Modeling the complete spatiotemporal spread of the COVID-19 epidemic in mainland China

Bisong Hu\textsuperscript{a,b}, Pan Ning\textsuperscript{a}, Jingyu Qiu\textsuperscript{c}, Vincent Tao\textsuperscript{c}, Adam Thomas Devlin\textsuperscript{a}, Haiying Chen\textsuperscript{d}, Jinfeng Wang\textsuperscript{b,*}, Hui Lin\textsuperscript{a,*}

\textsuperscript{a}School of Geography and Environment, Jiangxi Normal University, Nanchang 330022, China

\textsuperscript{b}State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

\textsuperscript{c}Wayz AI Technology Company Limited, Shanghai 201210, China

\textsuperscript{d}Nanchang Center for Disease Control and Prevention, Nanchang 330038, China

\textsuperscript{*}Corresponding authors.

E-mail addresses:
hubisong@jxnu.edu.cn (B. Hu), ningpan@jxnu.edu.cn (P. Ning), jingyu.qiu@wayz.ai (J. Qiu), vincent.tao@wayz.ai (V. Tao), atdevlin@jxnu.edu.cn (A. T. Devlin), nccdcchy@126.com (H. Chen), wangjf@lreis.ac.cn (J. Wang), huilin@cuhk.edu.hk (H. Lin).
Present addresses:

aSchool of Geography and Environment, Jiangxi Normal University, No. 99, Ziyang Rd., Nanchang 330022, Jiangxi, China

bState Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, No. 11A Datun Rd., Chaoyang District, Beijing 100101, China

cWayz AI Technology Company Limited, No. 58, Xiangke Rd., Pudong District, Shanghai 201210, China

dNanchang Center for Disease Control and Prevention, No 833, Lijing Rd, Honggutan New District, Nanchang 330038, Jiangxi, China
Graphical abstract

Highlights

- The mean latent/removed periods of COVID-19 epidemic in China were 5.40/2.13 days.
- There was 1.72 latent/infected persons per 10,000 Wuhan travelers until January 20th.
- Space-time $R_0$ estimates indicated an initial value between 2 and 3.5 in most cities.
- There were 196 cities which had an estimate of $R_0<1$ on February 29th.
- An obvious descent in temporal $R_0$ started 1 week after the outbreak and lasted 2 weeks.
Abstract

Objectives: The outbreak of novel coronavirus (COVID-19) epidemic is reaching its final phase in China. The complete epidemic data are available for a complete assessment of epidemiological parameters in all regions and time periods.

Methods: This study aims to present a spatiotemporal epidemic model based on spatially stratified heterogeneity (SSH) to simulate the epidemic spread. A susceptible-exposed/latent-infected-removed (SEIR) model was constructed for each SSH-identified stratum (each administrative city) to estimate the spatiotemporal epidemiological parameters of the outbreak.

Results: We estimated that the mean latent and removed periods were 5.40 and 2.13 days, respectively. There was an average of 1.72 latent or infected persons per 10,000 Wuhan travelers to elsewhere until January 20th, 2020. The space-time basic reproduction number ($R_0$) estimates indicate an initial value between 2 and 3.5 in most cities on this date. The mean period for $R_0$ estimates to decrease to 80% and 50% of initial values in cities were an average of 14.73 and 19.62 days, respectively.

Conclusions: Our model estimates the complete spatiotemporal epidemiological characteristics of the outbreak in a space-time domain, and the findings will help enhance a comprehensive understanding of the outbreak and inform the strategies of prevention and control in other countries worldwide.

Keywords
COVID-19; **spatially stratified heterogeneity**; SEIR model for a stratum; space-time

$R_0$: latent and infection ratio; mainland China

**Introduction**

As of this writing, the novel coronavirus (COVID-19) outbreak is considered to be almost over in China after an intensive spread lasting over two months. The complete data of the COVID-19 epidemic in China enables us to make a complete and accurate estimation of the epidemiological parameters in the country. Early confirmed cases were mainly linked to a seafood wholesale market in Wuhan, Hubei province, China, starting from late November, 2019 (Li et al., 2020; Zhu et al., 2020). The epidemic transmission in the early phase was associated with the movement of populations from the seafood market and away from Wuhan. Massive human movements via railways and airlines from Wuhan due to the annual Chinese (Lunar) New Year holiday migrations enabled the virus to spread nationwide and worldwide (Peeri et al., 2020). Travel restrictions and quarantine measures in Wuhan were effective in delaying the overall epidemic progression in mainland China and reduced the exportation of cases to international locations (Chinazzi et al., 2020). Strict control strategies activated by governments and individuals (e.g., transportation restrictions, holiday extensions and self-isolation measures) were effective in weakening the outbreak trend in China (Lin et al., 2020).

Previous studies of the epidemiological characteristics of COVID-19 focused on clinical and descriptive statistics (Chan et al., 2020; Guan et al., 2020; Huang et al.,
Modeling studies of COVID-19 transmission patterns include forecasting future spreads (Al-qaness et al., 2020; Alsayed et al., 2020; Wu et al., 2020), exploring characteristics and determinants (Li et al., 2020; Liu et al., 2020; Xiong et al., 2020), risk assessment (Boldog et al., 2020; Jung et al., 2020; Tang et al., 2020), control/restriction measure effects (Chinazzi et al., 2020; Tang et al., 2020), estimation of epidemiological parameters (e.g., basic reproduction number, $R_0$) (Chen et al., 2020; Mollalo et al., 2020; Wang et al., 2020; Zhang et al., 2020; Zhao et al., 2020a) and others. Epidemiological and modeling studies indicate that the COVID-19 epidemic has a $R_0$ value of 2–3 (Li et al., 2020; Wu et al., 2020; Zhang et al., 2020; Zhao et al., 2020a), which is lower than that of the 2003 severe acute respiratory syndrome (SARS) outbreak (Lipsitch, 2003; Riley et al., 2003).

