SPRINGER NATURE Link

Log in

Menu   Search   Cart

Home   SN Computer Science   Article

Analysis of COVID-19 Datasets Using Statistical Modelling and Machine Learning Techniques to Predict the Disease

Original Research   Published: 10 January 2024

Volume 5, article number 181, (2024)   Cite this article

Download PDF ⤓

Access provided by Swami Keshvanand Institute of Technology Management and Gramothan

SN Computer Science

Aims and scope

Submit manuscript

Senthil Kumar Nramban Kannan, Bhanu Prakash Kolla ✉, Sudhakar Sengan, Rajendiran Muthusamy, Raja Manikandan, Kanubhai K. Patel & Pankaj Dadheech

Senthil Kumar Nramban Kannan

Department of Computer Science and Engineering, School of Computing, Institute of Science and Technology (Deemed to Be University), Vel Tech Rangarajan Dr. Sagunthala R&D, Chennai, Tamil Nadu, 600062, India

View author publications

You can also search for this author in

PubMed | Google Scholar

## Abs

Susta
Coro
pand
chall
aims
diag

investigation of the literature and a research study is conducted to find suitable mathematical methods and measure how they impact prediction models. This paper analyses COVID-19 pandemic deaths, confirmed cases, and recovered individuals using Time Series Analysis (TSA) to study the disease's impacts and understand the TAS. A forecasting model can predict future COVID cases by analyzing trends in time-series and connecting global changes with government restrictions. Since higher predictive accuracy is the limitation of ensemble learning algorithms, better ML approaches are proposed here. Autocorrelation plots clearly showed the results executed for the considered objectives. The hybrid ARIMA algorithm proposed in this work proved adequate results.

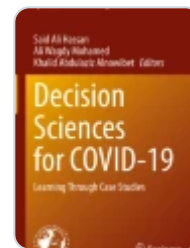## Similar content being viewed by others
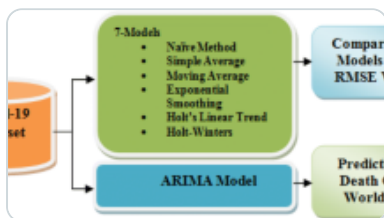
Application of machine learning time series analysis for prediction COVID-19 pandemic

Article | 24 October 2020
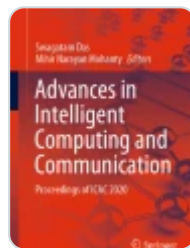
COVID-19 Pandemic Analysis and Prediction Using Machine Learning Approaches in India

Chapter | © 2021

A Multi-Step Predictive Model for COVID-19 Cases in Nigeria Using Machine Learning

Chapter | © 2022

Use our pre-submission checklist →

Avoid common mistakes on your manuscript.

Introduction

When the Coronavirus Disease 2019 (COVID-19) pandemic stuck, many countries were clueless about dealing with it. Many countries were fast enough to close their borders to restrict the spread of the disease. International borders were sealed. Many countries imposed a nationwide lockdown where people were allowed to move out of their houses only in the case of a medical emergency. In public places, only people with masks are permitted by the rule of social distancing. This pandemic has exposed the fragility of

healthcare systems throughout the world. The statistics relating to COVID-19, such as the total number of confirmed, recovered, and fatal cases, are being investigated to take proactive measures to reduce the spread of this disease. These cases reflect the number of people who have died from the viral infection.

Time-series data, such as the number of COVID-19 cases, can be analyzed in several approaches to find developments while developing forecasts. This analysis is done using Time Series Analysis (TSA), so that the data trend can be determined and future trends can be found. Time-series helps to understand the trend and nature of the given data, which will be helpful in forecasting. Because of the extreme seriousness of COVID-19, its quick spread, and the logistical challenges associated with monitoring and diagnosing patients, the TSA tests are not recommended for researching infectious diseases. Are these techniques still effective in the COVID-19 scenario? Is there a more efficient method for illustrating disease trends and forecasting future trends? Internet and open-source data and modelling utilities benefit society without sustainability concerns.

This study aims to identify the TSA approach that provides the most accurate predictions of COVID-19 case trends. Several individual objectives need to be succeeded in before we can reach our primary objective:

    a) Select appropriate TSA methods, COVID-19 data formats, and case trend modelling parameters.

    b) The test selected methods for predicting case trends, comparing their performance to the baseline and each other, and selecting the most accurate method.

The article continues with the following structure: the section "Literature survey" focuses on the literature surrounding TSA models. Then, in the section "Methodology", develop the recommended *ARIMA* Model for TSA. The section "Results and analysis" contains both the findings and the discussion of those findings. The conclusion of the work is presented at the end of the article in the section "Conclusion and future work", along with a few ideas for possible further research.

Literature Survey

Many researchers have focused on the different dimensions of human life that were affected by this pandemic. There was a considerable shift in the teaching–learning process for the students and the teachers. A case study was conducted and accepted for publication [1, 2] to examine and evaluate the pandemic's impact on the faculty members at engineering schools and the students who attended classes in those educational institutions. Rumors and fake posts have created panic during this uncertain pandemic. For this, the post was collected from Facebook using the Natural Language Processing (NLP) module. The regression algorithm used for this purpose results in less accuracy as it is challenging to get Facebook post data. Also, it is not easy to evaluate posts in regional languages [3]. Some researchers have also used the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm to check for fake text from almost 50000 research papers [4].

Climate changes that occurred during COVID-19 are also studied, and it is observed that due to the lockdown where movement was restricted, pollution was drastically reduced. [5]. The effect of COVID-19 on the Indian Power sector and the impact of sectorial parameters, such as consumer demand, the delivery process, and financial resources, has been investigated in [6].

A study evaluated how the community complies with COVID-19 restrictions using Twitter data [7]. For this analysis, Twitter data were collected from Indonesia, and words like COVID-19 panic were searched. These data are analyzed for two classes: obedient and non-obedient. The logistic regression and the forest are more accurate than the decision tree.

This pandemic has also affected the stock market trends. In [8], the telecommunication stock trend during COVID-19 is studied using regression techniques from the Indonesian stock market. K nearest regression tree gives the best results. Omar [9] proposed sharing an economic model for the current pandemic. It addresses two questions: how sharing can be done to avoid an infection and which sharing practices limit the COVID-19 spread.

The severity of the pandemic can be reduced to a considerable level if people with respiratory diseases, pregnant women, and older people who are more prone to catching the infection are monitored, and alerts are given to all such people. For this purpose,

Machine Learning (ML) algorithms like Deep Neural Networks (DNN), eXtreme Gradient Boosting (XGBOOST), and Logistic Regression (LR) can be used [10].

Josimar studied the effects of COVID-19 in Paris by using Twitter data. Many countries in Europe share boundaries. Italy was number one in COVID-19 cases, followed by France. Data were collected from the Twitter Application Programming Interface (API) for this analysis. After analyzing the data, the conclusion is that many tweets were found during the control phase about health issues and emergencies. After this lockdown phase, tweets tended towards economic emergencies [11].

The COVID-19 data are non-stationary data and can consist of noisy signals. For accurate prediction of the trend from such data in Italy, entropy is used [12]. Chiara [13] proposed a compartmental model for this epidemic and studied the effects of lockdown, so that hospital capacity can be quantified. This model uses a Bayesian model called Conditional Robust Calibration (CRC).

Bangladesh was also worst affected due to the COVID-19 pandemic. To predict the future trend of this pandemic in Bangladesh, data are collected constantly. For prediction, linear regression and K nearest neighbors are used [14].

The SEIR model is proposed to study the pandemic behavior in Indonesia with parameters like distribution, cure rate, mortality rate, communication rate, and movement. To understand the pandemic behavior in Indonesia, its COVID data are compared with other countries. SEIR is a type of dynamic modelling in Epidemiology. '$S(t)$' represents the number of people who are susceptible to the disease at the time 't', '$E(t)$' symbolizes the number of people who are subjected to the disease at the time 't', '$I(t)$' embodies the number of individuals who are spreading at the time 't', and '$R(t)$' stands for the number of people who are either immune to the disease or died at the time 't'. Most epidemics have an identifiable general form [15].

As can be noticed from the data presented in the preceding Fig. 1, the rate of increase or decrease in the number of new diseases is directly proportional to the number of people who are both highly susceptible to the disease and have been subjected to it. In terms of mathematical information, Eq. (1)

Fig. 1
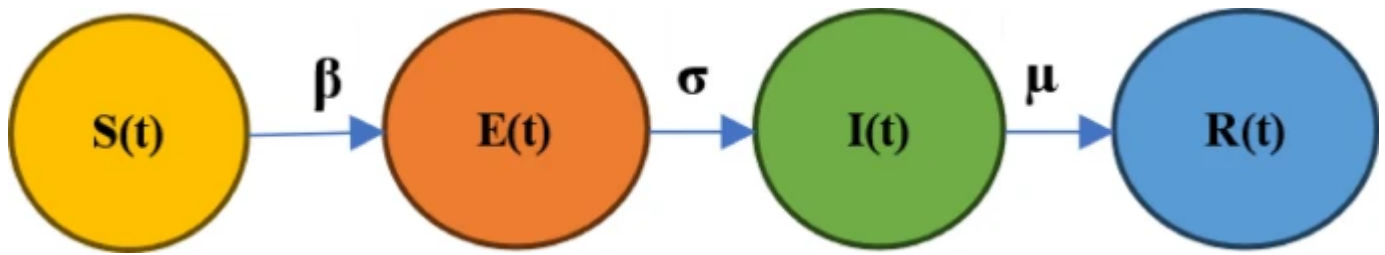
SEIR epidemiological model

$$\frac{ds}{dt}=-\beta SI.$$

(1)

The value of 'B' sets the rate at which infection occurs (per diagnosed individual per at-risk individual), Eq (2)

$$\frac{dI}{dt}=-\beta SI-\mu I.$$

(2)

The value of 'μ' determines the rate of recovery. What this implies is that an average infectious period is '$1/_\mu$'.

$$\frac{dR}{dt}=\mu I.$$

(3)

The drawback of the Susceptible-Exposed-Infectious-Recovered (SEIR) model was that the number of pre-symptomatic and asymptomatic individuals was not considered separately in the Exposed classification. In [16], a modified SEIR model is used to model a pandemic situation in Thailand. This model has ordinary differential equations.

To study the impact of this pandemic, the TSA of the countrywide and statewide data is considered. As this pandemic started in Wuhan in China, the trend of deaths, positive

cases, and recovered cases are considered for the TSA.

TSA examines data points collected regularly to uncover underlying patterns and trends. It refers to studies recorded on an actual variable called time. The intervals can be years, months, days, or hours. To perform the TSA of any problem, a model representing a time-series is first built, and then, the model is validated. After these two steps, a model can be used for predicting future trends.

