Received: 8 June 2021 | Revised: 10 July 2021 | Accepted: 25 July 2021

DOI: 10.1111/tbed.14263

Transboundary and Emerging Diseases WILEY

ORIGINAL ARTICLE

Discovering dynamic models of COVID-19 transmission

Jinwen Liang[1] | Xueliang Zhang[2] | Kai Wang[2] | Manlai Tang[3] | Maozai Tian[1,2]

[1] Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, China

[2] Department of Medical Engineering and Technology, Xinjiang Medical University, Urumqi, China

[3] Department of Mathematics, College of Engineering, Design & Physical Sciences, Brunel University, London, UK

Correspondence
Tian Maozai, Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, 100872, China.
Email: mztian@ruc.edu.cn

Abstract

Existing models about the dynamics of COVID-19 transmission often assume the mechanism of virus transmission and the form of the differential equations. These assumptions are hard to verify. Due to the biases of country-level data, it is inaccurate to construct the global dynamic of COVID-19. This research aims to provide a robust data-driven global model of the transmission dynamics. We apply sparse identification of nonlinear dynamics (SINDy) to model the dynamics of COVID-19 global transmission. One advantage is that we can discover the nonlinear dynamics from data without assumptions in the form of the governing equations. To overcome the problem of biased country-level data on the number of reported cases, we propose a robust global model of the dynamics by using maximin aggregation. Real data analysis shows the efficiency of our model.

KEYWORDS
COVID-19, global transmission, maximin aggregation, sparse identification of nonlinear dynamics (SINDy)

1 | INTRODUCTION

The outbreak of a new virus named SARS-CoV-2 was initially identified in mid-December 2019 in Wuhan, Hubei Province, China. COVID-19, the disease caused by this coronavirus, was characterized as a pandemic by WHO on 11 March 2020. As of 5 April 2021, the number of confirmed cases rose to 131,419,173 worldwide, and the number of deaths reached 2,854,842.

Research on the epidemic has sprung up in the past year. Prediction of the infectious cases, modelling the dynamics including differential equations and identifying relationships between the infectious cases and other factors, such as environmental factors, are the main concerns of statisticians. Studying the patterns of transmission can simulate the development of the current epidemic and predict the trend of the epidemic in the future, further, help governments make a decision. We focus on mechanistic equations for disease dynamics from case notification data for COVID-19 in this paper.

It is a huge challenge to discover the governing equations from real data. The data on major communicable diseases are more open, and the dynamic model of its transmission is relatively traditional and fixed. Classical mathematical models describing the spread of infectious diseases include the susceptible, infective, and recovered (SIR) model (Beretta & Takeuchi, 1995), susceptible, exposed, infective, and recovered (SEIR) model (Ma et al., 2004), susceptible, exposed, infective, diagnosed, and recovered model (C. Liu et al., 2004) and susceptible, infective, recovered and dead model (Rui & Tian, 2021), etc. The idea is to divide the population into susceptible (S), exposed (E), infected (I), confirmed (C) and recovered (R) populations, then reveal the law of epidemic transmission through the infection mechanism that how individuals move between the compartments. And these models are used to study the spread of infectious diseases such as measles, smallpox, rabies, Ebola viruses, etc. Due to the characteristic that COVID-19 has a relatively long incubation period and quarantine measures are implemented, it is hard to describe with the existing epidemic models. Researchers have been expanded these models in many ways to account for observed epidemic patterns (Chowdhury et al., 2020; Frenkel & Schwartz, 2021; Mandal et al., 2020; Paiva et al., 2020; Rajendran & Jayagopal, 2021; Zhao & Chen, 2020).

The above models all prespecify assumptions of the mechanism of dynamic transmission of COVID-19. These assumptions are hard to verify, which reduces the possibility to receive accurate results. So we hope the dynamic transmission mechanism of COVID-19 is not

assumed in advance, and the governing equations are determined by the statistical method according to recorded time series data. Horrocks and Bauch (2020) showed that sparse identification of nonlinear dynamics (SINDy) can be applied to epidemiological data to yield models that describe observed epidemic patterns. The SINDy framework was proposed by Brunton et al. (2016) to extract governing equations from data. SINDy can be used for the discovery of new models (see Kaiser et al., 2017; Mangan et al., 2016, 2017; Markus et al., 2018; Rudy et al., 2016; Tran & Ward, 2017). So far, SINDy has not yet been tested in model discovery from empirical data on COVID19 dynamics. For one specific country or region, we can use SINDy as in Horrocks and Bauch (2020). But one difficulty we face is that country-level or regional-level data are heterogeneous, if we fit a regression model using global data with a mixture of all country or regional level data, there would be a significant error. Estimating regressions over the pool of all available data are likely to estimate effects that might be strong for one country but very weak for another country, resulting in too many effects that are not replicable. For example, L. Liu et al. (2021) used a dynamic panel data model to generate density forecasts for daily active COVID-19 infections for a panel of countries/regions. Siwiak et al. (2020) provided a high-resolution global model of the pandemic but prespecified assumptions of the mechanism.

