medRxiv preprint doi: https://doi.org/10.1101/2020.02.18.20024570; this version posted February 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Modeling and Prediction of the 2019 Coronavirus Disease Spreading in China Incorporating Human Migration Data

Choujun Zhan[a], Chi K. Tse[b,*], Yuxia Fu[c], Zhikang Lai[c], Haijun Zhang[d]

[a]*School of Computing, South China Normal University, Guangzhou, China*
[b]*Department of Electrical Engineering, City University of Hong Kong, Hong Kong.*
[c]*School of Computer Science and Engineering, Nanfang College of Sun Yat-Sen University, China*
[d]*Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China*

Abstract

This study integrates the daily intercity migration data with the classic Susceptible-Exposed-Infected-Removed (SEIR) model to construct a new model suitable for describing the dynamics of epidemic spreading of Coronavirus Disease 2019 (COVID-19) in China. Daily intercity migration data for 367 cities in China are collected from Baidu Migration, a mobile-app based human migration tracking data system. Historical data of infected, recovered and death cases from official source are used for model fitting. The set of model parameters obtained from best data fitting using a constrained nonlinear optimization procedure is used for estimation of the dynamics of epidemic spreading in the coming weeks. Our results show that the number of infections in most cities in China will 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.

Keywords: Epidemic spreading, COVID-19, new coronavirus, human migration, intercity travel.

1. Introduction

The Coronavirus Disease 2019 (COVID-19) (known earlier as New Coronavirus Infected Pneumonia) began to spread since December 2019 from Wuhan, which has been widely regarded as the epicenter of the epidemic, to almost all provinces throughout China and 28 other countries. Human-to-human transmission had been found to occur in some early Wuhan cases in mid December [1], and the high volume and frequency of movement of people from Wuhan to other cities and between cities is thus an obvious cause for the wide and rapid spread of the disease throughout the country. The Susceptible-Exposed-Infected-Removed (SEIR) model has traditionally been used to study epidemic spreading with various forms of networks of transmission which define the contact topology [2], such as scalefree networks [3, 4, 5], small-world networks [6, 7], Oregon graph [8, 9], and adaptive networks [10]. Moreover, in most studies, the contact process assumes that the contagion expands at a certain rate from an infected individual to his/her neighbor, and that the spreading process takes place in a single population (network). The COVID-19 outbreak, however, began to occur and escalate in a special holiday period in China (about 20 days surrounding the Lunar New Year), during which a huge volume of intercity travel took place, resulting in outbreaks in multiple regions connected by an active transportation network. Thus, in order to understand the COVID-19 spreading process in China, it is essential to examine the human migration dynamics, especially between the epicenter Wuhan and other Chinese cities. A recent study has also revealed the risk of transmission of the virus from Wuhan to other cities [11].

In this paper, we utilize the human migration data collected from Baidu Migration [12], which provides historical indicative daily volume of travellers to/from and between 367 cities in China. To demonstrate the impact of intercity traffic on the COVID-19 epidemic spreading, we plot in Figure 1 the number of infected individuals in different cities versus the inflow traffic volume from Wuhan, which clearly shows that for cities farther away from Wuhan,

[*] Corresponding author
Email address: chitse@cityu.edu.hk (Chi K. Tse)

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.02.18.20024570; this version posted February 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

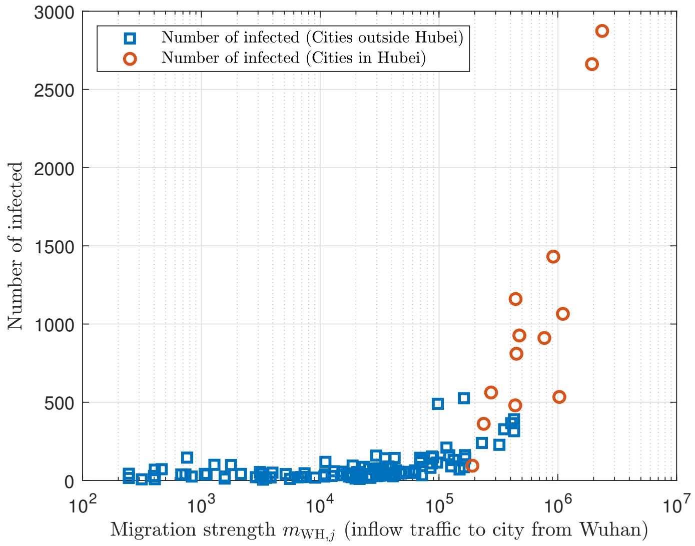


Figure 1: Number of infected individuals in various cities on 13/Feb/2020 versus the city's inflow traffic from Wuhan. Inflow traffic of each city from Wuhan is quantified by migration strength from Wuhan extracted from Baidu Migration data.

the number of infected individuals almost increases linearly with the inflow traffic from Wuhan. In view of the importance of human migration dynamics to the disease spreading process, we combine, in this study, intercity travel data collected from Baidu Migration [12] with the traditional SEIR model [2] to build a new dynamic model for the spreading of COVID-19 in China. Using official historical data of infected, recovered and death cases in 367 cities, we perform fitting of the data to estimate the best set of model parameters, which are then used to estimate the number of individuals exposed to the virus in each city and to predict the extent of spreading in the coming months. Our study shows that the number of infected cases in various Chinese cities will 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.