The study of COVID-19 datasets analyzes the data with TSA to predict the future effect of the Coronavirus globally or separately. Four prediction models are analyzed: Naïve, Holt's Linear trend method, Holt's winter Seasonal method, and Autoregressive Integrated Moving Average (ARIMA).

The TSA is done to study the effects of lockdown in India. The ARIMA model is used for forecasting the trend in the COVID-19. The number of positive cases and tests conducted was measured for investigation [16].

The model developed contains linear and non-linear time-series models and neural network autoregressive models to predict deaths accurately, recovered cases, and vaccination cases.

The coronavirus originated in Italy and Spain, which were also the very first nations in Europe to be affected. To identify the purpose of modelling the cumulative incidence of COVID-19 and computing approximate values of the primary development number, rate of increase, and doubling a period, the highly susceptible Infection-causing Recovered Model and a log-linear regression model have been used as modelling tools. The results of this research demonstrate that the predictive value of log-linear regression presents an improved fit. In Indonesia, tries are being made to predict the future development of coronavirus disease using ML models [17]. The accuracy of different datasets that include cases that have been confirmed, loss of life, and cases that were recovered is analyzed in this study using Facebook's PROPHET Prediction and ARIMA models. The comparison of the performance of these two models shows that PROPHET model accuracy is better than ARIMA [18].

In North Africa, a statistical relation was developed between the number of deaths resulting from COVID-19 and the number of cases that were confirmed. This linear statistical relationship was developed between death and confirmed cases in Morocco, Algeria, and Tunisia in three African regions. The COVID-19 database was determined so that proper corrective actions can be taken. It is observed that despite corrective actions, there are essential clusters of COVID-19 present. The statistical relation between the number of confirmed, recovered, and death cases was established for selecting a proper prediction technique. It is observed that a linear relation exists between confirmed cases and death cases, and there is a non-linear relation between confirmed cases and recovered cases [19].

Prediction models, such as exponentially increasing, smoothing, Autoregressive Integrated Moving Average (ARIMA), and Seasonal Autoregressive Integrated Moving Average (SARIMA), can project the number of COVID-19 cases in India. In India, the effects of holidays have been associated with an increase in COVID-19 cases, and the SARIMAX model is used to account for this phenomenon. It is observed that the ARIMA model gives the best accuracy. To assess the accuracy of these models, the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are used.

Epidemiology is the study of infectious diseases in a population by considering all aspects of a pandemic, like spread, control, and vaccination strategy. The SIR model of mathematical epidemiology is used for modelling COVID-19 data in Morocco. As infections have random behavior, stochastic modelling of this data gives results remarkably close to accurate data (see Table 1).

Table 1 Literature survey

Methodology

Data Set

Johns Hopkins University's Center for Systems Science and Engineering (CSSE) is the TSA's data repository. Every day, the current information is added to this dataset. The dashboard provides information about the spread of disease around the world. In addition, it provides information on the total number of vaccination doses administered across all

countries and states, as well as the number of people who contracted the disease, recovered from it, and passed away as a result.

The dataset is maintained since 21st January 2020. This paper attempts to implement the TSA for deaths, recovered, and infected cases in China, USA, and Australia. Table 2 shows the sample dataset of the US.

Table 2 The sample dataset of the United States (US)

Data Repair

As observed from the dataset, there is more than one administrative region in each State. Because we plan to conduct the study statewide, we have determined the total number of accidental deaths, recovered cases, and newly identified infections in each state.

Algorithm: ARIMA Model

Step 1. *Data Preparation:* Collect and preprocess Time Series Data (TSD), ensuring that it is stationary if necessary.

Step 2. *Order Selection:* Choose the appropriate p, d, and q values.

Step 3. This was done through techniques like ACF and PACF plots and tools like the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).

Step 4. *Model Fitting:* Use the chosen p, d, and q values to fit the ARIMA model to your data using software libraries like stats models in Python.

Step 5. *Model Evaluation:* Evaluate the model's performance using residual analysis and Mean Squared Error (MSE) techniques.

Step 6. *Forecasting*: Once the model is evaluated and deemed satisfactory, we can use it to make future predictions by providing the required lagged values.

Algorithm for Time Series Autocorrelation

Step 1.

Data Preparation: Collect and organize TSD. This could be daily, weekly, or any other data interval related to COVID-19 cases.

Step 2. *Calculate Autocorrelation Function (ACF)*: A correlation between a time-series and its previous values can be measured using the autocorrelation function (ACF).

Calculate the ACF: Follow These Steps:

- Compute the mean (μ) of the TSD.

- For each lag (k), calculate the autocovariance between the original series and the series shifted by k time units.

- Normalize the autocovariance by dividing it by the variance of the original series. This gives you the autocorrelation coefficient for that lag.

- The autocorrelation coefficient at lag k is given by:

- ACF(k) = Variance of the original series/Autocovariance at lag k.

- *Interpret the ACF Plot*: Make a graph in which the x-axis and the autocorrelation coefficients indicate that the y-axis embodies the lags. This plot is known as the autocorrelation plot. Interpret the plot to identify significant autocorrelation values and patterns. Peaks or spikes in the ACF plot can indicate periodic behavior in the data.

- *Partial Autocorrelation Function (PACF)*: The limited autocorrelation function, or PACF, finds the correlation between a time-series and its lagged values while considering the impact of temporary lags. It helps identify the direct effect of each lag on the current value. Calculate and plot the PACF like ACF to understand the relationships between different lags.

- *Interpret the PACF Plot*: Like the ACF plot, the PACF plot helps you identify the most influential lags for predicting future values. Significant spikes in the PACF plot can guide your choice of lag for predictive modelling.

- *Model Selection and Prediction*: We can choose an appropriate prediction model based on the insights from the ACF and PACF plots. For example, if we observe a high autocorrelation at lag 7 (weekly pattern), we might use a lag of 7 to predict the following week's value. The current time step's value is predicted using the previous time step's value; this is a fundamental autoregressive model.

AutoARIMA Parameter Settings

AutoARIMA is a Python function that automatically selects optimal parameters for an ARIMA model based on TSD, with standard parameter settings included.

Step 1. Start_p, Start_q: These parameters define the range of possible values for the order of autoregressive (p) and moving average (q) terms. By default, they are set to '2'.

Step 2. Max_p, Max_q: These parameters define the maximum values for (p, q). By default, they are set to '5'.

Step 3. The parameter 'd', set to None by default, specifies the distinguished order for time-series papers, allowing AutoARIMA to determine the optimal value automatically.

Step 4. The parameter 'seasonal' indicates whether to include seasonal components in the model, which is default set to 'False'.

Step 5. The seasonal parameter 'm', set to '1', indicates the number of periods in each season when the seasonal option is True.

Step 6. Start_P, Start_Q: These parameters define the range of possible values for seasonal autoregressive (P) and moving average (Q) terms when seasonal is True. By default, they are set to 1.

Step 7. Max_P, Max_Q: These parameters define the maximum values for P and Q when seasonal is True. By default, they are set to 2.

Step 8. D: When the seasonal flag is set to True, the value of this parameter symbolizes the order of changing seasons that must be performed on the time-series before it can be considered stationary. By default, it is set to None.

Step 9. Trace: This parameter controls whether to print debugging information during model fitting. By default, it is set to False.

These are just some of the commonly used parameters for AutoARIMA. Adjust these settings based on your specific requirements and the characteristics of test TSD.

Results and Analysis

Specifications Regarding Hardware and Software

Clarifying the hardware and software used in the experiments is required to provide a visual representation of the experiment setting to allow users to replicate the results and create an environment identical to or like the one used in the procedures. As the primary workstation, an iMac 21.5" "Core i5" 2.3 is used for this research. Tables 3 and 4 list the required technical details for the hardware and the software versions.

Table 3 Hardware technical specifications

Table 4 Software versions

Analysis of Confirmed Cases in China

In an analysis of confirmed cases in China, we can see that the data for confirmed cases are displayed province-wide. The data set of confirmed cases in China is provided here in Table 5. The following data frames present a comparison between the number of confirmed cases, the number of deaths, and the number of cases that were recovered. Table 6 indicates the data set for recovered cases, and Table 7 highlights the data set for the examined death cases.

Table 5 Data set of confirmed cases in China

Table 6 Data set of recovered cases

Table 7 Data set of death cases

Like the state of China, the number of cases that have been verified and deaths and recoveries is more significant than usual. As a result, the total number of cases throughout the province is shown in Table 8. As this number is less, the data set is summarized countrywide for all other countries.


Table 8 Data set of all confirmed cases in China

While dealing with a time-series dataset, the data may contain the Date, Month, Day, and time in any format.

To study the time-series more clearly, subplots are plotted for every province. From the plots and time limits shown in Figs. 2 and 3, we can conclude that Hubei province has the highest number of confirmed COVID-19 cases. The second largest number of cases is observed in Guangdong and Shanghai city.

Fig. 2
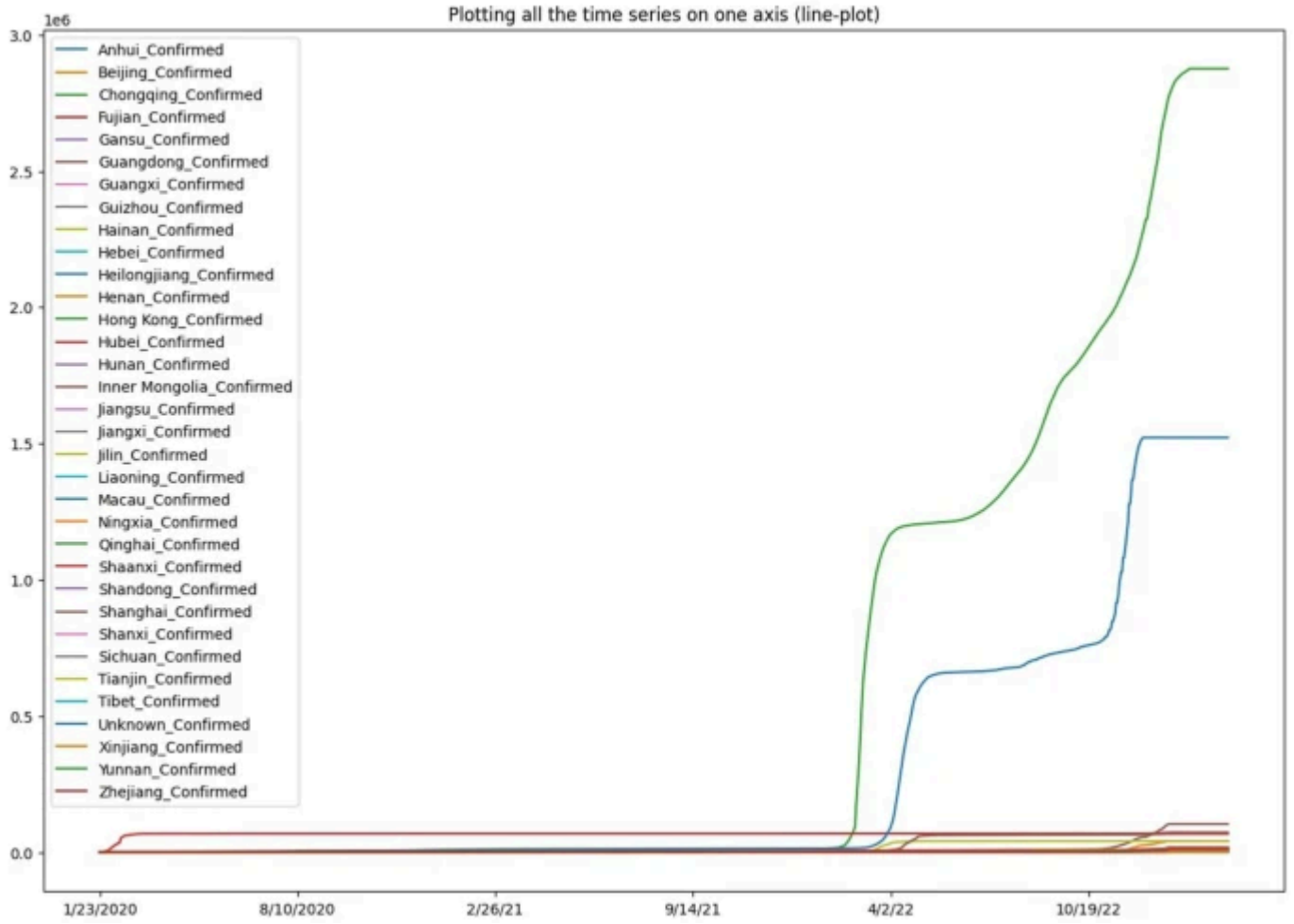
Time Series Plot (TSP) of confirmed cases in China

