Modeling and Monitoring COVID-19 Monthly Infected Cases and Deaths

RAJARATHINAM ARUNACHALAM (arrathinam@yahoo.com)
Manonmaniam Sundaranar University https://orcid.org/0000-0002-3245-3181

Research Article

Keywords: Panel Regression Model, Least Squares Dummy Variable Model, Fixed Effect model, Random Effect Model, Restricted F-test, Hausman Test

Posted Date: June 14th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-612366/v1

MODELING AND MONITORING COVID-19 MONTHLY INFECTED CASES AND DEATHS

Rajarathinama,A[a] and Anju, J.B.[b]
[a,b]Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli – 627 012, Tamil Nadu State, INDIA

Emails: rajarathinam@msuniv.ac.ina and anjudwaraka71@gmail.comb

CORRESPONDING AUTHOR
Dr.Rajarathinam, A., Professor, Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli – 627 012, Tamil Nadu State, INDIA
Email: rajarathinam@msuniv.ac.in

ABSTRACT

The effects of the novel coronavirus (COVID-19) pandemic could not have been more profound, with the world encountering health crises as well as enormous economic crises. In this paper, the relationships, and trends in the number of COVID-19 infected new cases and the number of deaths due to COVID-19 in all 37 districts of Tamil Nadu state, India, during the period, 3rd July 2020 to 31st March,2021 based on a panel regression model.

KEYWORDS: Panel Regression Model, Least Squares Dummy Variable Model, Fixed Effect model, Random Effect Model, Restricted F-test, Hausman Test

1 INTRODUCTION

1.1 Background of the study

The COVID-19 pandemic that began in 2019 has received much attention, affecting most of the world economies and leading to countless deaths. In the absence of antiviral drugs and vaccines, the number of new COVID-19-infected cases has increased tremendously and has caused many deaths. The deployment of various methodologies to analyze pandemic data has become a particularly important research area to forecast coronavirus infected cases and deaths.

1.2. Review of literature

A few of the research work carried out by various authors in modeling of COVID-19 data are reviewed in sequence as below.

Al-Rousan and Al-Najjar,2020 pointed out that on 1 st of February 2020, CoVID-19 coronavirus outbreak was announced to the public and it was classified as epidemic. Although the disease was discovered and concentrated in Hubei province, China, but it was exported to all other Chinese provinces and spread globally. Until this moment, all plans failed to contain the novel coronavirus disease, and it continued spreading to the entire world to exceed 98000 cases globally with 80000 cases exist in mainland China. This manuscript aims to study and interpret the effect of environment and metrological variables on coronavirus disease spreading in 30 Chinese provinces. Besides, to investigate of the impact of new China regulations and plans to mitigate of CoVID-19 on spreading the disease. This article forecasts the size of the disease spreading based on time series forecasting models including Brown, Holt, Simple, and Auto Regressive Integrated Moving Average (ARIMA). The growing size for CoVID-19 in China for the next 210 days is estimated by predicting the expected confirmed and recovered cases. The results revealed that weather conditions have strong effect on coronavirus spreading in most of the Chinese provinces. Increasing both temperature and shortwave radiation variables would increase the number of confirmed, death, and recovered cases.

Al-Rousan and Al-Najjar,2020 stated that the Coronavirus epidemic caused announcing emergency case in South Korea. The virus started with one infected case by January 20, 2020, where 9583 announced cases were reported by March 29, 2020. This indicates that the number of confirmed cases is increasing rapidly, which can cause national crises for South Korea. The aim of this study is to fill a gap between previous studies and the current development of CoVID-19 spreading, by extracting a relationship between independent variables and dependent variable. This research statistically analyzed the effect of sex, region, infection reasons, birth year, and released or diseased date on the reported numbers of recovered and deceased cases. The results found that sex, region, and infection reasons affected on both recovered and deceased cases, while birth year only affected on deceased cases. Besides, no deceased cases are reported for released cases, while 11.3% of deceased cases positive confirmed after their deceased. Unknown reason of infection is the main variable that detected in South Korea with more than 33% of total infected cases.

Arumugam and Rajathi (2020) asserted that, the Markov chain model is mainly used for business, manpower planning, share market and many different areas. Because the prediction of the any ideas based on the Markov chain the result needs to be efficient. Now, the infection of corona virus COVID-19 is a large task for the human being as well as the government. This paper is focusing tool for prediction of corona virus infection with a Markov chain model. Markov chain model had been used to predict the corona virus (COVID-10) based at the secondary data as on 13th March 2020. The 1st order Markov models had been used to predict the impact of corona virus using probability matrices and Monte Carlo simulation. To present the applications of this model, 2020 corona virus pandemic in India by country and union territory become used as a case study. It will be useful for prediction of the corona virus COVID-19 in destiny.

Bertozzi et al. (2020) stated that the coronavirus disease 2019 (COVID-19) pandemic has placed epidemic modeling at the forefront of worldwide public policy making. Nonetheless, modeling and forecasting the spread of COVID-19 remains a challenge. Here, detailed three regional scale models for forecasting and assessing the course of the pandemic. This work demonstrates the utility of parsimonious models for early-time data and provides an accessible framework for generating policy-relevant insights into its course. Also to show how these models can be connected to each other and to time series data for a particular region. Capable of measuring and forecasting the impacts of social distancing, these models high light the dangers of relaxing nonpharmaceutical public health interventions in the absence of a vaccine or antiviral therapies.

Mahanty et al., (2020) presented a medical stance on research studies of COVID-19, wherein they estimated a time-series databased statistical model using prophet to comprehend the trend of the current pandemic in the coming future after July 29, 2020 by using data at a global level. Prophet is an opensource framework discovered by the Data Science team at Facebook for carrying out forecasting based operations. It aids to automate the procedure of developing accurate forecasts and can be customized according to the use case we are solving. The Prophet model is easy to work because the official repository of prophet is live on GitHub and is open for contributions and can be fitted effortlessly. The statistical data presented on the paper refers to the number of daily confirmed cases officially for the period January 22, 2020, to July 29, 2020. The estimated data produced by the forecast models can then be used by Governments and medical care departments

of various countries to manage the existing situation, thus trying to flatten the curve in various nations as we believe that there is minimal time to do this. The inferences made using the model can be clearly comprehended without much effort. Furthermore, it tries to give an understanding of the past, present, and future trends by showing graphical forecasts and statistics. Compared to other models, prophet specifically holds its own importance and innovativeness as the model is fully automated and generates quick and precise forecasts that can be tunable additionally

