Predicting the COVID-19 Prevalence Rate Using Data Mining

Fatemeh Ahouz
Behbahan Khatam Alanbia University of Technology, Behbahan, Iran
amin golabpour (a.golabpour@shmu.ac.ir)
Shahrood University of Medical Sciences https://orcid.org/0000-0001-7649-4033

Research article

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

Posted Date: April 22nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-21247/v1

Version of Record: A version of this preprint was published at BMC Public Health on June 7th, 2021. See the published version at https://doi.org/10.1186/s12889-021-11058-3.

Abstract

Background The high prevalence COVID-19 has made it a new pandemic. Predicting the prevalence and incidence of this disease throughout the world is crucial to helping health professionals make key decisions about the disease.

Methods The coronavirus dataset contains information on COVID-19 cases in 252 geographic regions since January 22 and is updated daily. Data are included in the analysis as of March 29, 2020, with 17,136 records and 4 variables: latitude, longitude, date, and records. In order to design the prevalence pattern for each geographic area, the information of the region and its neighborhoods in the past two weeks, has been used. Then, using a Boosting Classification algorithm, a method was developed to predict the prevalence rate for the next two weeks.

Results The model was presented for three groups with a prevalence of less than 200, ranging from 200 to 1000, and more than 1000 cases, and the model error rates were 9.42, 17.08, and 12.26, respectively. In addition, more than 1 million new cases are expected to become infected in the next two weeks (March 30 - April 12). The number of new cases are expected to be more than 19,000 new cases in Africa, over 100,000 in Asia, over 14,000 in Australia, over 600,000 in Europe and over 300,000 in the Americas.

Conclusion The increase in the prevalence growth rate of the disease will also be in the Southern Hemisphere, and the United States will have the highest prevalence worldwide.

Background

On December 8, 2019, the Chinese government reported the death of one patient and the hospitalization of 41 others with unknown etiology in Wuhan [1]. This cluster initiated the novel coronavirus (COVID-19) epidemic respiratory disease. While early cases of the disease were linked to the wet market, human-to-human transmission has led to widespread outbreak of the virus through China [2]. On January 30, the World Health Organization (WHO) announced the emergence of 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 of COVID-19 has 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 of China has exceeded of China, so after March 17, instead of providing patient statistics in and outside of China, WHO report the number of confirmed and dead cases on each continent.

According to the 70th WHO situation reports [7], by March 30, the number of people infected with COVID-2019 worldwide is 693282, of which 392815 (about 57%) are in European Region, 142081 (about 20%) in Region of the Americas, 103775 (about 15%) in Western Pacific Region, 46329 (about 7%) in Eastern Mediterranean Region, 4084 (about 0.5%) in South-East Asia Region, and 3486 (about 0.5%) in African

Region. Of these, 33106 globally died, of which 23962 (about 72% of all death) are in European Region, 3649 are in western pacific Region (about 11% of all death) and 5488 (about 17%) are collectively in other regions.

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

In [18], the exponential curve is proposed for forecasting the growth of new cases for two-week ahead, by March 30, based on the WHO situation reports and the heuristic method. The model has been tested for the 58th situation report based on previous reports. They reported 1.29% error. Then based on this assumption that the current trend can continue for the next 17 days, they predicted one million new cases outside of China by March 30.

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 is reached its peak on May 23, and the maximum number of infected individuals will be 425.066 million globally. Then it will start to drop around early July 2020 and reached under 10,000 on 14 Sep 2020.

In [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 models, which were used to forecast some previous infectious outbreaks.

Although some work has provided methods for predicting COVID-19 cases, on the basis of our knowledge at the time of writing this paper, none have been comprehensive and have not predicted the new cases in each geographical region as well as each continent. In this study, using the coronavirus dataset, we aim to predict COVID-19 infected people in each geographical regions included in the dataset as well as each continent in 2-week ahead. Predicting this situation in the current pandemic is very crucial to containment of the threat because it helps to make timely operations and medical decisions. These include equipping medical facilities, managing the resource allocation, sending more personnel to high-risk areas, deciding whether to close borders or resume traffic, suspend or resume community services.

