ORIGINAL ARTICLE

Identifying epidemic spreading dynamics of COVID-19 by pseudocoevolutionary simulated annealing optimizers

Choujun Zhan[1] · Yufan Zheng[2] · Zhikang Lai[2] · Tianyong Hao[1] · Bing Li[3]

Abstract

At the end of 2019, a new coronavirus (COVID-19) epidemic has triggered global public health concern. Here, a model integrating the daily intercity migration network, which constructed from real-world migration records and the Susceptible–Exposed–Infected–Removed model, is utilized to predict the epidemic spreading of the COVID-19 in more than 300 cities in China. However, the model has more than 1800 unknown parameters, which is a challenging task to estimate all unknown parameters from historical data within a reasonable computation time. In this article, we proposed a pseudo-coevolutionary simulated annealing (SA) algorithm for identifying these unknown parameters. The large volume of unknown parameters of this model is optimized through three procedures co-adapted SA-based optimization processes, respectively. Our results confirm that the proposed method is both efficient and robust. Then, we use the identified model to predict the trends of the epidemic spreading of the COVID-19 in these cities. We find that the number of infections in most cities in China has reached their peak from February 29, 2020, to March 15, 2020. For most cities outside Hubei province, the total number of infected individuals would be less than 100, while for most cities in Hubei province (exclude Wuhan), the total number of infected individuals would be less than 3000.

1 Introduction

Infectious diseases have been raging in the world many times in history. For example, the Black Death (also known as the Pestilence) in the fourteenth century lasted for 30 years in Europe, with more than 25 million deaths, accounting for about 1/3 of the European population at that

✉ Bing Li
11539@whut.edu.cn

Choujun Zhan
20185076@m.scnu.edu.cn

1 School of Computer, South China Normal University, Guangzhou 510631, China

2 Nanfang College of Sun Yat-Sen University, Guangzhou 510970, China

3 School of Economics, Wuhan University of Technology, Wuhan 430070, China

time [2]. In 1918, Spanish influenza, which initially outbreaks within the U.S. military, eventually swept the world, infected nearly 600 million people, and caused about 40–50 million death [3]. In 2003, SARS, with a fatality rate of 11%, spread from Guangdong Province to the whole country, bringing huge losses to China's national economy [4]. In 2009, the H1N1 flu outbreak spread to 214 countries and regions, causing 1, 220 deaths in a few months [5]. In 2014, the outbreak of the Ebola epidemic resulted in 28,637 infections and 11,315 deaths. At the end of 2019, a highly contagious disease, which is caused by infection of the SARS-CoV-2 virus and named the 2019 Coronavirus Disease (COVID-19), broke out and caused millions of infections [6, 7]. The spreading trends of the COVID-19, including when the peaks would occur, how many people would eventually be infected, the final infection rate of the population of each city, and which cities would run the risk of being out of control, are the core questions that need to be answered.

Scholars from various disciplines have participated in the research on epidemic transmission and control. The epidemic spreading model can be traced back to the

analysis of smallpox by Daniel Bernoulli in 1760 [8], while the most classic epidemic spreading model is the Susceptible–Infected–Removed (SIR) model proposed by Kermack and McKendrich in 1927 [9]. Based on the SIR model, scholars found that there exists an epidemic threshold depended on the infection and recovery rate. If the infection rate is greater than a threshold, epidemics will spread on a large scale in the population. In 1932, Kermack and McKendrich established the Susceptible–Infected–Susceptible (SIS) model [10], which is similar to the SIR model, except that infected individuals would return to a susceptible state instead of an immune state after recovery. In 1992, for infectious diseases with a limited immune period, J. Mena-Lorca and et al. improved and proposed a more complex Susceptible–Infected–Removed–Susceptible (SIRS) model [11]. Inspired by these pioneering studies, a lot of efforts have been devoted to investigating epidemic spreading under various circumstances [12, 13]. One of the basic problems in theoretical epidemiology is to study the epidemic threshold of classic epidemic spreading models, e.g., SIR, SIS, and SEIR, in various networks, such as scale-free network and its extension [12, 13], complex heterogeneous networks [14], the real-world Oregon graph [15, 16], adaptive networks [17], and the complete worldwide air travel network [18]. Previous literature has unveiled that the epidemic threshold is highly related to the spectral radius of the adjacent matrix of the network [19, 20]. Other researches focused on developing epidemic spreading models in different scenarios and the temporal evolution [21, 22]. However, to our best knowledge, few of them focus on developing models to describe and predict the dynamic of real epidemic spreading cases.

The migration of individuals, especially intercity migration, plays a core role in the spread of SARS-CoV-2 [7, 18]. Wuhan is a metropolis with a population of more than 11 million and is also one of the transportation hubs in China. During the Spring Festival in 2019, 5 million people set off from Wuhan to other cities in China. Large-scale migration greatly enhanced the spread of COVID-19 from outbreak areas to other cities in China. Therefore, the intercity migration data is also an important indicator for describing and predicting the spread of the virus. Here, daily intercity migration data for 367 cities in China are collected and utilized to construct intercity migration networks. Further, a model established by combining complex network theory and the classic SEIR model can be used to describe how COVID-19 spreads from Wuhan to other cities in China. This dynamic model has more than 1800 unknown parameters to be determined from the historical data. The inference of model parameter values from rare historical time-

course data can be reformulated as an optimization problem and is still one of the most challenging tasks [23, 24]. Evolutionary algorithms have outstanding performance in solving nonlinear optimization problem [25, 26]. Hence, it is worth developing evolutionary algorithms, which should be robust against noise, efficient in computation, and flexible enough to meet different constraints for estimating these 1800 unknown parameters. In this article, a novel pseudocoevolutionary simulated annealing algorithm is proposed to solve this problem [27]. Results show that the proposed algorithm successfully identified optimal parameter sets of this epidemic spreading model. Also, the model can fit the number of infected, recovered individuals, and the death toll of each city with a minor error.

Based on the model, we find that migration control was extremely effective in controlling the spread of the epidemic. If the government continues strict migration control, the infections numbers of most cities in China would peak between mid-February to early March 2020. The peak number of infections in most cities is smaller than 100, while the proportion of infected individuals in each cities population is smaller than 0.01%. However, if the epidemic spreading is out of control, it would infect about 1% of the population in Hubei province, while infecting about 0.3% population outside Hubei province. The peak number of infections in most cities would come at the end of April 2020. Evidence shows that China has controlled the spreading of COVID-19.

The main contributions of this study are as follows:

- First, we integrate daily intercity migration data and traditional SEIR model to develop an extended SEIR model;
- A novel pseudocoevolutionary simulated annealing algorithm is proposed. Additionally, we compared the estimation result with simulated annealing, particle swarm optimization, and pattern search algorithms. Results show that the proposed algorithm provides the best results;
- The pandemic situation of China has been investigated. Results show that this technique can accurately reflect the spread of COVID-19. Study shows that migration control is extremely effective in controlling the spread of the epidemic.

The rest of this paper is organized as follows. Section 2 reviews related works. The description of data is introduced in Sect. refsec:datapsdescription, including official released confirmed cases, recovered cases, death toll, and intercity migration data. In Sect. 4, the SEIR-migration model and pseudocoevolutionary simulated annealing algorithm are introduced. Then, the experimental design

and results are shown in Sect. 5. Finally, conclusion and future work are presented in Sect. 6.

