Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14 (2020) 1231–1240

Contents lists available at ScienceDirect

Diabetes & Metabolic Syndrome: Clinical Research & Reviews

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

Data-driven modelling and prediction of COVID-19 infection in India and correlation analysis of the virus transmission with socio-economic factors

Amit Kumar [a], Poonam Rani [a,*], Rahul Kumar [b], Vasudha Sharma [c], Soumya Ranjan Purohit [d]

[a] Agricultural and Food Engineering Department, Indian Institute of Technology Kharagpur, West Bengal, 721302, India
[b] Vignan's Foundation for Science Technology and Research, Guntur, Andhra Pradesh, 522213, India
[c] Department of Food Technology, Jamia Hamdard, New Delhi, 110062, India
[d] Amity Institute of Food Technology, Amity University Uttar Pradesh, Noida, 201313, India

ARTICLE INFO

Article history:
Received 6 June 2020
Received in revised form
2 July 2020
Accepted 4 July 2020

Keywords:
COVID-19
Coronavirus
Statistical model
India
Principal component analysis
Correlation

ABSTRACT

Aims: The current study attempts to model the COVID-19 outbreak in India, USA, China, Japan, Italy, Iran, Canada and Germany. The interactions of coronavirus transmission with socio-economic factors in India using the multivariate approach were also investigated.
Methods: Actual cumulative infected population data from 15 February to May 15, 2020 was used for determination of parameters of a nested exponential statistical model, which were further employed for the prediction of infection. Correlation and Principal component analysis provided the relationships of coronavirus spread with socio-economic factors of different states of India using the Rstudio software.
Results: Cumulative infection and spreadability rate predicted by the model was in good agreement with the actual observed data for all countries (R^2 = 0.985121 to 0.999635, and MD = 1.2–7.76%) except Iran (R^2 = 0.996316, and MD = 18.38%). Currently, the infection rate in India follows an upward trajectory, while other countries show a downward trend. The model claims that India is likely to witness an increased spreading rate of COVID-19 in June and July. Moreover, the flattening of the cumulative infected population is expected to be obtained in October infecting more than 12 lakhs people. Indian states with higher population were more susceptible to virus infection.
Conclusions: A long-term prediction of cumulative cases, spreadability rate, pandemic peak of COVID-19 was made for India. Prediction provided by the model considering most recent data is useful for making appropriate interventions to deal with the rapidly emerging pandemic.

1. Introduction

The outbreak of novel coronavirus officially named as "COVID-19" emerged as a distressing large-scale pandemic and has become a threat to global health. The newly discovered coronavirus is scientifically classified as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. The coronavirus was first originated in December 2019, in Wuhan, the capital of Hubei Province in mainland China and further spread to several countries including USA, Italy, and India [2]. So far, as on 28 May, total positive cases of the world was 5,900, 907, out of which 3,61,549 patients died, and 2,577,250 have been recovered [3]. As on 28 May, India has already reported a cumulative count of 1,65,386 positive cases, comprising 89,755 active cases, and still witnessing a continuously growing spread [4].

The spread of coronavirus is not a perfect exponential spread, due to which classic epidemiological approach used to model the spread of infectious diseases fails to explain the COVID-19 [5]. A few studies have reported COVID- 19 cases forecast in China [6], Italy [7], Mexico [8], Italy, UK and USA [9]. Prediction in India was performed by Tomar and Gupta [10] using the long short-term memory (LSTM) and curve fitting, considering the dataset stretching

* Corresponding author.
E-mail addresses: amit.fpe@gmail.com (A. Kumar), poonam.fpe@iitkgp.ac.in, poonam.fpe@gmail.com (P. Rani), rk_ft@vignan.ac.in (R. Kumar), kotpalvasudha@gmail.com, vasudhakatwal@gmail.com (V. Sharma), srpurohit.iitkgp@gmail.com (S.R. Purohit).

https://doi.org/10.1016/j.dsx.2020.07.008
1871-4021/

from 30 January to April 4, 2020. Further, Salgotra et al. [11] made a prediction of COVID-19 pandemic in India for 10 days till May 25, 2020. Recently, Goswami et al. [12] also projected the COVID-19 outbreak in India till 21 May, 2020. India has witnessed a drastic rise in COVID-19 cases in May 2020, and consideration of latest data of May in model development would provide a more meaningful and realistic prediction for the coming months so that appropriate interventions can be made to deal with the emerging pandemic.

Virus transmission in society is considered to be affected by environmental parameters like temperature, wind speed, relative humidity as well as, socio-economic factors like population, people's behavior, governmental measures etc. Coccia [13] underlined a strong association of air pollution of cities with dynamics of COVID-19 transmission. Bashir, Bilal and Komal [14] observed a significant correlation of environmental pollutants like PM10, PM2.5, SO_2, NO_2, and CO with the COVID-19 transmission in California and suggested to analyze further the interaction of socio-economic factors with the virus outbreak.

There is a lack of adequate literature which urges for investigation of newly discovered infection in different countries of the world, and no study has forecasted the long term outbreak of COVID-19 in India, where the virus is rapidly emerging. Models developed are region/country-specific; there is a need for the development of a simple and versatile model that can explain the COVID-19 pandemic in different countries. Moreover, no study till date is focused on investigating the relationship of socio-economic factors with coronavirus spread in India. The broad objectives of the present study are to develop and validate a simple predictive model for the analysis of the COVID-19 pandemic in 8 countries and forecast the COVID-19 outbreak and spreadability in India. A mathematical model with only three parameters is proposed, which made it possible to correlate the pandemic spread with different socio-economic factors. Thus, interactions of coronavirus transmission with socio-economic factors of different states of India also has been attempted.

2. Materials and method

2.1. Data procurement

The confirmed COVID-19 data of India, China, USA, Germany, Japan, Iran, Italy, Canada were collected from the online platform of World Health Organization [3]. The data sets of coronavirus cases witnessed in different states of India were collected from the online platform of the Ministry of Health and Family Welfare, India [4]. The source of data of India state-wise population, Gender ratio, Rural-urban ratio, literacy, GDP contribution is CENSUS 2011, provided by the Ministry of Home Affairs, India [15].

