Proceedings Paper:
Liu, X orcid.org/0000-0001-6354-2067, Li, K orcid.org/0000-0001-6657-0522, Yang, Z et al. (1 more author) (2021) A Regression Model for Short-Term COVID-19 Pandemic Assessment. In: Communications in Computer and Information Science. 6th International Conference on Life System Modeling and Simulation, LSMS 2020, and 6th International Conference on Intelligent Computing, 25 Oct 2020, Hangzhou, China. Springer Nature . ISBN 978-981-33-6377-9

https://doi.org/10.1007/978-981-33-6378-6_38

A Regression model for short-Term COVID-19 pandemic assessment

Xuan LIU[1][0000−0001−6354−2067], Kang LI[1][0000−0001−6657−0522], Zhile YANG[2][0000−0001−8580−534X], and Dajun DU[3]

[1] School of Electronic and Electrical Engineering, University of Leeds, Leeds, LS2 9JT, UK. k.li1@leeds.ac.uk

[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China zyang07@qub.ac.uk

[3] School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200072, China. ddj@shu.edu.cn

Abstract. COVID-19 has rapidly spread around the world in the past few months, researchers around the world are working around the clock to closely monitor and assess the development of this pandemic. In this paper, a time series regression model is built to assess the short-term progression of COVID-19 pandemic. The model structure and parameters are identified using COVID-19 pandemic data released by China within the time window from 22 January to 09 April 2020. The same model structure and parameters are applied to a few other countries for day ahead forecasting, showing a good fit of the model. This modeling exercise confirms that the underlying internal dynamics of this disease progression is quite similar. The differences in the impact of this pandemic on different countries are largely attributed to different eternal factors.

Keywords: COVID-19 · Regression Model · FRA · Data driven

1 Introduction

COVID-19 is an infectious disease which was first reported in December 2019 in Wuhan, China. This novel coronavirus rapidly swept the world in just a few months. Unlike most coronaviruses, COVID-19 is a new strain which has not been previously identified in humans. The symptoms of COVID-19 are mainly manifested in the respiratory system, such as coughing, shortness of breath, difficulty breathing and fever, etc [10]. COVID-19 is highly contagious, and it has been widely reported that the incubation period of the virus symptoms is long and asymptomatic infection is difficult to prevent [1]. For most common cold, the symptoms usually manifest within three days of infection. However, the effects of COVID-19 generally appear in 2 to 14 days. Up to April 20, 2020, there are more than 2.4 million people have been infected with COVID-19 and the total number of deaths has exceeded 165,000.

In order to assess and forecast the development of COVID-19 pandemic, great efforts have been made by researchers around the world to closely monitor and

assess the latest development of this pandemic. Among various approaches to model infectious diseases, compartmental models such as Susceptible-Infectious-Removed (SIR) model [9,4] and its derived ones are most popular [7,2]. However, the performance of these models are highly dependent on initial parameter settings and cannot address the causal factors in the development of epidemics [3]. Regression models [5,8] are also widely used for epidemics assessment and forecasting because the causal relationships between the dependent and independent variables could be inferred. In this paper, a time series linear regression model is developed for inferring the correlations among the epidemic data. Once the correlations are inferred, this model can be used for short-term epidemic forecasting without model adjustment, under the assumption that the underlying dynamics of this infectious disease is unchanged, and the differences in the outcomes are largely attributed by external factors such as intervention measures.

2 Modelling

Although the first outbreak of COVID-19 occurred in China, the spread of this infectious disease was then effectively controlled in March due to a series of intervention measures being implemented consequently. Therefore, public available data for the whole period can be analyzed as a reference for countries which are still suffering from this pandemic. It was noted that the number of counted daily new infections is highly dependent on the testing scale in a particular country. Therefore, it is highly likely that the number of counted daily new infections is much smaller than the actual infection number. In contrast, the counted number of daily deaths is more likely closer to the actual number than the confirmed infection cases, therefore it is more meaningful to analyze internal dynamics of this infectious disease using the number of daily reported deaths. Given the aforementioned considerations, the daily death number reported from China in the time window from 22 January to 09 April 2020 is used in this paper for building a time-series model.

Although the incubation period of COVID-19 and the time from symptom onset to final exit (death or recovery) vary for each infected individual, their expectations may be correlated to the daily death (DD) data. In other words, the DD value of day n, denoted as $Day\ (n)$ can be projected using the DD data of some specific days in the past. Thus, this problem can be converted to a regression modelling problem.

