Check for updates

RESEARCH ARTICLE

COVIDSpread: real-time prediction of COVID-19 spread based on time-series modelling

[version 1; peer review: 1 approved, 1 approved with reservations]

Siroos Shahriari[1,2], Taha Hossein Rashidi[1,2], AKM Azad[3], Fatemeh Vafaee[4,5]

[1]Research Centre for Integrated Transport Innovation, University of New South Wales, Sydney, NSW, Australia
[2]School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW, Australia
[3]Children's Medical Research Institute, University of Sydney, Westmead, NSW, Australia
[4]School of Biotechnology and Biomolecular Science, University of New South Wales, Sydney, NSW, Australia
[5]UNSW Data Science Hub, University of New South Wales, Sydney, NSW, Australia

v1 **First published:** 03 Nov 2021, **10**:1110
https://doi.org/10.12688/f1000research.73969.1
Latest published: 03 Nov 2021, **10**:1110
https://doi.org/10.12688/f1000research.73969.1

Abstract

A substantial amount of data about the COVID-19 pandemic is generated every day. Yet, data streaming, while considerably visualized, is not accompanied with modelling techniques to provide real-time insights. This study introduces a unified platform, COVIDSpread, which integrates visualization capabilities with advanced statistical methods for predicting the virus spread in the short run, using real-time data. The platform uses time series models to capture any possible non-linearity in the data. COVIDSpread enables lay users, and experts, to examine the data and develop several customized models with different restrictions such as models developed for a specific time window of the data. COVIDSpread is available here: http://vafaeelab.com/COVID19TS.html.

Keywords

COVID-19, SARS-CoV-2, prediction of spread, time series modelling

Open Peer Review

Approval Status ✓ ?

	1	2
version 1 03 Nov 2021	✓ view	? view

1. **Mahdieh Allahviranloo**, The City College of New York – CUNY, New York City, USA

2. **Ross Gore**, Old Dominion University, Norfolk, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Pathogens gateway.

This article is included in the Emerging Diseases and Outbreaks gateway.

Corresponding authors: Taha Hossein Rashidi (rashidi@unsw.edu.au), Fatemeh Vafaee (f.vafaee@unsw.edu.au)

Author roles: Shahriari S: Formal Analysis, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Hossein Rashidi T**: Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Azad A**: Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Vafaee F**: Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Resources, Software, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: SS and THR acknowledge the support from the Australian Research Council under Linkage Scheme (LP160100450). THR acknowledges the support from the Australian research Council under the DECRA Scheme (DE170101346).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Shahriari S *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Shahriari S, Hossein Rashidi T, Azad A and Vafaee F. **COVIDSpread: real-time prediction of COVID-19 spread based on time-series modelling [version 1; peer review: 1 approved, 1 approved with reservations]** F1000Research 2021, **10**:1110 https://doi.org/10.12688/f1000research.73969.1

First published: 03 Nov 2021, **10**:1110 https://doi.org/10.12688/f1000research.73969.1

Introduction

The pandemic of coronavirus disease 2019 (COVID-19), caused by acute respiratory syndrome coronavirus 2 (SARS-CoV-2) represents the most serious public health threat during the last century.[1] The global impact of COVID-19 has been profound. As of 12 September 2021, over 224 million cases and 4 million deaths have been confirmed globally.[2] Forecasting the imminent spread of COVID-19 informs policymaking and enables an evidence-based allocation of medical resources, arrangement of production activities and economic development.[3] Therefore, it is urgent to establish efficient trend prediction models, on the latest available data, to provide a point of reference for the governments to formulate adaptive responses based on reliable predictions on the impending progress of the pandemic.

The classical Susceptible-[Exposed]-Infected-Recovered (SEIR/SIR) epidemic models,[4] have been widely developed to simulate the transmission dynamics of COVID-19[5,6] and the impact of non-therapeutic interventions – e.g., travel and border restrictions,[7,8] quarantines and isolations,[5,9–11] or social distancing and closure of facilities – on the spread of the pandemic, and in some cases, on the healthcare demand.[5,9,11–13] These studies have been mostly focused on calibrating models for a specific country/region based on the data at the time of the model-development and assuming a multitude of parameters initialized upon prior knowledge such as social contact structure, rate of compliance with the policy and incubation or infection period among others. Complementing upon SEIR mathematical models, and owing to the increased amount of data and consistency of reports, some recent efforts have been focused on developing statistical[3,14] or machine learning methods[15] to predict the near-future spread of COVID-19 (in terms of the number of confirmed cases or deaths) based on the historical data.

While reliable predictions of the pandemic trend are essential for policymaking and resource-allocation, there is a lack of an adaptive real-time modelling platform which evolves as new data arrives. Here, we present, COVIDSpread, a time-series online platform for real-time modelling of the progression of COVID-19 using the Autoregressive Integrated Moving Average (ARIMA)[16] statistical analyses combined with different non-linear transformation approaches.[17] Our platform offers an interactive online dashboard which efficiently generates country-wise predictive models, in real-time, based on the latest report of COVID-19 cases worldwide.

The proposed modelling approach neither relies on strict modelling assumptions (e.g., linearity, stationarity, or existence of an epidemic steady state) nor on any initial parameters requiring a priori knowledge. It offers a transparent mathematical function to better understand the trend and to predict future points in the series. Different types of transformation have been examined to capture the nonlinearity in the time-series data followed by multiple differencing steps to eliminate the non-stationarity status.

The main objective of this study is to introduce an easy-to-use and readily available statistical tool to develop rigor models for time series data of COVID-19 as data becomes available on a real time basis. In this article, it is demonstrated that the proposed modelling tool is reliable to estimate accurate model parameters and predict the short-term spread of COVID-19 across different countries.

Methods

Autoregressive integrated moving average (ARIMA) for epidemic trend forecasting

The structure of the data and the autocorrelation between daily reported instances makes an intuitive case for time series analyses. Autocorrelation occurs when there is a correlation between ordered observations in time or space resulting in the covariance between the error terms being nonzero. In the context of the ordinary least square method, where the dependent variable (even if it is observed for multiple times) is regressed only against the explanatory variables rather than previous observation of the dependent variable, the estimated parameters remain unbiased (their expected value is equal to the true values), and asymptotically normally distributed. However, the coefficients are not any-more efficient (not having the minimum variance) which means that, the commonly used, t and F statistics are not any more reliable.