Methods

3.1 Dataset

The coronavirus dataset is provide by Rami Krispin [21]. This dataset is taken from coronavirus package that provides a tidy format dataset of the COVID-19 pandemic. The raw data pulled from the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) Coronavirus repository.

This dataset contains COVID-2019 case information from January 22 and is updated daily. In this study, data entered into the analysis by March 29, 2020, with 50660 records and 7 variables. Each record specifies a region. Variables include Province State, Country Region, latitude (Lat), longitude (Long), case record date (Date), number of cases (Cases), and types of cases (Type). By March 29, the dataset includes cases from 177 countries and 252 different geographic regions around the world. There are 720117 confirmed, 33925 death, and 149082 recovered cases in the dataset.

3.2 Preprocessing Step

Pre-processing runs on the dataset before training the proposed model. Figure 1 shows the preprocessing steps. The dataset is first examined for noise data. In this study, noise data are considered data which have negative values in Cases variable. This dataset contains 42 negative values in this variable. After deleting these values, the number of records were reduced to 50618.

After that, the Date variable is written in numerical format, and its name is changed to Day variable. To do this, January 22 will be considered as the beginning of the outbreak and the next days will be calculated in terms of distance from the origin. As a result, January 22 and March 29 are considered as Day 1 and Day 68, respectively.

Since each geographical region is uniquely identified by its latitude and longitude, the data for Province.Satate and Country.Region are excluded from the dataset. Also, since the goal of the study is predicting the prevalence in geographical regions, we consider only records that provided the information of confirmed cases, not death nor recovered ones. So, after preserving the records with "Confirmed" value in the Type variable, this variable is deleted from the database. In this study, the Cases is considered as dependent variable.

3.3 Constructing the Prediction Model

An ensemble method of regression learners is used to predict the prevalence of COVID-19 in different geographical regions. The model uses a set of individual Least-squares boosting (LSBoost) learners that tries to minimize the mean squared error (MSE).

Due to the recent major changes in the prevalence of COVID-19 worldwide over the past 2 weeks, we aim 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 Fig. 2.

Based on our idea that the prevalence in a geographical region may follow the pattern of prevalence in recent days in this region and in the neighboring regions, the model has been proposed that adds previous information of the region and its neighbors to the region's record in the dataset to predict the prevalence in that region.

Suppose we want to use the data from the last 14 to 20 days of the region and its two neighbors as well to predict prevalence in next two weeks. In this case, for each geographical region, the dataset will be reconstructed from the day 21. That is, for example, in day 21 for each geographical region, the Lat, Long,

Day and Cases are stored (same as the original coronavirus dataset). Then, the information of Cases and Day recorded in day 1 to day 7 of this region, are added to the information of the record as new variables, and store in the 5th through 18th columns of new dataset. For each neighbor its latitude and longitude and the number of confirmed cases in days 1 through 7 are stored. So, the latitude and longitude of the first neighbor are stored in columns 19 and 20 and the Date and corresponding Cases are stored in columns 21 through 34. The same process is accomplished for other neighbors. Therefore, in each iteration of the algorithm, the number of records is reduced by R, which is calculated from Eq. 1:

$$R = NumOfAllRecords - (C \times Zone) \quad (1)$$

Where *NumOfAllRecords* is the number of records in the coronavirus dataset, *Zone* specifies the number of unique geographic regions in the dataset, and *C* is the farthest day used in the forecast. On the other hand, by using the previous days' information for prediction, the number of features, F, stored for each record increases by Eq. 2:

$$F = 4 + N \times (2 \times D) + D \quad (2)$$

Where N is the number of neighbors and D is the number of previous days that participate in the forecast of the next 14 days. The value of N is multiplied by 2 because for each neighbor, latitude and longitude are added to the record information. Also for each previous day that are used in forecast, the Day and the Cases are added to the record information, so multiply D by 2. D has been added at the end of the Eq. (2) because the information of D previous days of that region is also added to its record. It should be noted that the dependent variable remains the Cases of current day.

Since the number of neighbors and the number of previous days that can be effective in forecasting is unknown, we assume these values to be unknown variables, and we obtain the most accurate model by examining all possible combinations of these variables in an iterative process.

The accuracy of the model is evaluated in terms of Mean Squared Error (MSE) and Mean Absolute Error (MAE). To evaluate the model, the information of the last two weeks of all geographical regions are considered as a testset and the model is trained on other records.