2 Literature review

Complex network theory is a powerful tool for researchers to study epidemic spreading. With the development of complex network theory, a large amount of work has investigated the effect of the structure of complex networks (such as degree of relevance, clustering coefficients, community structure, hierarchical structure, and edge weights) on the propagation properties of infectious diseases (such as spreading rate, scale of propagation, and epidemic threshold) [13, 28]. In 2001, Pastor-Satorras et al. studied the propagation model on scale-free networks and proved that scale-free networks are weak against infectious diseases and can maintain spreading at any small infection rate [12, 13]. In the same year, May and Lloyd investigate the effect of network scale on spreading behavior of scale-free networks and pointed out that there are positive propagation thresholds for limited scale-free networks [29]. In 2002, Newman investigated the SIR propagation model in a scale-free network and proved that there is an epidemic threshold when the cutoff value of the degree of nodes in the network is relatively small [30]. In 2007, Toroczkai et al. proposed the concept of dynamic proximity networks based on the premise of dynamic contact networks [31]. Ball et al. constructed a model that includes local adjacent and global accidental connections, and results showed that the degree distribution and accidental connections have a significant effect on the spreading of epidemic [32]. In 2018, Wang et al. proposed a dynamic epidemiological model based on complex routing in the form of multiple routes [33].

Epidemic threshold β_c plays an important role in epidemic spreading. For a large-scale system, when the infection rate $\beta > \beta_c$, the proportion of infected individuals will reach a limited proportion. Otherwise, if $\beta < \beta_c$, the proportion of infected people will reduce to almost zero. Therefore, to control the outbreak of epidemics, reducing the infection rate is one of the effective ways. Research shows that frequent absorption, wearing a mask, and disconnecting from the infectious individuals will reduce the probability of infection and then effectively control or slow the outbreak of an epidemic [34]. Additionally, studies have shown that quarantining, closing schools, and restricting individuals from attending public events can make people's contact networks sparse and reduce infection rate [35]. In 2011, Jin et al. developed an epidemiological model of influenza A, demonstrating that an immunization strategy targeting specific populations with given connectivity can greatly reduce epidemic spreading

[36]. Two years later, Guo et al. introduced a continuous-time adaptive susceptible–infectious–susceptible (ASIS) model, proving that the adaption of the topology can inhibit infection [37]. In the same year, Peng et al. investigate several epidemiological models, including susceptibility, infection, and incomplete vaccination segment models, on the Watts–Strogatz small world, Barabasi–Albert scale-free, and random scale-free networks, for analyzing the epidemic threshold and infection rate [38]. More information can be found in the review papers [39]. However, to our understanding, most of the literature about epidemiological models analyzing the propagation in a network is based on analytical methods and large-scale simulations, without the support of real-world data.

A common approach for explaining and analyzing real-world phenomena is to establish epidemiological models based on real-world observation data. These epidemiological models are always nonlinear dynamical models with high parameter dimension, which is often presented as a set of ordinary differential equations (ODEs) or discrete-time equations containing a large volume of unknown parameters [40]. The identification of a large volume of unknown parameters from historical observation data is critical for judging the performance of an epidemiological model. The method for identifying unknown parameters can be classified as "reverse engineering techniques," which usually formulate the problem of parameter identification into a nonlinear optimization problem that minimizes an objective function representing the fitness of the model with respect to the observation data [23, 24]. The identification results are highly dependent on the optimization algorithm. Due to the simplicity and ease of use, evolutionary algorithms are widely utilized to identify unknown parameters of nonlinear dynamic models [41–44]. However, for nonlinear dynamical systems with high parameter dimension, the objective function is always complicated and has tremendous local minima. Parameter estimation algorithms face a high possibility of converging at local optima but not global minima [45]. Additionally, one parameter identification trial always required tens of hours computation time [46, 47]. Therefore, the parameter estimation problem is still a challenging task and even a bottleneck for nonlinear dynamical models with high parameter dimension. Evolutionary algorithms have been extensively used in nonlinear optimization problems and shown that can provide satisfying results [25–27, 48]. In this article, a new pseudocoevolutionary algorithm is proposed to solve this hard engineering problem, and the more detail information about hard engineering can be found in [49–52].

3 DATA

3.1 Official data of COVID-19 cases

Testing is the only way to know whether a susceptible individual is infected or not. At present, there exist two kinds of tests (techniques) for testing COVID-19. One kind of tests (techniques) checks the presence of the COVID-19 virus, aiming to establish whether an individual is currently infected. The other kind of tests examines the presence of antibodies, which can figure out whether an individual has been infected in the past, even this individual has recovered and not carried COVID-19 virus now. A summary of the current state of testing technologies associated with their implementations can be found in [53]. Now, the most common way to perform a COVID-19 test is adopted by detecting the viral RNA through polymerase chain reaction (PCR). In this work, the official number of infected cases only contains individual who has a positive COVID-19 testing result. In China, Wuhan is received the first confirmed case of COVID-19 infection on December 8, 2019 [6]. Most other cities in China released data of COVID-19 infections around January 20, 2020. The data of COVID-19 infections, recovery, and death toll used in the study were derived from official data released by the National Health Commission of China. Hubei Province was the epicenter of the epidemic in China. Most of the infections occurred in Hubei Province (as shown in Fig.1), while the number of other provinces is relatively small. In this study, one of our aims is to develop an epidemiological model that accurately describes how the number of infections, recovery, and death toll change over time in various cities.

3.2 Intercity travel data

COVID-19 mainly spread through human-to-human transmission. In this case, the intercity migration of infected and exposed individuals has become the main driving force for COVID-19 to spread from one city to another. Chinese New Year (mid-January to early February 2020) is the most important holiday for Chinese people. In 2020, the period around during the Spring Festival holiday in 2019 is approximately from mid-January to early February 2020. Wuhan, as one of the most important transportation hubs in China and the world, is one of the cities with the largest flow of entry and exit around the Spring Festival. China's Ministry of Transport estimates that Wuhan has about 5 million trips, while China as a whole has about 3 billion trips during the Spring Festival holiday. We have collected daily intercity travel data for
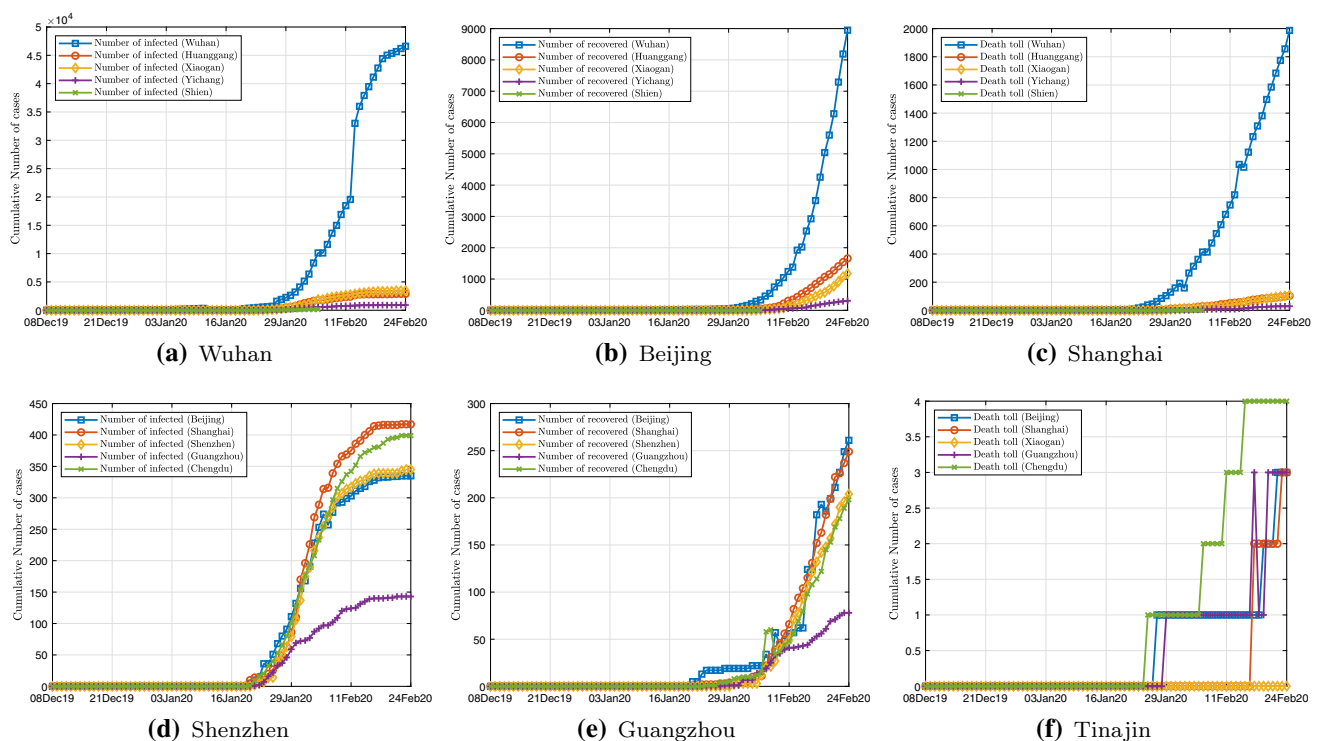

