. Data Analysis

Initially, we performed a pre-processing task which removed all negative values for the *new_cases* variable. These negative values occur when the Italian Civil Protection Department adjusted the daily data about total cases for some areas, resulting in a reduced number of infected people compared to the number of the previous day. This was probably due to errors in the positive cases count. In addition, all records with missing values were removed before the correlation analysis.

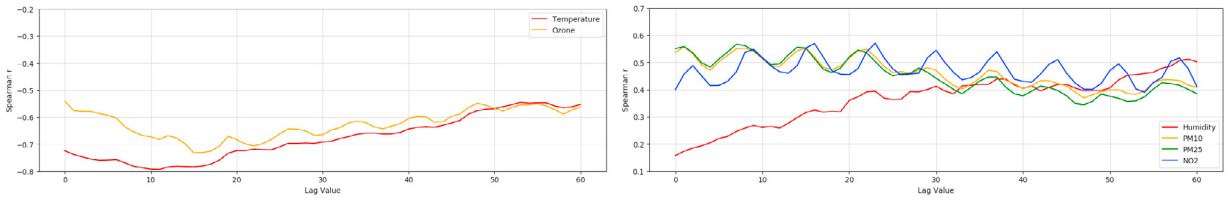
In order to perform an accurate correlation analysis, it is necessary to take into account a probable incubation period of the virus. It is important to notice that an additional delay time might be due to the bureaucracy related to the execution and analysis of the nasopharyngeal swab. This value was not known to us a priori.

In this study, cross-correlations are used for the analysis of time-lagged relationships between several environmental factors and their possible influence on the number of new positive cases. The use of the cross-correlation functions allows to assess the sensitivity and responsiveness at different time [9]. This is due to the fact that a specific environmental configuration may influence the spread of the virus, but consequences may be evident only some days later. For a specific factor evidence, the amount of days needed is not known a priori. Therefore, to find the best time-lag, we shift the number of new daily cases of i positions in the datasets (with i varying from 0 to 60). This is done to compare the environmental data of a given day with the number of new infected cases after i days. In this way, we looked for the maximum correlation value of each environmental parameter in a time window of two months in the past.

4.3. Correlation Analysis Results

In general, we observed a strong negative correlation with temperature and ozone; a moderate positive correlation with NO_2, $PM_{2.5}$, PM_{10}, and humidity; a poor positive correlation with SO_2. Depending on the area, the results are different. Among all the analyzed districts, here we report results for Brescia dataset and a brief discussion about the differences with other datasets. Due to the lack of space full results are not reported.

In the district of Brescia (which was one of the most compromised during the pandemic), we found a strong negative correlation with the temperature and a negative correlation with ozone. This means that as the daily temperature increases, we observe that the number of daily infections decreases. In particular, the correlation peak between temperature and new COVID-19 cases occurred for $i = 10$ ($r = -0.7799$, $p\_value < 0.001$). Similarly, ozone has a strong negative correlation with COVID-19 cases and the pick is for $i = 15$ ($r = -0.7358$, $p\_value < 0.001$). Therefore, an increase in the maximum concentration of ozone in the atmosphere is associated with a decrease in positive cases. Figure 2a depicts the different values for the correlation indexes when we shift the time-lag window from 0 to 60 days. As you can notice, after the peak the effect of these two factors on the virus decreases as expected.



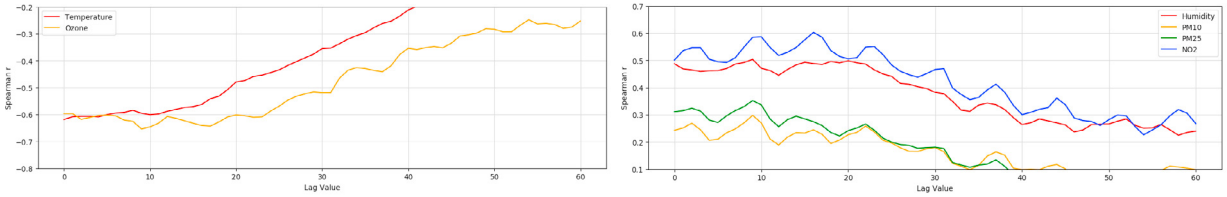
(a) (b)

Fig. 2: 2a Area of Brescia: cross-correlations values for temperature, ozone, and new daily infected cases, sliding time-window from 0 to 60. 2b Area of Brescia: cross-correlations values for NO_2, $PM_{2.5}$, PM_{10}, and new daily infected cases, sliding time-window in 0 to 60.