If we define Y_t as the dependent variable, i.e., the number of confirmed cases, and X_t as the explanatory variables, i.e., an intervention, γ being a parameter to be estimated, and ε_t as the residual at time t; in a typical linear regression models, we have $Y_t = \gamma X_t + \varepsilon_t$ which can be simply written as $Y_{t-1} = \gamma X_{t-1} + \varepsilon_{t-1}$ for one time interval earlier than t. This can be extended for s time intervals earlier as $Y_{t-s} = \gamma X_{t-s} + \varepsilon_{t-s}$. To examine the first order autocorrelation ($s = 1$, i.e., only the correlation between residuals are considered which can be simply translated to the dependency between Y_t and Y_{t-1}), the unobserved part of the error terms are correlated and denoted by $\varepsilon_t = \rho\varepsilon_{t-1} + \epsilon_t$, where ϵ_t is white noise series and ρ can be estimated which is the main factor for examining the strength of autocorrelation as well as correcting the estimated coefficients.

Time series models implicitly assume that the stochastic process is *stationary* which loosely implies that mean and variance of data do not change over time. Then by using **autoregressive** regression (AR, as explained in the previous paragraph) and a **moving average** (MA, as explained in the next paragraph) mechanism, unbiased parameters can be estimated. If the data is not stationary, there are ways to transform it to stationary data such as by differencing i times which is denoted as $I(i)$, **integrated** of order i.

When Y_t is defined for one side, i.e., dependent on the past, and it is weighted by parameters, say θ_i for the earlier s time intervals), a moving average model is constructed as:

$$Y_t = \mu + \sum_{i=-s}^{0} \theta_i \epsilon_{t+i},$$

where $\theta_0 = 0$. In this case, Y_t is observed and θ_i is estimated. The main difference between MA and the AR model is that several instances of white noise appear on the right-hand side of MA while past instances of the dependent variable do not appear on the right-hand side of the MA equation. In other words, in the MA model white noise of previous time intervals is scaled and carried over to the later time intervals while μ captures a drift in the number of cases in each time interval.

When the data includes both the impact of scaled white noise and previous instances of the dependent variable affecting the current situation and ARIMA model is used. Unlike the compartmental models in epidemiology (e.g., SIR/SIER), ARIMA does not require exogeneous information about the susceptible population and recovery patterns. Instead, it captures the declining or increasing pattern of the data by extracting information from the nonlinear trends observed in the previous time intervals.

Data and pre-processing

The platform retrieves daily number of confirmed cases from Coronavirus Resource Centre at John Hopkins University using *coronavirus* R package. Yet, it is independent of data source and can incorporate other major COVID-19 reporting parties (on countries, provinces and territories time-series), and can be readily extended to model number of deaths or recovered cases. John Hopkins reports latest available public data on COVID-19 on daily bases for all affected countries; latest data can be directly accessed from R environment. Countries with more than 30 reporting days from >50 cases were retained by the dashboard assuming that filtered countries do not hold enough data for reliable modelling/forecasting. At the time of submission 195 counties pass this constraint and modelled by COVIDSpread. The number of countries modelled increases continuously, as number of observations increases daily.

Time-series transformation

The data driven approach of this study employs three transformation operations including *Ratio* transformation (the ratio of observations in two consecutive days), *power* transformation (n^{th} root) and *logarithmic* (natural log) transformation to stabilize variance in raw data and adjust the historical data for a simpler forecasting task. Models on transformed data were compared against the models developed for the non-transformed data and the best overall model were selected.

Eliminating non-stationarity

Time series models often assume stationary time-series which implies that statistical properties such as mean, variance, autocorrelation, etc. are constant over time Other than transformation, differencing once or twice (i.e., differencing between the values of consecutive data points) helps estimating the speed of growth or the acceleration/deceleration of growth. If data is non-stationary based on augmented Dickey–Fuller (ADF) test, the differencing step were applied one or more times to eliminate the non-stationarity (ADF p-value > 0.05). The differencing step allows to develop a model that is comparable to models developed based on in the SIR\SIER models. The differencing will continue until the stationary status is obtained.

Model development and prediction

Once stationary data is obtained, the best ARIMA (p, i, q) model were fitted to each time-series by searching through different combinations of p which is the order of the autoregressive model, i is the degree of differencing (as previously discussed), and q is the order of the moving-average model. The 'auto.arima' function in the R 'forecast' package has been used for model optimisation using non-stepwise selection. The selection of the best model is based on the root mean square error (RMSE) value estimated based on an out of sample estimation process on the latest 20% part of the observations (as a rough estimate of out-sample RMSE). The best model would be then used for prediction of the disease spread in the next N days (defined by the user) using the parameters of the model. The developed models can be then used to forecast changes in the number of infected cases. The prediction algorithm simulates the expected total number of infected cases as well as a bandwidth around the expected values reflecting the 80%-95% confidence level which is estimated based on the significance of the estimated parameters.

Platform design and implementation

The whole pipeline including automated data retrieval, pre-processing and modelling, has been implemented in R, providing a unified platform for ease of reuse and maintenance. The online dashboard has been developed using R Shiny. Scheduled data updates were automated via a reactive file reader. Interactive line plots and maps were visualised using R-integrated Plotly and Leaflet JavaScript graphing libraries, respectively. We recommend clearing the browser autocomplete history to delete previous selections from the date-picker. The shiny web server can run in any modern web browser including Google Chrome, Mozilla Firefox, and Safari. Moreover, the COVIDSpread source code can be run using the RStudio IDE on a standard workstation (Windows/Linux/Mac) with an i5 processor and 8 GB RAM.

Results

Model development and performance

Multiple transformation operations are investigated to stabilise variance, coupled with recursive differencing until eliminating non-stationarity in the time-series data, i.e., p-value < 0.05 based on augmented Dickey–Fuller test.[16] Upon each transformation, the best ARIMA model is obtained for each country, according to Akaike information criterion (AIC) value using maximum likelihood estimation. The optimal model for each transformation is then recorded based on the overall model Root Mean Square Error (RMSE) on the last 20% of observations reported as a surrogate estimate of out-sample prediction performance where models are trained on the first 80% of data. The predictive power of the best model per country is compared against estimations provided through exponential growth in number of cases including, 1) doubling time of two days, 2) doubling time of three days and 3) doubling time of one week, as well as a conventional linear univariate regression on log-transformed data. Extended Table 1 shows the parameters of the optimal ARIMA model per country and the corresponding RMSE measures (of the last 20% of observations) compared with conventional trends using data obtained on 23rd April 2021 from Coronavirus Resource Centre at John Hopkins University (https://coronavirus.jhu.edu). While, the purpose of this study is not to develop the most accurate time-series predictive model, statistics of the Extended Table (Extended Table 1 shows that using a more sophisticated statistical model significantly improves the prediction accuracy of COVID-19 spread in the near future (Wilcoxon test p-value << 0.001 comparing distributions of residuals).

