Predicting the Incidence of COVID-19 Using Data Mining

Fatemeh Ahouz

Behbahan Khatam Alanbia University of Technology, Behbahan, Iran

amin golabpour

Shahrood University of Medical Sciences

Research article

Keywords: COVID-19, Predicting, Data Mining, Prevalence

Abstract

Background: The high prevalence of COVID-19 has made it a new pandemic. Predicting both its prevalence and incidence throughout the world is crucial to help health professionals make key decisions. In this study, we aim to predict the incidence of COVID-19 within a two-week period to better manage the disease.

Methods: The COVID-19 datasets provided by Johns Hopkins University, contain information on COVID-19 cases in different geographic regions since January 22 and are updated daily. Data from 252 such regions were analyzed as of March 29, 2020, with 17,136 records and 4 variables, namely latitude, longitude, date, and records . In order to design the incidence pattern for each geographic region, the information was utilized on the region and its neighboring areas gathered two weeks prior to the designing. Then, a model was developed to predict the incidence rate for the coming two weeks via a Least-Square Boosting Classification algorithm.

Results: The model was presented for three groups based on the incidence rate: less than 200, between 200 to 1000, and above 1000. The model evaluation error rates were 4.71%, 8.54%, and 6.13%, respectively. Also, comparing the forecast results with the actual values in the period in question showed that the proposed model predicted the number of globally confirmed cases of COVID-19 with a very high accuracy of 98.45%.

Conclusion: Using data from different geographical regions within a country and discovering the pattern of prevalence in a region and its neighboring areas, our boosting-based model was able to accurately predict the incidence of COVID-19 within a two-week period.

Background

On December 8, 2019 the Chinese government reported the death of one patient and hospitalization of 41 others with unknown etiology in Wuhan [1]. This cluster initiated the novel coronavirus (COVID-19) epidemic respiratory disease. While the early cases were linked to the wet market, human-to-human transmission had led to widespread outbreak of the virus nationwide[2]. On January 30, the World Health Organization (WHO) declared COVID-19 as a public health emergency with international concern (PHEIC) [3].

On the basis of the global spread and severity of the disease, on March 11, the Director-General of WHO officially declared the COVID-19 outbreak a *pandemic* [4]. The pandemic as such, entered a new stage with rapid spread in countries outside China [5]. According to the 56th WHO situation report [6], as of March 16, the number of COVID-19 confirmed cases outside China exceeded those inside. Consequently, after March 17, WHO began to report the number of confirmed and dead cases on each continent as opposed to merely providing patient statistics in and out of China.

According to the 70th WHO situation report[7], by March 30, the number of people infected with COVID-2019 worldwide were 693282. 392815 (about 57%) of whom were in Europe, 142081 (about 20%) in the Americas, 103775 (about 15%) in Western Pacific, 46329 (about 7%) in Eastern Mediterranean, 4084 (about 0.5%) in South-East Asia, and 3486 (about 0.5%) in Africa. Of that total, 33106died worldwide, 23962 of whom (around 72% of all death) were in Europe, 3649 (around 11%) in Western Pacific, and 5488 (around 17%) were in other regions collectively.

Due to the growing prevalence of COVID-2019 across the world, several works have examined different aspects of the disease. They involve identifying the source of the virus as well as analyzing its gene sequences [8, 9], patient information [10], early cases in the countries infected [11-13], methods of virus detection [14, 15], the epidemiological outbreak [16, 17], and predicting COVID-19 cases [2, 17-20].

In [18], using heuristic method and WHO situation reports, an exponential curve was proposed to predict the number of cases in the next two weeks by March 30. The model was then tested for the 58th situation report. The authors reported 1.29% error. Afterwards, on the assumption that the current trend could continue for the next 17 days, they predicted that by March 30, one million cases outside China would be reported in the 70/71th WHO situation report. Given that the number of confirmed cases outside China was 693176 on March 30 [21], their forecast error was 44.26%.

In [17], the CoronaTracker team proposed a Susceptible-Exposed-Infectious-Recovered (SEIR) model based on the queried data in their website, and made the 240-day prediction of COVID-19 cases in and out of China, started on 20 January. They predicted that the outbreak would reach its peak on May 23, and the maximum number of infected individuals would amount to 425.066 million globally. In addition, the authors stated that this number would start to drop around early July 2020 and reach below 10,000 on 14 Sep 2020. Given the information available now, these predictions were far from what really happened around the world.

Elsewhere [19], the authors examined some available models to predict 5 and 10-day ahead of cumulative cases in Guangdong and Zhejiang by February 23. They used generalized logistic growth, the Richards growth, and a sub-epidemic wave model, which were utilized to forecast some previous infectious outbreaks.

Although some works have proposed methods for predicting COVID-19 cases, to our knowledge at the time of writing this paper, none have been comprehensive and have not predicted the new cases in each geographical region along with each continent. In this study, using the COVID-19 Cases dataset provided by Johns Hopkins University [22], we aim to predict COVID-19 infected people in each geographical regions included in the dataset as well as each continent in the coming 2-week period. Predicting the situation in the current pandemic is very crucial to containment of the threat because it helps make timely medical measures e.g. equipping medical facilities, managing resource allocation, sending more personnel to high-risk areas, deciding whether to close borders or resume traffic, and suspending or resuming community services.

Methods

4.1 Dataset

COVID-19 epidemiological data have been compiled by the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) [22]. The data have been provided in three separate datasets for confirmed, recovered, and death cases since January 22 and are updated daily. In each of these datasets, there is a record (row) for every geographic region. The variables in each dataset are province/state, country/region, latitude, longitude, and the incremental dates since January 22. For each region, the value for any date indicates the cumulative number of confirmed/recovered/death cases from January 22.

