Predicting new cases of COVID-19 and the application to population sustainability analysis

Chengcheng Bei\textsuperscript{a}, Shiping Liu\textsuperscript{b}, Yin Liao\textsuperscript{c}, Gaoliang Tian\textsuperscript{a}, Zichen Tian\textsuperscript{d}

\textsuperscript{a}School of Management, Xi’an Jiaotong University, Xi’an, China
\textsuperscript{b}GBICC Global Business Intelligence Consulting Corp., Beijing, China
\textsuperscript{c}Department of Applied Finance, Macquarie University, Sydney, NSW, Australia
\textsuperscript{d}Department of Economics, University of Wisconsin–Madison, Madison, WI, USA

Abstract

We propose a new spatio-temporal point process model to predict infectious cases of COVID-19. We illustrate its practical use with data from six key cities in China, and we analyse the effects of natural and social factors on the occurrence and spread of COVID-19. We show that large-scale testing and strict containment are key factors for the successful suppression of the COVID-19 contagion. This study provides an effective tool to develop early warning systems for major infectious diseases, offering insights on how to develop prevention and control strategies to reduce the impact of disease and maintain population sustainability.

Key words: Control measures; COVID-19; Infectious cases; Spatio-temporal point process; Sustainability

JEL classification: D12, D83, D84, E32

doi: 10.1111/acfi.12785

1. Introduction

The impact of the COVID-19 pandemic on the global economy is unprecedented since the Great Depression in the 1930s and is inducing significant environmental changes to society. Meanwhile, COVID-19 highlights how vulnerable nations are to a sudden crisis, and brings a new focus to the sustainability of countries (Ahmed \textit{et al.}, 2018; Chan \textit{et al.}, 2020; Fong \textit{et al.}, 2020; Quilty \textit{et al.}, 2020). Thus, understanding the origin and transmission of
the COVID-19 virus is strikingly important, and the effective responses and actions towards the global crisis provide opportunities to reset and reshape the world to be more sustainable. Building upon epidemiological theories, we propose a new spatio-temporal point process model to capture the spatio-temporal emergence and spread of the COVID-19 virus. We employ big data and machine learning to analyse the transmission mechanism of COVID-19 in major cities in China (e.g., Beijing, Chongqing, Guangzhou, Shanghai, Shenzhen and Wuhan) and the effects of social and natural factors on the occurrence and development of the pandemic.

There are many mathematical modelling studies on epidemics in the literature. Among these models, the system dynamics model is widely used (Ahmed et al., 2018; Chan et al., 2020; Fong et al., 2020; Quilty et al., 2020). In system dynamics, the susceptible-infectious-removed (SIR) model is used to analyse the variation patterns in the number of infected people according to general infectious disease transmission mechanisms. It uses quantitative relationships to describe the transmission process of an infectious disease to reveal the progression trends. In this model, the population is divided into three categories, susceptible (S), infective (I) and removed (R). There are three basic assumptions in the model: (i) the total population in the study does not change with time and the natural birth rate and mortality rate are not considered; (ii) susceptible people are affected by the infectious disease and the change in their number is directly proportional to the number of infected people; and (iii) the growth rate of the number of quarantined people and removed people is directly proportional to the number of infected people (Ferguson et al., 2003; Riley et al., 2003; Klepac et al., 2009; Kucharski et al., 2020). The general equations for the system dynamics-based SIR models are very mature. However, the solutions to these models are simplified approximation solutions of the Taylor expansion and the physical significance of the parameters in the simplification process is unclear. Infectious disease transmission routes have their own characteristics. Hence, when the SIR model is used to simulate the transmission mechanisms of different infectious diseases, the model parameters are different. The model parameters for even the same type of infectious disease can also dynamically change as healthcare levels improve and governments implement prevention and control measures. Therefore, some parameters are greatly affected by subjective factors and empirical factors when the SIR model is used to identify the progression trends of an infectious disease.

Another modelling method for epidemics is the time series model (Mossong et al., 2008; Prem et al., 2017; Zhang et al., 2019; Riou and Althaus, 2020; Tian et al., 2020; Wu et al., 2020). The time series model reflects temporal dynamic dependence relationships and can identify variation patterns in quantitative relationships in an epidemic over time. The Box–Jenkins (BJ) model is a representative time series analysis and prediction method that is composed of four basic models: the autoregressive model, the moving average model, the autoregressive moving average (ARMA) model and the autoregressive
integrated moving average (ARIMA) model. The ARIMA model is applied in some cases where the data show evidence of non-stationary time series and is more widely used than other models. The time series model contains corresponding prediction methods for non-seasonal diseases, seasonal diseases and diseases with a cyclical pattern (linear, nonlinear or multiple curve-fitting models are generally used when there is no obvious seasonality in the time series; seasonal cycle analysis is usually employed if there is obvious seasonality; an analysis of variance (ANOVA) filter is mostly employed if the time series shows cyclical fluctuations).

There are still some challenges in infectious disease transmission simulation studies. Firstly, the temporal dynamics model for infectious disease spread is not synchronised with the spatial dynamics model. For example, the SIR model focuses on temporal changes but overlooks spatial processes. Secondly, there is a problem when combining infectious disease intensity predictions with transmission region predictions. The SIR and ARIMA models are strong predictors of the extent of an epidemic but are not good at determining the correlation of disease frequency. Thirdly, these models focus mainly on the utilisation of epidemic data but ignore the effects and predictive capabilities of external elements, such as social factors (e.g., population density, healthcare conditions and government prevention and control policies) and natural factors (e.g., wind speed and air quality) on disease transmission. The availability of big data provides opportunities for us to obtain multi-dimensional pandemic-related data. We use these data to validate our new model, which includes these influencing factors. Machine learning enables our quantitative model to select valid high-dimensional variables, thereby facilitating their combination with factors other than disease-related characteristics to provide a new tool for constructing an infectious disease model and pandemic predictions. In addition, it is difficult to incorporate variables that reflect external factors in models, such as the SIR model. Therefore, these models are not appropriate to examine the effects of external factors on COVID-19 transmission.