Effect of transformation

Different time-series transformation operations, namely power transformation, logarithmic transformation and ratio transformation, have been applied to pre-process the data prior to the differencing step. We have observed that the type of transformation can significantly improve the performance of a model (in terms of the estimated out-sample RSME) as there is no *a priori* knowledge about the best-performing transformation (except that power transformation always performs poorly). Figure 1 shows some countries, as case studies, whose ARIMA models (as of April 23, 2021) are significantly affected by the type of transformation. As Figure 1 shows models on countries such as Zimbawe and Burundi has better performance with logarithmic transformation. For Nepal, Argentina and Greece, ratio transformations provide superior results. The case of Eswatini, interestingly, demonstrates that the ratio and logarithmic transformations outperform without transformation because the model can capture rapid fluctuation better with those transformations. Overall, the results signify the value of a performance-driven transformation selection approach upon trying multiple operations, as implemented in this platform.

Dynamic model estimation

The nonlinear dynamic system underlying COVID-19 spread is producing a regularly disrupted pattern making static predictions increasingly unreliable. Accordingly, a powerful feature of the platform is *dynamic model estimation*, that is, all models are re-optimised temporally with availability of new daily observations. Accordingly, the latest reports on COVID-19 case numbers are reflected in model estimation which accounts for the impact of new interventions improving the reliability of the future forecasts. As a case study, we have chosen to show the value of this feature on prediction of future case numbers in Iran. Iran's trend shows significant fluctuations in the last 10 days (as of April 13, 2020) offering an interesting case study. We assumed that the model has access to data up to April 03, and then reported the next 10 days predictions and the RMSE of predicted number on April 13th. This procedure was repeated 9 times, where new observations became available to the model, one at the time. Figure 2 shows how such dynamic re-estimation adjusts the model with emerging pattern in time-series trend and improves prediction accuracy.

Online dashboard

We have developed an interactive online dashboard to facilitate real-time model development for lay users as well as data scientists (Figure 3). Users can select the country of interest from the left panel and observe an interactive visualisation on cumulative counts of confirmed cases in the middle panel. Upon pressing the 'Predict' button', the platform provides users with optimal models fitted to the latest reports of COVID-19 spared as provided by John Hopkins University Coronavirus Resource Centre. For any country of interest, the interactive user-interface enables users to

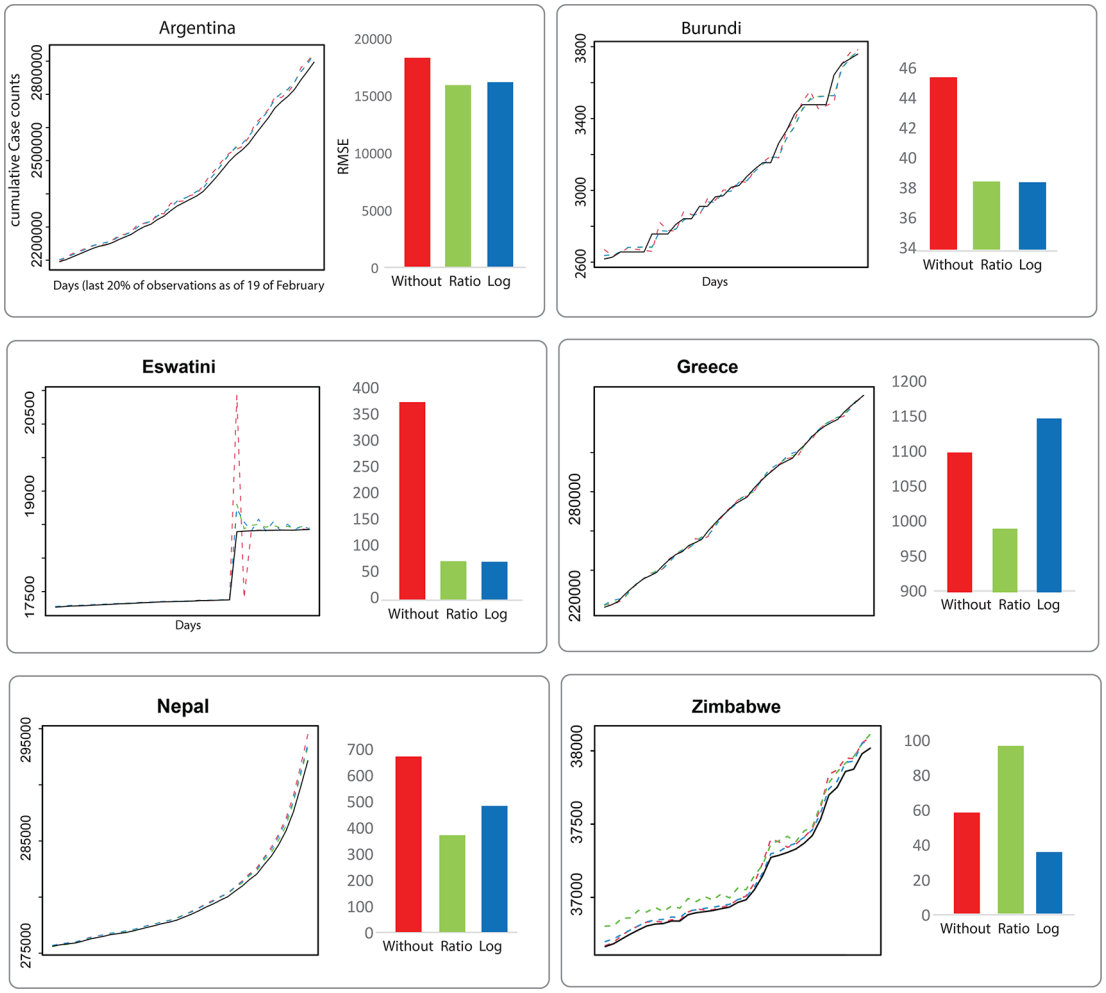


Figure 1. Effect of transformation on modelling performance. Six countries were selected as case studies to demonstrate the effect of ratio and logarithmic transformations on the model performance as measured by RMSE on last (most recent) 20% of time-series data. The solid line shows the observed trend and the dashed lines shows model fitted values without transformation (red) and after ratio (green) or logarithmic (blue) transformation. The bar plot beside each trend graph shows the corresponding RMSE estimations.

