MDPI

*Article*

# Is a COVID-19 Second Wave Possible in Emilia-Romagna (Italy)? Forecasting a Future Outbreak with Particulate Pollution and Machine Learning

**Silvia Mirri, Giovanni Delnevo and Marco Roccetti ***

Department of Computer Science and Engineering, University of Bologna, Bologna, 40127, Italy; silvia.mirri@unibo.it (S.M.); giovanni.delnevo2@unibo.it (G.D.)

* Correspondence: marco.roccetti@unibo.it

Received: 24 July 2020; Accepted: 20 August 2020; Published: date

**Abstract:** The Nobel laureate Niels Bohr once said that: "Predictions are very difficult, especially if they are about the future". Nonetheless, models that can forecast future COVID-19 outbreaks are receiving special attention by policymakers and health authorities, with the aim of putting in place control measures before the infections begin to increase. Nonetheless, two main problems emerge. First, there is no a general agreement on which kind of data should be registered for judging on the resurgence of the virus (e.g., infections, deaths, percentage of hospitalizations, reports from clinicians, signals from social media). Not only this, but all these data also suffer from common defects, linked to their reporting delays and to the uncertainties in the collection process. Second, the complex nature of COVID-19 outbreaks makes it difficult to understand if traditional epidemiological models, such as susceptible, infectious, or recovered (SIR), are more effective for a timely prediction of an outbreak than alternative computational models. Well aware of the complexity of this forecasting problem, we propose here an innovative metric for predicting COVID-19 diffusion based on the hypothesis that a relation exists between the spread of the virus and the presence in the air of particulate pollutants, such as $PM_{2.5}$, $PM_{10}$, and $NO_2$. Drawing on the recent assumption of 239 experts who claimed that this virus can be airborne, and further considering that particulate matter may favor this airborne route, we developed a machine learning (ML) model that has been instructed with: (i) all the COVID-19 infections that occurred in the Italian region of Emilia-Romagna, one of the most polluted areas in Europe, in the period of February–July 2020, (ii) the daily values of all the particulates taken in the same period and in the same region, and finally (iii) the chronology according to which restrictions were imposed by the Italian Government to human activities. Our ML model was then subjected to a classic ten-fold cross-validation procedure that returned a promising 90% accuracy value. Finally, the model was used to predict a possible resurgence of the virus in all the nine provinces of Emilia-Romagna, in the period of September–December 2020. To make those predictions, input to our ML model were the daily measurements of the aforementioned pollutants registered in the periods of September–December 2017/2018/2019, along with the hypothesis that the mild containment measures taken in Italy in the so-called Phase 3 are obeyed. At the time we write this article, we cannot have a confirmation of the precision of our predictions. Nevertheless, we are projecting a scenario based on an original hypothesis that makes our COVID-19 prediction model unique in the world. Its accuracy will be soon judged by history—and this, too, is science at the service of society.

**Keywords:** COVID-19; predictions; second wave; machine learning models; air pollution; Emilia-Romagna; Italy

## 1. Introduction

"The virus remains the public enemy number one", World Health Organization (WHO), Director General, Tedros Adhanom Ghebreyesus maintained at a recent press conference, and he also added that: "If basics are not followed, the only way the pandemic is going to go, it is going to get worse and worse and worse" [1]. These threatening words are justified in light of the current pandemic numbers. As of 17 July 2020, global COVID-19 cases exceed 13.5 million, and 584,940 people have died of it in almost seven months, with the current biggest rises in the United States, Brazil, India, and South Africa [2].

When COVID-19 first struck several provinces of Northern Italy in early 2020 (especially in Lombardy and in Emilia-Romagna), the conditions there made it a perfect storm. The virus outbreak spread with an unusual violence (in the period from late February to April 2020), with a catastrophic toll in terms of human deaths. Still now, several months after that last virus surge, and a severe subsequent lockdown period, the consequences are profound. Italy counts a total number of 252,235 registered infections, and as many as 35,231 human deaths (as of 13 August 2020) [3]. Not only that, but some recent financial studies also estimate that Italy's Gross Domestic Product (GDP) could drop significantly in 2020 due to the impact of the pandemic, with some industrial sectors severely hit, including textile, train and air transport, hotels, restaurants, entertainment, and automotive [4].

The proportion of this disaster is key to understanding why policymakers, health officials, and media in general have an increasing interest in making use of computational models that can forecast possible resurgences of the virus, in order to put in place containment measures [5]. Unfortunately, there are several problems here, primarily linked to collecting data, and then using them to feed an adequate forecasting model.

Along this line of reasoning, we propose a clear direction. We do believe that a relationship exists between particulate matter (of various types) and COVID-19 incidence, and that this favors the spread of the contagion. We have devoted a previous study to verifying the presence of such a possible correlation between the series of the new daily COVID-19 infections in the period February–April 2020 in Emilia-Romagna (Italy) and the correspondent series of the daily values of the $PM_{2.5}$, $PM_{10}$, and $NO_2$ pollutants [6]. A specific statistical hypothesis testing method was then employed (i.e., the Granger causality statistical methodology [7,8]), which returned a positive response to our question based on a complex set of experiments that extended before, during, and after the lockdown periods decided by the Italian Government on 8–10 March 2020.

Obviously, it is not our intention to run through, again, all the technical and epistemic issues behind this hypothesis here. The interested reader can refer to [6]; nonetheless, some of the basics that lie behind our decision to use this hypothesis to select the data used to make predictions on future COVID-19 outbreaks need to be discussed.

First, it is out of discussion that poor air quality easily brings one to a state of permanent inflammation and chronic respiratory difficulties, along with a hyper-activation of the immune system. All these circumstances make human lungs prone to be attacked by respiratory viral infections [9]. Owing to this condition, it has been demonstrated that humans living in highly polluted areas have a reduced respiratory capacity to react to virus attacks [10]. In addition to these general considerations, which are confirmed by an impressive wealth of recent literature [11–13], more interesting is the biological phenomenon at the center of the following controversy: Can particulate matter be a carrier for COVID-19?

To respond to this question, it would be enough to remind a recent claim of 239 experts who maintained that this virus can be airborne [14], united with the information of the presence of the COVID-19 RNA, found in the particulate matter of Bergamo (Italy) [15]. All these seem to confirm that this virus can create clusters with particulate matter, and that it can be carried by this type of microscopic pollutants.

To close this issue: Although we are aware that there is an ongoing scientific controversy, concerning the link between that first experimental finding (i.e., [15]) and the degree of severity with which a COVID-19 outbreak can spread [16], we believe that both our previous study and those

detailed in [17,18] provide a support to the hypothesis that the presence of COVID-19 on outdoor air samples can represent a potential early indicator of the diffusion of the virus in a given area.

Hence, based on the hypothesis that this virus can be airborne and assuming that particulate matter may favor this airborne route, we developed a machine learning model (ML, for short), with special attention to the Italian region of Emilia-Romagna (Italy). Our model was instructed with:

- All the COVID-19 infections that occurred in Emilia-Romagna, one of the most polluted areas in Europe, in the period of February–July 2020;
- The daily values of all the aforementioned particulates taken in the same period and in the same region; and finally,
- The chronology according to which restrictions were imposed by the Italian Government to human activities in the same period under observation.

Our ML model was then subjected to a classic testing procedure that has returned a promising accuracy value of approximately 90%. Finally, the model was used to predict a possible second wave of the virus for all the nine provinces of Emilia-Romagna, in the period of September–December 2020.

To get the predictions, input to our ML model were the daily measurements of the aforementioned pollutants registered in the periods of September–December 2017/2018/2019, along with the hypothesis that the mild containment measures taken by Italy in the so-called *Phase 3* are obeyed [19].

Having covered the reasoning behind our choice, we shall return now to possible alternatives.

Inspired by a wealth of recent literature, new techniques have been proposed to aggregate data that could predict the pandemic's next moves. For example, drawing on the use of new information technologies, including search engines, news reports, crowdsourced infoveillance, Twitter feeds, travel data, tele-traffic measurements, and many others again, the authors of [20] exploited a Bayesian model that calculates, in near-real time, the probability of an exponential growth or subsequent decay of the virus spread, based on data collected in the USA, between January and June 2020. Interesting in this kind of study is the fact that data from Twitter and Google searches emerge as the earliest uptrend signals to anticipate a virus surge (with a median earliness of 2–3 weeks), while UpToDate (an evidence-based clinical decision support system, developed by the health division of Wolters Kluwer [21]) was capable of providing early signals of uptrend for deaths (earliness of 4.5 weeks). Additionally, Google searches, united with the elaboration of some form of mobility data from citizens, provided early downtrend signals to anticipate a virus decay (median earliness of 2 weeks).

This type of proposal appears as an advancement to the state-of-the-art, especially if one considers that, as far as data are concerned, the problem is that virus case counts, hospitalized patients, number of deaths, reports from clinicians, etc. all suffer from reporting delays (as well as from uncertainties in the data collection process).

While we refrain from expressing non-positive comments on this research, we have to admit that it is certainly true that combining many streams of real time information may lead decision makers to be responsive to sudden changes; nevertheless, crucial remains the issue of how reliable and precise those streams of observations are when it comes to describing a pandemic spread, especially if no working hypothesis lies behind. Told in simpler words, the strength of the approach here is also its weakness: What the authors of [20] are doing is observing, without making any assumptions. This could be just a little bit extreme, since we all know that it is not the first time in the history of the science of data that one realizes, just at the end of the process, that too many data can be a bad thing. Making useful predictions requires something more than data, in fact—for example, some strong conceptual insights, as discussed at length in [22].

Let us talk now about forecasting models in more detail. Once, it was the SIR (susceptible, infectious, or recovered) model that dominated this scenery. A survey on this model is out of the scope of this paper, and the interested reader can refer to [23]. The actual value and importance of this traditional model is, obviously, out of the question in the epidemiological field; nonetheless, new proposals are emerging for modeling the COVID-19 pandemic that share similar goals, such as making predictions on the disease spread, yet adopting different computational methodologies.

Among these new proposals, the lion's share is played by machine learning models. The majority of the ML models used in practice are supervised. Learning, with supervision, involves

learning a function that maps an input to an output based on examples of input–output pairs [24]. Providing a very simple example: If we had a set of data, regarding children with age in the range of 0–10 years, along with their correspondent weight, we could implement a very simple supervised ML model that predicts the weight of a child, based on their exact age.

Returning to the use of ML models for predicting a COVID-19 emergence, exemplary is the case of the work done in [25]. There, the authors provide a comparative analysis of various ML models to predict COVID-19 outbreaks. After a study of different ML models, based on the collection of data on infectious cases for 30 days from five different countries (Italy, China, Iran, Germany, and the USA), their most prominent finding is that the multilayered perceptron (MLP) model delivers the most accurate results, in terms of predicting an outbreak, without the assumptions that epidemiological models typically require.

Nevertheless, a criticism that we pose to articles of similar tenor is that all these studies can be assimilated to a process that starts from a bunch of example data and learns to point to the most likely output; where the meaning of likely is usually vague or fuzzy [26–30] or stochastic at best [31]. While we agree on the fundamental role played by data in these models, our belief is that, at least, a conceptual hypothesis should exist that drives one in their choice and selection.

Returning to our approach, we would like to conclude this section certain that the reader has a clear vision of our position, before they proceed with the article. To this aim, we have already stated in our previous work that neither Granger nor any other statistical testing procedure can provide final evidence that the two phenomena, between which we conjecture a relationship (i.e., pollution and COVID-19 infection spread), are correlated in nature. Additionally, the same holds for the predictions of our ML model. Nonetheless, with our predictions, we are projecting a scenario based on an original assumption that makes our COVID-19 ML model unique, as it selects the data to be used based on a well-defined and unambiguous hypothesis. Whatever will happen in September–December 2020, we will have learnt an important truth about the validity of our hypothesis—and this, too, is science at the service of society.

The remainder of the paper is structured as follows. In the next section, we describe the methodology behind our approach. Section 3 presents and critically discusses the results we yielded. Finally, Section 4 concludes the paper, with some final considerations.

## 2. Methods

We now present, first, some preliminary information relevant to the present study, second, a description of the dataset we used, and third, the reasoning we used to precisely decide what kind of predictions we are looking for. Finally, we provide some reflections on how we have selected the ML model that could fit squarely into our COVID-19 scenario of interest.

### *2.1. Preliminary Assumptions*

As already anticipated, we have based this study on the precise idea that the correlation between air pollution (specifically, the $PM_{2.5}$, $PM_{10}$, and $NO_2$ pollutants) and the spread of COVID-19 infections in the Emilia-Romagna region is a very plausible hypothesis. Using that hypothesis, we consequently selected the data of interest.

We do not return, again, to this main assumption; it suffices here to remind that the presence of COVID-19 on air samples can represent an early indicator of the diffusion of the virus in a given geographical area [18]. To further summarize this concept, one should consider that, whatever the origin of this virus is, there are clear indications that COVID-19 transmission occurs from infected people, either through virus-laden droplets or aerosol transmissions. In this second case of airborne transportation, pollutants may help the diffusion, playing the role of additional carriers. All this is graphically summarized, with relative simplifications, in the following Figure 1, where it is crystal clear that the arrows in the figure should not be intended as a means to represent a direct causation, but they amount to a simple indication of a conceptual path; that is, infection propensity is favored by the transmission of droplets and aerosols, with air pollutants as further carriers.

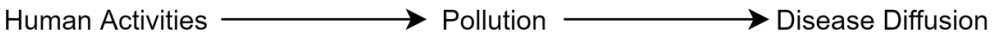


**Figure 1.** Infection propensity.

All this anticipated, it is also important to give some details about the Italian region we took into consideration in our studies: Emilia-Romagna. The region is situated in the northeast section of the country and is divided into nine provinces: Bologna, Ferrara, Forlì-Cesena, Modena, Parma, Piacenza, Reggio nell'Emilia, Rimini, and Ravenna. It is populated by almost 4,500,000 citizens and was one of the more seriously hit by this virus in Italy, with a total number of infections of 30,342, and as many as 4298 fatalities, as of 13 August 2020. Its death toll linked to the virus is second, in Italy, only to Lombardy, where, on the same date, more than 97,000 persons were registered as infected, and the fatalities had almost reached the number of 17,000.