Fig. 3
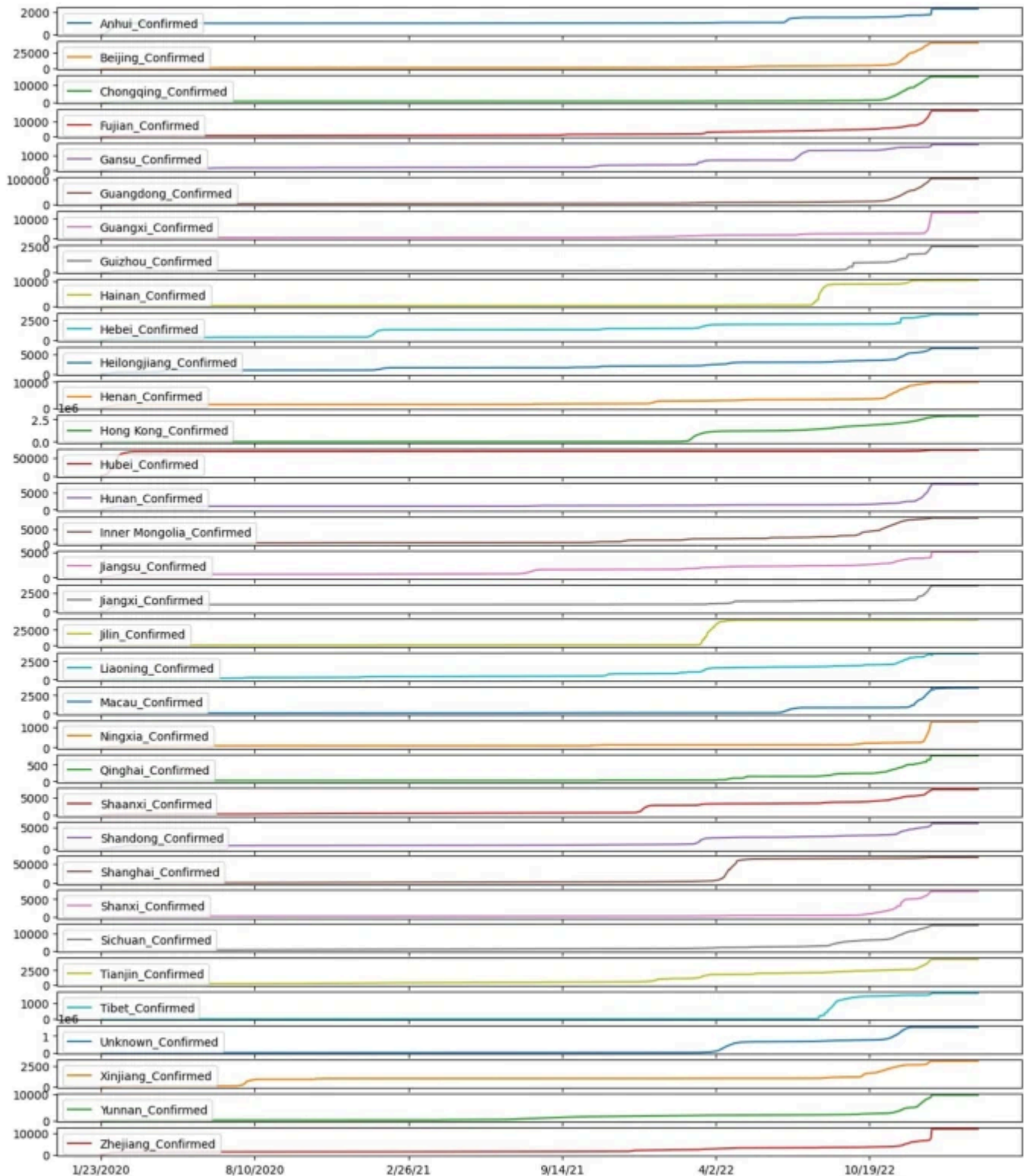
TSP of province-wide confirmed cases in China

Time-series has three essential components: trend, seasonality, and error. For the COVID-19 data, it can be observed that the number of cases varies irregularly due to imposed restrictions and relaxations given during the lockdown. Therefore, stochastic models can model them.

Analysis of Recovered Cases in China

The data form is plotted province-wide for recovered cases to analyze the TSD for recovered and death cases in China (Fig. 4).

Fig. 4
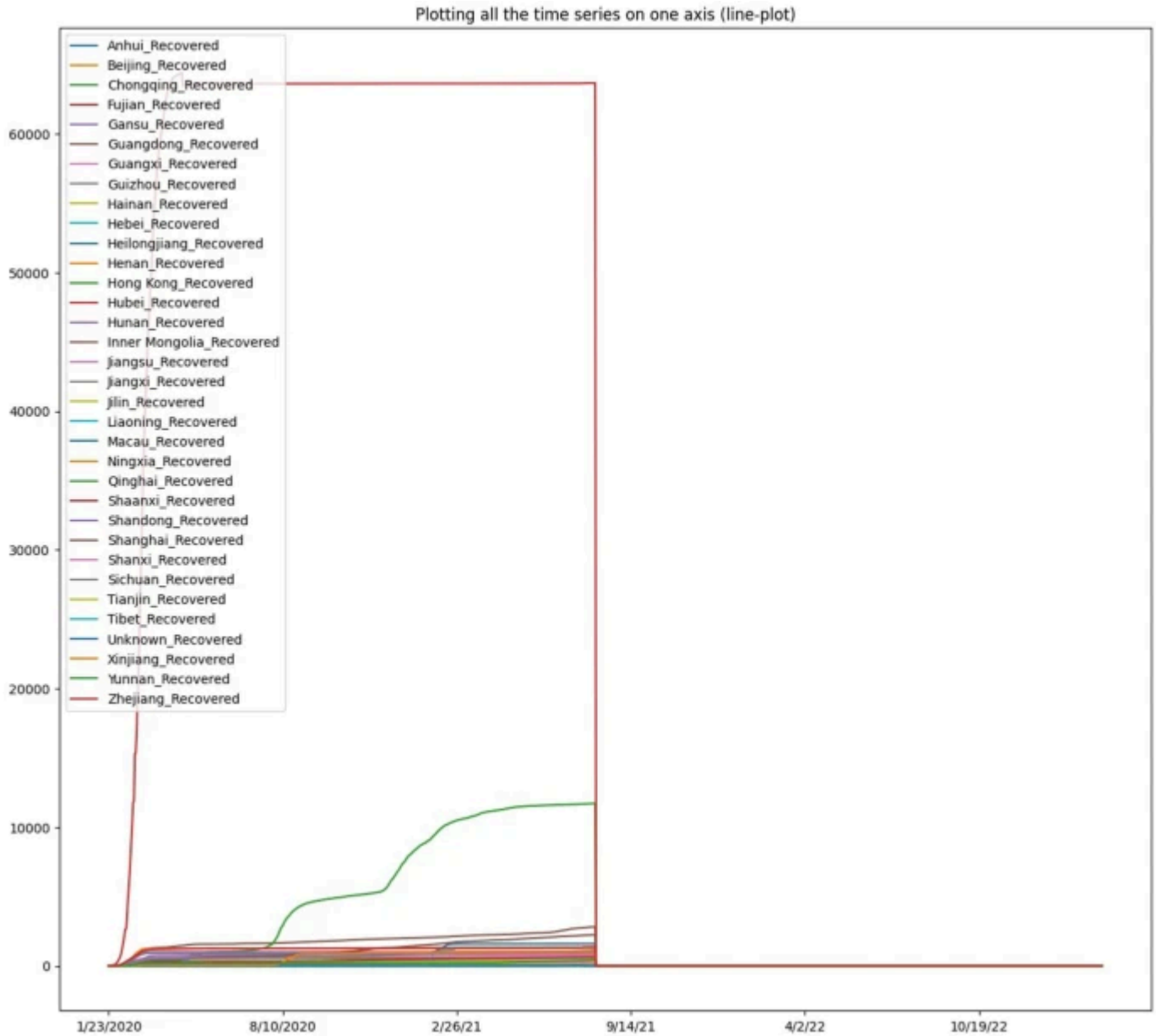
TSP of recovered cases in China

To analyze the data of recovered cases for every province in China, subplots are plotted for every province in China (Fig. 5).