In this study, according to the input requirements of the proposed model, we changed the data representation so that instead of three separate datasets for three groups of confirmed, recovered, and death cases, only one dataset containing the information of all three groups was arranged. In this new dataset, each record (or row) of the dataset contains information about the number of confirmed, recovered, or deaths per day for each geographic region. As a result, the variables in this new dataset are: Province / State, Country / Region, Latitude (Lat), Longitude (Long), Date (specifying a certain date), Cases (indicating the number of confirmed, recovered, or death cases on the certain date), and Type (specifying the type of cases, i.e. confirmed, recovered, or death) as suggested by Rami Krispin [23].

In this study, the data were applied into the analysis by March 29, 2020, with 50660 records and 7 variables. By March 29, the dataset consisted of cases from 177 countries and 252 different regions around the world. There were 720139 confirmed, 33925 death, and 149082 recovered cases in the dataset.

4.2 Preprocessing Step

Pre-processing was carried out on the dataset before training the proposed model. Figure 1 shows the preprocessing steps. The dataset was first examined for noise, since the noise data were considered as having negative values in Cases variable. The dataset contained 42 negative values in this variable. After deleting these values, the number of records were reduced to 50618.

Subsequently, the Date variable was written in numerical format and renamed into "Day" variable. To that effect, January 22 marked the beginning of the outbreak and the next days were calculated in terms of distance from the origin. As a result, January 22 and March 29 were considered as Day 1 and Day 68, respectively.

Since each region is uniquely identified by its latitude and longitude, the data for Province/Satate and Country/Region were excluded from the dataset. Moreover, as the study aimed at predicting the incidence in any geographical region, we considered only those records providing information on the confirmed cases (17179 records), but not on the dead or the recovered. So, after preserving the records with "Confirmed" value in the Type variable, it was deleted from the dataset. In this study, the "Cases" is considered as the dependent variable.

4.3 Constructing the Prediction Model

An ensemble method of regression learners was utilized to predict the incidence of COVID-19 in different regions. The idea of ensemble learning is to build a prediction model by combining the strengths of a collection of simpler base models called weak learners [24]. At every step, the ensemble fits a new learner to the difference between the observed response and the aggregated prediction of all learners grown previously. One of the most commonly used loss functions is least-squares (LS) error [25].

In this study, the model employed a set of individual Least-squares boosting (LSBoost) learners trying to minimize the mean squared error (MSE). The output of the model in step m, $F_m(x)$, was calculated using Equation 1:

$$F_m(x) = F_{m-1}(x) + \rho_m h(x; a_m) \tag{1}$$

where x is input variable and h(x;a) is the parameterized function of x, characterized by parameters a [25]. The values of ρ and a were obtained from Equation 2:

$$(\rho_m, a_m) = \arg\min_{a,\rho} \sum_{i=1}^{N} [\tilde{y}_i - \rho h(x_i; a)]^2 \tag{2}$$

Where N is the number of training data and is the difference between the observed response and the aggregated prediction up to the previous step.

Due to the recent major changes in the incidence of COVID-19 worldwide over the past 2 weeks, we aimed to predict the number of new cases as an indicator of prevalence over the next 2 weeks. The structure of the proposed method is shown in Figure 2.

Since the incubation period of COVID-19 can be 14 days [26], we assumed that we needed at least 14 days prior information to predict the incidence of Covid-19 in one day. Therefore, the proposed model examined all possible intervals between the first and the last 14 days to find the optimal time period to use its information to predict the number of cases in the coming days.

We hypothesized that the incidence in any region might follow the pattern of recent days in the same region and nearby. Therefore, after determining the optimal time period, the model added the information on confirmed cases in each region and nearby in the specified period to the same region's record in the dataset.

After setting the time interval, [A, B], and the number of neighbors, the data set was rearranged. In this case, the number of records was reduced from N to M, where M is calculated from Equation 3:

$$M = N - (B \times R) \tag{3}$$

Where R is the number of different regions in the data set and B is the last day of the time period. Similarly, the number of variables stored for each record increased from the first 4 variables (latitude, longitude, Day and Cases) to F, which is calculated from Equation 4:

$$F = 4 + 2 \times |B - A + 1| + NN \times (|B - A + 1| + 2) \tag{4}$$

Where NN is the number of neighbors and 4 is the number of variables in the original data set because for each geographical region, Lat, Long, Day and Cases are stored. |B-A+1| is the number of days within the period that participate in the forecast of the next 14 days. The value of NN is multiplied by 2 because for each neighbor, latitude and longitude are added to the record information. Furthermore, for each day within the period of forecast, the Cases were added to the record information, so NN was multiplied by|B-A+1|. For each region, the Day and Cases data during the period were added as well. Thus, |B-A+1| was multiplied by 2. It should be noted, however, that the dependent variable remained the Cases of current day.

Since the number of both the nearby regions and the previous days effective in forecasting were unknown, we assumed these values to be unknown variables and obtained the most accurate model by examining all possible combinations of such variables in an iterative process.

The accuracy of the model was evaluated in terms of Mean Squared Error (MSE) and Mean Absolute Error (MAE). To do so, the information of the last two weeks on all regions was considered as a validation set, and the model was trained using other information in the dataset.

4.4 Forecast Incidence in the Next Two Weeks

A new test set was created to predict incidence in the next two weeks (by April 12). The number of records in this dataset was equal to that of unique geographical regions in the COVID-19 dataset. Then, according to the best neighborhood and optimal time interval specified in the previous step, the necessary features were provided for each record. After that, the best model was created in the previous step was retrained on the entire dataset as a training set. Later on, this model was examined on the new test set to predict the incidence rate.

