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.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
A Statistical Non-Parametric data analysis for COVID-19 incidence data

R.I. Minu b, G. Nagarajan a, *

a Sathyabama Institute of Science and Technology, India
b SRM Institute of Science and Technology, India

ABSTRACT

Background: The impact of COVID-19 on the Global scale is tremendously drastic. There are several types of research going on across the world simultaneously to understand and overcome this dire pandemic outbreak. This paper is purely a statistical study on a distinct set of datasets regarding COVID-19 in India. The motivation of this study is to provide an insight into the rapid growth of confirmed COVID-19 cases in India.

Methods: The rapid growth of COVID-19 cases in India started in March 2020. The main objective of this paper is to provide a solid statistical model for the policymaker to handle this kind of pandemic situation in the near future with nonlinear data. In this paper, the data was got from 1st April to 29th November 2020. To come up with a solid statistical model, various nonlinear data such as confirmed COVID-19 cases, maximum temperature, minimum temperature, the total population (state-wise), the total area in km2 (state-wise), and the total rural and urban population count (state-wise) have been analyzed. In this paper, six different Generalized Additive Models (GAM) was identified after a thorough analysis of other researchers’ (Xie and Zhu, 2020; Prata et al., 2020) findings.

Results: In all perspectives, the results were identified and analyzed. The GAM model regarding total COVID-19 confirmed cases, total population, and the total rural population provides the best average fit of R2 value of 0.934. As the population value is quite high, the author has concise it using logarithm to provide the best p-value of 0.000542 and 0.001407 for a relation between the total number of COVID-19 cases regarding the total population and total rural population respectively.

© 2022 ISA. Published by Elsevier Ltd. All rights reserved.

1. Introduction

COVID-19 had converted the view of the perspective of human beings. Starting from the initial outbreak [1] till now there are over 500 research papers [2] published by researchers to provide an insight regarding the COVID-19 pandemic from their perspectives. Xie & Zhu [3] have claimed in their paper published on March 30th, 2020, that there is a definite impact of the temperature rise on the confirmed COVID-19 cases. Based on the Chinese datasets, their findings suggest that there might be a relationship between the temperature and the confirmed cases of COVID-19. Prata et al. [4] carried out the statistical analysis with the Brazilian dataset and declared that there is no such evidence at all their paper was published on April 25th, 2020. This paper provides a statistical overview of finding the relation between the population, an area in km2, an average temperature, and the number of confirmed COVID-19 cases in India. In India, even though the first case of COVID-19 was reported on 30th January 2020, the number of cases increased around Mid-March 2020.

India being a densely populated country, in order to control the spread of this contagious virus, the country went into lockdown. The paper is organized in such a way in Section 2, the identified study area, the source of dataset collection, and the general analysis of the raw data are discussed. The outcome of Section 2 is to justify the non-linearity of the dataset and the reason to go for non-parametric analysis. In Section 3, a detailed analysis of GAM, the proposal of 6 different GAM models, and their results are discussed.

2. Study area

In this statistical analysis, the study area in India, the second most highly populated country in the world. India has a total area of 3,287,263 km2 and a density of 406.3/km2 [5]. India's temperature is influenced by various factors due to its localization between the hills on the north and oceans on three sides [6]. India has 28 states and 8 Union territories. For this statistical study, 32 states (including a few union territories) where the major COVID-19 cases were reported were considered. With its diversity, India comprises six different types of climatic systems as per the Koppen system. Basically, from March till May, it would be the summer season. Even though the first outbreak
of COVID-19 was reported on 30 January 2020, the number of cases gradually increased from March 2020 [7] when the summer season starts in the country.

3. Data collection

To derive a statistical model to understand this pandemic situation, in this analysis three kinds of data set collections were collected and bound as per requirement. The first set of datasets is the population study based on the Indian census data [8], which is collected once in 10 years. From that dataset, the required data for the study — the total population, urban population, and the total area were collected. These data for 32 states were extracted as one dataset (named the population). The second set of data is the daily COVID-19 confirmed cases from each state — these live data were fetched from the Ministry of Health and Family Welfare - Government of India and The Karnataka Government Media Bulletins [9]. As the pandemic started to surge higher from March, we have considered the data from 1st April 2020 till 29th November 2020. From the source dataset, we extracted only the Total Case of the Indian Nation (TCIN) (named Covid). The last set of datasets was required to the record daily maximum and a minimum average temperature of the States. These data were obtained from the Indian Meteorological Department and AccuWeather websites (named tempmax and tempmin).

4. Dataset analysis

In this paper, three different kinds of data sets are used as explained in data collection. Among them, Covid, tempmax and tempmin are rearranged 32 × 244 as per the date of the collection (243 dates) and state name (32). The population dataset is of size 32 × 8 data. Fig. 1, shows the variation of temperature in India on two different dates in the year 2020. In India, there are states which have an average maximum temperature of 14 °C and more than 38 °C on the same day. So, a statistical analysis is applied to find the relation between the total number of confirmed cases with temperature variations. Fig. 2 shows the variation of Confirmed cases in India for sample dates around April till November.