China adopted unprecedented control and quarantine measures to prevent sustained transmission and spread of the disease during the COVID-19 pandemic (Abbott et al., 2020; Backer et al., 2020; Brooks et al., 2020; Kraemer et al., 2020). Therefore, we combine the model we proposed here with the SIR model in epidemiology to analyse the development of the pandemic when no measures are taken and when varying degrees of prevention and control measures are taken in China, including the proportion of infected people and the time for the entire population to become infected. Specifically, we propose a new spatio-temporal point process-based mathematical model to simulate new confirmed COVID-19 cases each day. The model can be used as an early warning tool for COVID-19 detection, as well as other infectious disease outbreaks.

Firstly, our model focuses on simultaneous simulation of the frequency and extent of the pandemic. As a marked point process, there are two aspects to
examine the number of confirmed cases of COVID-19: (i) the occurrence rate (i.e., the number of confirmed cases per spatio-temporal unit within a fixed spatio-temporal range), which can also be interpreted as the possibility of confirmed cases; and (ii) the extent of the pandemic (i.e., the number of confirmed cases within a fixed spatio-temporal range). In our model, the number of confirmed cases is considered to be a Hawkes process and linear function that is dependent on previous spatio-temporal distribution with temporal autocorrelation and spatial cross-correlation. At the same time, the model considers the interactions between these two functions and provides a new method for the early warning of infectious disease outbreaks from different perspectives, including the spatio-temporal point process, occurrence rate and extent. Mainstream models are typically focused on one of these aspects. For example, the ARIMA and SIR models focus mainly on the extent of an outbreak (i.e., confirmed cases) while the Poisson regression and logistic growth models focus mainly on the occurrence rate. In Poisson regression and logistic regression models used to determine the probability of infectious disease outbreaks, the occurrence rate is often assumed to be a time-invariant function. In contrast, our model focuses on the spatio-temporal transmission of infectious disease outbreaks to capture its evolution dynamics.

Secondly, we combine temporal and spatial simulation in our model. Early warning models of infectious disease focus mainly on using the temporal dimension to detect and analyse abnormalities in disease monitoring data. These models focus mainly on the temporal distribution or the various characteristics of monitoring markers in a specific region to reflect whether the probability of infectious disease occurrence has significantly increased or whether clusters have occurred in a specific time period. Examples include the autoregressive integrated moving average (ARIMA) model that uses a random time series, exponentially weighted moving average (EWMA) model, Poisson regression and logistic regression. Today, many early warning models include spatial dimensions. Spatial early warning models focus on monitoring the spatial distribution of indicators at one or more time points and observe whether there is any significant spatial aggregation. One early warning spatial model that has attracted attention is the space scan statistic proposed by Kulldorff (1997). In that model, geographical space is divided into small regions and those with statistically significant differences in disease occurrence when compared with normal levels are identified, and used to detect spatial aggregation. The spatio-temporal early warning model we propose simultaneously focuses on changes in the temporal and spatial dimensions of the pandemic so that the accuracy of the early warning signal is further improved. In China, early warning models that are widely used include the space-time scan statistic and prospective space-time permutation scan statistic proposed by Kulldorff et al. (2005). In contrast, our model provides a better simulation of the spatio-temporal dynamic changes and transmission routes of the pandemic, thereby improving the timeliness and accuracy of early warnings.

© 2021 Accounting and Finance Association of Australia and New Zealand
Thirdly, we employ machine learning theory so our model can reflect the temporal variability of viral transmission. Generally, large-scale infectious disease outbreaks can be classified into several stages: outbreak growth phase, flattening phase and decline phase. At various phases of a pandemic, changes in transmission speed, mechanisms and influencing factors occur, which is known as the temporal variability of disease spread. We use the results of a study by Imperial College London and a study on COVID-19 trends by Shanghai Jiao Tong University School of Medicine for phasic modelling: In stage 1 (11 January 2020 to 23 January 2020), Wuhan did not carry out stringent prevention and control measures and only case reporting was conducted. On 23 January, a lockdown was announced. In stage 2 (24 January onwards), which was the first day of lockdown in Wuhan, traffic control and quarantine systems were gradually established to limit population movement. In stage 3 (2 February to 24 February), community prevention and control were strengthened in Wuhan and the “no omission of households and individuals” community public health management model was proposed. At the same time, Huo Shen Shan, Lei Shen Shan, and cabin hospitals were successively constructed and activated. Large tertiary general hospitals from 20 provinces sent medical assistance teams to Wuhan, which further improved the prevention and control and medical levels in Wuhan (Wang et al., 2020).

Fourthly, our model is designed to better utilise big data sets from multiple sources. Conventional early warning models focus mainly on infectious disease transmission data for modelling and analysis. Such data include confirmed cases, suspected cases and close contacts. These models usually ignore other relevant data, such as social and environmental factors. The enrichment of infectious disease early warning data sources has resulted in the development of early warning technology based on multi-source data. The core of this technology is to use the interactions and restrictions between multiple factors to analyse the various patterns of events, thereby identifying abnormalities in infectious disease occurrence. We can incorporate other factors into conditional frequency and intensity functions for infectious disease occurrence outliers to successfully incorporate social factors (e.g., population density, population migration) and natural factors (e.g., air quality, wind speed) into our early warning model.

We use our model to fit recent COVID-19 transmission in Wuhan, the origin of the pandemic, and five other cities in China (Beijing, Shanghai, Guangzhou, Shenzhen and Chongqing) to validate the accuracy of model predictions. Compared with conventional models, we find that our model demonstrates better prediction capabilities, particularly at the outbreak growth phase. As the model simulates the pandemic in two dimensions (temporal and spatial) and considers the effects of external factors, it greatly mitigates the underestimation of the pandemic in these phases by conventional models. At the same time, we use our model to examine how the prevention and control measures employed by the Chinese government affect the spread of COVID-19 in China. After the
outbreak of COVID-19, the Central Committee of the Communist Party of China and the State Council of the People’s Republic of China activated national emergency responses. The results show that the above measures effectively suppress the rapid spread of COVID-19, and eliminate the risk of rapid transmission.