Tiwar et al., (2020) stated that COVID-19 is rapidly spreading in South Asian countries, especially in India. India is the fourth most COVID-19 affected country at present i.e., until July 10, 2020. With limited medical facilities and high transmission rate, the study of COVID-19 progression and its subsequent trajectory needs to be analyzed in India. Epidemiologic mathematical models have the potential to predict the epidemic peak of COVID-19 under different scenarios. Lockdown is one of the most effective mitigation policies adopted worldwide to control the transmission rate of COVID-19 cases. In this study, we use an improvised five compartment mathematical model, i.e., Susceptible (S)-Exposed (E)-Infected (I)-Recovered (R)-Death (D) (SEIRD) to investigate the progression of COVID-19 and predict the epidemic peak under the impact of lockdown in India. The aim of this study is to provide a more precise prediction of epidemic peak and to evaluate the impact of lockdown on epidemic peak shift in India. For this purpose, we examine the most recent data (from January 30, 2020 to July 10, 2020 i.e., 160 days) to enhance the accuracy of outcomes obtained from the proposed model. The model predicts that the total number of COVID-19 active cases would be around 5.8 × 105 on August 15, 2020 under current circumstances. In addition, our study indicates the existence of under-reported cases i.e., 105 during the post-lockdown period in India. Consequently, this study suggests that a nationwide public lockdown would lead to epidemic peak suppression in India. It is expected that the obtained results would be beneficial for determining further COVID-19 mitigation policies not only in India but globally as well.

Rosario et al., (2020), aimed to evaluate the relationship between weather factors (temperature, humidity, solar radiation, wind speed, and rainfall) and COVID-19 infection in the State of Rio de Janeiro, Brazil. Solar radiation showed a strong (-0.609, $p<0.01$) negative correlation with the incidence of novel coronavirus (SARS-CoV-2). Temperature (maximum and average) and wind speed showed negative correlation ($p < 0.01$). Therefore, in this studied tropical state, high solar

radiation can be indicated as the main climatic factor that suppress the spread of COVID-19. High temperatures, and wind speed also are potential factors. Therefore, the findings of this study show the ability to improve the organizational system of strategies to combat the pandemic in the State of Rio de Janeiro, Brazil, and other tropical countries around the word.

Zuo (2020) asserted that in the current scenario, the outbreak of a pandemic disease COVID-19 is of great interest. A broad statistical analysis of this event is still to come, but it is immediately needed to evaluate the disease dynamics in order to arrange the appropriate quarantine activities, to estimate the required number of places in hospitals, the level of individual protection, the rate of isolation of infected persons, and among others. In this article, we provide a convenient method of data comparison that can be helpful for both the governmental and private organizations. Up to date, facts and figures of the total the confirmed cases, daily confirmed cases, total deaths, and daily deaths that have been reported in the Asian countries are provided. Furthermore, a statistical model is suggested to provide a best description of the COVID-19 total death data in the Asian countries.

Chu (2021) reported that the novel coronavirus (COVID-19) that was first known at the end of 2019 has impacted almost every aspect of life as we know it. This paper focuses on the incidence of the disease in Italy and Spain—two of the first and most affected European countries. Using two simple mathematical epidemiological models—the Susceptible-Infectious-Recovered model and the log-linear regression model, we model the daily and cumulative incidence of COVID-19 in the two countries during the early stage of the outbreak, and compute estimates for basic measures of the infectiousness of the disease including the basic reproduction number, growth rate, and doubling time. Estimates of the basic reproduction number were found to be larger than 1 in both countries, with values being between 2 and 3 for Italy, and 2.5 and 4 for Spain. Estimates were also computed for the more dynamic effective reproduction number, which showed that since the first cases were confirmed in the respective countries the severity has generally been decreasing. The predictive ability of the log-linear regression model was found to give a better fit and simple estimates of the daily incidence for both countries were computed.

Gecili et al., (2021), states that the novel coronavirus (COVID-19) is an emergent disease that initially had no historical data to guide scientists on predicting/ forecasting its global or national impact over time. The ability to predict the progress of this pandemic has been crucial for decision making aimed at fighting this pandemic and controlling its spread. In this work we considered four different statistical/time series models that are readily available from the 'forecast' package in R. We performed novel applications with these models, forecasting the number of infected cases (confirmed cases and similarly the number of deaths and recovery) along with the corresponding 90% prediction interval to estimate uncertainty around pointwise forecasts. Since the future may not repeat the past for this pandemic, no prediction model is certain. However, any prediction tool with acceptable prediction performance (or prediction error) could still be very useful for public-health planning to handle spread of the pandemic, and could policy decision-making and facilitate transition to normality. These four models were applied to publicly available data of the COVID-19 pandemic for both the USA and Italy. We observed that all models reasonably predicted the future numbers of confirmed cases, deaths, and recoveries of COVID-19. However, for the majority of the analyses, the time series model with autoregressive integrated moving average (ARIMA) and cubic smoothing spline models both had smaller prediction errors and narrower prediction intervals, compared to the Holt and Trigonometric Exponential smoothing state space model with Box-Cox transformation (TBATS) models. Therefore, the former two models were preferable to the latter models. Given similarities in performance of the models in the USA and Italy, the corresponding prediction tools can be applied to other countries grappling with the COVID-19 pandemic, and to any pandemics that can occur in future.

Rajarathinam and Tamilselvan, (2021) studied the short- and long-term cointegration relationships between the cumulative number of COVID-19 infections and the cumulative numbers of deaths due to COVID-19 are studied by employing an autoregressive distributed lag model and bound cointegration tests. The stability of the estimated model is also assessed. The cumulative sum of the recursive residuals test and the cumulative sum of recursive residuals squares tests are used to assess the consistency of the model's parameters.

Rajarathinam and Tamilselvan,(2021) estimated the dynamic relationship between the number of cases of new COVID-19 infections and the number of deaths due to COVID-19 was assessed using the Johnsen-Fisher cointegration test, vector error correction model and Granger causality test. The daily number of new cases of COVID-19 infections and deaths due to COVID-19 in the United States, Canada, the Ukraine and India were collected from the website for the period from 01 April 2020 to 26 December 2020. The summary statistics revealed that the highest number of COVID-19 infected cases were registered in the United States, followed by India, Canada and Ukraine; the highest number of deaths due to COVID-19 were registered in the United States, followed by India, Ukraine and Canada. The death percentage is exceedingly high in Canada, followed by the United States, Ukraine and India. The Johnsen-Fisher cointegration test results reveal the existence of one cointegration equation. The vector error correction model and Granger causality test reveal that long-term and short-term causality exists between cases of COVID-19 infections and deaths. The speed of adjustment is found to be 9.9% (Rajarathinam and Tamilselvan, 2021).

1.3 Objectives of the present study

Based on the above discussion, the present study aimed to study the relationships and trends in the number of new COVID-19 infections and the number of deaths due to COVID-19 in all 37 districts of Tamil Nadu state, India, during the months of 3rd July-2020 to 31st March,2021 based on panel regression model with the number of deaths (DEATH) due to COVID-19 as the dependent variable and the number of new positive COVID-19 infected new cases (NCASE) as the independent variable.