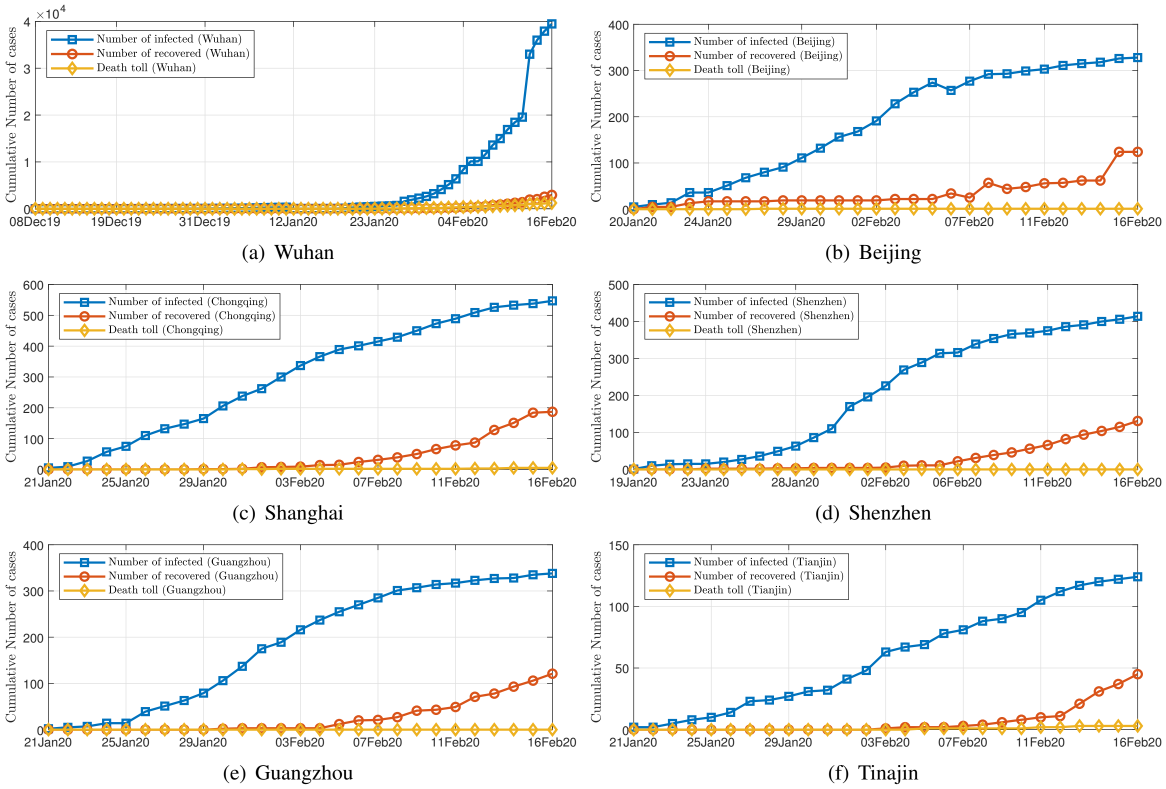
In the remainder of the paper, we first introduce the official daily infection data and the intercity migration data used in this study. The SEIR model is modified to incorporate the human migration dynamics, giving a realistic model suitable for studying the COVID-19 epidemic spreading dynamics. Historical data of infected, recovered and death cases from official source and data of daily intercity traffic (number of travellers between cities) extracted from Baidu Migration are used to generate the model parameters, which then enable estimation of the propagation of the epidemic in the coming months. We will conclude with a brief discussion of our estimation of the propagation and the reasonableness of our estimation in view of the measures taken by the Chinese authorities in controlling the spreading of this new disease.

2. Data

2.1. Official Data of COVID-19 Cases

The availability of official data of infected cases in China varies from city to city. Wuhan, being the epicenter, has the first confirmed case of COVID-19 infection on December 8, 2019 [1]. Most other cities in China began to report cases of COVID-19 infections around mid January 2020. Our data of daily infected and recovered cases, and death tolls, are based on the official data released by the National Health Commission of China, and the daily data used in our study are from January 24, 2020, to February 16, 2020, including the daily total number of confirmed cases in each city, daily total cumulative number of confirmed cases in each city, daily cumulative number of recovered cases in each city, and daily cumulative death toll in each city. It should be emphasized that the official data may not be the actual (true) data. Although the earliest confirmed case appeared on December 8, 2019, subsequent missing cases were expected to be significant in Hubei Province in the early stage of the epidemic outbreak. Systematic updates of

medRxiv preprint doi: https://doi.org/10.1101/2020.02.18.20024570; this version posted February 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .



(a) Wuhan

(b) Beijing

(c) Shanghai

(d) Shenzhen

(e) Guangzhou

(f) Tinajin

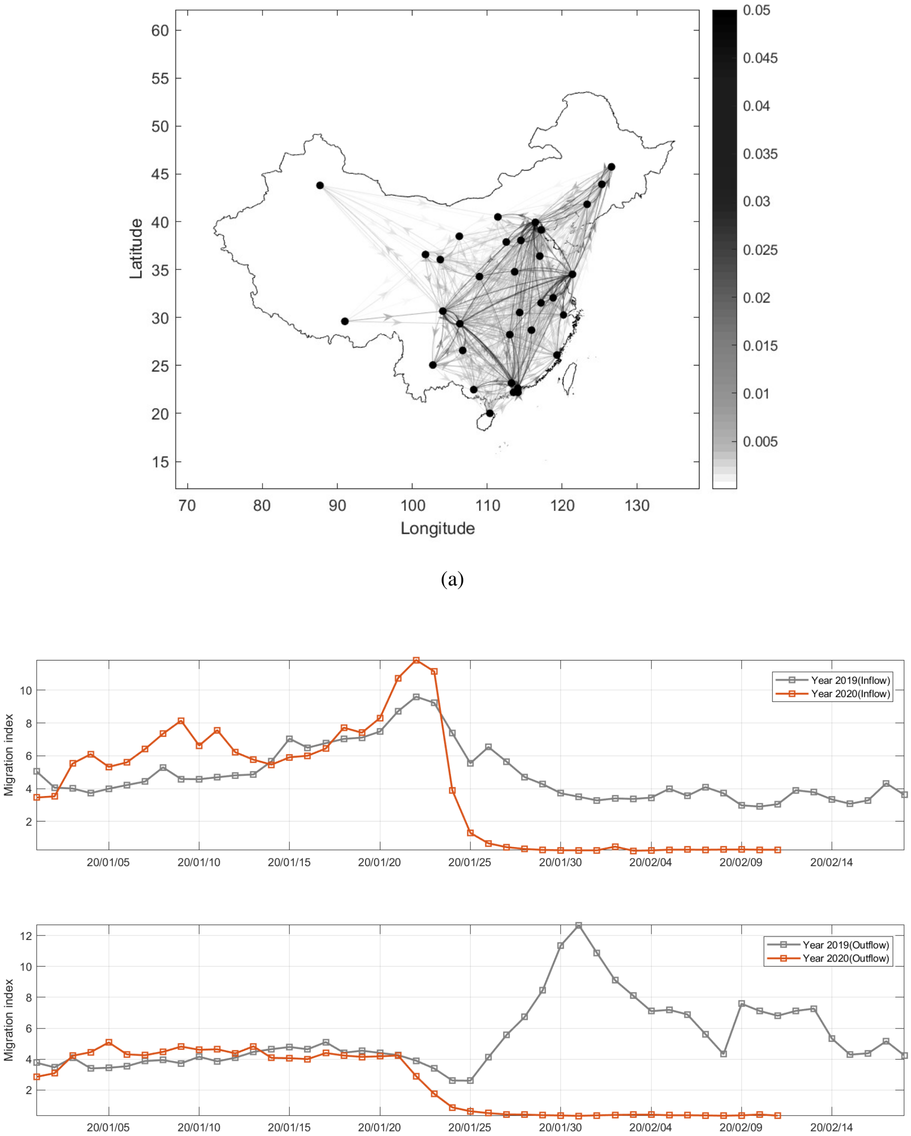
Figure 2: Daily data of COVID-19 infections in six Chinese cities from December 8, 2019 to February 13, 2020. (a) Wuhan (available from December 8, 2019); (b) Beijing (available from January 20, 2020); (c) Chongqing (available from January 20, 2020); (d) Shenzhen (available from January 19, 2020); (e) Guangzhou (available from January 21, 2020); (d) Tianjin (available from January 21, 2020).