Fifthly, our analysis of the initial phase of the pandemic in China and its current status in the world finds that information asymmetry, untimely data and insufficient risk prediction are some of the factors that limit effective control of the disease. Therefore, high-quality application of data and accurate disease transmission prediction and early warning are prerequisites for ensuring effective pandemic decisions, reducing disease spread and safeguarding lives, and are also key to formulating effective governance systems to achieve coordinated social and economic development. Our model is an effective tool that can be employed for early warning and timely prevention and control of major infectious disease outbreaks. It is useful for the quantitative analysis of the effects of infectious disease prevention and control measures on the economy in the post-COVID-19 era.

Finally, there are signs that the impacts of COVID-19 on the global economy will be more intense and long-lasting than those felt during the 2008–2009 global financial crisis. These impacts pose a serious threat to the development prospects of less industrialised nations, and to the realisation of the United Nations’ sustainable development goals (SDGs) by 2030. This work develops a new toolkit to track the origin and transmission of the pandemic that should be able to help with reducing the economic impact of the pandemic on achieving the UN’s SDGs.

2. Data

2.1. Sample selection and data sources

In this study, provincial capitals and sub-provincial cities are used as study samples to test the model. The period we study is from 11 January 2020 to 24 February 2020. We remove samples with missing variables to obtain daily observation values for 881 cities. We divide the data into three parts: Part 1 includes pandemic-related data, which we obtain from the official websites of health bureaus in various cities in China, including official websites, official Sina Weibo accounts and official WeChat public accounts. Part 2 consists of data on natural factors that may affect disease spread. We obtain city-level atmospheric data from the China Meteorological Administration’s website, and air quality data from the China National Environmental Monitoring Center. Geographical data are from Baidu Maps. Part 3 consists of data on social factors that may affect disease spread. Among these factors, population migration data are from the Baidu Qianxi website, macroeconomic and healthcare condition data are from the Wind database, and government
efficiency data are from the “2018 Research Report of Local Governments’ Efficiency in China.” Here, we only report results for several main cities, including Wuhan, which is the origin of the COVID-19 in China, and five other major cities (e.g., municipalities) in China. The results for other cities are available upon request.

2.2. Description of variables

The disease-related variables include the cumulative number of confirmed cases (Confirm), cumulative death toll (Dead) and cumulative number of people who recovered (Heal). The natural factors include wind speed (WindSpeed), which is the diurnal wind speed in the city; temperature (Temperature) is the maximum diurnal temperature in the city; altitude (Altitude) is the mean altitude of the city; linear distance (Distance) is the distance of the city centre from the city centre of Wuhan; and road distance (RoadDistance) is the road distance of the city centre from the city centre of Wuhan. The social factors include immigration rate (ImmigrationRate), which is the number of people who left Wuhan and went to that city over the total number of people who left Wuhan from 10 January to 24 January. Cities that are not among the top 100 cities are represented by data from the 100th city. Data in 2018 are used for control variables, including population of the city (Population), population density (PopulationDensity), gross domestic product (GDP), number of hospitals (NumberofHospitals), number of medical staff (NumberofMedical) and number of hospital beds (NumberofBeds). We used the 2018 rather than 2019 data as controls because our COVID-19 infected cases data is from the beginning of 2019. The data from 2018 better reflects the economic or demographic conditions during that time. Also, most of the macroeconomic data for 2019 were not released at the beginning of the year. The municipal government efficiency (CityEfficiency) is an integrated efficiency value of municipal governments in the above reports. The data from four direct-controlled municipalities are integrated efficiency values in the provincial government. As the report only shows the data of the top 100 cities, cities that are not among the top 100 are represented by data from the 100th city. Details are in Table 1. Table 2 shows the descriptive statistics of the major variables.

3. Theoretical model

3.1. Spatio-temporal point process model

We use a marked point process to simulate the number of new confirmed COVID-19 cases at a specific time point and in a specific region. The marked point process model is a popular framework for analysing statistical events. The model is frequently employed to simulate earthquakes or extreme events in financial markets. The core of the marked point process model is to use the
intensity of an event occurrence per unit time to predict the future rate of occurrence. Specifically, in our spatio-temporal point process model framework, we use \( x = \{ x_i = (t_i, s_i) \}_{i=1}^N \) to define new confirmed cases at a time \( t \) and in geographical space \( \{ s \}_{i=1}^N \). In an information set \( H_t = \{ x_1, \ldots, x_{t-1} \} \) of all events that occur before time \( t \), the frequency of new confirmed cases at time \( t \) and geographical space \( s \) is defined as:

\[
\lambda(t, s) = \lim_{|dt| \to 0, |ds| \to 0} \frac{E[N(dt \times ds)]}{|dt||ds|},
\]

where \( |dt| \) denotes a very small time interval around time \( t \) and \( |ds| \) denotes a small geographical region around the geographical position \( s \). \( N(dt \times ds) \)