4.5 Evaluation the actual performance of the proposed model

Given that the actual number of confirmed cases within March 30-April 12, 2020 period was available at the time of review, the performance of the proposed model was measured based on percent error between the predicted and the actual values. The percent error was calculated from Equation 5:

$$\delta = \left(\frac{|v_A - v_E|}{v_A}\right) \times 100 \qquad (5)$$

Where δ is percent error, va is the actual observed value and   is the expected (predicted) value. Furthermore, according to the predicted and actual confirmed cases in 252 geographical regions in the dataset, the continental incidence rate was calculated using Equation 6:

$$Continental\ incidence\ rate = \left(\frac{I_C}{I_W}\right) \times 100 \qquad (6)$$

where I_C is the incidence in each continent and I_W is the global incidence of COVID-19 from March 30 to April 12.

The experimentation platform is Intel® Core ™ i7-8550U CPU @ 1.80GHz 1.99 GHz CPU and 12.0 GB of RAM running 64-bits OS of MS Windows 10. The pre-processing and model construction has been implemented in MATLAB.

Results

5.1 Model Construction

The number of neighbors ranged from zero to 10. The value of 10 was obtained by trial and error. Euclidean distance based on latitude and longitude was used to calculate nearest neighbors. Given that the dataset contains data from January 22 to March 29, for the day we want to predict the incidence, the nearest and farthest days were selected as 14 and 54, respectively. Because the number of confirmed cases varies greatly from region to region, the proposed algorithm was implemented for 3 different groups of regions: for regions with less than 200 confirmed cases per day (16825 records), those with 200 to 1000 cases per day (220 records), and those with over 1000 cases per day (152 records).

Table 1 shows the results of the best proposed model with regard to the different composition of the neighborhood and the days before. In order to predict the incidence of COVID-19 in regions with more than 1000 confirmed cases per day, the proposed model demonstrated the best performance with 6.13% error, considering the information of the last 14 to 17 days of the region and its two neighboring areas. In the dataset, the number of cases records in these regions varied from 1019 to 19821.

Table 1. The results of the best models evaluated on COVID-19 dataset (January 22 to March 29).

Maximum Number of confirmed cases in a day		Number of Neighbors	Interval of days [min,max]	MSE		MAE	
				Value	Percent	Value	Percent
<200	Train	-	[14,34]	1.86	0.005%	0.52	0.29%
	Test			407.47	1.04%	9.12	4.71%
[200,1000)	Train	9	[14,20]	1.71	0.002%	0.62	0.07%
	Test			1.59e+04	1.87%	79.01	8.54%
≥1000	Train	2	[14,17]	140.62	0.00003%	5.89	0.03%
	Test			7.14e+06	1.79%	1.2e+03	6.13%

For regions with 200 to 1000 cases per day, the proposed model performed best with respect to the 9 nearest neighboring areas and with data from the last 14 to 20 days, with an error of 8.54% on the validation set. For regions with fewer than 200 cases per day, on the other hand, the proposed model performs best with a 4.71% error, taking into account the region data for the last 14 to 34 days.

5.2 Prediction of Incidence by April 12

Figure 3 shows the prevalence of the COVID-19 from the first week to the tenth week in different regions, based on the information provided by the COVID-19 epidemiological dataset [22]. In this Figure, the diameter of the circles is proportional to the prevalence in those regions and the center of each circle matches the geographical coordinates of the region.

Table 2 shows the results of the forecast as to the number of new cases per day on different continents. By April 12, 1134018 new cases worldwide were expected to be on record. Of these, Europe with 687665 (60.64%), North America with 272957 (24.07%) and Asia with 107,000 (9.44%) new cases were the most prevalent, whereas Australia with 14526 (1.28%), Africa with 19131 (1.69%) and South America with 32.739 (2.89%) new cases were the least incidence. Africa, Europe and South America had the highest rates of COVID-19 incidence, with 283%, 221.23%, and 178.87%, respectively. Asia was the only continent that had slowed its growth with an incidencerate of -34.

Table 2. Forecast the COVID-19 new cases for the next two weeks.

Date	Continents						Total number of confirmed cases
	Africa	Asia	Australian	Europe	North America	South America	
22 Jan~29 Mar	4995	161986	4522	385097	150877	11740	719217
30-Mar	635	7720	802	37853	19269	1906	68185
31-Mar	820	7227	722	37433	16890	2000	65092
1-Apr	472	7533	338	38512	19625	1508	67988
2-Apr	1046	6438	981	44047	18435	1955	72902
3-Apr	1047	6790	780	53087	19802	2359	83865
4-Apr	1015	9739	872	51954	19302	2258	85140
5-Apr	1014	10563	1226	47352	19579	2490	82224
6-Apr	1447	6867	1015	48562	19060	2530	79481
7-Apr	1636	8027	1057	51192	20191	2768	84871
8-Apr	2087	6786	1444	56826	19546	2550	89239
9-Apr	2157	7749	1270	55316	20475	2685	89652
10-Apr	1976	5818	1430	54377	20819	2573	86993
11-Apr	1849	8962	1390	56284	19627	2351	90463
12-Apr	1930	6781	1199	54870	20337	2806	87923
Total	19131	107000	14526	687665	272957	32739	1134018
Prevalence growth rate	**283.00**	-33.94	221.23	78.57	80.91	178.87	57.67

Figure 4 shows the prediction of incidence rates in different regions. Accordingly, the prevalence would decrease over the next two weeks in the Middle East, yet it would increase in North America and Europe. Outbreak forecasts for 244 geographic regions are provided in Appendix.

5.3 Comparison of Predicted and Actual Cases From March 30 to April 12

Table 3 shows the total number of daily cases in the 252 regions surveyed between March 30 and April 12. As shown, the daily percent error is below 20%. The best accuracy of the proposed model in predicting the incidence of COVID-19 was obtained on April 10 with 99.6 %, and the worst on April 11 with 81.3%. Data analysis of the two-week continental incidence rates are also shown in Figure 5. The best predicted continental incidence rates were found in South America and Asia with 18.15% and 21.04% percent error, respectively. The worst cases, still, were observed in Africa and Australian with more than 80% percent errors.