infection data in other cities began after January 17, 2020. Figure 2 shows the number of confirmed infected cases, recovered cases and death tolls of six major Chinese cities.

2.2. *Intercity Travel Data*

As human-to-human transmission has been confirmed to occur in the spreading of COVID-19, gatherings of people and intercity travel of infected and exposed individuals within China have been the main drives that escalate the spreading of the virus. The period (around 20 days) surrounding the Lunar New Year (mid January to early February in 2020) is the most important holiday period in China. Migrant workers and students travel from major cities to country towns for family reunions, and return to the cities at the end of the holiday period. Holiday goers also travel to and from tourist cities. China's Ministry of Transport estimates around 3 billion trips to be taken during this period. Wuhan, being a major transport hub and having a large number of higher education institutions as well as manufacturing plants, is among the cities with the largest outflow and inflow traffic before and after the Chinese New Year festival. Our study aims to incorporate these important human migration dynamics in the construction of the spreading model. We collect daily intercity travel data in China from Baidu Migration, which is a mobile-app based big data system recording movements of mobile phone users. Specifically, we have collected Baidu Migration data for 367 cities (or administrative regions) in China over the period of January 1, 2020, to February 13, 2020. The data provide the migration strengths of cities which are indicative measures of the human traffic volume moving in and out of individual cities and administrative regions, as depicted by the inflow and outflow networks shown in Figure 3(a).

medRxiv preprint doi: https://doi.org/10.1101/2020.02.18.20024570; this version posted February 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .



(a)

(b)

Figure 3: (a) Illustration of travel network of major cities in China with arrows indicating direction of travel and darkness of lines indicating migration strengths (relative traffic volumes); (b) total inflow/outflow of travellers to/from Wuhan from/to other Chinese cities using Baidu Migration data. Inflow and outflow data of each city with individual cities are also collected to form m_{ij}.

Based on the collected data, we construct the *migration matrix*, which is given as

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

medRxiv preprint doi: https://doi.org/10.1101/2020.02.18.20024570; this version posted February 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

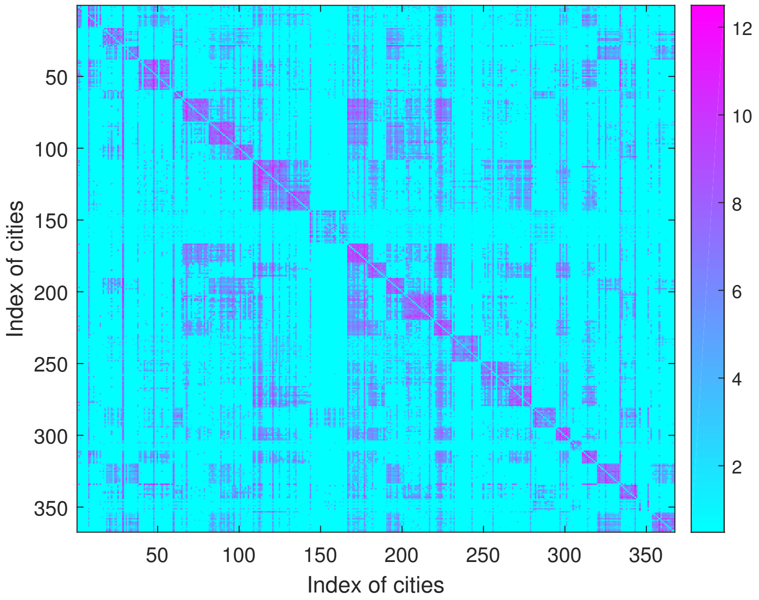


Figure 4: Visualization of migration volume ($\log(m_{ij})$) from city i (x-axis) to city j (y-axis) on January 23, 2020.

where K is the number of the cities or administrative regions ($K = 367$ in this study), 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, as shown in Figure 3(a). Several properties of M are worth noting:

- M records migration from one city to another. Movement within a city is not counted, i.e., $m_{ii}(t) = 0$ for all i.
- M is non-symmetric as traffic from one city to another is not necessarily reciprocal at any given time, i.e., $m_{ij}(t) \neq m_{ji}(t)$.
- Number of outflow migrants of city i at time t is

$$m_i^{(\text{out})}(t) = \sum_{i=j}^{K} m_{ij}(t). \tag{2}$$

- Number of inflow migrants of city i at time t is

$$m_i^{(\text{in})}(t) = \sum_{j=1}^{K} m_{ji}(t). \tag{3}$$

The charts in Figure 3(b) plot the daily total inflow and outflow migration strengths of Wuhan, showing the abrupt decrease of migration strengths after the city shut down all inbound and outbound traffic from January 24, 2020. A visualization of migration matrix M is given in Figure 4, where x-axis and y-axis correspond to city i and city j, respectively, and the color intensity indicates the migrant volume.