### Table 1

| Variable category | Variable name | Unit | Definition |
|-------------------|---------------|------|------------|
| Epidemic-related  | Confirm       | 1    | Cumulative number of confirmed cases |
|                   | Dead          | 1    | Cumulative death toll |
|                   | Heal          | 1    | Cumulative number of patients who recovered |
| Natural factor    | WindSpeed     | Grade | Wind speed, represented by diurnal wind speed in the city |
|                   | Temperature   | ° C  | Temperature, represented by maximum diurnal temperature in the city |
|                   | AQI           |      | Air quality index |
|                   | Altitude      | m    | Mean altitude of city |
|                   | Distance      | km   | Linear distance of city centre from the city centre of Wuhan |
| Social factor     | ImmigrationRate|      | Proportion of people who migrated out of Wuhan to that city over the total number of people who migrated out of Wuhan from 10 January to 24 January. Cities that are not among the top 100 cities are represented by data from the 100th city. |
|                   | Population    | 10,000 people | Total population |
|                   | Population Density | People/km² | Population density |
|                   | GDP           | 100 million RMB | Gross domestic product |
|                   | Number of Hospitals | Hospitals | Total number of hospitals |
|                   | Number of Medical | 10,000 people | Total number of medical staff |
|                   | Number of Beds | 10,000 beds | Total number of beds |
|                   | City Efficiency |      | City government efficiency. Cities that are not among the top 100 cities are represented by data from the 100th city. |
denotes the region with $N$ confirmed cases in the time range of $|dt|$ and geographical area of $|ds|$. Simulation of dynamic changes in $\lambda(t,s)$ can effectively predict the spatio-temporal points where new confirmed cases will appear. When external factors are not considered, the spatio-temporal point process is often considered to follow the spatial-temporal Hawkes process (Hawkes, 1971). In a time series, the frequency of occurrence of new confirmed cases will increase with the increase in the number of new confirmed cases in the past, particularly the most recent cases, which is a self-exciting process. Spatially, the occurrence of new confirmed cases in a region will accelerate the frequency of the occurrence of new confirmed cases in neighbouring regions (i.e., mutually exciting). Therefore, the frequency of the occurrence of new confirmed cases can be expressed as:

$$\lambda(t,s|H_t) = \mu + \int_S \int_0^t g(s-s',t-t')dN(s',t'),$$

where $\mu$ is the unconditional rate of occurrence of confirmed cases and $g(s-s',t-t')$ denotes the spatio-temporal excitation function of the occurrence of confirmed cases.

In addition, the frequency and intensity of confirmed cases interact with each other (i.e., frequency increases as the number of confirmed cases increases, while intensity also increases as frequency increases). At the same time, many external factors will affect the frequency of disease occurrence, such as

| Variable       | Observed values | Mean       | Standard deviation | Minimum | First quartile | Median | Third quartile | Maximum |
|----------------|-----------------|------------|--------------------|---------|---------------|--------|---------------|---------|
| Confirm        | 881             | 107.7367   | 116.4493           | 1       | 28            | 53     | 155           | 572     |
| Dead           | 881             | 0.4892     | 1.0214             | 0       | 0             | 0      | 1             | 6       |
| Heal           | 881             | 30.2656    | 50.0655            | 0       | 2             | 10     | 34            | 316     |
| WindSpeed      | 881             | 2.5227     | 1.3372             | 1.5     | 1.5           | 1.5    | 3.5           | 6.5     |
| Temperature    | 881             | 10.2054    | 8.1784             | -18     | 6             | 11     | 16            | 31      |
| AQI            | 881             | 69.3848    | 30.2163            | 0       | 49            | 63     | 84            | 250     |
| Altitude       | 881             | 336.9527   | 561.8478           | 3.3     | 29.8          | 72.2   | 259.1         | 2261.2  |
| Distance       | 881             | 950.4658   | 483.3611           | 256.4   | 695.7         | 837.1  | 1146.4        | 2765.8  |
| RoadDistance   | 881             | 1113.7289  | 564.1278           | 346.3   | 829.9         | 987.6  | 1339.1        | 3311.5  |
| ImmigrationRate| 881             | 0.0033     | 0.0031             | 0.0009  | 0.0011        | 0.0019 | 0.005         | 0.0127  |
| Population     | 881             | 1011.7417  | 674.3247           | 225.06  | 595.2         | 820.2  | 1302.66       | 3124.32 |
| PopulationDensity| 881   | 1139.8952  | 1340.0919          | 179.1902| 388.4721      | 728.2702| 1280.735      | 6521.55 |
| GDP            | 881             | 1.1772     | 0.9963             | 0.1286  | 0.5207        | 0.7857 | 1.4104        | 3.8155  |
| NumberofHospitals| 881     | 306.3598   | 200.2954           | 54      | 162           | 269    | 349           | 892     |
| NumberofMedical| 881            | 9.085      | 6.2774             | 2.11    | 4.49          | 7.74   | 10.44         | 28.2    |
| NumberofBeds   | 881             | 6.5408     | 4.5873             | 1.03    | 3.7455        | 5.3    | 7.52          | 22.04   |
| CityEfficiency | 881             | 0.3655     | 0.4651             | 0.0972  | 0.0972        | 0.2109 | 0.4043        | 2.3575  |
population density, healthcare conditions and air quality. Lastly, the effects of
the various factors on the frequency of disease outliers show temporal
variability. For example, at the outbreak growth phase, the number of infected
people continuously increases and the probability of spatio-temporal disease
transmission also increases greatly. As disease transmission peaks, some
prevention and control measures, such as quarantining and the development,
production and usage of vaccines, will greatly decrease the probability of
transmission. Therefore, the following equation is based on the standard
Hawkes spatio-temporal frequency function:

$$\lambda(t, s | H_t) = \mu + \int_0^t \int K(\lambda_{t,s}) C(v) g(s - s', t - t') \rho(m|s-s', t-t') dN(s', t').$$

(3)

In the equation, $\mu > 0$ is the occurrence rate of disease outliers, which is
the unconditional probability of outliers when all influencing factors are not
considered. $g(s - s', t - t')$ is the spatio-temporal trigger function of disease
occurrence. We use this to describe the temporal autocorrelation and spatial
cross-correlation of outliers. Overall, $g(s - s', t - t')$ must be a non-negative
function. Therefore, we employ the commonly used form of $g(s - s', t - t') > 0$ in
the literature and split $g(s - s', t - t')$ into the product of two equations:
$g(s - s', t - t') = f(s - s') h(t - t')$. We also use the power law attenuation
function and the kernel function to describe $h(t - t')$ and $f(s - s')$:

$$h(t - t') = \int_0^t \omega(t - t') dN(t') = \sum_{t' < t} e^{-\delta(t-t')},$$

(4)

$$f(t - t') = \int_0^s \kappa(s - s') dN(s') = \sum_{s' < s} e^{-\rho \sqrt{(s-s')^2}},$$