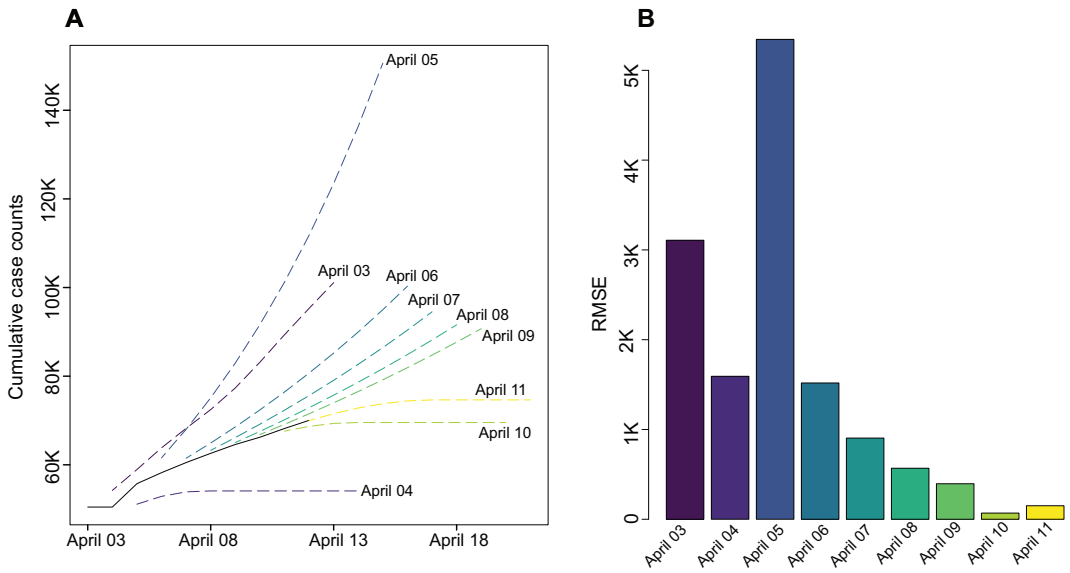


Figure 2. Dynamic model estimation. A. Predictions of next 10 days for COVID trend in Iran, assuming that the last available date was April 03 2020 to April 11 2020 as marked on the plot (last obsereved date at the time of analysis: April 13). Solid line shows observed cases. B. RMSE comparing predictions with obsereved data at April 13.

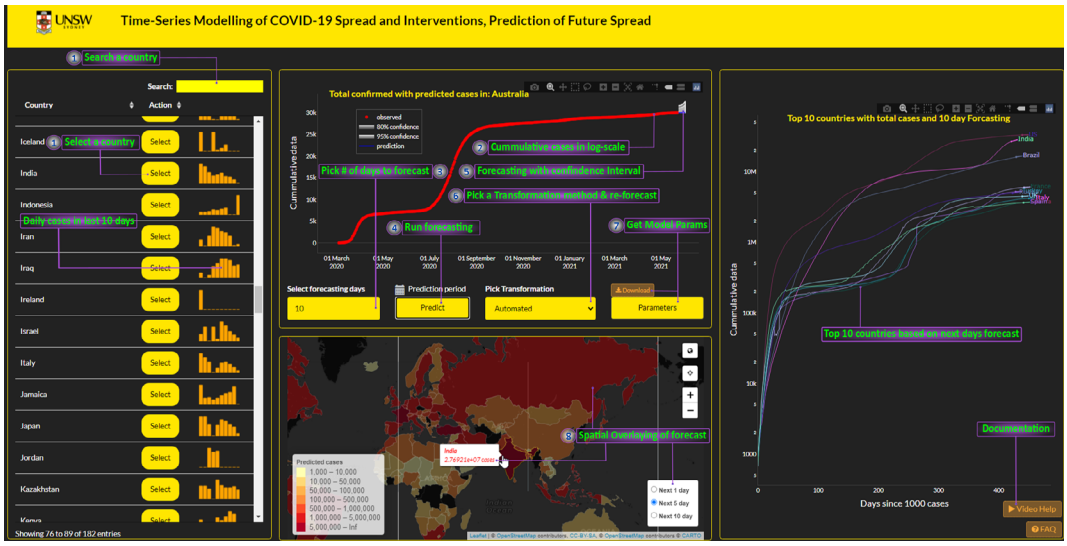


Figure 3. An overview of COVIDSpread dashboard representing different panels to select the country of interest (left panel), perform time series prediction for the selected country (top middle panel), and compare the distribution and rate of confirmed cases across different countries (bottom middle panel and left panel).

re-estimate models by customising the range of days to be included in the model. The right panel visualises the cumulative number of confirmed cases since the 1000th case of top 10 countries in terms of total number of cases, plus predictions of growth trajectories in the next 10 days. Similarly, the middle-button panel shows the world map color-coded with predicted number of cases per 100K, together providing a global comparative view of the forthcoming COVID-19 spread. The dashboard back-end, i.e., data mining, pre-processing, and model development were implemented in R with several R packages including *forecast*, *tseries*, *tsir*, *imputeTS*, and *coronavirus*. The front-end of the dashboard was implemented in R shiny with several R packages, including *rplotly*, *ggplot2*, *ggiraph*, *leaflet*, *DT*, *sparkline*, *data.table*, *survival*, *tidyr*, and *shinyWidgets*. Having a single codebase for the whole framework is useful, especially in the context of reproducibility and ongoing maintenance. This dashboard offers users to not only view information in an interactive manner (e.g., on mouse hover), but also allows to download the parameters of the selected model used for the forecasting.

Discussion

COVID-CDR contribution and limitations compared with related studies

Real-time COVID-19 data analytics have been mainly focused on visualizing the spread[18] with comparatively less effort in developing models to dynamically analyze the data. Epidemiological models, i.e., SIR/SIER models, have a strong foundation in analyzing epidemic growth/decline, and have been substantially explored for modelling the speed of infectious disease progression. Yet, such models are often offline/static, require assumptions for the parametric formulation of the model and rely on multitude of initial parameters.

Aside from SIR/SIER models, several models have been used to predict COVID-19 cases, including ARIMA, nonlinear autoregression neural network (NARNN), support vector regression (SVR), Prophet, and different deep neural network-based models such as long short-term memory (LSTM)/Stacked LSTM, Convolutional LSTM, and Bidirectional LSTM.