3. Method

3.1. Model

In the SEIR model, each individual in a population may assume one of four possible states at any time in the dynamic process of epidemic spreading, namely, susceptible (S), exposed (E), infected (I) and recovered/removed

medRxiv preprint doi: https://doi.org/10.1101/2020.02.18.20024570; this version posted February 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

(R). The dynamics of the epidemic can be described by the following set of equations: $\dot{S}(t) = -\beta S(t)I(t)$, $\dot{E}(t) = \beta S(t)I(t) - kE(t)$, $\dot{I}(t) = \kappa E(t) - \gamma I(t)$, and $\dot{R}(t) = \gamma I(t)$, where $S(t)$, $E(t)$, $I(t)$ and $R(t)$ are, respectively, the number of people **s**usceptible to the disease, **e**xposed (being able to infect others but having no symptoms), **i**nfected (diagnosed as confirmed cases), and **r**ecovered (including death cases); β is the exposition rate (infection rate of susceptible individuals); κ is the infection rate of exposed individuals; and γ is the recovery rate. For simplicity, recovered individuals include patients recovered from the disease and death tolls. In discrete form, the SEIR model can be represented by

$$\begin{aligned} \Delta S(t) &= -\beta S(t-1)I(t-1), \\ \Delta E(t) &= \beta S(t-1)I(t-1) - \kappa E(t-1), \\ \Delta I(t) &= \kappa E(t-1) - \gamma I(t-1), \\ \Delta R(t) &= \gamma I(t-1) \end{aligned} \tag{4}$$

where $\Delta S(t) = S(t) - S(t-1)$, $\Delta E(t) = E(t) - E(t-1)$, $\Delta I(t) = I(t) - I(t-1)$, and $\Delta R(t) = R(t) - R(t-1)$, with t being a daily count. As the incubation period for COVID-19 can be up to 14 days, the number of exposed individuals (who show no symptom but are able to infect others) plays a crucial role in the spreading of the disease. The state E, which is not available from the official data, is thus an important state in our model. Furthermore, combining death toll with the recovered number as state R will simplify the computation without affecting the accuracy of our data fitting and subsequent estimation.

Suppose, for city i, the four states are $S_i(t)$, $E_i(t)$, $I_i(t)$ and $R_i(t)$, at time t. Here, we also define a total susceptible population, N_i^s, which is the eventual number of infected individuals in city i. 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

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

The classic SEIR model would give ΔI_i as the difference between the number of exposed individuals who become infected and the number of removed individuals. However, the onset of the COVID-19 epidemic has occurred in a special period of time in China, during which a huge migration traffic is being carried among cities, leading to a highly rapid transmission of the disease throughout the country. In view of this special migration factor, the SEIR model should incorporate the human migration dynamics in order to capture the essential features of the dynamics of the spreading. In particular, for city i, in addition to the abovementioned classic interpretation, the daily increase in the number of infected cases should also include the inflow of infected individuals from other cities, less the outflow of removed cases from city i. In reality, inflow and outflow of exposed individuals to and from the city are also important and to be estimated in the model. Thus, if $m_{ij}(t)$ people move from city i to city j on day t, and the population of city i is $P_i(t)$, then the number of infected individuals moving from city i to city j is

$$\Delta I_{ij}^{\text{in}}(t) = \frac{I_i(t)m_{ij}(t)}{P_i}. \tag{6}$$

Also, the number of migrants leaving from city j is $\sum_{i=1}^{N} m_{ji}$, and the number of infected cases that have migrated out of city j is

$$\Delta I_j^{\text{out}}(t) = \frac{I_j(t)\sum_{i=1}^{N} m_{ji}(t)}{P_j}, \tag{7}$$

where $P_j(t)$ is the population of city j on day t. Thus, the increase in infected cases on day t in city j is given by

$$\begin{aligned} \Delta I_j(t) &= \kappa_j(t)E_i(t) - \gamma(t)I_j(t) + \sum_{i=1}^{N} \Delta I_{ij}^{\text{in}}(t) - \Delta I_j^{\text{out}}(t) \\ &= \kappa_j(t)E_i(t) - \gamma_j(t)I_j(t) + \sum_{i=1}^{N} \left(\frac{I_i(t)m_{ij}(t)}{P_i(t)} \right) - \frac{I_j(t)\sum_{i=1}^{N} m_{ji}(t)}{P_j(t)}, \end{aligned} \tag{8}$$

where $\Delta I_j(t) = I_j(t+1) - I_j(t)$ and $\kappa_j(t)$ is the infection rate in city j on day t, i.e., the rate at which exposed individuals become infected. Moreover, infected individuals, once confirmed, would unlikely be able to migrate to another city.

medRxiv preprint doi: https://doi.org/10.1101/2020.02.18.20024570; this version posted February 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

We thus implement this condition by writing (8) as

$$\Delta I_j(t) = \kappa_j(t)E_i(t) - \gamma_j(t)I_j(t) + k_I \left(\sum_{i=1}^{N} \left(\frac{I_i(t)m_{ij}(t)}{P_i(t)} \right) - \frac{I_j(t) \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)} \right), \tag{9}$$

where $0 < k_I \ll 1$ is a constant representing the possibility of an infected individual moving from one city to another.

Likewise, incorporating the migrant dynamics, the increase in exposed individuals on day t in city j is

$$\Delta E_j(t) = \frac{\beta_j(t)}{N_j^s(t)} I_j(t)S_j(t) + \frac{\alpha_j(t)}{N_j^s(t)} E_j(t)S_j(t) - \kappa_j(t)E_i(t) + \sum_{i=1}^{N} \left(\frac{E_i(t)m_{ij}(t)}{P_i(t)} \right) - \frac{E_j(t) \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)}, \tag{10}$$

where $\Delta E_j(t) = E_j(t+1) - E_j(t), \beta_j$ is the infection rate of susceptible individuals in city j, and α_j is the infection rate of exposed individuals in city j. In a likewise fashion, we have

$$\Delta S_j(t) = -\frac{\beta_j(t)}{N_j^s(t)} I_j(t)S_j(t) - \frac{\alpha_j(t)}{N_j^s(t)} E_j(t)S_j(t) + \sum_{i=1}^{N} \left(\frac{S_i(t)m_{ij}(t)}{P_i(t)} \right) - \frac{S_j(t) \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)}, \tag{11}$$

where $\Delta S_j(t) = S_j(t+1) - S_j(t)$. Finally, we have

$$\Delta R_j(t) = \gamma_j(t)I_j(t), \tag{12}$$

where $\Delta R_j(t) = R_j(t+1) - R_j(t)$. In the above derivation, we should note that

- the recovered individuals are assumed to stay in city j;
- the recovery rates in different cities are assumed to be different due to varied quality of treatments and availability of medical facilities;
- the recovery rates increase as time goes, as treatment methods are expected to improve gradually (i.e., taking $\gamma_j(t)$ as a monotonically increasing function);
- the evetual recovery rates in all cities will converge to the same constant $\Gamma \approx 1$.