On the other hand, the number of new infected cases and the concentration of the main air pollutants (i.e., NO_2, $PM_{2.5}$, and PM_{10}) show a moderate correlation at close but different lag values. Specifically, the median concentrations of NO_2, $PM_{2.5}$ and PM_{10} are positively correlated and the values of the correlation coefficients are respectively: for NO_2 after $i = 23$ days, $r = 0.5699$ ($p\_value < 0.001$); for $PM_{2.5}$ after $i = 21$ days, $r = 0.5639$ ($p\_value < 0.001$); and for PM_{10} after $i = 22$ days, $r = 0.5617$ ($p\_value < 0.001$). These results indicate that a higher median daily concentration of these pollutants is associated with a greater number of people contracting the infection after more or

less 20 days. Figure 2b depicts the different values for correlation indexes when we shift the time-lag window from 0 to 60 days. The oscillating behaviour that can be observed in Figure 2b might be due to traffic emissions or industrial productions which are higher during working days of the week and should decrease in the week-end (also due to the restrictions imposed). In fact, oscillations have a period of 7 days. Similar results were also obtained with data regarding the areas of Milan, Bologna, Parma and Modena.

The results obtained for the datasets of Naples, Prato, and Rome are much weaker. The moderate correlation with temperature and ozone might be caused by a set of co-factors probably not considered in this study. For instance, all the other considered areas are in the Po Valley which is a geographical area surrounded by the Alps and the Apennines. The wind is rare and the air is colder in the plains than in the mountains, causing emissions stay above Po Valley and making harder for natural and artificial emissions to be dissolved. This may be one of the reasons why the atmospheric pollutants seem to have more effects on the spread of the virus in areas of Po Valley than in other Italian areas which instead are close to the sea like Rome or Naples. Although the sign of the correlation coefficients are in line with those of the other northern provinces.



(a) (b)

Fig. 3: 3b Area of Rome: cross-correlations values for temperature, ozone, and new daily infected cases, sliding time-window from 0 to 60. 3b Area of Rome: cross-correlations values for temperature, ozone, and new daily infected cases, sliding time-window from 0 to 60.

The correlation peak occurs at distance of different days depending on the parameter analyzed: atmospheric particulate matter, median nitrogen dioxide and ozone are more related to new cases registered after 22-23 days from the environmental measurements, while temperature is more related to cases identified 9-10 days later. This difference could be due to a different effect of these factors on the virus. Indeed, atmospheric pollutants could take a few days to decrease the environmental resistance of SARS-CoV-2, while the contribution of the temperature to the spread of the virus could be more immediate.

4.4. Machine Learning Models

Given the statistically significant correlation of some variables with COVID-19 cases, we developed and trained some regressors for predicting the number of new infected cases based on the values of the environmental parameters. The machine learning algorithms used in this work estimate the impact of the environmental factors on the spread of the COVID-19 pandemic. The aim of these models is to predict the number of confirmed cases given the measurements of the atmospheric variables.

The simulations are developed in Python 3.7. We adopted *RandomForestRegressor* by Scikit-learn, *XGBRegressor* by XGBoost and *Sequential Model* by Keras for Neural Network.

We develop a shallow neural network with two hidden layers and 6,871 total trainable parameters. For this model data were standardized using *StandardScaler()* by Scikit-learn. Each model was trained separately for each of the eight districts for both the value of i corresponding to the pollutant correlation peak (e.g., $i = 22$) and for that corresponding to the temperature correlation peak (e.g., $i = 10$). This was done to compare the results from different trainings in order to determine which is the best delay time for predicting new cases in each district.

In order to estimate the generalization performance, we adopted the holdout approach, i.e. the dataset was divided into two disjoint sets, called training sets and test sets. The training set contains 70% of the instances of the original dataset and it was used to train the model. The test dataset contains the remaining 30% of the samples, and it was used to test the generalization level of the regressor. The split of the instances was done randomly.

For *RandomForestRegressor* a tuning phase of the hyperparameters was performed in order to obtain the best possible accuracy. To do that, we adopted a grid search approach (i.e., a list of allowed values is specified for each

hyperparameters and then they are evaluated through a 5-fold cross validation to determine the best combination). We chose to optimize the following parameters: i)*max_depth*, that is the maximum depth that each tree can have. Values for this parameter were searched in the interval [3,7). ii) *n_estimators*, that is the number of trees. Values for this parameter were searched in {10, 50, 100, 1000}.