1.4 Panel data model

Panel data are a type of data that contain observations of multiple phenomena collected over different time periods for the same group of individuals, units or entities. In short, econometric panel data are multidimensional data collected over a given period.

A simple panel data regression model is specified as

$$Y_{it} = \alpha + \beta X_{it} + \nu_{it}$$

Here, Y is the dependent variable, X is the independent or explanatory variable, $\alpha \, and \, \beta$ are the intercept and slope, i stands for the i^{th} cross-sectional unit and t for the t^{th} month, and X is assumed to be non-stochastic and the error term to follow the classical assumptions, namely,

$E(\nu_{it}) = N(0, \sigma^2)$. In this study, i, that is, the number of cross-sections (districts), is 37 (i=1, 2, 3, …, 37), and t=1, 2, 3, …, 9.

Detailed discussions of panel data modeling can be found in, viz., Baltagi (2001), Gujarati et al. (2017), and Hsiao (2003).

By combining time series of cross-sections of observations, panel data provide "more informative data, more variability, less collinearity among variables, more degrees of freedom and more efficiency" (Baltagi, 2001).

2 MATERIALS AND METHODS

2.1 Materials

The COVID-19 dataset was collected from the official Tamil Nadu government website (www.stopcorona.tn.gov.in) from 3rd July, 2021 to 31st March, 2021 (the study period). Different econometric tools related to panel data regression modeling were employed to investigate the research questions of the present study. Several methodologies for panel data regression modeling are discussed in the methods section. EViews Ver.11. was used for the calculations.

2.2 Methods

Panel data models describe individual behavior both across time and across individuals. There are three types of models: Pooled Regression Model (PRM), Fixed Effects (FE) models and Random effects (RE) models. The Polled Regression Model is also know are Constant Coefficient Model (CCM).

2.2.1 Unit root tests

Unit roots in panel data can be tested for using either the Levin et al. (2002) test or the Hadri (2000) Lagrange multiplier (LM) stationarity test. The null hypothesis is that the panels contain unit roots, and the alternative hypothesis is that the panels are stationary. In the results, if the p value is less than 0.05, then one can reject the null hypothesis and accept the alternative hypothesis.

2.2.2 Pooled Regression OLS model or Constant Coefficient Model

The pooled model with constant coefficients (the usual assumption for cross-sectional analysis) is specified as

$$Y_{it} = \alpha + \beta X_{it} + \nu_{it}$$

Here, i = 1, 2, 3, …, 37, and t = 1, 2, 3 … 9, where i stands for the i^{th} cross-sectional units (Districts) and t for the t^{th} month period, and it is assumed that X (the independent variable) is non-stochastic and that the error term follows the classical assumptions, namely,

$$E(\nu_{it}) \sim N(0, \sigma^2)$$

2.2.3 Individual-specific effects model

We assume that there is unobserved heterogeneity across individuals captured by α_i. The main question is whether the individual-specific effects α_i are correlated with the regressor. If they are correlated, we have an FE model. If they are not correlated, we have a RE model.

2.2.4 FE least squares dummy variable (LSDV) model (Gujarati et al., 2017)

The term fixed effects is used because although the intercept may vary across districts, each entity's intercept does not vary over time; that is, it is time invariant.

$$y_{it} = \alpha_i + x_{it}\beta + \gamma_{it}$$

One can recover the individual-specific effect after estimation as $\hat{\alpha}_i = \overline{y}_i - \overline{x}_i\hat{\beta}$

In other words, the individual-specific effects are the leftover variation in the dependent variable that cannot be explained by the regressor. By using the dummy variable technique, one can allow the fixed effects intercept to vary among the districts.

2.2.5 RE model

The RE model assumes that the individual-specific effects α_i are distributed independently of the regressor and includes α_i in the error term. Each unit has the same slope parameters and a composite error term $\varepsilon_{it} = \alpha_i + v_{it}$.

$$y_{it} = x_{it}\beta + (\alpha_i + v_{it})$$

Here, $\text{var}(\varepsilon_{it}) = \sigma_\alpha^2 + \sigma_v^2$ and $\text{cov}(\varepsilon_{it}, \varepsilon_{is}) = \sigma_\alpha^2$ so $\rho_\varepsilon = cor(\varepsilon_{it}, \varepsilon_{is}) = \dfrac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}$.

Rho is the interclass correlation of the error, that is, the fraction of the variance in the error due to the individual-specific effects. It approaches 1 if the individual effects dominate the idiosyncratic error.

2.2.6 Hausman test (Hausman, 1978)

The null hypothesis of the Hausman test is that the RE model is preferred and the alternative hypothesis that the FE model is preferred. It tests whether the unique error (α_i) is correlated with the regressor, and the null hypothesis is that they are not correlated. The RE estimator is more efficient, so one should use it if the Hausman test supports this choice. The Hausman test statistic, which can be calculated only for the time-varying regressors, is

$$H = \left(\hat{\beta}_{RE} - \hat{\beta}_{FE}\right)' \left(V(\hat{\beta}_{RE}) - V(\hat{\beta}_{FE})\right)\left(\hat{\beta}_{RE} - \hat{\beta}_{FE}\right)$$

2.2.7 Restricted F-test (Bhaumik,2017)

In the Restricted F-test, The null hypothesis is

$$H_N = \alpha_2 = \alpha_3 = \ldots = \alpha_N$$

To test the validity of H_N we compute

$$F^* = \frac{\left(R^2_{FEM} - R^2_{CCM}\right)/\left(N-1\right)}{\left(1-R^2_{FEM}\right)/\left((NT-N-k\right)} \sim F\left(N-1, NT-N-k\right)$$

Where

R^2_{FEM} = Computed R^2 value from estimated Fixed Effects Model (called unrestricted regression)

R^2_{CCM} = Computed R^2 value from the estimated Constant Coefficient Model (restricted regression)

N = Number of intercepts in Fixed Effects Model (equal to number of cross-sectional units)

NT = total number of observations

k – number of explanatory variables in the Fixed Effect Model

The decision rule is : If $F^* \rangle F_\lambda \left(N-1, NT-N-k\right)$, i.e., computed-F is greater than the theoretical-F at chosen level of significance λ and degrees of freedom (N-1) for the numerator and (NT-N-k) for denominator, we reject the null hypothesis ($\left(H_N\right)$ and conclude that, compared with the Constant Coefficient Model, the Fixed Effects Model is more appropriate in the context of our pooled or panel data set. This also means that the fixed effects are present and the intercepts of cross-sectional units are statistically significantly different from each other.

3 RESULTS AND DISCUSSIONS

3.1 Unit root tests

In analyses of time series data, it is important that the study variables are stationary, which means that the means and variances of the variable data are the same. Accordingly, Levin-Lin-Chu unit root tests were carried out to test the stationarity of the study variables, viz., the number of COVID-19-infected patients (NCASE) and of deaths (DEATH) due to COVID-19. The results are reported in Tables 1.