Table 3. comparison of predicted and actual daily incidence of COVID-19.

| Date | Across all 252 geographic regions | | Percent error |
	Predicted	Actual	
30-Mar	68185	65321	4.38%
31-Mar	65092	76799	15.24%
1-Apr	67988	76657	11.31%
2-Apr	72902	81340	10.37%
3-Apr	83865	83272	0.71%
4-Apr	85140	80392	5.91%
5-Apr	82224	71994	14.21%
6-Apr	79481	73285	8.45%
7-Apr	84871	77773	9.13%
8-Apr	89239	84275	5.89%
9-Apr	89652	86461	3.69%
10-Apr	86993	87520	0.60%
11-Apr	90463	76217	18.69%
12-Apr	87923	95353	7.79%
Total number of confirmed cases	1134018	1116659	1.55%

Discussion

Data mining is capable of presenting a predictive model and extracting new knowledge from retrospective data. The way data is processed, as well as the variables selected, had a significant impact on knowledge discovery. There are various data mining techniques used to predict an outbreak. As an actual global health concern, COVID-19 had already developed into one of the world's major emergencies. The present study proposed to investigate its outbreak worldwide during a two-week period via a predictive model based on retrospective data. It was concluded that such a model could be presented with acceptable error rates.

The study made use of a coronavirus dataset to design an incidence of COVID-19 prediction model. According to the incidence rate per day, the model was trained based on three groups of below 200, 200-1000 and above 1,000 cases. One-way ANOVA results showed that there was a statistically significant difference between the prevalence rates in the three groups (p-value <0.001). For each group, the prediction model was implemented and the incidence was predicted for the next two weeks. The proposed model achieved 10% error (90% similarity) for the group of less than 200 cases, 18% error (82% similarity) for the group of 200-1000 cases, and 14% error (86% similarity) for that exceeding 1,000 cases.

In this study, as the incidence of COVID-19 was evaluated for 68 days worldwide, and a prediction model presented for the two-week period (i.e., March 30-April 12), more than 1,000,000 people were expected to contract the disease within the next two weeks, which was statistically up 58% compared to 700,000 of the outbreak so far.

The study found that adjacent regions with a prevalence of less than 1,000 had similar incidence, so the incidence of each of these regions could be determined from information on neighboring areas.

Given that the proposed model was trained using only 68-day data (which was the most up-to-date information at the time of writing), the accuracy of predicting the incidence above 81% was deemed acceptable for such an unknown disease. Further, according to the results shown in Table 3, the model prediction error for a total of 12 days for 252 regions was less than 2%. Therefore, if the data of each country were stored more precisely using more geographical regions, it was promising that we could create an accurate model for predicting the incidence of covid-19 over a two-week period in each country. While many unknowns would be expected of a new pandemic, having this information can guide planning and resource allocation for prevention, treatment, and palliative care.

One of the limitations of the study was that the data set did not provide sufficient information from all continents. Given that the disease did not occur simultaneously on all continents, and the continental prevalence was in most cases after the 40th day of the first case in China, 68 days of data did not seem sufficient to predict the prevalence of such an unknown disease.

In Africa, the first case was reported in more than 80% of the 45 geographical regions since the 50th day. The number of confirmed cases since then was 4682, which was 97.83% of the total 4783 confirmed cases in Africa. In Australian, the first case was reported in more than 45% of the 11 geographical regions from the 40th day onwards. Also, out of a total of 4504 cases on the continent, 4478 cases (99.4%) were confirmed then.

In Europe, the first case was reported in 60 of the 69 geographic regions in the dataset from the 40th day onwards. Out of a total of 385735 cases, information on 384268 cases (i.e. 99.62%) has also been entered since that day. Similarly, South America confirmed its first case after the 40th day in 16 out of 17 regions. It is noteworthy that out of a total of 11,642 cases, 11,542 (14.99%) were confirmed from day 50 onwards.

In contrast, 88% of the North American regions had their first cases confirmed since day 50. In addition, of the 46 confirmed cases by March 29 on the continent, 38 were reported since day 50 (82.61%) And 41 were confirmed from day 40 onwards (89.13%).

Due to insufficient information on some continents as a result of their prevalence later than the declared beginning of the outbreak, the effect of measures such as increasing the number of tests taken per day as well as quarantine restrictions in some continents such as Europe, begin in place from March 30 to April 12, were not reflected in the dataset.

Nevertheless, the inaccurate prediction of the number of cases in Africa could be attributed, in turn, to the insufficient information about the continent in the dataset. In 80% of the African regions, the first confirmed case was recorded 50 days into the outbreak. Out of a total of 4786 cases there, up until the 68th day, 4682 cases (more than 97%) were reported since day 50.

Also, due to the fact that latitude and longitude are two important indicators in the data set, the non-uniformity of recording these information for different geographical regions is another limitation of the work; for some areas, the information is about one state of a country and for some areas it is for the whole country. For example, in the data set for USA, all cases are provided in terms of only one latitude and longitude, but for Netherlands, the data of COVID-19 cases are provided for four different latitude and longitude pairs.

Another limitation of this study was the use of data from all countries coping with in COVID-19 with their own protocols for testing and identifying patients. However, in general, this is the only global dataset for COVID-19 that has been used in other studies [16, 17]. Besides, early information on each country was taken into account in the proposed model to predict the incidence in that country so as to reduce the mentioned limitation.

It is worth noting that the model rests on both the info provided by the dataset and the current measures taken in dealing with the disease. Hence, if government's' policies to tackle the disease change, so will the accuracy of the information.

Conclusions

Since epidemiological models such as SIR failed to accurately predict COVID-19 cases, as stated in [17, 27, 28], the current study relied on data from January 22 to March 29 provided by Johns Hopkins University and proposed a more complex model based on machine learning methods. The mean absolute error of the proposed model was 6.13% in predicting the incidence of COVID-19 in the two-week period of March 16-29 for regions with more than 1000 cases per day. The error was 8.45% and 4.71% for regions with a daily incidence rate between 200 and 1000 cases and less than 200 cases, respectively. An accuracy of more than 91% on the evaluation set confirms our perception that the pattern of incidence of a region is influenced by the pattern of disease in recent days in the same region and neighboring areas.