(5)

where $\omega(\cdot)$ is a decreasing function that changes with time interval $t - t'$. This
function ensures the historical events in the same region affect the probability
density function of future events but these effects will decrease with time. $\kappa(\cdot)$ is
a decreasing function that changes with geographical distance. This function
ensures that events in other regions affect the probability density function of
events in a region at the same time and these effects will decrease as
geographical distance increases. $\rho(m_j|s-s', t-t') = \gamma_0 + \gamma_1 m_j$ is a function of
the effects of the extent of the pandemic on $m_j$. Specifically, if the occurrence of
confirmed cases is defined to be a marked point process, a characteristic can be
observed in addition to the time and location, which is the intensity of
occurrence outliers ($m$) (i.e., number of confirmed cases). In our model, the
frequency of occurrence will increase as the intensity of outlier occurrences
increases. At the same time, the degree of outlier occurrences will change with
spatio-temporal distribution. Therefore, we use $\rho(m|s-s',t-t')$ to describe the spatio-temporal distribution of outlier occurrence intensities. We also add this to $\lambda(t,s|H_t)$ to describe the effects of the extent of disease spread on the frequency of occurrence. $(v) = \beta'v$ is a linear function about the variable set $v$, and $\beta$ is the corresponding coefficient. We use this function to describe the effects of external factors (e.g., natural and social), on disease occurrence frequency. In a region $s$, the frequency of occurrence of new confirmed cases at time $t$ is expressed as:

$$
\lambda(t,s|H_t) = \mu_{t,s} + \beta'v + \sum_{t'<t} \gamma_{0,t} + \gamma_{1,t}m' + \sum_{t'<t} e^{-\delta_{t-t'}} + \sum_{s'<s} e^{-\rho_{t}(s-s')^2}
$$

\[6\]

3.2. Model estimation

As our spatio-temporal point process model with external factors has high-dimensional attributes, we employ regularised maximum likelihood estimation (MLE) to estimate the parameters. The parameter set is $\Theta = \{\mu_{t,s}, \beta'v, \gamma_{0,t}, \gamma_{1,t}, \delta_{t}, \rho_{t}\}$, the logarithmic likelihood function corresponding to the time window is $[0,T]$, and $N$ regions in Equation (4) are defined as:

$$
\log(L(\Theta)) = \sum_{i=1}^{N} \log(\lambda_i) - \sum_{u=1}^{u_{x}} \int_{0}^{T} \lambda_u dt ds.
$$

\[7\]

We introduce two aspects of regularisation to reduce the dimensionality of parameters and discover temporal changes in disease spread in a timely manner:

$$
\log(L(\Theta)) = \sum_{i=1}^{N} \log(\lambda_i) - \sum_{u=1}^{u_{x}} \int_{0}^{T} \lambda_u dt ds + \phi \| \Theta \|, \tag{8}
$$

where we include $\phi \| \Theta \|$ to normalise parameters. This means that the variable does not significantly affect the frequency of disease occurrence at a certain time corresponding to that parameter. This normalisation can effectively decrease the number of parameter dimensions in the model. We also attempted to employ principal component analysis to reduce the dimensionality of extrinsic variables. We generate a principal component to include all variable information for every external element. In this way, the external variables can be represented by two principal components: social and natural.
3.3. Model predictions

Our model is designed to predict two aspects of disease occurrence: (1) the spatio-temporal distribution of confirmed cases; and (2) the regions and times of occurrence of confirmed cases. Specifically, we specify a future time interval $[T, T + \Delta T]$. We combine simulation with parameter estimation to obtain the frequency and intensity distribution of new confirmed cases in various spatial regions during the future time interval. We use the means of these frequencies and intensities to predict disease occurrence frequency and intensity. Specifically, we assume that the number of new cases in various regions is observed in the past (until time $T$), $\{X_1, \ldots, X_{T-1}\}$, and $X_t = \{x_i = (t_i, s_i)\}_{i=1}^N$. We use the following steps to simulate the trajectory of new future cases, thereby predicting the spatio-temporal point, frequency and intensity of its occurrence.

Firstly, we combine the estimated parameters and use Equation (6) to simulate the frequency ($\lambda_i, i = 1, \ldots, N$) of new cases in various regions in the future $[T, T + \Delta T]$.

Secondly, we predict the spatio-temporal point of disease spread. We repeat similar simulations 1,000 times to obtain the mean frequency of disease occurrence. We use the following equation to obtain the expected mean number of confirmed cases ($N_i(t, s)$) in region $i$ at a future point in time:

$$N_i(t, s) = \int_{t \times s} \lambda_i \, dt \, ds.$$  \hspace{1cm} (9)

In summary, our model uses the algorithm in Table 3 to generate the expected number of new cases (i.e., the mean values of predicted new cases in various regions at time $[T, T+\Delta T]$).

3.4. Prediction results

3.4.1. Prediction evaluation standards

In order to validate the predictive capabilities of our spatio-temporal point process model, we compare our model with common epidemiological models in the literature, including the ARIMA model (Unkel et al., 2012), the logistic growth model (O’Hara and Kotze, 2010), and the ordinary differential equation SIR model (Koelle and Pascual, 2004; Koelle et al., 2005).

**ARIMA model.** The autoregressive integrated moving average (ARIMA) model ($p, d, q$), where $p$ is the autoregression item, $q$ is the moving average item and $d$ is the number of differencing transformations, is used to obtain a stationary time series. The basic concept of the model is that the data series generated from the predicted number of new confirmed cases with time is considered to be a random series. The model considers time series dependence.
and the interference due to random fluctuations. We use the Akaike information criterion (AIC) and Bayesian information criterion (BIC) models to select the optimal $p$, $q$ and $d$ values. When $d = 1$, the model is generally expressed as:

$$
(N_t - N_{t-1})^d = \varphi_1 N_{t-1} + \ldots + \varphi_p N_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_p \varepsilon_{t-q}. \quad (10)
$$

**Logistic growth model.** We modify the logistic growth model based on the Malthusian population model in which the population growth rate is set as $r(1 - \frac{P}{K})$ and $K$ is defined as the maximum population permitted by the environment. When population $P$ approaches $K$, the growth rate decreases (i.e., the linearity of the population growth rate decreases as the population increases).

$$
\frac{dP}{dt} = r \left(1 - \frac{P}{K}\right) P. \quad (11)
$$

Solving the differential equation enables us to obtain the function of population changes with time:
\[
P(t) = \frac{K}{1 + \left(\frac{K}{P_0} - 1\right)e^{-rt}}
\]  

(12)

where \(P_0\) is the population at period 0. We use the logistic growth model for nonlinear least squares fitting of the cumulative number of confirmed cases in different cities.