For inhomogeneous data, Bühlmann and Meinshausen (2015) showed a different type of aggregation can still lead to consistent estimation of the effects which are common in all heterogeneous data, the so-called maximin effects (Meinshausen & Bühlmann, 2015). The maximin aggregation, also called magging, is simple and general. Applying maximin effects to the SINDy framework, we can provide a robust global transmission model.

In this article, we apply SINDy to time series data from COVID-19 to discover dynamic models that govern its epidemic patterns on each grouped data, then use maximin aggregation to get a global dynamic model. One advantage of this research is we create a global model of the early stages of the pandemic that would overcome the problem of the heterogeneous data on the number of notification cases; the other advantage is we determine governing equations according to recorded time series data.

The remainder of this paper is structured as follows. In Section 2, we introduce SINDy and magging estimating procedures. In Section 3, we do some descriptive analysis of the data of the Johns Hopkins University dataset. In Section 4, a systematical analysis based on SINdy and magging is carried out. Finally, in Section 5, we give concluding remarks and future research proposals.

2 | MATERIALS AND METHODS

For given time t, $t = 1, \ldots, T$ and N_t, S_t, I_t, R_t, D_t represent the cumulative number of the total population, the number of the susceptible individuals, the number of the infected cases, the cumulative number of the recovery cases and the cumulative number of the death cases at time t, respectively.

The SINDy algorithm (Brunton et al., 2016) performs the intractable brute force search for all possible model functional terms. The historical data $x(t)$, which are sampled at several times $t_1, t_2, \ldots, t_n$, are arranged into two large matrices:

$$X = \begin{bmatrix} x^T(t_1) \\ x^T(t_2) \\ \vdots \\ x^T(t_n) \end{bmatrix} = \begin{bmatrix} x_1(t_1) & x_2(t_1) & \cdots & x_p(t_1) \\ x_1(t_2) & x_2(t_2) & \cdots & x_p(t_2) \\ \vdots & \vdots & \ddots & \cdots \\ x_1(t_n) & x_2(t_n) & \cdots & x_p(t_n) \end{bmatrix},$$

$$\dot{X} = \begin{bmatrix} \dot{x}^T(t_1) \\ \dot{x}^T(t_2) \\ \vdots \\ \dot{x}^T(t_n) \end{bmatrix} = \begin{bmatrix} \dot{x}_1(t_1) & \dot{x}_2(t_1) & \cdots & \dot{x}_p(t_1) \\ \dot{x}_1(t_2) & \dot{x}_2(t_2) & \cdots & \dot{x}_p(t_2) \\ \vdots & \vdots & \ddots & \cdots \\ \dot{x}_1(t_n) & \dot{x}_2(t_n) & \cdots & \dot{x}_p(t_n) \end{bmatrix},$$

where $\dot{x}(t)$ is the derivative of $\boldsymbol{x}(t)$.

Then, we construct an augmented library $\Theta(X)$ consisting of candidate nonlinear functions of the columns of X. For example, $\Theta(X)$ may consist of constant, polynomial and trigonometric terms

$$\Theta(X) = \begin{bmatrix} | & | & | & | & & | & | & | & | \\ 1 & X & X^{P_2} & X^{P_3} & \cdots & \sin(X) & \cos(X) & \sin(2X) & \cos(2X) \cdots \\ | & | & | & | & & | & | & | & | \end{bmatrix}.$$

Here, higher polynomials are denoted as X^{P_2}, X^{P_3}, etc. For example, X^{P_2} denotes the quadratic nonlinearities in the state variable x, given by

$$X^{P_2} = \begin{bmatrix} x_1^2(t_1) & x_1(t_1)x_2(t_1) & \cdots x_2^2(t_1) & x_2(t_1)x_3(t_1) & \cdots \\ x_1^2(t_2) & x_1(t_2)x_2(t_2) & \cdots x_2^2(t_2) & x_2(t_2)x_3(t_2) & \cdots \\ \vdots & \vdots & \ddots \vdots & \vdots & \cdots \\ x_1^2(t_n) & x_1(t_n)x_2(t_n) & \cdots x_2^2(t_n) & x_2(t_n)x_3(t_n) & \cdots \end{bmatrix}.$$

A sparse regressionproblem is set up with the sparse vectors of coefficients $\Xi = [\xi_1\ \xi_2\ \ldots\ \xi_p]$,

$$\dot{X} = \Theta(X)\Xi + \eta Z.$$

where Z is modelled as a matrix of independent identically distributed Gaussian entries with zero mean, and noise magnitude η. The matrix $\Theta(X)$ has dimensions $n \times m$, where m is the number of candidate nonlinear functions ($n \gg m$). And Ξ has dimensions $m \times p$.