In order to determine the structure and parameters of the time-series model, a subset selection method namely Fast Recursive Algorithm (FRA) [6] is applied in this case to select and determine the model regression terms and parameters simultaneously. For a linear-in-the-parameter model to represent a time series, its discrete time form generally can be represented as

$$y = \Phi\ \Theta + \Xi \tag{1}$$

where $y = [y(1), ...y(m)]^T$ are the observations, $\Phi = [\phi_1, ...\phi_n]$ is the regression matrix and each $\phi_i = [\phi(1), ...\phi(m)]^T$ $(i = 1,...\text{n})$ contains all candidate regression

terms. $\Theta = [\theta_1, ...\theta_n]^T$ is the set of unknown parameters to be identified. $\Xi = [\xi_1, ...\xi_m]^T$ is the residual matrix of the model. In FRA, there are two predefined recursive matrices, M_k and R_k to fulfil the forward model selection procedure as

$$M_k = \Phi_k^T \ \Phi_k \tag{2}$$

$$R_k = I - \Phi_k M_k^{-1} \Phi_k^T \tag{3}$$

where Φ_k contains the first columns of the full regression matrix. Then, it has

$$R_{k+1} = R_k - \frac{R_k \phi_{k+1} \phi_{k+1}^T R_k^T}{\phi_{k+1}^T R_k \phi_{k+1}}, \qquad k = 0, 1...(n-1) \tag{4}$$

Define E_k as the cost function, as the first k columns in Φ are selected, E_k can be expressed as

$$E_k = y^T R_k y \tag{5}$$

Then, using Eq.(4) and (5), it has

$$E_{k+1} = y^T R_{k+1} y = E_k - \frac{y^T R_k \phi_{k+1} \phi_{k+1}^T R_k^T y}{\phi_{k+1}^T R_k \phi_{k+1}} \tag{6}$$

Therefore, the net contribution of the selected model term ϕ_{k+1} to the cost function can be calculated as

$$\Delta E_{k+1} = -\frac{(y^T \phi_{k+1}^k)^2}{(\phi_{k+1}^k)^T \phi_{k+1}^k} = \frac{(a_{k+1,y}^T)^2}{a_{k+1,k+1}} \tag{7}$$

The net contribution of each term can be calculated and the terms with maximum contributions will be selected. Finally, the model parameters can be identified by the procedure as

$$\hat{\theta_j} = \frac{aj, y - \sum_{i=j+1}^{k} \hat{\theta}_i a_{j,i}}{a_{j,j}} \qquad j = k, k-1, ...1 \tag{8}$$

The last step is to determine how many of the most contributing terms should be selected. If too few terms are selected, some important candidates may be abandoned which may lead to underfitting. If too many terms are selected, the terms which have little contribution will be selected. This not only increases computing cost, but may also lead to overfitting the model. In order to address this problem, the DD data in China is divided into two parts. 80% of the data (22 Jan 2020 to 25 Mar 2020) is used for model training and the other 20% of data is used for model validation. Therefore, a set of models that have various number of most contributing terms can be generated. These models will be validated on the validation data and their root mean square errors (RMSE) will be compared as the an indicator. The model which produces the minimum RMSE in validation will be considered as the most suitable model in this paper.

3 Results and Discussion

3.1 Model analysis based on Pandemic data from China

As mentioned earlier, the models generated in Section 2 are trained with 80% of the DD data and validated with the other 20% of the DD data from China. The validation results indicate that the model which selected 3 most important terms ($\phi(n-1)$, $\phi(n-2)$ and $\phi(n-6)$) has the best performance. While $\phi(n-1)$ and $\phi(n-2)$ imply the persistence of the pandemic progression, $\phi(n-6)$ implies the inherent latency effect of this disease. The model parameters are identified as θ_1 = 0.6728, θ_2 = 0.4076 and θ_3 = -0.1114.

In order to determine the system stability, Z-transform technique is applied to the developed model to convert the discrete-time expression into a frequency domain representation. The Z-transform of the developed model can be expressed as

$$z^6 = 0.6728 - z^5 + 0.4076 * z^4 - 0.1114 \tag{9}$$

Therefore, the poles of the system can be calculated as shown in Table 1, and their locations are presented in Figure 1.

Table 1. System poles

Poles	P_1	P_2	P_3	P_4	P_5	P_6
	0.7523	0.96104	0.05873 + 0.59618i	0.05873 - 0.59618i	-0.579 + 0.30677i	-0.579 - 0.30677i

All of the system poles are located within the unit circle, therefore, the internal stability and convergence of the system can be guaranteed. It worth noting that there is a pole (P_2) very close to the unit circle, which implies the stability margin of the system is small. The system may approach to its critical point of stability at the peak region, which may lead to oscillatory fluctuations. The oscillatory fluctuations are particularly visible at the inflecting point and after the external interventions are introduced. Moreover, two complex-conjugate pole pairs (P_3 and P_4, P_5 and P_6) represent the system oscillating behaviours due to the external disturbances. The locations of P_3 and P_4 are almost on the imaginary axis, there will be little damping in their generated oscillations. P_5 and P_6 have larger phase angles with the real axis than the ones of P_3 and P_4, which implies they may cause oscillatory fluctuations with higher frequency.