2.2. Predictive mathematical model development

Similar to any bacteria or microorganisms, a virus spreading locally or globally undergoes three phases, i.e. lag phase, exponential phase and stationary phase and follows "S"-shaped curves depending upon the surrounding environmental conditions. The previous study has shown that coronavirus has not shown perfectly exponential growth [5]. Thus, in the present study, a nested exponential model, expressed in Eq. (1), was developed to define and predict the cumulative number of coronavirus cases against time.

$$Y_p = \alpha.expK.(1 - exp(-\beta.(t - t_i))) + Y_o \quad (1)$$

Where Y_p is the predicted cumulative infected population after time "t" (days), Y_o is the infected population at the first day, t_i is the inflection point (days), and K is the precautionary measurement constant. The model parameters α and β represent the inflection population, i.e., the infected population at the inflection point and a shape factor for the spreadability curve, respectively.

Thus the proposed model has three model parameters (α, β, t_i) and one model constants (Y_0 and K). The proposed model was fitted to the actual COVID-19 infected population (cases) data using Origin Pro.2019 software (OriginLab, USA). Initially, regression constants of the equation were obtained by considering $K = 1$ and then evaluated values of constants were used to predict the future trend of the outbreak. The future outbreak predictions can be adjusted by manipulating K value, and a higher K value indicates a higher number of infected people.

2.2.1. Terminologies used

Population (Y): It is the cumulative number of total infections or infected individual in the study zone.

Spreadability or spreadability rate (R_y): It is the rate of increase of a cumulative number of infected populations and is expressed as the total number of infections per unit time (infections/day). Mathematically, it is defined as the first derivative of the cumulative population with time, given by Eq. (2).

$$R_y = \frac{dY_p}{dt} = Y_p \times \beta \times [exp\{-\beta.(t - t_i)\}] \quad (2)$$

Time (t): Time is the duration or the total number of days after first confirmed infection in the study zone. In this study, the first confirmed infection was considered to be on February 15, 2020 for all countries, except China (January 22, 2020) and Iran (February 19, 2020).

Inflection time (t_i): It is the instance of time when the spreadability rate reaches its maximum value, or the rate of spreadability attains zero value.

Lag time (t_l): This is the time duration between the first observed case and the time when spreadability become significant. The data during the lag phase can be omitted during predictive modelling, and the term (t_l) must be incorporated by replacing time "t" with ($t + t_l$) in the model.

2.2.2. Assumptions

1. Spreadability rate (cases/day) is assumed to increase up to a certain time to attain maximum value, and then it follows a reducing trend.
2. The cumulative population in the study zone follows an "S"-shaped curve that passes through an inflection point of time where the spreadability rate will be maximum.

2.3. Principal component analysis

The phenomenon of virus spread in society differs from microbial growth due to various influencing factors like the nature of the virus, climatic condition, geographical factors, socio-economic factors, population behavior, etc. Each of these factors may have some possible impact on pandemic spread, thus, needs consideration during pandemic assessment. Principal component analysis (PCA), a data dimension reduction technique, was employed to understand the possible contribution of different socio-economic variations in different states of India. The state-wise socio-economic factors considered for assessing the COVID-19 spread in India were COVID-19 confirmed cases (Cases), populations (population), average literacy rate (Literacy), percent contribution of gross domestic production (GDP), rural to urban population ratio (Rural/Urban) and Gender ratio (W/M). The correlation and