Last but not least, despite numerous limitations of the dataset, lack of knowledge about such an unknown disease and changes in disease control policies in different countries during the period under scrutiny, the proposed model proved effective in predicting the global incidence of COVID-19 in the two-week period of March 30 and April 12 with 98.45% accuracy. In addition, the accuracy of the proposed model in predicting daily cases in a worst-case scenario was 81.31%.

This model is written in general and can be run for different intervals. It is suggested that the model be implemented for future data as well.

List Of Abbreviations

WHO: World Health Organization; PHEIC: Public Health Emergency with International Concern; SEIR: Susceptible-Exposed-Infectious-Recovered; JHUCCSE: Johns Hopkins University Center for Systems

Science and Engineering; Lat: Latitude; Long: Longitude; LSBoost: Least-squares boosting; MSE: Mean Squared Error; MAE: Mean Absolute Error.

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: The dataset analyzed during the current study is public and it is available in the [https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases] and in [https://codeload.github.com/RamiKrispin/coronavirus-csv/zip/master].

Competing interests: The authors declare that they have no competing interests.

Funding: Not applicable.

Authors' contributions: 'FA' and 'AG' equally contributed to the conception, design of the work, analysis and interpretation of data. In addition, they read and approved the final manuscript.

Acknowledgements: Not applicable

References

1. Nkengasong, J., *China's response to a novel coronavirus stands in stark contrast to the 2002 SARS outbreak response.* Nature Medicine.
2. Roosa, K., et al., *Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020.* Infect Dis Model, 2020. **5**: p. 256-263.
3. Eurosurveillance Editorial, T., *Note from the editors: World Health Organization declares novel coronavirus (2019-nCoV) sixth public health emergency of international concern.* Eurosurveillance, 2020. **25**(5): p. 2-3.
4. *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020.* 11 march 2020; Available from: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.
5. Bedford, J., et al., *COVID-19: towards controlling of a pandemic.* 2020.
6. who, *World Health Organization, Coronavirus disease 2019 (COVID-19)Situation Report −60.* 2020.
7. *World Health Organization,Coronavirus disease 2019 (COVID-19)Situation Report −70.* 2020.
8. Ji, W., et al., *Cross-species transmission of the newly identified coronavirus 2019-nCoV.* Journal of Medical Virology, 2020. **92**(4): p. 433-440.
9. Paraskevis, D., et al., *Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event.* Infect Genet Evol, 2020. **79**:

p. 104212.

10. Huang, C., Y. Wang, and X. Li, *Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China (vol 395, pg 497, 2020).* Lancet, 2020. **395**(10223): p. 496-496.

11. Kim, J.Y., et al., *The First Case of 2019 Novel Coronavirus Pneumonia Imported into Korea from Wuhan, China: Implication for Infection Prevention and Control Measures.* Journal of Korean Medical Science, 2020. **35**(5).

12. Bernard Stoecklin, S., et al., *First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020.* Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin, 2020. **25**(6).

13. Giovanetti, M., et al., *The first two cases of 2019-nCoV in Italy: Where they come from?* Journal of Medical Virology.

14. Corman, V.M., et al., *Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR.* Eurosurveillance, 2020. **25**(3): p. 23-30.

15. Zhang, N.R., et al., *Recent advances in the detection of respiratory virus infection in humans.* Journal of Medical Virology, 2020. **92**(4): p. 408-417.

16. Dey, S.K., et al., *Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach.* Journal of Medical Virology.

17. Binti Hamzah, F.A., et al., *CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction.* 2020.

18. Koczkodaj, W.W., et al., *1,000,000 cases of COVID-19 outside of China: The date predicted by a simple heuristic.* Global Epidemiology, 2020: p. 100023.

19. Roosa, K., et al., *Short-term Forecasts of the COVID-19 Epidemic in Guangdong and Zhejiang, China: February 13-23, 2020.* J Clin Med, 2020. **9**(2).

20. Nishiura, H., et al., *The Extent of Transmission of Novel Coronavirus in Wuhan, China, 2020.* Journal of Clinical Medicine, 2020. **9**(2).

21. Organization, W.H. *Coronavirus disease 2019 (COVID-19)Situation Report – 70.* 30 March 2020; Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200330-sitrep-70-covid-19.pdf?sfvrsn=7e0fe3f8_4.

22. CCSE), J.H.U.C.f.S.S.a.E.J. *Novel Coronavirus (COVID-19) Cases Data.* 2020; Available from: https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases.

23. Krispin, R. *coronavirus.* 2020 1 march 2020]; Available from: https://github.com/RamiKrispin/coronavirus.

24. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning, second edition.* Springer Series in Statistics. 2008: Springer-Verlag New York.

25. Friedman, J., *Greedy Function Approximation: A Gradient Boosting Machine.* The Annals of Statistics, 2000. **29**.

26. Organization, w.H. *Transmission of SARS-CoV-2: implications for infection prevention precautions*. 9 July 2020; Available from: https://www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions#:~:text=The%20incubation%20period%20of%20COVID,to%20a%20confirmed%20case.

27. Postnikov, E.B., *Estimation of COVID-19 dynamics "on a back-of-envelope": Does the simplest SIR model provide quantitative parameters and predictions?* Chaos, Solitons & Fractals, 2020. **135**: p. 109841.

28. Cooper, I., A. Mondal, and C.G. Antonopoulos, *A SIR model assumption for the spread of COVID-19 in different communities.* Chaos, Solitons & Fractals, 2020. **139**: p. 110057.