In addition, due to intercity migration, the population of city j on day t would increase or decrease according to

$$\Delta P_j(t) = \sum_{i=1}^{N} \left(\frac{P_i(t)m_{ij}(t)}{P_i(t)} \right) - \frac{P_j(t) \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)} = \sum_{i=1}^{N} m_{ij}(t) - \sum_{i=1}^{N} m_{ji}(t), \tag{13}$$

where $\Delta P_j(t) = P_j(t+1) - P_j(t)$. Thus, the total susceptible population should be

$$\begin{aligned} \Delta N_j^s(t) = k_I & \left(\sum_{i=1}^{N} \left(\frac{I_i(t)m_{ij}(t)}{P_i(t)} \right) - \frac{I_j(t) * \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)} \right) + \sum_{i=1}^{N} \left(\frac{E_i(t)m_{ij}(t)}{P_i(t)} \right) - \frac{E_j(t) * \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)} \\ & + \sum_{i=1}^{N} \left(\frac{S_i(t)m_{ij}(t)}{P_i(t)} \right) - \frac{S_j(t) \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)}. \end{aligned} \tag{14}$$

where $\Delta N_j^s(t) = N_j^s(t+1) - N_j^s(t)$.

medRxiv preprint doi: https://doi.org/10.1101/2020.02.18.20024570; this version posted February 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

In summary, our modified SEIR model with consideration of human migration dynamics, for city j, is given by

$$\begin{cases} \Delta I_j(t) = \kappa_j(t)E_i(t) - \gamma_j(t)I_j(t) + k_I\left(\sum_{i=1}^{N}\left(\frac{I_i(t)m_{ij}(t)}{P_i(t)}\right) - \frac{I_j(t) * \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)}\right), \\ \Delta E_j(t) = \frac{\beta_j(t)}{N_j^s(t)} I_j(t)S_j(t) + \frac{\alpha_j(t)}{N_j^s(t)} E_j(t)S_j(t) - \kappa_j(t)E_i(t) + \sum_{i=1}^{N}\left(\frac{E_i(t)m_{ij}(t)}{P_i(t)}\right) \\ \qquad - \frac{E_j(t) * \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)}, \\ \Delta S_j(t) = -\frac{\beta_j(t)}{N_j^s(t)} I_j(t)S_j(t) - \frac{\alpha_j(t)}{N_j^s(t)} E_j(t)S_j(t) + \sum_{i=1}^{N}\left(\frac{S_i(t)m_{ij}(t)}{P_i(t)}\right) - \frac{S_j(t)\sum_{i=1}^{N} m_{ji}(t)}{P_j(t)}, \\ \Delta R_j(t) = \gamma_j(t)I_j(t), \\ \Delta P_j(t) = \sum_{i=1}^{N} m_{ij}(t) - \sum_{i=1}^{N} m_{ji}(t), \\ \Delta N_j^s(t) = k_I\left(\sum_{i=1}^{N}\left(\frac{I_i(t)m_{ij}(t)}{P_i(t)}\right) - \frac{I_j(t) * \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)}\right) + \sum_{i=1}^{N}\left(\frac{E_i(t)m_{ij}(t)}{P_i(t)}\right) - \frac{E_j(t) * \sum_{i=1}^{N} m_{ji}(t)}{P_j(t)} \\ \qquad + \sum_{i=1}^{N}\left(\frac{S_i(t)m_{ij}(t)}{P_i(t)}\right) - \frac{S_j(t)\sum_{i=1}^{N} m_{ji}(t)}{P_j(t)}. \end{cases} \tag{15}$$

where subscript j denotes the city itself, and subscript i denotes another city from/to which people migrate on day t. Letting $X_j(t)$ be the extended state vector, i.e., $X_j(t) = [S_j(t)\ E_j(t)\ I_j(t)\ R_j(t)\ P_j(t)\ N_j^s(t)]^T$, we write the above difference equation as

$$\Delta X_j(t) = f(X_j, X_i, \mu_i) \tag{16}$$

where $f(x)$ is the right side of (15), and μ_j is the set of parameters including $\alpha_j, \beta_j, \gamma_j, \kappa_j$ and δ_j. For computational convenience, we write (15) as

$$X_j(t+1) = X_j(t) + f(X_j, X_i, \mu_i) \tag{17}$$

In performing the data fitting, we assume $\alpha_j(t)$, $\beta_j(t)$, $\gamma_j(t)$, $\kappa_j(t)$, and δ_j are constants throughout the period of spreading, and the spreading begins at t_0, at which $N_j^s(t_0) = \delta_j P_j(t_0)$.

3.2. Parameter Identification

The model represented by (17) describes the dynamics of the epidemic propagation with consideration of human migration dynamics. The parameters in model (17) are unknown and to be estimated from historical data. We solve this parameter identification problem via constrained nonlinear programming (CNLP), with the objective of finding an estimated growth trajectory that fits the data. An estimated number of infected cases of each city can be generated from (15) with unknown set θ_j, i.e.,

$$\theta_j = \{\alpha_j, \beta_j, \gamma_j, \kappa_j, \delta_j, I_{j,0}\} \tag{18}$$

where $I_{j,0} = I_j(t_0)$ is the initial number of infections in city j, and $\{\alpha_j, \beta_j, \gamma_j, \kappa_j, \delta_j\}$ are parameters that determine the rates of spreading and recovery in city j. Then, the unknown set is $\Theta = \{\theta_1, \theta_2, \cdots, \theta_K\}$ essentially has $5K$ unknowns, where K is the number of cities, thus requiring an enormous effort of computation. Here, to gain computational efficiency, we assume that

- all cities share one parameter set $\theta = \{\alpha, \beta, \kappa, \gamma\}$;
- the numbers of initial infected and exposed individuals in city i are $\lambda_I I_i(t_0)$ and $\lambda_E I_i(t_0)$, respectively, where λ_I and λ_E are constant;
- each city has an independent δ_i.

medRxiv preprint doi: https://doi.org/10.1101/2020.02.18.20024570; this version posted February 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Then, the size of the unknown set becomes computationally manageable, i.e.,

$$\Theta = \{\alpha, \beta, \kappa, \gamma, \delta_i, \lambda_I, \lambda_E\}.$$

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

$$\begin{aligned} &\mathrm{P}_0\colon \min_{\Theta} \sum_{j=0}^{N} \left\| w_j(I(t_j) - \hat{I}(t_j)) \right\|_l \\ &s.t. \begin{cases} (i) & \hat{x}(t+1) = \hat{x}(t) + F(\hat{x}(t)), \\ (ii) & \Theta_U \geq \Theta \geq \Theta_L, \end{cases} \end{aligned} \tag{19}$$

where $F(\cdot)$ represents model (15) and $\hat{x}(t) = [\hat{I}(t), \hat{R}(t), E(t), S(t), P(t), N^s(t)]$ is the set of estimated variables, with unknown set Θ, which 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 (19).