(a) Wuhan   **(b)** Beijing   **(c)** Shanghai

(d) Shenzhen   **(e)** Guangzhou   **(f)** Tinajin

Fig. 1 Daily data of COVID-19 infections, recovery, and death toll in 5 cities in Hubei province and 5 metropolis in China from December 8, 2019, to February 13, 2020. **a** Cumulative number of infections of 5 cities in Hubei; **b** cumulative number of recovery of 5 cities in Hubei; **c** cumulative number of death toll of 5 cities in Hubei; **d** cumulative number of infections of 5 metropolis; **e** cumulative number of recovery of 5 metropolis; **f** cumulative number of death toll of metropolis (color online)

367 cities in China. The data provide the intensity of population migration and also indicate the strength of the population in and out of various cities. Based on these data, we can develop the migration networks (shown in Fig. 2). After the outbreak of COVID-19, the Chinese government has rapidly restricted the intercity migration since January 23, so the strength of intercity migration has dramatically reduced since January 23. Figure 3 shows the total inflow/outflow of travelers of 6 metropolis in Chinese. Note that after migration control, the strength of the intercity migration of Wuhan has almost reduced to zero. The control measure effectively reduced the speed of virus transmission and ultimately successfully controlled the further spread of the virus.

Base on this data, we can construct the *migration matrix*, which is given as

$$M(t) = \begin{bmatrix} m_{11}(t) & m_{12}(t) & \cdots & m_{1K}(t) \\ m_{21}(t) & m_{22}(t) & \cdots & m_{2K}(t) \\ \vdots & \vdots & \ddots & \vdots \\ m_{N1}(t) & m_{N2}(t) & \cdots & m_{KK}(t) \end{bmatrix}, \quad (1)$$

where $K = 367$ is the number of the cities, and $m_{ij}(t)$ is the migrant volume from city i to city j at time t. Migration matrix M thus effectively describes the network of cities with human movement constituting the links of the network. Figure 3 plots the daily total inflow and outflow migration strengths of Wuhan, showing the abrupt decrease in migrant strength after the city shut down all inflow and outflow traffic from February 01, 2020.

4 The SEIR-migration model and pseudocoevolutionary simulated annealing algorithm

First, we will give a brief description of human contact networks with multiple sub-networks representing a city or administrative regions for epidemic spreading propagation. A real human contact network consists of multiple sub-networks, just as a country consisting of many cities, towns, and villages. Here, we consider a human contact network $G = (V, E)$ contains K sub-networks $\{G_1, G_2, \ldots, G_M\}$. V stands for the set of nodes, and E is the set of edges. Here, each node represents an individual. If two individuals/nodes v_i and v_j have contacts, there will be a link $e_{i,j}$ between them, otherwise, no connection (shown in Fig. 4a). Note that nodes in the same sub-network have plenty and strong connections with network neighbors, which results in a highly clustered sub-network. However, nodes belonging to different sub-networks have less and weak connections. In this work, a sub-network can be treated as a city, while nodes in a sub-network stand for citizens. The K sub-network (city) forms a huge contact network of a country (shown in Fig. 4b).

In the classic SIR model, each individual can be in three different states: infected (I), susceptible (S), and recovered (R). In an epidemic spreading case, infected individuals (I) can infect susceptible individuals (S) through human contact network, while an infected individual can be cured and turn into recovery state (R). Once an infectious disease starts to spread in a certain sub-network, due to the

Fig. 2 Intercity travel network of main cities in China on February 10, 2020. Node size represents the inflow volume, while arrows show direction. Color of lines indicates migration strength (color online)
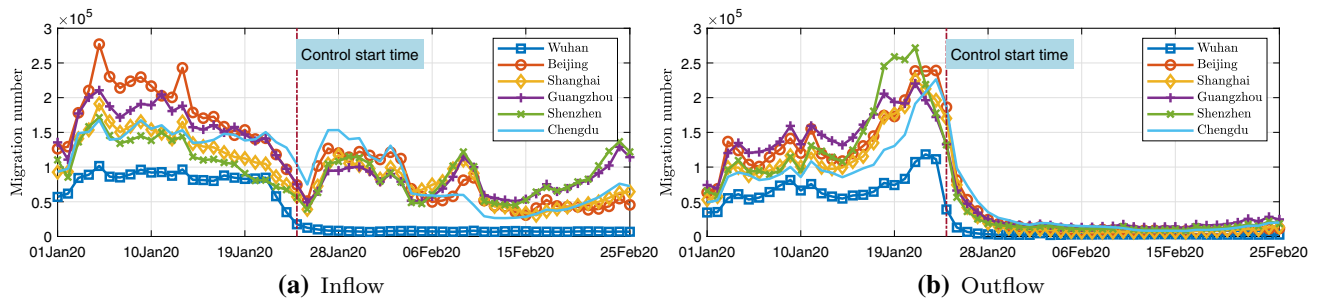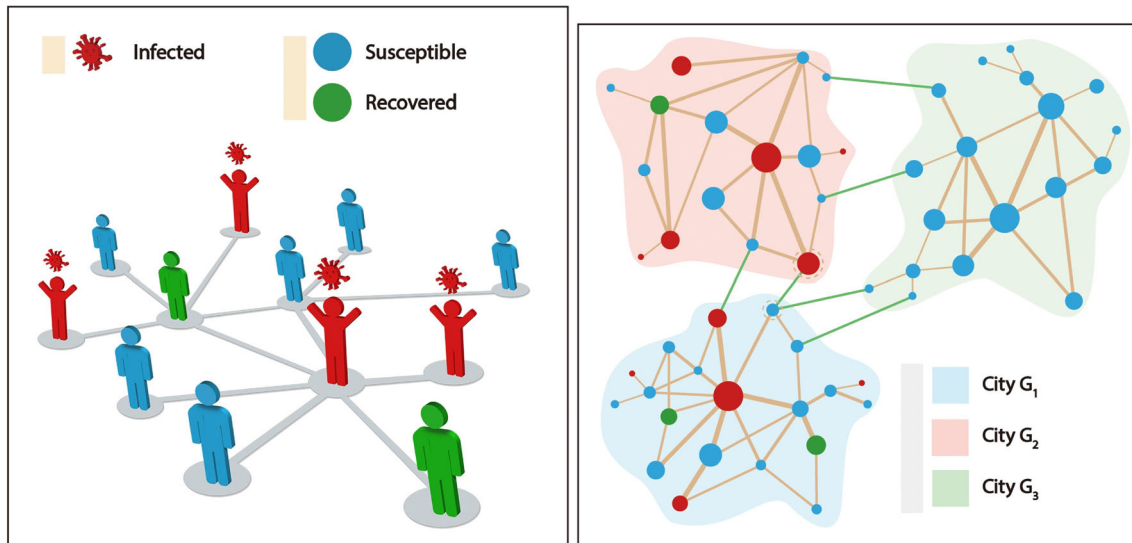
(a) Inflow   **(b)** Outflow
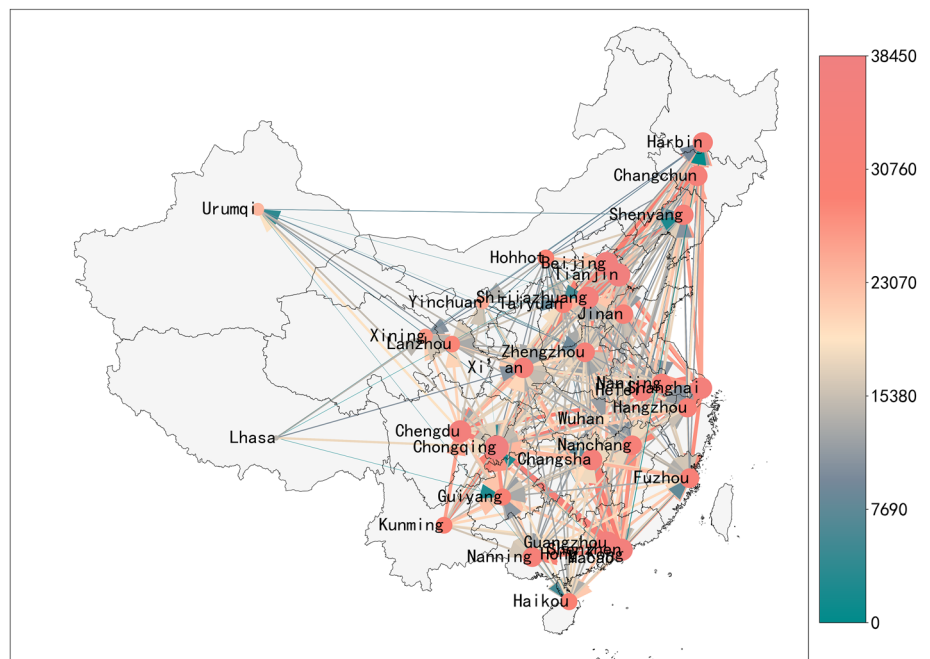
Fig. 3 Total inflow/outflow of travelers of 6 metropolis in Chinese. **a** Travelers to these 6 metropolis; **b** travelers from these 6 metropolis (color online)