Figures

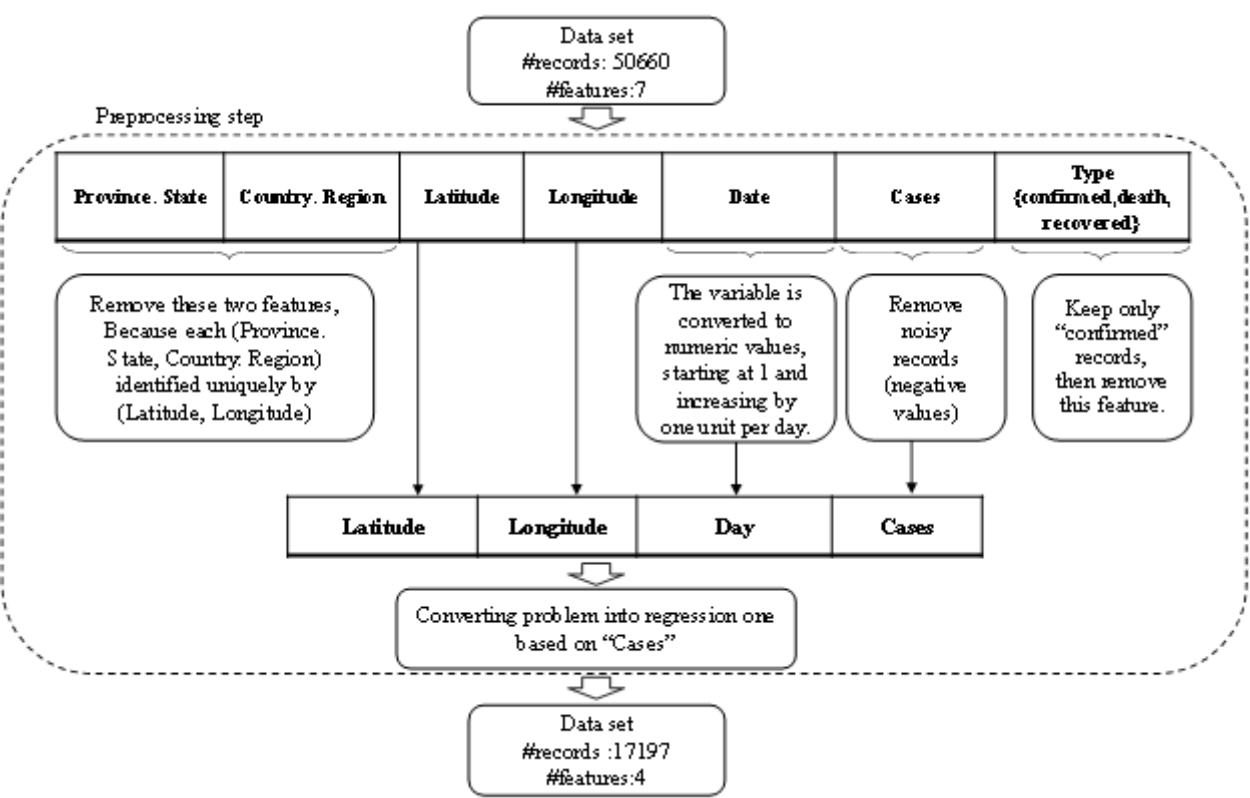



Figure 1

Preprocessing steps on coronavirus dataset.