The Root Mean Square Percentage Error (RMSPE) is adopted as the criterion, i.e., fitting error, to measure the difference between the number of infected individuals generated by the model and the official daily infection data.

$$\mathrm{RMSPE} = \sqrt{\frac{1}{K} \sum_{i=1} \sum_{j=1} \left(\frac{\hat{I}_i(t_j) - I_i(t_j)}{I_i(t_j)} \right)^2} \times 100\%, \tag{20}$$

where K is the number of cities to be evaluated.

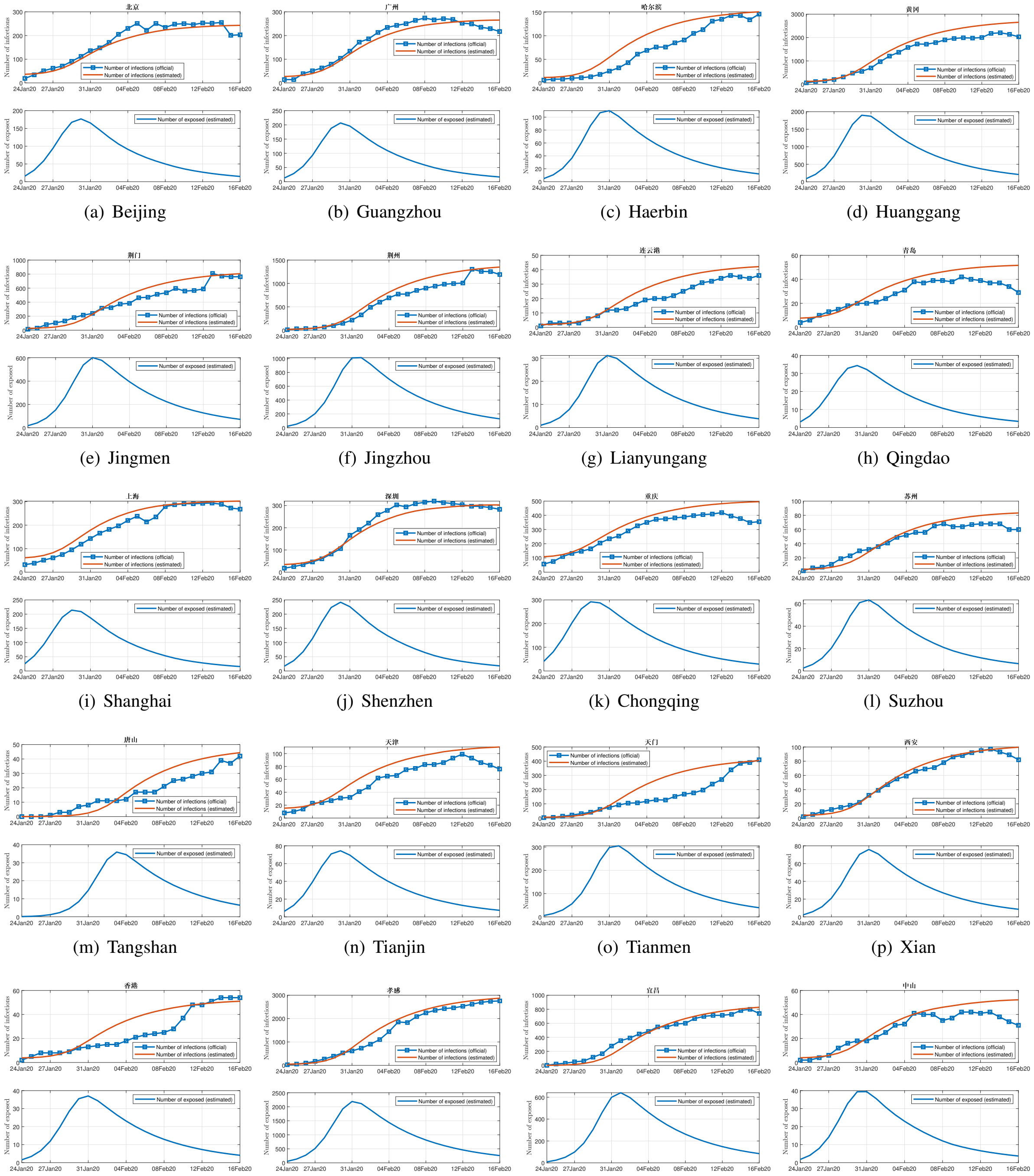
4. Results and Discussions

We perform data fitting of the model, described by (17), using historical daily infection data provided by the National Health Commission of China, from January 24, 2020 to February 13, 2020. Our approach, as described in the previous section, is to apply constrained nonlinear programming to find the best set of estimates for the unknown parameters and states. Data fitting for all 367 cities are performed. Values are updated iteratively in the optimization process. Moreover, since all parameters, like infection rates, are to be generated by fitting data with the model, the integrity of the data becomes crucial. As the official Wuhan data are expected to deviate from the true values quite significantly during the early outbreak stage due to uncertainty in diagnosis and other issues related to reporting of the epidemic by the local government, we have allowed the fitting errors for Wuhan to expand over a reasonable range, while the fitting errors for most other cities remain small. In addition, as the epidemic propagates in time, effective control measures and improved public education would reduce the infection rates for the susceptible and exposed individuals, making these parameters time varying in reality. Nonetheless, our fitting assumes these parameters being constant during the short fitting period for computational simplicity.