Relevant for considering this process from the right temporal perspective is also the chronology according to which restrictions were first imposed to human activities in those provinces, and then released after a substantial decay of the virus incidence. In particular, we can count four subsequent phases:

- Phase 0: Prior to 8 March 2020, no specific restriction was imposed, which was valid for all the nine provinces of Emilia-Romagna, except for some local control measures (for example, for schools and universities);
- Phase 1: A full lockdown was first imposed to the provinces of Modena, Parma, Piacenza, Reggio nell'Emilia, and Rimini, as of 8 March 2020 [32], and then extended to the remaining provinces of Bologna, Ferrara, Forlì-Cesena, and Ravenna on 10 March, 2020 [33];
- Phase 2: On 4 May 2020, the lockdown was partially released, though with several commercial and industrial activities still suspended, as well as the obligation for people to stay in quarantine if found or suspected ill, wear cloth face covering in public settings, wash hands frequently, etc., and where a social distancing of at least 1 meter and a half was difficult to maintain [34];
- Phase 3: On 14 June 2020, the lockdown was almost completely removed, with almost all activities resuming, provided that the personal protection measures mentioned above were obeyed [19].

*2.2. Dataset Description*

Based on the hypothesis that a relation exists between pollutants and infections, at least in Emilia-Romagna, the data we used to instruct our ML model were essentially of two types:

- The measurements of the particulate pollutants: $PM_{2.5}$, $PM_{10}$, and $N0_2$; taken on a daily basis, for all the aforementioned provinces (Bologna, Ferrara, Forlì-Cesena, Modena, Parma, Piacenza, Reggio nell'Emilia, Rimini, and Ravenna).
- The number of the daily COVID-19 infections, again for all the provinces mentioned above.

The amount of daily infections was collected using the GitHub repository of the Italian Civil Protection, for the entire period, starting on 24 February and closing on 7 July 2020 [35].

The daily values of the pollutants, by contrast, were collected using the website of the Regional Environmental Protection Agencies (ARPA) of the Emilia-Romagna region for all the nine provinces [36]. With various ARPA monitoring stations distributed over each province, an average of the values returned by each station was computed on a daily/provincial basis. The period along which those data were collected was from 10 February up to 30 June 2020.

The two periods have the same temporal length, but there is a discrepancy in the starting/closing dates. This is due to the fact that a typical COVID-19 infection can be subjected to an incubation period, whose duration can range from a few days to almost 14 before a human begins to manifest

some symptoms. Many papers provide evidence of this fact, concluding, at the end, that 99% of the infected population develop symptoms within 14 days [37].

Following this reasoning, the period during which we measured the particulates started on 10 February and closed on 30 June 2020, while the infections were registered in the period of 24 February–7 July, 2020. Simply told, if we want to instruct an ML model with a function that maps input (particulates) into output (infections), based on examples of the input–output pairs we have collected, an offset has to be introduced that temporally separates these two time-series. This stems from the simple consideration that all that can happen on a given day, say $x$, in terms of augmented spread of the virus due to pollution, may have its effects in terms of manifestations of the infections up to day $x + 14$.

This is not still enough, though: In fact, the function that our ML model has to learn is a little bit more complex than usual, as we need to also take into consideration the specific period during which a given event (for example, an infection) has occurred. It makes a great difference, in fact, whether we consider events occurring during either Phase 0, or Phase 1, or Phase 2, or finally Phase 3. In conclusion, in addition to the data that are part of the relationship between particulates and infections, input to the ML model should also be the various phases through which the management of the spread of the infections has passed.

To conclude and summarize this complex situation, the following three figures show the entire dataset we have used, in a graphical form. All these graphs simultaneously show the curves for both the pollutants (black, measured in micrograms per cubic meter) and COVID-19 infections (gray, measured in units). On the x axis of all graphs reported are the timelines for the pollutants' series (in black) and for the infections (in gray). Important to note is the temporal offset explained above. Figures 2–4 are, respectively, relative to $PM_{2.5}$, $PM_{10}$, and $NO_2$. Moreover, in each graph, with the colors orange, yellow, light blue, and green, the passage through the different four phases we have mentioned is demarcated.

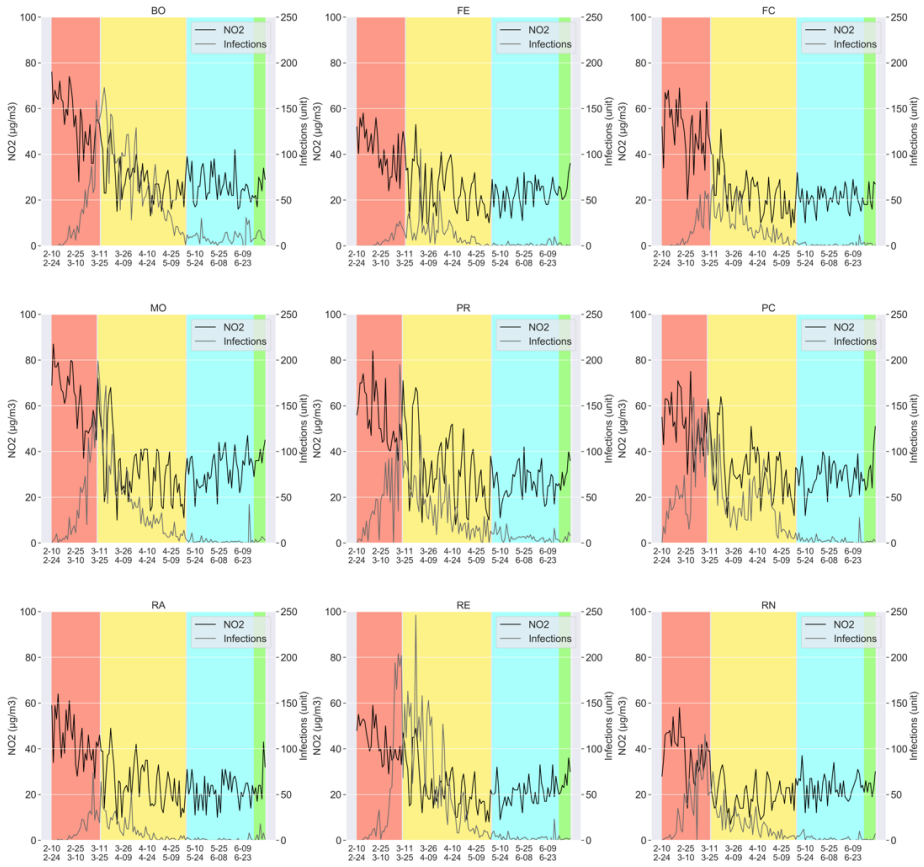


**Figure 2.** The dataset (I): infections, pollutant ($NO_2$), and phases.

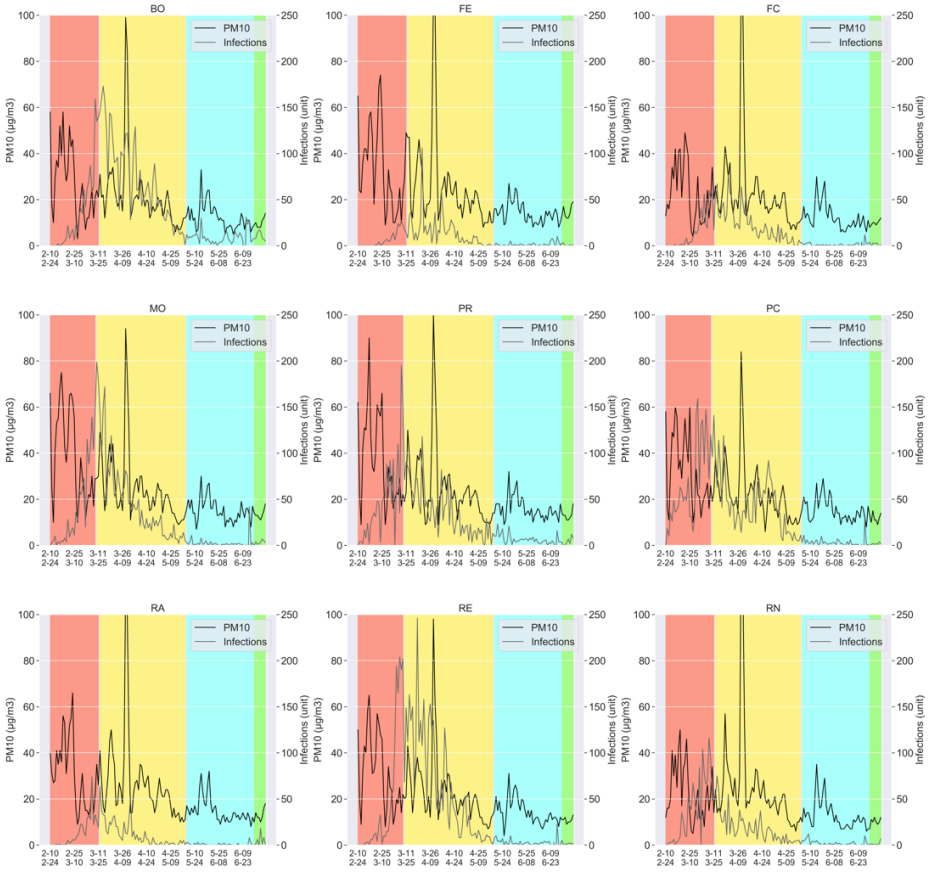


**Figure 3.** The dataset (II): infections, pollutant ($PM_{10}$), and phases.

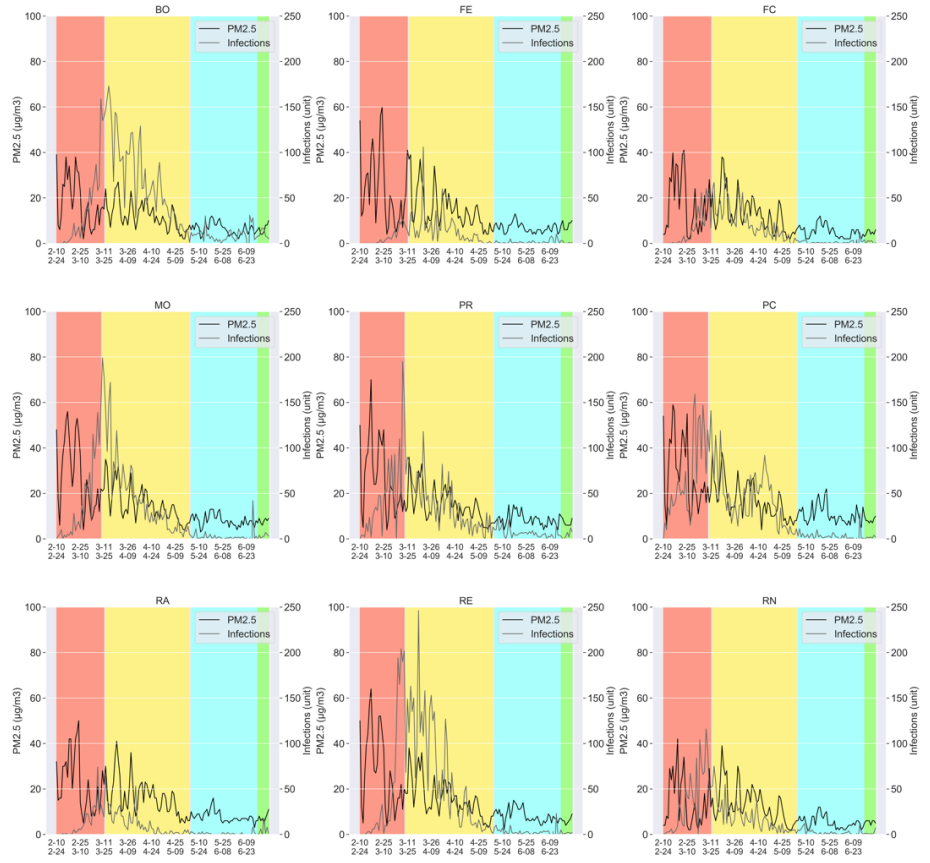


**Figure 4.** The dataset (III): infections, pollutant ($PM_{2.5}$), and phases.

*2.3. What Kind of Predictions Are We Looking for?*

In essence, the point is: If we have an ML model that has learnt from data the possible relationship between the presence of particulate in the air and the incidence of the virus, what kind of predictions should we ask of our model?

Let us say that the situation could become complex. In fact, while it is true that we are interested in knowing whether a second wave of COVID-19 could hit the provinces of the Emilia-Romagna region, we cannot ignore that trying to extract, from an ML model, a prediction on the exact number of new infected people, per each province, on a daily basis, is something more like a puzzle, rather than a scientific investigation.

To simplify this problem, we resorted to a more effective procedure which was as follows. The idea was to count the number of daily infections registered per each province, in all the nine provinces of Emilia-Romagna, during the four days that preceded the lockdown decision taken by the Italian Government on 8–10 March 2020 (the specific day depends on the specific province).

Once those infections counts were obtained, we computed an average value of those daily numbers on a per-province basis for those 4 days. We then got nine numbers that were finally aggregated on a regional basis, under the form of a further average count, thus yielding the average number of infections per-province on a regional basis in Emilia-Romagna. The result was 17 (from now on, the so-called threshold). Told differently, the daily number of infected people, in Emilia-Romagna, averaged over those four days, amounts to 17 times 9 = 153.

Now, please follow the reasoning. If the Italian Government, using its own decisional models, opted for a lockdown decision, as soon as the average regional number of daily infections on a per-province basis in Emilia-Romagna had surpassed the threshold of 17, then we could use that number as a key to design the predictions scheme of our ML model. Not to forget also the fact that Emilia-Romagna was, at that time, the region with the largest number of infections after Lombardy. Hence, the number of infections that occurred in this region has had an important role in that lockdown decision.

