Chapter 15
Predictive Analysis of COVID-19 Transmission: Mathematical Modeling Study

K. Anitha

Abstract A novel COVID-19 caused by the virus SARS-CoV-2 was identified at China in December 2019, and it was declared as global pandemic on 11th March 2020 by the World Health Organization. In India on 30th January 2020, the first COVID-19 case was confirmed at Kerala. As of 6th June 2020, 6,702,662 cases have been reported in India even though stringent nationwide lockdown has been implemented. It is essential to identify the rate of disease transmission and its characterization. ICMR stated that 80% of COVID patients in India are without symptoms. In that case, we should focus on identifying the optimal attributes that will give the rate of transmission of this disease. Statistical machine learning is a powerful tool to fight against COVID-19 which gives the potential models for identifying the pattern of this disease, predictive analysis, and rate of transmission. This chapter aims to construct the predictive model to identify the severity of this disease transmission and also find the topographical spread of the contagion.

Keywords Predictive analysis · Parametric and nonparametric modeling · Estimation of parameters · Regression

Mathematics Review Subject Classification 92B05, 92B10, 97M60.

15.1 Introduction to Mathematical Epidemiology

In human history, communicable diseases play major revolution and drastic fatality rate. From Spanish flu to COVID-19, it is essential to identify the rate of transmission. In 1916, Greenwood JR stated that remarkable results have been to control the
disease with the help of mathematical methods in epidemics. Daniel Bernoulli first gave the mathematical analysis for smallpox in 1760. Mathematical models incorporated with computational tools bring the new insight to key issues like drug discovery, vaccination, transmission rate based on the various parameters, and treatment methods. The process of constructing the mathematical model that will improve the optimality in prediction is known as predictive analysis. Global search engine Google, Netflix, and financial sectors handling credit cards are widely using this predictive analysis. Always communicable diseases play vital role in human history. One of the best incidents is the 1918 Spanish flu which created more than 50,000,000 deaths. In 1980, more than 2,600,000 were occurred by measles, but it was drastically reduced in 2011 with the help of measles vaccine. The efficient understanding of the particular disease is more important to develop the drug. For this understanding, mathematical modeling helps the epidemiologist to identify the various factors involving the disease like age, rate of transmission, rate of mortality, demographic structure, etc.

The first mathematical model was initiated by Daniel Bernoulli against smallpox in the year 1760. In his model, he focused the competitive risk about smallpox, and he focused on how to control the death rate. Another remarkable spatial model was introduced by John snow in the year 1855 for cholera (London). For constructing the mathematical model for communicable diseases, it is essential to maintain some basic assumptions about its rate of transmission. The transmission rate always depends on the number of susceptible and infective persons [10]. This idea is the building block for compartmental model. In 1911, Ross proposed the simple compartmental model for malaria transmission, and he stated that the population of mosquitoes below the critical level will minimize the infection. Kermack and Mckendrik consecutively published three papers for basic compartmental model in 1927, 1932, and 1933. The formation of their epidemic model is represented by the following equations:

\[ v(t) = -s'(t) \]

\[ s'(t) = -s(t) \left[ \int_0^t N_1(u) v(t-u) \, du + N_1(t) K_0 \right] \]

\[ r'(t) = \int_0^t N_2(u) v(t-u) \, du + N_2(t) K_0 \]

\[ w(t) = \int_0^t N_3(u) v(t-u) \, du + N_3(t) K_0 \]

where

\[ s(t) = \text{No. of susceptibles}, \]
\[ w(t) = \text{No. of infected people}, \]
\[ r(t) = \text{No. of recovered people}. \]
\[
N_3(u) = e^{-\int \alpha(u) du}, \quad \alpha(u) \to \text{Rate of recovery at the infection age } u
\]

\[
N_1(u) = \beta(u)N_3(u), \quad \beta(u) \to \text{Rate of recovery when the infection age is } u
\]

Simple mathematical model which describes the transmission level is being defined by individual compartments to capture the status of individual, and this model gives the significant result. SIR (Susceptible-infectious-recovered) frame is the basic structure for all disease modeling. In this model, any individual can be in single compartment based on impact of infection. Julie C. Blackwood analyzed the basic framework of SIR frame. Compartmental models first divide the population into different clusters or compartments in which each compartment has the same level. They can move from one compartment to another compartment over a time with respect to their level of infection. Number of Infectious, Number of susceptible individuals, rate of recovery and lifelong immunity are some of the variables used in this model.