In order to evaluate and compare the different models, we compute the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), and the R^2 score on the test set. RMSE and MAE are measures of error, therefore when we compare two regression models on the same dataset, the one with the lowest values is the one with the best predictions. In contrast, R^2 score or coefficient of determination represents the proportion of variance of y that has been described by the independent variables of the model. This metric provides an indication of the goodness of fit and thus it is a measure of the likelihood that samples never seen by the model are predicted correctly. The best possible score is 1 and it occurs when it is possible to predict exactly what the value of the target variable will be, knowing the values of the independent variables. A constant model that always predicts the expected value of y, ignoring the input features, has a R^2 score equal to 0. The value of R^2 can also be negative as the model can be arbitrarily worse than a constant model. Therefore, if we compare two regression models on the same dataset, the model with the greater R^2 score will be the one with the highest predictive power. Furthermore, the Mean 5-Fold Cross Validation score was computed over the entire dataset. This latter approach of evaluation combines the 5-fold cross validation technique with the R^2 score in order to obtain a more generalized result.

4.5. Prediction Results

As expected, the results obtained for each district are different. The models have greater predictive capacity for the area of Milan (Table 3), followed by Bologna. The maximum performance using Milan dataset was achieved with the Random Forest algorithm. The best delay time due to the identification of the disease was found to be 10 days (although good results were obtained even with i = 15).

Table 2: Performance of the models on the test set of the province of Milan.

i value	*Model*	R^2 *score*	*RMSE*	*MAE*	*5-fold CV score*
10	XGBoost	0,55	530,08	280,62	0,66 ± 0,10
10	Neural Network	0,67	454,51	277,60	0,67 ± 0,10
10	**Random Forest**	**0,74**	**405,36**	**241,27**	**0,73 ± 0,02**
15	XGBoost	0,71	447,70	265,24	0,65 ± 0,13
15	Neural Network	0,75	441,84	262,80	0,65 ± 0,10
15	Random Forest	0,66	518,65	283,73	0,72 ± 0,07
21	XGBoost	0,36	719,91	364,29	0,54 ± 0,20
21	Neural Network	0,41	684,29	374,53	0,49 ± 0,16
21	Random Forest	0,46	660,96	338,90	0,53 ± 0,18

Instead, for Bologna district the maximum Cross Validation accuracy was 70% ($DS = 5\%$) by adopting XGBoost as predictive model and 16 days as delay time. On the other hand, performances of Brescia, Modena, Naples, Parma, Prato and Rome are weaker, with Cross Validation accuracy values ranging between 36 % and 63 %. In general, the models that perform better are XGBoost and Random Forest, while the neural network performs worst.

The low accuracy values of the models could hide a dependence of the target variable, that is the number of daily COVID-19 cases in Italy, on many factors, probably not observable from the available data. Indeed, it must be taken into account that the main modality of transmission of the virus is direct contact between people. Consequently, the behavior of the population and containment measures could affect the number of infections. Another important factor is certainly the number of swabs carried out daily in each district. Unfortunately this information is not available at district level.

5. Discussion

In this paper we tried to understand, on the basis of current knowledge and available data, whether air quality may play a role in the spread of the COVID-19 pandemic and, in particular, on the number of recorded daily cases in several Italian districts. Of course, the spread of a viral infection is a complex and multi-factor system. Therefore, our analysis includes some limitations: i) **Open-source datasets**: open source datasets containing official information at district level are very little, poor or completely missing for some areas. Although numerous research studies on the COVID-19 pandemic have been published so far, in most cases the databases used have not been made available or they include data only at national level, such as the well-known dataset by Johns Hopkins University [5]. Furthermore, almost all the information at provincial level about the spread of the pandemic in Italy is stored in a platform accessible only to the Istituto Superiore della Sanità (ISS) and other authorized entities. ii) **Data accuracy**: the Civil Protection Department has repeatedly corrected past data, published in the repository, modifying the daily data. Of course, the inaccuracy in the number of documented infected can cause a significant increase in the uncertainty of the estimate provided by the prediction models based on historical data. In addition, environmental measurements may also be subject to error. Indeed, the World Air Quality Index project team has underlined that not all data have been validated. iii) **Missing dimensions**: the employed machine learning models does not take into account (or at least very marginally) the time dimension. We only consider the data for a given day and we are not taking into account any restriction that the government enacted during the pandemic. We presume that the information of previous days may influence the daily data. Thus considering the time dimension might improve the performance of regressors. iv)**Confounding factors**: factors capable of generating spurious associations, which could have altered the results. For instance, the restrictive measures and the rigor with which they have been observed is a possible confounding factor. As well as the number of daily swabs which varied considerably over the time. v) **Lack of knowledge**: this work is based on a still uncertain understanding of the phenomenon. There are many questions that the research has yet to answer. For example, it is not clear to what extent surfaces and aerosols favor the transmission of the virus and whether it is actually possible that SARS-CoV-2 can travel incorporated into air pollution particles while maintaining its vitality. vi) **Data availability, quality and representativeness**: machine learning algorithms require a large amount of data to be able to accurately learn the relationships between variables. A small dataset could be poorly representative of the variability of interactions and could consequently lead to low predictive performance. The limited observation period, due to the recent discovery of the virus, could therefore represent a further limitation of the empirical study.