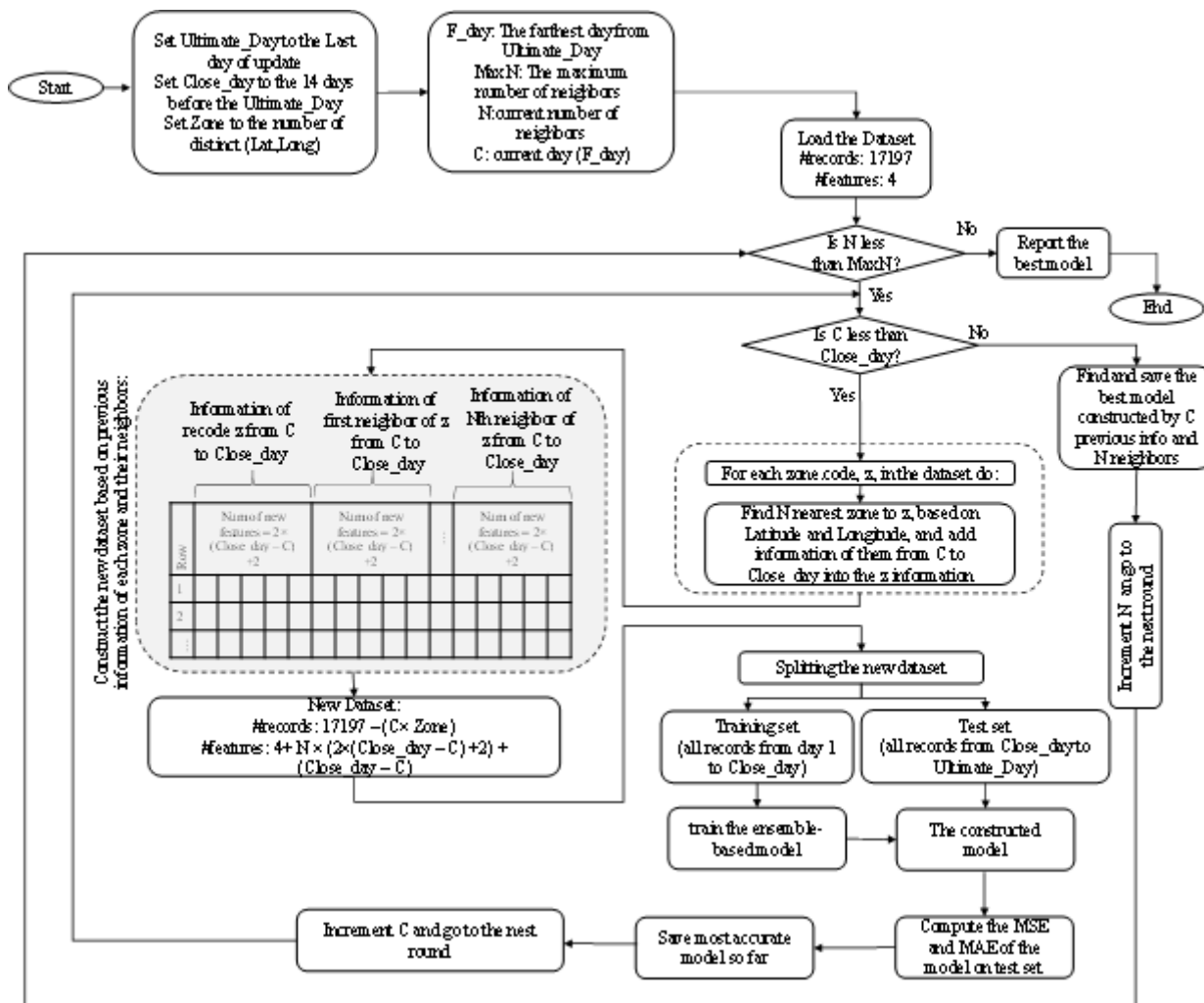



Figure 2

The structure of the Proposed model.