The test results presented in Tables 1 reveals the two variables under study, NCASE and DEATH, to be stationary in level, since the Levin, Lin and Chu t-statistics are found to be highly significant ($p<0.0000$). Hence, the variables under study are found to be stationary.

Table 1. Unit root test results for the variables NCASE and DEATH

Method	Variable	Statistic	Prob.**
Levin, Lin & Chu t*	NCASE	-52.6381	0.0000
	DEATH	-122.017	0.0000
** Probabilities are computed assuming asympotic normality			

3.2 Summary statistics

Fig. 1 depicts the number of COVID-19-positive patients registered in different districts of Tamil Nadu during the months of 3rd July-2020 to 31st March,2021.

Further the of COVID-19-positive patients registered follows the following third-degree polynomial with the value of R^2 is equal to 99 %. The model is highly significant, and the parameters values are also significant at 5% level. The residuals due to this model are normally distributed because the Shapiro-Wilk's test (test for normality) statistic is non-significant. Also, the residuals are independent as the run-test statistic value is also non-significant. Hence the model is well defined one and the results obtained due to this model are statistically valid.

$$y = 87753.53^{*} + 92779.55^{*}t - 28848.64^{*}t^{2} + 1997.04^{*}t^{3}$$

(* indicates 5 % level of significance)

The highest numbers of new COVID-19 infections were registered in the month of August- 2020 (181817). The lowest numbers of new COVID-19-infections were registered in the month of February 2021. Overall, during the study period 786990 (3rd July 2020 to 31st March 2021) COVID-19 infections were registered all over Tamil Nadu.

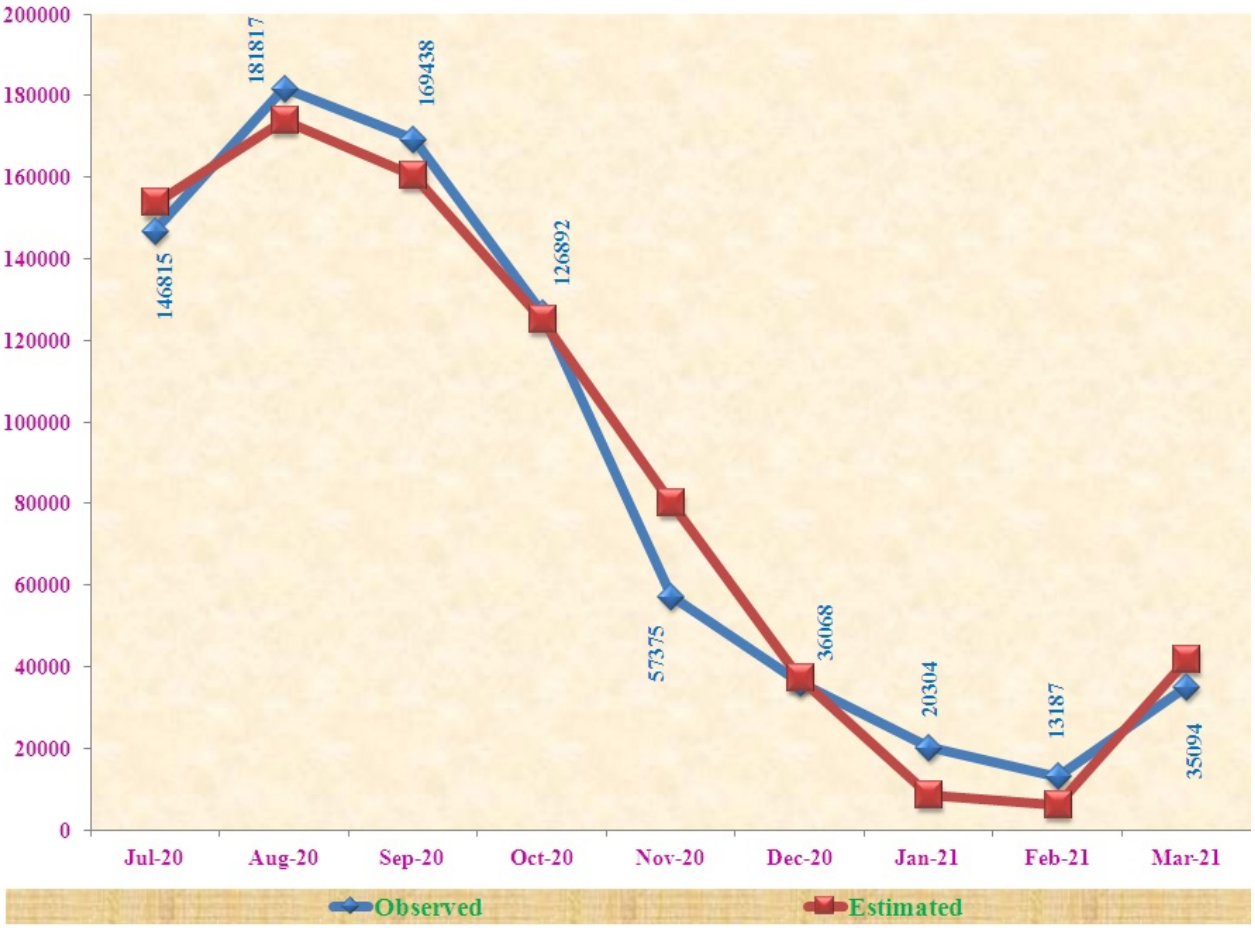

Fig.1. The number of COVID-19-positive patients registered in different districts of Tamil Nadu during the months of 3rd July-2020 to 31st March,2021.

Fig. 2 depicts the number of deaths due to COVID-19 registered in different districts of Tamil Nadu during the period 3rd July-2020 to 31st March, 2021.

$$5515.98^{**}\exp(-0.409437^{**})$$

Further the death due to COVID-19 registered follows the following exponential model with the value of R^2 is equal to 95 %. The model is highly significant and the parameters values are also significant at 5% level. The residuals due to this model are normally distributed because the Shapiro-Wilk's test (test for normality) statistic is non-significant. Also the residuals are independent as the run-test statistic value is also non-significant. Hence the model is well defined one and the results obtained due to this model are statistically valid.

The highest number of deaths due to COVID-19 occurred in the month of August – 2020 (3387) and the lowest number (140) of deaths was in the month of February, 2020. In total, in during the study period 11022 number of deaths were registered due to COVID-19 in Tamil Nadu.