**SIR model.** The SIR model constructed by Kermack and McKendrick (1927) is mainly used in infectious disease dynamics. The SIR model divides the total population into three categories: (1) susceptibles, which is recorded as \(s(t)\) and represents the number of uninfected people at time \(t\) that may become infected; (2) infectives, which is recorded as \(i(t)\) and represents the number of infected people that are infectious at time \(t\); and (3) recovered, which is recorded as \(r(t)\), and represents the number of people that are removed from the infected population at time \(t\). The total population is recorded as \(N(t)\), i.e., \(N(t) = s(t) + i(t) + r(t)\). We construct the SIR model based on the following three assumptions:

Firstly, births, deaths, movement and other population dynamic factors are not considered. The population remains constant (i.e., \(N(t) \equiv K\)).

Secondly, once an infected individual contacts a susceptible person, there is a high likelihood of some infectivity. Assume that at time \(t\) the number of susceptibles that can be infected by one person is directly proportional to the total number of susceptibles in the environment \(s(t)\) and the proportion coefficient is \(\beta\). The number of people that are infected by all infected individuals at time \(t\) is thus \(\beta s(t)i(t)\).

Thirdly, at time \(t\), the number of people removed from the infected population per unit time is directly proportional to the number of infected individuals and the proportion coefficient is \(\gamma\). The number of infected individuals removed per unit time is thus \(\gamma i(t)\).

Based on our three assumptions, the infection mechanism is as follows:

\[
\begin{pmatrix}
S(i) + I(j) \rightarrow \\
I(i) \rightarrow R(i)
\end{pmatrix}
\]

\[
I(i) + I(j)
\]  

(13)

Under the three basic assumptions, it can be seen that when infected and susceptible individuals are mixed, the growth rate of infectives is \(\beta i(t)s(t) - \gamma i(t)\), the rate of decrease in susceptible individuals is \(\beta i(t)s(t)\), and the growth rate of recovered individuals is \(\gamma i(t)\). We separate the susceptibles from infected individuals as follows:
\[
\begin{align*}
\frac{ds(t)}{dt} &= -\beta_i(t)s(t) \\
\frac{di(t)}{dt} &= -\beta_i(t)s(t) - \gamma i(t) \\
\frac{dr(t)}{dt} &= \gamma i(t),
\end{align*}
\]

(14)

The solution of Equation (14) is \( I = (S_0 + I_0) - S + \frac{1}{\sigma} \ln \frac{S_0}{S} \) (\( S_0 \) and \( I_0 \) is the initial value) and \( \sigma \) is the number of contacts \( \sigma = \frac{\beta}{\gamma} \) during the infectious stage.

We employed mean absolute error (MAE) and root-mean-square error (RMSE) to assess the accuracy of model predictions. MAE and RMSE measure the differences between the predicted and actual values, and are expressed as:

\[
MAE = \frac{\sum_{T=1}^{T} \sum_{s=1}^{S} |N_{t,s} - \tilde{N}_{t,s}|}{TS}, \quad (15)
\]

\[
RMSE = \sqrt{\frac{\sum_{T=1}^{T} \sum_{s=1}^{S} (N_{t,s} - \tilde{N}_{t,s})^2}{TS}}, \quad (16)
\]

where \( N_{t,s} \) is the predicted value and \( \tilde{N}_{t,s} \) is the actual number of new confirmed cases. \( T \) and \( S \) are the duration and number of regions involved in the prediction, respectively.

3.4.2. Prediction accuracy analysis

We examine six key cities in China (Beijing, Chongqing, Guangzhou, Shanghai, Shenzhen and Wuhan) in our phasic modelling of the pandemic: in stage 1 (11 January 2019 to 23 January 2020), stringent prevention and control measures were not carried out in Wuhan and only case reporting was conducted. It can be regarded as the onset phase of the epidemic. In stage 2 (24 January onwards after a lockdown was announced on 23 January), traffic control and quarantine systems were gradually established to limit population movement, which would effectively flatten the affected cases of the epidemic. In stage 3 (2 February to 24 February), community prevention and control measures were strengthened in Wuhan and the “no omission of households and individuals” community public health management model was proposed. At the same time, the Huo Shen Shan hospital, the Lei Shen Shan hospital, and cabin hospitals were successively constructed and activated. Large hospitals from 20 provinces sent medical assistance teams to Wuhan, which improved the prevention and control and medical levels in Wuhan. We indeed observe the decline of the infected cases in this stage.
Table 4 shows the MSE and RMSE of our spatio-temporal point process model, as well as the three models at the three different stages. All models demonstrate good prediction results for cities with less severe disease spread at stage 1 (Shenzhen and Shanghai). However, there are few prediction differences between the models in stage 1. In contrast, in stages 2 and 3, our spatio-temporal point process model has relatively more accurate predictions than the other three models. For example, the RMSE of the ARIMA model, logistic model and SIR model for new cases in Wuhan in stage 2 are 119.3, 140.2 and