(a) An illustrated human contact network with susceptible, infected, and recovered individuals.

(b) A human contact network with 3 sub-networks and an epidemic spreading from one sub-network to another.

Fig. 4 An illustrative example of epidemic spreading a human contact network including three sub-networks (cities) G_1, G_2 and G_3. **a** Virus spread from person to person through a human contact network. A susceptible individual may become an infection if he/she contacts with an infection. A red man with virus icon on the head represents an infection who can spread virus to susceptible neighbors (light blue

man), and the solid line between two individuals means they have closely contacted and virus can transmit from one person to the other. An infected individual can be cured and then become a recovered individual (light green mean); **b** a human contact network with three highly clustered communities (cities) of infected, susceptible, and recovered individuals. (Color online)

denseness of the sub-network nodes and the short distance between neighboring nodes, the epidemic will outbreak within the sub-network in a short time (G_1 in Fig. 4b). With the epidemic spread, plenty of susceptible nodes in the network transit into infected nodes. Some infected nodes have connections with nodes in other sub-network. Then, the virus can spread from one sub-network to another sub-network through the nodes connecting two sub-networks and eventually spread to the entire network (G_2 in Fig. 4b).

4.1 SEIR-Migration model

Studies reveal that the median incubation period of COVID-19 to be 5.6 days (95% CI 4.8-6.3) [54]. Exposed individuals can also infect other individuals during the incubation period. Each node in human contact networks may assume one of four possible states in the epidemic spreading process, namely susceptible (S), exposed (E), infected (I), and recovered/removed (R). For sub-network (city) j, the number of nodes in the four states $S_j(t)$, $E_j(t)$, $I_j(t)$, and $R_j(t)$, at time t.

Here, $m_{ij}(t)$ represents the volume of individuals moving from city i to city j at time t. Here, we assume the population of city j is P_j. Then, the number of infected individuals moving from city i to city j is

$$\Delta I_{ij}^{\text{in}}(t) = \frac{I_i(t)m_{ij}(t)}{P_i}. \tag{2}$$

Also, the volume of individual migrating from city j is $\sum_{i=1}^{N} m_{ji}$. Then, the number of infected individuals moving out of city j is

$$\Delta I_j^{\text{out}}(t) = \frac{I_j(t)\sum_{i=1}^{N} m_{ji}(t)}{P_j}, \tag{3}$$

Thus, the dynamic change of the number of infected cases in city j at time t is given by

$$\Delta I_j(t) = \kappa_j E_i(t) - \gamma_j I_j(t) \\ + k_I \left(\sum_{i=1}^{N} \left(\frac{I_i(t)m_{ij}(t)}{P_i(t)} \right) \\ - \frac{I_j(t)\sum_{i=1}^{N} m_{ji}(t)}{P_j(t)} \right), \tag{4}$$

where $\Delta I_j(t) = I_j(t+1) - I_j(t)$. Moreover, if city i has a population of P_i and the eventual percentage of infection is δ_i, then $N_i^s = \delta_i P_i$. Thus, we have

$$N_i^s(t) = S_i(t) + E_i(t) + I_i(t) + R_i(t), \tag{5}$$

where N_i^s is the eventual number of infections. Similarly, the dynamic changes of infected, susceptible, recovered individuals, and the population of a city can be obtained. Then, we have the modified SEIR model with consideration of human migration dynamics as follows:

$$\Delta I_j(t) = \kappa_j E_i(t) - \gamma_j I_j(t) \\ + k_I \left(\sum_{i=1}^{N} \left(\frac{I_i(t)m_{ij}(t)}{P_i(t)} \right) \\ - \frac{I_j(t)\sum_{i=1}^{N} m_{ji}(t)}{P_j(t)} \right),$$

$$\Delta E_j(t) = \frac{\beta_j}{N_j^s(t)} I_j(t)S_j(t) \\ + \frac{\alpha_j}{N_j^s(t)} E_j(t)S_j(t) - \kappa_j E_i(t) \\ + \sum_{i=1}^{N} \left(\frac{E_i(t)m_{ij}(t)}{P_i(t)} \right) \\ - \frac{E_j(t) * \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)},$$

$$\Delta S_j(t) = -\frac{\beta_j}{N_j^s(t)} I_j(t)S_j(t) \\ - \frac{\alpha_j}{N_j^s(t)} E_j(t)S_j(t) + \sum_{i=1}^{N} \left(\frac{S_i(t)m_{ij}(t)}{P_i(t)} \right) \\ - \frac{S_j(t)\sum_{i=1}^{N} m_{ji}(t)}{P_j(t)},$$

$$\Delta R_j(t) = \gamma_j I_j(t),$$

$$\Delta P_j(t) = \sum_{i=1}^{N} m_{ij}(t) - \sum_{i=1}^{N} m_{ji}(t),$$

$$\Delta N_j^s(t) = k_I \left(\sum_{i=1}^{N} \left(\frac{I_i(t)m_{ij}(t)}{P_i(t)} \right) \right. \tag{6}$$
$$\left. - \frac{I_j(t)\sum_{i=1}^{N} m_{ji}(t)}{P_j(t)} \right) + \sum_{i=1}^{N} \left(\frac{E_i(t)m_{ij}(t)}{P_i(t)} \right) \\ - \frac{E_j(t)\sum_{i=1}^{N} m_{ji}(t)}{P_j(t)} \\ + \sum_{i=1}^{N} \left(\frac{S_i(t)m_{ij}(t)}{P_i(t)} \right) \\ - \frac{S_j(t)\sum_{i=1}^{N} m_{ji}(t)}{P_j(t)},$$

where $\Delta E_j(t) = E_j(t+1) - E_j(t)$, $\Delta S_j(t) = S_j(t+1) - S_j(t)$, $\Delta R_j(t) = R_j(t+1) - R_j(t)$, $\Delta N_j^s(t) = N_j^s(t+1) - N_j^s(t)$, and $\Delta P_j(t) = P_j(t+1) - P_j(t)$. The physical meaning of each parameter of model (6) is presented in Table 1, while a detailed description of the model is given in [1]. Note that we assume the recovered individuals are assumed to stay in the city j.