However, the COVID-19 epidemic spread in China produced various regional outbreaks with diversified epidemiological characteristics which varied by region and human movements from the epidemic center. Moreover, epidemiological parameters/indicators (e.g., $R_0$, infective rate and removed rate) generally varied by region and time during the epidemic period because of the various control strategies activated by multiple governments and other entities.

A variety of epidemics indicate the region-varying and time-varying parameter characteristics (McCallum and Partridge, 2010), and a global model is inappropriate for all regions in such a large-scale area (e.g., mainland China). Control strategies (e.g., transportation control programs and 2-week self-isolation measures) in various regions might cause heterogeneous interactions between subpopulations to the region in
varying degrees. A global model would be confounded if the population exhibits spatially stratified heterogeneity (SSH) (Buttle et al., 2016; Wang et al., 2016; Xu et al., 2011). Traditional mathematical models of infectious diseases, such as susceptible-infected-removed (SIR) and susceptible-exposed/latent-infected-removed (SEIR) models, are appropriate and applicable under mass interaction assumption (i.e., subpopulations mix fully and homogeneously, and have identical interactions with one another) (Kermack and Mckendrick, 1991). Control strategies of the restriction of human migration between regions may cause the disagreement of a global model due to this assumption. Moreover, epidemic outbreak strongly associated with human movement from the epidemic source (e.g., the COVID-19 epidemic in China), generally indicates a spatiotemporal heterogeneity typical of human movements. Therefore, to better estimate the spatiotemporal epidemiological characteristics of an outbreak, SSH-based models can indicate similar characteristics within strata (e.g., region groups) and different characteristics between strata (Wang et al., 2010b).

In view of the above considerations, here we provide a SEIR model with a time-varying infective rate for each SSH-identified stratum to simulate the spatiotemporal epidemic spread of the COVID-19 outbreak in geographical strata of mainland China, based on the SSHs (Wang et al., 2016, 2010b) of the epidemic spread and human movements from infection sources. This study estimates the spatiotemporal epidemiological characteristics of the outbreak (e.g., space-time varying $R_0$) all over the country, spatial distribution of the imported latent and infected populations from the epidemic source and several other indicators (e.g., latent and removed periods) in strata.
(administrative cities in mainland China).

Materials and Methods

**SSH \(q\) statistic**

The SSH refers to ubiquitous phenomena (those within strata are more similar than those between strata), implies potential distinct mechanisms by stratum, and enforces the applicability of statistical inferences (Wang et al., 2016). The geographical detector (GeoDetector) \(q\) statistic is generally applied to quantitatively evaluate the SSH of an explained variable (Wang et al., 2016, 2010b), and assess the determinant power of explanatory variables and their interactions, without linear assumptions (Yin et al., 2019). The fundamental formula of the \(q\) statistic is given by:

\[
q = 1 - \frac{\sum_{h=1}^{L} N_h \sigma_h^2}{N \sigma^2}
\]

(1)

where \(q\), with a value ranging from 0 to 1, is the SSH measure of an explained variable or the determinant power of a factor to the objective. \(N\) is the number of explained variable observations and \(\sigma^2\) indicates the variance of all the observations. The explained variable is stratified into \(L\) strata, denoted by \(h = 1, 2, \ldots, L\), which is determined by the prior knowledge, the determinant factor or a classification algorithm. \(N_h\) is the number of observations and \(\sigma_h^2\) is the corresponding variance within stratum \(h\).

There must be at least two subregions in each stratum for the variance calculation within strata if the explained variable distributes spatially (Wang et al., 2010b). However, a stratum containing only one subregion is allowable when the explained...
variable observations are constructed in a space-time domain. There might be one or more subregions in a specific stratum under a given geographical stratification. The variance is calculated according to the spatiotemporal data in a specific stratum with multiple subregions and temporal observations. More specifically, if a stratum has only one subregion, its variance can be calculated based on the temporal observations. Multiple geographical stratification solutions in a large-scale area can be comparatively analyzed to identify an appropriate one to indicate the significant SSH of the epidemic spread and human movements from an infection source. Various epidemic models in strata can be combined to introduce a model of a complete spatiotemporal spread of the epidemic.

The GeoDetector software is accessible at http://geodetector.cn/, and the $q$ statistics in this study were performed with the use of the R software package (R Foundation for Statistical Computing).

**SEIR model for a stratum**

After determining a specific geographical stratification solution, the study area is stratified into $L$ strata denoted by $h=1, 2, \ldots, L$. The epidemic SEIR model is separately calibrated in each of the various strata. The population in stratum $h$ is divided into four subpopulations: the susceptible ($S_h$), the exposed/latent ($E_h$), the infected ($I_h$) and the removed/isolated/recovered/dead ($R_h$). The numbers of four subpopulations in stratum $h$ are denoted by $S_h(t)$, $E_h(t)$, $I_h(t)$ and $R_h(t)$, respectively, at time $t$. And the population amount in stratum $h$ is assumed to be constant and can be denoted by
In the current stratified model, SEIR models in various strata have a set of similar differential equations as follows:

\[
\frac{dS_h}{dt} = -\beta_h(t) \frac{I_h S_h}{N_h}
\]

\[
\frac{dE_h}{dt} = \beta_h(t) \frac{I_h S_h}{N_h} - \lambda_h E_h
\]

\[
\frac{dI_h}{dt} = \lambda_h E_h - \gamma_h I_h
\]

\[
\frac{dR_h}{dt} = \gamma_h I_h
\]

where \( \lambda_h \) is the latent rate at which exposed individuals become infectious in stratum \( h \) (\( 1/\lambda_h \) indicates the estimated latent period, \( T_{l,h} \)), and \( \gamma_h \) is the removed rate at which infectious individuals are removed (\( 1/\gamma_h \) denotes the estimated removed period, \( T_{r,h} \)).