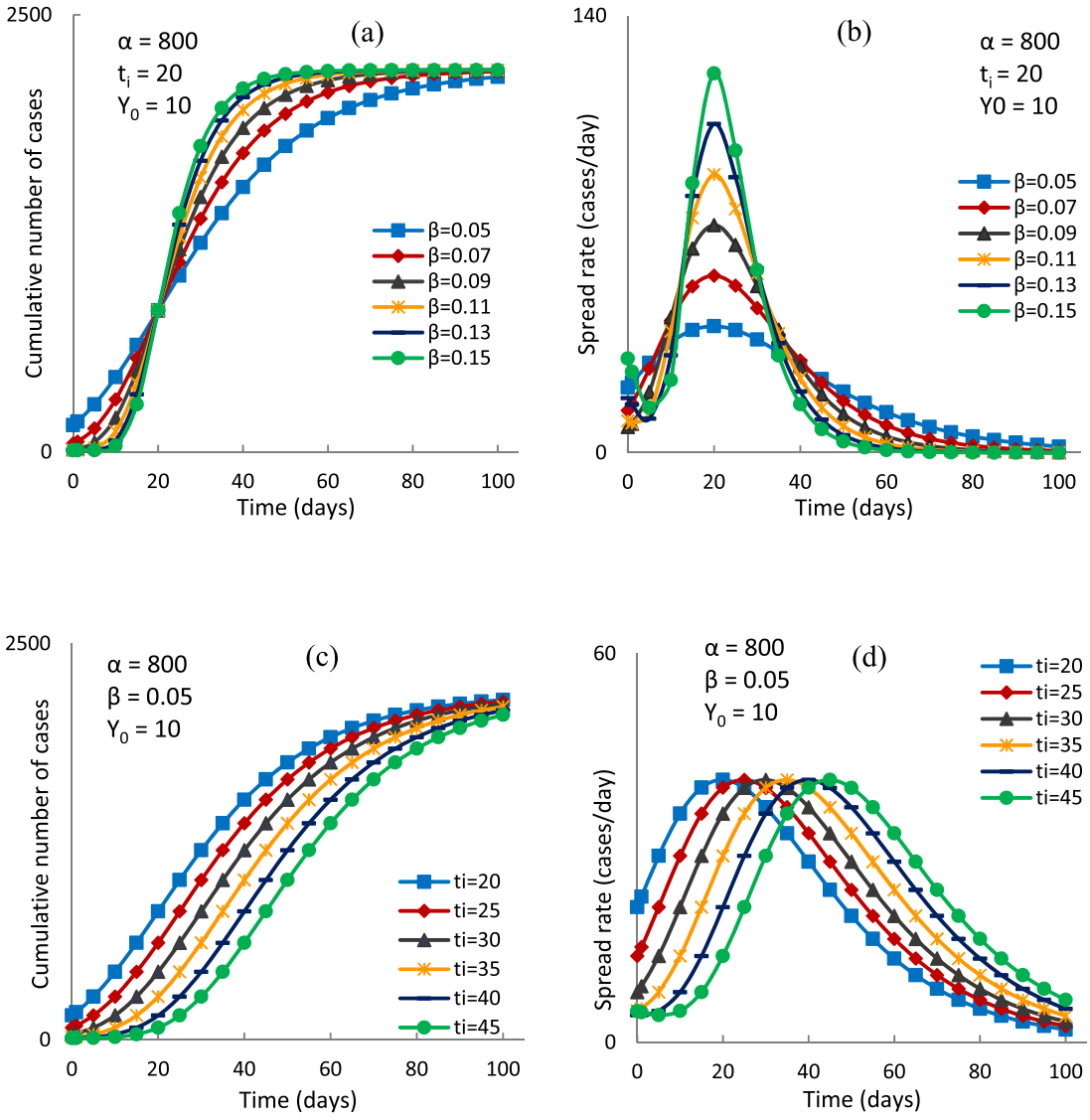

Fig. 1. Variation in shape of the model curve (a) effect of change in spreadability constant (β) on cumulative infected population, (b) effect of change in spreadability constant (β) on spreadability rate, (c) effect of change in inflection time (t_i) on cumulative infected population, (d) effect of change in inflection time (t_i) on spreadability rate.

principal component analysis were performed using the Rstudio software, and associations among the socio-economic factors and COVID-19 for different Indian states were summarised using a biplot.

3. Results and discussion

An attempt was made to establish an empirical mathematical model based upon the already recorded confirmed COVID-19 cases from mid-February to May. The developed model has been used to forecast the pandemic spread for coming days and approximate estimation of spreadability. In the general form, the model function is a nested exponential function, containing only three parameters, which makes the model quite simple and versatile. Since only a few parameters define the developed model, these were accounted during the examination of correlation of infection outbreak with socio-economic factors. The virus infected cumulative population forms an S-shaped curve with time (days) of spread depending upon the model parameters which can be fitted with cumulative population data for almost all the states or countries worldwide.

3.1. Effect of model parameters on the shape of the curve

Before analyzing the actual results, efforts were undertaken to analyze the effect of model parameters on the shape of infection curves. The variation in the shape of prediction curves as a result of changes in the model parameters, β and t_i, is presented in Fig. 1, considering constant initial population, $Y_0 = 10$ and inflection population, $\alpha = 800$. Fig. 1a and b depicts the impact of the model parameter (β) on the cumulative population and spreadability rate, respectively. It was identified that change in β considerably influenced the shape of the curve for the cumulative population as well as spreadability rate, therefore, defining β as "shape factor" for the model. Increase of β from 0.05 to 0.15 caused a faster and abrupt rise of cumulative population. The curve with a higher β value exhibited higher steepness as a consequence of an early and sudden rise in the infected population. Further, the increase of shape factor also demonstrated an intensified and well-defined peak of the spreading rate, as presented in Fig. 1b. It can be inferred that the higher value of β is a representative of rapid enlargement of COVID-19 outbreak.