Effects of control measures like social distancing in public places, patterns of outbreak, positive cases of local and foreign nationals, health history, age, and migration details have to be reported. Kiesha Prem et al. stated that mixing pattern of social distancing will reduce the number of infections and they addressed the limitations of quantifying the contact persons in public places. This chapter proposes biological waves of COVID-19 in Sect. 15.2, construction of SIRD model along with its analyzation in Sect. 15.3, testing and validation in Sect. 15.4, and followed by experimental results and conclusion in Sects. 15.5 and 15.6.

**15.2 Biological Waves of COVID-19**

Pandemics are framed by bacteria and viruses. The COVID-19 pandemic is new to humans and it attacks the respiratory system; hence human body needs strong immune system. The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has started its impact in India on 30th January 2020 at Kerala. It is a single-stranded enveloped RNA Virus-1 which has a spherical shape with glycoprotein projections. It has many subtypes like alpha-, beta-, gamma-, and delta-type coronavirus. It has been found in bats, mice, cats, pigs, and some bird family. Figure 15.1 describes the genome and structure of COVID-19.

As per the scientists point of view in India, COVID-19 may have plateaued and could fall after certain period of lockdown is lifted. So far, 6,702,662 positive cases, 393,210 deaths, and 3,045,695 recovered cases have been reported globally. In India, 226,770 positive cases, 6363 deaths, and 109,462 recovered cases have been reported as on first week of June 2020. The duration for peak will depend on the maintenance of social distancing even after the relaxation of lockdown. Once lockdown is lifted, the second wave of COVID-19 will start again. In that case, curve
will reach the peak based on social distancing. Before that, vaccine or drug has to be initiated. More numbers of mathematical models are being developed to forecast and control COVID-19 worldwide.

The pathogen defined the significant challenges for health, planning about the pandemic, and the healthcare systems. Until the identification of vaccine, some of the non-pharmaceutical measures have to be considered. Infective individual’s contact tracing and isolation will minimize the level of outbreak. Epidemiological models should reduce the morbidity and mortality.

The following pictures show the data of COVID-19 status in India up to 16th May 2020 (Fig. 15.2).

### 15.3 Construction of Mathematical Model: SIRD

Let \( N(t) \) be the population at a time \( t \) based on the status of infection. The population is categorized as Susceptible (\( S \)), Infected (\( I \)), Recovered (\( R \)), and Death (\( D \)). The following flow chart exhibits the working method of this model (Fig. 15.3).

Assume that the constant population of the state = ten million

\[
N(t) = S + I + R - D
\]

Then, the rate of change of population is considered as
where \( B = S + I + R \), which is the exponential growth.

\[
\frac{dN}{dt} = (S + I + R)N - DN
\]

\[
N(t) = N_0 e^{(B-D)t}
\]  

(15.1)
For the process of self-limiting, Eq. (15.1) can be modified as follows:
\[
\frac{dN}{dt} = rN\left(1 - \frac{N}{F}\right) \text{ (logistic growth)}
\]

The solution of the above equation is (Fig. 15.4)

\[
N(t) = \frac{N_F F^{(r)t}}{F + N_F (e^{(r)t} - 1)} \to F \text{ as } t \to \infty \quad (15.2)
\]

The differential equation of susceptible population is given by

\[
\frac{ds}{dt} = N(t) - \alpha_1 S - \alpha_2 S - \alpha_3 S - d_1 S \quad (15.3)
\]

where \(\alpha_1\)—quarantined individuals, \(\alpha_2\)—recovered individuals, \(\alpha_3\)—not confirmed, and \(d_1\)—natural death.