In any kind of prediction analysis, before coming up with a statistical model, we need to understand the distribution of those data. As all those datasets have too much information, to have an exact analysis, the data has to be studied properly. For convenience, some of the required data columns are extracted and created as a subset to perform statistical analysis in the R tool. The columns are named as specified in Table 1.

| Actual label        | Changed label |
|---------------------|---------------|
| TCIN                | y             |
| Maximum temperature | x1            |
| Minimum temperature | x2            |
| Total population (of a state) | x3     |
| Area (of a state in km2) | x4 |
| Rural population (of a state) | x5 |
| Urban population (of a state) | x6 |

The y, x1, x2, x3, x4, x5, and x6 values of all 244 days (1st April 2020 till 29th November 2020) are extracted as 243 different subsets. In this paper, we have shown the results and discussions of 10 such subsets. From Fig. 2 (data shown on April 2nd, May 16th, and September 22nd, 2020) it is clear that these data are not normally distributed, and also these are not categorical data. This would have narrowed the usage of Non-parametric statistical tests. Also, there is no negative correlation between these data as most of the values range between 0.3 to 0.5

To test the hypothesis behind the underlying concept that temperature plays a vital role in COVID-19 spread, a simple graphical analysis was done as shown in Fig. 4. In this analysis, three important variables were considered y, x1, and x2. It can be seen that the COVID-19 spread is not related to the average temperature of the state. The P-value of this variable is more than 0.05 as shown in the graph Fig. 5, which just disproves the hypotheses stated in [3].

After seeing the correlation between the data, we thought of determining the dependencies of these values. So, the p-value of y (confirmed COVID-19 cases) with respect to all other data was determined and the test analysis is shown in Fig. 4 for 10 test cases (on 10 different days). With this, it can be concluded that y is dependent on x3(population) and x4(Area). In the graph, in Fig. 4 we can see that the P-value with respect to an area falls even very less on an average of 0.009157. Fig. 5 shows the smooth linear regression curve derived using Eq. (1).

\[ y = \alpha + \alpha_2 \times 4 \]  

Here, y is the total number of confirmed COVID-19 cases, x4 is the area of the state, and (\( \alpha \) and \( \alpha_2 \)) are intercepted. Fig. 6(a) was derived from the initial stage of spread (on April 5th, 2020) and Fig. 6(b) was derived from the date when there were more cases (on May 24th, 2020). These were calculated for all 32 states.

As India is a densely populated country this value is highly dependent. Thus, to verify the hypothesis there is a need for a statistical model which needs to find the relation between the COVID-19 confirmed cases, temperature, the total population, rural population, and area data. The Generalized Additive Model (GAM) [10] will provide a way to analyze the nonlinearity in the data. GAM will also provide a way to fit the non-parametric data. For this analysis, a number of R packages were used, some of them are mgm, visibly, and ggplot2 [11].

5. Statistical analysis with results discussion

To create a generalized linear model with all sets of these three different datasets, in this paper GAM is used. Through this statistical analysis, we determine the nonlinear function to fit these three different datasets. Linear regression was applied to different sets of combinational values to find the dependency and R2 value. The intensity of the relationship between the linear regression model and the dependent variable is measured by R2. It is the ratio between the Variance of the model by Total variance. With that insight, GAM was calculated with different data to find the relationship between them. After understanding the relationship between different data through linear regression, the GAM was calculated for 6 different scenarios. The equations used are shown in Eqs. (2) to (7)

\[ \text{GAM1: } y = \alpha + s(x1) + s(x3) + s(x4) \]  
\[ \text{GAM2: } y = \alpha + s(x1) + s(x3) \]  
\[ \text{GAM3: } y = \alpha + s(x3) + s(x4) \]  
\[ \text{GAM4: } y = \alpha + s(x3) + s(x5) \]  
\[ \text{GAM5: } y = \alpha + s(x1) + s(x4) \]  
\[ \text{GAM6: } y = \alpha + s(x4) + s(x5) \]
The following analysis was derived by analyzing this model on all 55 dates. This paper, it shows the visual output for 2 different dates — one on 5th April 2020 and another on 24th May 2020. Fig. 7 shows the smooth GAM curve derived from all these 6 models. The proposed model can be analyzed using a different parameter. In this paper, the author wants to study if the data fit with less error. To show the determined results, arbitrarily 10 dates were picked and shown in Table 2 instead of all entries. Five different dates were picked from these data.