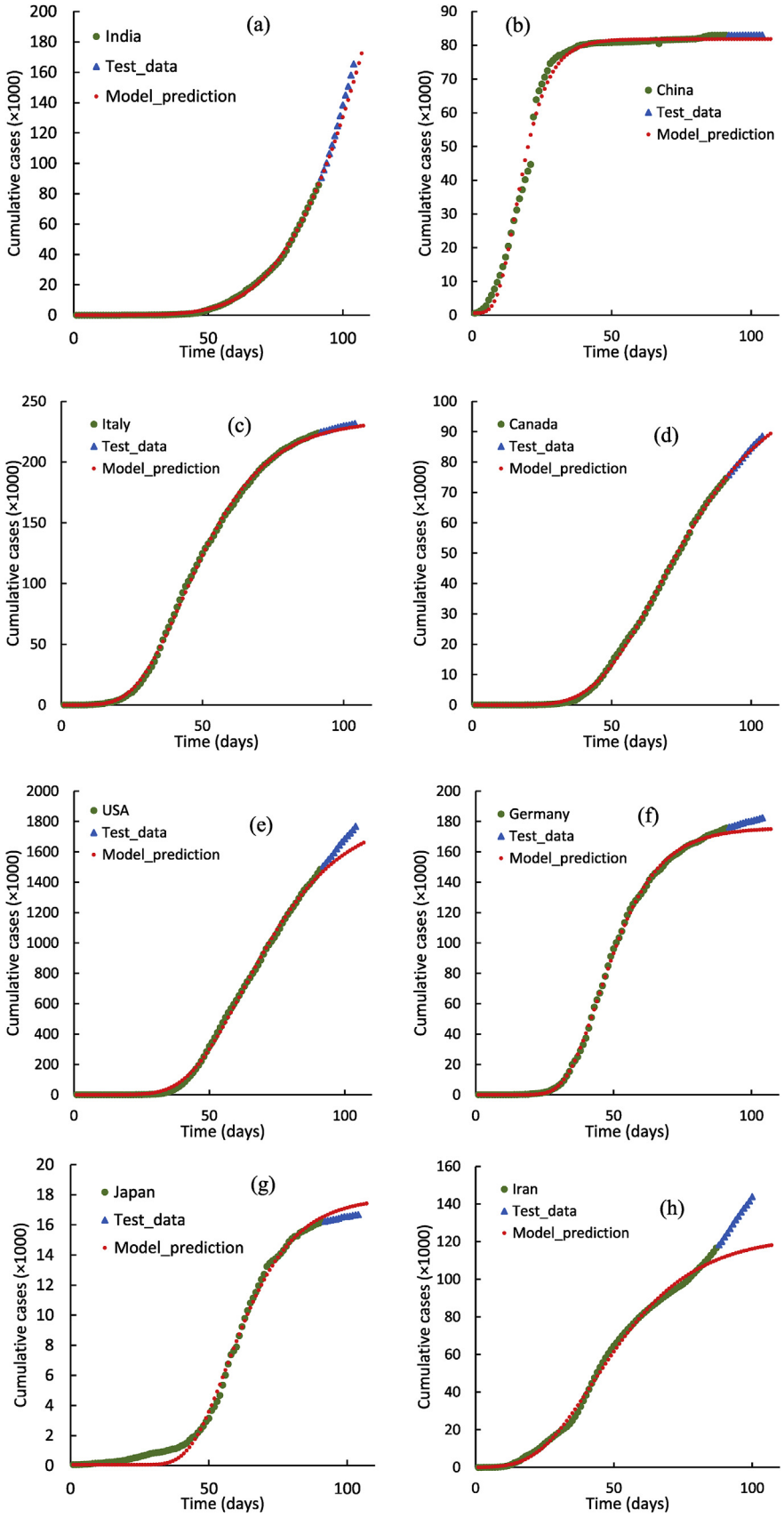

Fig. 2. Actual and model predicted cumulative infected population for (a) India, (b) China, (c) Italy, (d) Canada, (e) USA, (f) Germany, (g) Japan, (h) Iran, tested data: (15.05.2020–28.05.2020).

Table 1
Statistical model parameters for different countries obtained by regression modelling considering the data up to 15 May, 2020.

Countries	Starting date	Model parameters				
		Y_0	α	β	t_i	R^2
India	February 15, 2020	3	483417	0.018357	145.59	0.9994
China	January 22, 2020	571	29892.41	0.152632	15.47	0.9937
Italy	February 19, 2020	3	86335.43	0.060355	42.73	0.9995
Canada	February 15, 2025	8	39323.2	0.043275	67.23	0.9995
USA	February 15, 2020	15	682228.7	0.049143	62.22	0.9992
Germany	February 15, 2021	25	64701.81	0.083874	44.59	0.9996
Japan	February 15, 2022	53	6540.975	0.074024	56.51	0.9956
Iran	February 15, 2023	2	45338.53	0.049268	42.58	0.9972

Where, Y_0: Infected population at zero day; α: Cumulative infected population at inflection point; β: Shape factor; t_i: Inflection time (days).

Table 2
Validation of the developed COVID-19 model for different countries.

Country	R^2 value	MD (%)
India	0.999635	7.07
China	0.985121	1.41
USA	0.997118	7.76
Germany	0.987608	4.24
Japan	0.987081	3.77
Iran	0.996316	18.38
Italy	0.996885	1.2
Canada	0.998726	1.48

R^2: Degree of goodness of predicted value with actual test data from 16 May to 28 May 2020; MD: Maximum percent deviation of predicted data from actual test data (%).

Effect of inflection time (t_i) on the shape of the infection curves is shown in Fig. 1c and d. It is evident that on increasing of the inflection time, shape of the cumulative population curve as well as spreadability rate curve almost remains the same; however, the whole curve makes a shift along the time axis. Therefore, inflection time did not manifest considerable effect on the shape of the curve, but substantial changes in the position of the curve were distinguished. It is noteworthy to highlight that model function can provide different types of shapes depending upon the model parameters, shape factor and inflection time. The developed nested exponential model has the potential to define and predict ever-increasing trend like cumulative infected population (Y_p), and bell-shaped curves governing spreadability rate model (R_y).

3.2. Analysis and model validation for COVID-19 outbreak

The actual data of period 15 February to May 15, 2020 of different countries were used for parameter estimation of the model. Subsequently, the actual data observed during the period from 16 May to May 28, 2020, indicated by blue triangles in Fig. 2, was applied as the test data for validation of the mathematical model. The values of the model parameters obtained for different countries using regression modelling are presented in Table 1.

3.2.1. Cumulative infected population

The developed model was used to predict the COVID-19 cumulative infected population for countries India, China, USA, Germany, Japan, Iran, Italy, and Canada. Fig. 2 presents the actual and model-predicted cumulative population of infected people for the eight countries. It is implied from Fig. 2a that; currently, coronavirus spread in India is emerging with a fast-growing upward trajectory, while Canada (Fig. 2d) and USA (Fig. 2e) are now presenting depreciation of outbreak and appears to move toward flattening of the curve. On the other hand, the infection in China (Fig. 2b) demonstrates a characteristic S-shape and finally flattened curve indicating attenuation of a pandemic. Moreover, Italy, Germany, Japan (Fig. 2c, f, g) tend to form an S shape curve, which signals the fading of the pandemic in these countries. The outbreak appears to have begun later in India as compared to other countries. Till February 15, 2020, India had witnessed only 3 confirmed cases, which increased to 1397 on March 31, 2020. The slower pace of outbreak, as observed in March and April, can be attributed to the total lockdown in the country in conjunction with a lower rate of infection testing. This provided more time to India for preparation and expansion of the healthcare system and medical facilities in order to deal with the pandemic. As on April 30, 2020, total confirmed cases of coronavirus patients were reported to be 34,863. Thereafter in May, India witnessed a visible abrupt expansion in the outbreak and the total number of confirmed cases reached 1,65,386 on May 28, 2020. The significant rise of spread reported in May may be attributed to the increased rate of testing along with the relaxations given in the lockdown leading to movement of travellers in different parts of the country.