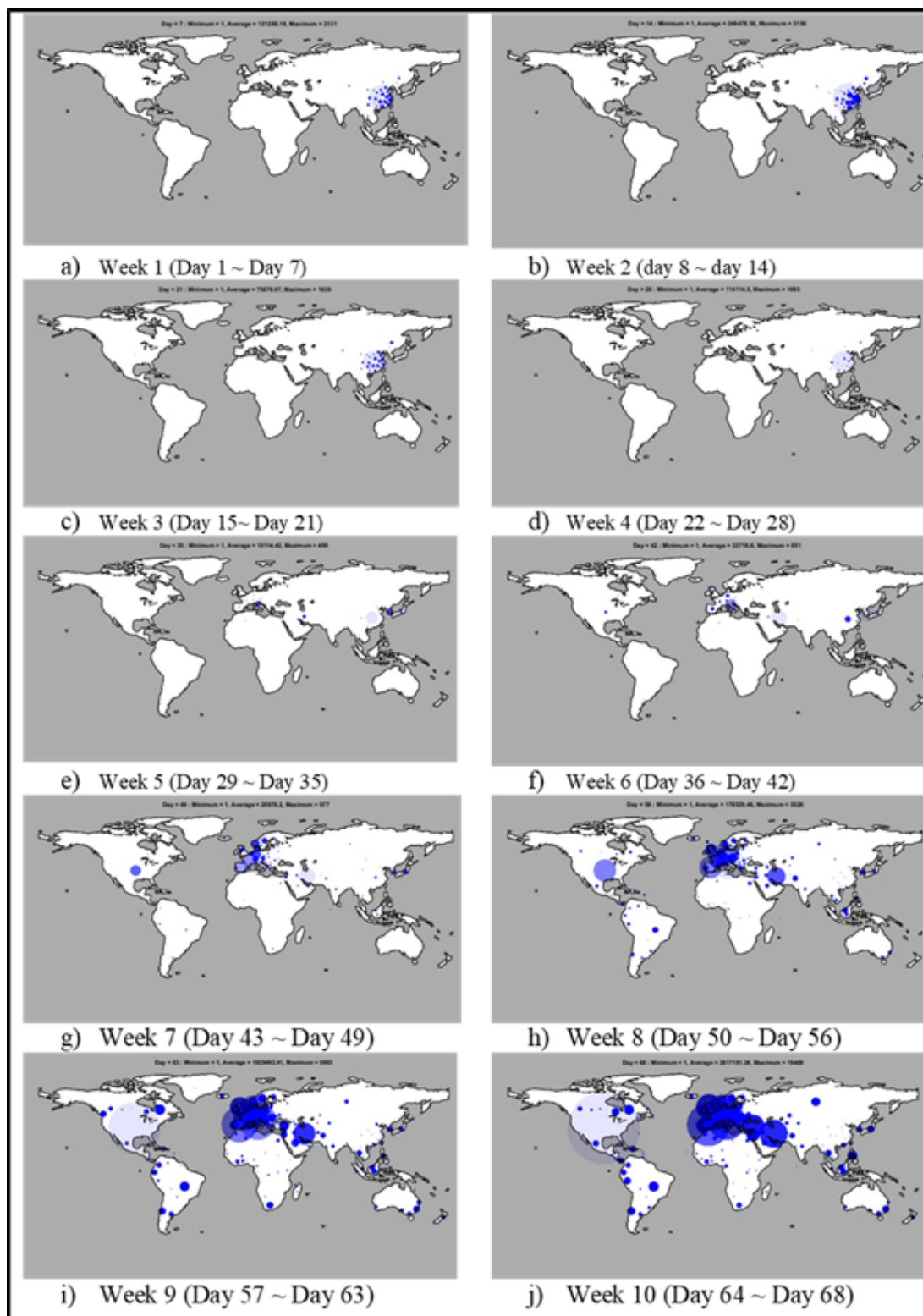



Figure 3

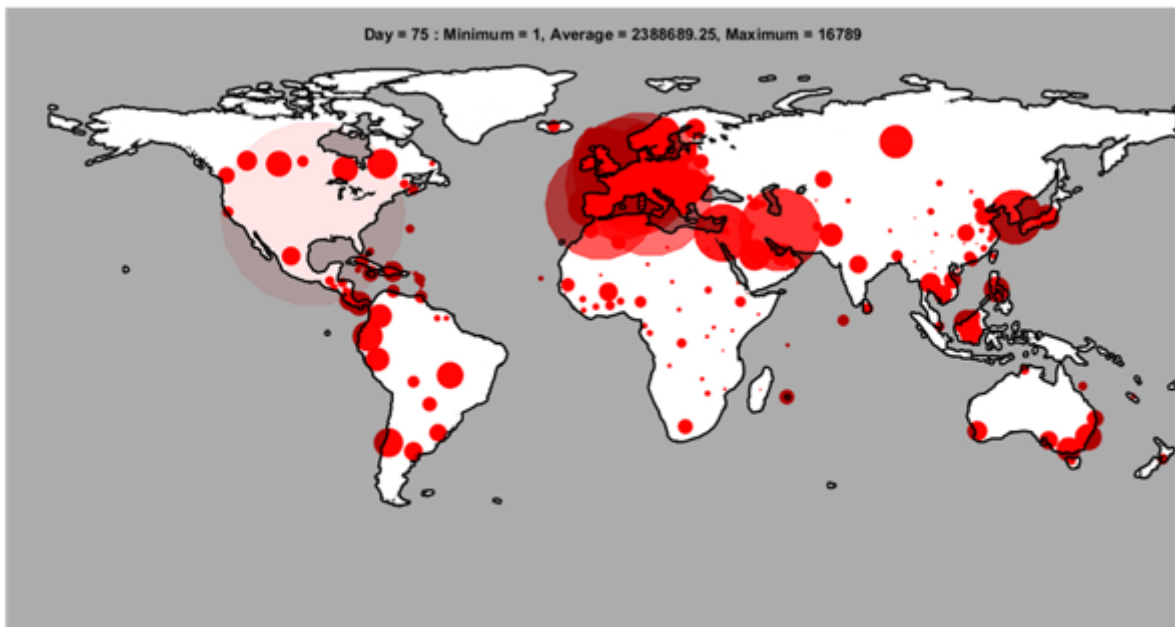
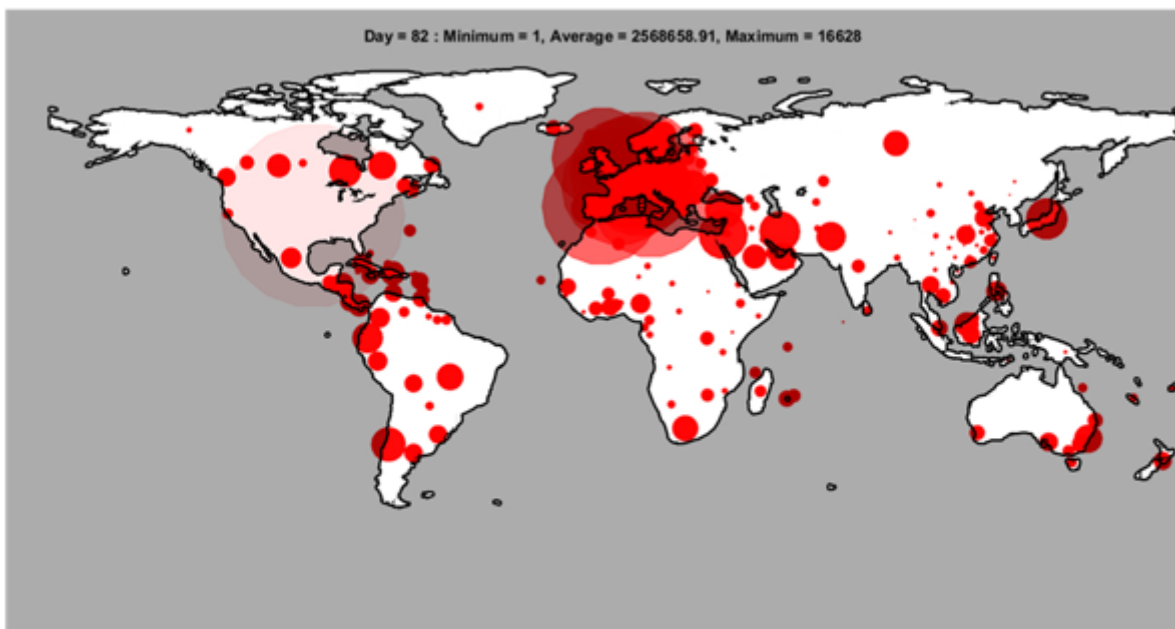
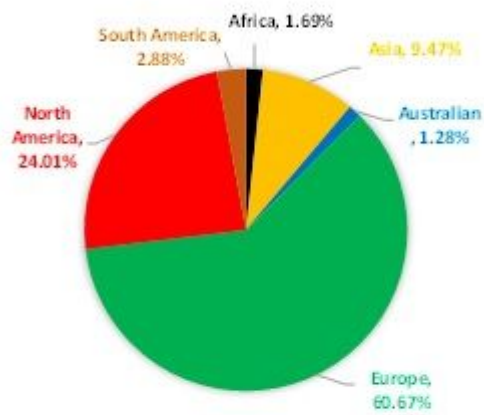
Visualize the outbreak over the days. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

a) Prediction of the prevalence on day 75 (end of week 10).

b) Prediction of the prevalence on day 82 (week 11).

Figure 4

Prediction of the prevalence in week 10 and 11. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

a) Predicted incidence rate b) Actual Incidence rate

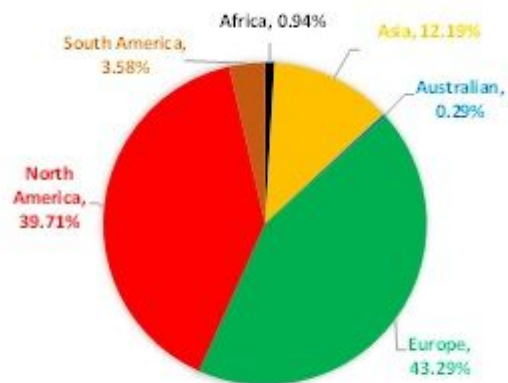



Figure 5

Comparison of predicted and actual continental incidence rates between March 30 and April 12.