Fig. 5
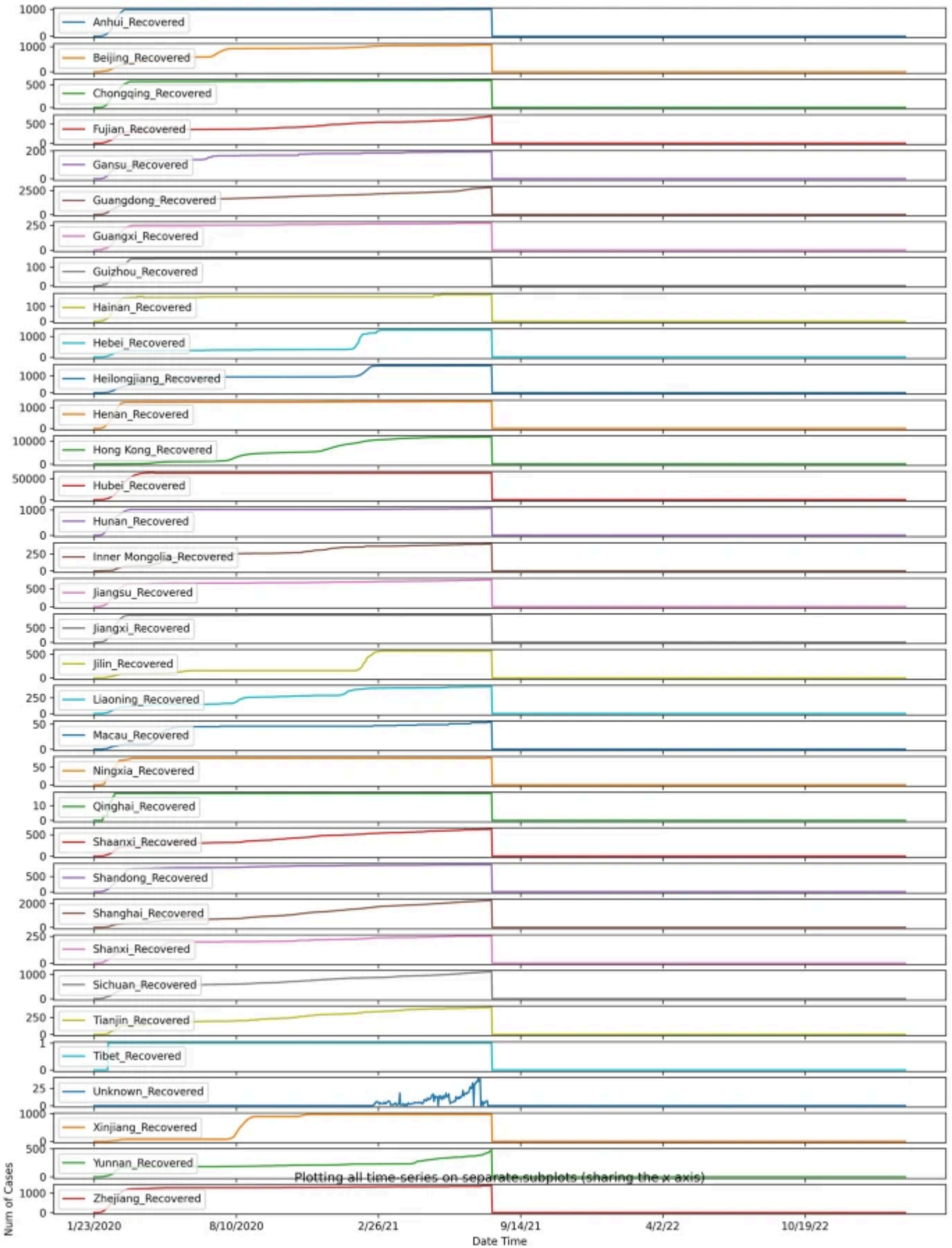
Province-wide TSP for recovered cases in China

From the above plots, we can conclude that the highest number of recovered cases are in Hubei province, followed by Guangdong and Shanghai City.

Analysis of Death Cases in China

To analyze the number of deaths in China according to each province, TSP for the number of death cases is plotted. There is initially a steep rise in the number of cases (Fig. 6).

Fig. 6
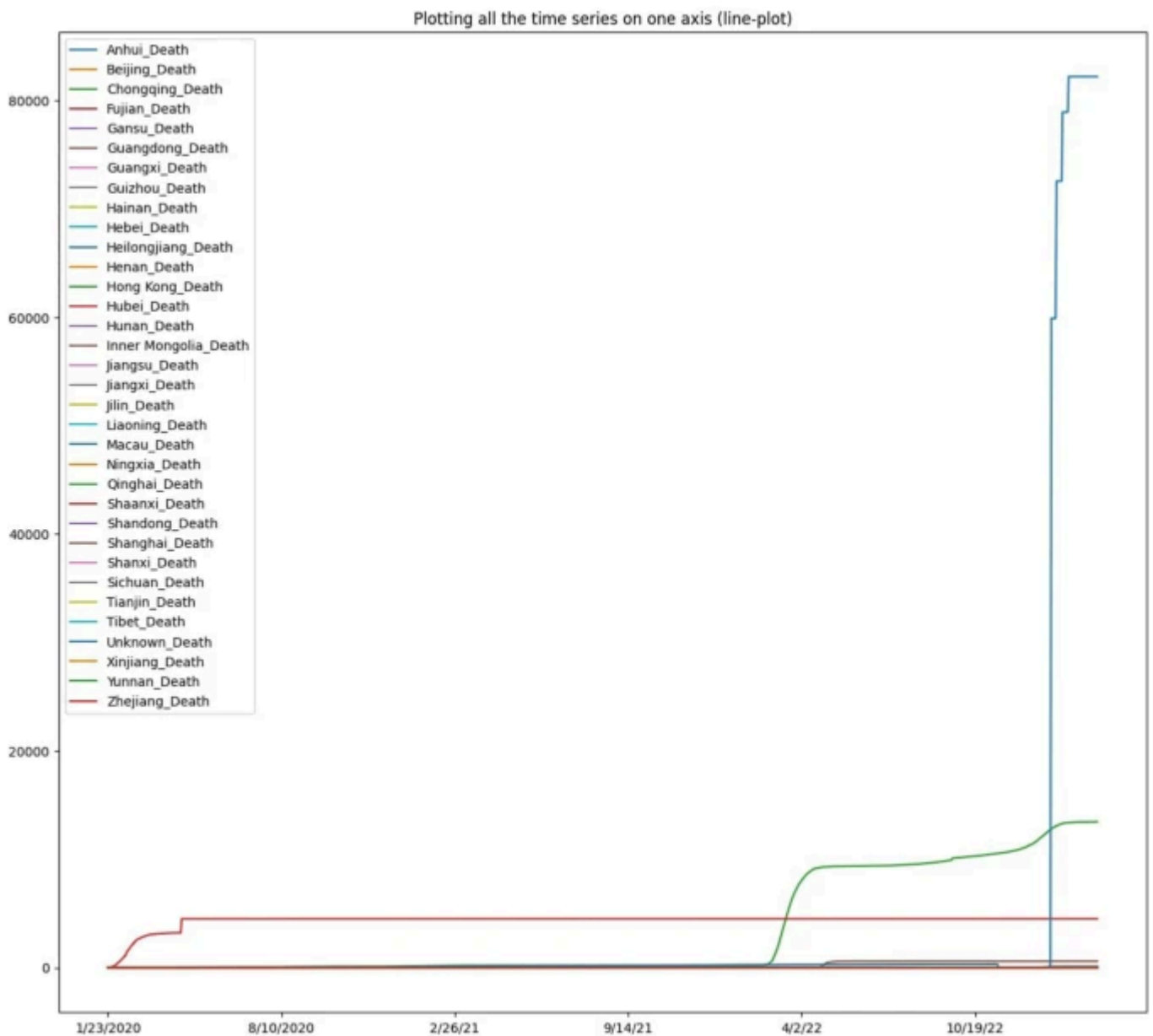
TSA of the number of death cases in China for each province

Figure 7 clearly shows the number of deaths in each province.

Fig. 7
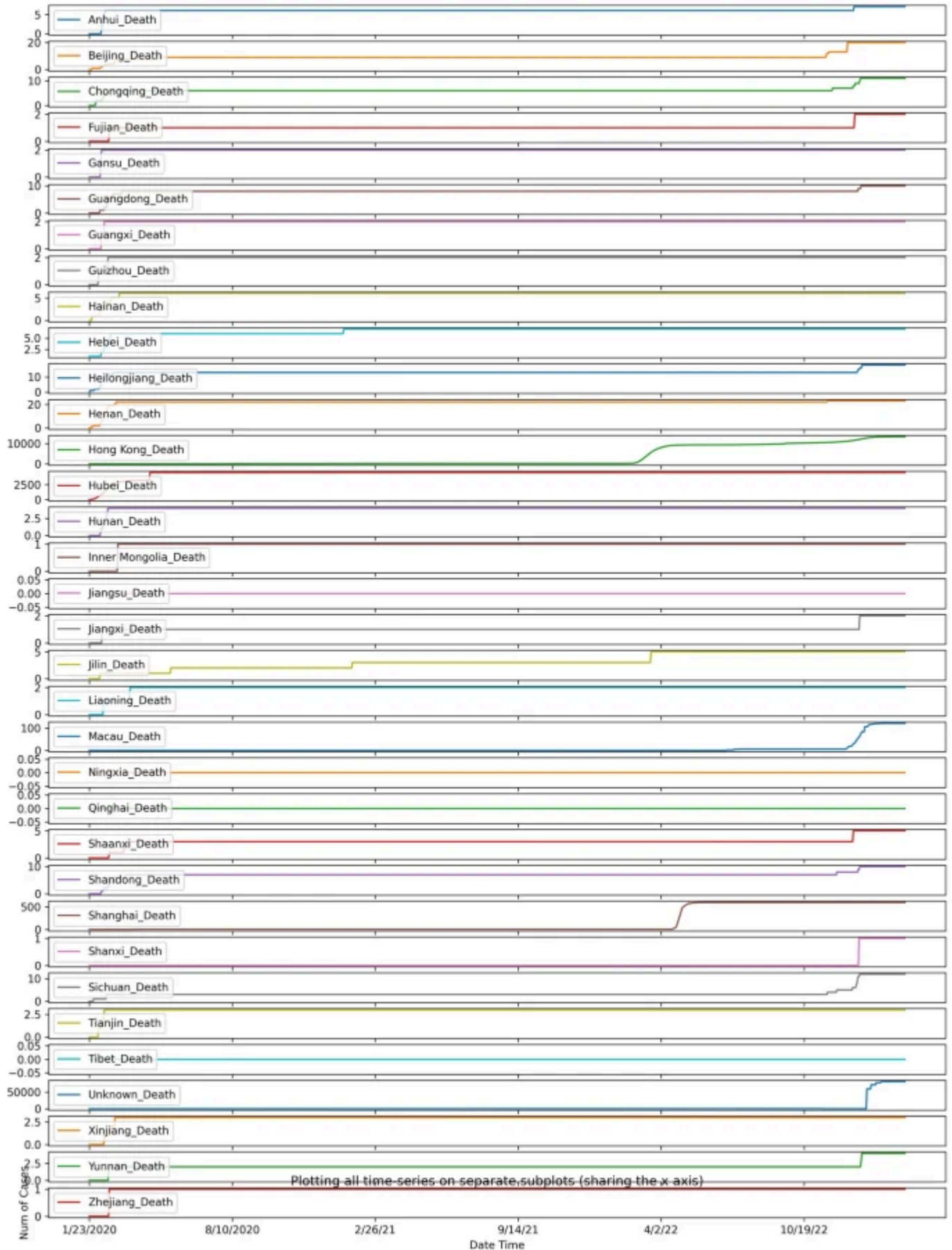
TSP for death cases in China according to each province

As a result of having the highest number of reported cases, the state of Hubei also has the highest number of deaths in all of China's rural areas.