Sparse regression techniques, such as LASSO (Tibshirani, 1996), adaptive LASSO ((Hui, 2006)), smoothly clipped absolute deviation (Fan & Li, 2001), elastic-net (Hui & Hastie, 2005), etc have been researched in a large number of studies. In this article, we use sequential least-squares proposed in Brunton et al. (2016); e algorithm is given below.

The kth row of the dynamical system is reconstructed as follows:

$$\dot{x}_k = \Theta\left(x_k^{\top}\right)\xi_k, \quad k = 1 \ldots, p.$$

ALGORITHM 1 SINDy algorithm

Input: $\dot{X}$, $\boldsymbol{X}$, λ is a sparsification knob.

1. Initial guess: $\hat{} = (\Theta(X))^{+}\dot{X}$.
2. Find small coefficients and threshold and regress dynamics onto remaining terms to find sparse $\hat{\Xi}$.
3. Repeat step 2 until convergence.

Output: $\hat{\Xi}$.

Applying SINDy to discover a continuous-time model involves determining the derivative vector $\dot{x}_t$. Using the discrete system, the response vector is $\dot{x}_t = x_t - x_{t-1}$. Take the SIR model for an example, in its most elementary version it can be written in discrete time as follows:

$$\dot{S}_t = S_t - S_{t-1} = \beta S_{t-1}\,(I_{t-1}/N),$$

$$\dot{I}_t = I_t - I_{t-1} = \beta S_{t-1}\,(I_{t-1}/N) - \gamma I_{t-1},$$

$$\dot{R}_t = R_t - R_{t-1} = \gamma I_{t-1},$$

where N is the (fixed) size of the population, β is the average number of contacts per person per time and γ is the rate of recovery or mortality.

Now we add country or regional level $g, g = 1 \ldots, G$. The kth row of the dynamical system is as follows:

$$\dot{x}_{gk} = \Theta\left(x_{gk}^{\top}\right) = \xi_{gk}, \quad g = 1, \ldots, G, \quad k = 1 \ldots, p.$$

As estimates for the ξ_{gk}, we use the sequential threshold least-squares estimator

$$\hat{\xi}_{gk} = \arg\ min_{\xi_{gk} \in R^m} \left\| \dot{x}_{gk} - \Theta\left(x_{gk}^{\top}\right)\xi_{gk} \right\|_2^2,$$

$$\text{s.t.}\ \left|\xi_{gkj}\right| > \lambda, \quad j = 1, \ldots, m,$$

where λ is a tuning parameter, the calculating process is illustrated in Algorithm 1, and Akaike's information criterion (AIC; Bozdogan, 1987) is used to select λ. Otherwise, the sparse regression technique as we mentioned above is also a good choice.

Here, let

$$\Sigma_g = n_g^{-1}\left(\Theta\left(x_{gk}^{\top}\right)_g^{\top}\Theta\left(x_{gk}^{\top}\right)_g\right),$$

the empirical explained variance in group g is

$$\hat{V}_{\xi gk}^{g} := \frac{2}{n_g}\xi_{gk}^{\top}\Theta(x_{gk}^{\top})_g^{\top}\dot{x}_{gk} - \xi_{gk}^{\top}\hat{\sum}_{g\xi gk}. \tag{1}$$

So the estimator for $\xi_{k\text{maximin}}$ according to Bühlmann and Meinshausen (2015) is

$$\xi_{k\text{maximin}} = \arg\min_{\xi_{gk} \in R^m}\ \max_{g\,=\,1,\,\ldots,\,G}(-\hat{V}_{\xi_k}^{g}).$$

$\xi_{k\text{maximin}}$ is approximately a combination of the weighted estimators of each group, and the weights are computed by quadratic programming. The optimization and computation can be implemented in a very efficient way; refer to Bühlmann and Meinshausen (2015) for more details. The steps in the Matlab environment are as follows:

Step 1. Calculate the empirical covariance matrix $^{\circ} = \frac{1}{n}\Theta(X)^{\top}\Theta(X)$, $n = n_1 + \cdots + n_G$.

Step 2. $H = \xi_k^{\top}\Sigma\xi_k$, $\xi_k = (\xi_{1k}, \ldots, \xi_{Gk})$.