The infective rate in stratum \( h \) denoted by \( \beta_h \) indicates the average infection number per infectious individual per timestep (e.g., one day), and is due to the temporal change of the effects of control programs and others, therefore, it is considered to depend on time:

\[
\beta_h(t) = a_h + \frac{b_h}{1 + e^{c_h(t-d_h)}}
\]

where \( a_h, b_h, c_h \) and \( d_h \) are the coefficients of an inverse logistic function to describe the temporal \( \beta_h(t) \) in stratum \( h \). Control measures have no explicitly representative parameters included in the model. However, they are reflected in the observed data, and thus, are adopted in the parameters in the modeling.

The SEIR model with time-varying infective rates is constructed stratum by stratum according to a specific stratification solution. The SEIR model for each SSH-identified stratum can describe the spatiotemporal variation of the epidemic spread and be used to estimate the epidemiological characteristics in a space-time domain.

In each stratum, the SEIR model has a total of eight parameters to be estimated,
including four coefficients of the $\beta_h(t)$ function ($a_h, b_h, c_h, d_h$), $\gamma_h, \lambda_h, E_h(1)$ and $I_h(1)$. The latter two are the estimated numbers of the latent and infected subpopulations in stratum $h$, respectively, which are imported from the infection source at time $t=1$ (the initial time of the simulation). The initial removed subpopulation is assumed to be zero and the susceptible subpopulation is calculated by $S_h(t) = N_h - I_h(t) - E_h(t) - R_h(t)$.

Using spatiotemporal data of the COVID-19 epidemic in China, we can fit the proposed SEIR model, estimate the model parameters in all the strata, and then estimate the spatiotemporal epidemiological characteristics. The mean latent and removed periods, respectively, can be estimated as follows:

$$T_l = \frac{1}{L} \sum_{h=1}^{L} T_{l,h} = \frac{1}{L} \sum_{h=1}^{L} \frac{1}{\hat{\lambda}_h}$$

$$T_r = \frac{1}{L} \sum_{h=1}^{L} T_{r,h} = \frac{1}{L} \sum_{h=1}^{L} \frac{1}{\hat{\gamma}_h}$$

where $L$ is the number of strata, $\hat{\lambda}_h$ is the estimated latent rate and $\hat{\gamma}_h$ is the estimated removed rate in stratum $h$.

During the modeling process, the latent and infected subpopulations imported from the infection source (Wuhan for the COVID-19 epidemic) are estimated at the initial time in all strata. Following this, we can then calculate the latent and infected ratio (L&I ratio) imported from Wuhan at $t=1$ in stratum $h$ as follows:

$$R_{L&I,h} = \frac{\hat{E}_h(1) + \hat{I}_h(1)}{n_h(1)}$$

where $R_{L&I,h}$ is the L&I ratio at $t=1$ in stratum $h$, $\hat{E}_h(1)$ and $\hat{I}_h(1)$ are the estimated imported latent and infected subpopulations, respectively, and $n_h(1)$ denotes the cumulative number of Wuhan travelers to stratum $h$ from January 1st, 2020 at $t=1$. Note
that the former is the date when the seafood market was closed, and human movements from Wuhan to elsewhere in the country contributed to the regional variation of the initial epidemic status by stratum.

Finally, the space-time $R_0$ in all the strata during the time period of the simulation can be estimated by:

$$R_{0,h}(t) = \frac{\beta_h(t)}{\gamma_h} = \frac{\hat{a}_h + \hat{b}_h/(1 + e^{\hat{c}_h(t-\hat{d}_h)})}{\gamma_h}$$

(10)

where $R_{0,h}(t)$ is the $R_0$ estimate at time $t$ in stratum $h$, and $\hat{a}_h$, $\hat{b}_h$, $\hat{c}_h$ and $\hat{d}_h$ are the parameter estimates of the infective rate function. The estimated $R_0$ function depends on time, varies by stratum and describes the complete spatiotemporal characteristics of the COVID-19 epidemic spread in mainland China.

**Data and stratification**

We collected the spatiotemporal data of daily new COVID-19 confirmed cases in administrative cities of mainland China from the daily bulletins of the National Health Commission of the People’s Republic of China (NHC) and various Provincial/Municipal Health Commissions. The final epidemic dataset was comparatively verified through the public platform of the 2019-nCoV-infected pneumonia epidemic from the Chinese Center for Disease Control and Prevention (China CDC) (China CDC, 2020). Additionally, demographic data was collected from the 2019 China Statistical Yearbook to identify the populations in strata.

The Huanan seafood wholesale market and Wuhan were considered as the primary and secondary epidemic centers. Human movements of populations from these two
sources were associated with the spatiotemporal epidemic spread. We used the data of location-based service (LBS) requests of mobile devices to indicate human migrations from the sources to elsewhere. The datasets of LBS requests, which cover over 80% mobile devices supported by the three telecommunication operators in China, were provide by Wayz Inc., Shanghai, China, and applied to identify the generations of the COVID-19 epidemic spreads in mainland China during the early phase (Hu et al., 2020).

The LBS-requesting statistics are implemented every two hours with high-resolution location information and private individual information was deleted from the raw data of the mobile devices.

Based on these above data sources, we evaluated the SSHs of the spatiotemporal data of COVID-19 confirmed cases and the LBS-requesting data of mobile devices from two epidemic sources. Three stratification solutions were carried out for a comparative analysis. First, considering the distance from the epidemic source, the study area (mainland China) was stratified into four subareas of increasing distance (strata): Wuhan, Hubei province excluding Wuhan, provinces adjacent to Hubei and the rest of the country (Solution 1). Second, provinces, municipalities and autonomous regions were considered as the strata (Solution 2). Third, all administrative cities were considered as the strata (Solution 3).