3.4 Forecast Prevalence in the Next Two Weeks

A new test set is created to predict prevalence in the next two weeks (by April 12). The number of records in this dataset is equal to the number of unique geographical regions in the coronavirus dataset. Then, according to the best neighborhood and the number of days specified in the previous step, the necessary features are provided for each record. Then, the best model created in the previous step is trained on the entire dataset. After that, this model is evaluated on the new test set to predict the prevalence rate.

Results

The experimentation platform is Intel® Core ™ i7-8550U CPU @ 1.80 GHz 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.

4.1 Model Construction

The number of neighbors ranged from zero to 10. The value of 10 was obtained by trial and error. Euclidean distance between 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 prevalence, the nearest and farthest days were selected as 14 and 54, respectively. Since the number of confirmed cases is very different in different regions, the proposed algorithm was implemented for 3 different intervals: for areas with less than 200 cases per day over all 68 days (16825 records), for areas with 200 and less than 1000 cases per day (220 records) and for areas with more than 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 prevalence in regions with more than 1000 confirmed cases per day, the proposed model has the best performance with 6.13% error, considering the information of the last 14 to 17 days of the region and its two neighbors. 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 coronavirus dataset (January 22 to March 25).

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 performs best with respect to the 9 nearest neighbors and with data from the last 14 to 20 days, with an error of 8.54% in the test set. For regions with fewer than 200 cases per day, the proposed model performs best with a 4.71% error, taking into account the region data for the last 14 to 34 days.

4.2 Prediction of Prevalence 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 coronavirus dataset. In this Figure, the diameter of the circles is proportional to the prevalence in those areas and the center of each circle is on the geographical coordinates of the region.

Table 2 shows the results of the forecast of the number of new cases per day on different continents. By April 12, 1134018 new cases worldwide are expected to be registered. Of these, Europe with 687665 (60.64%), North America with 272957 (24.07%) and Asia with 107,000 (9.44%) new cases will be most prevalent, and Australia with 14526 (1.28%), Africa with 19131 (1.69%) and South America with 32.739 (2.89%) new cases will have the lowest prevalence. Africa, Europe and South America had the highest rates of COVID-19 prevalence, with 283%, 221.23%, and 178.87%, respectively. Asia is the only continent that has slowed its growth with a prevalence rate 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 prevalence rates in different geographic regions. Accordingly, the prevalence will decrease over the next two weeks in the Middle East. But it will increase in North America and Europe. Outbreak forecasts for 244 geographic regions are provided in Appendix.

The prevalence plot of each continent from day 1 (January 22) to day 82 (April 12) is shown in Fig. 5. The light blue polygon points to the predicted values and the dark circles to the actual values based on the coronavirus data. Given the figure, the slope of the plot in Asia will decrease slightly in the coming days.

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, has a significant impact on knowledge discovery. There are various data mining techniques used to predict the outbreak. COVID-19 is a global health concern and has become one of the world's major emergencies. The present study investigated the prevalence of COVID-19 worldwide and based on retrospective data, a predictive model for its prevalence has been presented. It was found that the predictive model for COVID-19 prevalence worldwide could be presented with acceptable error rates.

The study used a coronavirus dataset to design a COVID-19 prevalence prediction model. Based on the prevalence rate per day, the model was trained based on three groups of less than 200, between 200 and 1000 and more than 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 prevalence was predicted for the next two weeks. The proposed model achieved 10% error (90% similarity) for the prevalence group of less than 200 cases, 18% error (82% similarity) for the prevalence group between 200 and 1000 cases, and 14% error (86% similarity) for the prevalence group of more than 1,000 cases.

In this study, the prevalence of COVID-19 was evaluated for 68 days worldwide, and a prediction model for the next two weeks (March 30 ~ April 12) was presented. More than 1,000,000 people are expected to infect the disease within the next two weeks, up 58% compared to 700,000 of the outbreak so far.