The propagation profiles, in terms of the number of infected individuals and estimated number of exposed individuals, for all 367 cities are estimated. As limited by space, we only show in Figure 5 the results for 20 selected cities. This model can also provide projections of the number of infected and exposed individuals in the next 200 days, as shown in Figure 6, which clearly show that the daily infection would reach a peak sooner or later. By running the identification algorithm, we identify the optimal parameter set as $\alpha = 0.5869$, $\beta = 0.8949$, $\kappa = 0.1008$, and $\gamma = 0.0602$. From the estimated propagation profiles of the COVID-19 epidemic for all 367 cities, we have the following findings:

- For most cities in China, the infection numbers will peak between mid February to early March 2020, as shown in Figure 7(a).
- The peak number of infected individuals will be between 1,000 to 5,000 for cities in Hubei, and that outside Hubei will be below 500, as shown in Figure 7(b).
- At the end, about 0.8%, less than 0.1% and less than 0.01% of the population will get infected in Wuhan, Hubei Province and the rest of China, respectively, as graphically presented in Figure 7(c).

medRxiv preprint doi: https://doi.org/10.1101/2020.02.18.20024570; this version posted February 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .



(a) Beijing

(b) Guangzhou

(c) Haerbin

(d) Huanggang

(e) Jingmen

(f) Jingzhou

(g) Lianyungang

(h) Qingdao

(i) Shanghai

(j) Shenzhen

(k) Chongqing

(l) Suzhou

(m) Tangshan

(n) Tianjin

(o) Tianmen

(p) Xian

(q) Hong Kong

(r) Xiaogan

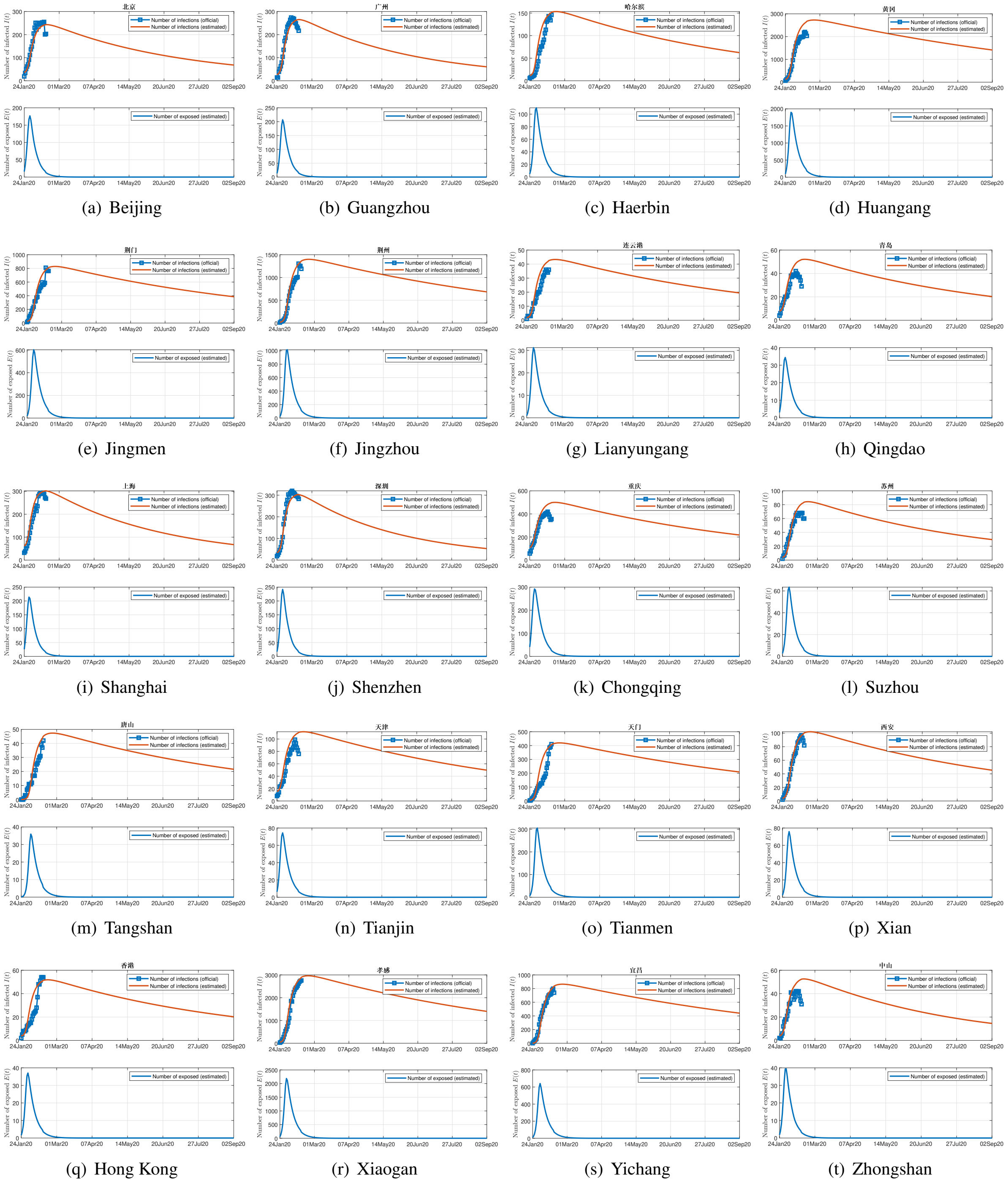
(s) Yichang

(t) Zhongshan

Figure 5: Official number of infected individuals and estimated number of infected individuals in 20 selected cities in China (upper), and estimated number of exposed individuals (lower).

Opinions diverge in the social media on the estimated extent of the outbreak of the new coronavirus disease (COVID-19). While there are pure speculations, there are also predictions based on rigorous study of the spreading dynamics. Different models used for prediction and different assumptions made regarding the transmission process

medRxiv preprint doi: https://doi.org/10.1101/2020.02.18.20024570; this version posted February 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .



(a) Beijing (b) Guangzhou (c) Harbin (d) Huangang

(e) Jingmen (f) Jingzhou (g) Lianyungang (h) Qingdao

(i) Shanghai (j) Shenzhen (k) Chongqing (l) Suzhou

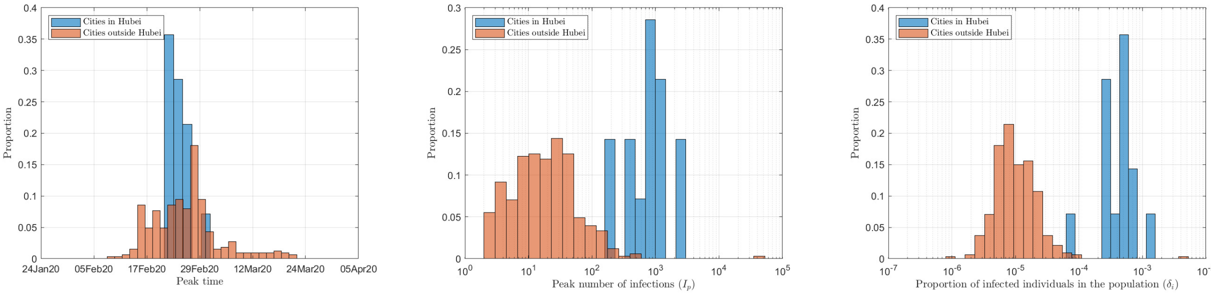
(m) Tangshan (n) Tianjin (o) Tianmen (p) Xian

(q) Hong Kong (r) Xiaogan (s) Yichang (t) Zhongshan

Figure 6: Prediction of the number of infected (upper) and exposed individuals (lower) in 20 selected cites in China for the next 200 days Chongqing for the next 200 days.

would lead to different results and quite diverged conclusions. For instance, an AI-powered simulation run had forecasted 2.5 billion people to be infected in 45 days [13]. Academics in Hong Kong expected 1.4 million eventually infected in the city of 7.5 million people. Our results, however, do not seem to agree with such forecasts. In fact, our

medRxiv preprint doi: https://doi.org/10.1101/2020.02.18.20024570; this version posted February 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .



(a) (b) (c)

Figure 7: (a) Distribution of peak time; (b) distribution of peak number of infections; (c) distribution of proportion of the population eventually infected in a city.

results are expected to be conservative, under normal circumstances, in the sense that the severity and duration of the epidemic could be over-estimated. This is because stringent measures have been taken to limit travel and hence to curb the spreading of the virus since late January 2020. The infection rates should be progressively reduced, with the effect of reducing the number of infected individuals as well as mitigating of the peak times. Furthermore, climate factors may also play a role in the epidemic spreading. As temperature generally rises from February to May, the infection peaks would unlikely occur later than the predicted times as the new coronavirus epidemic is generally known to subside in warm weather. Based on our findings and assuming no drastic change in infection rates (e.g., schools and workplaces remain closed and intercity travel limited), we can conclude, with optimism, that the COVID-19 epidemic would end around April 2020.

5. Conclusion

We employ human migration data to provide information on intercity travel that is crucial to the transmission of the new coronavirus disease from its epicenter Wuhan to other parts of China. Our model for the disease spreading is essentially the classic SEIR model, with intercity travel data supplying the essential information about the number of infected, exposed and recovered individuals moving between different cities. All parameters of the model, including infection rates, recovery rates, and eventual percentage of infected population for 367 cities in China, are identified by fitting the official data with the model using a constrained nonlinear programming procedure. Using these parameters, estimates of the number of exposed individuals in 367 cities as well as projections into the next 200 days are also found. Our model, however, does not consider the contact network topology that would be necessary if details of the transmission process, such as superspreading events, are to be captured. Nonetheless, our model provides a very reasonable estimation of the propagation of average numbers of exposed, infected and recovered individuals, despite missing details of fluctuation (e.g., sudden surge due to a superspreading event). The main conclusion of our study is that the COVID-19 epidemic spreading will 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.

Acknowledgment

This work was supported by National Science Foundation of China Project 61703355 and Guangdong Youth University Innovative Talents Project 2016KQNCX223, and City University of Hong Kong under Special Fund 9380114 provided by VP(RT).

References

[1] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong, et al., Early transmission dynamics in Wuhan China of novel coronavirus-infected pneumonia, New England Journal of Medicine (2020) DOI: 10.1056/NEJMoa2001316.

medRxiv preprint doi: https://doi.org/10.1101/2020.02.18.20024570; this version posted February 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

[2] O. Diekmann, H. Heesterbeek, T. Britton, Mathematical Tools for Understanding Infectious Disease Dynamics, Princeton University Press, Princeton, 2013.

[3] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Physical Review Letters 86 (14) (2001) 3200.

[4] M. Boguná, R. Pastor-Satorras, A. Vespignani, Absence of epidemic threshold in scale-free networks with degree correlations, Physical Review Letters 90 (2) (2003) 028701.

[5] M. Small, C. K. Tse, D. Walker, Super-spreaders and the rate of transmission of the SARS virus, Physica D 215 (2006) 146–158.

[6] M. Small, C. K. Tse, Small world and scale free network model of transmission of SARS, International Journal of Bifurcation and Chaos 15 (5) (2005) 1745–1756.

[7] M. Small, C. K. Tse, Clustering model for transmission of the SARS virus: Application to epidemic control and risk assessment, Physica A 351 (2005) 499–511.

[8] Y. Wang, D. Chakrabarti, C. Wang, C. Faloutsos, Epidemic spreading in real networks: An eigenvalue viewpoint, in: Proceedings of International Symposium on Reliable Distributed Systems, IEEE, 25–34, 2003.

[9] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, C. Faloutsos, Epidemic thresholds in real networks, ACM Transactions on Information and System Security (TISSEC) 10 (4) (2008) 1.

[10] T. Gross, C. J. D. D'Lima, B. Blasius, Epidemic dynamics on an adaptive network, Physical review letters 96 (20) (2006) 208701.

[11] Z. Du, L. Wang, S. Cauchemez, X. Xu, X. Wang, B. J. Cowling, L. A. Meyers, Risk of transportation of 2019 novel coronavirus disease from Wuhan to other cities in China, Emerging Infectious Disease 26 (5) (2020) DOI: 10.3201/eid2601.200146.

[12] Baidu Migration, URL `https://http://qianxi.baidu.com/`, 2020.

[13] J. Koetsier, AI predicts coronavirus could infect 2.5 billion and kill 53 million, URL `https://www.healthexec.com/topics/care-delivery/ai-predicts-25b-coronavirus-infections`, 2020.