6. Conclusion

In this work, we investigate possible relationships among environmental parameters, geographical distribution and the spread of the COVID-19 pandemic in different Italian areas. The analysis highlights a possible diagnostic delay period. It has also shown that machine learning techniques can be applied to make useful predictions on the number of COVID-19 cases per day as a function of environmental data measurements. For instance, the possibility of predicting future new infected cases could be useful to make adequate decisions on the management of the pandemic, avoiding the overload of the health system.

This work can be seen as another step towards the understanding of a complex system, which deserves to be investigated through in-depth scientific studies. Future epidemiological investigations should be based on sufficiently extensive and comprehensive data. In addition, further studies aimed at investigating the possible mechanisms of interaction of environmental factors with COVID-19 are needed.

In the future, the analysis will be extended to other geographical areas and additional co-factors will be included in the dataset in order to improve the performance of the models. We plan to investigate the application of new ensemble methods (e.g., [8]) to improve performance, and recurrent neural networks to consider also the time dimension in the analysis.

[5] https://coronavirus.jhu.edu/map.html - Last visited on March 30 2022

References

[1] Becchetti, L., Conzo, G., Conzo, P., Salustri, F., 2020. Understanding the heterogeneity of adverse COVID-19 outcomes: the role of poor quality of air and lockdown decisions. Available at SSRN 3572548 .

[2] Benesty, J., Chen, J., Huang, Y., Cohen, I., 2009. Pearson correlation coefficient, in: Noise reduction in speech processing. Springer, pp. 1–4.

[3] Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

[4] Chen, X., Yao, L., Zhou, T., Dong, J., Zhang, Y., 2021. Momentum contrastive learning for few-shot COVID-19 diagnosis from chest CT images. Pattern Recognition 113, 107826. doi:https://doi.org/10.1016/j.patcog.2021.107826.

[5] Chin, A.W., Chu, J.T., Perera, M.R., Hui, K.P., Yen, H.L., Chan, M.C., et al., 2020. Stability of SARS-CoV-2 in different environmental conditions. The Lancet Microbe 1, e10.

[6] Ciencewicki, J., Jaspers, I., 2007. Air pollution and respiratory viral infection. Inhalation toxicology 19, 1135–1146.

[7] Collivignarelli, M.C., Abbà, A., Caccamo, F.M., Bertanza, G., Pedrazzani, R., Baldi, M., Ricciardi, P., Miino, M.C., 2021. Can particulate matter be identified as the primary cause of the rapid spread of COVID-19 in some areas of northern italy? Environmental Science and Pollution Research , 1–13.

[8] Cornelio, C., Donini, M., Loreggia, A., Pini, M.S., Rossi, F., 2021. Voting with random classifiers (VORACE): theoretical and experimental analysis. Autonomous Agents and Multi-Agent Systems 35, 22. doi:10.1007/s10458-021-09504-y.

[9] Derrick, T.R., Thomas, J.M., 2004. Time series analysis: the cross-correlation function .

[10] Franzese, M., Iuliano, A., 2019. Correlation analysis, in: Encyclopedia of Bioinformatics and Computational Biology. Academic Press, Oxford, pp. 706 – 721.

[11] Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics , 1189–1232.

[12] Friedman, J.H., 2002. Stochastic gradient boosting. Computational statistics & data analysis 38, 367–378.

[13] Grigg, J., 2018. Air pollution and respiratory infection: an emerging and troubling association.

[14] Hudson, J.B., Sharma, M., Vimalanathan, S., 2009. Development of a practical method for using ozone gas as a virus decontaminating agent. Ozone: science & engineering 31, 216–223.

[15] La Rosa, G., Bonadonna, L., Lucentini, L., Kenmoe, S., Suffredini, E., 2020. Coronavirus in water environments: Occurrence, persistence and concentration methods-a scoping review. Water Research , 115899.