Analysis of Cross-comparison of Recovered, Confirmed, and Death Cases

As the Hubei province in China had the most considerable number of recovered, confirmed, and death cases, this region was chosen for cross-comparison in these three categories (Figs. 4, 8).

Fig. 8
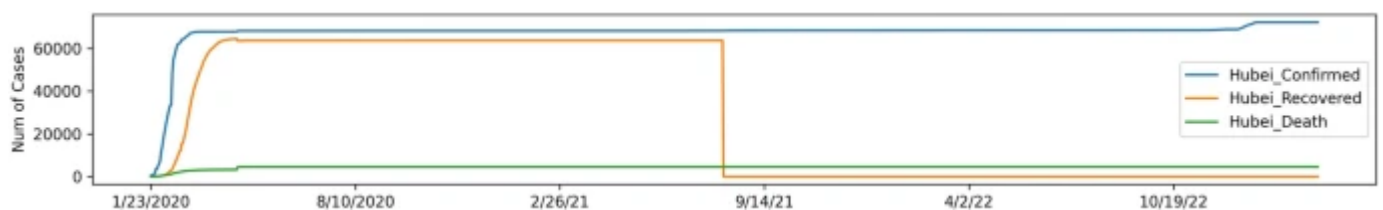
Cross-comparison of three categories in one frame

From the above plots, we can conclude that in Hubei province, the number of confirmed cases increased after January 2020 and remained constant throughout. The same is the case for recovered cases. However, the number of recovered cases dropped after June 2021. Also, the number of death cases is constant throughout this period. Confirmed and death cases are calculated better to compare the rate of recovered cases (Fig. 9).
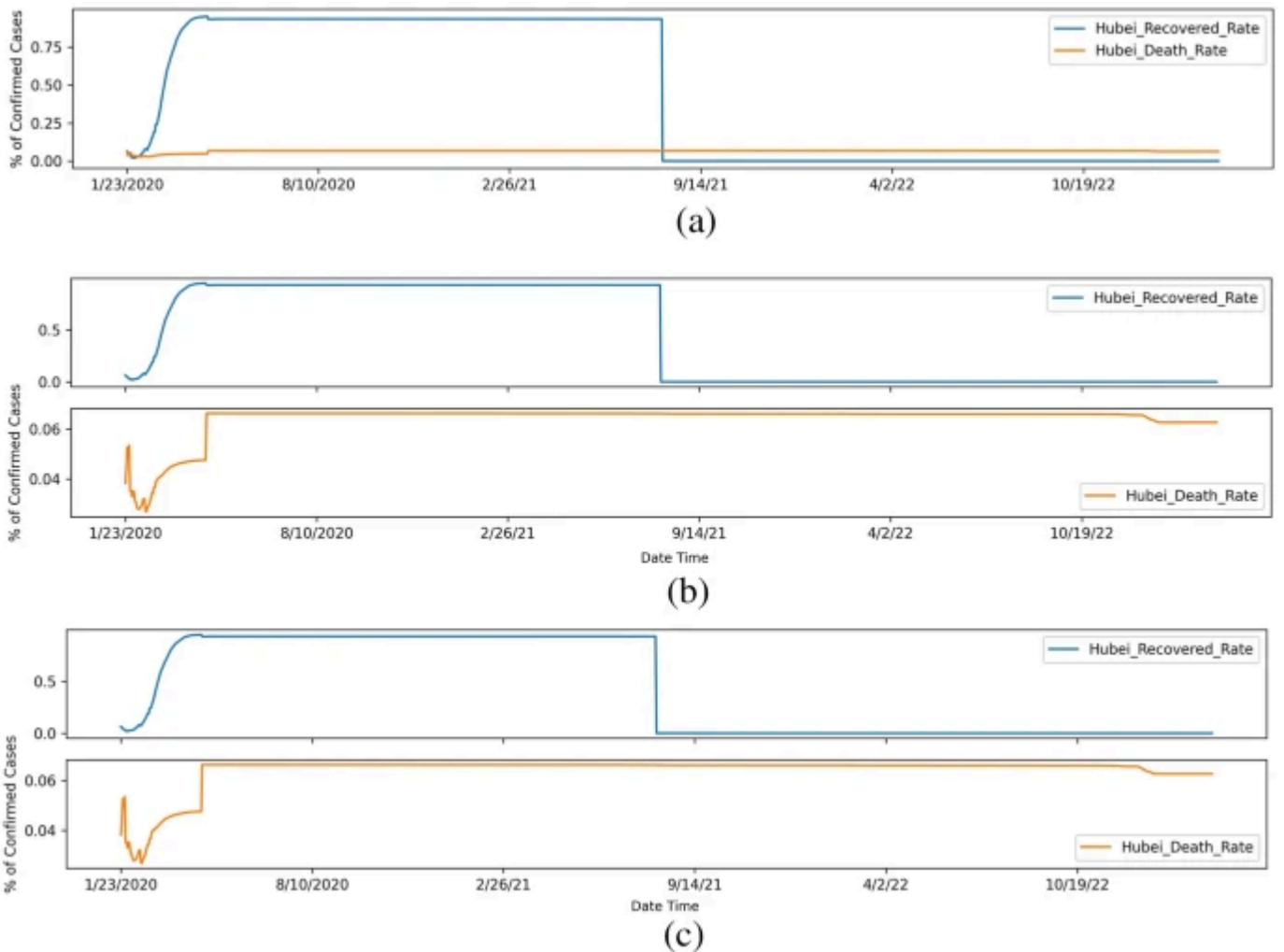
Fig. 9
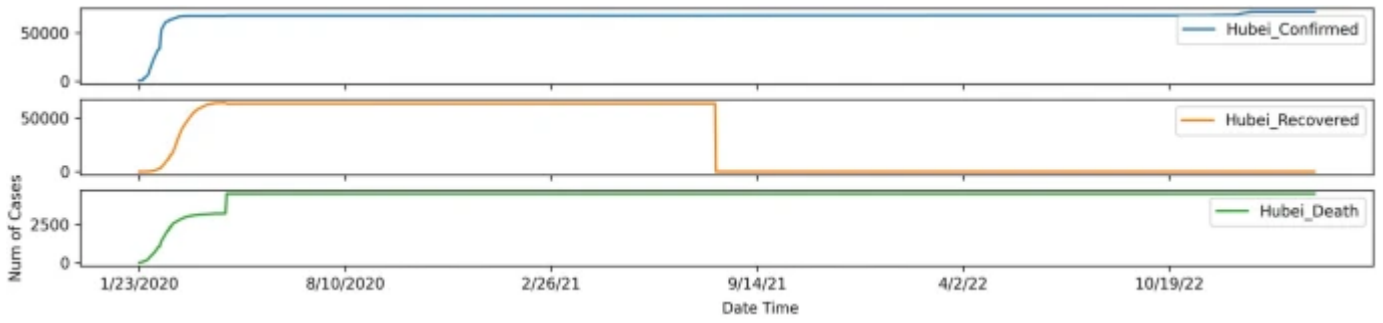
Cross-comparison of three categories in different frames

From the plots, we can conclude that for Hubei province, the recovered rate has surpassed death after 2 May 2020 (Fig. 10a−c).

Fig. 10

(a)

(b)

(c)

**a−c.** Plots for cross-comparison of rates per province

Cross-comparison of Confirmed, Recovered, and Death Cases in China's Hubei, Guangdong, and Zhejiang Province

As Hubei, Guangdong, and Zhejiang have the highest number of confirmed, recovered, and death cases, these regions are chosen for cross-comparison. From the data frames, we can conclude that the number of confirmed cases in Guangdong and Shanghai is almost similar, but Hubei has a different pattern. [26].