4.2 Parameter identification problem

Model (6) has a large volume of unknown parameter. The parameter estimation problem can be transformed into a nonlinear optimization problem (NLP). The purpose of optimization is to find a set of suitable parameters to make the estimated growth trajectory that matches historical data. Here, we define $I_{j,0} = I_j(t_0)$ and $E_{j,0} = E_j(t_0)$, which are the initial number of infected and exposed individuals in city j, respectively. Note hat Wuhan is the epicenter of the COVID-19 pandemic in China, with Hubei province being the region immediately surrounding it. Therefore, it is reasonable to assume that Wuhan has initially infected individuals. Then, $I_{\text{WH}}(t_0), E_{WH}(t_0) \neq 0$, and $I_j(t_0), E_j(t_0) = 0$ for all other cities, where $I_{\text{WH}}(t_0)$ and $E_{WH}(t_0)$ represent the initial number of infected and exposed individuals in Wuhan, respectively. For city j, there exist a set of unknown parameters, i.e.,

$$\theta_j = \{\alpha_j, \beta_j, \gamma_j, \kappa_j, \delta_j\}. \tag{7}$$

Table 1 Parameter set of model (6)

β_j:	The rate at which the infected individuals infect the susceptible individuals in city j
α_j:	The rate at which the exposed individuals infect the susceptible individuals in city j
κ_j:	The rate at which exposed individuals become infected in city j
γ_j:	The recovery rate in city j
k_I:	The possibility of an infected individual moving from one city to another
δ_j:	The eventual percentage of infections in city j
$I_{j,0}$:	The initial number of infected individuals in city j
$E_{j,0}$:	The initial number of individuals in city j

Then, the unknown parameter set is $\Theta = \{I_{WH,0}, E_{WH,0}, k_I, \theta_1, \theta_2, \ldots, \theta_K\}$. Totally, there exist $5K + 3$ unknown parameters, where K is the number of cities. Thus, an enormous effort of computation is required to estimate the suitable parameters.

Let $X_j(t)$ be the extended state vector, i.e., $X_j(t) = [S_j^T(t)\ E_j^T(t)\ I_j^T(t)\ R_j^T(t)\ P_j^T(t)\ (N_j^s(t))^T]^T$, then, we define:

$$X(t) = [X_1^T(t)\ X_2^T(t)\ \cdots\ X_K^T(t)]^T \qquad (8)$$

Model (6) can be reformulated as

$$\Delta X_j(t_i) = f(X_j(t_i), X(t_i)|\Theta), \qquad (9)$$

where $f(x)$ is the right side of (6), and Θ is the set of unknown parameters. Note that $\Delta X_j(t) = X(t_i + 1) - X(t_i)$, then, Eq. (9) can be reformulated as:

$$X_j(t_i + 1) = X_j(t_i) + f(X_j(t_i), X(t)|\Theta). \qquad (10)$$

Finally, the parameter estimation problem can be formulated as the following constrained nonlinear optimization problem:

$$P_0: \min_{\Theta} \sum_{j=1}^{K} \sum_{i=1}^{N} w_{ij}(I(t_i) - \hat{I}(t_i|\Theta))^2$$

$$s.t. \begin{cases} (i) & X_j(t_i + 1) = X_j(t_i) + f(X_j(t_i), X(t_i)|\Theta). \\ (ii) & \Theta_U \geq \Theta \geq \Theta_L, \\ (iii) & j = 1, 2, \ldots, K, \end{cases}$$

$$(11)$$

where $\hat{I}(t_i|\Theta)$ represents the estimated number of infected individuals at time t_i with parameter set Θ and initial condition $\{I_{WH,0}, E_{WH,0}\}$. w_{ij} stands for the weighted coefficient. The unknown parameter set is bounded between Θ_L and Θ_U. In this work, an inverse approach is taken to find the unknown parameters and states by solving (11).

4.3 Proposed pseudocoevolutionary simulated annealing algorithm

---

Algorithm 1 Sub-algorithm for calculating time series $I_j(t_i)$

Input: Parameter set and initial number of infected and exposed individuals of each city $\Theta = \{I_{WH,0}, E_{WH,0}, k_I, \theta_1, \theta_2, \cdots, \theta_K\}$, where $\theta_j = \{\alpha_j, \beta_j, \gamma_j, \kappa_j, \delta_j\}$ and $j = 1, 2, \cdots K$;

Output: A time series $\{I(t_0), I(t_1), \cdots, I(t_N)\}$;

    Initialization :

1: Set $I_{WH})(t_0) = I_{WH,0}$ and $E_{WH})(t_0) = E_{WH,0}$, otherwise $I_j)(t_0) = 0$ and $E_j(t_0) = 0$; $R_j(t_0) = 0$; $N_j^s(t_0) = \delta_j P_j(t_0)$, $S_j(t_0) = N_j^((t_0) - I_j(t_0) - E_j(t_0)$; $X(t_0) = [X_1^T(t_0), X_2^T(t_0), \cdots, X_K^T(t_0)]^T$ and $X_j(t_0) = [S_j^T(t_0), E_j^T(t_0), I_j^T(t+0), R_j^T(t_0, P_j^T(t_0), (N_j^s(t_0))^T]^T$.

    LOOP Process

2: **for** $i = 0$ to $N - 1$ **do**

3:     Calculate $X(t_i + 1)$ from model (6), namely,

$$X_j(t_i + 1) = X_j(t_i) + f(X_j(t_i), X(t)|\Theta).$$

    Since $I_j(t_i) \in X_j(t_i)$ and $X_j(t_i) \in X(t_i)$, we can derive $I_j(t_i)$ from $X(t_i)$.

4: **end for**

5: **return** $\{I(t_0), I(t_1), \cdots, I(t_N)\}$

Algorithm 2 Sub-algorithm for estimating the optimal parameter set $\bar{\Theta}^*$

Input: The initial parameters set of all the K cities Θ_0 and the set of index ($\Phi = \{a_1, a_2, \cdots, a_n\}$, $1 \leq a_j \leq K$) of the city, of which the parameters need to be optimized.

Output: Optimal parameter set $\bar{\Theta}^*$;

    Initialisation :

1: Initialize temperature T and set

$$\Theta_{current} = \Theta_0$$

    Compute the value of objective function from Algorithm-1 and objective function (11) with parameter set $\Theta_{current}$

$$cost_{current} = \sum_{j=1}^{K} \sum_{i=1}^{N} w_{ij} (I(t_i) - \hat{I}(t_i \,|\, \bar{\Theta}_{current}))^2,$$

    Initialize temperature T and random starting point

$$\bar{\Theta}_0 = \Theta_L + k_{rand} * (\Theta_U - \Theta_L),$$

    where k_{rand} is a randomly generated real number between 0 and 1.

    LOOP Process

2: **for** $i_{iter} = 0$ to i_{max} **do**

3:

$$i_{iter} = i_{iter} + 1,$$
$$temp_{iter} = 0,$$
$$\Theta_{previous} = \Theta_{current}$$
$$cost_{previous} = cost_{current}$$

4:     **while** $temp_{iter} \leq n_{rep}$ **do**

5:

$$temp_{iter} = temp_{iter} + 1.$$

6:       **if** $j \in \Phi$ **then**

7:         Select a new set of parameters (θ_j) from the neighborhood,

$$\theta_j \leftarrow \theta_j$$

8:       **else**

9:         Keep $\theta_j = \theta_{j,0}$.

10:       **end if**

        Compute the value of objective function (11) and derive δ

$$\delta = cost_{current} - cost_{previous}.$$

11:       **if** $\delta < 0$ **then**

12:         Accept new parameter set.

13:       **else**

14:         Accept new parameter set with probability $\exp(-\delta/T)$.