The suggested nested exponential mathematical model proved to be efficient by explaining the COVID-19 outbreak with a high goodness of fit (>0.99) in all countries except Iran (Fig. 2h). The maximum percent deviation of predicted data from actual test data for the period 15 May to May 28, 2020 for different countries is tabulated in Table 2. Remarkably, the model was fitted best to India, China, Italy, Canada, and the USA with the goodness of fit greater than 0.999 and percent deviation of the predicted data from actual values varied from 1.2% to 7.76%. This presented the ability of the model for an accurate approximation of future coronavirus outbreak with a slight deviation from actual values. Apart from this, the model was able to fit actual data of different countries exhibited by different shapes of curves of the infection population, with great accuracy. Japan showed a significant deviation of model-predicted values from the actual values for the test data during the early phase of virus spread, which might be because of the constant population of infected patients observed during the onset phase of virus spread. This initial stagnant stretch might be due to low testing rate during the early phase of the outbreak and can be termed as the lag period of the virus. It is expected that omitting of some initial data from this lag period can define the COVID-19 outbreak more accurately along with improving the future approximation in countries like Japan, USA and Germany, where the model predicted values showed a deviation from actual values. The shape factor (β) obtained for curves of different countries are enlisted in Table 1. The highest value of the shape factor, 0.152632, was observed to be of China, which clearly presents a sudden abrupt rise in the cumulative cases of infections, as compared to other countries. This was followed by Germany, Japan, Italy, USA, and Canada. Among all the countries, India showed the lowest value of shape factor, i.e., 0.018357, which is still at a rapidly increasing pace of cumulative cases.

3.2.2. Spreadability rate

The change in spreadability rate expressed as a number of cases per day noted for different countries is presented in Fig. 3. The model predicted values (indicated by the red circle) were obtained by putting model parameters (Table 1) in spreadability rate model (R_y). The predicted values (indicated by the red circle) were followed for several days from 16 May to May 28, 2020, and it was found that these values were close to the actual observed values for all the countries except Iran. Therefore, it can be stated that the model developed in the present study presented a good approximation of the spread rate also. India reported 7300 new confirmed cases on 28 May. Fig. 3a demonstrates an intensification of infection rate in India for the current situation. The coronavirus disease was

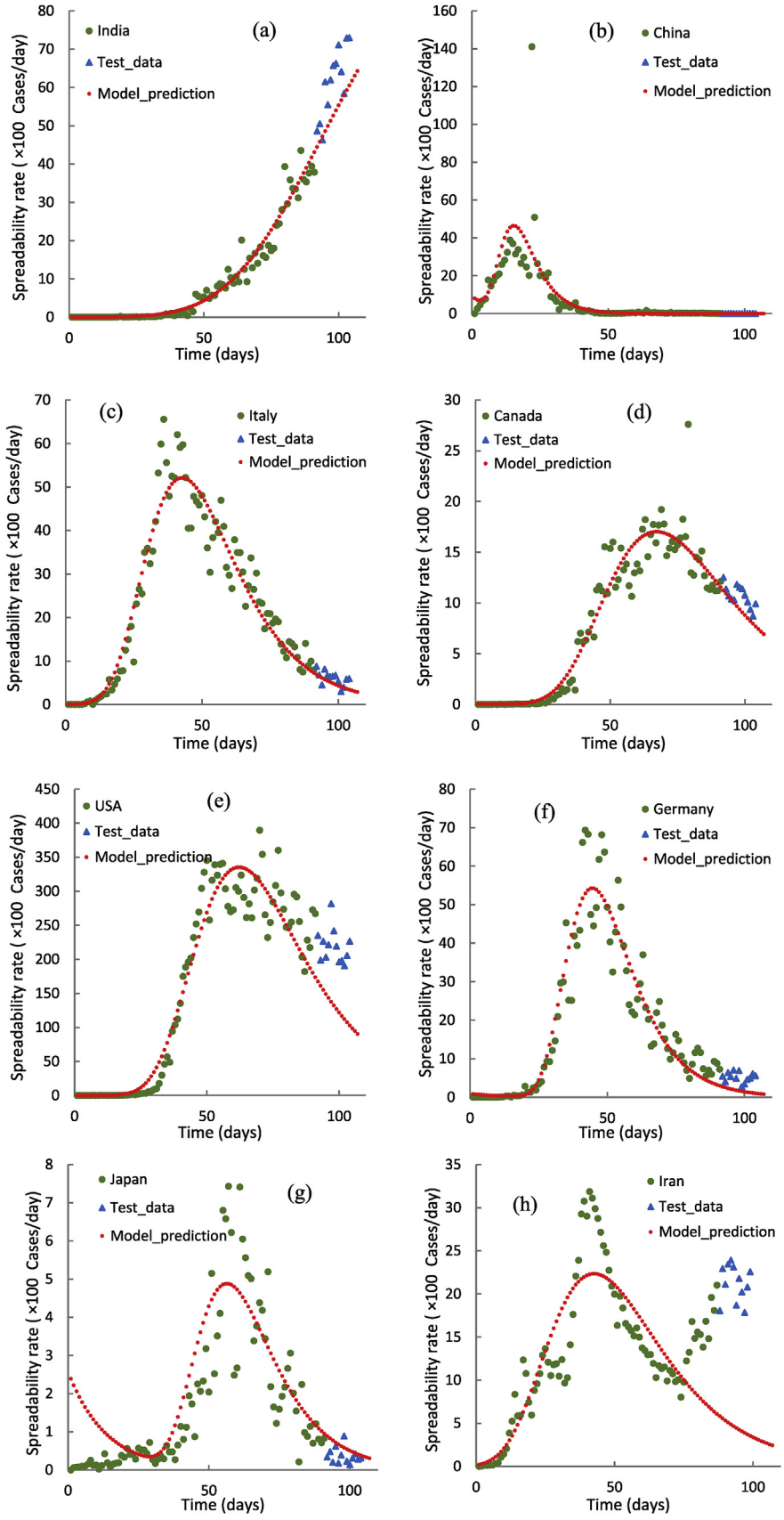

Fig. 3. Actual and predicted COVID-19 spreadability rate for (a) India, (b) China, (c) Italy, (d) Canada, (e) USA, (f) Germany, (g) Japan, (h) Iran, tested data: (15.05.2020–28.05.2020).