| Table 4 |
| Assessment of model predictions at various stages (major cities) |
| | Wuhan | Beijing | Shanghai | Guangzhou | Shenzhen | Chongqing |
| A: Stage 1 (MAE) (11/01/2020–23/01/2020) | | | | | | |
| ARIMA model | 3.8 | 2 | 1.1 | 0.5 | 0.3 | 2.3 |
| Logistic model | 4.8 | 1.5 | 0.6 | 0.3 | 0.2 | 2.0 |
| SIR model | 3.9 | 1.5 | 0.5 | 0.3 | 0.2 | 1.9 |
| Spatio-temporal point process model | 3.5 | 1.4 | 0.3 | 0.1 | 0.1 | 1.6 |
| A: Stage 1 (RMSE) (11/01/2020–23/01/2020) | | | | | | |
| ARIMA model | 5.0 | 2 | 1.2 | 0.5 | 0.2 | 2.2 |
| Logistic model | 5.5 | 1 | 0.8 | 0.3 | 0.2 | 1.3 |
| SIR model | 5.6 | 1 | 0.7 | 0.3 | 0.1 | 1.5 |
| Spatio-temporal point process model | 4.9 | 0.6 | 0.5 | 0.1 | 0.1 | 0.9 |
| B: Stage 2 (MAE) (24/01/2020–02/02/2020) | | | | | | |
| ARIMA model | 88.1 | 2.7 | 2.1 | 1.9 | 3.5 | 2.5 |
| Logistic model | 95.5 | 1.2 | 0.8 | 2.1 | 3.5 | 1.7 |
| SIR model | 63.7 | 1.4 | 0.7 | 1.9 | 3.0 | 1.5 |
| Spatio-temporal point process model | 35.1 | 1.0 | 0.7 | 1.0 | 2.8 | 1.1 |
| B: Stage 2 (RMSE) (24/01/2020–02/02/2020) | | | | | | |
| ARIMA model | 119.3 | 4.6 | 2.7 | 2.3 | 6.1 | 5.3 |
| Logistic model | 140.2 | 4.1 | 2.7 | 2.3 | 5.3 | 5.0 |
| SIR model | 92.2 | 3.8 | 2.3 | 2.3 | 4.6 | 4.9 |
| Spatio-temporal point process model | 50.1 | 2.3 | 1.5 | 1.5 | 4.0 | 4.6 |
| C: Stage 3 (MAE) (03/02/2020–24/02/2020) | | | | | | |
| ARIMA model | 612.4 | 2.1 | 1.3 | 1.4 | 1.5 | 2.9 |
| Logistic model | 466.1 | 3.2 | 0.8 | 1.3 | 1.6 | 2.7 |
| SIR model | 509.2 | 3.5 | 0.8 | 1.1 | 1.3 | 2.4 |
| Spatio-temporal point process model | 489.7 | 2.7 | 0.6 | 0.5 | 1.0 | 2.2 |
| C: Stage 3 (RMSE) (03/02/2020–24/02/2020) | | | | | | |
| ARIMA model | 1130.9 | 3.8 | 1.8 | 2.8 | 2.5 | 4.1 |
| Logistic model | 686.6 | 3.2 | 1.8 | 2.3 | 2.3 | 3.8 |
| SIR model | 895.7 | 2.3 | 1.4 | 1.9 | 1.8 | 3.5 |
| Spatio-temporal point process model | 506.3 | 1.5 | 1.1 | 1.1 | 1.5 | 3.3 |

© 2021 Accounting and Finance Association of Australia and New Zealand
92.2, respectively, while the RMSE of the spatio-temporal point process is 50.1. Similar model rankings are observed for the other five cities.

Figure 1 shows the time series graphs for new confirmed cases in each city (blue) and the predicted sequence graphs of the various models, such as
ARIMA (black), logistic model (green), SIR model (purple), and the spatio-temporal point process model (red). The graphs show the natural logarithmic values of the number of new confirmed cases in the six cities. Overall, our spatio-temporal point process model has the best fit, particularly when there are large fluctuations in new cases. For example, there is a large increase in the number of confirmed cases in Wuhan on 15 February, which is due in part to changes in the test criteria. As our model has a higher capability of evaluating the effects of external factors, it thus generated better predictions when there are large fluctuations in the data. The model with the next best fit is the SIR model, followed by the logistic model and the ARIMA model. At the second stage in Wuhan, the prediction from the ARIMA model significantly underestimated the actual number of new confirmed cases. From 23 January 2020 onwards, its prediction conformed to disease trends but estimated fewer confirmed cases. In the third stage (3 February 2020), the underestimation by the ARIMA model becomes more significant. The logistic model shows relatively better performance but there is still underestimation, particularly during the inflection points when the pandemic escalates to the next stage. For example, from 3 February 2020 onwards, the logistic model does not show a good fit with actual values. The SIR model and our spatio-temporal point process model demonstrate better predictive capabilities. The predictive capabilities of various models in these graphs match the MAE and RMSE evaluation results in Table 4.

Figure 2 shows the time series graphs of the cumulative confirmed cases and the prediction graphs from each model for each city. Similarly, the original values underwent natural logarithmic transformation. The advantages of our spatio-temporal point process model are more apparent. Using Wuhan as an example, our model demonstrates good fit with actual values in stages 1 and 2; in stage 3, the model prediction is slightly underestimated. However, the other three models show more significant underestimation. This is particularly so for the ARIMA model, which shows significant underestimation from stage 2 onwards. Both the logistic model and SIR models demonstrate good performance but also greatly underestimate the cumulative number of confirmed cases with time; these models exhibit the same performance in the other five cities.

In summary, compared with conventional models, our new spatio-temporal point process model demonstrates better prediction capabilities. At the same time, accurate prediction becomes more difficult as the number of new confirmed cases increases. Firstly, infectious diseases show some seasonality. Therefore, it is difficult to use data from periodic confirmed cases to predict infectious diseases. In particular, the ARIMA model is not suitable for this type of analysis. In addition, infectious diseases show temporal and spatial dynamic transmission, and conventional models, such as the SIR model, focus mainly on temporal changes and ignore spatial processes. Thirdly, there is a problem with combining the frequency and intensity of infectious disease occurrence.
Conventional models mainly use the number of confirmed cases for modelling but ignore the frequency of occurrence. Our evidence shows that our spatio-temporal point process model significantly improves these aspects, thereby significantly improving prediction accuracy.