The prevalence of the disease was analyzed in each continent. The number of new cases in Africa is estimated to reach 19131 at 46 points, up 283% from the 4995 affected so far. In Asia, it is estimated that 107,000 people will be affected in the next two weeks, about 76% less than the 161986 cases so far. Australian will have a prevalence of 14526 at 12 points, an increase of 221 percent from the 4522 affected so far. Europe will have a prevalence of 687655 at 51 points, an increase of 78 percent from the 385097 affected so far. New cases in the North American continent will reach 272,957 at 43 points, an increase of 81 percent over the 150,877 people who have ever been infected. The South American continent has a prevalence of 32,739 at 19 points, an increase of 178 percent from the 1,740 people affected so far.

According to our analysis, within the next 2 weeks, among the 252 different geographic regions, the United States has the highest prevalence of the disease with a prevalence of 239680, an increase of 70% compared to 140886 and across all continents, Africa has the highest prevalence rate with 283%.

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

The biggest limitation of this study is that the analysis is based on data recorded from the coronavirus dataset, and in some continents, such as Africa, very little information is recorded. Therefore, based on the information provided by the dataset and the current operations in dealing with the disease, the

prediction model is presented. If government's' policies to tackle the disease change, the accuracy of this information will change.

Another limitation of this study is the use of data from all countries involved in COVID-19, while each country has its own protocol 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]. Also, in the proposed model, the past information of each country has been used to predict the prevalence of that country, which reduces the mentioned limitation.

Conclusions

In this study, based on the COVID-19 data from January 22 to March 29, a prevalence prediction model is developed with an error of 17%, which can predict the prevalence of COVID-19 in all infected areas by April 12, with an accuracy of 83%. In the proposed model, the prevalence for each continent was calculated separately and it was found that Europe and America would have the highest prevalence of the disease. The United States will also have the highest prevalence in the world in the next two weeks, with more than 200,000 new cases. In addition, over the next two weeks, more than 1 million new cases are expected to be confirmed worldwide which indicates an increasing trend in the prevalence of the disease. Therefore, we have not yet reached the peak of this disease, so it is recommended that this model be continuously implemented with daily data of this disease to determine when the prevalence of COVID-19 will begin to decrease.

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 analysed during the current study is public and it is available in the [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 –56*. 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. Krispin, R. *coronavirus*. 2020 1 march 2020]; Available from: https://github.com/RamiKrispin/coronavirus