Step 3. Use '**quadprog**' to calculate the weight vector **x**.

[x,fval,exitflag,output,lambda] = quadprog(H,d,[],[],Aeq,beq,lb), d = zeros(G,1); Aeq = ones(1,G);beq = 1; lb = zeros(G,1).

Step 4. The combination of the weighted estimators is $\xi_{k\text{maximin}} = x_1 * \xi_{1k} + \cdots + x_G * \xi_{Gk}$.

3 | DATA DESCRIPTION

The data we use in this paper were collected by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. Original global data are updated once a day. The reported time series started on 22 January 2020. The infected population is calculated as $I_t = C_t - R_t - D_t$, where C_t represents the cumulative confirmed patients at time t, R_t is the cumulative recovered patients at time t and D_t denotes the cumulative deaths cases at time t. Adding up regional data from the same country except for 'Diamond Princess', there are 191 countries in the global daily reports. There is a comparison between the numbers of cumulative confirmed cases on 10 March 2020, and 10 March 2021, as illustrated in Figure 1. The global epidemic is getting worse and worse.

By 10 March 2021, the top 10 countries with the largest number of confirmed cases are shown in Figure 2.

The susceptible population means more easily infected group and is an indispensable part of compartment models. But original dataset does not include susceptible cases. Here we add an assumption that the total population N_t is fixed with constraint $N_t = S_t + I_t + R_t + D_t$, which results in $N_t = S_t + C_t$. So we can replace S_t with C_t in constructing differential equations.

Accordingly, $x_t = (C_t, I_t, R_t, D_t)$ holds in our real data analysis. SINDy was then applied to the CSSE dataset using a function library consisting of all polynomials involving C_t, I_t, R_t, and D_t up to second order. That is

$$\Theta(x_t) = \left[1, C_t, I_t, R_t, D_t, C_t^2, C_tI_t, C_tR_t, C_tD_t, I_t^2, I_tR_t, I_tD_t, R_t^2, R_tD_t, D_t^2\right],$$

which is 15 dimensional. And the discrete system is as follows:

$$C_t - C_{t-1} = \Theta(x_t)\,\xi_1,$$

$$I_t - I_{t-1} = \Theta(x_t)\,\xi_2,$$

$$R_t - R_{t-1} = \Theta(x_t)\,\xi_3,$$

$$D_t - D_{t-1} = \Theta(x_t)\,\xi_4.$$

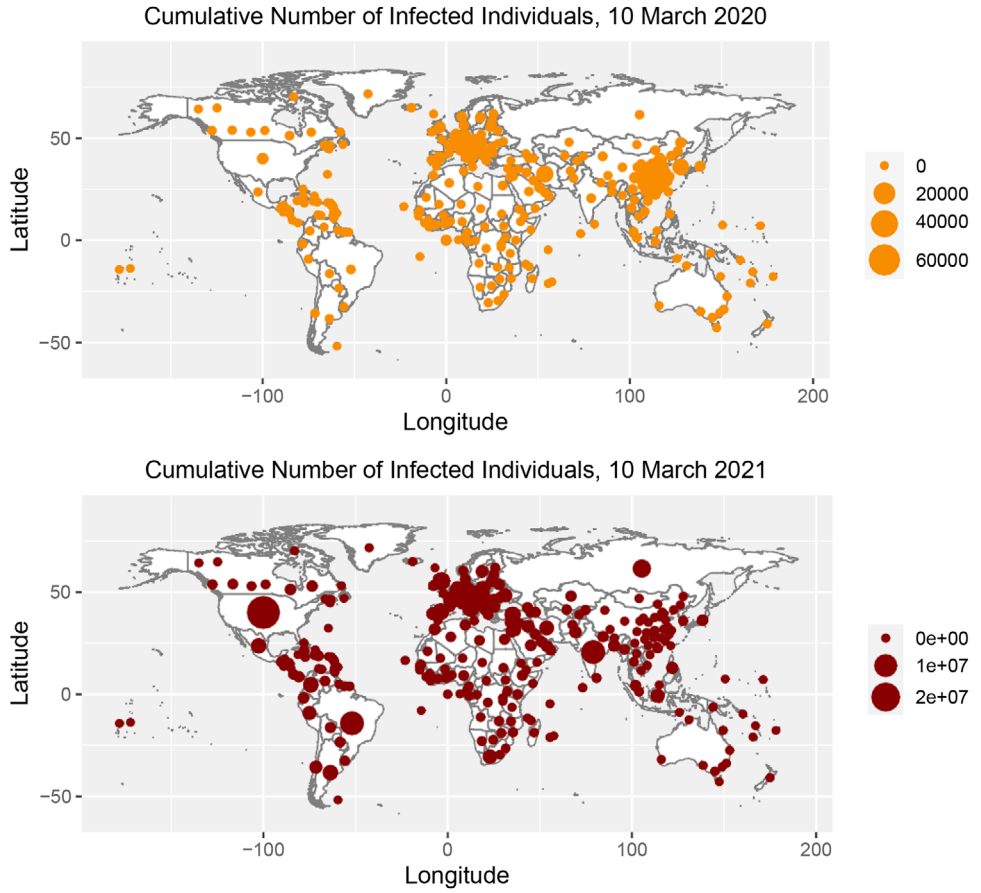

FIGURE 1 Bubble plot of numbers of cumulative confirmed cases on 10 March 2020, and 10 March 2021. The larger the bubble, the severer the situation

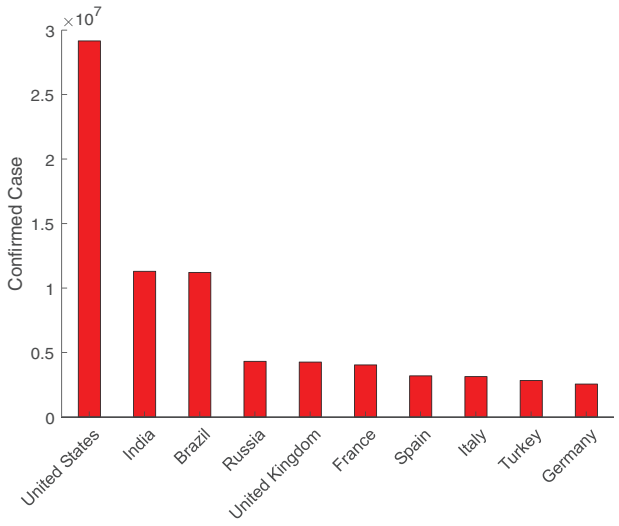

FIGURE 2 Number of cumulative confirmed cases for selected top ten countries on 10 March 2021

Remark: Polynomial terms often appear in traditional compartmental epidemic models, so we choose polynomials as library functions. When a third-order library is used instead, the results will be different due to the number of increasing parameters. Furthermore, there are severe noises and errors in the CSSE dataset, the final estimation is terrible. We only show our baseline analysis using a second-order polynomial library.

The AIC criterion is used to choose the optimal tuning parameter λ, which balances sparsity and overfitting. Models yielding the lowest AIC score across the grid will finally be selected.

The evaluation of the global transmission of COVID-19 is based on the top 10 countries with the largest number of confirmed cases. We choose dates $t_i, i = 1, \ldots, 10$, whose deaths number are greater than 20 as the start of the event time and dates $t_i + 56, i = 1, \ldots, 10$ as the end in our analysis.

Another way to group the global population is according to belonging to continents. As we know, the world is divided into seven continents, namely Asia, Europe, Africa, North America South America, Oceania, and Antarctica. No Antarctic data are reported in the table. Summing up data of the same continent, we get six summary data of continents, respectively. Different from country-level data, the start of the event time $t_i, i = 1, \ldots, 6$ is when deaths number exceeds 350, and the end is $t_i + 56, i = 1, \ldots, 6$.

We also analyze global summary data adding up all country and regional data as a comparison. The period is from 18 February 2020 to 13 April 2020, a 56-day interval.

TABLE 1 Coefficients comparison between three different models using a function library of polynomials up to second order

Methods	Terms	1	C	I	R	D	C2	CI	CR	CD	I2	IR	ID	R2	RD	D2
Country based	C Eq.	2722	−1.06	1.14	1.06	0.383	0	0	0	0	0	0	0	0	0	0
	I Eq.	2399	0.0018	−0.0059	−0.0195	−0.0923	0	0	0	0	0	0	0	0	0	0
	R Eq.	178.0	0.241	−0.233	−0.254	−0.217	0	0	0	0	0	0	0	0	0	0
	D Eq.	0	0.0565	0.0565	0.0565	0.0565	0	0	0	0	0	0	0	0	0	0
Continent based	C Eq.	2700	0.0069	−0.0147	−0.0218	0.0633	0	0	0	0	0	0	0	0	0	0
	I Eq.	−861.0	0.802	−0.81	−0.784	−0.32	0	0	0	0	0	0	0	0	0	0
	R Eq.	1611.0	0.0041	−0.0036	−0.0121	0.0458	0	0	0	0	0	0	0	0	0	0
	D Eq.	24.2	−0.0073	0.0082	0.0074	0.0186	0	0	0	0	0	0	0	0	0	0
Global based	C Eq.	0	−0.397	0.488	0.39	0	0	0	0	0	0	0	0	0	0	0
	I Eq.	0	−0.415	0.477	0.411	0	0	0	0	0	0	0	0	0	0	0
	R Eq.	0	0.0267	0	−0.0286	0	0	0	0	0	0	0	0	0	0	0
	D Eq.	0	−0.0127	0.0161	0.012	0	0	0	0	0	0	0	0	0	0	0