In each stratum, the spatiotemporal observations of an explained variable were used to calculate the corresponding variance. That is to say, the variance in calculating the SSH \( q \) value is about the number of confirmed cases or device traces. Note that the variances in Solution 3 were calculated based on temporal observations since each
stratum has only one subregion (city). The variances of the spatiotemporal observations within all strata were compared with the variance of all the observations, and according to Equation (1), the SSH $q$ value of the explained variable can be calculated under a specific stratification solution. There were three stratification solutions for the evaluations of the SSHs of the explained variables. The SSH $q$ values of each variable can be compared under different solutions. As shown in Table 1, spatiotemporal COVID-19 cases and two categories of device traces from the market and Wuhan had significantly weak SSH in the strata of Solution 2 ($q$ statistics < 0.1). However, they had strong SSH in the strata of both Solution 1 and Solution 3. The $q$ statistics of device traces from Wuhan and the seafood market (after January 1st, 2020) were higher than 0.86 and 0.95, respectively. The spatiotemporal COVID-19 cases had $q$ statistic values over 0.26. It is thus apparently inappropriate to construct the SEIR model for strata by provinces. The $q$ statistics in strata of cities were generally higher than those in the four strata separated by increasing distance from the epidemic source. In addition, more strata reveal more details of the epidemic in the space-time domain. We therefore selected Solution 3 to implement the stratification, i.e., the proposed SEIR model of simulating the spatiotemporal COVID-19 epidemic spread was constructed in strata of administrative cities in mainland China.

**Experimental setups**

Each administrative city in mainland China was considered as one single stratum and the SEIR model for a stratum was correspondingly constructed to simulate the epidemic
spread. The experimental time period was set from January 20\textsuperscript{th} to February 29\textsuperscript{th}, 2020, with a timestep of one day (\(t=1, 2, \ldots, 41\)). The population amount in stratum \(h\) at the end of 2018 was assigned to \(N_h\), which was set as a constant during the modeling process. The infected subpopulation, \(I_h\), had an initial value of the cumulative confirmed cases in the city at \(t=1\) (January 20\textsuperscript{th}), and was assumed not to exceed \(N_h\). The latent subpopulation, \(E_h\), was assumed to have a value range from zero to the cumulative number of Wuhan travelers to the city at \(t=1\), and had an initial importation ratio of 1/1000. Note that the initial values of these two variables were also set as the model parameters and would be re-estimated during the modeling process. The removed subpopulation, \(R_h\), had an initial value of zero at \(t=1\), and the susceptible subpopulation, \(S_h\), can be simply calculated. The latent rate, \(\lambda_h\), and the removed rate, \(\gamma_h\), were assumed to have bounds of 1/14–1/3 (day\(^{-1}\)) and 1/9–1 (day\(^{-1}\)), respectively. That is to say, the latent and removed periods were estimated, respectively, from 3 to 14 days and from 1 day to 9 days, which covered the corresponding values considered in most previous studies (e.g., Guan et al., 2020; Lin et al., 2020; Wu et al., 2020).

Based on the control of the spatiotemporal COVID-19 epidemic data, the proposed SEIR model can be solved according to Equations (2)–(6), and the parameter estimates of the SEIR models in strata were obtained afterwards. Furthermore, the spatiotemporal epidemiological parameters of the COVID-19 outbreak in mainland China could be calculated according to Equations (7)–(10). The root mean square error (RMSE), coefficient of determination (\(R^2\)) and adjusted \(R^2\) were selected to evaluate the model performance (goodness-of-fit, see Appendix A). The calculation was implemented by
comparing the observed and estimated (modeled) numbers of the cumulative cases. The calculations of the proposed SEIR model were implemented using MATLAB programming (MathWorks). Thematic mapping to display the geographical distributions of epidemiological parameters was implemented in the ArcGIS platform (ESRI).

Results

Model performance

The SEIR model for a stratum was constructed amongst a total of 315 applicable administrative cities in mainland China. The observations of the cumulative confirmed cases were applied to be compared with the modeled values for the evaluation of model performance. Table 2 shows the details of the model performance. Regarding the modeling accuracy, the mean values of $R^2$ and adjusted $R^2$ of the presented model were 0.9663 and 0.9592, respectively, and the corresponding values of one standard deviation (1-StdDev) were 0.0396 and 0.0480, respectively. Although several cities had slightly low-value $R^2$ and adjusted $R^2$ (lower than 0.7), over 75% cities had adjusted $R^2$ values over 0.9. The proposed SEIR model indicated a satisfactory overall performance in reducing the estimation errors and improving accuracy. Several cities with extremely high numbers of confirmed cases (e.g., Wuhan and several cities in Hubei), expanded the range of the RMSE value range (maximum value reached 1423.55 persons). However, the mean RMSE value was 8.9190 persons (1-StdDev=81.15 persons), and the RMSE “box” was located around the low values from 1 to 2 persons (Figure 1),
indicating an overall point-by-point accuracy of the model in most cities. Furthermore, as shown in Figure 1, under the condition of consistent coordinate-axis ranges, the “box” of adjusted $R^2$ values was slightly wider than $R^2$ and exhibited a slight sinking which makes the “box” nearer to the horizontal axis. The adjusted $R^2$, however, still had a median value of 0.9751 with only a few lower outliers. The proposed SEIR model was further confirmed to have a very good accuracy in modeling the COVID-19 outbreak in mainland China.