Comparing Confirmed Cases in Australia, Canada, and the US

From the plots below, we can conclude that the Australian Capital Territory has the highest number of confirmed cases in all states in Australia (Fig. 11).
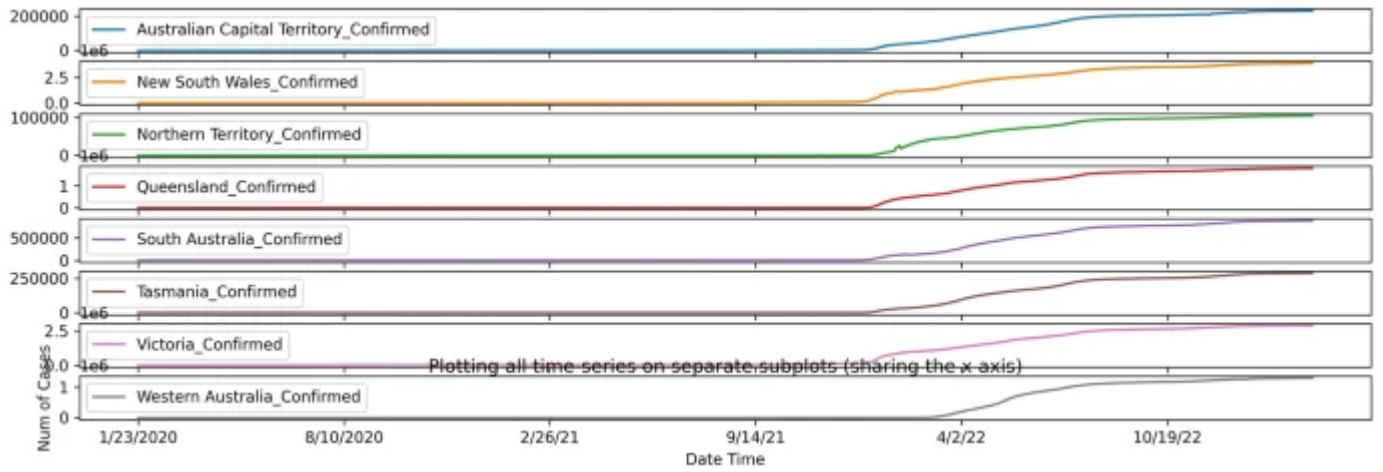
Fig. 11
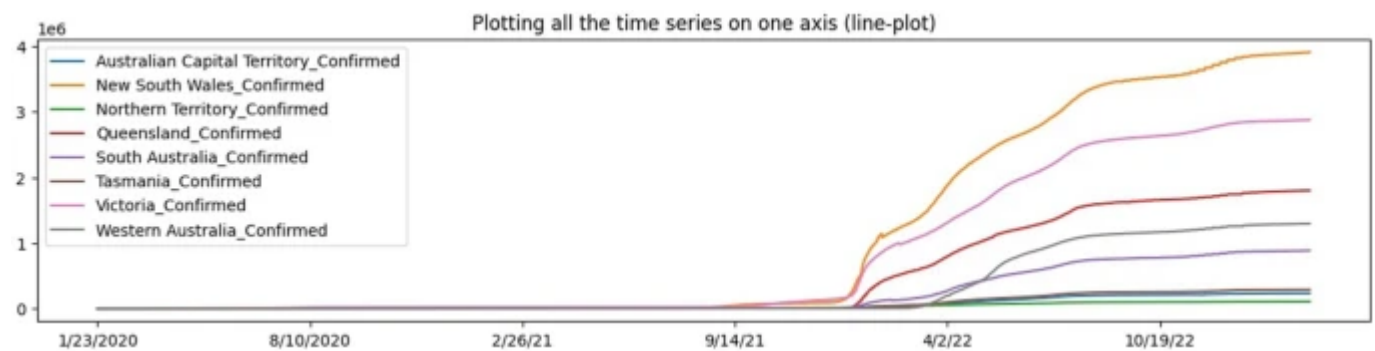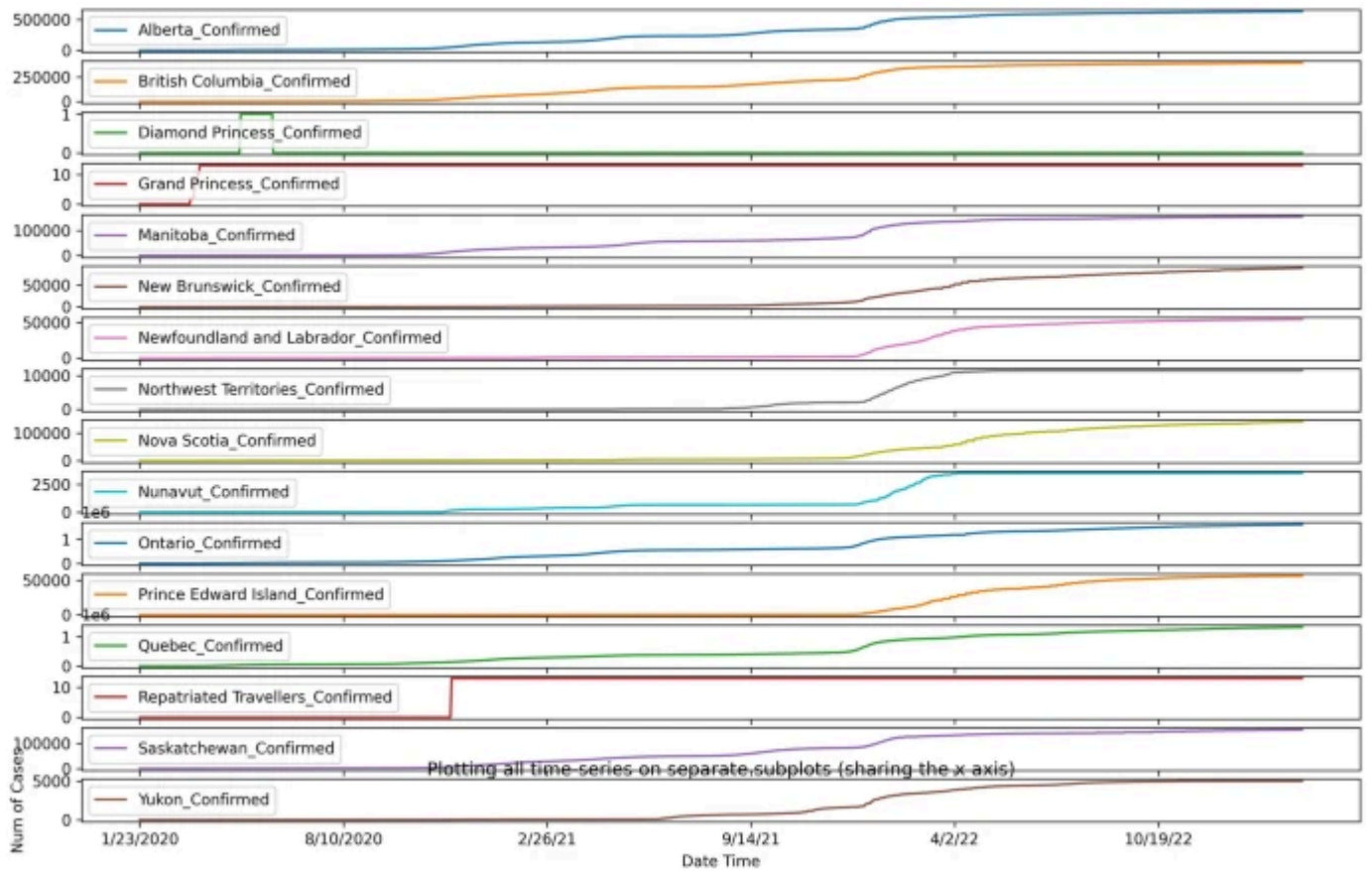

Confirmed cases in Australia

Alberta State has 500,000 cases in Canada, the highest among all the states in Canada (Figs. 12 and 13, 14 California, Florida, and New York have the highest number of confirmed COVID-19 cases (Figs. 13).
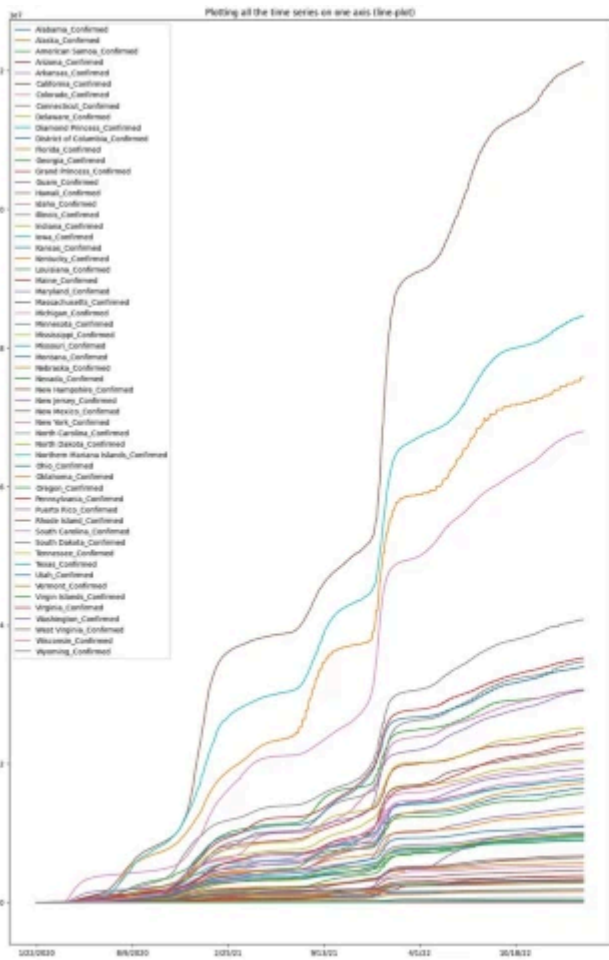
Fig. 12

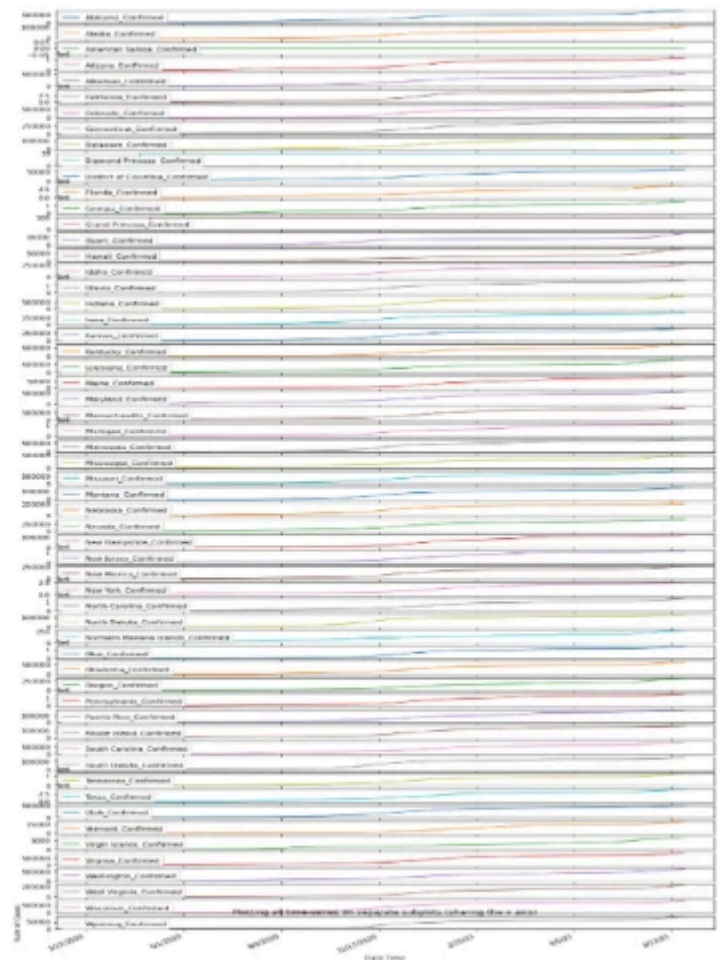Statewide confirmed cases in Australia

Fig. 13

Statewide confirmed cases in Canada

Fig. 14

(a)   (b)

a and **b** Statewide TSP of confirmed cases in the U.S

Cross-Comparison of Confirmed Cases Reported on Cruise Ships in the US and Other Countries

The Cruise ships are also listed under country/region in the dataset. The confirmed cruise ship cases are plotted on TSPs for a complete data analysis. Figure 15 shows that the number of confirmed cases on the Diamond Princess Cruise ship has rapidly increased in mid-February 2020.

Fig. 15
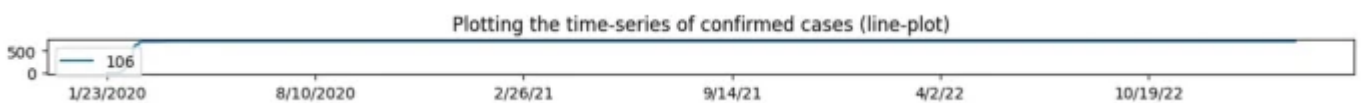
TSP of confirmed cases on 'Diamond Princess' Cruise ship .