Tomar and Gupta[19] and Chimmula and Zhang[20] are two recent studies that concentrate on the LSTM model. In a similar vein, Shastri, *et al.*[21] developed LSTM, Stacked LSTM, Bidirectional LSTM, and Convolutional LSTM models for COVID-19 cases prediction. Another area of study is the use of a recurrent neural network (RNN) to predict possible COVID-19 cases which was followed by Arora, *et al.*[22] with proposed models based on RNN, LSTM, Bi-LSTM. Similarly, Hawas[23] used RNN for daily COVID-19 infection predictions.

Aside from LSTM and RNN, one of the most commonly used models is the ARIMA time series model. Several studies focused on ARIMA models for predicting future cases including Alzahrani, *et al.*[24] with forecasting the spread of COVID-19 cases based on the ARIMA model and Shahid, *et al.*[25] focusing on ARIMA, SVR, LSTM and Bi-LSTM models.

Along the same vein, Ribeiro, *et al.*[26] compared ARIMA, cubist regression (CUBIST), random forest (RF), ridge regression (RIDGE), SVR, and stacking-ensemble learning for the prediction of COVID-19 cumulative cases. Similarly,

Kırbaş, *et al.*[27] focused on ARIMA, nonlinear autoregression neural network (NARNN) and LSTM for future case prediction. Likewise, Papastefanopoulos, *et al.*[28] provided a comparison on six different forecasting methods including ARIMA, the Holt-Winters additive model (HWAAS), TBAT, Facebook's Prophet and deep AR for active case prediction and Devaraj, *et al.*[29] evaluated ARIMA, LSTM and SLSTM for COVID-19 prediction.

We developed a time-series based statistical model to dynamically predict the future trend of COVD-19 spread. It is coupled with the capacity of time-series models in 1) considering higher orders of derivatives of the number of cases in previous time intervals, 2) accounting for the impact of residuals of the previous time intervals. The literature demonstrates a diverse set of short-term forecasting models, although the focus of this research is on ARIMA models, other models may be considered if the developed model does not perform adequately. For example, as shown by the results, ARIMA (0,2,0) was the best-fitting ARIMA model for India, with high RMSE values. This means that the model does not include any autoregressive or moving average terms and does not provide a suitable representative of COVID-19 cases in India at the investigated time period.

Different models for analyzing COVID-spread in India can be used as an alternative as several studies have been conducted to model COVID-19 spread in India using other models. For instance, Tomar and Gupta[19] used the LSTM model and Arti and Bhatnagar[30] proposed a tree-based model to model spread the disease in the community. Bherwani, *et al.*[31] used the SEIR model to model the spread of COVID-19 while Mahajan, *et al.*[32] used the compartmental epidemic model SIPHERD for modelling spread COVID-19 cases in India. In the same line, Roy and Roy Bhattacharya[33] proposed A mathematical model based on a differential equation to demonstrate how the number of asymptomatic patients grows over time and Kumari, *et al.*[34] proposed multiple linear regression with autoregression used to predict the possible number of cases in the future.

Similarly, models based on countries like Brazil and Italy do not work well with high RMSE values, meaning that other models can be used instead. Other studies with focus on Brazil used models such as SIR,[35] Holt[36] and artificial intelligence (AI) models.[37] Studies on Italy focused on extended SIR (eSIR) models[38] and mathematical models with a Gaussian error function type.[39]

As shown in the results, ARIMA (5,2,5) was the best-fitting ARIMA model for Panama, Iran, and Spain when data was not transformed. Despite having the same model structure, the established model for Panama outperforms Iran and Spain with a lower RMSE values. The explanation may be due to a fluctuation in the number of cases in Iran, such as the rapid spread of COVID-19 in Iran at the start of the pandemic, and misunderstandings that led to ignoring the issue of social distance while eliminating travel restrictions one by one in April 2020 resulting in the reappearance of the virus.[40] Another reason may be that the Iran COVID-19 data behavior is more complicated due to the geographical correlation of cases in Iran[41] and the variation caused by sudden rises in the number of cases in different parts of the country at different times.

Another reason for the models' disparities in performance could be the effect of weather factors on COVID-19 cases, which is being investigated by Fernández-Ahúja and Martínez.[42] While Panama has a tropical maritime climate, Spain is home to four distinct climates. Climate variables clarify some key aspects of COVID-19 spread in Spain[42] that are not captured by ARIMA models, whereas the influence of such variables in Panama will be less due to more consistent weather. In the same line, Gupta, *et al.*[43] investigated the impact of weather on COVID-19 spread in the United States, while the derived model for the United States of America has a high RMSE value. Other models has been used for Iran such as LSTM,[44] Recursive-based prediction model, Boltzmann function-based prediction model and Beesham's prediction model.[45] LSTM[46] and SEIR[47] have been employed to model the cases in USA and Spain.

While the developed model works well in African countries like Uganda, Congo, and Algeria, the model for South Africa has a high RMSE value, indicating that different models should be used for this country which is investigated by Ding, *et al.*[48] with SIR model and Reddy, *et al.*[49] with a set of nonlinear growth models and Nadim and Chattopadhyay[50] with mathematical model considering the imperfect lockdown effect. Similarly, the developed model for Ethiopia shows high RMSE value. Other studies researched on COVID-19 cases in Ethiopia with machine learning algorithms[51] and mathematical models based on susceptible, exposed, symptomatically infected, asymptomatically infected, hospitalized and recovered/immune compartments.[52]

As results show, developing models for a number of countries, including China and Eswatini, performs better on transformed results. Both countries have seen a rapid fluctuation in the number of cases, which can be due to a change in how the virus is diagnosed and the number of diagnostic tests conducted. While data transformation in ARIMA models can help to improve model performance by removing skewness and fluctuation from the original data, other data processing methods such as machine learning can be used for data preprotein as discussed by Pinter, *et al.*[53] to improve

Table 1. Countries and literature on modelling COVID-19 cases.

Country	Alternative models used in literature
India	LSTM model,[19] Tree-based model,[30] SEIR model,[31] Compartmental epidemic SIPHERD model,[32] Mathematical model based on a differential equation,[33] Multiple linear regression with autoregression[34]
Brazil	SIR model,[35] Holt model,[36] Artificial intelligence (AI) models,[37] Data driven models[54]
Italy	extended SIR (eSIR) model,[38] Mathematical models with Gauss error function type[39]
Iran	LSTM model,[44] Recursive-Based prediction model, Boltzmann function-based prediction model, Beesham's prediction model[45]
Spain	LSTM model,[46] SEIR model[47]
United states	LSTM model,[46] SEIR model[47]
South Africa	SIR model,[48] Nonlinear growth models,[49] Mathematical model considering the imperfect lockdown effect[50]
China	Hybrid AI model[55]
Ethiopia	Machine learning models,[51] Mathematical model based on compartmental approach of susceptible, exposed, symptomatically infected, asymptomatically infected, hospitalized and recovered/immune compartments[52]

the performance of models such as SIR models. In the same line, Other models for predicting new cases in China and Brazil that are developing AI and data-driven models include a data pre-processing phase.[54,55] Table 1 summarises the section and the studies conducted on countries where the developed ARIMA model does not provide adequate performance.