To conclude this reasoning, our intention is to replace the initial idea to predict if, in a given future day, Emilia-Romagna is under the risk of a second wave of a COVID-19 resurgence with the more concrete and effective prediction of whether the number of infected people will surpass that threshold of 17, on that day, on a per-province, regional basis. More precisely, we ask our ML model to compute the probability that, in a given future day, each province in Emilia-Romagna will count a number of infections larger than 17—and, then, we look at the regional picture with all its nine provinces, and the probability that the number of infections for each exceeds 17. The higher this probability is, the higher the risk of a second regional wave will be, especially if various provinces simultaneously surpass that threshold on a certain given day.

For the sake of completeness, in Figure 5, we provide a graph with the cumulative quantities of infected people, per day, for all the nine provinces of interest, plus the cumulative values of the regional and the national averages, registered during the four days prior to 8–10 March 2020.

In Figure 5, one can read: Bologna, bo; Ferrara, fe; Forlì-Cesena, fc; Modena, mo; Parma, pr; Piacenza, pc; Ravenna, ra; Reggio nell'Emilia, re; Rimini, rn; Emilia-Romagna, er; and Italy, ita.

Important to note is the fact that, in Figure 5, our regional infection average, being cumulative over those four days, amounts to 17 times 4 = 68 (as read at the rightmost end of the figure).

By contrast, if one takes into consideration the national average, they can notice that the following value of 8 times 4 = 32 can be computed (as read at the rightmost end of the figure). This smaller quantity at a national level is due to the fortunate fact that many regions in Southern Italy were not severely affected by the virus, thus providing a smaller contribution to the national average.

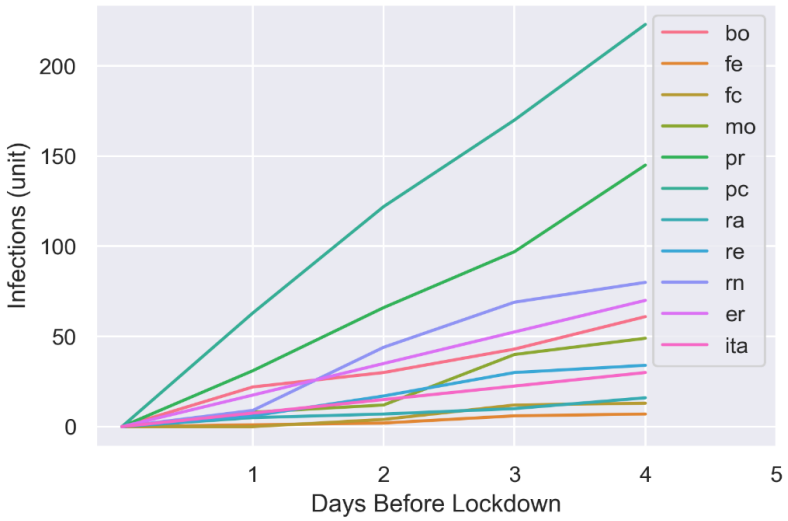


**Figure 5.** 4–7 March 2020—cumulative number of infections: Modena, Parma, Piacenza, Reggio nell'Emilia, and Rimini; 7–10 March 2020—cumulative number of infections: Bologna, Ferrara, Forlì-Cesena, and Ravenna; cumulative regional and national infections averages.

Interesting to remind is also the average number of infections per day in Lombardy, computed in a similar way (i.e., the average per-province number of infections, on a regional basis), which was as high as 38. This latter number is important. It is well known that in that period Lombardy was really the hardest-hit Italian area, thus becoming a sort of hotspot for COVID-19 diffusion in Italy. This is why we decided not to choose this number (i.e., 38) in our scheme. It would have been somewhat misleading, especially in consideration of the fact that we want predictions that are valid for the Emilia-Romagna region.

At this point, it is important to mention that, using the value of 17, we have essentially split our initial dataset into two separate portions:

i) The former, with all those days with a number $y$ of daily infections, equal or smaller than 17; and
ii) The latter, with those days registering a number of new infected people larger than 17.

Not only this, but also, to properly manage the hypothesis of a relationship between pollutants and infection spread, crucial is also the concept of *lag*. In particular, with *lag*, we account for the following fact: On a given day, say $z$, we may have registered a certain number of infections, say $y$. Those $y$ infections could have manifested themselves after exactly fourteen days, since the original contagion happened exactly on the day: $z$ - 14. Nonetheless, we also know that there is a degree of uncertainty, affecting the exact number of days that should be taken into account for this count.

To take this fact into account, with a *lag* equal to 4, for example, we reason as if all the $y$ infections, which occurred on day $z$, originated from the contribution of pollutants that were in the air during a longer time interval of length 4 (starting from day: $z - 14$), in this specific case, our interval would go from day $z - 14$ up to day $z - (14 - 4 + 1)$, that is, day: $z - 11$.

This is an important fact, giving rise to an important implication: With the concept of *lag*, which can range from 1 to 8 in our model, we try to mitigate the uncertainty concerning the exact day when people get infected (as also discussed in [37]).

To conclude, considering the nine provinces of Emilia-Romagna, each one observed for a period of 135 days, the number of examples we counted where the number of daily infections was equal to or smaller than 17 amounted to 789. By contrast, the number of examples where the number of daily infections was larger than 17 amounted to 426.

Finally, it is important also to note that **80%** of all these data were employed for instructing our ML models, while the remaining 20% were retained to test the performance those models could reach upon completion of the learning phase.

*2.4. Model Selection*

We have reached the following point: We have collected a special set of data, based on the hypothesis of a relation between pollutants and infections. Those data represent each single day, with its infections and quantities of measured pollution in the air. Then, we have divided them into two separate sets of examples, specifically: (i) the first one comprising those days with a number of infections equal to or smaller than 17, and (ii) the second one with those days with a number of infected people larger than 17. We have also introduced the concept of *lag* and have exactly computed how large the two aforementioned sets of examples are (**789** vs. **426**).

What we still need to decide now is the ML model to adopt that should be instructed with all those data.

To choose our ML model, we proceeded as follows. Without any initial preference, we tried to instruct a wide range of possible ML models, suitable to learn the function pollutants/infections. We started with the following ML models:

- K nearest neighbor (KNN) [38];
- Classification and regression tree (CART) [39];
- Support vector machine (SVM) [40];
- Multilayer perceptron (MLP) [41];
- Ada boosting with decision tree (AB) [42];
- Gradient boosting (GB) [43];
- Random forest (RF) [44];
- Extra tree (ET) [45].

The procedure with which we selected our ML model went through two separate and subsequent phases, aimed at measuring their performance in terms of accuracy of the predictions they made, more precisely, a ten-fold cross-validation and a testing phase.

First, we allowed all the eight models mentioned above to learn the function we described before, and then we subjected each one to a classic ten-fold cross-validation procedure, yielding an F1 score. Before we proceed, we shall briefly remind what a ten-fold cross-validation procedure and a F1-score are.

Simply put, cross-validation is a procedure that evaluates predictive models by partitioning the original dataset into two portions. With the training portion, the model is trained, while with the validation portion, the model is evaluated. In a ten-fold cross-validation, the original dataset is randomly partitioned into 10 subsamples of equal size. Of the 10 portions, a single portion is kept separate to validate the model, while the model is trained with the remaining nine portions of data. We use the term cross as this validation procedure is reiterated 10 times, with each of the 10 portions used exactly once to validate the model. The ten obtained results coming from the validation portions can then be averaged to produce a final evaluation.

As regards the F1 score, in a classic classification problem (comprising true and false positives, and true and false negatives), it is intended to be the harmonic mean of the precision and recall values, where such a score reaches its best at 1. In turn, precision is the number of true positives divided by the number of true positives plus the number of false positives, while recall is the number of true positives divided by the number of true positives plus the number of false negatives (i.e., all the samples that should have been identified as positive).

All this anticipated, in Table 1 , we show the results we have obtained with **80%** of our data, and a ten-fold cross-validation conducted with all the eight ML models mentioned before. All the results are in terms of the F1 score, which was measured on average, plus its standard deviation.

Important to note is the fact that we allowed the models to learn our function both with each single pollutant (i.e., $PM_{2.5}$, $PM_{10}$, and $NO_2$) in isolation, and then with all the pollutants considered together; not only this, but we also varied the *lag*, as already anticipated, from 1 to 8.

In essence, each cell in Table 1 tells us how accurate, on average, the prediction was that a given model has made that the threshold of 17 infections was either surpassed or not, for a given day, with a certain amount of pollutants in the air.

If one accurately analyzes Table 1, they can find that almost all the ML models have comparable performances, except for CART and AB (highlighted with the red color). This convinced us to proceed with the next step of testing, at the end of which only one model was to be selected to be used to make COVID-19 predictions for the period of September–December 2020, excluding the CART and the AB candidates.

| Pollution | Lag | KNN | CART | SVC | MLP | AB | GB | RF | ET |
|---|---|---|---|---|---|---|---|---|---|
| All | 1 | 0.81 ± 0.03 | 0.75 ± 0.05 | 0.81 ± 0.04 | 0.84 ± 0.03 | 0.78 ± 0.05 | 0.81 ± 0.04 | 0.81 ± 0.03 | 0.78 ± 0.04 |
| | 2 | 0.81 ± 0.03 | 0.81 ± 0.04 | 0.83 ± 0.03 | 0.84 ± 0.04 | 0.79 ± 0.05 | 0.83 ± 0.03 | 0.84 ± 0.05 | 0.83 ± 0.04 |
| | 3 | 0.83 ± 0.05 | 0.78 ± 0.05 | 0.83 ± 0.04 | 0.84 ± 0.04 | 0.81 ± 0.04 | 0.83 ± 0.02 | 0.85 ± 0.03 | 0.85 ± 0.03 |
| | 4 | 0.82 ± 0.04 | 0.78 ± 0.05 | 0.84 ± 0.04 | 0.84 ± 0.05 | 0.81 ± 0.03 | 0.84 ± 0.03 | 0.84 ± 0.03 | 0.85 ± 0.03 |
| | 5 | 0.85 ± 0.03 | 0.82 ± 0.05 | 0.85 ± 0.03 | 0.86 ± 0.04 | 0.82 ± 0.05 | 0.86 ± 0.04 | 0.87 ± 0.04 | 0.86 ± 0.03 |
| | 6 | 0.86 ± 0.03 | 0.82 ± 0.05 | 0.87 ± 0.03 | 0.86 ± 0.03 | 0.82 ± 0.04 | 0.86 ± 0.03 | 0.85 ± 0.04 | 0.88 ± 0.03 |
| | 7 | 0.86 ± 0.03 | 0.83 ± 0.04 | 0.87 ± 0.03 | 0.87 ± 0.02 | 0.84 ± 0.03 | 0.85 ± 0.03 | 0.86 ± 0.03 | 0.87 ± 0.04 |
| | 8 | 0.87 ± 0.02 | 0.84 ± 0.04 | 0.89 ± 0.03 | 0.89 ± 0.02 | 0.82 ± 0.03 | 0.86 ± 0.03 | 0.86 ± 0.04 | 0.90 ± 0.03 |
| PM2.5 | 1 | 0.79 ± 0.03 | 0.77 ± 0.04 | 0.80 ± 0.04 | 0.81 ± 0.03 | 0.78 ± 0.06 | 0.81 ± 0.03 | 0.76 ± 0.03 | 0.77 ± 0.04 |
| | 2 | 0.80 ± 0.04 | 0.78 ± 0.05 | 0.82 ± 0.04 | 0.82 ± 0.04 | 0.78 ± 0.04 | 0.82 ± 0.04 | 0.81 ± 0.03 | 0.79 ± 0.03 |
| | 3 | 0.81 ± 0.03 | 0.79 ± 0.05 | 0.82 ± 0.03 | 0.82 ± 0.04 | 0.81 ± 0.03 | 0.84 ± 0.03 | 0.83 ± 0.03 | 0.82 ± 0.03 |
| | 4 | 0.80 ± 0.03 | 0.81 ± 0.04 | 0.85 ± 0.02 | 0.83 ± 0.03 | 0.81 ± 0.03 | 0.85 ± 0.04 | 0.85 ± 0.03 | 0.83 ± 0.03 |
| | 5 | 0.84 ± 0.03 | 0.82 ± 0.04 | 0.86 ± 0.03 | 0.85 ± 0.04 | 0.81 ± 0.04 | 0.86 ± 0.04 | 0.87 ± 0.03 | 0.85 ± 0.02 |
| | 6 | 0.85 ± 0.04 | 0.84 ± 0.04 | 0.87 ± 0.03 | 0.87 ± 0.03 | 0.82 ± 0.04 | 0.87 ± 0.04 | 0.87 ± 0.03 | 0.87 ± 0.03 |
| | 7 | 0.85 ± 0.04 | 0.85 ± 0.03 | 0.87 ± 0.03 | 0.86 ± 0.02 | 0.82 ± 0.05 | 0.86 ± 0.04 | 0.88 ± 0.03 | 0.88 ± 0.03 |
| | 8 | 0.88 ± 0.03 | 0.84 ± 0.04 | 0.88 ± 0.03 | 0.87 ± 0.03 | 0.83 ± 0.05 | 0.86 ± 0.03 | 0.87 ± 0.04 | 0.89 ± 0.04 |
| PM10 | 1 | 0.79 ± 0.05 | 0.77 ± 0.03 | 0.80 ± 0.04 | 0.81 ± 0.04 | 0.78 ± 0.05 | 0.81 ± 0.04 | 0.78 ± 0.05 | 0.79 ± 0.04 |
| | 2 | 0.81 ± 0.04 | 0.79 ± 0.05 | 0.81 ± 0.04 | 0.82 ± 0.04 | 0.78 ± 0.04 | 0.82 ± 0.03 | 0.82 ± 0.05 | 0.82 ± 0.04 |
| | 3 | 0.80 ± 0.03 | 0.77 ± 0.03 | 0.82 ± 0.03 | 0.83 ± 0.04 | 0.80 ± 0.03 | 0.83 ± 0.03 | 0.83 ± 0.04 | 0.83 ± 0.03 |
| | 4 | 0.83 ± 0.03 | 0.78 ± 0.03 | 0.84 ± 0.03 | 0.84 ± 0.04 | 0.80 ± 0.02 | 0.85 ± 0.04 | 0.84 ± 0.02 | 0.84 ± 0.03 |
| | 5 | 0.84 ± 0.04 | 0.81 ± 0.06 | 0.85 ± 0.03 | 0.86 ± 0.03 | 0.80 ± 0.05 | 0.87 ± 0.04 | 0.86 ± 0.03 | 0.85 ± 0.03 |
| | 6 | 0.85 ± 0.03 | 0.82 ± 0.04 | 0.86 ± 0.03 | 0.87 ± 0.03 | 0.82 ± 0.04 | 0.86 ± 0.04 | 0.86 ± 0.04 | 0.85 ± 0.04 |
| | 7 | 0.87 ± 0.03 | 0.85 ± 0.04 | 0.88 ± 0.03 | 0.87 ± 0.03 | 0.83 ± 0.05 | 0.87 ± 0.04 | 0.87 ± 0.03 | 0.88 ± 0.03 |
| | 8 | 0.87 ± 0.02 | 0.85 ± 0.03 | 0.88 ± 0.03 | 0.88 ± 0.03 | 0.82 ± 0.04 | 0.88 ± 0.04 | 0.87 ± 0.04 | 0.89 ± 0.03 |
| NO2 | 1 | 0.80 ± 0.04 | 0.78 ± 0.03 | 0.81 ± 0.03 | 0.81 ± 0.04 | 0.78 ± 0.04 | 0.80 ± 0.04 | 0.77 ± 0.03 | 0.78 ± 0.03 |
| | 2 | 0.79 ± 0.03 | 0.76 ± 0.04 | 0.81 ± 0.02 | 0.82 ± 0.02 | 0.79 ± 0.05 | 0.81 ± 0.03 | 0.80 ± 0.04 | 0.79 ± 0.04 |
| | 3 | 0.82 ± 0.03 | 0.77 ± 0.03 | 0.82 ± 0.03 | 0.83 ± 0.03 | 0.80 ± 0.03 | 0.82 ± 0.04 | 0.83 ± 0.02 | 0.83 ± 0.02 |
| | 4 | 0.85 ± 0.02 | 0.80 ± 0.02 | 0.83 ± 0.03 | 0.84 ± 0.04 | 0.80 ± 0.04 | 0.83 ± 0.03 | 0.86 ± 0.03 | 0.85 ± 0.02 |
| | 5 | 0.86 ± 0.02 | 0.81 ± 0.03 | 0.84 ± 0.03 | 0.85 ± 0.04 | 0.82 ± 0.04 | 0.83 ± 0.03 | 0.87 ± 0.03 | 0.85 ± 0.02 |
| | 6 | 0.86 ± 0.03 | 0.83 ± 0.04 | 0.85 ± 0.03 | 0.86 ± 0.04 | 0.80 ± 0.04 | 0.84 ± 0.04 | 0.87 ± 0.03 | 0.86 ± 0.03 |
| | 7 | 0.85 ± 0.03 | 0.82 ± 0.05 | 0.85 ± 0.02 | 0.85 ± 0.03 | 0.81 ± 0.04 | 0.84 ± 0.04 | 0.87 ± 0.04 | 0.87 ± 0.03 |
| | 8 | 0.86 ± 0.03 | 0.81 ± 0.03 | 0.86 ± 0.02 | 0.86 ± 0.03 | 0.81 ± 0.04 | 0.85 ± 0.04 | 0.87 ± 0.03 | 0.88 ± 0.02 |