firstly originated in China (Fig. 3b), where the peak of spreadability rate was detected around 4 February, after which diminishing of infection rate was identified which is in good approximation with predicted inflection time of 15 days for China. The values of the inflection time (t_i) and the cumulative infected population (α) at the inflection for different countries as predicted by the model are enlisted in Table 1. The inflection time predicted by the model for USA (Fig. 3e) was 63 days from the start of the initial point of the model, i.e. 15 February 2020, which implies the value of maximum spreadability on April 18, 2020. The actual peak spreadability for the USA was identified around 24 April when the cumulative infected population was 925232 [3]. Currently, the infection rate of the virus in the USA is manifesting a declining trend. Based upon the inflection time predicted by the model, the peak of the spreadability for Italy, Canada, Germany and Japan (Fig. 3d, e, g, and h), should be around 1 April, 23 April, 31 March, 12 April, respectively. Actual data demonstrated infection rate peak for these mentioned countries around 22 March, 3 May, 27 March, 11 April, respectively, which was found to be very close to corresponding predicted model values. Since these countries had already witnessed the peak spreadability, therefore, currently, spreadability is dropping with time. Surprisingly, Iran (Fig. 3h) presented an exceptional case of spreadability curve as it had started to present slowing down of spread after attaining a peak, but the outbreak was accelerated second time rendering an increased number of cases in the country leading to the upward trajectory of spreadability rate. Therefore, the present model failed to explain this observation of the second time expansion in the spread rate of the virus. It is worth mentioning that spreadability rate predicted by the model was observed to be in good agreement with the actual observed data (except for Iran) and can be considered as an additional validation of the proposed model. The model further enforced for the approximation of the spreading rate in India, where coronavirus outbreak has not yet attained inflection point.

3.3. Future prediction of COVID-19 outbreak in India

The spread in India has appeared to be delayed with respect to other counties; hence, the spread rate of the virus is expected to be augmented in coming days, as visualized from increasing predicted trend shown in Fig. 3a. The long-term prediction of COVID-19 spread for India, considering a confidence interval of ±5% of inflection population (α) is presented in Fig. 4. Accounting the data obtained till date, the increasing curve of the cumulative population of infected patients is likely to move towards flattening possibly in October (Fig. 4a). Spreadability rate of infection in India may rise more than 9000 new cases per day in June. According to prevailing conditions of May, the inflection time for India is estimated to be 145 days from February 15, 2020, indicating the peak of the spreadability to be attained in July. However, it is likely to change due to increase of virus spreadability rate and ease on lockdown restrictions in June. The cumulative number of confirmed infected population at inflection is approximated to be more than 5 lakhs. Therefore, it is predicted that in June and July, India is expected to experience an increase in the outbreak of COVID-19 till it attains peak spreading rate. Moreover, total coronavirus cases may cross 12 lakhs till the flattening of the outbreak is attained. The effectiveness of lockdown, governmental policies and other socio-economic factors may affect the virus spread, and actual values may deviate from the model predicted values. It is noteworthy to mention that presented forecast is based on the present situation and may be affected in future with a change of government policies in the direction and other socio-economic factors. Therefore, it is suggested that revision of forecast in coming months with consideration of updated data should be carried out for better approximation. A revision of model parameters was attempted by considering the actual data from 15 February to May 28, 2020. Model parameters and prediction of cumulative infected population and spreadability rate for June for all the 8 countries are provided in Appendix A. Since currently, India is undergoing a steady rise in infection rate, measures like increased testing rate, robust virus tracing and strict enforcement of social distancing and restricted social gathering are crucial steps toward attaining downward trajectory of the spread rate.

Despite the strong fitting of predicted data with actual observed data, the model has some limitations. It does not account the effect of influential parameters and control strategies like lockdown, social distancing, testing, tracing, personal hygiene, socio-economic

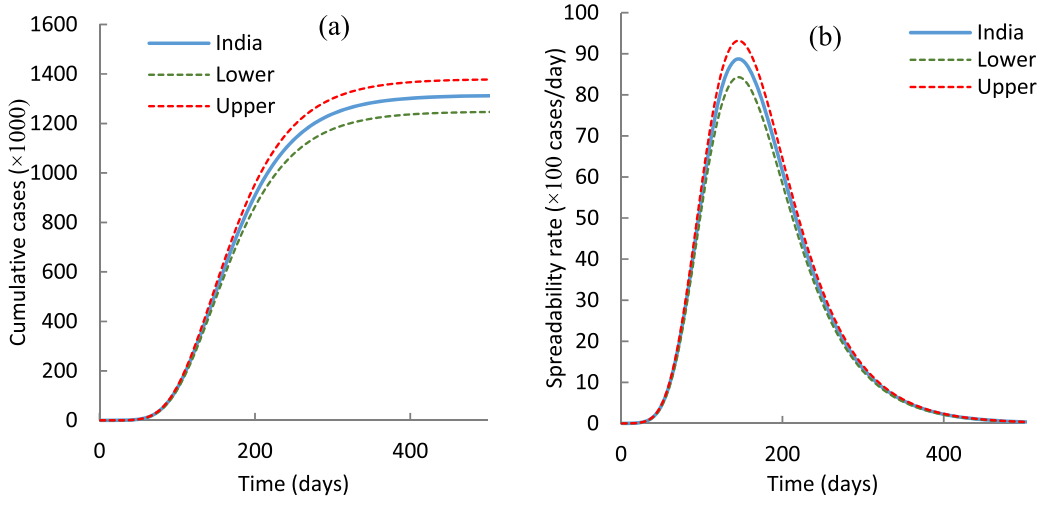

Fig. 4. Long term prediction for India up to 500 days (29 June 2021) (a) Cumulative infected population, (b) Spreadability rate, where α is the cumulative infected population at inflection point.