C, confirmed cases; CD, product of confirmed cases and death cases, C eq., equation of confirmed cases; CI, product of confirmed cases and infected cases; CR, product of confirmed cases and recovered cases; D, death cases, D eq., equation of death cases; I, infected cases; I eq., equation of infected cases; ID, product of infected cases and death cases; IR, product of infected cases and recovered cases; R, recovered cases; RD, product of recovered cases and death cases; R eq., equation of recovered cases.

4 | RESULTS

The calculating coefficients are illustrated in Table 1. Three models all remain one order terms and second-order terms make no difference.

To see the results more clearly, we predict the confirmed cases, infected cases, recovered cases, and deaths cases in 30 days of the early period. The confirmed cases, infected cases, recovered cases, and death case on 18 January 2020, was 557, 510, 30, and 17, respectively, which is the start point of the differential equations. Figure 3 displays the complete results. Global-based predictive cases are much less than country-based and continent-based predictive cases. To better illustrate the results, we use biaxial coordinates. Though the red line and the green line are closer, global-based predictions deviate most from true data because there is a mixture of all country-level biases. Continent-based models perform better in earlier confirmed and infected cases prediction, while country-based models perform better in latter prediction. For both recovered and death cases prediction, it is clear that country-based models are more accurate than continent-based models.

5 | DISCUSSION

In this paper, we demonstrate that SINDy can discover dynamical models from COVID-19 data.

The data have some limitations. Case notifications are typically under-reported and biased even when they are available. Infectious cases are measured with error because individuals who are asymptomatic are not captured in the official statistics. It is hard to decide whether death cases are dying with or dying of COVID-19. Also, counts of the number of recovered individuals are often inaccurate.

Due to state policies towards the disease are different, country-level data are heterogeneous. Fitting model curves on global data bear a significant error, as these data are a mixture of all country-level biases.

Estimating regressions over the pool of all available data is likely to estimate effects that might be strong for one country but very weak for another country, resulting in too many effects that are not replicable. To overcome the problem of biased and heterogeneous country-level data on the number of cases, we provide a robust global model of the dynamics by using the maximin aggregation technique. This model is simple and conservative, can reduce the complexity of data sources and extract a structure with the common contribution to all country data.

Regression algorithms and tuning parameter selection too influence the final results. Our approach replaces the susceptible time series with the confirmed time series based on the assumption. According to the research from Eastin and Eastin (2020), almost everyone is susceptible to COVID-19, which increases the difficulty for counts of accurate susceptible populations.

One problem is how to construct a library of functions, we can incorporate domain knowledge and other related methods, for example, compartment models. We select polynomials functions as a basis due to forms of compartment models. Deep methods (Long et al., 2018, 2019) are another technique that can provide more flexible representations to discover differential equations, but lack interpretations.

Regression coefficients in our model are fixed, they may be varying due to seasonal factors, etc. In our future research, we will consider regression coefficients are functions too.

In conclusion, we have shown that SINDy and maximin aggregation can be applied to COVID-19 data to yield models that describe global epidemic patterns.