Similarly, the differential equations of infected and recovered population are defined by

\[
\frac{dI}{dt} = N(t) - \beta_1 S - \beta_2 S - \beta_3 S - d_2 S - d_3 S \quad (15.4)
\]

\(d_3\)—infected death

\[
\frac{dR}{dt} = N(t) + \gamma_1 Q(t) + \gamma_2 I - d_1 S - d_2 S \quad (15.5)
\]

\(Q(t)\)—quarantined

Combining Eqs. (15.2)–(15.4) with initial conditions \(S(t) > 0, I(t) \geq 0, Q(t) \geq 0, R(t) \geq 0\), the solution is defined as

Fig. 15.4 Logistic growth of infected population
\[ \frac{dx}{dt} \geq -\omega S(t) \quad \omega = \alpha_1 + \beta_1 + d_1 \]

\[ \frac{dI}{dt} \geq -\varphi R(t) \quad \varphi = \alpha_2 + \beta_2 + d_1 \]

The SIRD model structure is locally asymptotically stable under the condition when \( R(0) < 1 \) and becomes unstable when \( R(0) < 1 \).

The Jacobian of the above system is given by

\[
J(E) = \begin{bmatrix}
-\omega & -\frac{\alpha_1 \omega}{S} & 0 & 0 \\
0 & -\frac{\alpha_1 \omega}{S} - \varphi & 0 & 0 \\
\beta_1 & \beta_2 & 0 & 0 \\
0 & \alpha_1 & \alpha_2 & -d_1
\end{bmatrix}
\]

The characteristic Equation is \( \det(J(E) - \lambda I) = 0 \)

where \( \lambda \) is the eigen value and the eigen values are \(-\omega, -\frac{\alpha_1 \omega}{S}, -\frac{\alpha_1 \omega}{S} - \varphi, \alpha_2\).

### 15.3.1 Analysis and Optimization Process of SIRD

The process of constructing the mathematical model or the particular model which generates the accurate prediction is called predictive modeling. Splitting of data, data prediction, performance analysis, and comparative study are the key roles of predictive modeling. The software R—has the package named applied predictive model package. For the implementation of above set of equations, we need to analyze and optimize with some set of data. Data splitting is the initial process in modeling. We have to check how well the derived model extrapolates for different population. Furthermore, it is essential to find how much data should be considered for testing and validating. Scaling and centering the predictor variables is the most significant process in data transformation. The predictor variable has zero mean in centering, while each variable is divided by its standard deviation in scaling process. These values increase the numerical stability of the calculation. For the transformation process, Box and Cox proposed the following transformation function:

\[
f(x) = \begin{cases} 
x^\theta - 1 & \text{if } \theta \neq 0 \\
\log(x) & \text{if } \theta = 0
\end{cases}
\]  

(15.6)
The value of $\theta$ can be identified using the training data. Principal component analysis (PCA) is one of the common methods to reduce the irrelevant information from the data. In this method, linear combination of predictor variables is considered as principal components. Mathematical formation of $i$th component of PC is written by

$$PC_i = \sum_{j=1}^{P} a_{ij} P_j$$  (15.7)

where $P$ is the number of predictors.

The technical term collinearity is the technical term which is being used when the pair of predictor variables have correlated between them. Multiple correlations are also available between the variables. In this process, the first step is formation of correlation matrix between the predictor variables. Next select the pair of highly correlated variables. Let the pair be $C_1$ and $C_2$. Now calculate the average correlation between $C_1$ and other variables and similarly for $C_2$ and other variables. If $C_1$ has the larger correlated value, then remove the variable $C_1$; otherwise remove $C_2$. Continue the process until no absolute correlation is found.

### 15.4 Testing and Validation

COVID-19 statewise status of India has been taken from [https://www.mohfw.gov.in/](https://www.mohfw.gov.in/) which displays the COVID-19 affected people from all the states of India up to June 6, 2020, and the values are displayed here. Figures 15.5 and 15.6 give the numerical values of number of active cases, number of death, and total number of affected people.

The predictive model has been verified with this data. For implementation process, Weka classifier is being used for analysis. The results are being displayed here.

### 15.5 Experimental Results (Fig. 15.7)

| Test mode: | Tenfold cross-validation |
| Classifier model for full training set | Locally weighted learning |
| Using classifier: | Decision tree, Linear Weighting Kernels and all Neighbour with Cross Validation |
| Co-efficient of Correlation | 0.7315 |
| Mean absolute error | 11,200.3906 |
COVID-19 STATEWISE REPORT-I