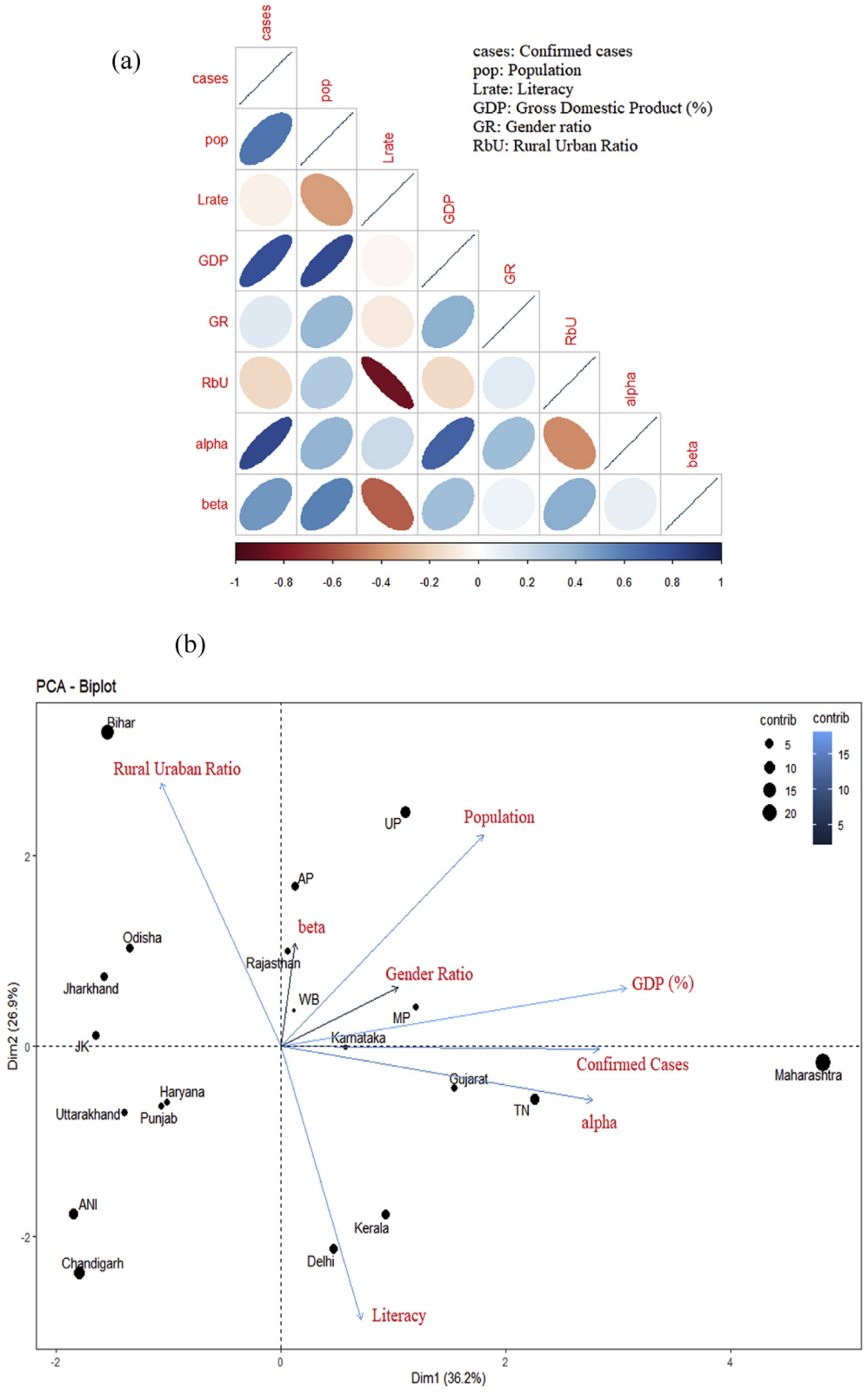

Fig. 5. (a) Pairwise spearman correlation analysis among confirmed coronavirus cases and different socio-economic factors, (b) Biplot showing scores of different states and loadings of different socio-economic factors on first two principal components.

factors which considerably affect the spread of infectious coronavirus disease. Moreover, the model did not fit well to analyze the second wave of infection, as observed in the case of Iran.

3.4. Principal component analysis for states of India

Principal component analysis (PCA) was done to understand the possible contribution of different socio-economic variation in different Indian states. Fig. 5 a elucidates the pairwise Spearman correlation among different socio-economic factors, model parameters and coronavirus outbreak. The cumulative confirmed cases had a strong direct relationship with the model parameters inflection population (α) and shape factor (β). Apart from this, the infected population was also found to be strongly positive with state GDP contribution and population. The direct relation of population density with spreading of coronavirus disease was also observed in Iran [16]. The inflection population also presented a positive correlation with population and GDP(%), but the relationship between inflection population and GDP contribution was observed to be stronger, indicating state with higher GDP contribution witnessed a severe outbreak of COVID-19. Further, the shape factor was observed to be negatively correlated with literacy rate, which implies that states having higher literacy rate attained lower shape factor. The literacy rate presented an inverse association with rural to urban ratio, which also negatively correlated with inflection population. It can be suggested that states with lower rural to urban ratio witnessed a higher spread of coronavirus disease.

The data were used in non-normalized form for comparative visualization of the contribution of each factor in confirmed cases. The whole data set formed a relatively spherical data set, and a total of 5 principal components (PC) were found significant (Eigenvalue > 1.0). The data has huge variation due to geographical variation of different states, and the first two PC's were able to explain 54.6% of the total variance in the data. From the PCA biplot (Fig. 5b), it can be inferred that states like Maharashtra, Tamil Nadu, Gujarat contributing higher percentages of GDP were more susceptible to coronavirus infection. High GDP contribution can be related to higher economic and commercial activities leading to increase of urbanization and international connectivity. Maharashtra is the worst affected state by the outbreak of coronavirus, which could be justified with its high industrialization contributing higher GDP and larger urban population indicating higher international connectivities and movement. Further, states like Jharkhand, Bihar, Odisha, Uttarakhand, Haryana, Punjab were found to be lying far away from the GDP contribution line and close to lower rural to urban ratio, thus, presented a lesser number of cases of virus infection.