For the detailed modeling information in strata (cities), the temporal estimates of the cumulative cases can be plotted by stratum. Figure 2 demonstrates the information of the temporal estimates in Wuhan (the most severe city of the COVID-19 epidemic). Blue circles correspond to the cumulative cases, the solid red line denotes the estimate curve and the dashed lines depict the 95% confidence intervals (CIs). Note that clinically diagnosed cases were reported as confirmed ones on February 12th, 2020, leading to a very rapid increase of the numbers. The presented model was able to deal with such a sudden shift condition and still reached an overall satisfactory performance. See Figures S1 and S2 for more modeling results in severe cities inside/outside Hubei province. Note that another sudden shift condition (a community transmission in prison caused a very rapid increase of the numbers), occurred in Jining, Shandong province on February 20th, 2020. The model received a less satisfactory modeling performance to match the reported numbers in this city but can still in part reflect the temporal epidemiological characteristics.
Epidemiological parameter estimates

Table 3 shows the statistics of the epidemiological parameter estimates in strata (cities). The mean latent and removed periods, $T_l$ and $T_r$, of the COVID-19 epidemic in China were estimated to be 5.40 days (1-StdDev=1.29, 95% CI: 5.26–5.54) and 2.13 days (1-StdDev=0.31, 95% CI: 2.10–2.17), respectively. Most cities exhibited a clustering trend, with low 1-StdDev values of $T_l$ and $T_r$ estimates, as well as the latent and removed rates, $\lambda$ and $\gamma$. When considering the importations from Wuhan to other cities, the imported latent and infected subpopulations in strata at $t=1$ (January 20th, 2020) were also estimated, and the mean values of $\hat{E}_h(1)$ and $\hat{I}_h(1)$ were 32.18 persons (1-StdDev=157.75, 95% CI: 14.69–49.67) and 0.88 persons (1-StdDev=12.11, 95% CI: 0.00–2.22), respectively. Based on the cumulative number of Wuhan travelers to various cities until $t=1$, we further estimated the L&I ratios in strata at $t=1$. The mean $R_{LI,h}$ value was $1.72\times10^{-4}$ (1-StdDev=2.12×10^{-4}, 95% CI: 1.49×10^{-4}–1.96×10^{-4}). In other words, until January 20th, there were an average of 1.72 latent or infected persons per 10,000 Wuhan travelers to elsewhere.

As shown in Figure 3, the geographical distributions of the latent and removed periods in strata indicated a spatial homogeneity of the virus characteristics. We detected no obvious clustering trends of the latent and removed rates distributed in cities. However, the COVID-19 epidemic spread was strongly linked to human movements from the infection center, and the geographical distributions of the latent populations and L&I ratios from Wuhan at $t=1$ indicated a strong spatial heterogeneity. The latent populations and L&I ratios generally decreased by distance from the epidemic source.
(Figure 4). Several outliers appeared in the geographical distribution of the L&I ratios might be caused by the low-value cumulative number of Wuhan travelers in these cities.

**Space-time $R_0$ estimates**

Figure 5 depicts the geographical distributions of the $R_0$ estimates at $t=1$ (January 20th, 2020), $t=8$ (January 27th), $t=15$ (February 3rd), $t=22$ (February 10th), $t=29$ (February 17th), $t=36$ (February 24th), and $t=41$ (February 29th, the experimental end date), respectively. The experimental start date was set as January 20th (the next day after the first case was reported outside Hubei). Most cities had an initial $R_0$ value between 2 and 3.5 at $t=1$. After 2 weeks of the outbreak ($t=15$), the $R_0$ values in over 50% of cities had decreased to lower than 2, approximately 20% cities had $R_0$ values lower than 1.5, and the $R_0$ values in other severely affected cities ranged from 2 to 3. After 3 weeks of the outbreak ($t=22$), 21 out of 315 cities had $R_0$ estimates lower than 1 and only a few severely affected cities maintained $R_0$ values larger than 2. The number of cities with estimates of $R_0<1$ after 4 weeks of the outbreak ($t=29$) was 87, and the $R_0$ values in most cities decreased to lower than 1.5. At $t=36$ (February 24th), over 50% (184 out of 315) cities had estimates of $R_0<1$, and this number was 196 (over 60%) at $t=41$ (February 29th).

Temporal boxplots of infective rate ($\beta_h$) and $R_0$ estimates in the administrative cities of mainland China are demonstrated in Figure 6, respectively. They both indicated a mirrored “S” characteristic. The $\beta_h$ estimates started with an initial value of about 1.4 (i.e., an average of 1.4 infections per day per infectious individual), showed an obvious
descending trend after 1 week of the outbreak, and after a descending period of 2 weeks, maintained a relatively stable value of about 0.5. The spatial heterogeneity of the $\beta_h$ estimates in strata was more significant during the descending period ("boxes" are obviously narrower in stable periods). Similar spatial heterogeneity appeared as well in the $R_0$ estimates in strata. Additionally, an obvious descending trend appeared in temporal $R_0$ estimates, starting after 1 week of the outbreak and lasting 2 weeks to maintain a relatively stable value. The mean periods of $R_0$ estimates decreasing to 80% and 50% of the initial values in strata were 14.73 and 19.62 days, respectively. At the end of the experiment (February 29th), nearly all cities had $R_0$ estimates lower than 30% of the initial values.

As shown in Figure 7, the time evolution of the $R_0$ estimates in Hubei’s cities indicated several features (see Figure S3 for details). The initial $R_0$ values were between 2.6 and 3.2 (except Xiangyang). The start date of the obvious descending trend varied by city, in many cities this started after about 1 week into the outbreak, but the date could also be delayed by about 3 weeks (e.g., Wuhan and Tianmen). The descending periods also varied by city and most cities maintained a descending trend for about 2 weeks (Figures 7 and S3). However, the descending period could be shortened to about 1 week (e.g., Wuhan and Shennongjia) or expanded to about 4 weeks (e.g., Qianjiang). Figure S4 depicts the temporal curves of $R_0$ estimates in the 9 most severely affected cities outside Hubei. The curves indicate a general consistency around the start date and a consistency of the descending trend. The epidemic spread had a strong spatial heterogeneity in Hubei and the heterogeneity decreased outside Hubei (Figure 7). An
exception should be noted about the temporal curve of $R_0$ estimates in Jining, Shandong province. The specific community transmission in a prison on February 20th, 2020 caused high-value $R_0$ estimates before that date. However, after the reports of the community outbreak, the $R_0$ estimates were corrected by the modeling.