Figures

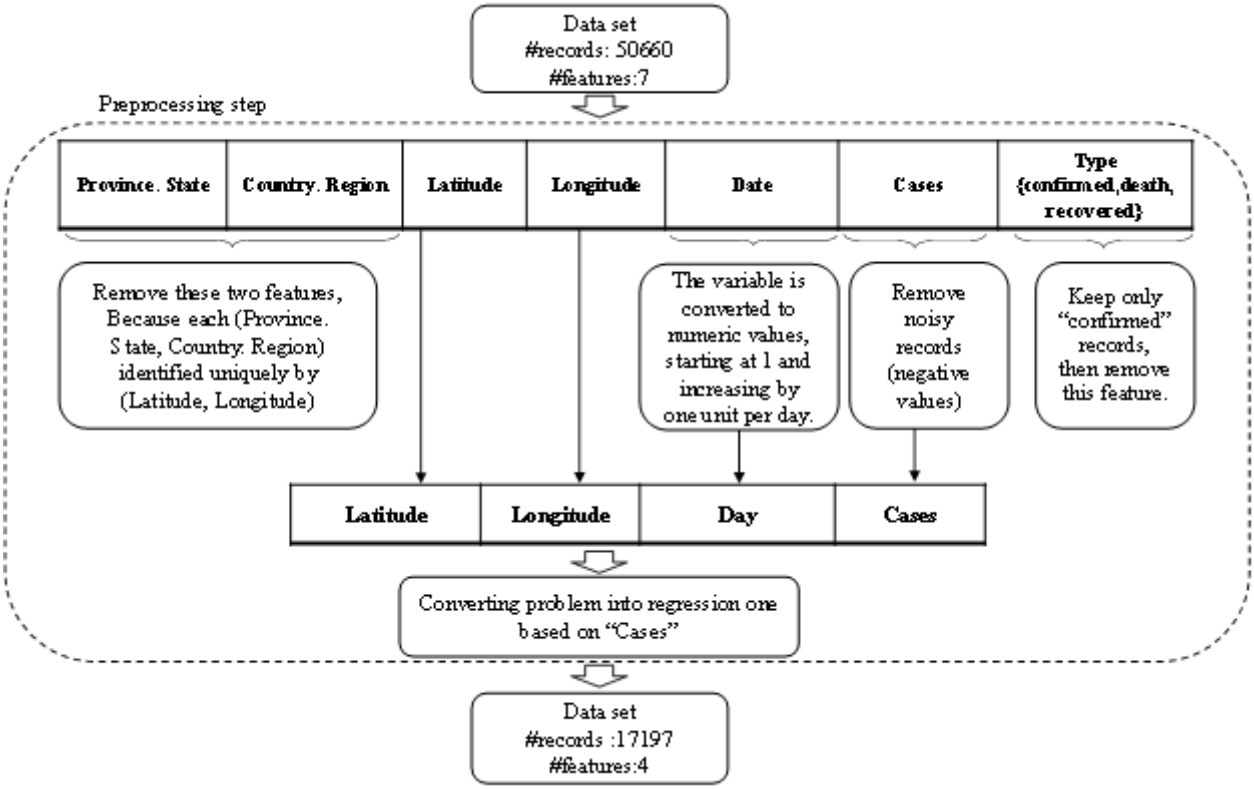


Figure 1

Preprocessing steps on coronavirus dataset.