The model predictions compared with the real data from China is shown in Fig. 2. The RMSEs of model training and validation are 15.8306 and 2.1756 respectively. It is shown that the output of the model is in good agreement with the actual data. The mean absolute error (MAE) is 0.962 which is less than 2.3% in relative error. However, it is worth noting that the maximum absolute error (MaxAE) is 54 which occurs on 24 February. The reason for the large error on this day is that the data on this day has a 52.67% sharp drop compared to the

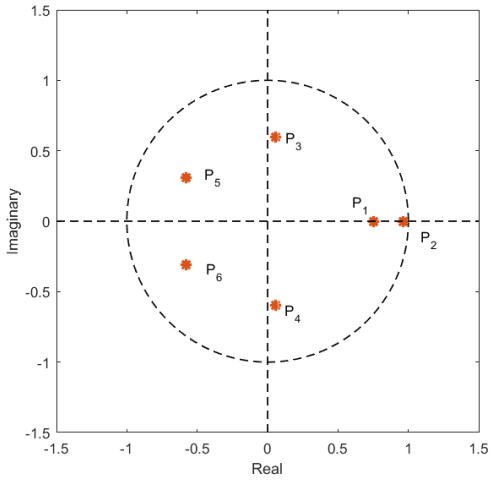

Fig. 1. The locations of the system poles in Z-plane

previous day. According to Figure 2, it is worth noting that large oscillations occur at the peak region, and oscillations are also observed at the downhill region. As discussed earlier, the oscillatory fluctuations are likely caused by the latency effect of the internal process.

From the system perspective, relatively strong oscillations occur in the region near the peak (12 to 27 February). The model parameters are more sensitive to measurement errors in this region than others. Thus, it may be inferred that when the system generates large oscillations, it is a sign of the apex.

3.2 Applications of the trained model

To validate our assumption that the pandemic progress in other countries has a similar internal dynamics as being observed in China, we applied the trained model presented in section 3.1 directly to other six countries which include UK, Italy, Germany, Spain, France and USA. The same model structure and parameters in Section 3.1 are applied to the data (15 February to 08 April 2020) of these countries. The model outputs and the actual DD data of these countries are presented in Figure 3. Four indicators namely root mean square error (RMSE), mean absolute error, maximum absolute error and coefficient of determination (R^2) are used to evaluate the model performance. The results of model performance are summarized in Table 2.

It is shown that the developed model performs well for these countries. The coefficient of determination for all of these countries are above 0.5 and most of them are greater than 0.87. This implies that the developed model has a good day-ahead forecasting performance, and internal dynamics of this pandemic progression is quite similar. Thus, the model developed using the data reported