**Table 1.** Ten-fold cross-validation for all the eight ML models: F1 score (average and standard deviation).

In this second step were, hence, included only the six models that exhibited good comparable performances during the ten-fold cross-validation, namely: KNN, SVC, MLP, GB, RF, and ET.

All these six models were subjected to this final testing step, conducted with 20% of the dataset we had retained for this specific aim. Results from this final testing phase are reported in Table 2.

As expected, all the six models under consideration yielded a reasonably good performance; nonetheless, the one with the best F1 score was GB, gradient boosting, as Table 2 reveals. The simple reason we expected quite good performances from almost all those six models is that they had already done well during the phase of the ten-fold cross-validation. Nonetheless, gradient boosting (GB) manifested as the best model in this specific circumstance (with its F1 score equal to 0.893). In other words, GB is the model that better learned the function pollution/infections on which our hypothesis is based. Consequently, it is the best candidate for making accurate predictions for the future.

| Algorithm | Testing | | | |
|---|---|---|---|---|
| | Class | Precision | Recall | F1 score |
| KNN | <=17 | 0.90 | 0.85 | 0.845 |
| | >17 | 0.75 | 0.82 | |
| SVC | <=17 | 0.95 | 0.87 | 0.890 |
| | >17 | 0.80 | 0.92 | |
| MLP | <=17 | 0.93 | 0.90 | 0.890 |
| | >17 | 0.82 | 0.87 | |
| GB | <=17 | 0.92 | 0,91 | 0.893 |
| | >17 | 0.84 | 0.86 | |
| RF | <=17 | 0.93 | 0.87 | 0.878 |
| | >17 | 0.79 | 0.88 | |
| ET | <=17 | 0.91 | 0.91 | 0.881 |
| | >17 | 0.83 | 0.84 | |

**Table 2.** Testing phase: gradient boosting (GB) selected by virtue of its F1 score.

In addition to the F1 score, where the GB surpasses all the other five candidates, with regard to the other metrics of precision and recall, it is interesting to note that it achieves a better accuracy (91–92%) in predicting whether a given future day will be classified in the class of those days with a number of infections equal or smaller than 17, and a slightly lower accuracy (84–86%) in predicting whether a given day will be classified in the class of those days with a number of infections larger than 17. This slight difference is probably due to the fact that it has been instructed with a larger number of examples of the former class.

To conclude this section, a simple explanation on how the GB model computationally works is in order now.

In very simple words, the gradient boosting method tries to find an approximation to the function $\hat{F}$ that we are letting our model learn (i.e., the relationship of pollutants vs. infections). To do that, a value is computed based on a weighted sum of $M$ functions $h_i$, which are, in some sense, the estimators of the number $y$ of infected people we expect to have for each given day, given that we have registered a certain value $x$ of some pollutant. All this is based on the following formula (where $a_i$ is the additional parameter to be learnt, and $\varepsilon$ is a given predefined constant value).

$$\hat{F}(x) = \sum_{i=1}^{M} \alpha_i h_i(x) + \varepsilon. \quad (1)$$

Technically speaking, we are minimizing loss function $L$, given a training set composed of couples of known values of $x$ (pollutant) and $y$ (infections), where the final target is to make the estimation as close as possible to the real value of $y$. All this is based on the following minimization procedure:

$$\hat{F}(x) = \ min_F \, E_{x,y}[L(y, F(x)]. \quad (2)$$

With this clarified, in the next section, we present the predictions that our GB model has made, regarding the plausibility of a second wave of COVID-19 infections in Emilia-Romagna.

## 3. Results: Predictions

We come now to the final step. Upon completion of the activities that led us to instruct our GB model using data from the period of February–July 2020, selected based on the assumption of a relationship between pollutants and infections, we now need to ask our model to make the predictions, on a daily and provincial basis, for the Emilia-Romagna region, for all the future days from 21 September to 31 December 2020.

The motivation behind the choice of this precise prediction period is obvious. We are all worried about the possibility that a second wave of COVID-19 will coincide with the end of the summer period (21 September), when many human activities will be resumed in Italy, including schools and universities, for example. As for the closing period for our predictions, we deem it natural not to extend the scope too much, thus reaching the end of the current year 2020.

Nonetheless, one element is still missing, which is relevant to our prediction activity. Our model has learnt the function that maps pollutant values into the number of infected people. Nevertheless, if we want it to try to predict what can happen on, say, day $z$, (with, for example, $z$ = 21 September 2020, in the province of Bologna), we need to give our model as an input the value of the pollutants circulating on that day $z$ in Bologna.

Obviously, at the time we write our article, we do not have the precise value of those pollutants for that future day $z$. What we can do to mitigate that factor is to try to estimate those values, based on the amounts of pollutants circulating in the air in Bologna, as measured on the same day $z$ some given years ago, for example, 21 September 2019.

Put simply, we have exploited the amount of the pollution registered in some previous years in Emilia-Romagna to have an estimate of those pollutants that need then to be given as an input to the ML model. We have done this following two alternative strategies. In the first case, we used all the values of the pollutants registered in the period 21 September–31 December 2019. In the second case, we used all the values of the pollutants measured in the period 21 September–31 December, yet averaged on three different previous years, namely: 2017–2019.

We present the obtained results in the following two subsections, in isolation.

Before we proceed, it is very important to remind that *all* the predictions presented in the two following subsections were made under the hypothesis that all the control/containment measures of the so-called Phase 3 are strictly obeyed. If those measures are not obeyed (or even partially disregarded), our model would return predictions very different from those shown in Sections 3.1. and 3.2.

### *3.1. Predictions: 2019 -> 2020*

We report in Table 3 the predictions that our GB model has made based on all the assumptions described in the previous sections, including that Phase 3 is obeyed.

Important are the following instructions to better read those results. Along the columns, we have all the nine provinces (Bologna, Ferrara, Forlì-Cesena, Modena, Parma, Piacenza, Reggio nell'Emilia, Rimini, and Ravenna), while on the rows, the prediction is given for each single day. A mixture of the values of the pollutants $PM_{2.5}$, $PM_{10}$, and $NO_2$ was considered as input to the model, in this case measured in the period 21 September–31 December 2019.

Each cell in the table shows the value of the probability that the number of infections, on that day per that province, exceeds the quantity of 17 (with the maximum probability value set equal to 1). The higher that probability, the higher the risk that we will have a number of infected people surpassing 17, on that day in that province, thus raising the relative concerns.

In a red color, we highlighted those days, on a per-province basis, where that threshold is surpassed. A quick comment to this table is that concerns arise, especially for two provinces (Parma and Piacenza), in the two periods of mid-October/mid-November 2020, as well as at the end of

November–end of December 2020. Again, we insist on the fact that these predictions were obtained based on the assumption that the personal protection measures of Phase 3 are respected.

We have deliberately moved a more detailed discussion of these results to the final section of this paper.

Here, we just input evidence that the following provinces seem to have the following total number of crucial days during the observed period, whose length is 135 days:

- Bologna (0);
- Ferrara (1);
- Forlì-Cesena (2);
- Modena (1);
- Parma (16);
- Piacenza (23);
- Ravenna (0);
- Reggio Emilia (1);
- Rimini (1).

Moreover, to allow the reader to a have a simpler and more comprehensive view of the results presented above, we have also reported them under an alternative format. In particular, in Figure 6, we present the same results as those of Table 3, but portrayed as a heatmap. Simply put, high probability values turn lean toward red, while low probability values are depicted in white. Different orange color gradations represent intermediate situations. Predictions are grouped on a weekly basis, per each province in the region.