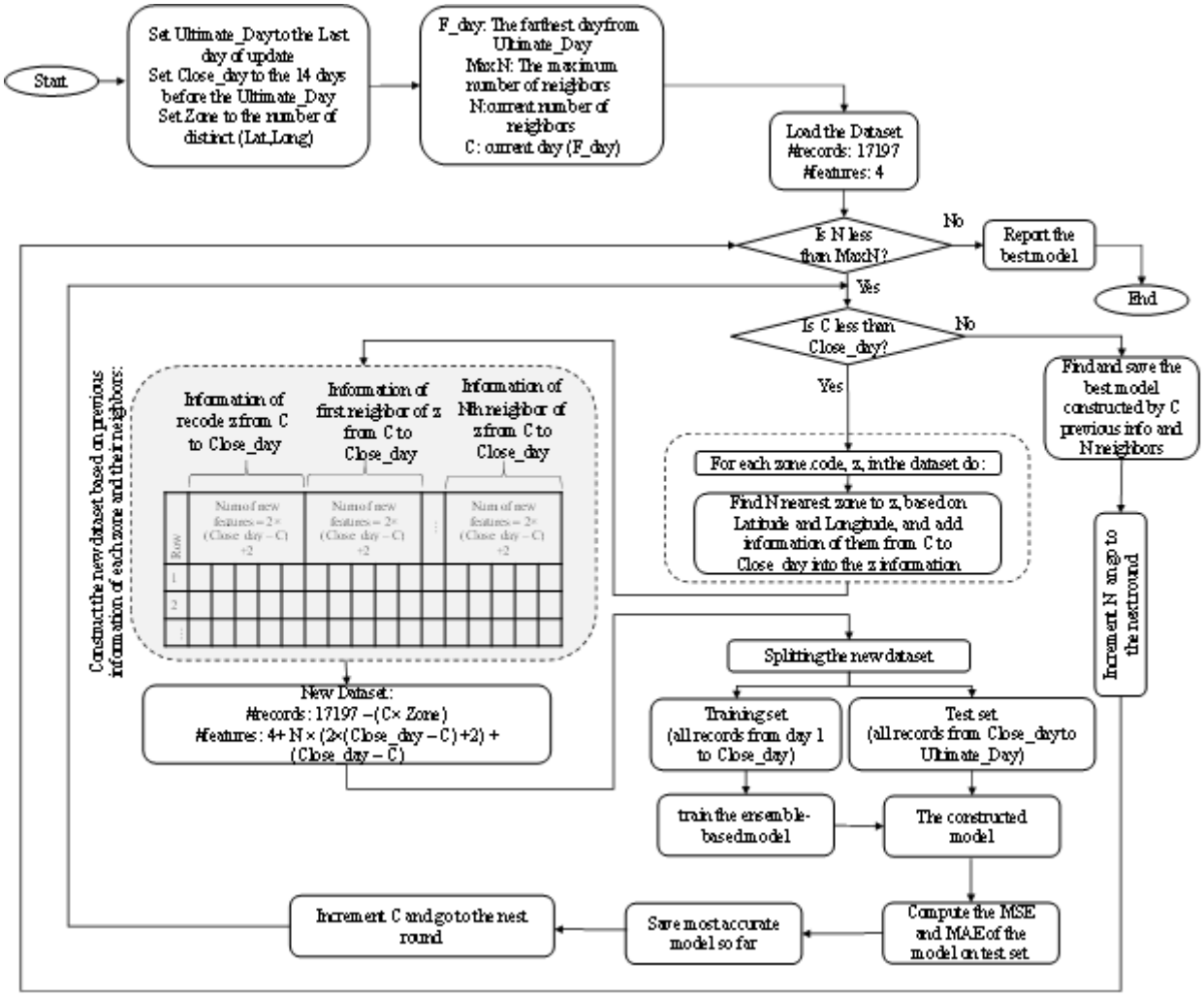


Figure 2

The structure of the Proposed model.