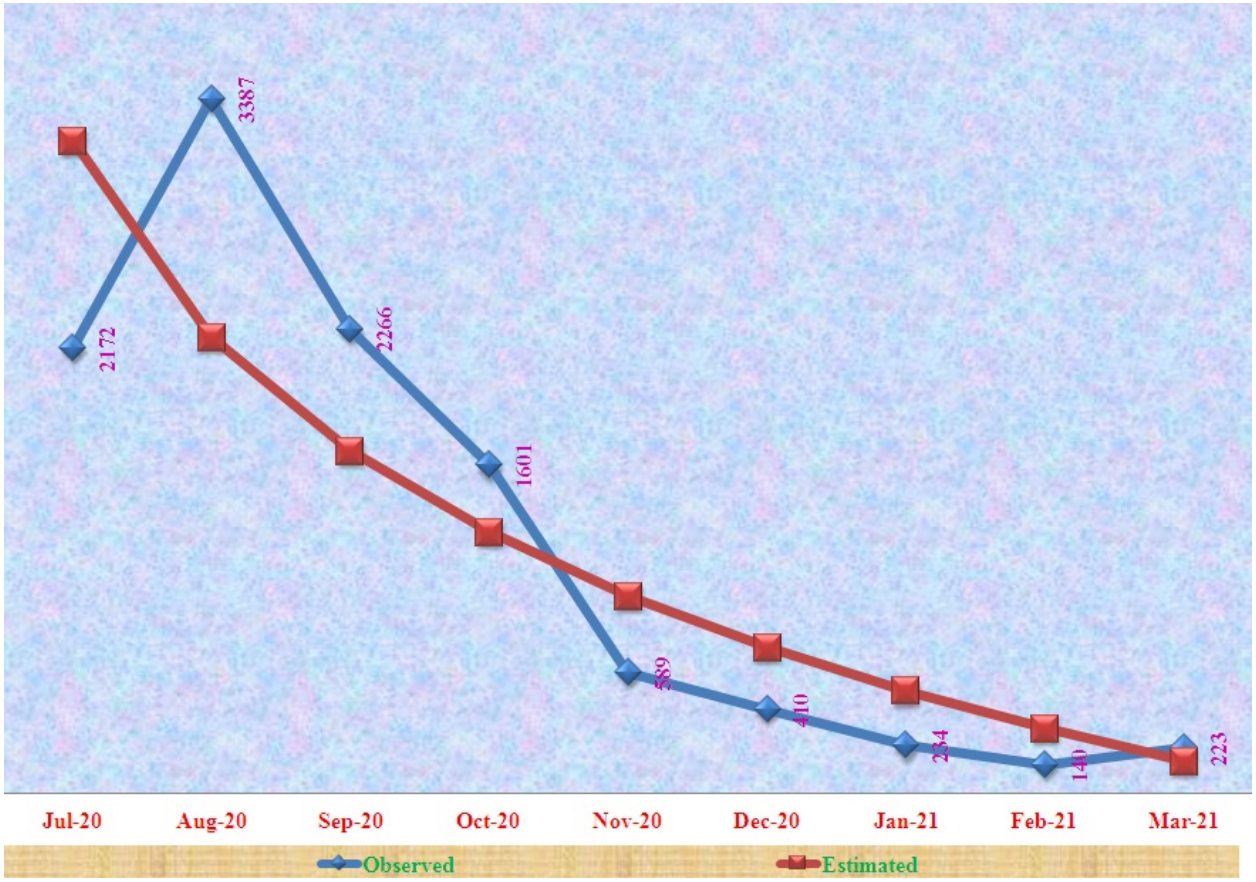

Fig. 2. The number of COVID-19-positive patients registered in different districts of Tamil Nadu during the months of 3rd July-2020 to 31st March,2021.

3.3 Variations between months

To determine the variations across the months under during the study period due to the number of COVID-19-positive infected cases and deaths due to COVID-19, ANOVA tests were carried out for each of the study variables, NCASE and DEATH, and the results are presented in Table 2. The results presented in Table 2 reveal that since the ANOVA tests are highly significant ($p<0.0000$) for both study variables, and highly significant between the months at 1% significance. This means that the differences in the number of positive infections registered in different months are highly significant at 1 % level of significant.

Table 2. Results of test for equality of means of number of COVID-19 infections.

Variables	Method	Df	Value	Probability
NCASE	Anova F-test	(8, 324)	7.572864	0.0000
	Welch F-test*	(8, 130.604)	8.657395	0.0000
DEATH	Anova F-test	(8, 324)	8.115342	0.0000
	Welch F-test*	(8, 130.179)	7.499838	0.0000
***Test allows for unequal cell variances**				

3.4 Pooled OLS regression Model or Constant Coefficients Model

The panel least squares method is employed with the number of deaths due to COVID-19 as the dependent variable and the number of new COVID-19-infected patients as the independent variable. The regression results based on EViews, Version 11, are presented in Table 3.

The estimated model is

DEATH = -4.915224 +0.0161 NASE **(R^2= 92%)**

Table 3. Results of pooled OLS regression or constant coefficients model

Variable	Coefficient	Std. Error	t-Statistic	Prob.
NCASE	0.016085	0.000260	61.80560	0.0000
C	-4.915244	1.292304	-3.803474	0.0002
R-squared	0.920259	Mean dependent var		33.09910
Adjusted R-squared	0.920018	S.D. dependent var		73.33554
S.E. of regression	20.74009	Akaike info criterion		8.908002
Sum squared resid	142380.1	Schwarz criterion		8.930874
Log likelihood	-1481.182	Hannan-Quinn criter.		8.917123
F-statistic	3819.933	Durbin-Watson stat		1.698423
Prob(F-statistic)	0.000000			

The results reveal that the intercept and slopes are very highly significant, and the model F-statistic is also highly significant, with a remarkably high R^2 of 92%. This model explains 92% variations in death by the regressor NCASE. Additionally, for every unit increase in NCASE, DEATH increases by 0.02%, as indicated earlier.

The major problem with this model is that it does not distinguish between the months, nor does it tell us whether the response of total COVID-19 deaths to the explanatory variable over time is the same for all months.

3.5 FE least squares dummy variable (LSDV) model

This FE model is implemented with the dummy variable technique. The model is written as

$$Y_{it} = \alpha_1 + \alpha_2 D_2 + \alpha_3 D_3 + \alpha_4 D_4 + ... + \alpha_{37} D_{37} + \beta X_{it} + \nu_{it}$$

where D_{2i}=1 if the observation belongs to Chengalpattu district and 0 otherwise, D_{3i}=1 if the observation belongs to Chennai and 0 otherwise, D_{4i}=1 if the observation belongs to Coimbatore and 0 otherwise, and so on. Here, the district Ariyalur is considered the baseline, or reference, category. Thus, the intercept α_1 represents the intercept value of the Ariyalur district, and the other α coefficients represent how much the intercept values of the other districts differ from the intercept value of the Ariyalur district. Thus, α_2 shows how much the intercept value of the second district, Chengalpattu, differs from α_1. The sum $(\alpha_1 + \alpha_2)$ gives the actual value of the intercept for Chengalpattu. The intercept values of the other districts can be computed similarly.