| Day | Bo | Fe | Fc | Mo | Pr | Pc | Ra | Re | Rn |
|---|---|---|---|---|---|---|---|---|---|
| 9/21 | 0,00 | 0,00 | 0,00 | 0.02 | 0.01 | 0.35 | 0,00 | 0.01 | 0,00 |
| 9/22 | 0.04 | 0,00 | 0.04 | 0.11 | 0.11 | 0.38 | 0.03 | 0.03 | 0,00 |
| 9/23 | 0.1 | 0.08 | 0.07 | 0.18 | 0.3 | 0.15 | 0.08 | 0.22 | 0.07 |
| 9/24 | 0.14 | 0.17 | 0.17 | 0.27 | 0.11 | 0.31 | 0.14 | 0.12 | 0.11 |
| 9/25 | 0.01 | 0.01 | 0.01 | 0.19 | 0.24 | 0.09 | 0,00 | 0.07 | 0.01 |
| 9/26 | 0,00 | 0.01 | 0.01 | 0.01 | 0.11 | 0.12 | 0,00 | 0.03 | 0,00 |
| 9/27 | 0,00 | 0.01 | 0,00 | 0.1 | 0.06 | 0.07 | 0.03 | 0.02 | 0.01 |
| 9/28 | 0,00 | 0.02 | 0.01 | 0.05 | 0.2 | 0.09 | 0.02 | 0.06 | 0.01 |
| 9/29 | 0.14 | 0,00 | 0.04 | 0.23 | 0.04 | 0.02 | 0.16 | 0.04 | 0.02 |
| 9/30 | 0.03 | 0.01 | 0.01 | 0.04 | 0.01 | 0.02 | 0.06 | 0.01 | 0.02 |
| 10/1 | 0.01 | 0.01 | 0.01 | 0.02 | 0.01 | 0.04 | 0.01 | 0.02 | 0,00 |
| 10/2 | 0,00 | 0.01 | 0.05 | 0.23 | 0.53 | 0.07 | 0,00 | 0.12 | 0,00 |
| 10/3 | 0,00 | 0.02 | 0.03 | 0.13 | 0.54 | 0.46 | 0.02 | 0.25 | 0.01 |
| 10/4 | 0.04 | 0.04 | 0.01 | 0.17 | 0.16 | 0.19 | 0,00 | 0.05 | 0.09 |
| 10/5 | 0.01 | 0.03 | 0,00 | 0.03 | 0.02 | 0.15 | 0.02 | 0,00 | 0,00 |
| 10/6 | 0.01 | 0.02 | 0,00 | 0.02 | 0.16 | 0.19 | 0,00 | 0.03 | 0,00 |
| 10/7 | 0.01 | 0,00 | 0.1 | 0.25 | 0.59 | 0.73 | 0,00 | 0.03 | 0,00 |
| 10/8 | 0.01 | 0,00 | 0,00 | 0.12 | 0.09 | 0.39 | 0.02 | 0.19 | 0.02 |
| 10/9 | 0,00 | 0,00 | 0.01 | 0.02 | 0.21 | 0.15 | 0.03 | 0,00 | 0.01 |
| 10/10 | 0,00 | 0,00 | 0,00 | 0.01 | 0.01 | 0.07 | 0,00 | 0,00 | 0,00 |
| 10/11 | 0.07 | 0.01 | 0,00 | 0.01 | 0.01 | 0.04 | 0,00 | 0,00 | 0,00 |
| 10/12 | 0.03 | 0.18 | 0.02 | 0.21 | 0.04 | 0.42 | 0.02 | 0.01 | 0,00 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10/13 | 0.01 | 0.01 | 0.01 | 0.06 | 0.08 | 0.37 | 0.01 | 0.04 | 0,00 |
| 10/14 | 0,00 | 0,00 | 0.01 | 0.01 | 0.03 | 0.02 | 0,00 | 0,00 | 0,00 |
| 10/15 | 0.01 | 0.01 | 0.01 | 0.05 | 0.16 | 0.02 | 0,00 | 0.01 | 0,00 |
| 10/16 | 0.08 | 0.04 | 0.4 | 0.35 | 0.29 | 0.1 | 0.1 | 0.15 | 0,00 |
| 10/17 | 0.11 | 0.37 | 0.09 | 0.39 | 0.47 | 0.45 | 0.25 | 0.18 | 0.08 |
| 10/18 | 0.06 | 0.01 | 0.05 | 0.04 | 0.02 | 0.02 | 0.02 | 0.04 | 0.06 |
| 10/19 | 0.01 | 0.04 | 0.01 | 0.16 | 0.45 | 0.14 | 0.2 | 0.09 | 0.02 |
| 10/20 | 0.04 | 0.4 | 0.04 | 0.2 | 0.74 | 0.81 | 0.06 | 0.38 | 0.07 |
| 10/21 | 0.07 | 0.03 | 0.06 | 0.07 | 0.4 | 0.64 | 0.09 | 0.07 | 0.09 |
| 10/22 | 0.01 | 0.02 | 0.01 | 0.08 | 0.45 | 0.51 | 0.02 | 0.1 | 0.01 |
| 10/23 | 0.03 | 0.07 | 0.01 | 0.14 | 0.44 | 0.46 | 0.02 | 0.04 | 0.02 |
| 10/24 | 0.04 | 0.06 | 0.09 | 0.17 | 0.56 | 0.49 | 0.02 | 0.03 | 0.02 |
| 10/25 | 0.03 | 0.02 | 0.02 | 0.22 | 0.14 | 0.35 | 0.01 | 0.04 | 0.01 |
| 10/26 | 0.03 | 0.04 | 0.09 | 0.17 | 0.26 | 0.64 | 0.04 | 0.1 | 0,00 |
| 10/27 | 0.01 | 0.01 | 0,00 | 0.01 | 0.06 | 0.07 | 0.01 | 0,00 | 0,00 |
| 10/28 | 0.01 | 0.02 | 0.04 | 0.19 | 0.53 | 0.13 | 0.01 | 0.04 | 0.01 |
| 10/29 | 0.03 | 0.59 | 0.68 | 0.33 | 0.79 | 0.74 | 0.28 | 0.35 | 0.18 |
| 10/30 | 0.03 | 0.07 | 0.04 | 0.04 | 0.19 | 0.09 | 0.02 | 0.03 | 0.05 |
| 10/31 | 0.02 | 0.01 | 0.02 | 0.02 | 0.42 | 0.44 | 0.02 | 0.11 | 0.03 |
| 11/1 | 0.12 | 0.34 | 0.12 | 0.09 | 0.22 | 0.13 | 0.45 | 0.03 | 0.05 |
| 11/2 | 0.06 | 0.33 | 0.05 | 0.4 | 0.82 | 0.71 | 0.1 | 0.3 | 0.07 |
| 11/3 | 0.03 | 0.24 | 0.02 | 0.17 | 0.38 | 0.78 | 0.06 | 0.03 | 0.03 |
| 11/4 | 0.14 | 0.06 | 0.08 | 0.25 | 0.42 | 0.55 | 0.09 | 0.03 | 0.03 |
| 11/5 | 0.09 | 0.17 | 0.14 | 0.13 | 0.62 | 0.4 | 0.09 | 0.16 | 0.15 |
| 11/6 | 0.04 | 0.06 | 0.11 | 0.05 | 0.12 | 0.15 | 0.05 | 0.02 | 0.07 |
| 11/7 | 0.05 | 0.08 | 0.25 | 0.08 | 0.72 | 0.51 | 0.13 | 0.14 | 0.13 |
| 11/8 | 0.03 | 0.06 | 0.08 | 0.08 | 0.45 | 0.32 | 0.09 | 0.02 | 0.04 |
| 11/9 | 0.13 | 0.03 | 0.04 | 0.13 | 0.19 | 0.08 | 0.35 | 0.06 | 0.21 |
| 11/10 | 0.03 | 0,00 | 0.03 | 0.06 | 0.03 | 0.01 | 0.03 | 0.01 | 0.1 |
| 11/11 | 0.01 | 0,00 | 0,00 | 0,00 | 0.05 | 0.07 | 0.01 | 0.06 | 0.02 |
| 11/12 | 0.04 | 0,00 | 0.07 | 0.02 | 0.18 | 0.09 | 0.03 | 0.05 | 0.13 |
| 11/13 | 0.04 | 0.05 | 0.03 | 0.01 | 0.02 | 0.03 | 0.01 | 0.02 | 0.05 |
| 11/14 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.03 | 0.01 | 0.01 | 0,00 |
| 11/15 | 0,00 | 0,00 | 0,00 | 0.01 | 0.01 | 0.08 | 0,00 | 0.01 | 0,00 |
| 11/16 | 0,00 | 0,00 | 0.01 | 0.23 | 0.22 | 0.22 | 0.01 | 0.22 | 0,00 |
| 11/17 | 0.06 | 0.03 | 0.02 | 0.1 | 0.05 | 0.07 | 0.02 | 0.01 | 0,00 |
| 11/18 | 0.32 | 0.1 | 0.22 | 0.06 | 0.09 | 0.03 | 0.39 | 0.05 | 0.06 |
| 11/19 | 0.14 | 0,00 | 0.11 | 0.08 | 0.06 | 0.05 | 0.02 | 0.52 | 0.01 |
| 11/20 | 0,00 | 0,00 | 0.01 | 0.16 | 0.08 | 0.12 | 0.01 | 0.01 | 0,00 |
| 11/21 | 0.09 | 0.03 | 0.17 | 0.01 | 0.02 | 0.07 | 0.12 | 0.01 | 0,00 |
| 11/22 | 0.15 | 0.06 | 0.11 | 0.12 | 0.04 | 0.05 | 0.11 | 0.02 | 0.06 |
| 11/23 | 0.01 | 0.01 | 0.01 | 0.02 | 0.03 | 0.01 | 0,00 | 0.02 | 0,00 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 11/24 | 0.03 | 0.03 | 0.02 | 0.09 | 0.1 | 0.57 | 0.01 | 0.04 | 0.01 |
| 11/25 | 0.02 | 0,00 | 0.55 | 0.03 | 0.01 | 0.54 | 0.08 | 0.01 | 0.16 |
| 11/26 | 0.03 | 0.02 | 0.09 | 0.01 | 0.02 | 0.04 | 0.01 | 0,00 | 0.03 |
| 11/27 | 0.29 | 0.13 | 0.15 | 0.11 | 0.12 | 0.1 | 0.03 | 0.11 | 0,00 |
| 11/28 | 0.08 | 0.05 | 0.03 | 0.11 | 0.1 | 0,00 | 0.09 | 0.01 | 0.01 |
| 11/29 | 0.01 | 0,00 | 0.03 | 0.07 | 0.24 | 0.05 | 0.01 | 0.07 | 0.01 |
| 11/30 | 0.02 | 0,00 | 0.37 | 0.02 | 0.02 | 0.16 | 0.01 | 0.13 | 0.04 |
| 12/1 | 0.03 | 0.05 | 0.03 | 0.04 | 0.13 | 0.16 | 0.05 | 0.01 | 0.03 |
| 12/2 | 0.05 | 0.02 | 0.02 | 0.05 | 0.47 | 0.04 | 0.01 | 0.01 | 0.01 |
| 12/3 | 0.01 | 0.12 | 0.08 | 0.21 | 0.15 | 0.07 | 0.04 | 0.07 | 0.16 |
| 12/4 | 0.01 | 0,00 | 0.37 | 0.03 | 0.33 | 0.02 | 0.02 | 0.24 | 0.01 |
| 12/5 | 0.01 | 0.01 | 0.02 | 0.04 | 0.02 | 0.01 | 0.01 | 0,00 | 0.01 |
| 12/6 | 0.06 | 0.03 | 0.01 | 0.21 | 0.08 | 0.01 | 0.16 | 0.01 | 0.01 |
| 12/7 | 0.01 | 0.12 | 0.48 | 0.17 | 0.39 | 0.28 | 0.1 | 0.05 | 0.29 |
| 12/8 | 0.05 | 0.05 | 0.2 | 0.2 | 0.41 | 0.19 | 0.05 | 0.25 | 0.04 |
| 12/9 | 0.07 | 0.06 | 0.03 | 0.02 | 0.07 | 0.06 | 0.05 | 0.04 | 0,00 |
| 12/10 | 0.02 | 0.01 | 0.23 | 0.11 | 0.21 | 0.05 | 0.02 | 0.02 | 0.16 |
| 12/11 | 0.13 | 0.06 | 0.05 | 0.29 | 0.38 | 0.61 | 0.05 | 0.1 | 0.01 |
| 12/12 | 0.07 | 0.18 | 0.05 | 0.11 | 0.12 | 0.2 | 0.07 | 0.03 | 0.02 |
| 12/13 | 0.23 | 0.14 | 0.13 | 0.13 | 0.03 | 0.26 | 0.21 | 0.1 | 0.05 |
| 12/14 | 0.14 | 0.03 | 0.29 | 0.3 | 0.47 | 0.57 | 0.1 | 0.08 | 0.03 |
| 12/15 | 0.25 | 0.09 | 0.25 | 0.16 | 0.69 | 0.84 | 0.3 | 0.34 | 0.47 |
| 12/16 | 0.12 | 0.18 | 0.08 | 0.14 | 0.52 | 0.78 | 0.18 | 0.07 | 0.4 |
| 12/17 | 0.09 | 0.1 | 0.11 | 0.08 | 0.46 | 0.63 | 0.1 | 0.11 | 0.08 |
| 12/18 | 0.06 | 0.05 | 0.04 | 0.18 | 0.21 | 0.8 | 0.06 | 0.04 | 0.02 |
| 12/19 | 0.18 | 0.16 | 0.15 | 0.12 | 0.44 | 0.87 | 0.19 | 0.19 | 0.18 |
| 12/20 | 0.17 | 0.3 | 0.07 | 0.28 | 0.39 | 0.92 | 0.28 | 0.04 | 0.26 |
| 12/21 | 0.04 | 0.13 | 0.17 | 0.31 | 0.61 | 0.69 | 0.15 | 0.27 | 0.28 |
| 12/22 | 0.14 | 0.34 | 0.14 | 0.15 | 0.61 | 0.77 | 0.45 | 0.42 | 0.05 |
| 12/23 | 0.21 | 0.23 | 0.23 | 0.13 | 0.46 | 0.49 | 0.21 | 0.39 | 0.14 |
| 12/24 | 0.32 | 0.12 | 0.08 | 0.26 | 0.48 | 0.46 | 0.2 | 0.08 | 0.06 |
| 12/25 | 0.09 | 0.13 | 0.08 | 0.16 | 0.4 | 0.33 | 0.15 | 0.16 | 0.02 |
| 12/26 | 0.1 | 0.04 | 0.09 | 0.18 | 0.55 | 0.57 | 0.11 | 0.11 | 0.09 |
| 12/27 | 0.06 | 0.04 | 0.01 | 0.04 | 0.33 | 0.48 | 0.03 | 0.02 | 0.01 |
| 12/28 | 0.07 | 0.02 | 0.12 | 0.12 | 0.27 | 0.21 | 0.03 | 0.02 | 0.07 |
| 12/29 | 0.38 | 0.14 | 0.36 | 0.55 | 0.71 | 0.19 | 0.15 | 0.26 | 0.53 |
| 12/30 | 0.4 | 0.32 | 0.27 | 0.44 | 0.23 | 0.29 | 0.04 | 0.14 | 0.29 |
| 12/31 | 0.02 | 0.01 | 0.01 | 0.01 | 0.04 | 0.4 | 0.1 | 0.01 | 0,00 |

**Table 3.** Predictions (2019->2020): probability values.

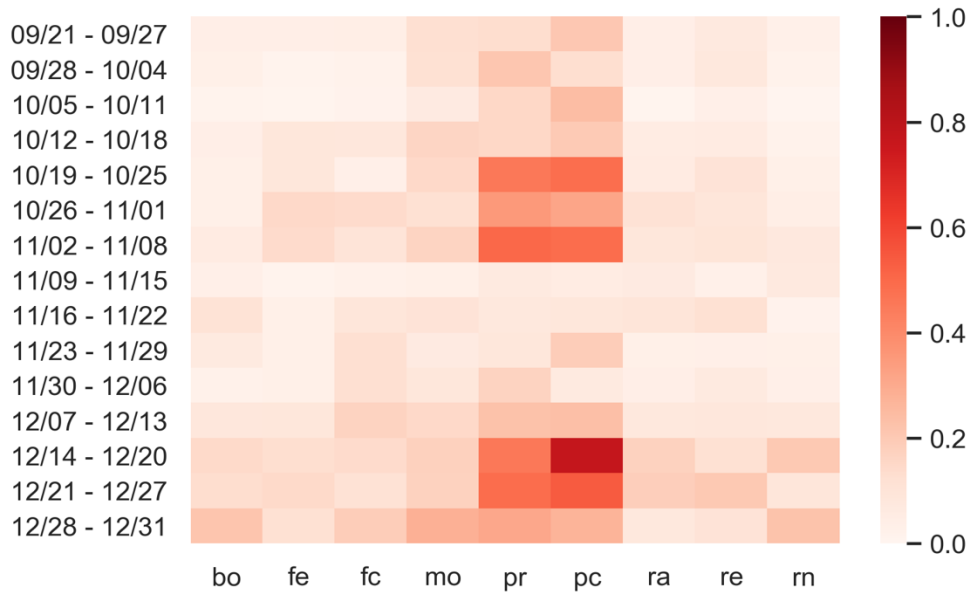


**Figure 6.** Predictions (2019->2020): *heatmap*.

*3.2. Predictions: 2017–2019 -> 2020*

In this section, we report the predictions our GB model has made in the case that the pollutants, at the basis of the relationship that we assume them to have, are considered as a mixture obtained with an average of the three previous years, more specifically, 2017, 2018, and 2019, and finally provided as input to the model. As before, the predictions are given in Table 4 on both a daily and a per-province basis.