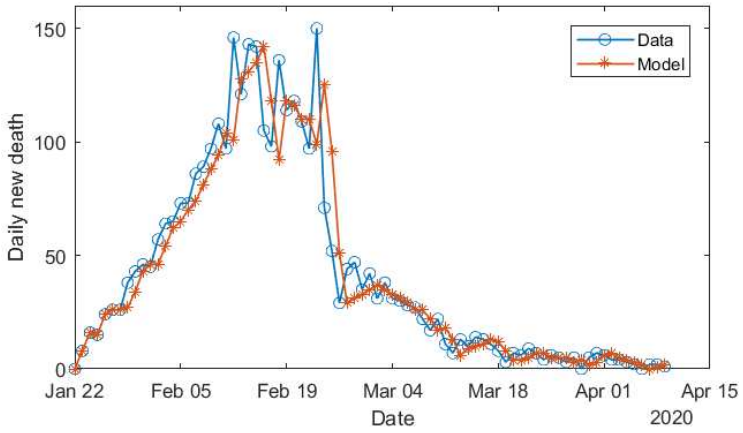

Fig. 2. Model outputs based on the data from China

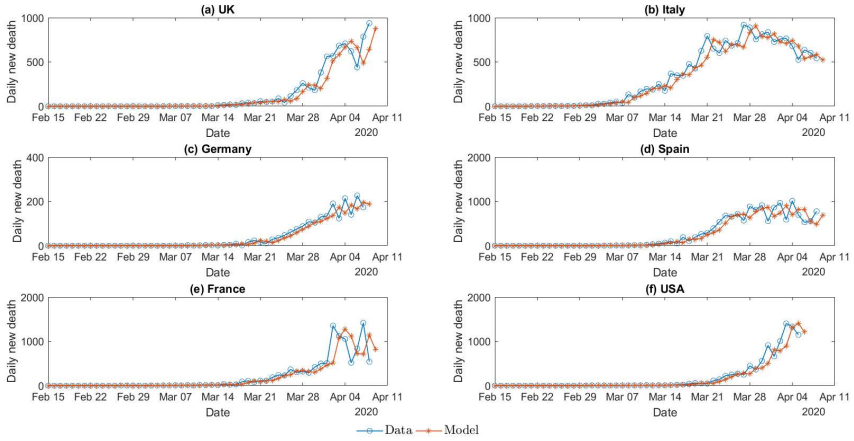

Fig. 3. One-day-ahead daily death toll forecasting of COVID-19. (a) UK, (b) Italy, (c) Germany, (d) Spain, (e) France, (f) USA.

Table 2. Results of model performance

Country	RMSE	Mean absolute error	Maximum absolute error	R^2
UK	83.4715	37.8551	300.2394	0.879
Italy	78.3751	48.5672	249.9158	0.941
Germany	18.6790	9.4191	65.8328	0.9095
Spain	119.8564	66.9247	319.1592	0.872
France	194.9673	73.6698	842.4610	0.6874
USA	112.3940	45.3986	507.3930	0.9016

in China can be directly transferable to other countries. Through the internal dynamics is similar, but the outcomes of this pandemics vary significant in different countries. Some countries have high fatality rate, while others have less. This is largely due to external factors, such as the differences in the culture, social background, health care systems, and infections disease control and intervention measures introduced by different governments. Different levels of system oscillations are also observed in these six countries at the peak and down-hill regions.

4 Conclusion

In this paper, a time series regression model is developed for assessing the internal dynamics of the COVID-19 pandemic. The model structure and parameters are determined using the FRA technique and are based on the data from China. The regression terms $\phi(n-1)$ and $\phi(n-2)$ are favored to reflect the persistence of the pandemic progression, and the other regression term $\phi(n-6)$ is selected, implying the inherent latency effect of the pandemic. Since all the system poles locate inside the unit circle in the Z-plane, the internal stability and convergence of the system can be ensured. The position characteristics of the system poles can reflect the negative latency effect which may causes oscillatory fluctuations around the inflecting point or due to measurement errors and system disturbances. Since the oscillatory fluctuations are particularly evident at the inflection point, the developed model may be used as an indicator to analyze if the pandemic progression is approaching the inflection point. This model is directly applied to 6 other countries. The application results reveal that the model fits fairly well with the data from other countries, and can be used for short-term forecasting of the pandemic progression. The model coefficients of determination (R^2) of all the simulated countries are greater 0.5 and most of them are greater than 0.87. Moreover, as the pandemic reaches the peak, large oscillations are observed.

5 Disclaimer

All the data used in the paper are based on publicly available resources, including references to official and professional websites and peer-reviewed journals. The model, data and discussions presented in this paper are for research only. The model may not represent the real situation, and it may fail due to inadequate model elements.

References

1. Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D.Y., Chen, L., Wang, M.: Presumed asymptomatic carrier transmission of covid-19. Jama (2020)
2. Caccavo, D.: Chinese and italian covid-19 outbreaks can be correctly described by a modified sird model. medRxiv (2020)

3. Chen, D.: Modeling the spread of infectious diseases: A review. Analyzing and modeling spatial and temporal dynamics of infectious diseases pp. 19–42 (2014)
4. Fanelli, D., Piazza, F.: Analysis and forecast of covid-19 spreading in china, italy and france. Chaos, Solitons & Fractals **134**, 109761 (2020)
5. Harrell Jr, F.E., Lee, K.L., Matchar, D.B., Reichert, T.A.: Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer treatment reports **69**(10), 1071–1077 (1985)
6. Li, K., Peng, J.X., Irwin, G.W.: A fast nonlinear model identification method. IEEE Transactions on Automatic Control **50**(8), 1211–1216 (2005)
7. Peng, L., Yang, W., Zhang, D., Zhuge, C., Hong, L.: Epidemic analysis of covid-19 in china by dynamical modeling. arXiv preprint arXiv:2002.06563 (2020)
8. Wang, Y., Beydoun, M.A.: The obesity epidemic in the united states—gender, age, socioeconomic, racial/ethnic, and geographic characteristics: a systematic review and meta-regression analysis. Epidemiologic reviews **29**(1), 6–28 (2007)
9. Yang, Z., Zeng, Z., Wang, K., Wong, S.S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao, Z., Mai, Z., et al.: Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. Journal of Thoracic Disease **12**(3), 165 (2020)
10. Zheng, Y.Y., Ma, Y.T., Zhang, J.Y., Xie, X.: Covid-19 and the cardiovascular system. Nature Reviews Cardiology pp. 1–2 (2020)