[16] Lavezzo, E., Franchin, E., Ciavarella, C., Cuomo-Dannenburg, G., Barzon, L., Del Vecchio, C., Rossi, L., Manganelli, R., Loregian, A., Navarin, N., et al., 2020. Suppression of a SARS-CoV-2 outbreak in the italian municipality of vo'. Nature 584, 425–429.

[17] Lolli, S., Chen, Y.C., Wang, S.H., Vivone, G., 2020. Impact of meteorological conditions and air pollution on COVID-19 pandemic transmission in italy. Scientific reports 10, 1–15.

[18] Magurano, F., Baggieri, M., Marchi, A., Rezza, G., Nicoletti, L., Group, C..S., et al., 2020. SARS-CoV-2 infection: the environmental endurance of the virus can be influenced by the increase of temperature. medRxiv .

[19] Moriyama, M., Hugentobler, W.J., Iwasaki, A., 2020. Seasonality of respiratory viral infections. Annual Review of Virology 7, 83–101. doi:10.1146/annurev-virology-012420-022445. pMID: 32196426.

[20] Murray, B.K., Ohmine, S., Tomer, D.P., Jensen, K.J., Johnson, F.B., Kirsi, J.J., Robison, R.A., O'Neill, K.L., 2008. Virion disruption by ozone-mediated reactive oxygen species. Journal of virological methods 153, 74–77.

[21] Myers, L., Sirois, M.J., 2004. Spearman correlation coefficients, differences between. Encyclopedia of statistical sciences .

[22] Setti, L., Passarini, F., De Gennaro, G., Barbieri, P., Perrone, M.G., Borelli, M., Palmisani, J., Di Gilio, A., Torboli, V., Fontana, F., et al., 2020a. Sars-cov-2rna found on particulate matter of bergamo in northern italy: First evidence. Environmental Research , 109754.

[23] Setti, L., Passarini, F., de Gennaro, G., Di Gilio, A., Palmisani, J., Buono, P., Fornari, G., Perrone, M.G., Piazzalunga, A., Barbieri, P., et al., 2020b. Evaluation of the potential relationship between particulate matter (pm) pollution and COVID-19 infection spread in italy , 1–6.

[24] Shaban, W.M., Rabie, A.H., Saleh, A.I., Abo-Elsoud, M., 2021. Accurate detection of COVID-19 patients based on distance biased naïve bayes (DBNB) classification strategy. Pattern Recognition 119, 108110. doi:https://doi.org/10.1016/j.patcog.2021.108110.

[25] Shalev-Shwartz, S., Ben-David, S., 2014. Understanding machine learning: From theory to algorithms. Cambridge university press.

[26] Sloan, C., Moore, M.L., Hartert, T., 2011. Impact of pollution, climate, and sociodemographic factors on spatiotemporal dynamics of seasonal respiratory viruses. Clinical and translational science 4, 48–54.

[27] Sobral, M.F.F., Duarte, G.B., da Penha Sobral, A.I.G., Marinho, M.L.M., de Souza Melo, A., 2020. Association between climate variables and global transmission of sars-cov-2. Science of The Total Environment 729, 138997.

[28] Van Doremalen, N., Bushmaker, T., Morris, D.H., Holbrook, M.G., Gamble, A., Williamson, B.N., Tamin, A., Harcourt, J.L., Thornburg, N.J., Gerber, S.I., et al., 2020. Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. New England Journal of Medicine 382, 1564–1567.

[29] Wang, J., Tang, K., Feng, K., Lin, X., Lv, W., Chen, K., Wang, F., 2021. Impact of temperature and relative humidity on the transmission of COVID-19: a modelling study in china and the united states. British Medical Journal Publishing Group 11. doi:10.1136/bmjopen-2020-043863.

[30] Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., et al., 2020. Cord-19: The COVID-19 open research dataset. ArXiv .

[31] Xing, Y.F., Xu, Y.H., Shi, M.H., Lian, Y.X., 2016. The impact of pm2. 5 on the human respiratory system. Journal of thoracic disease 8, E69.

[32] Yongjian, Z., Jingu, X., Fengming, H., Liqing, C., 2020. Association between short-term exposure to air pollution and COVID-19 infection: Evidence from china. Science of the total environment , 138704.

[33] Zhang, J.m., ZHENG, C.y., XIAO, G.f., ZHOU, Y.q., GAO, R., 2004. Examination of the efficacy of ozone solution disinfectant in inactivating sars virus. Chinese Journal of Disinfection 1.