All the considerations we have already anticipated in the previous subsections are all still valid, including the one on the personal protection measures.

It is worth noting here that extending to the three previous years (2017–2019) has not brought to more positive predictions. This can be due to several factors, including the fact that 2019 may have been a quite favorable year, in terms of registered pollution. In any case, the general trend of our predictions comes confirmed. Two provinces, in particular, Parma and Piacenza, seem to run larger risks in terms of the number of infections that exceed the threshold of 17. This is clearly visible both in the probability values reported in Table 4 and in the heatmap of Figure 7. Again, we have input evidence that during the observed period, as long as 135 days, the numbers of crucial days, on a per-province basis, are as follows:

Bologna (1);
Ferrara (0);
Forlì-Cesena (0);
Modena (0);
Parma (29);
Piacenza (43);
Ravenna (0);
Reggio Emilia (1);
Rimini (0).

| Day | Bo | Fe | Fc | Mo | Pr | Pc | Ra | Re | Rn |
|---|---|---|---|---|---|---|---|---|---|
| 9/21 | 0,00 | 0.01 | 0,00 | 0.02 | 0.02 | 0.33 | 0,00 | 0.12 | 0,00 |
| 9/22 | 0,00 | 0.03 | 0,00 | 0.12 | 0.05 | 0.12 | 0.02 | 0.04 | 0,00 |
| 9/23 | 0.11 | 0.02 | 0.07 | 0.08 | 0.09 | 0.08 | 0.03 | 0.07 | 0.05 |
| 9/24 | 0.04 | 0.01 | 0.15 | 0.1 | 0.09 | 0.08 | 0.01 | 0.02 | 0.1 |
| 9/25 | 0,00 | 0.02 | 0,00 | 0.06 | 0.06 | 0.06 | 0.01 | 0.02 | 0,00 |
| 9/26 | 0.04 | 0.01 | 0,00 | 0.11 | 0.08 | 0.05 | 0.01 | 0.01 | 0,00 |
| 9/27 | 0.01 | 0.01 | 0.01 | 0.05 | 0.05 | 0.03 | 0.03 | 0.04 | 0.05 |
| 9/28 | 0,00 | 0.03 | 0,00 | 0.12 | 0.04 | 0.09 | 0.03 | 0.06 | 0,00 |
| 9/29 | 0.01 | 0.07 | 0,00 | 0.15 | 0.06 | 0.03 | 0.02 | 0.04 | 0,00 |
| 9/30 | 0.02 | 0.01 | 0.01 | 0.03 | 0.01 | 0.13 | 0,00 | 0.01 | 0,00 |
| 10/1 | 0,00 | 0.01 | 0,00 | 0.26 | 0.05 | 0.07 | 0.04 | 0.04 | 0,00 |
| 10/2 | 0.01 | 0.09 | 0.1 | 0.08 | 0.09 | 0.14 | 0.02 | 0.04 | 0,00 |
| 10/3 | 0.07 | 0.02 | 0.06 | 0.1 | 0.15 | 0.19 | 0.01 | 0.07 | 0.01 |
| 10/4 | 0.14 | 0.03 | 0.11 | 0.22 | 0.24 | 0.15 | 0.02 | 0.03 | 0.12 |
| 10/5 | 0.01 | 0.04 | 0.01 | 0.11 | 0.05 | 0.19 | 0.05 | 0.01 | 0.01 |
| 10/6 | 0,00 | 0.06 | 0,00 | 0.06 | 0.08 | 0.05 | 0,00 | 0.02 | 0,00 |
| 10/7 | 0,00 | 0,00 | 0.02 | 0.12 | 0.4 | 0.12 | 0.01 | 0.09 | 0,00 |
| 10/8 | 0.07 | 0.04 | 0.13 | 0.06 | 0.24 | 0.23 | 0.04 | 0.06 | 0.02 |
| 10/9 | 0.1 | 0.07 | 0.11 | 0.09 | 0.22 | 0.3 | 0.08 | 0.12 | 0.05 |
| 10/10 | 0.03 | 0.07 | 0.01 | 0.22 | 0.19 | 0.19 | 0.03 | 0.06 | 0.01 |
| 10/11 | 0.01 | 0.04 | 0.01 | 0.13 | 0.32 | 0.18 | 0.05 | 0.07 | 0.01 |
| 10/12 | 0.01 | 0.13 | 0.15 | 0.36 | 0.28 | 0.21 | 0.04 | 0.09 | 0.02 |
| 10/13 | 0,00 | 0.01 | 0.03 | 0.04 | 0.26 | 0.04 | 0.05 | 0.07 | 0,00 |
| 10/14 | 0.03 | 0.04 | 0.08 | 0.06 | 0.21 | 0.52 | 0.12 | 0.08 | 0,00 |
| 10/15 | 0.04 | 0.11 | 0.04 | 0.16 | 0.44 | 0.47 | 0.05 | 0.06 | 0.06 |
| 10/16 | 0.07 | 0.08 | 0.23 | 0.37 | 0.51 | 0.34 | 0.05 | 0.12 | 0.04 |
| 10/17 | 0.16 | 0.09 | 0.12 | 0.19 | 0.41 | 0.28 | 0.09 | 0.23 | 0.04 |
| 10/18 | 0.02 | 0.21 | 0.04 | 0.06 | 0.14 | 0.43 | 0.15 | 0.05 | 0.07 |
| 10/19 | 0.03 | 0.47 | 0.05 | 0.22 | 0.67 | 0.35 | 0.11 | 0.6 | 0.23 |
| 10/20 | 0.11 | 0.11 | 0.13 | 0.16 | 0.56 | 0.66 | 0.1 | 0.24 | 0.15 |
| 10/21 | 0.06 | 0.16 | 0.05 | 0.05 | 0.28 | 0.86 | 0.07 | 0.07 | 0.03 |
| 10/22 | 0.05 | 0.09 | 0.13 | 0.1 | 0.38 | 0.82 | 0.07 | 0.07 | 0.03 |
| 10/23 | 0.05 | 0.12 | 0.06 | 0.19 | 0.62 | 0.81 | 0.03 | 0.18 | 0.02 |
| 10/24 | 0.12 | 0.21 | 0.21 | 0.2 | 0.53 | 0.9 | 0.22 | 0.25 | 0.03 |
| 10/25 | 0.34 | 0.27 | 0.11 | 0.18 | 0.36 | 0.92 | 0.1 | 0.19 | 0.07 |
| 10/26 | 0.18 | 0.11 | 0.04 | 0.18 | 0.49 | 0.81 | 0.08 | 0.19 | 0.05 |
| 10/27 | 0.16 | 0.1 | 0.4 | 0.17 | 0.62 | 0.86 | 0.23 | 0.16 | 0.11 |
| 10/28 | 0.27 | 0.18 | 0.26 | 0.15 | 0.59 | 0.85 | 0.28 | 0.22 | 0.19 |
| 10/29 | 0.17 | 0.3 | 0.05 | 0.19 | 0.84 | 0.82 | 0.05 | 0.26 | 0.04 |
| 10/30 | 0.02 | 0.08 | 0.07 | 0.18 | 0.77 | 0.6 | 0.06 | 0.11 | 0.02 |
| 10/31 | 0.04 | 0.27 | 0.25 | 0.13 | 0.4 | 0.64 | 0.14 | 0.12 | 0.01 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 11/1 | 0.04 | 0.14 | 0.11 | 0.23 | 0.68 | 0.91 | 0.09 | 0.12 | 0.02 |
| 11/2 | 0.09 | 0.06 | 0.15 | 0.18 | 0.43 | 0.64 | 0.15 | 0.1 | 0.04 |
| 11/3 | 0.07 | 0.47 | 0.14 | 0.2 | 0.33 | 0.7 | 0.25 | 0.21 | 0.04 |
| 11/4 | 0.23 | 0.16 | 0.21 | 0.38 | 0.64 | 0.84 | 0.2 | 0.33 | 0.05 |
| 11/5 | 0.17 | 0.17 | 0.06 | 0.15 | 0.53 | 0.48 | 0.25 | 0.13 | 0.02 |
| 11/6 | 0.03 | 0.02 | 0.04 | 0.15 | 0.62 | 0.35 | 0.04 | 0.03 | 0.01 |
| 11/7 | 0.05 | 0.05 | 0.03 | 0.02 | 0.3 | 0.43 | 0.03 | 0.03 | 0.03 |
| 11/8 | 0.03 | 0.05 | 0.06 | 0.11 | 0.42 | 0.27 | 0.01 | 0.08 | 0.02 |
| 11/9 | 0.02 | 0.07 | 0.03 | 0.28 | 0.68 | 0.23 | 0.02 | 0.05 | 0.03 |
| 11/10 | 0.07 | 0.06 | 0.03 | 0.07 | 0.41 | 0.18 | 0.1 | 0.12 | 0.07 |
| 11/11 | 0.04 | 0.01 | 0.17 | 0.15 | 0.49 | 0.22 | 0.04 | 0.14 | 0.01 |
| 11/12 | 0.06 | 0.09 | 0.05 | 0.28 | 0.67 | 0.07 | 0.04 | 0.18 | 0.05 |
| 11/13 | 0.02 | 0.02 | 0.02 | 0.01 | 0.14 | 0.1 | 0.04 | 0.01 | 0.02 |
| 11/14 | 0.03 | 0.01 | 0.03 | 0.02 | 0.13 | 0.15 | 0.01 | 0.02 | 0.05 |
| 11/15 | 0.01 | 0.02 | 0.01 | 0.08 | 0.47 | 0.1 | 0.02 | 0.05 | 0.05 |
| 11/16 | 0.08 | 0,00 | 0.01 | 0.09 | 0.41 | 0.03 | 0.06 | 0.03 | 0,00 |
| 11/17 | 0.01 | 0.01 | 0.02 | 0.03 | 0.09 | 0.05 | 0.02 | 0.03 | 0.01 |
| 11/18 | 0.01 | 0.09 | 0.03 | 0.16 | 0.22 | 0.03 | 0.02 | 0.05 | 0.01 |
| 11/19 | 0.01 | 0.08 | 0.02 | 0.32 | 0.52 | 0.22 | 0.13 | 0.18 | 0.01 |
| 11/20 | 0.03 | 0.04 | 0.16 | 0.04 | 0.17 | 0.05 | 0.07 | 0.04 | 0.04 |
| 11/21 | 0.05 | 0.02 | 0.08 | 0.21 | 0.44 | 0.18 | 0.05 | 0.01 | 0.09 |
| 11/22 | 0.03 | 0.01 | 0.03 | 0.09 | 0.13 | 0.04 | 0.03 | 0.02 | 0.04 |
| 11/23 | 0.04 | 0.04 | 0.02 | 0.08 | 0.26 | 0.1 | 0.02 | 0.05 | 0.01 |
| 11/24 | 0.01 | 0.04 | 0.03 | 0.16 | 0.4 | 0.13 | 0.02 | 0.07 | 0,00 |
| 11/25 | 0.07 | 0.06 | 0.03 | 0.16 | 0.4 | 0.26 | 0.09 | 0.05 | 0.05 |
| 11/26 | 0.14 | 0.02 | 0.08 | 0.22 | 0.52 | 0.21 | 0.13 | 0.07 | 0.08 |
| 11/27 | 0.05 | 0.06 | 0.04 | 0.15 | 0.15 | 0.19 | 0.06 | 0.04 | 0.05 |
| 11/28 | 0.22 | 0.28 | 0.03 | 0.24 | 0.46 | 0.29 | 0.25 | 0.08 | 0.07 |
| 11/29 | 0.12 | 0.04 | 0.14 | 0.23 | 0.31 | 0.51 | 0.15 | 0.22 | 0.11 |
| 11/30 | 0.48 | 0.15 | 0.15 | 0.13 | 0.19 | 0.38 | 0.23 | 0.11 | 0.15 |
| 12/1 | 0.03 | 0.07 | 0.09 | 0.11 | 0.39 | 0.16 | 0.12 | 0.05 | 0.12 |
| 12/2 | 0.05 | 0.09 | 0.03 | 0.09 | 0.52 | 0.63 | 0.07 | 0.1 | 0.02 |
| 12/3 | 0.04 | 0.04 | 0.03 | 0.04 | 0.32 | 0.67 | 0.03 | 0.05 | 0.04 |
| 12/4 | 0.04 | 0.08 | 0.02 | 0.09 | 0.38 | 0.32 | 0.07 | 0.04 | 0.02 |
| 12/5 | 0.14 | 0.08 | 0.06 | 0.27 | 0.76 | 0.66 | 0.13 | 0.04 | 0.01 |
| 12/6 | 0.14 | 0.08 | 0.22 | 0.1 | 0.38 | 0.5 | 0.12 | 0.06 | 0.07 |
| 12/7 | 0.01 | 0.36 | 0.35 | 0.25 | 0.66 | 0.45 | 0.25 | 0.17 | 0.3 |
| 12/8 | 0.03 | 0.32 | 0.19 | 0.38 | 0.38 | 0.48 | 0.2 | 0.08 | 0.03 |
| 12/9 | 0.15 | 0.22 | 0.03 | 0.13 | 0.35 | 0.55 | 0.2 | 0.08 | 0.01 |
| 12/10 | 0.12 | 0.12 | 0.17 | 0.1 | 0.45 | 0.56 | 0.09 | 0.05 | 0.34 |
| 12/11 | 0.24 | 0.08 | 0.11 | 0.09 | 0.36 | 0.73 | 0.08 | 0.08 | 0.06 |
| 12/12 | 0.1 | 0.06 | 0.04 | 0.18 | 0.22 | 0.77 | 0.09 | 0.05 | 0.03 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 12/13 | 0.23 | 0.19 | 0.25 | 0.19 | 0.53 | 0.9 | 0.2 | 0.22 | 0.09 |
| 12/14 | 0.09 | 0.12 | 0.18 | 0.22 | 0.39 | 0.79 | 0.22 | 0.19 | 0.12 |
| 12/15 | 0.15 | 0.27 | 0.18 | 0.24 | 0.72 | 0.84 | 0.23 | 0.26 | 0.07 |
| 12/16 | 0.07 | 0.38 | 0.1 | 0.29 | 0.73 | 0.85 | 0.29 | 0.13 | 0.07 |
| 12/17 | 0.05 | 0.18 | 0.21 | 0.12 | 0.36 | 0.82 | 0.16 | 0.09 | 0.12 |
| 12/18 | 0.12 | 0.18 | 0.08 | 0.26 | 0.53 | 0.86 | 0.18 | 0.18 | 0.05 |
| 12/19 | 0.16 | 0.36 | 0.42 | 0.38 | 0.55 | 0.92 | 0.24 | 0.33 | 0.06 |
| 12/20 | 0.13 | 0.05 | 0.18 | 0.13 | 0.42 | 0.71 | 0.24 | 0.15 | 0.07 |
| 12/21 | 0.18 | 0.28 | 0.33 | 0.2 | 0.51 | 0.71 | 0.27 | 0.34 | 0.17 |
| 12/22 | 0.06 | 0.48 | 0.15 | 0.23 | 0.37 | 0.44 | 0.4 | 0.17 | 0,00 |
| 12/23 | 0.59 | 0.08 | 0.41 | 0.13 | 0.57 | 0.5 | 0.25 | 0.19 | 0.04 |
| 12/24 | 0.12 | 0.14 | 0.14 | 0.08 | 0.43 | 0.66 | 0.1 | 0.08 | 0.15 |
| 12/25 | 0.07 | 0.11 | 0.08 | 0.11 | 0.24 | 0.48 | 0.14 | 0.12 | 0.06 |
| 12/26 | 0.09 | 0.23 | 0.1 | 0.18 | 0.52 | 0.8 | 0.1 | 0.08 | 0.16 |
| 12/27 | 0.1 | 0.1 | 0.11 | 0.18 | 0.49 | 0.82 | 0.11 | 0.16 | 0.2 |
| 12/28 | 0.05 | 0.15 | 0.03 | 0.21 | 0.49 | 0.63 | 0.1 | 0.12 | 0.02 |
| 12/29 | 0.09 | 0.22 | 0.13 | 0.19 | 0.45 | 0.86 | 0.18 | 0.15 | 0.04 |
| 12/30 | 0.27 | 0.22 | 0.19 | 0.13 | 0.27 | 0.7 | 0.15 | 0.12 | 0.1 |
| 12/31 | 0.12 | 0.24 | 0.05 | 0.12 | 0.55 | 0.5 | 0.12 | 0.04 | 0.06 |