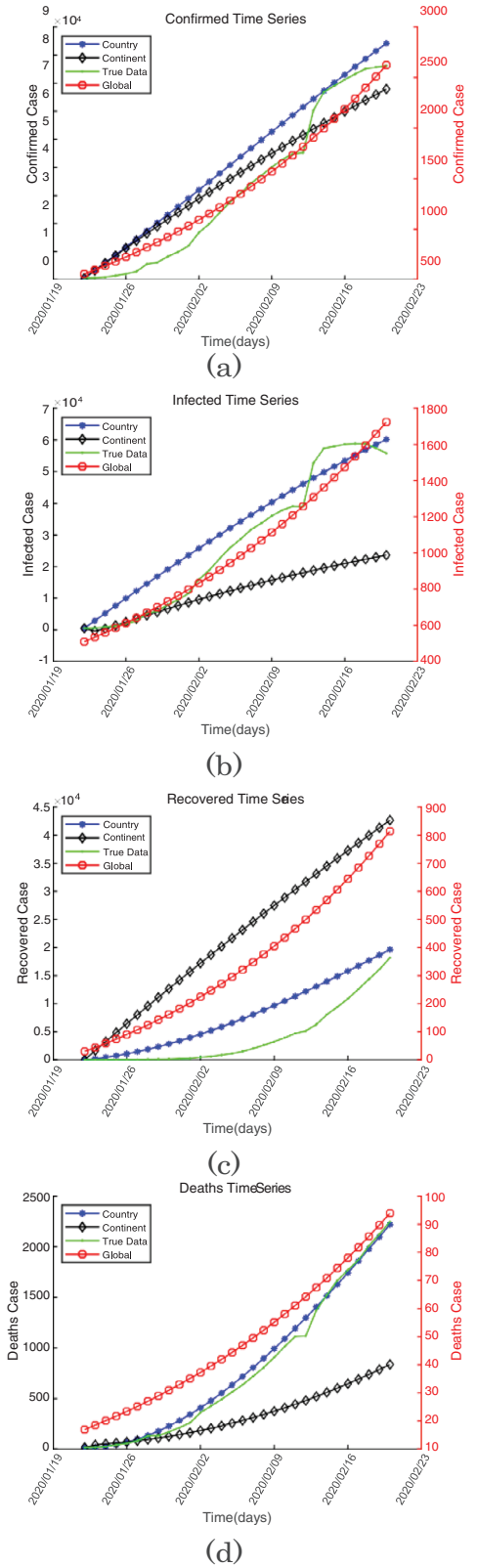

FIGURE 3 Comparison of the predictive number of confirmed cases, infected cases, recovery cases and death cases, respectively, in the first 30 days. Country based results (blue star line), continent based results (black rhombus), global-based results (red ring line) and true data (green dot line)

ACKNOWLEDGEMENTS

This work was supported by the Outstanding Innovative Talents Cultivation Funded Programs 2020 of Renmin University of China. The work of Man-Lai Tang was partially supported through grants from the Research Grant Council of the Hong Kong Special Administrative Region [UGC/FDS14/P01/16, UGC/FDS14/P02/18 and The Research Matching Grant Scheme (RMGS)] and a grant from the National Natural Science Foundation of China (Grant 11871124). Professor Tian's work was partially supported by the National Natural Science Foundation of China (No.11861042), and the China Statistical Research Project (No.2020LZ25). The computing facilities/software were supported by SAS Viya and the Big Data Intelligence Centre at the Hang Seng, University of Hong Kong.

CONFLICTS OF INTEREST

The authors declare no conflict of interest with respect to the research, authorship and/or publication of this article.

AUTHOR CONTRIBUTIONS

All authors conceived the study, carried out the analysis, discussed the results, drafted the manuscript, critically read the manuscript. Specifically, Dr. Jinwen Liang and Professor Xueliang Zhang contributed to all of the following: (1) conception and design of the work, acquisition of data; and (2) drafting the article or revising it critically for important intellectual content; Professors Kai Wang, Manlai Tang and Tian Maozai contributed to all of the following: data analysis and interpretation of data; final approval of the version to be published and agreement to be accountable for all aspects of the work.

ETHICAL STATEMENT

The data of daily reports were collected from CSSE at Johns Hopkins University and thus neither ethical approval nor individual consent was not applicable.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from CSSE at Johns Hopkins University online at https://github.com/CSSEGISandData/COVID-19.

ORCID

Maozai Tian https://orcid.org/0000-0002-0515-4477

REFERENCES

Beretta, E., & Takeuchi, Y. (1995). Global stability of an sir epidemic model with time delays. *Journal of Mathematical Biology*, *33*, 250–260. https://doi.org/10.1007/BF00169563

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370. https://doi.org/10.1007/BF02294361

Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 3932–3937. https://doi.org/10.1073/pnas.1517384113

Bühlmann, P., & Meinshausen, N. (2015). Magging: Maximin aggregation for inhomogeneous large-scale data. *Proceedings of the IEEE*, *104*, 126–135.

Chowdhury, R., Heng, K., Shawon, M., Goh, G., & Franco, O. H. (2020). Dynamic interventions to control covid-19 pandemic: A multivariate prediction modelling study comparing 16 worldwide countries. *European Journal of Epidemiology*, *35*, 389-399. https://doi.org/10.1007/s10654-020-00649-w

Eastin, C., & Eastin, T. (2020). Clinical characteristics of coronavirus disease 2019 in china. *Journal of Emergency Medicine*, *58*, 711–712. https://doi.org/10.1016/j.jemermed.2020.04.004

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360. https://doi.org/10.1198/016214501753382273

Frenkel, G., & Schwartz, M. (2021). Modeling social distancing and "spontaneous" infection in an epidemic outbreak phase application to the 2020 pandemic. *Physica A: Statistical Mechanics and its Applications*, *567*, 125727. https://doi.org/10.1016/j.physa.2020.125727

Horrocks, J., & Bauch, C. T. (2020). Algorithmic discovery of dynamic models from infectious disease data. *Scientific Reports*, *10*, 7061. https://doi.org/10.1038/s41598-020-63877-w

Hui, Z. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.