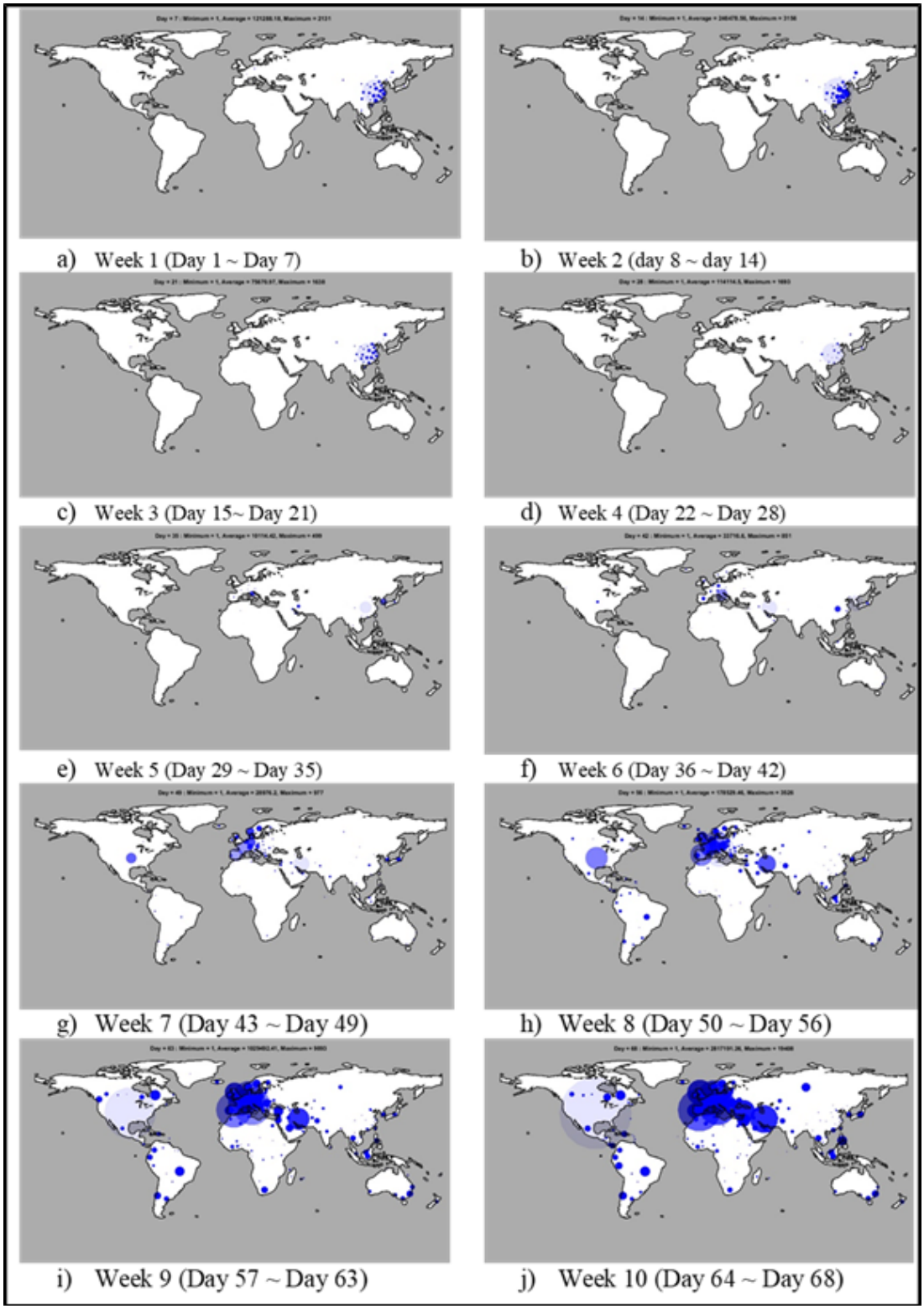
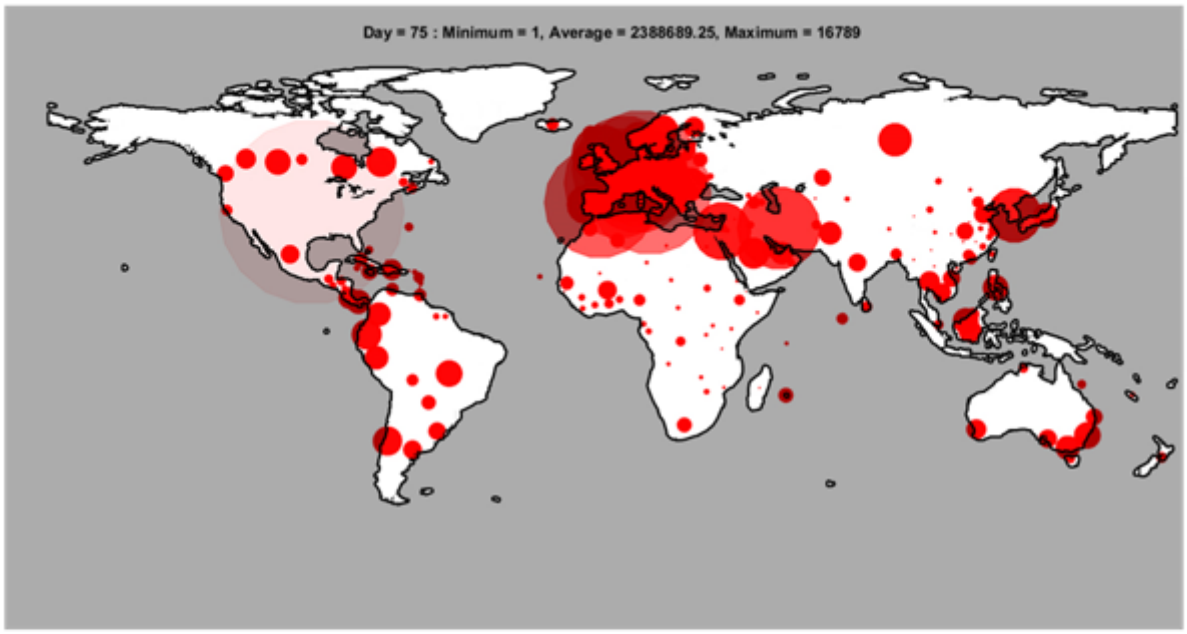
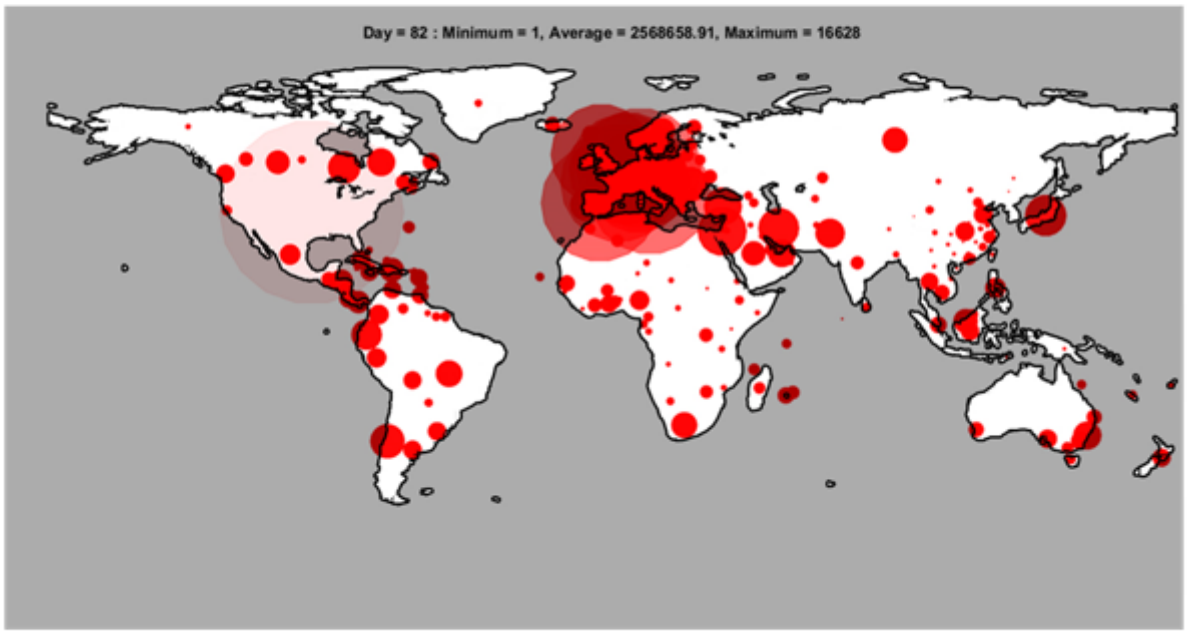