TSP for confirmed cases of Grand Princess indicates that this number increased rapidly in mid-March and the end of March 2020 (Fig. 16)

Fig. 16

TSP of confirmed cases on "Grand Princess" Cruise ship

Plotting Time Series Lag Scatter Plots

Scatter plots are used to observe the relationships between the observation and the previous observation, *i.e.,* lags. The data in the scatter plot represent individual data points and patterns used to identify correlation relations. Scatter plots also indicate the presence of data lags or outlier points, so that data segmentation can be done quickly.

Relations deduced from scatter plots can be positive or negative. If data point clusters are from the bottom left to the top right, it indicates a positive relation. On the other hand, if data points cluster from the top right to the bottom left, then it indicates a negative relation. If more data points are closer to the diagonal line, it indicates a more substantial relation; if data points are scattered, it indicates a weak relation.

As in the COVID-19 data, there are simultaneous records of confirmed, recovered, and death cases over time; the resultant time-series is multivariate time-series. It can be represented as

$$y\left(n \right)\,\, = \,\,\left({y\left(1 \right) \ldots \ldots ..y\left(l \right)} \right)^{T} .$$

(4)

Figure 17 shows a scatter plot for confirmed cases in Hubei province in China. As most of the points are remarkably close to the diagonal, we can say that it has a strong relation, and because the points are increasing from bottom left to top right along the diagonal, we can conclude that this relation is a positive correlation relation. There appears to be a

strong linear pattern in confirmed cases at 't' and '$t + 1$', which confirms the first-order autoregression model.

Fig. 17
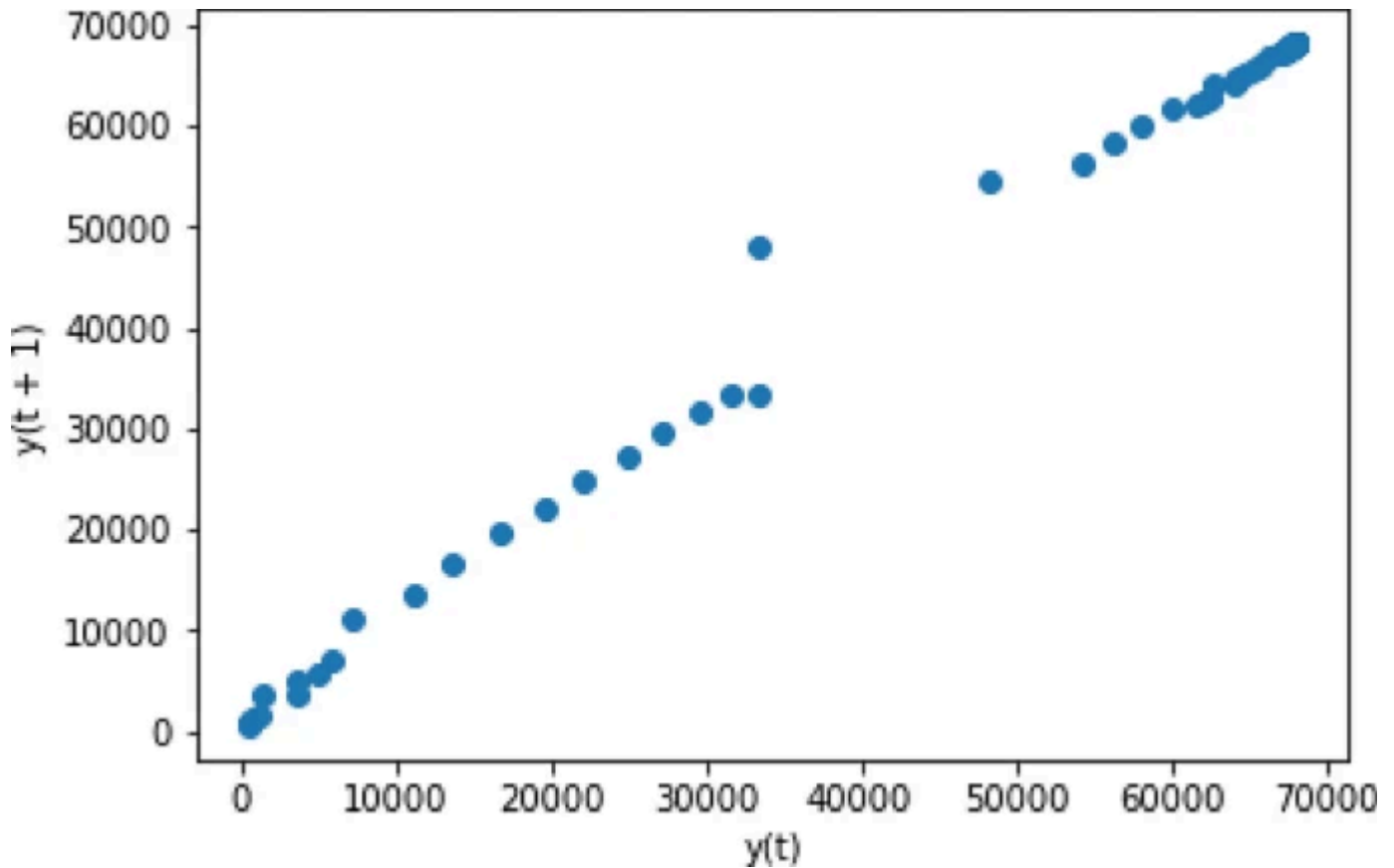
Scatter plot of confirmed cases in Hubei

The scatter plot can be tested for observations in the previous 7 days and the last month.

Figure 18 subplots indicate a strong positive correlation with each value in the last week. From all the scatter plots for different time limits, there are very few outliers in the data. All the data points are very close to the diagonal, which indicates a linear relation.

Fig. 18
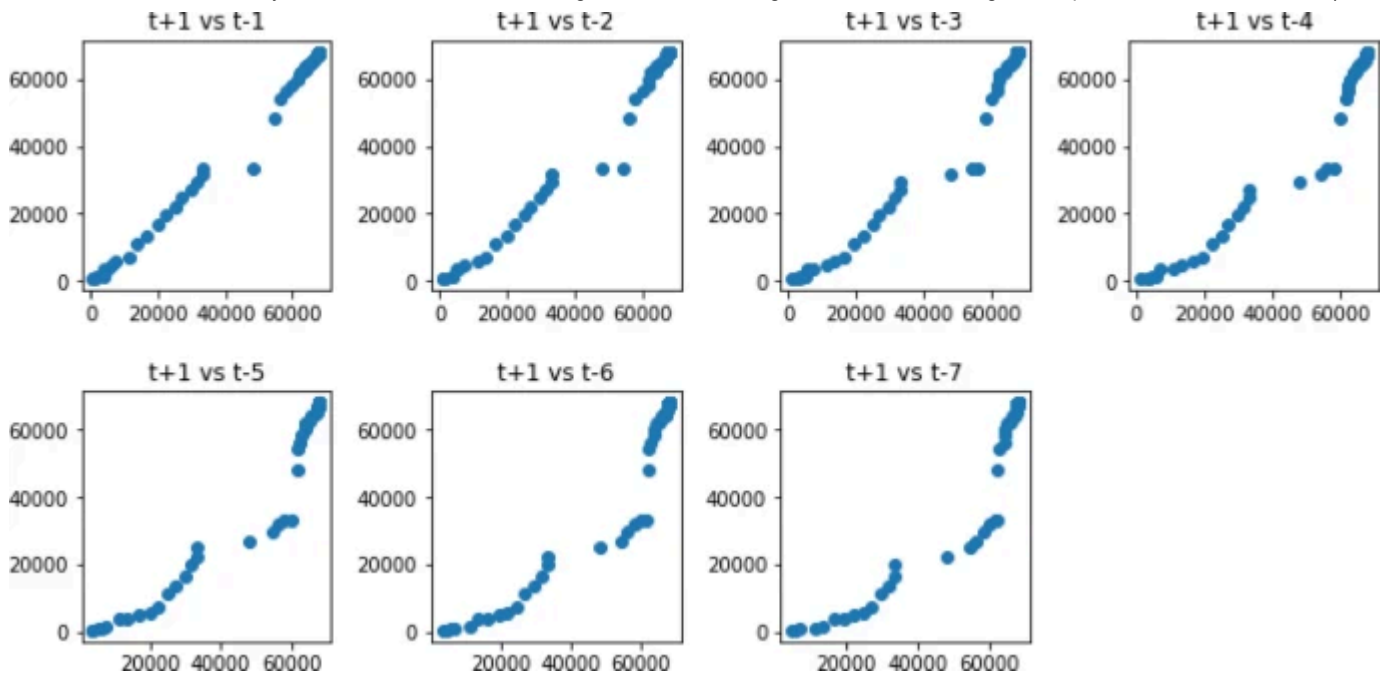
Scatter plot for different time limits