The results presented in Table 4 reveal that the FE model is highly significant, with a high R^2 of 93%. The slope coefficient for new COVID-19 infections is also found to be highly significant, which shows that new COVID-19 infections exhibit significant variations in deaths due to COVID-19. All the dummy variable coefficients are found to be non-significant indicating that the pooled regression model values may be informative and appropriate. Additionally, the values of the slope coefficients in Table 4 are also almost same and highly significant in both the model. These two inferences indicates that CCM seems to be better fit than the FE model.

Table 4. Results of FE or LSDV regression model

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2.722354	6.893230	-0.394932	0.6932
NCASE)	0.016846	0.000400	42.14810	0.0000
Root MSE	19.45635	R-squared		0.929401
Mean dependent var	33.09910	Adjusted R-squared		0.920546
S.D. dependent var	73.33554	S.E. of regression		20.67152
Akaike info criterion	9.002452	Sum squared resid		126057.0
Schwarz criterion	9.437015	Log likelihood		-1460.908
Hannan-Quinn criter.	9.175737	F-statistic		104.9599
Durbin-Watson stat	1.796987	Prob(F-statistic)		0.000000

3.7 RE model

The RE model is employed, keeping the number of deaths due to COVID-19 as the dependent variable and the number of new COVID-19 infections as the independent variable, and the test results are presented in Table 5. The results reveal since the value of Rho is 0, the absence of random effect is confirmed.

Table 5. Characteristics of Fitted RE model

Variable	**Coefficient**	**Std. Error**	**t-Statistic**	**Prob.**
C	-4.915244	1.288031	-3.816091	0.0002
NCASE	0.016085	0.000259	62.01063	0.0000
Effects Specification			**S.D.**	**Rho**
Cross-section random			0.000000	0.0000
Idiosyncratic random			20.67152	1.0000

3.8 Hausman test

The Hausman test evaluates whether there is a significant difference between the FE and RE estimators. The results presented in Table 6 reveal that since the estimated chi-square value is significant, we reject the hypothesis that there is no significant difference in the estimated coefficients of the two models. It seems there is correlation between the error term and one or more regressor. Hence, we can reject the random effects model in favor of the fixed effect model. Note, however, as the last part of the Table – 6 shows, not all coefficients differ in the two models (Fixed and Random).

Table 6. Hausman test results (Test cross-section random effects)

Test Summary		Chi-Sq. Statistic	Chi-Sq. d.f.	Prob.
Cross-section random		6.269545	1	0.0123
Cross-section random effects test comparisons:				
Variable	**Fixed**	**Random**	**Var(Diff.)**	**Prob.**
NCASE	0.016846	0.016085	0.000000	0.0123

The cross-section fixeds (as deviations from common intercept) in the context of Fixed effect model are given in the Table – 7. Since all the cross-section fixed effects are non-zero, the presence of fixed effect confirmed. Thus, compared with Constant Coefficients Model, Fixed Effects Model appears to have provided a better model specification. However, it should be confirmed restricted F – test.

Table – 7. Cross-Section Fixed Effects Values

Sr.No.	DISTRICT	Effect
1	Ariyalur	3.992391
2	Chengalpattu	-8.948207
3	Chennai	-16.35356
4	Coimbatore	-25.41786
5	Cuddalore	-7.785048
6	Dharmapuri	0.810342
7	Dindigul	7.866965
8	Erode	-6.307846
9	Kallakurichi	-0.846331
10	Kancheepuram	-3.069413
11	Kanniyakumari	3.763264
12	Karur	1.928434
13	Krishnagiri	3.434316
14	Madurai	18.20761
15	Nagapattinam	5.833080
16	Namakkal	-3.617397
17	Nilgiris	-3.637570
18	Perambalur	5.069764
19	Pudukottai	2.476425
20	Ramanathapuram	10.38651
21	Ranipet	-2.679962
22	Salem	-2.544961
23	Sivagangai	7.761527
24	Tenkasi	8.474297
25	Thanjavur	0.326877

26	Theni	-1.735380
27	Thirupathur	6.444579
28	Thiruvallur	-4.739029
29	Thiruvannamalai	4.605357
30	Thiruvarur	5.405969
31	Thoothukudi	-6.805670
32	Tirunelveli	1.560978
33	Tiruppur	-3.564851
34	Trichy	-0.655398
35	Vellore	8.381224
36	Villupuram	-9.079694
37	Virudhunagar	1.058276

To confirm this, the Redundant Fixed Effects Test has performed, and the results are presented in Table 8. The test results reveal that both the Cross-section F and Cross-section Ch-square tests are non-significant indicating that Constant Coefficients model appears to be more appropriate than the Fixed effect.

Table – 8 : Results of Redundant Fixed Effects test

Effects Test	Statistic	d.f.	Prob.
Cross-section F	1.061102	(36,295)	0.3806
Cross-section Chi-square	40.548276	36	0.2766

Aging to confirm this the inference obtained through Redundant Fixed Effects Test, the Restricted F-test has been carried out and discussed below.

3.9 The Restricted F-test

The Restricted F-test discussed in the section 2.2.7 has been employed and the calculated F^* is found to be less than the F-table value indicating that absence of Fixed Effects and the intercepts of cross-sectional units are non-significant at 5% level of significance as per the results given in the Table 8 and hence the CCM is better than the FE model. So, in the Panel Regression model fails to find the variations in death due to number of cases infected whereas the CCM found suitable for the same. Since both the tests viz. Redundant Fixed Effect test and the Restricted F-test values are non-significant, the Fixed effect model is rejected over the CCM.

The estimated Pooled OLS regression model or Constant Coefficient Model is,

$$\mathbf{DEATH = -\ 4.915244^{**} + 0.016085^{**}\ NCASE} \qquad \mathbf{(R^2=92\ \%)}$$

(** indicates $p<0.0000$)

This implies that if every units increases of NCASE the death rate would be increased by 1.6 %.

4.CONCLUSIONS

During the study period (3rd July-2020 to 31st March 2021), the highest numbers of infections,181817 and the highest number of deaths, 3387 were registered in the month of August – 2020, the lowest were in the month of February-2021.Overall during the study period, 78,6,990 infected cases and 11,022 deaths were registers In Tamil Nadu. The interesting results obtained in this paper is that even though the data is Panel type, none of the panel regression model was found suitable whereas the Constants Coefficient Model (Pooled Regression Model) was found suitable to study the relationships between number of covid infects and deaths. The average death due to COVID-19 was about 1.6 %.

ACKNOWLEDGEMENTS

Authors are highly thankful to Dr. K. Senthamari Kannan, Senior Professor and Head, Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli-621 012, Tamil Nadu State, India, for helpful discussion and encouragement while preparing this paper.

REFERENCES

[1]. Al-Rousan, N., Al-Najjar, H., Data analysis of coronavirus COVID-19 epidemic in South Korea based on recovered and death cases. Journal of Medical Virology,2020,92, 1603-1608.