**Discussion**

Epidemiological parameters usually exhibit stratified heterogeneity in regions and in time, due to various reasons, e.g., the various temporal characteristics of the influenza-A 2009 pandemics in temperate zones, tropical zones and Pacific islands (McCallum and Partridge, 2010). A global model is inappropriate when epidemiological parameters vary in regions and by time. SSH can account for the universal features of phenomena (Wang et al., 2016), and is adequate to evaluate the local heterogeneity (non-homogeneity) in strata. Effective stratifications (e.g., zoning and discretization) can reduce the errors and improve the accuracy of modeling and estimation (Cao et al., 2013; Wang et al., 2010a, 2009). The COVID-19 outbreak in mainland China was obviously influenced by human movements from infection sources and presents a significant SSH in regions and by distance (in part caused by the much more significant SSHs of human movements from Wuhan to elsewhere). Therefore, based on the SSHs of the COVID-19 epidemic spread and human movements, modeling in strata are more appropriate to improve modeling of the spread of the epidemic in mainland China. Moreover, a comprehensive understanding of the epidemic in China will help inform the measures and strategies of prevention and control in other countries worldwide, many of which
are currently experiencing concurrent outbreaks.

In this study, we presented a SEIR model for each SSH-identified stratum to simulate a complete spread of the COVID-19 epidemic, and accurately estimate the spatial and spatiotemporal epidemiological parameters in mainland China. The model had an overall satisfactory performance with a median adjusted $R^2$ value of 0.9751 and the estimations had the consistency with most previous studies. We estimated that the mean latent and removed periods were 5.40 and 2.13 days, respectively, which are very close to the previous estimates (Guan et al., 2020; Li et al., 2020; Linton et al., 2020; Sun et al., 2020). Furthermore, an added value of this study is the estimations of the imported latent and infected subpopulations in strata from Wuhan on the initial experimental date (January 20th, 2020). The results showed the average imported latent and infected subpopulations from Wuhan to other cities were 32.18 and 0.88 persons, respectively, due to the annual Chinese New Year holiday migrations. Further estimation of the imported L&I ratios indicated that there was an average of 1.72 latent or infected persons per 10,000 Wuhan travelers to elsewhere until January 20th.

Another main contribution of the current study is the estimation of the space-time characteristics of $R_0$ during the COVID-19 epidemic spread in mainland China. To our knowledge, this is the first study to construct simultaneously time-varying and region-varying $R_0$ estimates of epidemic spreads. The results indicated that the initial $R_0$ values were between 2 and 3.5 in most cities on January 20th, 2020, which are very close to the previous estimates (Chen et al., 2020; Jung et al., 2020; Li et al., 2020; Wang et al., 2020; Wu et al., 2020; Zhang et al., 2020; Zhao et al., 2020b, 2020a). The $R_0$ values in
over 50% cities decreased lower than 2 after 2 weeks of the outbreak and only a few severely affected cities maintained $R_0$ values larger than 2 after 3 weeks of the outbreak. There were 87 cities with estimates of $R_0<1$ after 4 weeks of the outbreak (February 17th), and on February 29th, the number of cities was 196 (over 60% of all cities considered). The mean periods of $R_0$ estimates to decrease to 80% and 50% of the initial values in strata were 14.73 and 19.62 days, respectively. The temporal curves of $R_0$ estimates indicated stronger spatial heterogeneity in Hubei than the rest of mainland China. In most cities, an obvious descending trend in temporal $R_0$ estimates started 1 week into the outbreak and lasted 2 weeks to maintain a relatively stable value. The start date of the obvious descending trend of $R_0$ estimates in Hubei could be delayed about 2–3 weeks, and the descending period could be shortened to about 1 week or expanded to about 4 weeks.

To evaluate the SSHs of the COVID-19 epidemic spread and human movements from the infection sources, we comparatively analyzed three stratification solutions of mainland China (four distance-dependent subareas, provinces and administrative cities). Stratifications with both higher $q$ statistic values and interpretably practical implications are considered as better solutions to construct the stratified models. Therefore, we selected a relatively better stratification solution in this study, instead of the strictly “best” one (e.g., the one with the highest $q$ statistic value). This might be a limitation of the current study but remains a flexible extension of the proposed SEIR model and its applications for modeling the spread of other infectious diseases. A general and comprehensive solution to identify the stratification with higher $q$ statistic
value and practical implications is one of the main future works. The SEIR model for each SSH-identified stratum was verified to have a good regression accuracy with much smaller residuals in modeling the COVID-19 outbreak in mainland China. The localization of the SEIR model identified by SSH can bring out better modeling performance. The epidemic spreads in most cities had a relatively consistent trend with various spatiotemporal epidemiological characteristics, especially in the early-phase outbreak lasting two months. Further prospective applications request the cross validations for the proposed model. Moreover, during the modeling process, the modeling results are sensitive to the initial parameter values and their bounds. In this study, the initials and bounds of the parameter values were set up according to the observations and previous studies. It is important to implement the further analysis of the corresponding parameter sensitivities in future. Another future study direction is the further validation of the model universality to the COVID-19 epidemic in other large-scale areas (e.g., globally) and to other infectious diseases. We also intend to apply the proposed SEIR model in countries outside China and enhance the understanding of the COVID-19 epidemic worldwide.