Time Series Autocorrelation Plots

The strength of the relationship between observations, their lag values, and their type can be quantified in time-series using auto correlation. Autocorrelation measures the extent to which lagged values are related, *i.e.*, $y(t)$ and $y(t-k)$. To determine the direction and trend from individual correlation coefficients, all the correlation values are plotted. The autocorrelation is given by the ratio of the auto covariance by Variance.
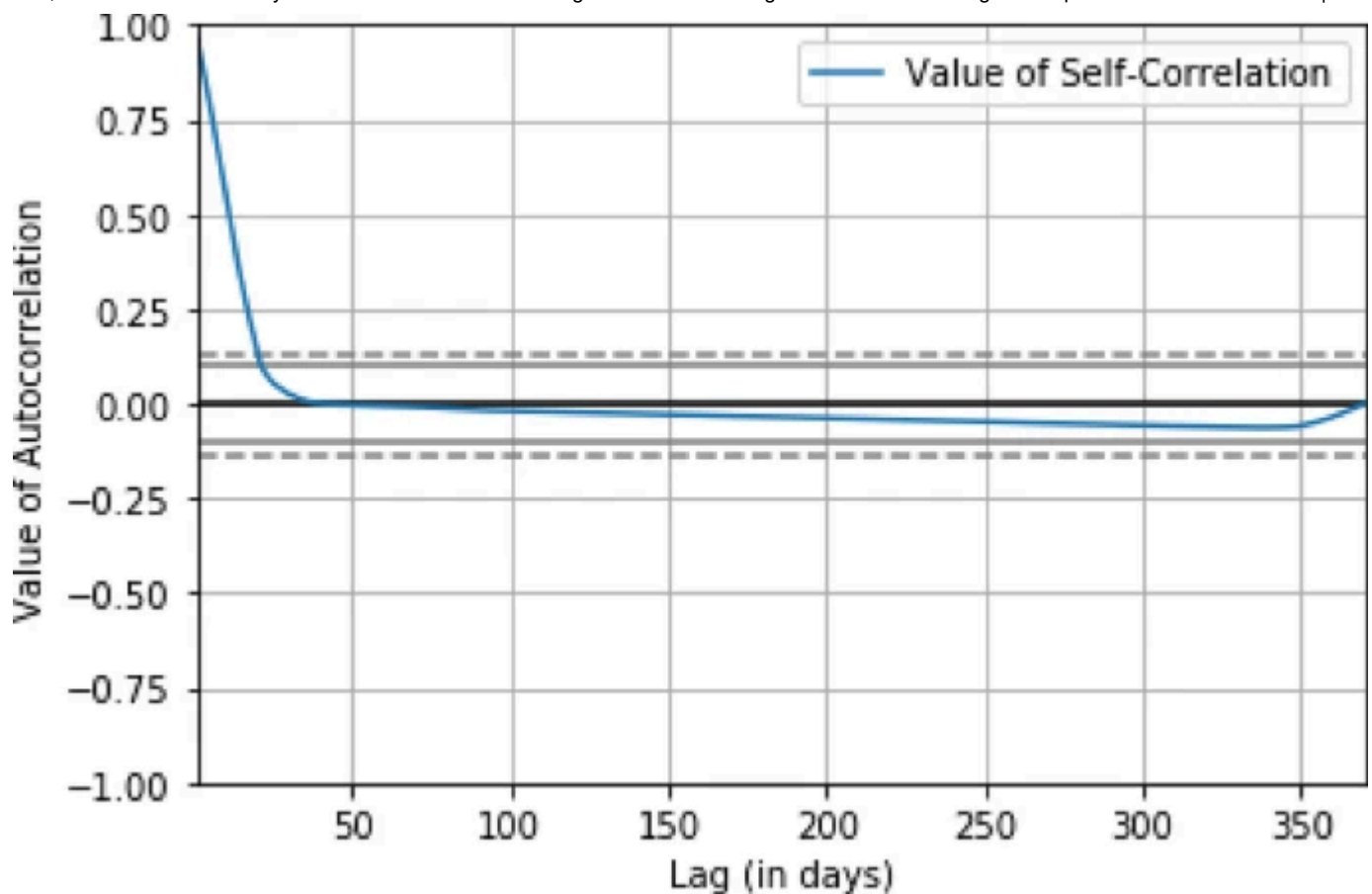
Given the time-series $(Y_1, Y_2...Y_N)$, the autocorrelation function at the lag 'k' is given as

$$rk \,\, = \,\,\,\,\,\frac{{\mathop \sum \nolimits_{i = 1}^{N - k} \,\,\left({Y_{i + k} - \overline{Y}} \right) \,\,\left({Yi - \overline{Y}} \right)}}{{\mathop \sum \nolimits_{i = 1}^{N} \,\,\,\left({Yi - \overline{Y}} \right)}}$$

(5)

To identify the appropriate TSA model, the autocorrelation function is plotted in Fig. 19 which shows an autocorrelation plot for confirmed cases in Hubei province.

Fig. 19

Autocorrelation plot for confirmed cases in Hubei

The above plot shows a lag in terms of the number of days on the x-axis and the correlation value on the y-axis. The dotted lines indicate that correlation values beyond these lines are statistically significant. In Hubei province, a strong correlation has been seen before day 6.

Conclusion and Future Work

The lesson all countries learned from this pandemic is to monitor the situation and take precautionary measures closely. Using data analysis and computational modelling, this paper proves the feasibility of performing COVID-19 analysis. Suitable COVID-19 patient prediction algorithms are identified through a systematic literature review. In this paper, the effects of the pandemic are studied in China, Australia, the US, and Cruise ships. A cross-comparison of recovered, confirmed, and death cases is done, and areas severely affected by this virus are identified. The accuracy of ML models is evaluated using training with patient record sets and TSA performance with accuracy measurement metrics. COVID-19 prediction features have been tested using different algorithms in available real-time patient data.

Future work will focus on calibrated and advanced hybrid ensemble methods for faster problem-solving and better results. Getting huge real-time datasets is a big challenge for performing analysis. Because many hospitals do not support real-time patient data, devices with sensors and features can detect diseases using AI techniques. Scatter demonstrates strong linear relations, aiding regression models in forecasting future cases with improved accuracy. Higher predictive accuracy of coronavirus existence in human bodies with many new variants like omicron will be the scope of future work. Quantum deep learning approaches will solve this problem in the near future.

Data availability

Not applicable.

Code availability

Not applicable.

References

1. Amit Kumar, Shivani Malhotra, Anuj Katoch, Ashish Sarathkar, Aman Manocha,' Webinars: An assistive tool used by higher education educators during Covid19 case study', 12th International Conference on Computational Intelligence and Communication Networks

2. Shivani Malhotra, Rubina Dutta, Amit Kumar Daminee, Sagar Mahna,' Paradigm Shift in Engineering Education During COVID-19: From Chalkboards to Talk Boards', 12th International Conference on Computational Intelligence and Communication Networks

3. Amit Kumar, Shivani Malhotra, Anuj Katoch, Ashish Sarathkar, Aman Manocha,' Webinars: An assistive tool used by higher education educators during Covid19 case study', 12th International Conference on Computational Intelligence and Communication Networks.

4. S. Panja, A. P. James, ' Belief Index for Fake COVID-19 Text Detection', 2020 IEEE Recent advances in intelligent computational systems (RAICS) | December 03-05, 2020 | Trivandrum

5. Subhra Debdas, Khushi Roy, Aniket Saha, Sayantan Kundu,' Analysis and Prediction of Climate Change in PostCovid19 India', IEEE Digital

6. Dheeraj Verma, Anup Shukla, Prerna Jain, 2020) 'COVID-19: Impact on Indian Power Sector', 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE) | 978-1-7281-8867-6/20/$31.00©2020 IEEE.

7. Nugroho Setio Wibwo, Rendy Mahardika,Kusrini, 'Twitter data analysis using machine learning to evaluate community compliance in preventing the spread of covid-19' , 2020 2nd international Conference on cybernetics and intelligent system (ICORIS) | 978-1-7281-7257-6/20/$31.00 ©2020 IEEE.

8. Kartika Maulida, Hindrayani,Tresna Maulana Fahrudin,Prismahardi Aji R,Eristya Maya Safitri,' Indonesian stock price prediction including Covid19 era using decision tree regression', 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) | 978-1-7281-8406-7/20/$31.00 ©2020 IEEE | DOI: https://doi.org/10.1109/ISRITI51436.2020.9315484

9. Omar Souissi, Latifa Ibrahimi, Mohamed Assellaou, Mourad Oubrich, 'Sharing Economy in a context of pandemic propagation: Case of the COVID19', IEEE Digital

10. Prema Gawade,Sarang Joshi, 'Personification and Safety during pandemic of COVID19 using Machine Learning', Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA-2020) IEEE Xplore Part Number: CFP20J88-ART; ISBN: 978-1-7281-6387-1.

11. Josimar Edinson Chire Saire, Jimy Frank Oblitas Cruz, 'Study of Coronavirus Impact on Parisian Population from April to June using Twitter and Text Mining Approach', 2020 International Computer Symposium (ICS) | 978-1-7281-9255-0/20/$31.00 ©2020 IEEE | DOI: https://doi.org/10.1109/ICS51289.2020.00056

12. Francesco Benedetto, Gaetano Giunta, Chiara Losquadro, Luca Pallotta,'Covid-19 Signal Analysis: Effect of Lockdown and Unlockdowns on Normalized Entropy in Italy ', 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE.

13. Chiara Antonini, Sara Calandrini, Fabrizio Stracci, Claudio Dario, Fortunato Bianconi,' Dynamical modelling, calibration and robustness analysis of COVID-19 using Italian data', 2020 IEEE 20th International Conference on BioInformatics and BioEngineering (BIBE)

14. Abir Abdullha, Sheikh Abujar, 'COVID-19: Data Analysis and the Situation Prediction Using Machine Learning Based on Bangladesh perspective', 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP) | 978-1-6654-1554-5/20/$31.00 ©2020 IEEE.

15. Nana Ramadijanti, Mu'arifin, Achmad Basuki,' Comparison of Covid-19 Cases in Indonesia and Other Countries for Prediction Models in Indonesia Using Optimization in SEIR Epidemic Models', IEEE Digital.

16. Isarapong E,' Monitoring the COVID-19 Situation in Thailand', 2020 1st International Conference on Big Data Analytics and Practices (IBDAP).

17. Vatsa D, Yadav A, Singh P, et al. An analytical insight of discussions and sentiments of indians on omicron-driven third wave of COVID-19. SN Comput Sci. 2023;4:791. https://doi.org/10.1007/s42979-023-02269-z.

Article **Google Scholar**

18. Negreiros RRB, Silva IHS, Alves ALF, et al. COVID-19 diagnosis through deep learning techniques and chest X-Ray images. SN Comput Sci. 2023;4:613. https://doi.org/10.1007/s42979-023-02043-1.

Article **Google Scholar**

19. Seth R, Sharaff A. Sentiment data analysis for detecting social sense after COVID-19 using hybrid optimization method. SN Comput Sci. 2023;4:568. https://doi.org/10.1007/s42979-023-02017-3.

Article **Google Scholar**

Funding

Author information

Authors and Affiliations

Department of Computer Science and Engineering, School of Computing, Institute of Science and Technology (Deemed to Be University), Vel Tech Rangarajan Dr. Sagunthala R&D, Chennai, Tamil Nadu, 600062, India

Senthil Kumar Nramban Kannan

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, 522302, India

Bhanu Prakash Kolla

Department of Computer Science and Engineering, PSN College of Engineering and Technology, Tirunelveli, Tamil Nadu, 627152, India

Sudhakar Sengan

Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, 600123, India

Rajendiran Muthusamy

Department of Electronics and Communication Engineering, K. Ramakrishnan College of Technology, Trichy, Tamil Nadu, 621112, India

Raja Manikandan

Charotar University of Science and Technology, Gujarat, 388421, India

Kanubhai K. Patel

Department of Computer Science and Engineering, Management and Gramothan (SKIT), Swami Keshvanand Institute of Technology, Jaipur, Rajasthan, 302017, India

Pankaj Dadheech

Corresponding author

Correspondence to Bhanu Prakash Kolla.

Ethics declarations

Conflict of interest

Not applicable.

Additional information

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This article is part of the topical collection "Soft Computing in Engineering Applications" guest edited by Kanubhai K. Patel.

Rights and permissions

Reprints and permissions

About this article

Cite this article

Nramban Kannan, S.K., Kolla, B.P., Sengan, S. *et al.* Analysis of COVID-19 Datasets Using Statistical Modelling and Machine Learning Techniques to Predict the Disease. *SN COMPUT. SCI.* **5**, 181 (2024). https://doi.org/10.1007/s42979-023-02464-y

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Provided by the Springer Nature SharedIt content-sharing initiative

Keywords

Time-series analysis   COVID-19   Time-series forecasting

Autocorrelation   Machine learning   Analysis and visualization