Conclusion

In this study, we presented an automated modelling platform that delves into multiple layers of information in the COVID-19 time series data to find the best fit with the aim of providing robust forecasts. COVIDSpread was shown to be effective in estimating the trend of the pandemic for each country. We elaborated the importance of data transformation as a preprocessing step and shown that there is no transformation operation which consistently provides the best fit to the data. Hence, exploring multiple options are recommended to stabilize variations prior to modelling using conventional econometrics formulations. A unique aspect of the presented platform is that it facilitates real-time model development incorporating latest reported data into modelling. We have shown that such adaptive model estimation significantly improves the prediction power and therefore, forecasting reliability.

Data availability

Underlying data

The platform retrieves daily number of confirmed cases from Coronavirus Resource Centre at John Hopkins University using *coronavirus* R package.

Extended data

Zenodo: VafaeeLab/COVIDSpread: First release of COVIDSpread, https://doi.org/10.5281/zenodo.5587835.[56]

This project contains the following extended data:

- Extended-Table-1.xlsx (represents model specifications and performance across 162 countries as of the time of submission)

Software availability

COVIDSpread is available online: http://vafaeelab.com/COVID19TS.html

All the codes, including Shiny app is available at the project GitHub Repository: https://github.com/VafaeeLab/COVID-Spread

Archived code as at time of publication: https://doi.org/10.5281/zenodo.5587835[56]

License: Apache License, Version 2.0

Acknowledgments

SS and THR acknowledge the support from the Australian Research Council under Linkage Scheme (LP160100450). THR acknowledges the support from the Australian research Council under the DECRA Scheme (DE170101346). An earlier version of this article can be found on medRxiv (doi: https://doi.org/10.1101/2020.04.24.20078923).

References

1. Fetzer T, *et al.*: **Pandemics and social capital: From the Spanish flu of 1918-19 to COVID-19.** 2020.
2. *COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). ArcGIS.* Johns Hopkins University; 12 September 2021.
3. Petropoulos F, Makridakis S: **Forecasting the novel coronavirus COVID-19.** *PLoS One.* 2020; **15**: e0231236.
PubMed Abstract | Publisher Full Text | Free Full Text.
4. Liu Y, Gayle AA, Wilder-Smith A, *et al.*: **The reproductive number of COVID-19 is higher compared to SARS coronavirus.** *J. Travel Med.* 2020; **27**.
Publisher Full Text
5. Ferguson N, *et al.*: **Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand.** 2020.
6. Kucharski AJ, *et al.*: **Early dynamics of transmission and control of COVID-19: a mathematical modelling study.** *Lancet Infect. Dis.* 2020; **20**: 553–558.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Chinazzi M, *et al.*: **The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak.** *Science.* 2020; **368**: 395–400.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Wells CR, *et al.*: **Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak.** *Proc. Natl. Acad. Sci.* 2020; **117**: 7504–7509.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Hellewell J, *et al.*: **Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts.** *The Lancet Glob. Health.* 2020; **8**: e488–e496.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Lopez LR, Rodo X: **A modified SEIR model to predict the COVID-19 outbreak in Spain: simulating control scenarios and multi-scale epidemics.** *medRxiv.* 2020.
11. Tang B, *et al.*: **The effectiveness of quarantine and isolation determine the trend of the COVID-19 epidemics in the final phase of the current outbreak in China.** *Int. J. Infect. Dis.* 2020.
12. Moghadas SM, *et al.*: **Projecting hospital utilization during the COVID-19 outbreaks in the United States.** *Proc. Natl. Acad. Sci.* 2020; **117**: 9122–9126.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Prem K, *et al.*: **The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study.** *Lancet Infect. Dis.* 2020.
14. Anastassopoulou C, Russo L, Tsakris A, *et al.*: **Data-based analysis, modelling and forecasting of the COVID-19 outbreak.** *PLoS One.* 2020; **15**: e0230405.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Hu Z, Ge Q, Jin L: **Artificial intelligence forecasting of covid-19 in china.** *arXiv.* 2020; **07112**.
16. He Z, Tao H: **Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: A nine-year retrospective study.** *Int. J. Infect. Dis.* 2018; **74**: 61–70.
PubMed Abstract | Publisher Full Text
17. Shahriari S, Ghasri M, Hossein Rashidi T: **Ensemble of ARIMA: Combining Parametric and Bootstrapping Techniques for Traffic Flow Prediction.** *Transportmetrica.* 2020; **16**: 1552–1573.
Publisher Full Text
18. Muhareb R, Giacaman R: **Tracking COVID-19 responsibly.** *Lancet.* 2020.
Publisher Full Text
19. Tomar A, Gupta N: **Prediction for the spread of COVID-19 in India and effectiveness of preventive measures.** *Sci. Total Environ.* 2020; **728**: 138762.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Chimmula VKR, Zhang L: **Time series forecasting of COVID-19 transmission in Canada using LSTM networks.** *Chaos, Solitons Fractals.* 2020; **135**: 109864.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Shastri S, Singh K, Kumar S, *et al.*: **Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study.** *Chaos, Solitons Fractals.* 2020; **140**: 110227.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Arora P, Kumar H, Panigrahi BK: **Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India.** *Chaos, Solitons Fractals.* 2020; **139**: 110017.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Hawas M: **Generated time-series prediction data of COVID-19' s daily infections in Brazil by using recurrent neural networks.** *Data Brief.* 2020; **32**: 106175.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Alzahrani SI, Aljamaan IA, Al-Fakih EA: **Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions.** *J. Infect. Public Health.* 2020; **13**: 914–919.
PubMed Abstract | Publisher Full Text | Free Full Text
25. Shahid F, Zameer A, Muneeb M: **Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM.** *Chaos, Solitons Fractals.* 2020; **140**: 110212.
PubMed Abstract | Publisher Full Text | Free Full Text
26. Ribeiro MHDM, da Silva RG, Mariani VC, *et al.*: **Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil.** *Chaos, Solitons Fractals.* 2020; **135**: 109853.
PubMed Abstract | Publisher Full Text | Free Full Text
27. Kırbaş İ, Sözen A, Tuncer AD, *et al.*: **Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches.** *Chaos, Solitons Fractals.* 2020; **138**: 110015.
PubMed Abstract | Publisher Full Text | Free Full Text
28. Papastefanopoulos V, Linardatos P, Kotsiantis S: **Covid-19: A comparison of time series methods to forecast percentage of active cases per population.** *Appl. Sci.* 2020; **10**: 3880.
Publisher Full Text
29. Devaraj J, *et al.*: **Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant?.** *Results in Physics.* 2021; **21**: 103817.
PubMed Abstract | Publisher Full Text | Free Full Text
30. Arti M, Bhatnagar K: **Modeling and predictions for COVID 19 spread in India.** 2020.no. April.
31. Bherwani H, Gupta A, Anjum S, *et al.*: **Exploring dependence of COVID-19 on environmental factors and spread prediction in India.** *npj Climate and Atmospheric Science.* 2020; **3**: 1–13.
Publisher Full Text
32. Mahajan A, Sivadas NA, Solanki R: **An epidemic model SIPHERD and its application for prediction of the spread of COVID-19 infection in India.** *Chaos, Solitons Fractals.* 2020; **140**: 110156.
PubMed Abstract | Publisher Full Text | Free Full Text
33. Roy S, Roy Bhattacharya K: **Spread of COVID-19 in India: A mathematical model.** 2020. *SSRN 3587212.*
34. Kumari R, *et al.*: **Analysis and predictions of spread, recovery, and death caused by COVID-19 in India.** *Big Data Mining and Analytics.* 2021; **4**: 65–75.
Publisher Full Text
35. Bastos SB, Cajueiro DO: **Modeling and forecasting the early evolution of the Covid-19 pandemic in Brazil.** *Sci. Rep.* 2020; **10**: 1–10.
Publisher Full Text
36. Martinez EZ, Aragon DC, Nunes AA: **Short-term forecasting of daily COVID-19 cases in Brazil by using the Holt's model.** *Rev. Soc. Bras. Med. Trop.* 2020; **53**.
Publisher Full Text
37. da Silva RG, Ribeiro MHDM, Mariani VC, *et al.*: **Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables.** *Chaos, Solitons*