Conclusions

In this paper, we introduce a SEIR model for each SSH-identified stratum based on the SSHs of the COVID-19 epidemic spread and human movements from infections sources. The proposed model estimates the spatial and spatiotemporal epidemiological parameters of the COVID-19 outbreak in the space-time domain, such as the
geographical distribution of L&I ratios and the space-time $R_0$ estimates. The following conclusions were achieved:

(1) The mean latent and removed periods of the COVID-19 epidemic were 5.40 days (95% CI: 5.26–5.54) and 2.13 days (95% CI: 2.10–2.17), respectively. The geographical distributions of these two epidemiological parameters indicated a spatial homogeneity amongst cities in mainland China.

(2) Due to the annual Chinese New Year holiday migrations, on January 20\textsuperscript{th}, 2020 (the next day after the first case was reported outside Hubei), the average imported latent and infected subpopulations travelling from Wuhan were estimated as 32.18 (95% CI: 14.69–49.67) and 0.88 (95% CI: 0.00–2.22) persons, respectively.

(3) There was an average of 1.72 (95% CI: 1.49–1.96) latent and infected persons per 10,000 Wuhan travelers to elsewhere until January 20\textsuperscript{th}. The geographical distributions of the imported latent and infected subpopulations amongst cities from Wuhan indicated a spatial heterogeneity decreasing by distance.

(4) The space-time $R_0$ estimates indicated an initial value between 2 and 3.5 in most cities on January 20\textsuperscript{th}. There were 87 cities which had an estimate of $R_0<1$ on February 17\textsuperscript{th} (four weeks after the large-scale outbreak), increasing to 196 cities by February 29\textsuperscript{th}. The mean period for $R_0$ estimates to decrease to 80% and 50% of the initial values in cities were an average of 14.73 and 19.62 days, respectively.

(5) An obvious descending trend in temporal $R_0$ estimates in most cities started 1 week after the outbreak and lasted 2 weeks, maintaining a relatively stable value. In the outbreak center (Hubei province, China), however, the start date could be delayed 2–3
weeks and the lasting period could be shortened to 1 week or expanded to 4 weeks.

In short, the findings will help enhance a comprehensive understanding of the epidemic spread in China and inform the strategies of prevention and control of the epidemic for policymakers. The proposed SEIR model can describe the complete spatiotemporal spread of the COVID-19 epidemic in China, conclude the space-time epidemiological characteristics and parameters, and is easily for the extension of the modeling applications in other countries worldwide.

Appendix A. Model performance indicators

Three indicators were selected to evaluate the performance of the proposed SEIR model in this study, including the root mean square error (RMSE), the coefficient of determination ($R^2$) and the adjusted $R^2$. In stratum $h$, the performance of the specific SEIR model was assessed by

$$RMSE_h = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (y_{h,t} - \hat{y}_{h,t})^2}$$  \hspace{1cm} (A1)

$$R^2_h = 1 - \frac{\sum_{t=1}^{T} (y_{h,t} - \hat{y}_{h,t})^2}{\sum_{t=1}^{T} (y_{h,t} - \bar{y}_{h,t})^2}$$  \hspace{1cm} (A2)

$$\tilde{R}^2_h = 1 - \frac{(T - 1) (1 - R^2_h)}{T - k - 1}$$  \hspace{1cm} (A3)

where $y_{h,t}$ and $\hat{y}_{h,t}$ are the observed and estimated (modeled) numbers of the COVID-19 cumulative cases, respectively, in stratum $h$ and at time $t$, and $\bar{y}_{h,t}$ is the mean of all the observed values. $T$ is the total number of considered timesteps ($T=41$ in this study). $\tilde{R}^2_h$ denotes the adjusted $R^2$ in stratum $h$. $k$ is the number of the independent regressors (estimated parameters) in the SEIR model. These measures can be separately
calculated by stratum and combined together to evaluate the overall performance of the proposed SEIR model.

Appendix B. Supplementary material

The Supplementary material to this article includes: Figure S1 (Estimates of the temporal cumulative cases in the 9 most severely affected cities of Hubei Province excluding Wuhan); Figure S2 (Estimates of the temporal cumulative cases in the 9 most severely affected cities of mainland China excluding Hubei); Figure S3 (Temporal $R_0$ estimates in Hubei’s cities); and Figure S4 (Temporal $R_0$ estimates in the 9 most severely affected cities of mainland China excluding Hubei).

Ethics approval and consent to participate

No individual data was collected and the ethical approval or individual consent was not applicable.

Availability of data and materials

The LBS-requesting mobile device data were provided by Wayz Inc., Shanghai, China and are not available for distribution due to the constraint in the consent. The dataset of the COVID-19 cases is available from multiple public sources. The MATLAB codes that support the proposed SEIR model computing in this study are available from the corresponding authors on reasonable request.
Funding

This work was supported by the National Natural Science Foundation of China (42061075 and 41531179), the National Science and Technology Major Project (2016YFC1302504) and the Science and Technology Major Project of Jiangxi Province, China (2020YBBGW0007).

Disclaimer

The funders had no role in study design and conduct; data collection, management, analysis and interpretation; manuscript preparation, writing and review; decision to submit the manuscript for publication.

Authors’ contributions

Conceptualization: Hui Lin, Jinfeng Wang and Bisong Hu; Resources and Data curation: Vincent Tao, Jingyu Qiu and Hui Lin; Formal analysis: Bisong Hu, Pan Ning, Adam Thomas Devlin and Haiying Chen; Methodology: Bisong Hu and Jinfeng Wang; Software: Jinfeng Wang; Visualization: Bisong Hu, Pan Ning, Adam Thomas Devlin and Haiying Chen; Validation, Bisong Hu and Jinfeng Wang; Writing—original draft: Bisong Hu; Writing—review and editing: Bisong Hu, Jinfeng Wang and Hui Lin; all the other authors contributed to the writing. All authors have read and agreed to the published version of the manuscript, and have final responsibility for the decision to submit for publication.
Conflict of interests

We declare no competing interests.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to acknowledge the support for the data collection from Wayz Inc., Shanghai, China. The authors would like to appreciate the relevant government departments in China and multiple public sources for providing the open COVID-19 datasets, the socioeconomic statistical data and the geographical datasets. The authors are very grateful to the editors and reviewers for their valuable comments.