15:       **end if**

16:     **end while**

17:

$$T = \alpha * T, \quad (0 < \alpha < 1).$$

18: **end for**

19: **return** $\bar{\Theta}^*$

10

Algorithm 3 Pesudocoevolutionary algorithm for estimating the optimal parameter set Θ^*

Input: The set of unknown parameters and initial number of infected and exposed individuals of each city
$\Theta = \{I_{WH,0}, E_{WH,0}, k_I, \theta_1, \theta_2, \cdots, \theta_K\}$, where $\theta_j = \{\alpha_j, \beta_j, \gamma_j, \kappa_j, \delta_j\}$ and $j = 1, 2, \cdots K$;

Output: Optimal parameter set Θ^*;

Initialisation : Initialize temperature T, and random starting point

$$\Theta_0 = \Theta_L + k_{rand} * (\Theta_U - \Theta_L),$$

The index of adopted cities is $\Phi = 1, 2, \cdots, K$.

1: Apply Algorithm-2 to optimize parameter set to achieve $\bar{\Theta}^*$.

$$\Theta_0 \leftarrow \bar{\Theta}^*$$

LOOP Process

2: **for** $i = 0$ to M **do**
3:     **for** $j = 1$ to K **do**
4:         Set the model parameter as Θ_0 and apply Algorithm-1 to derive $\{\hat{I}(t_0 | \Theta_0), \hat{I}(t_1 | \Theta_0), \cdots, \hat{I}(t_N | \Theta_0)\}$.
5:         Using evaluation criteria (13) to derive RMSPE$_j$ for each city.
6:         Find the index of the $M1$ largest RMSPE$_j$ and set $\Phi = \{a_1, a_2, \cdots, a_{M1}\}$, where a_i represents the index of the i-th largest RMSPE$_j$.
7:         Apply Algorithm-2 to optimize parameter set to achieve $\bar{\Theta}^*$.

$$\Theta_0 \leftarrow \bar{\Theta}^*$$

8:         Randomly generate $M2$ unique integers a_i between 1 and K and set $\Phi = \{a_1, a_2, \cdots, a_{M2}\}$.
9:         Apply Algorithm-2 to optimize parameter set to achieve $\bar{\Theta}^*$.

$$\Theta_0 \leftarrow \bar{\Theta}^*$$

10:     **end for**
11: **end for**

$$\Theta^* \leftarrow \Theta_0$$

12: **return** Optimal parameter set Θ^*.

Note that the cumulative number of infections varies widely by cities. For example, Wuhan, the city with the largest number of infected people in China, is infected by more than 50,000 people. However, in some small cities, only a dozen people are infected, or even no one has been infected yet. Therefore, the weighted coefficient of the objective function 11 should be carefully adopted. In this work, w_{ij} is defined as follows:

$$w_{ij} = \frac{a}{\sum_i^N (I_j(t_i))}. \tag{12}$$

If the city is Wuhan, we have $a = 5000$; otherwise, $a = 1$. In our model, each city has a set of unknown variables $\theta_j = \{\alpha_j, \beta_j, \gamma_j, \kappa_j, \delta_j\}$, which controls the size of infected population, spreading rate, and the death rate.

Note that this model has $5K + 3$ parameters, namely the proposed model has a high-dimension unknown parameter set, which should be optimized. The search space for these optimization problems may be highly nonlinear and contain many local minima. Evolutionary algorithms have been extensively used in nonlinear optimization [25, 26, 48]. In this article, we proposed a new pseudo-coevolutionary algorithm to solve this inverse engineering problem. This main procedure tunes all the parameters, while the other two processes tune part of the parameters. The parameter estimation problem is separated into three co-adapted SA-based optimization processes. This main procedure tunes all the parameters, while the other two processes tune part of the parameters.
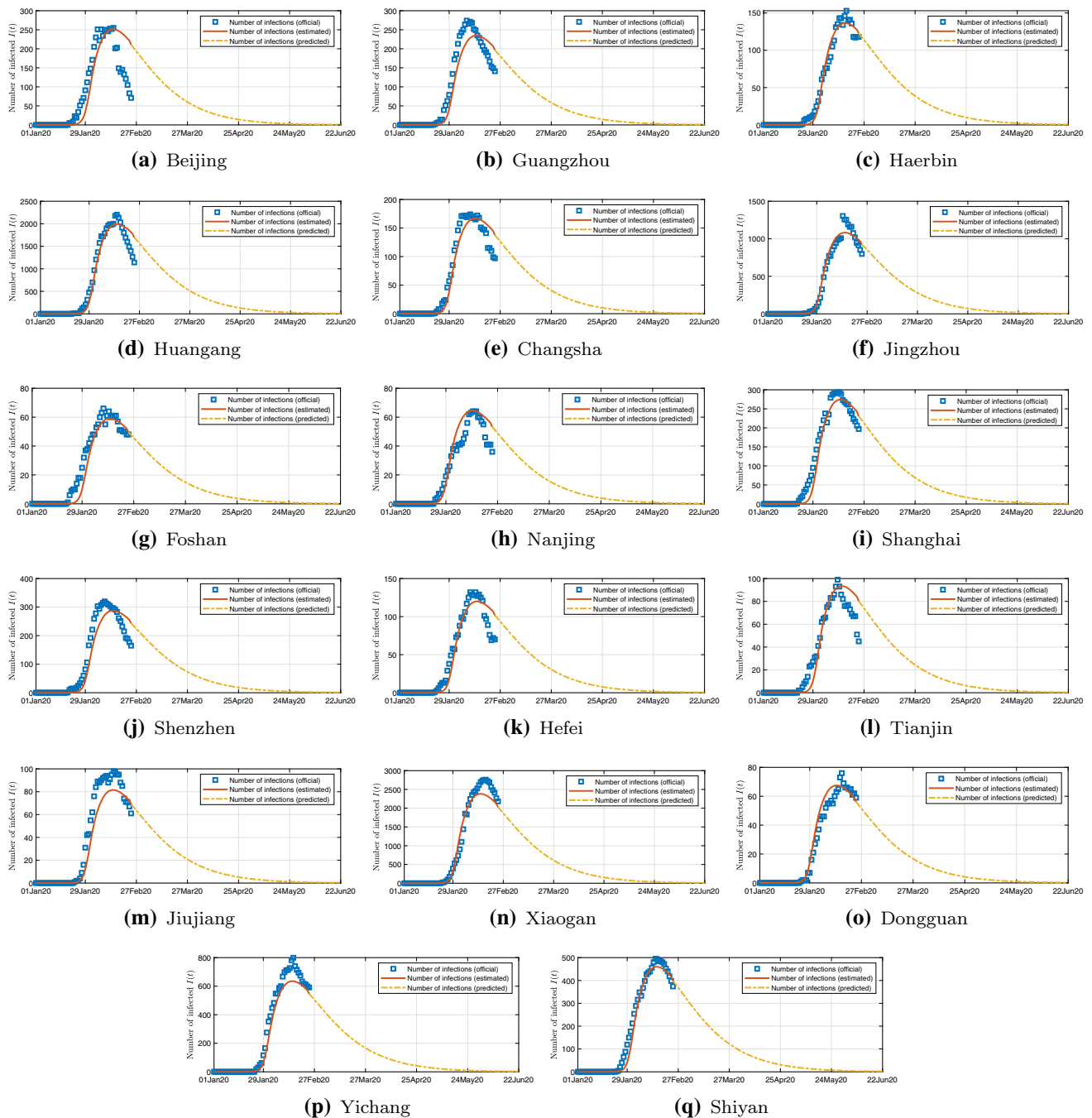
Fig. 5 Estimated historical data and prediction of the number of infected individuals in 17 selected cities in China for the next 150 days

1. In the main procedure, we tune the all the $5K + 3$ parameters. Then, we adopt root mean square percentage error (RMSPE) as follows:

$$\text{RMSPE}_j = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\frac{\hat{I}_j(t_i | \Theta) - I_j(t_i)}{I_j(t_i)} \right)^2}, \quad (13)$$

Here, RMSPE_j is utilized as the criterion to measure the difference between the real daily infection data and

the estimated infected individuals generated by this extended SEIR-migration model with an optimal parameter set Θ;

2. In this process, we find the index of the M_1 largest RMSPE_j and only tune the parameter sets of the corresponding cities.