**Table 4.** Predictions (2017-2018-2019->2020): probability values.

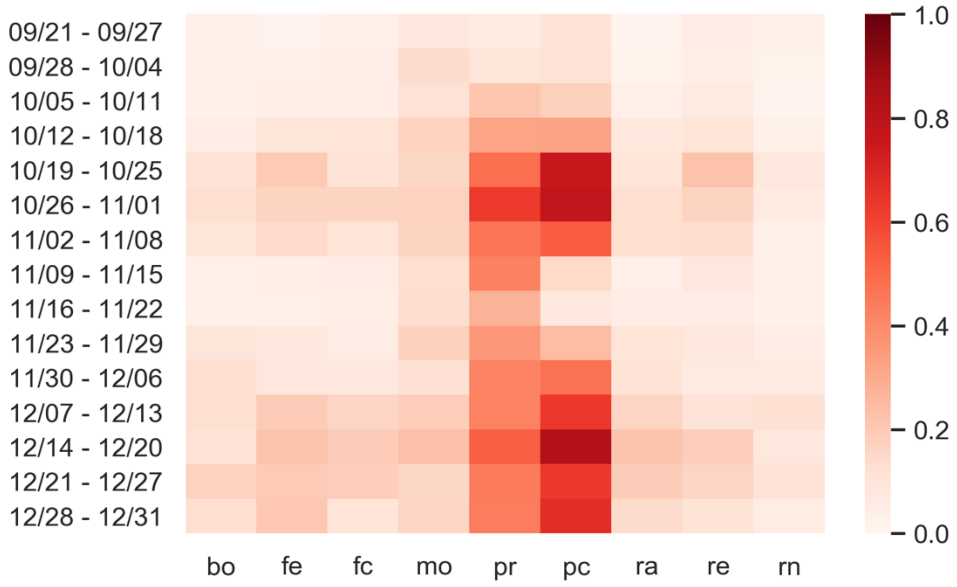


**Figure 7.** Predictions (2017-2018-2019->2020): heatmap.

*3.3. Predictions: What Happens if Personal Protection Measures Are not Respected?*

In this final section, we report on the predictions our GB model has made in the case that the personal protection measures indicated by the Italian Government are not respected.

In some sense, this is a special kind of sensitivity analysis where we have varied the unique model parameter that can be significantly touched (i.e., the personal protection measures).

In this specific case, we present the predictions, returned by our model, only under the form of heatmaps, where again, exactly like before, a lot of red color in the map corresponds to a very likely occurrence of a resurgence of the virus, on that week, in that province.

Once again, we have provided two separate sets of predictions and two correspondent heatmaps. The first one is that where we have used only the pollution measured in the year 2019 (Figure 8), while the second one averages the pollutants over three different years, 2017–2019 (Figure 9).

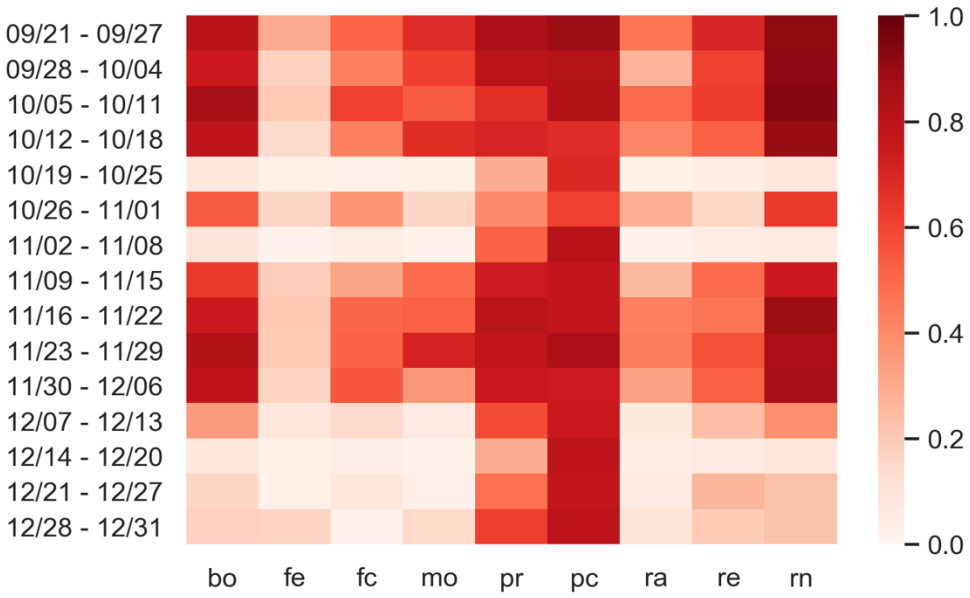


**Figure 8.** Predictions: no personal protection measure (2019->2020).

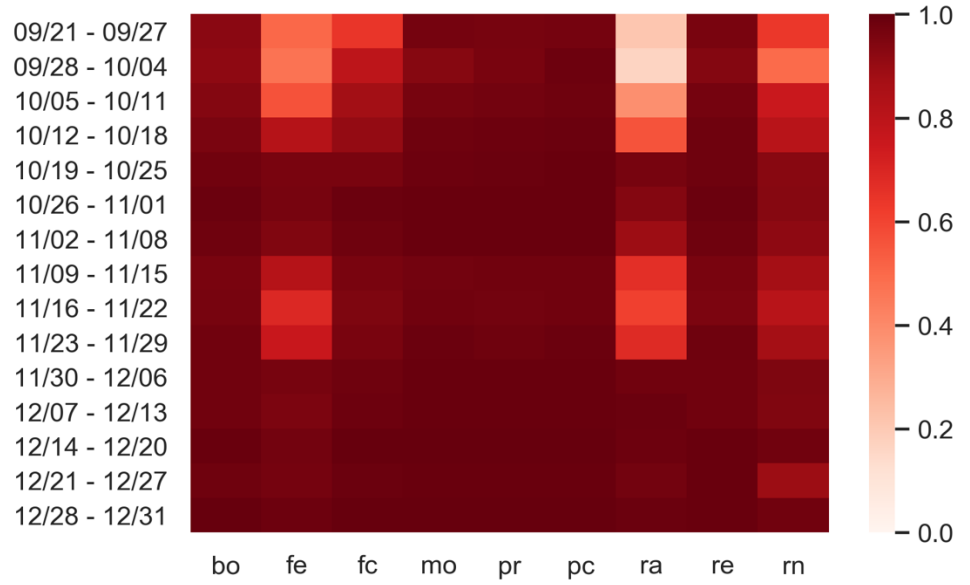


**Figure 9.** Predictions: no personal protection measure (2017-2018-2019->2020).

We believe that no other comment is needed here, as the plausibility of a resurgence of the virus is highly evident under the given circumstances. We could only add the consideration that one could recognize that the current rising infective trend of August 2020 could be just the trigger of a new virus explosion that those heatmaps clearly display.

## 4. Discussion and Conclusions

We have developed a scientific study that aims at making predictions on a possible resurgence of a COVID-19 incidence in the Italian region of Emilia-Romagna (which was one of the most hardly hit during the first phase of contagion in the period of February–April 2020).

We have based our study on a precise, given hypothesis, the most important being that of a correlation existing between the presence of circulating pollutants in the air, such as $PM_{2.5}$, $PM_{10}$, and $NO_2$, and the number of infected people. Believing in the existence of this relationship, data were collected, on a daily basis, for a period as long as mid-February 2020–end of July 2020.

These data amounted both to the measurements of the values of the aforementioned pollutants, as well as to the registered number of infections. This was carried out for all the nine provinces comprising the Italian region of Emilia-Romagna. Not only that, but data which were useful to instruct our predictive models were also represented by the restrictions that were imposed to the region of Emilia-Romagna by the Italian Government, during four different and subsequent phases, which happened during that period.

Upon completion of the data collection activity, we moved on to the selection of a computational model. Among many possible alternatives, we resorted to machine learning (ML) models, suitable for learning the function we believe can be at the basis of our hypothesis. After having conducted a numerical comparative study among several ML models, based on the available data, we found that the gradient boosting (GB) model was the one that fit squarely to the situation under observation, reaching an accuracy of almost 90% in a preliminary testing phase.

With that model, we then moved to the predictions, considering as possible estimates of the pollution than could happen in the future period of 21 September–31 December 2019 the values of the pollutants measured in previous years, namely, 2017–2019 (for the same temporal period of interest).

Relevant is also the consideration that the predictions were made by inputting to our model the situation demarcated by the measures decided on Phase 3 by the Italian Government.

At the end of all this long process, we have got our predictions provided under the form of a probability value. In essence, our model predicts the probability of surpassing a threshold of infected people in a given province, and on a certain day. Based on those probability values, we finally depicted heatmaps that could better give a general picture of the possible COVID-19 resurgence in the region of interest.

To summarize these results, the risk of a very strong second wave of COVID-19 in Emilia-Romagna seems moderate, even if those predictions also express the concern that at least two single provinces (namely, Parma and Piacenza) could be subjected to a more complex situation.

To conclude the set of our predictions, we also conducted a special kind of sensitivity analysis where our model was run, yet with a variation on the parameter concerning the use of personal protection measures. In such a case (i.e., the personal protection measures are not adopted by people), the situation becomes very different, and a risk of a resurgence of the virus becomes very plausible, for almost all the nine provinces in Emilia-Romagna.

Before we can conclude this paper, we feel the duty to finally discuss possible fallacies and limitations of our investigation regarding, at least, the three following points: (i) the scientific methodology we adopted, (ii) the choice of the data, the adopted model, and the decisional threshold of 17 infections, and finally (iii) the extensibility of the model to different COVID-19 situations.

As far as the scientific method is concerned, we have already discussed, at length, in a previous paper on the hypothesis of the existence of a correlation between pollutants and infections in our region [6]. We do not have any intention to retrace here the entire scientific path that led us to believe to this hypothesis. It suffices here to remind that we have already subjected that assumption to a statistical testing procedure (i.e., a Granger causality testing), whose results were essentially confirmative. Moreover, we know very well that, while this scientific issue is still at the center of a controversy [46], it is also true that various papers (and numerous researchers) have claimed that this virus can be airborne, and that particulate matter may further favor an airborne route, as various already cited papers have confirmed.

To move on to the second point, we would like to discuss, first, the issue of the employed data. We understand the reasons valued researchers have decided to resort to multiple sources of data to be used as early indicators of a second wave of the virus (see [20], for example). Nonetheless, we, as experienced data scientists, believe in the actual validity of data only when they are accompanied by a well-defined hypothesis. This always brings to a positive result. If experimental data provide confirmative results, in fact, one gets a kind of confirmation that can also be extended, by some measure, to the theory in general; otherwise, the hypothesis needs to be rejected, or at least revised. On the other side, with a lot of data, yet without a working hypothesis, one could also get good/bad results in some circumstances, but they would ignore what the real motivations are behind that success/failure. This justifies our approach as to the choice of our data.

As for the computational model, it should be clear that with our work, we do not want to refuse to acknowledge the importance of more traditional predictive methodologies, such as SIR, for example. They are well-founded epidemiologic models whose validity is out of discussion [47]. Nonetheless, the incidence of a quite unknown virus, like COVID-19, has put all of us into the difficult position of dealing with new alternatives. From this point of view, we are confident that ML models can provide great help, provided that they are used by experts, who are perfectly aware of all the implications they carry [48].