Fig. 15.5 COVID-19 statewise report-I

COVID-19 STATEWISE REPORT-II

Fig. 15.6 COVID-19 statewise report-II
Test mode: Tenfold cross-validation
Classifier model for full training set
Locally weighted learning
Root mean squared error 36,790.2434
Relative absolute error 59.0362%
Root relative squared error 87.7999%
Total number of instances 37
Attributes: 6
1. Number of hits
2. Name of state/UT
3. Active cases
4. Cured/discharged/migrated
5. Deaths
6. Total confirmed cases

Test mode: evaluate on training data
Simple K-Means—Euclidean Distant—R first last
@attribute 'Name of State / UT' { 'Andaman and Nicobar Islands', 'Andhra Pradesh', 'Arunachal Pradesh', 'Assam', 'Bihar', 'Chandigarh', 'Chhattisgarh', 'Dadar Nagar Haveli', 'Delhi', 'Goa', 'Gujarat', 'Haryana', 'Himachal Pradesh', 'Jammu and Kashmir', 'Jharkhand', 'Karnataka', 'Kerala', 'Ladakh', 'Madhya Pradesh', 'Maharashtra', 'Manipur', 'Meghalaya', 'Mizoram', 'Nagaland', 'Odisha', 'Puducherry', 'Punjab', 'Rajasthan', 'Sikkim', 'Tamil Nadu', 'Telangana', 'Tripura', 'Uttarakhand', 'Uttar Pradesh', 'West Bengal', 'Cases being reassigned to states', 'Total#'}
@attribute 'Active Cases*' numeric
@attribute Cured/Discharged/Migrated* numeric
@attribute Deaths** numeric
@attribute 'Total Confirmed cases*' numeric
Number of iterations: 3

Fig. 15.7 Time series plot for infected people
Test mode: evaluate on training data

Within cluster sum of squared errors: 40.04935720953203

Initial starting points (random):

Cluster 0: 12, Haryana, 1439, 2134, 24, 3597

Cluster 1: 25, Odisha, 996, 1604, 8, 2608

| Correlation coefficient | 0.7534 |
|-------------------------|--------|
| Mean absolute error     | 10,527.7943 |
| Root mean squared error | 35,716.3567 |
| Relative absolute error | 55.491% |
| Root relative squared error | 85.2371% |
| Total number of instances | 37 |

Parameters for five states

| Parameters | Maharashtra | Tamil Nadu | Gujarat | Uttar Pradesh | Kerala |
|------------|-------------|------------|---------|---------------|--------|
| $S(t)$     | 25,600      | 17,000     | 12,000  | 10,000        | 5389   |
| Death due to COVID-19 | 0.36 | 0.24 | 0.37 | 0.33 | 0.12 |

**15.5.1 Age-Parameter Compartmental Model**

Equations (15.2)–(15.5) can be reframed for age-related model. From https://www.kaggle.com/covid19-in-india, age group data is taken. This data contains ten unique values which represent the age group from 0 to 9, 10–19, … up to ≥80 along with missing data, and the percentage of cases affected in the age group (Figs. 15.8 and 15.9).

![AGE GROUP DATA SET](image-url)

*Fig. 15.8 Age group data*
Regression equations for the above data are $y = -0.475x + 98.026$.
Correlation coefficient between age group and affected people is $-0.214$.

15.6 Conclusion

In this chapter, predictive epidemic model of COVID-19 has been framed. The confirmation cases are increasing rapidly for each and every day. The prediction factor of every infected individual is more important to avoid the transmission. The various parameters like age, health factors, respiratory syndrome level, and density of the population should be focused properly. It has been observed that when the value of $R(0) < 1$, then the system is locally stable. Based on the numerical results, it has been shown that there is the correlation between infected and susceptible individual. Hence, social distancing and invention of drug are more important. In future, the predict model can be extended to Bayesian network model which gives the result based on categorical responses. It is the acyclic graph which analyzes the various factors and constructs the network which will give the optimal result in decision-making.