3. In order to avoid that the parameters of some cities have not been adopted and adjusted individually during
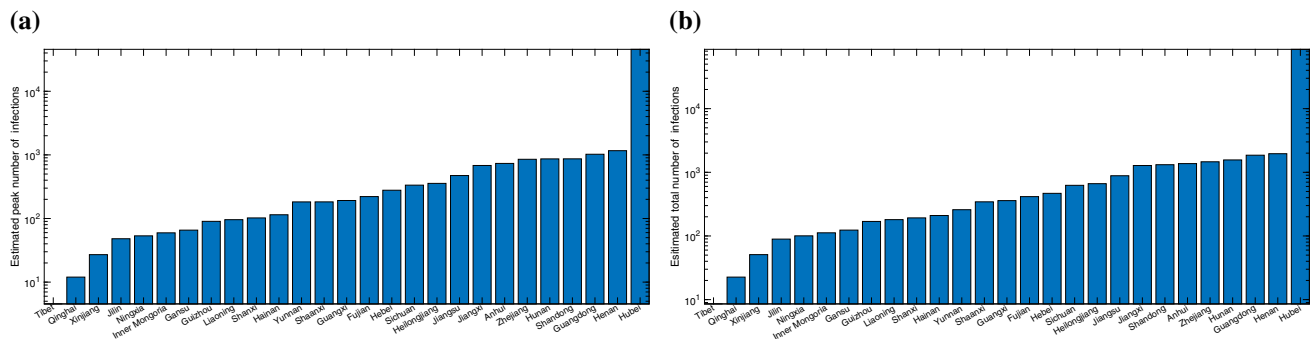
(a)

(b)

Fig. 6 **a** Peak number of infections in each province; **b** estimated total number of infected individuals eventually infected in a province

the whole identification process, we randomly select M_2 cities and adjust their parameters.

The whole optimization procedure is summarized in Algorithm-3, while Algorithm-2 is utilized for searching optimal parameters in subspace in each step.

5 Experimental results

The National Health Committee of China has published data on the spread of the COVID-19 epidemic from January 20, 2020. We use these historical data for parameter estimation of the SEIR-migration model (6). The pseudo-coevolutionary simulated annealing algorithm, as described in Sect. 4.3, is adopted to find the optimal parameter set of this model. Since the parameters of the model are all estimated from historical data through reverse engineering techniques, the accuracy and integrity of the data are essential. In the early stage of COVID-19, people do not know much about this virus, and the diagnostic techniques are limited. Therefore, during the early stage of the outbreak of COVID-19, the historical data of Wuhan City are possible to deviate from the true value in a wide range. Hence, we reduce the weighting coefficients corresponding to Wuhan data in the objective function. In addition, after the outbreak of the epidemic, Chinese government has adopted effective quarantine measures and promoted epidemic prevention knowledge. These measures can effectively reduce the infection rate in each city. Therefore, the parameters in the model should be time-varying. Nevertheless, to simplify the calculation, we assume that these parameters are constant over the whole process. We also applied traditional simulated annealing, particle swarm optimization, genetic algorithm, and pattern search method to estimate the parameters of the model. However, these methods cannot provide a satisfied result or cannot even converge in an acceptable computation time (such as one day), while the proposed method can converge to the global optima in two hours.

This model can estimate the daily number of infected, exposed, and recovered individuals in all 367 cities. Due to space limitations, we only show the results of 17 cities in Fig. 5. Assuming that the migration control measures, infection rate, and recovery rate will remain unchanged for a period of time in the future, this model can provide a prediction of the amount of actively infected individuals in each city, as shown in 5. Results clearly show that the number of actively infected individuals has reached or peaked in most cities in China from February 20 to March 15, 2020. The prediction results show that there will be few new confirmed cases from early March, and the number of actively infected individuals will gradually decrease. The pseudocoevolutionary simulated annealing algorithm successfully finds the optimal parameter set: $\alpha = 0.5631 \pm 0.0161$, $\beta = 0.2742 \pm 0.0178$, $\gamma = 0.0509 \pm 0.0007$, $\kappa = 0.1310 \pm 0.0035$. Figure 6a shows the estimated peak number of infected individual in each province, while Fig. 6b reveals the estimated total number of infected individuals in each province. Results show that the number of infections in most provinces is smaller than 1000. It is reasonable to assume that our knowledge of the COVID-19 will gradually increase. Then, the medical treatment will improve and the recovery rate will increase each day. We assume the recovery rate will increase 0.0005 each day, namely every 20 days, the number of daily recovered individuals increases by 1% of the total number of infected individuals. Then, most cities in China will have almost zero infections before July 2020. Therefore, we can claim that China has already controlled the spreading of COVID-19 in China.

6 Conclusion

The novel Coronavirus Disease 2019 (COVID-19) epidemic has caused 75,204 confirmed cases and 2,006 death utile February 19, 2020, which triggers global public health concern. Peak prediction informing social and non-

pharmaceutical prevention interventions is illuminating but remains difficult to achieve accuracy. Intercity migration is one of the essential factors in the spread of the disease. We construct migration networks of 367 cities in China. An SEIR-migration model with more than 1800 unknown parameters is utilized to model the spreading of COVID-19 in the 367 cities in China. We proposed a pseudocoevolutionary simulated annealing algorithm to identify these unknown parameters from historical data of the number of infected, recovered, and death toll. From this model, we can achieve all the essential information about the epidemic spreading, including infection rates, recovery rates, and eventual percentage of the infected population for 367 cities in China. The main conclusion of our study is that the COVID-19 epidemic spreading would peak between mid-February to early March 2020, with about 0.8%, less than 0.1%, and less than 0.01% of the population eventually infected in Wuhan, Hubei Province, and the rest of China, respectively. Results indicate that the COVID-19 epidemic has been controlled. This work provides a method for estimating the proportion of infected people. However, only seroprevalence studies may actually estimate the proportion of infected individuals.

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

Ethical approval All authors have known that the manuscript is submitted to your journal. And all of us confirm that the content of the manuscript has not been published elsewhere.