The issue regarding the threshold value of 17 infections may be the source of much controversy. Nevertheless, first, we would like to work with a parameter that was both simple to calculate and also a clear direct indicator of how many people got infected on a daily per-province basis. Following this reasoning, one could suggest working with a separate prediction model for each province, based on the average value of those infections that occurred only in that province. Nonetheless, In Emilia-Romagna, our experience was that we had provinces (such as Ferrara, for example) with a constantly low value of that average, even during the hardest part of the COVID-19 outbreak, while the situation in the region was generally very bad, hence our decision to use a unique value, computed as an average over all the total number of infected people in the region, yet to be applied to each province, as an early indicator. To strengthen this argument, one should consider as generally alarming, and needing to be taken into serious consideration, a situation where many provinces in a region simultaneously reach, or surpass, a given predefined value of infections. Less worrying, by contrast, would be that situation where just a very few of them surpass even a high value of infected people. In this case, more plausible is the occurrence of a local isolated outbreak, whose management is usually easier. In simpler words, this latter situation would not raise any serious concerns about the plausibility of a second wave running over the majority of the region, and over all its provinces.

Finally, allow us to address the extensibility of our model to different COVID-19 scenarios. In some sense, we recognize that this could be one of the main weaknesses of our approach. Just to cite one, for example, the hypothesis that links infections with pollution can be applied just to those geographical areas that mostly suffer from this unpleasant condition. Nevertheless, it is also true that pieces of evidence have begun to emerge that this virus hits harder in those geographical areas where the general climatic and environmental conditions are somewhat complex.

This said, our model could be generalized, provided that our GB algorithm is instructed, validated, and finally tested with the values of the pollutants and the number of infections coming from the region of interest. Not only this: As one of the inputs of the model is also the type of personal protection measures which are adopted (or even enforced) in that given region, this parameter is also needed to allow the model to make the predictions.

Again, one could criticize our study on the basis of the fact that there are multiple possible factors that have led to the devastations brought by the virus in many areas in the world. Nonetheless, we respond to this criticism with the consideration that many traditional studies have been already conducted that have proven to be a very poor proxy for understanding the extension of this contagion. Our investigation, by contrast, is projecting a new scenario based on an original hypothesis that makes our prediction model unique in the world. At the time we write this article, we cannot have a confirmation of the precision of our predictions, but they will be soon confirmed/rejected by history—and this, too, is science at the service of society.

We want to conclude with a final, but important, consideration. All the experiments we have conducted are reproducible using the data available in the public repositories we have mentioned.

**Author Contributions:** Conceptualization, M. Roccetti. and G. Delnevo.; methodology, M. Roccetti. and S. Mirri.; software, G. Delnevo.; validation, M. Roccetti., S. Mirri. and G. Delnevo.; formal analysis, M. Roccetti.; investigation, G. Delnevo.; resources, S. Mirri.; data curation, G. Delnevo.; writing—original draft preparation, M. Roccetti.; writing—review and editing, M. Roccetti.; visualization, G. Delnevo.; supervision, S. Mirri.; project administration, S. Mirri.; funding acquisition, M.Roccetti. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

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

## References

1. Jamieson, A. Coronavirus crisis may get worse and worse and worse, warns WHO. Available online: https://www.independent.co.uk/news/uk/home-news/coronavirus-cases-deaths-who-infection-rate-global-latest-a9616366.html?fbclid=IwAR1rTs52bD1jZBjNEYNt63OuN_DweUkCHlB5oQAAExD2JAR-TXpc5pL2-QA (accessed on 18 July 2020).
2. World Health Organization (WHO). Coronavirus disease (COVID-2019) situation reports. Available online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (accessed on 18 July 2020).
3. Dipartimento della Protezione Civile, Italia. Aggiornamento casi COVID-19. Available online: http://opendatadpc.maps.arcgis.com/apps/opsdashboard/index.html#/b0c68bce2cce478eaac82fe38d4138b1 (accessed on 18 July 2020).
4. Mehta, D. Moody's: Italy's GDP to contract 9.3% in 2020. FXStreet, 2020. Available online: https://www.fxstreet.com/news/moodys-italys-gdp-to-contract-93-in-2020-202004300541 (accessed on 18 July 2020).
5. Carey, B. Can an algorithm predict the pandemic's next moves? The New York Times. 2020. Available online: https://www.nytimes.com/2020/07/02/health/santillana-coronavirus-model-forecast.html?smid=fb-share&fbclid=IwAR15B7tGHRL8oyL1NHgjXyGojTSYbHpoO0ww8hG85B2bN7NVMxJVK2da5wU (accessed on 18 July 2020).
6. Delnevo, G.; Mirri, S.; Roccetti, M. Particulate Matter and COVID-19 disease diffusion in Emilia-Romagna (Italy). Already a cold case? *Computation* **2020**, *8*, 59.
7. Granger, C. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **1969**, *37*, 424–438.
8. Maziarz, M. A review of the Granger-causality fallacy. *J. Philos. Econ.* **2015**, *8*, 86–105.
9. Conticini, E.; Frediani, B.; Caro, D. Can atmospheric pollution be considered a co-factor in extremely high level of SARS-CoV-2 lethality in Northern Italy? *Environ. Pollut.* **2020**, *261*, 114465.
10. Becchetti, L.; Conzo, G.; Conzo, P.; Salustri, F. Understanding the Heterogeneity of Adverse COVID-19 Outcomes: The Role of Poor Quality of Air and Lockdown Decisions. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3572548 (accessed on 18 July 2020).
11. Li, Q.; Guan, X.; Wu, P.; Wang, X.; Zhou, L.; Tong, Y.; Ren, R.; Leung, K.S.M.; Lau, E.H.Y.; Wong, J.Y.; et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New Engl. J. Med.* **2020**, *382*, 1199–1207.
12. Zhou, P.; Yang, X.L.; Wang, X.G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.R.; Zhu, Y.; Li, B.; Huang, C.L.; et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **2020**, *579*, 270–273.
13. Nunez Soza, L.; Jordanova, P.; Nicolis, L.; Strelec, L.; Stehlik, M. Small sample robust approach to outliers and correlation of Atmospheric Pollution and Health Effects in Santiago de Chile. *Chemom. Intell. Lab. Syst.* **2019**, *185*, 73–84.
14. Morawska, L.; Milton, D.K. It is time to address airborne transmission of COVID-19. *Clin. Infect. Dis.* **2020**, doi:10.1093/cid/ciaa939
15. Setti, L.; Passarini, F.; De Gennaro, G.; Barbieri, P.; Perrone, M.G.; Borelli, M.; Palmisani, J.; Di Gilio, A.; Priscitelli, P.; Miani, A. SARS-Cov-2RNA found on particulate matter of Bergamo in Northern Italy: T First evidence. *Environ. Res.* **2020**, *188*, 109754.

16. Ferrarotti, M.J.; Cavalli, A. On the hypothesis of particulate matter as carrier of SARS-CoV-2. Numerical findings with a novel package for Smoluchowski aggregation via direct Monte Carlo. Preprint, n 1, Department of Chemistry, University of Bologna **2020**.
17. Setti, L.; Passarini, F.; De Gennaro, G.; Barbieri, P.; Perrone, M.G.; Borelli, M.; Palmisani, J.; Di Gilio, A.; Priscitelli, P.; Miani, A. Airborne Transmission Route of COVID-19: Why 2 Meters/6 Feet of Inter-Personal Distance Could Not Be Enough. *Int. J. Environ. Res. Public Health* **2020**, *17*, 2932.
18. Setti, L.; Passarini, F.; De Gennaro, G.; Barbieri, P.; Perrone, M.G.; Borelli, M.; Palmisani, J.; Di Gilio, A.; Priscitelli, P.; Miani, A. Searching for SARS-COV-2 on Particulate Matter: A Possible Early Indicator of COVID-19 Epidemic Recurrence. *Int. J. Environ. Res. Public Health* **2020**, *17*, 2986.
19. Gazzetta Ufficiale della Repubblica Italiana. Decreto del Presidente del Consiglio dei Ministri 11 giugno 2020. Available online: https://www.gazzettaufficiale.it/eli/id/2020/06/11/20A03194/sg (accessed on 19 July 2020)
20. Kogan, N.E.; Clemente, L.; Liautaud, P.; Kaashoek, J.; Link, N.B.; Nguyen, A.T.; Lu, F.S.; Huybers, P.; Resch, B.; Havas, C.; et al. An early warning approach to monitor COVID-19 activity with multiple digital traces in near real-time. *arXiv* **2020**, arXiv:2007.00756.
21. Fox, G.N.; Moawad, N.S. UpToDate: A comprehensive clinical database. *J. Family Pract.* **2003**, *52*, 706–710.
22. Buchanan, M. The limits of machine prediction. *Nat. Phys.* **2019**, *15*, 304.
23. Kermack, W.O.; McKendrick, A.G. Contributions to the mathematical theory of epidemics—I. *Bull. Math. Boil.* **1991**, *53*, 33–55.
24. Russell, S.J.; Norvig, P. Artificial Intelligence: A Modern Approach. *Prentice Hall* **2010**.
25. Ardabili, S.F.; Mosavi, A.; Ghamisi, P.; Ferdinand, F.; Varkonyi-Koczy, A.R.; Reuter, U.; Rabczuck, T.; Atkinson, P.M. COVID-19 outbreak prediction with machine learning. *medRxiv preprint* **2020**. Available online: https://www.medrxiv.org/content/10.1101/2020.04.17.20070094v1.full.pdf (accessed on 18 July 2020).
26. Tuli, S.; Tuli, S.; Tuli, R.; Gill, S.S. Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing. *Internet Things* **2020**, *100222*.
27. Liu, Y.; Wang, Z.; Ren, J.; Tian, Y.; Zhou, M.; Zhou, T.; Ye, K.; Zhao, Y.; Qiu, Y.; Li, J. A COVID-19 Risk Assessment Decision Support System for General Practitioners: Design and Development Study. *J. Med. Internet Res.* **2020**, *22*, e19786.
28. Nguyen, T.T. Artificial intelligence in the battle against coronavirus (COVID-19): A survey and future research directions. Preprint 2020. Available online: https://figshare.com/articles/Artificial_Intelligence_in_the_Battle_against_Coronavirus_COVID-19_A_Survey_and_Future_Research_Directions/12127020 (accessed on 18 July 2020).
29. Alimadadi, A.; Aryal, S.; Manandhar, I.; Munroe, P.B.; Joe, B.; Cheng, X. Artificial intelligence and machine learning to fight COVID-19. *Physiol. Genomics* **2020**, *52*, 200–202.
30. Naudé, W. Artificial Intelligence against COVID-19: An Early Review. Available online: https://www.iza.org/publications/dp/13110/artificial-intelligence-against-covid-19-an-early-review (accessed on 18 July 2020).
31. Kucharski, A.J.; Russell, T.W.; Diamond, C.; Liu, Y.; Edmunds, J.; Funk, S.; Eggo, R.M. Early dynamics of transmission and control of COVID-19: A mathematical modelling study. *Lancet Infect. Dis.* **2020**, *20*, 553–558.
32. Gazzetta Ufficiale della Repubblica Italiana. Decreto del Presidente del Consiglio dei Ministri 8 Marzo 2020. Available online: https://www.gazzettaufficiale.it/eli/id/2020/03/08/20A01522/sg (accessed on 18 July 2020).
33. Governo Italiano Presidenza del Consiglio dei Ministri. Available online: http://www.governo.it/it/articolo/firmato-il-dpcm-9-marzo-2020/14276 (accessed on 18 July 2020).
34. Gazzetta Ufficiale della Repubblica Italiana. Decreto del Presidente del Consiglio dei Ministri 26 aprile 2020. Available online: https://www.gazzettaufficiale.it/eli/id/2020/04/27/20A02352/sg (accessed on 18 July 2020).
35. COVID-19 Italia—Monitoraggio Situazione. Available online: https://github.com/pcm-dpc/COVID-19 (accessed on 18 July 2020).
36. Arpae Emilia-Romagna. Available online: https://arpae.it/mappa_qa.asp?idlivello=1682&tema=stazioni (accessed on 18 July 2020).
37. Lauer, S.A.; Grantz, K.H.; Bi, Q.; Jones, F.K.; Zheng, Q.; Meredith, H.R.; Azman, A.S.; Reich, N.G.; Lessler, J. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Ann. Intern. Med.* **2020**, *172*, 577–582.

38. K Keller, J.M.; Gray, M.R.; Givens, J.A. A fuzzy k-nearest neighbor algorithm. *IEEE Trans. Syst. Man Cybern.* **1985**, *4*, 580–585.
39. Lawrence, R.L.; Wright, A. Rule-based classification systems using classification and regression tree (CART) analysis. *Photogramm. Eng. and Remote Sens.* **2001**, *67*, 1137–1142.
40. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.
41. Hastie, T.; Tibshirani, R.; Friedman, J.; Franklin, J. The elements of statistical learning: Data mining, inference and prediction. *Math. Intell.* **2005**, *27*, 83–85.
42. Dietterich, T.G. Ensemble methods in machine learning. In Proceedings of the First International Workshop on Multiple Classifier Systems, Cagliari, Italy, 21–23 June 2000; Springer: Berlin, Germany, 2000; pp. 1–15
43. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378.
44. Barandiaran, I. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell*. **1998**, *20*, 1–22.
45. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42.
46. Cohen, A.J.; Brauer, M.; Burnett, R.; Ross Anderson, H.; Frostad, J.; Estep, K.; Balakrishnan, K.; Brunekreef, B.; Dandona, L.; Dandona, R.; et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases. *Lancet* **2017**, *389*, 215.
47. Carbone, M.; Green, J.B.; Bucci, E.M.; Lednicky, J.A. Coronaviruses: Facts, Myths, and Hypotheses. *J. Thorac. Oncol.* **2020**, *15*, 675–678.
48. Delnevo, G.; Roccetti, M.; Mirri, S. Modeling Patients' Online Medical Conversations: A Granger Causality Approach. In Proceedings of the 2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies, Washington, DC, USA, 26–28 September 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 40–44.