Bibliography

1. Ahmed, F., Zviedrite, N., Uzicanin, A.: Effectiveness of workplace social distancing measures in reducing influenza transmission: a systematic review. BMC Public Health. 18, 518 (2018).  
https://doi.org/10.1186/s12889-018-5446-1

2. Tuite, A.R., Fisman, D.N., Greer, A.L.: Mathematical MODELLING of COVID-19 transmission and mitigation strategies in the population of Ontario, Canada. CMAJ. 192, E497 (2020)
3. Andreasen, V.: The final size of an epidemic and its relation to the basic reproduction number. Bull. Math. Biol. 73(2011), 2305–2232 (2011)

4. Bernoulli, D.: Reflexions sur les avantages de l’inoculation. Merc. France, 173 (1760)

5. Bichara, D., Kang, Y., Castillo-Chavez, C., Horan, R., Pettina, C.: SIS and SIR epidemic models under virtual dispersal. Bull. Math. Biol. 77, 2004–2034 (2015)

6. Castillo-Chavez, C.: Mathematical and statistical approaches to AIDS epidemiology. In: Lecture Notes in Biomathematics, vol. 83, pp. 289–300. Springer, Berlin (1989)

7. Breuer, F.: Mathematical epidemiology: past, present and future. Infect. Dis. Model. 2(2), 113–127 (2017)

8. Fong, M.W., Gao, H., Wong, J.Y., et al.: No pharmaceutical measures for pandemic influenza in no healthcare settings—social distancing measures. Emerg. Infect. Dis. 26, 976 (2020). https://doi.org/10.3201/eid2605.190995

9. Greenwood Jr., M.: The application of mathematics to epidemiology. Nature. 97, 243–244 (1916)

10. Hamer, W.H.: Epidemic disease in England - the evidence of variability and of persistence. Lancet. 167, 733–738 (1906)

11. Blackwood, J.C., Childs, L.M.: An introduction to compartmental MODELLING for the budding infectious disease modeler. Lett. Biomath. 5(1), 195–221 (2018). https://doi.org/10.1080/23737867.2018.1509026

12. Kermack, W.O., McKendrick, A.G.: A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. A Math. Phys. Eng. Sci. 115, 700–721 (1927)

13. Prem, K., Liu, Y., Russell, T.W., Kucharski, A.J., Eggo, R.M., Davies, N.: The effect of control strategies that reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China. Lancet Public Health. 5, e261 (2020)

14. Kermack, W.O., McKendrick, A.G.: Contributions to the mathematical theory of epidemics, Part II. Proc. R. Soc. Lond. 138, 55–83 (1932)

15. Kermack, W.O., McKendrick, A.G.: Contributions to the mathematical theory of epidemics, Part III. Proc. R. Soc. Lond. 141, 94–112 (1933)

16. Abraham, P., Aggarwal, N., Babu, G.R., Barani, S., Bhargava, B., Bhatnagar, T., Dhama, A.S., Gangakhedkar, R.R., Giri, S., Gupta, N., Kurup, K.K., Manickam, P., Murhekar, M., Potdar, V., Praharaj, I., Khan, R., Reddy, D.C.S., Saravankumar, V., Shah, N., Singh, H., Thangaraj, J.W.V., Yadav, N.: Laboratory surveillance for SARS-CoV-2 in India: performance of testing & descriptive epidemiology of detected COVID-19, January 22 - April 30, 2020. Indian J. Med. Res. 151, 424 (2020). https://doi.org/10.4103/ijmr.IJMR_1896_20044

17. Ross, R.A.: The Prevention of Malaria (with Addendum). John Murray, London (1911)

18. Stehlé, J., Voiron, N., Barrat, A., et al.: Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. BMC Med. 9, 87 (2011)

19. Kim, S., Seo, Y.B., Jung, E.: Prediction of COVID-19 transmission dynamics using a mathematical model considering behavior changes in Korea. Epidemiol. Health. 42, e2020026 (2020)

20. Wallinga, J., Teunis, P., Kretzschmar, M.: Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. Am. J. Epidemiol. 164, 936–944 (2006)

21. Woelfel, R., Corman, V., Guggemos, W., Seilmaier, M., Zange, S., Mueller, M., et al.: Clinical presentation and virological assessment of hospitalized cases of coronavirus disease 2019 in a travel-associated transmission cluster. medRxiv. (2020). https://doi.org/10.1101/2020.03.05.20030502

22. Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., et al.: Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. JAMA. 323, 1061 (2020)

23. Li, Y., Liang, Z., Wu, J., Wang, X., Han, X., et al.: The effect of control strategies that reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China. Lancet Public Health. 5, e261 (2020)

24. Zhang, J., Klepac, P., Read, J.M., et al.: Patterns of human social contact and contact with animals in Shanghai, China. Sci. Rep. 9, 1–11 (2019)