References


Boldog P, Tekeli T, Vizi Z, Dénes A, Bartha FA, Röst G. Risk Assessment of Novel


Peeri NC, Shrestha N, Rahman MS, Zaki R, Tan Z, Bibi S, et al. The SARS, MERS and


Wang J, Haining R, Cao Z. Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. Int J Geogr Inf Sci


Zhang S, Diao M, Yu W, Pei L, Lin Z, Chen D. Estimation of the reproductive number...


Figure 1. Boxplots of the accuracy indicators found from modeling the COVID-19 spread in administrative cities from January 20th to February 29th, 2020: (A) RMSE; (B) $R^2$; (C) adjusted $R^2$. Note that several very large scatters in the RMSE boxplot are not shown for conciseness.
**Figure 2.** Estimates of the temporal cumulative cases in Wuhan. Day 1 is January 20\textsuperscript{th}, 2020.

**Figure 3.** Geographical distributions of the epidemiological period estimates: (A) latent periods (days); (B) removed periods (days).

**Figure 4.** Geographical distributions of the estimates of the imported (A) latent
subpopulations (persons) and (B) L&I ratios (1/10,000) from Wuhan at $t=1$ (January 20th, 2020).
Figure 5. Geographical distribution of $R_0$ estimates in the administrative cities of mainland China, at: (A) $t=1$ (January 20th, 2020); (B) $t=8$ (January 27th, 2020); (C) $t=15$ (February 3rd); (D) $t=22$ (February 10th, 2020); (E) $t=29$ (February 17th); (F) $t=36$ (February 24th, 2020); (G) $t=41$ (February 29th, the end of the experimental time period).
Figure 6. Temporal boxplots of the epidemiological characteristics: (A) $\beta_h$ estimates; (B) $R_0$ estimates. Day 1 is January 20th, 2020.
**Figure 7.** Time evolution of $R_0$ estimates in Wuhan, Hubei excluding Wuhan, provinces adjacent to Hubei and the rest of mainland China.

**Table 1.** $q$ statistic values of COVID-19 cases and device traces with various stratification solutions$^1$.

<table>
<thead>
<tr>
<th>Stratification</th>
<th>Cases$^2$</th>
<th>Market devices$^3$</th>
<th>Wuhan devices</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st</td>
<td>1st</td>
<td></td>
</tr>
<tr>
<td>Solution 1</td>
<td>0.2682</td>
<td>0.7155</td>
<td>0.9524</td>
</tr>
<tr>
<td>Solution 2</td>
<td>0.0677</td>
<td>0.0598</td>
<td>0.0865</td>
</tr>
<tr>
<td>Solution 3</td>
<td>0.2742</td>
<td>0.7158</td>
<td>0.9530</td>
</tr>
</tbody>
</table>

$^1$ For all the $q$ statistic values, $p<0.001$. $^2$ The COVID-19 cases were cumulatively...
summed until February 29th, 2020.³ Mobile device data from the seafood market was divided into two subsets separated by the date of January 1st, 2020, for the $q$ statistic calculations.

### Table 2. Performance of the SEIR model for a stratum.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Mean</th>
<th>Min.</th>
<th>Max.</th>
<th>StdDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>8.9190</td>
<td>0.1659</td>
<td>1423.55</td>
<td>81.15</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.9663</td>
<td>0.7256</td>
<td>0.9985</td>
<td>0.0396</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.9592</td>
<td>0.6674</td>
<td>0.9982</td>
<td>0.0480</td>
</tr>
</tbody>
</table>

### Table 3. Estimates of the epidemiological parameters of the SEIR model for a stratum.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Mean</th>
<th>Min.</th>
<th>Max.</th>
<th>StdDev</th>
<th>95% CIs</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$ (day⁻¹)</td>
<td>0.19</td>
<td>0.07</td>
<td>0.30</td>
<td>0.03</td>
<td>[0.1887, 0.1959]</td>
</tr>
<tr>
<td>$T_l$ (days)</td>
<td>5.40</td>
<td>3.32</td>
<td>14.00</td>
<td>1.29</td>
<td>[5.26, 5.54]</td>
</tr>
<tr>
<td>$\gamma$ (day⁻¹)</td>
<td>0.48</td>
<td>0.19</td>
<td>0.71</td>
<td>0.05</td>
<td>[0.4703, 0.4820]</td>
</tr>
<tr>
<td>$T_r$ (days)</td>
<td>2.13</td>
<td>1.40</td>
<td>5.32</td>
<td>0.31</td>
<td>[2.10, 2.17]</td>
</tr>
<tr>
<td>$\hat{E}_h$ (1) (persons)</td>
<td>32.18</td>
<td>0.00</td>
<td>2452.56</td>
<td>157.75</td>
<td>[14.69, 49.67]</td>
</tr>
<tr>
<td>$\hat{I}_h$ (1) (persons)</td>
<td>0.88</td>
<td>0.00</td>
<td>214.14</td>
<td>12.11</td>
<td>[0.00, 2.22]</td>
</tr>
<tr>
<td>$R_{LLH}$ ($\times10^{-4}$)</td>
<td>1.72</td>
<td>0.03</td>
<td>20.34</td>
<td>2.12</td>
<td>[1.49, 1.96]</td>
</tr>
</tbody>
</table>