The developed model is simple and contains only three important model parameters, which were easily considered for analyzing the relationship with socio-economic factors. This is the additional advantage of the model in addition to accurate approximation and one of the original point of the present study. The insights and forecast of COVID-19 outbreak for India as explained by the statistical model are expected to be helpful for higher authorities in preparing future plans and robust policies to face the outbreak as well as to reduce the virus spread. Considering the pandemic rise in India in coming months as predicted by the model, it is anticipated that expansion of medical facilities, quarantine centres, closure of schools, social gathering places are crucial intervention to be followed.

4. Conclusion

A nested exponential model showing S-shaped curve was developed and validated using the actual data to define and predict the cumulative infected population and spreadability against time for countries like India, USA, China, Japan, Italy, Iran, Canada and Germany. The developed statistical model contained only three model parameters and was found to be best fitted to actual data of countries except, Iran. The higher value of the model parameter, β indicates higher spreadability and more abrupt rise in the virus outbreak, while, inflection time caused substantial changes in the position of the curve. Currently, coronavirus spread in India is following a fast-growing upward trajectory, while USA and Canada appear to move towards flattening of the curve. Besides, the virus outbreak in China showed a definite S-shaped flattened curve indicating a pandemic attenuation, and Germany, Japan, Italy tends to form an S-shape curve, signaling favourable movement toward attenuation. The model predicted COVID-19 spreading in India is likely to increase in June and July till the pandemic attains the peak spreadability. Further, correlation and principal component analysis were used to examine the relationships among the coronavirus spreading and socio-economic factors of different states of India. The cumulative confirmed cases were strongly positively correlated with the population at inflection point and shape factor, in addition to state GDP contribution and population. Findings revealed that states like Maharashtra, Tamil Nadu, Gujarat contributing higher percentages of GDP, witnessed a more severe outbreak of coronavirus. The forecast made by the model is based upon current prevailing situation and data, which may be influenced by the government's decisions and people's response. Therefore, it is suggested that revision of forecast in the coming months with consideration of updated data should be carried out for better approximation of the outbreak. It is anticipated that the findings of the present study will be beneficial in making the best decisions to curb the COVID-19 spread.

Declaration of competing interest

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dsx.2020.07.008.

References

[1] Bansal M. Cardiovascular disease and COVID-19. Diabetes Metab Syndr Clin Res Rev 2020;14:247–50. https://doi.org/10.1016/j.dsx.2020.03.013.
[2] Ghosal S, Sengupta S, Majumder M, Sinha B. Prediction of the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020). Diabetes Metab Syndr Clin Res Rev 2020;14:311–5. https://doi.org/10.1016/j.dsx.2020.03.017.
[3] WHO. World Health Organization coronavirus disease (COVID-19). https://covid19.who.int/; 2020. accessed May 23, 2020.
[4] MoHFW. Ministry of Health and Family Welfare. Government of India; 2020. https://www.mohfw.gov.
[5] Ziff AL, Ziff RM. Fractal kinetics of COVID-19 pandemic (with update 3/1/20). medRxiv 2020:1–15. https://doi.org/10.1101/2020.02.16.20023820.
[6] Sun T, Wang Y. Modeling COVID-19 epidemic in Heilongjiang Province, China. Chaos, Solitons Fractals 2020. https://doi.org/10.1016/j.chaos.2020.109949.
[7] Giordano G, Blanchini F, Bruno R, Colaneri P, Di Filippo A, Di Matteo A, et al. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. 2020. https://doi.org/10.1038/s41591-020-0883-7.
[8] Torrealba-Rodrigueza O, Conde-Gutiérrezb RA, Hernández-Javierc AL. Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models. Chaos, Solit Fractals 2020. https://doi.org/10.1016/j.chom.2020.04.004.
[9] Utkucan Ş, Tezcan Ş. Forecasting the cumulative number of confirmed cases of COVID-19 in Italy, UK and USA using fractional nonlinear grey Bernoulli model. Chaos, Solit Fractals 2020. https://doi.org/10.1016/j.chaos.2020.109948.
[10] Tomar A, Gupta N. Prediction for the spread of COVID-19 in India and

effectiveness of preventive measures. Sci Total Environ 2020;728:138762. https://doi.org/10.1016/j.scitotenv.2020.138762.
[11] Salgotra R, Gandomi M, Gandomi AH. Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming. Chaos, Solit Fractals 2020. https://doi.org/10.1016/j.chaos.2020.109945.
[12] Goswami K, Bharali S, Hazarika J. Projections for COVID-19 pandemic in India and effect of temperature and humidity. Diabetes Metab Syndr Clin Res Rev 2020. https://doi.org/10.1016/j.dsx.2020.05.045.
[13] Coccia M. Factors determining the diffusion of COVID-19 and suggested strategy to prevent future accelerated viral infectivity similar to COVID. Sci Total Environ 2020;729:138474. https://doi.org/10.1016/j.scitotenv.2020.138474.
[14] Bashir MF, Bilal BM, Komal B. Correlation between environmental pollution indicators and COVID-19 pandemic: a brief study in Californian context. Environ Res 2020;187:109652. https://doi.org/10.1016/j.envres.2020.109652.
[15] CENSUS. Ministry of Home Affairs. Government of India; 2011.
[16] Ahmadi M, Sharifi A, Dorosti S, Jafarzadeh Ghoushchi S, Ghanbari N. Investigation of effective climatology parameters on COVID-19 outbreak in Iran. Sci Total Environ 2020;729:138705. https://doi.org/10.1016/j.scitotenv.2020.138705.