References

1. Zhan C, Chi KT, Fu Y, Lai Z, Zhang H Modeling and prediction of the 2019 coronavirus disease spreading in china incorporating human migration data, medRxiv
2. List of epidemics (2020). https://en.wikipedia.org/wiki/List_of_epidemics
3. Mackowiak PA (2007) Post mortem: solving history's great medical mysteries. ACP Press, Maryland
4. The world health report 2003 (2003). https://www.who.int/whr/2003/en/
5. Ding H, Santibanez TA, Jamieson DJ, Weinbaum CM, Euler GL, Grohskopf LA, Lu P-J, Singleton JA (2011) Influenza vaccination coverage among pregnant women-national 2009 h1n1 flu survey (nhfs). Am J Obstetr Gynecol 204(6):S96–S106
6. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KS, Lau EH, Wong JY et al (2020) Early transmission dynamics in wuhan, china, of novel coronavirus-infected pneumonia. N Engl J Med 382:1199–1207. https://doi.org/10.1056/NEJMoa2001316
7. Wu JT, Leung K, Leung GM (2020) Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. The Lancet 395(10225):689–697
8. Dietz K, Heesterbeek JAP (2002) Daniel Bernoulli's epidemiological model revisited. Math Biosci 180(1–2):1–21
9. Kermack WO, McKendrick AG (1991) Contributions to the mathematical theory of epidemics—I. Bullet Math Biol 53(1–2):33–55
10. Kermack WO, McKendrick AG (1932) Contributions to the mathematical theory of epidemics. ii.-the problem of endemicity. In: Proceedings of the Royal Society of London. Series A, containing papers of a mathematical and physical character 138(834) pp 55–83
11. Mena-Lorcat J, Hethcote HW (1992) Dynamic models of infectious diseases as regulators of population sizes. J Math Biol 30(7):693–716
12. Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. Phys Rev Lett 86(14):3200
13. Pastor-Satorras R, Vespignani A (2001) Epidemic dynamics and endemic states in complex networks. Phys Rev E 63(6):066117
14. Moreno Y, Nekovee M, Pacheco AF (2004) Dynamics of rumor spreading in complex networks. Phys Rev E 69(6):066130
15. Wang Y, Chakrabarti D, Wang C, Faloutsos C (2003) Epidemic spreading in real networks: an eigenvalue viewpoint. In: Proceedings of International Symposium on Reliable Distributed Systems, IEEE, pp 25–34
16. Chakrabarti D, Wang Y, Wang C, Leskovec J, Faloutsos C (2008) Epidemic thresholds in real networks. ACM Trans Inf Syst Secur (TISSEC) 10(4):1
17. Gross T, D'Lima CJD, Blasius B (2006) Epidemic dynamics on an adaptive network. Phys Rev Lett 96(20):208701
18. Colizza V, Barrat A, Barthélemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. Proc Natl Acad Sci USA 103(7):2015–2020
19. Valdano E, Ferreri L, Poletto C, Colizza V (2015) Analytical computation of the epidemic threshold on temporal networks. Phys Rev X 5(2):021005
20. Prakash BA, Chakrabarti D, Valler NC, Faloutsos M, Faloutsos C (2012) Threshold conditions for arbitrary cascade models on arbitrary networks. Knowl Inf Syst 33(3):549–575
21. Newman ME (2002) Spread of epidemic disease on networks. Phys Rev E 66(1):016128
22. Sanz J, Xia C-Y, Meloni S, Moreno Y (2014) Dynamics of interacting diseases. Phys Rev X 4(4):041005
23. Zhan C, Yeung LF (2011) Parameter estimation in systems biology models using spline approximation. BMC Syst Biol 5(1):14
24. Zhan C, Situ W, Yeung LF, Tsang PW-M, Yang G (2014) A parameter estimation method for biological systems modelled by ode/dde models using spline approximation and differential evolution algorithm. IEEE/ACM Trans Comput Biol Bioinf 11(6):1066–1076
25. Liu X-F, Zhan Z-H, Gao Y, Zhang J, Kwong S, Zhang J (2018) Coevolutionary particle swarm optimization with bottleneck objective learning strategy for many-objective optimization. IEEE Trans Evolut Comput 23(4):587–602
26. Yang Q, Chen W-N, Da Deng J, Li Y, Gu T, Zhang J (2017) A level-based learning swarm optimizer for large-scale optimization. IEEE Trans Evolut Comput 22(4):578–594

27. Zhang J, Chung HS-H, Lo W-L (2006) Pseudocoevolutionary genetic algorithms for power electronic circuits optimization. IEEE Trans Syst Man Cybern Part C Appl Rev 36(4):590–598

28. Lü L, Chen D-B, Zhou T (2011) The small world yields the most effective information spreading. New J Phys 13(12):123005

29. May RM, Lloyd AL (2001) Infection dynamics on scale-free networks. Phys Rev E 64(6):066112

30. Moore C, Newman ME (2000) Epidemics and percolation in small-world networks. Phys Rev E 61(5):5678

31. Toroczkai Z, Guclu H (2007) Proximity networks and epidemics. Phys A Statist Mech Appl 378(1):68–75

32. Ball F, Neal P (2008) Network epidemic models with two levels of mixing. Math Biosci 212(1):69–87

33. Wang Y, Cao J, Li X, Alsaedi A (2018) Edge-based epidemic dynamics with multiple routes of transmission on random networks. Nonlinear Dyn 91(1):403–420

34. Fenichel EP, Castillo-Chavez C, Ceddia MG, Chowell G, Parra PAG, Hickling GJ, Holloway G, Horan R, Morin B, Perrings C et al (2011) Adaptive human behavior in epidemiological models. Proc Natl Acad Sci 108(15):6306–6311

35. Barabási A-L et al (2016) Network science. Cambridge University Press, Cambridge

36. Jin Z, Zhang J, Song L-P, Sun G-Q, Kan J, Zhu H (2011) Modelling and analysis of influenza a (h1n1) on networks. BMC Public Health 11(S1):S9

37. Guo D, Trajanovski S, van de Bovenkamp R, Wang H, Van Mieghem P (2013) Epidemic threshold and topological structure of susceptible-infectious-susceptible epidemics in adaptive networks. Phys Rev E 88(4):042802

38. Peng X-L, Xu X-J, Fu X, Zhou T (2013) Vaccination intervention on epidemic dynamics in networks. Phys Rev E 87(2):022813

39. Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A (2015) Epidemic processes in complex networks. Rev Mod Phys 87(3):925

40. Frauenthal JC (2012) Mathematical modeling in epidemiology. Springer, Berlin

41. Huang Y, Gao L, Yi Z, Tai K, Kalita P, Prapainainar P, Garg A (2018) An application of evolutionary system identification algorithm in modelling of energy production system. Measurement 114:122–131

42. Cuevas E, Gálvez J, Avalos O (2017) Parameter estimation for chaotic fractional systems by using the locust search algorithm. Computación y Sistemas 21(2):369–380

43. Samsuri SFM, Ahmad R, Zakaria MZ, Zain MZM (2019) Parameter tuning for comparing multi-objective evolutionary algorithms applied to system identification problems. In: 2019 IEEE International conference on smart instrumentation, measurement and application (ICSIMA), IEEE, pp 1–6

44. Fan Q, Zhang Y (2016) Self-adaptive differential evolution algorithm with crossover strategies adaptation and its application in parameter estimation. Chem Intell Lab Syst 151:164–171

45. Tsai K-Y, Wang F-S (2005) Evolutionary optimization with data collocation for reverse engineering of biological networks. Bioinformatics 21(7):1180–1188

46. Kimura S, Ide K, Kashihara A, Kano M, Hatakeyama M, Masui R, Nakagawa N, Yokoyama S, Kuramitsu S, Konagaya A (2005) Inference of s-system models of genetic networks using a cooperative coevolutionary algorithm. Bioinformatics 21(7):1154–1163

47. Koh G, Teong HFC, Clément M-V, Hsu D, Thiagarajan P (2006) A decompositional approach to parameter estimation in pathway modeling: a case study of the akt and mapk pathways and their crosstalk. Bioinformatics 22(14):e271–e280

48. Shi W, Chen W-N, Lin Y, Gu T, Kwong S, Zhang J (2017) An adaptive estimation of distribution algorithm for multipolicy insurance investment planning. IEEE Trans Evolut Comput 23(1):1–14

49. Li Q, Wu Z, Zhang H (2020) Spatio-temporal modeling with enhanced flexibility and robustness of solar irradiance prediction: a chain-structure echo state network approach. J Clean Product 261:121151. https://doi.org/10.1016/j.jclepro.2020.121151

50. Arora S, Singh S (2017) Node localization in wireless sensor networks using butterfly optimization algorithm. Arab J Sci Eng 42(8):3325–3335

51. Wu Z, Li Q, Wu W, Zhao M (2019) Crowdsourcing model for energy efficiency retrofit and mixed-integer equilibrium analysis. IEEE Trans Ind Inf 16(7):4512–4524

52. Dai P, Liu K, Feng L, Zhang H, Lee VCS, Son SH, Wu X (2018) Temporal information services in large-scale vehicular networks through evolutionary multi-objective optimization. IEEE Trans Intell Transp Syst 20(1):218–231

53. Zhan C, Chi KT, Fu Y, Lai Z, Zhang H (2020) Humanity tested. Nat Biomed Eng 4:355–356

54. Backer JA, Klinkenberg D, Wallinga J (2020) Incubation period of 2019 novel coronavirus (2019-ncov) infections among travellers from wuhan, china, 20–28 January 2020. Eurosurveillance 25(5):2000062

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