Hui, Z., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301–320.

Kaiser, E., Kutz, J. N., & Brunton, S. L. (2017). Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proceedings of The Royal Society A Mathematical Physical and Engineering Sciences*, *474*, 20180335. https://doi.org/10.1098/rspa.2018.0335

Liu, C., Ding, G., Gong, J., L. Wang, Cheng, K., & Zhang, D. (2004). Studies on mathematical models for SARS outbreak prediction and warning. *Chinese Science Bulletin*, *49*, 2245–2251.

Liu, L., Moon, H. R., & Schorfheide, F. (2021). Panel forecasts of country-level Covid-19 infections. *Journal of Econometrics*, *220*, 2–22. https://doi.org/10.1016/j.jeconom.2020.08.010

Long, Z., Lu, Y., & Dong, B. (2019). Pde-net 2.0: Learning pdes from data with a numeric-symbolic hybrid deep network. *Journal of Computational Physics*, *399*, 108925. https://doi.org/10.1016/j.jcp.2019.108925

Long, Z., Lu, Y., Ma, X., & Dong, B. (2018). Pde-net: Learning pdes from data. In *35th International Conference on Machine Learning* (ICML 2018) (pp. 5067–5078). Proceedings of Machine Learning Research, Vol. 80. International Machine Learning Society.

Ma, E., W, W., Zhou, Y., & Jin, Z. (2004). *Mathematical models and dynamics of infectious diseases* (1st edn.). China Science Press.

Mandal, M., Jana, S., Nandi, S. K., Khatua, A., & Kar, T. K. (2020). A model-based study on the dynamics of covid-19: Prediction and control. *Chaos Solitons & Fractals*, *136*, 109889.

Mangan, N. M., Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, *2*, 52–63. https://doi.org/10.1109/TMBMC.2016.2633265

Mangan, N. M., Kutz, J. N., Brunton, S. L., & Proctor, J. L. (2017). Model selection for dynamical systems via sparse regression and information criteria. *Proceedings Mathematical Physical & Engineering Sciences*, *473*, 2017009.

Markus, Q., Markus, A., Nathan, K. J., & Brunton, S. L. (2018). Sparse identification of nonlinear dynamics for rapid model recovery. *Chaos An Interdisciplinary Journal of Nonlinear Science*, *28*, 063116.

Meinshausen, N., & Bühlmann, P. (2015). Maximin effects in inhomogeneous large-scale data. *Annals of Statistics*, *43*, 1801–1830. https://doi.org/10.1214/15-AOS1325

Paiva, H. M., Afonso, R. J. M., de Oliveira, I. L., & Garcia, G. F. (2020). A data-driven model to describe and forecast the dynamics of covid-19 transmission. *Plos One*, *15*, 1–16. https://doi.org/10.1371/journal.pone.0236386

Rajendran, S., & Jayagopal, P. (2021). Accessing Covid19 epidemic outbreak in Tamil Nadu and the impact of lockdown through epidemiological models and dynamic systems. *Measurement*, *169*, 108432. https://doi.org/10.1016/j.measurement.2020.108432

Rudy, S. H., Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Data-driven discovery of partial differential equations. *Science Advances*, *3*.

Rui, R., & Tian, M. (2021). Joint estimation of case fatality rate of COVID-19 and power of quarantine strategy performed in Wuhan, China. *Biometrical Journal*, *63*(1), 46–58. https://doi.org/10.1002/bimj.202000116

Siwiak, M., Szczesny, P., & Siwiak, M. (2020). From the index case to global spread: The global mobility based modelling of the Covid-19 pandemic implies higher infection rate and lower detection ratio than current estimates. *PeerJ*, *8*, e9548. https://doi.org/10.7717/peerj.9548

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *58*, 267–288.

Tran, G., & Ward, R. (2017). Exact recovery of chaotic systems from highly corrupted data. *Multiscale Modeling & Simulation*, *15*, 1108–1129.

Zhao, S., & Chen, H. (2020). Modeling the epidemic dynamics and control of Covid-19 outbreak in China. *Quantitative Biology* 1–9. Advanced online publication. https://doi.org/10.1007/s40484-020-0199-0

How to cite this article: Liang, J., Zhang, X., Wang, K., Tang, M., & Tian, M. (2022). Discovering dynamic models of COVID-19 transmission. *Transboundary and Emerging Diseases*, *69*, e64–e70. https://doi.org/10.1111/tbed.14263