Fractals. 2020; **139**: 110027.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Wangping J, *et al.*: **Extended SIR prediction of the epidemics trend of COVID-19 in Italy and compared with Hunan, China.** *Front. Med.* 2020; **7**: 169.
Publisher Full Text

39. Ciufolini I, Paolozzi A: **Mathematical prediction of the time evolution of the COVID-19 pandemic in Italy by a Gauss error function and Monte Carlo simulations.** *The European Physical Journal Plus.* 2020; **135**: 355.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Ghanbari B: **On forecasting the spread of the COVID-19 in Iran: The second wave.** *Chaos, Solitons Fractals.* 2020; **140**: 110176.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Ramírez-Aldana R, Gomez-Verjan JC, Bello-Chavolla OY: **Spatial analysis of COVID-19 spread in Iran: Insights into geographical and structural transmission determinants at a province level.** *PLoS Negl. Trop. Dis.* 2020; **14**: e0008875.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Fernández-Ahúja JML, Martínez JLF: **Effects of climate variables on the COVID-19 outbreak in Spain.** *Int. J. Hyg. Environ. Health.* 2021; **234**: 113723.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Gupta S, *et al.*: **Effect of weather on COVID-19 spread in the US: A prediction model for India in 2020.** *Sci. Total Environ.* 2020; **728**: 138860.

44. Wang P, *et al.*: **Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran.** *Chaos, Solitons Fractals.* 2020; **140**: 110214.

45. Niazkar M, *et al.*: **Assessment of Three Mathematical Prediction Models for Forecasting the COVID-19 Outbreak in Iran and Turkey.** *Comput. Math. Methods Med. 2020.* 2020.

46. Kumar M, *et al.*: **Spreading of COVID-19 in India, Italy, Japan, Spain, UK, US: A prediction using ARIMA and LSTM model.** *Digital Government: Research and Practice.* 2020; **1**(4): 1–9.

47. Efimov D, Ushirobira R: **On an interval prediction of COVID-19 development based on a SEIR epidemic model.** *Annu. Rev. Control.* 2021.

48. Ding W, *et al.*: **Analysis and prediction of COVID-19 epidemic in South Africa.** *ISA Trans.* 2021.

49. Reddy T, *et al.*: **Short-term real-time prediction of total number of reported COVID-19 cases and deaths in South Africa: A data driven approach.** *BMC Med. Res. Methodol.* 2021; **21**(1): 1–11.

50. Nadim SS, Chattopadhyay J: **Occurrence of backward bifurcation and prediction of disease transmission with imperfect lockdown: A case study on COVID-19.** *Chaos, Solitons Fractals.* 2020; **140**: 110163.

51. Khakharia A, *et al.*: **Outbreak prediction of COVID-19 for dense and populated countries using machine learning.** *Annals of Data Science.* 2021; **8**(1): 1–19.

52. Deressa CT, Duressa GF: **Modeling and optimal control analysis of transmission dynamics of COVID-19: The case of Ethiopia.** *Alex. Eng. J.* 2021; **60**(1): 719–732.

53. Pinter G, Felde I, Mosavi A, *et al.*: **COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach.** *Mathematics.* 2020; **8**: 890.

54. Pereira IG, *et al.*: **Forecasting Covid-19 dynamics in Brazil: A data driven approach.** *Int. J. Environ. Res. Public Health.* 2020; **17**: 5115.

55. Zheng N, *et al.*: **Predicting COVID-19 in China using hybrid AI model.** *IEEE Transactions on Cybernetics.* 2020; **50**: 2891–2904.

56. VafaeeLab: **VafaeeLab/COVIDSpread: First release of COVIDSpread (Version v1).** *Zenodo.* 2021.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ?

Version 1

Reviewer Report 04 March 2024

https://doi.org/10.5256/f1000research.77666.r247250

© **2024 Gore R.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Ross Gore
Old Dominion University, Norfolk, Virginia, USA

While reliable predictions of the pandemic trend are essential for policymaking and resource-allocation, there is a lack of an adaptive real-time modelling platform which evolves as new data arrives. ... Our platform offers an interactive online dashboard which efficiently generates country-wise predictive models, in real-time, based on the latest report of COVID-19 cases worldwide .. The main objective of this study is to introduce an easy-to-use and readily available statistical tool to develop rigor models for time series data of COVID-19 as data becomes available on a real time basis." This motivation could be improved by describing exactly how this platform will fit into a policymaking and resource-allocation workflow so that decisions can be made regularly by policy makers. Specifically, what current unanswered or ambiguous questions have been posed by policymakers that this platform actionably answers and what does the work/action flow look like for policy makers in the specific modeled countries when these questions/issues arise.