The overall average R2 value is shown in Fig. 8. The GAM4 shows a very good fit value of 0.934, this GAM4 is the GAM relation between the COVID-19 case count with the total population and total rural population. The fit value for GAM1 and GAM2 was around 0.66. The common factor in both cases is the maximum
temperature. The GAM1 finds the relationship between COVID-19 case count with temperature, population, and area, whereas GAM2 finds the relationship between COVID-19 case counts with temperature and population. The GAM1 and GAM2 were not good fits. To simplify it further, the author tried to reduce the value of $x_3$ and $x_4$ using log2. That was achieved by modifying the GAM4 as shown in Eq. (8). The fit curve as shown in Fig. 9 was determined. Also, the P-value of $x_3$ (Total population) is 0.000542, and $x_5$ (Total rural population) is 0.001407.

$$y = \alpha + s(\log_2(x_3)) + s(\log_2(x_5))$$ (8)

This model concludes that population and rural populations play a major role in the spread of COVID-19. In, India the population distribution is not even among all States and Union Territories.
Fig. 3. Distribution of data.
6. Real-time scenario-based result discussion

According to India’s census detail [8], the heavily populated states are Uttar Pradesh, Maharashtra, Bihar, West Bengal, Madhya Pradesh, Tamil Nadu, and so on. Among these the States, with the higher Area in km2 are Rajasthan, Madhya Pradesh, Maharashtra, Uttar Pradesh, Gujarat and so. Fig. 10(a) & (b) show that there is no even distribution of population within the Area. One of the major factors influencing the population of any country is the industrial opportunities available in that specific geographical location. So, when all put together as shown in Fig. 10. The major inference and suggestion through this finding are that before setting up industries, the Government policymaker should take the necessary parameters like population and area density
The heavy flow of a rural population can be controlled and proportionally the spread of any kind of disease can be avoided or controlled in near future. For applying our statistic inference, we have considered 9 different states with mixed COVID-19 cases. As per statistics given from the Indian Government portal [8]. According to [9], Maharashtra has a greater number of active COVID-19 cases, was from Fig. 10(c) and (d) we can see that number of persons engaged in factories is more than the state like Madhya Pradesh, where it has a smaller number of COVID-19 cases, Factories and its total overall population. But as per Fig. 10(b), the area available in Madhya Pradesh is equivalent to Maharashtra.

7. Conclusion

The purpose of this paper is to provide a statistical prediction model to study the COVID-19 situation. The motivation of this paper is to analyze the previous hypothesis and try to conclude the results with respect to Indian data. One of the major findings is that Average temperature may not influence the spread of
COVID-19 as suggested. In linear regression, it was mathematically shown that the maximum or minimum average temperature has a negative correlation with COVID-19 confirmed case count. The average R2 value of GAM4 (i.e.) the relation between the COVID-19 cases with total population and total rural population is about 0.934. This model may provide the Government agency with a statistical analysis that shows population and rural populations have a greater impact on COVID-19 spread. Maybe these models would be geographically dependent. The statistical analysis of these data will not end with this conclusion. This analysis needs to be integrated and further studied with Global Geographical data to understand this 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.
**Fig. 10.** Overall analysis.

(a) Total population

(b) Area measurement

(c) Factories operating

(d) Total number person engaged in Factories

(e) COVID-19 confirmed cases till 26th March 2021

References

[1] Shi H, Han X, Jaeg N, Cao Y, Alwalid O, Gu J, et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. Lancet Infect Dis 2020;20:425–34. http://dx.doi.org/10.1016/S1473-3099(20)30086-4.

[2] Shuja J, Alanazi E, Alasmary W, Alashaikh A. COVID-19 open source data sets: a comprehensive survey. Appl Intell 2021;51:1296–325. http://dx.doi.org/10.1007/s10489-020-01862-6.

[3] Xie J, Zhu Y. Association between ambient temperature and COVID-19 infection in 122 cities from China. Sci Total Environ 2020;724:138201. http://dx.doi.org/10.1016/j.scitotenv.2020.138201.

[4] Prata DN, Rodrigues W, Bermejo PH. Temperature significantly changes COVID-19 transmission in (sub)tropical cities of Brazil. Sci Total Environ 2020;729:138862. http://dx.doi.org/10.1016/j.scitotenv.2020.138862.

[5] Ministries/Departments IG. National portal of India. National Informatics Centre (NIC), Ministry of Electronics & Information Technology, Government of India; 2020, URL https://www.india.gov.in/india-glance/states-india.

[6] S. JR. Indian temperature scenario: “no global warming trend”. Open Access J Biog Sci Res 2020;1. http://dx.doi.org/10.46718/JBGSR.2020.01.000006.

[7] Centre NI. Ministry of health and family welfare India. Ministry of Health and Family Welfare, Government of India, URL https://www.mohfw.gov.in.

[8] of India G. India census data. National Informatics Centre (NIC), Ministry of Electronics & Information Technology, Government of India; 2020, URL https://data.gov.in/.

[9] Siva Athreya SB, Gadhivala N, Khakra J, Kanekar R, Mishra A, Mishra S, Nandi S, Nihesh Rathod AV, Sarath RS, Veeraraghavan S, Akshaya I. Covid-19 states of India and Karnataka district timeline 20-21. Karnataka: ISI-IISC; 2020.

[10] Grego JM. Generalized additive models. In: Encyclopedia of environmetrics. Wiley; 2001, http://dx.doi.org/10.1002/9780470057339.vag007.

[11] Wickham H. ggplot2. Wiley Interdiscip Rev Comput Stat 2011;3:180–5. http://dx.doi.org/10.1002/wics.147.