[2]. Al-Rousan, N., Al-Najjar, H., The impact of Chinese Government Plans on Coronavirus CoVID-19 Spreading and its Association With Weather Variables in 30 Chinese Provinces: Investigation, European Review for Medical and Pharmacological Sciences, 2020,24 (8), DOI:10.26355/eurrev_202004_21042

[3]. Arumugam, R., and M.Rajathi, A Markov Model for Prediction of Corona Virus COVID-19 in India - A Statistical Study, Journal of Xidian University, 2020, 14(4), 1422-1426, http://xadzkjdx.cn/ https://doi.org/10.37896/jxu14.4/164

[4]. Baltagi, B.H., Econometric Analysis of Panel Data, 2001,2nd edition. Wiley, New York, NY.

[5]. Bertozzi, A.L., E.Franco, G. Mohler, M.B. Short, and D. Sledge, The challenges of modeling and forecasting the spread of COVID-19, Proceedings of the National Academic

Services of the United States of America, 2020, 117 (29) : 16732-16738, https://doi.org/10.1073/pnas.2006520117

[6]. Breusch, T., and Pagan, A.R., The Lagrange Multiplier Test and its application to model specification in Econometrics. Review of Economic Studies, 1980,47, 239-253.

[7]. Bhaumik,S.K., Principles of Econometrics,2017,2nd Edition, Oxford University Press.

[8]. Chu ,J., A statistical analysis of the novel coronavirus (COVID-19) in Italy and Spain. PLoS ONE, 2021, 16(3): e0249037. https://doi.org/ 10.1371/journal.pone.0249037

[9]. Gondauri, D., Batiashvili, M., The study of the effects of mobility trends on the statistical models of the COVID-19 virus spreading. Electronic Journal of General Medicine, 2020,17, em243.

[10]. Gondauri, D., Mikautadze, E., Batiashvili, M., Research on COVID-19 virus spreading statistics based on the examples of the cases from different countries. Electronic Journal of General Medicine,2020, 17, em209.

[11]. Gujarati, D.N., Porter, D.C., Sangeetha, G., Basic Econometrics, 2017, 5th edition. McGraw Hill Education, New York, NY.

[12]. Gecili E, A.Ziady, and R.D. Szczesniak ,Forecasting COVID-19 confirmed cases, deaths and recoveries: Revisiting established time series modeling through novel applications for the USA and Italy. PLoS One. 2021,16(1), doi: 10.1371/journal.pone.0244173

[13]. Hadri, K., Testing for stationarity in heterogeneous panel data. The Econometrics Journal,2000, 3, 148-161.

[14]. Hausman, J.A., Specification tests in econometrics. Econometrica,1978, 46, 1251-1271.

[15]. Hsiao, C.,Analysis of Panel Data,2003, 2nd edition. Cambridge University Press, Cambridge, MA.

[16]. Katris, C., A time series-based statistical approach for outbreak spread forecasting: application of COVID-19 in Greece. Expert Systems with Applications,2021, 166, 114077.

[17]. Kumar, A., Roy, R., Application of mathematical modeling in public health decision making pertaining to control of COVID-19 pandemic in India. Epidemiology International, 2020, 05, 23-26.

[18]. Levin, A., Lin, C.F., Chu, C.S.J., Unit root tests in panel data: asymptotic and finite-sample properties. Journal of Econometrics,2002, 108, 1-24.

[19]. Mittal, S., An exploratory data analysis of COVID-19 in India. International Journal of Engineering and Technical Research,2020, 9, 580-584.

[20]. Mahanty, M., K. Swathi, K.S.Teja, P. H. Kumar, A. Sravani, Forecasting the Spread of COVID-19 Pandemic with Prophet,2021, Revue d'Intelligence Artificielle, 35(2),115-122. https://doi.org/10.18280/ria.350202

[21]. Rajarathinam, A., and P.Tamilselvan, Autoregressive Distributed Lag Model of COVID-19 Cases and Deaths, Appl. Math. Inf. Sci,2021 (in press).

[22]. Rajarathinam, A., and P.Tamilselvan, Vector error correction modeling of covid-19 infected cases and deaths, Journal of Statistics Applications & Probability,2021 (in press).

[23]. Tiwari, V., N.Deyal, and N.S.Bisht, Mathematical Modeling Based Study and Prediction of COVID-19 Epidemic Dissemination Under the Impact of Lockdown in India. Front. Phys. 2020, 8:586899. doi: 10.3389/fphy.2020.586899

[24]. Ogundokun, R.O., Lukman, A.F., Kibria, G.B.M., Awotunde, J.B., Aladeitan, B.B., Predictive modelling of COVID-19 confirmed cases in Nigeria. Infectious Disease Modelling,2020, 5, 543-548.

[25]. Rosario, D.K.A. , Y.S. Mutz, P.C. Bernardes , C.A. Conte-Junior, Relationship between COVID-19 and weather: Case study in a tropical country, International Journal of Hygiene and Environmental Health, 2020,229,113587 https://doi.org/10.1016/j.ijheh.2020.113587,

[26]. Takele, R., Stochastic modelling for predicting COVID-19 prevalence in East Africa countries. Infectious Disease Modelling,2020, 5, 598-607.

[27]. Zuo,M.,S.K. Khosa, Z.Ahmad, and Z.Almaspoor, Comparison of COVID-19 Pandemic Dynamics in Asian Countries with Statistical Modeling, Computational and Mathematical Methods in Medicine, 2020, vol. 2020, https://doi.org/10.1155/2020/4296806,