The paper would benefit from additional statistical analysis of the results presented in the paper. While the most effective model for each country is identified it is unclear if the best performing model is statistically significantly different from the presented alternatives. Conducting statistical tests to identify which model is the statistically significantly superior (or at least those with the same level of statistically significant performance) as the alternatives would benefit the paper.

The paper would benefit from a discussion of the limitations of the analysis. As it currently stands all the results at are the same geographic level (i.e. country). Identifying research (Ref [1,2,3,4]) that has shown that statistics related to the prevalence of respiratory viruses can depend on the geographic level of evaluation (i.e country vs. city vs. county vs. neighborhood). For example, different geographic areas experience different transmission rates of respiratory viruses. Similarly, there can be stark differences in the vaccination levels at the country vs. city vs. county vs. neighborhood level. Furthermore, the vaccines (and their associated efficacy) available to individuals in different countries can differ. Identifying exactly which the context in which the authors expect these results to generalize and would improve the validity of the paper and make it more actionable for public health professionals. References related to the importance of geographic context in understanding respiratory virus transmission (including SARS-CoV-2) are

added to the review.

There are a nontrivial number of readers who suffer from red/green color blindness. Using a color-blindness safe color palette in the graphics in the web application, and figures 1, 3, would improve the paper. A series of color blind safe palettes are available here (https://davidmathlogic.com/colorblind/#%23D81B60-%231E88E5-%23FFC107-%23004D40) and larger text fonts would improve the readability of figures.

References

1. Cocco P, Meloni F, Coratza A, Schirru D, et al.: Vaccination against seasonal influenza and socio-economic and environmental factors as determinants of the geographic variation of COVID-19 incidence and mortality in the Italian elderly.*Prev Med*. 2021; **143**: 106351 PubMed Abstract | Publisher Full Text
2. Francis AI, Ghany S, Gilkes T, Umakanthan S: Review of COVID-19 vaccine subtypes, efficacy and geographical distributions.*Postgrad Med J*. 2022; **98** (1159): 389-394 PubMed Abstract | Publisher Full Text
3. Lynch CJ, Gore R: Short-Range Forecasting of COVID-19 During Early Onset at County, Health District, and State Geographic Levels Using Seven Methods: Comparative Forecasting Study.*J Med Internet Res*. 2021; **23** (3): e24925 PubMed Abstract | Publisher Full Text
4. Mao L, Yang Y, Qiu Y, Yang Y: Annual economic impacts of seasonal influenza on US counties: spatial heterogeneity and patterns.*Int J Health Geogr*. 2012; **11**: 16 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: data science, predictive analytics, modeling and simulation

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have

significant reservations, as outlined above.

Reviewer Report 31 March 2022

https://doi.org/10.5256/f1000research.77666.r123128

© **2022 Allahviranloo M.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Mahdieh Allahviranloo
Department of Civil Engineering, The City College of New York – CUNY, New York City, NY, USA

The paper presents an interactive dashboard which models the real-time spread of COVID-19 using time-series analysis, specifically Autoregressive Integrated moving average (ARIMA). The analysis does not rely on heavy assumption and simply generates the spread which can be utilized to predict short term spread. The paper is very well-written and easy to read. It also addresses an ongoing problem in the world, which is a plus. The design of the platform is very well thought and is user-friendly.

However, three main comments can be made to the paper:

1. The first comment is related to the spread of COVID-19. Spread of COVID depends on the proximity of the people and also the density of the population in a given area (crowdedness). How spatial parameters are factored in here? In fact, the impacts of density and concentration of crowd in areas, the attributes of the built environment (e.g. percentage of indoor, outdoor areas in the analysis site) are other factors that can be taken into account or at least being factored as a coefficient in the types of models. A discussion on further expansion of the study to incorporate these parameters is preferred.

2. I think depending on the scale of the regions, (neighborhood level, to city level to regional level), a tuning step is necessary in the model to account for these changes. Can the spread be illustrated and updated in different spatial granularity level? What modifications to the models are required to address different scenarios?

3. As the time has passed, the size of collected data on COVID-19 and its variants has increased. I would suggest authors to revise their models according to latest data from the available data repositories.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Transportation, Big data, Machine learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 13 Jul 2022

Fatemeh Vafaee

We would like to thank the reviewer for her time and valuable comments.
The comments are addressed below:

1. *"The first comment is related to the spread of COVID-19. Spread of COVID depends on the proximity of the people and also the density of the population in a given area (crowdedness). How spatial parameters are factored in here? In fact, the impacts of density and concentration of crowd in areas, the attributes of the built environment (e.g. percentage of indoor, outdoor areas in the analysis site) are other factors that can be taken into account or at least being factored as a coefficient in the types of models. A discussion on further expansion of the study to incorporate these parameters is preferred."*

   Response: We acknowledge the reviewer's comment. We would like to stress that due to the nature of time series models each prediction is based on the previous lagged observations and residual error from a moving average model applied to lagged observations. One difference from standard linear regression is that the data are not necessarily independent. As a result, the impact of spatial parameters such as density, the concentration of crowd in areas, and the attributes of the built environment (e.g. percentage of indoor, and outdoor areas in the analysis site) are included in the model for each prediction, spatial parameters had their impact on the previous observations (as long as no changes applied to these variables). The model should be altered using exogenous regressors to account for the influence of each parameter in the event that any change occurs for the parameters (such as concentration of crowd in particular places).

2. *"I think depending on the scale of the regions, (neighborhood level, to city level to regional level), a tuning step is necessary in the model to account for these changes. Can the spread be illustrated and updated in different spatial granularity level? What modifications to the models are required to address different scenarios?"*

Response: We appreciate receiving this comment. As mentioned in the previous comment, due to the nature of time series models, the scale of spread would be incorporated into the model (regressor of lagged observations). However, for converting the spread (calculated for a specific region) to another region, scaling factors could be defined based on variables (such as the ratio of population portion of two regions).

3. *"As the time has passed, the size of collected data on COVID-19 and its variants has increased. I would suggest authors to revise their models according to latest data from the available data repositories."*

   Response: There are examples provided in the manuscript regarding the model application (such as the effect of transformation and dynamic model estimation). The focus of such examples is to show model performance and features using a subset of data. As a result, updating the model with the latest data would not alter the point of those examples. On the other hand, the dashboard is constantly updated with the latest collected data on COVID-19.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research