Figure 3

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.

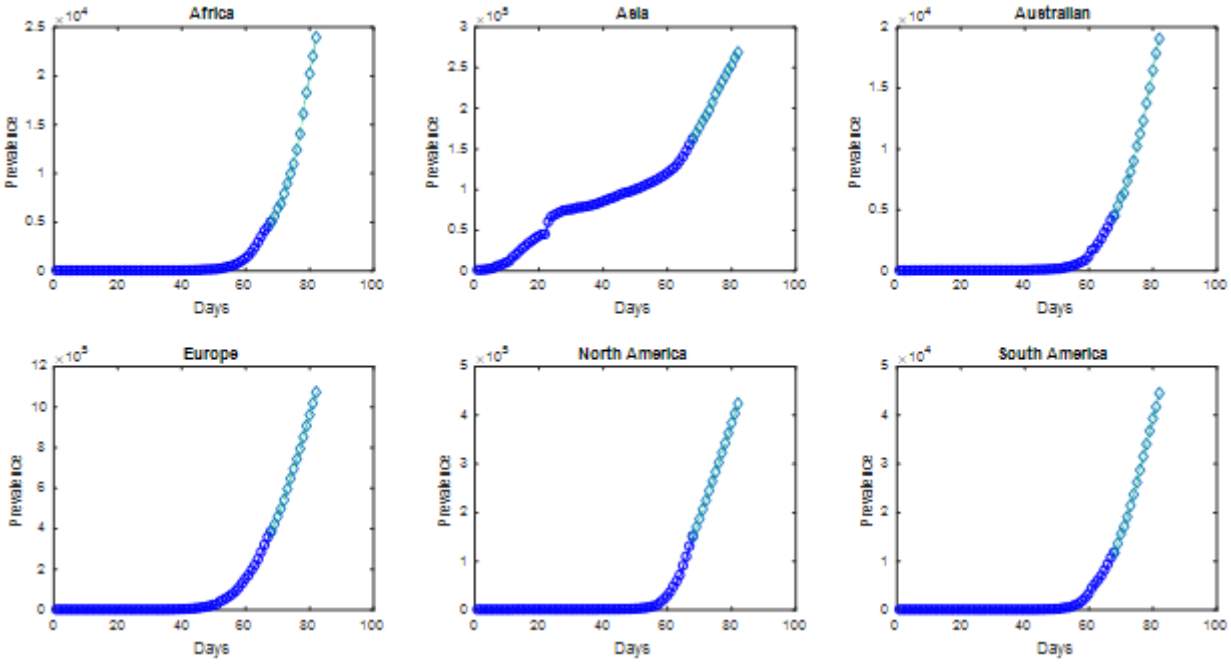


Figure 5

Plot of COVID-19 prevalence over days.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixB.pdf
- Appendix.pdf