Figures

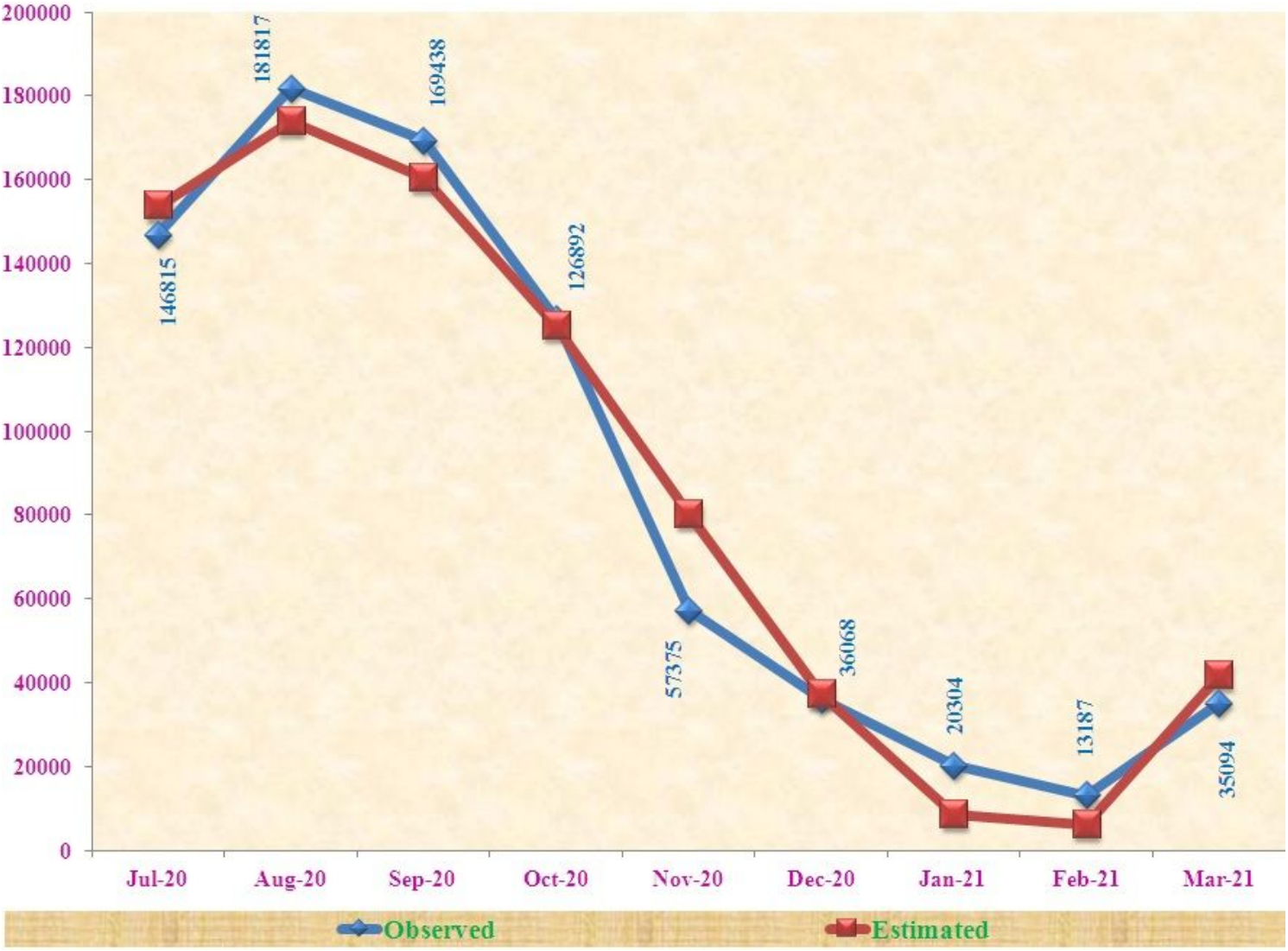

Figure 1

The number of COVID-19-positive patients registered in different districts of Tamil Nadu during the months of 3rd July-2020 to 31st March,2021.

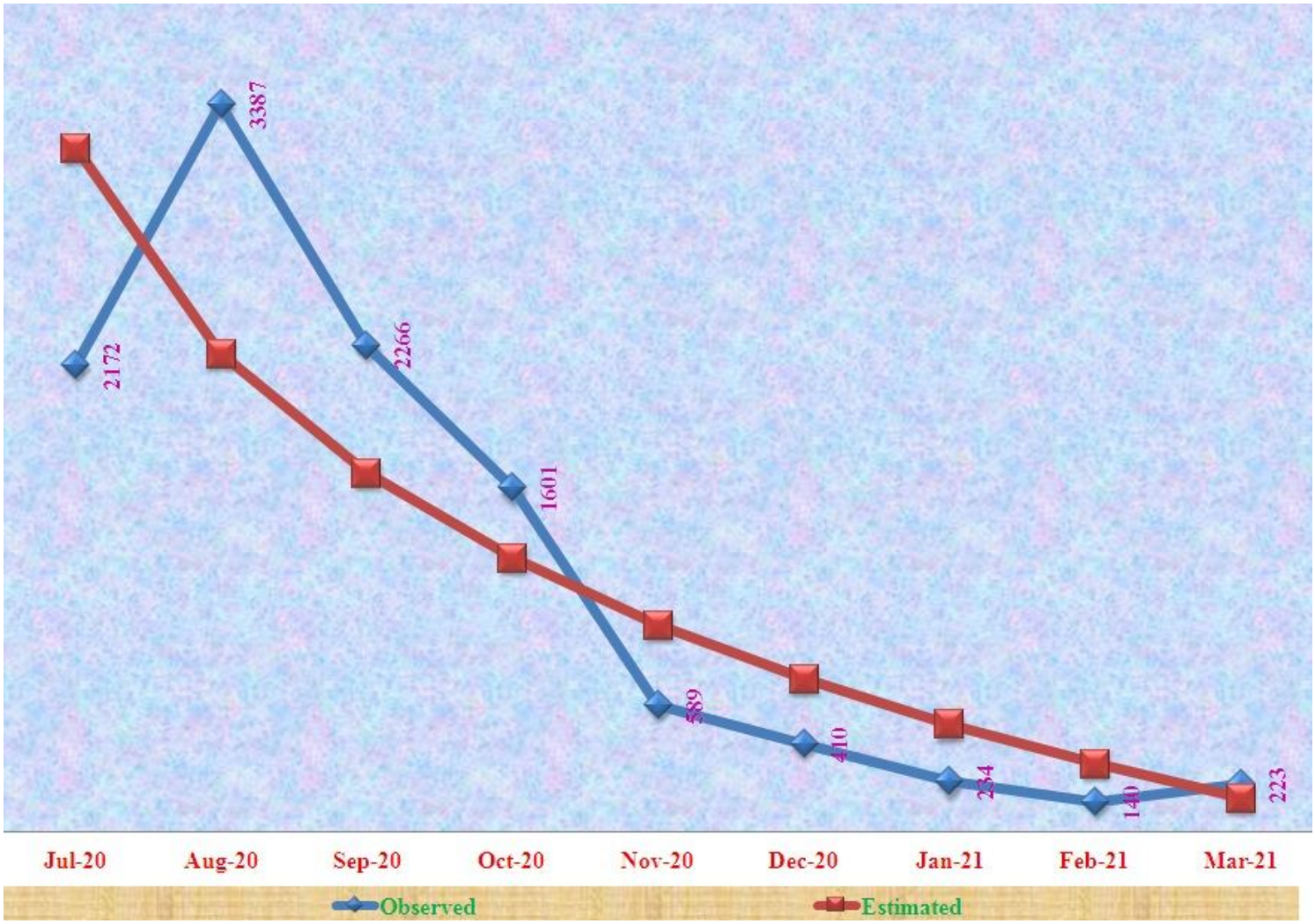

Figure 2

The number of COVID-19-positive patients registered in different districts of Tamil Nadu during the months of 3rd July-2020 to 31st March,2021.