© 2021 Accounting and Finance Association of Australia and New Zealand
4. Application of the model in analysis of population sustainability

After the outbreak of COVID-19, many countries employed stringent prevention and control measures at various stages. From 20 January 2020 onwards, the Chinese government included COVID-19 as a legally notifiable class B infectious disease and border quarantine infectious disease. A temperature monitoring and health declaration system was implemented and legal monitoring and transportation hub inspections began. On 23 January, Wuhan implemented strict traffic restriction measures, improved diagnosis, treatment and prevention and control protocols, and strengthened quarantine and resuscitation measures. Quarantine and medical observations were carried out for close contacts and for people from regions with disease spread. The Chinese New Year holiday was extended, traffic restrictions and capacity control measures were implemented, population movements were reduced, and gatherings were cancelled. Prevention and control information were released and public risk communication and health advocacy were strengthened. Coordinated deployment of medical resources, construction of new hospitals, activation of reserve beds and corresponding sites were used, and people were hospitalised if needed. The prices of daily necessities were stabilised, and measures enacted to maintain social stability. After 5 February, testing capacity was strengthened in China. China also employed big data and artificial intelligence to strengthen the management of close contacts and key populations. Health insurance payments, settlement in other regions, and financial catch-call health insurance policies were introduced. All of China focused on supporting Wuhan and other affected regions in Hubei province. Comprehensive testing, drugs, vaccines, disease spectrum, contact tracing and other emergency research were carried out. Did the above measures effectively suppress the rapid spread of COVID-19 and change the risk of rapid transmission?

Table 5 shows the estimated parameters at various stages from our spatio-temporal point process model. It can be seen that at stage 1, the daily number of new confirmed cases is low in the six cities during our sample period (see the estimated value for parameter $\mu$). At the same time, disease transmission is not high between these cities (see the estimated value for parameter $\rho$). The results show that compared with social factors, natural factors have greater effects on disease spread. However, these results may be due to insufficient testing capacity and the reported confirmed case data do not accurately reflect the number of infected people. At stage 2, the daily number of new cases increased dramatically in Wuhan and the other five cities. In addition, the effects of temporal self-excitation of disease occurrence frequency in the same city and mutual-excitation between different cities are greatly increased (see the estimated value for parameters $\delta$ and $\rho$). Social factors play a greater role in disease transmission than natural factors in stage 2. In stage 3, disease transmission is not controlled in Wuhan while it is controlled in the other five
cities, where the daily number of new confirmed cases significantly decreased. In addition, mutual-excitation in regions with disease spread is significantly decreased, while self-excitation in the same region is significantly increased. These results show that the series of strict prevention and control measures and increased testing capacity that were carried out since Wuhan was locked down could effectively control the spatio-temporal transmission of COVID-19.

We employ the SIR model to conduct further quantitative analysis of the pandemic control results due to the strict control and quarantine measures and increased testing capacity implemented by the Chinese government. The basic reproduction number ($R_0$) is an important parameter in the SIR model that shows the transmission speed and the final scale of the number of infected people for an infectious disease. We use the numerical definitions of $R_0$ at various stages in China, particularly in Wuhan, that are common in the literature (Wang et al., 2020). We calculate the time required to infect an entire population if no disease prevention, control and medical measures are carried out. We adopt changes in the proportion and peak value of the number of

| Table 5: Parameter estimates of the model at each stage (major cities) |
|----------------|----------------|----------------|----------------|----------------|----------------|
|                | Wuhan          | Beijing        | Shanghai       | Guangzhou      | Shenzhen       | Chongqing      |
| A: Stage 1 (11/01/2020–23/01/2020) |               |                 |                 |                 |                 |                 |
| $\mu$          | 35.214         | 10.523         | 6.021          | 2.341          | 0.478          | 10.898         |
| $\beta$ (Epidemic factor) | 1.235          | 0.378          | 0.146          | 0.078          | 0.014          | 0.765          |
| $\beta$ (Social factor) | 0.897          | 0.215          | 0.079          | 0.023          | 0.008          | 0.433          |
| $\beta$ (Natural factor) | 2.134          | 0.786          | 0.567          | 0.235          | 0.124          | 0.876          |
| $\gamma$0      | 0.021          | 0.007          | 0.004          | 0.001          | 0.0007         | 0.024          |
| $\gamma$1      | 0.542          | 0.156          | 0.076          | 0.054          | 0.023          | 0.179          |
| $\delta$       | 0.052          | 0.043          | 0.033          | 0.028          | 0.011          | 0.048          |
| $\rho$         | 0.031          | 0.028          | 0.025          | 0.020          | 0.015          | 0.029          |
| B: Stage 2 (24/01/2020–02/02/2020) |               |                 |                 |                 |                 |                 |
| $\mu$          | 469.214        | 18.126         | 17.987         | 19.056         | 22.897         | 27.886         |
| $\beta$ (Epidemic factor) | 5.145          | 2.780          | 2.565          | 3.098          | 4.206          | 4.565          |
| $\beta$ (Social factor) | 6.879          | 4.284          | 4.146          | 4.043          | 4.674          | 4.897          |
| $\beta$ (Natural factor) | 5.125          | 3.789          | 3.567          | 3.232          | 3.455          | 3.666          |
| $\gamma$0      | 0.105          | 0.095          | 0.088          | 0.076          | 0.098          | 0.099          |
| $\gamma$1      | 1.877          | 1.453          | 1.345          | 1.233          | 1.554          | 1.656          |
| $\delta$       | 0.156          | 0.104          | 0.102          | 0.112          | 0.134          | 0.123          |
| $\rho$         | 0.189          | 0.156          | 0.144          | 0.158          | 0.168          | 0.177          |
| C: Stage 3 (03/02/2020–24/02/2020) |               |                 |                 |                 |                 |                 |
| $\mu$          | 1910.446       | 8.098          | 5.198          | 5.592          | 6.893          | 10.897         |
| $\beta$ (Epidemic factor) | 23.654         | 0.989          | 0.786          | 0.876          | 0.887          | 1.246          |
| $\beta$ (Social factor) | 10.129         | 2.256          | 2.098          | 2.234          | 2.656          | 3.346          |
| $\beta$ (Natural factor) | 5.223          | 1.353          | 1.098          | 1.898          | 1.235          | 2.442          |
| $\gamma$0      | 0.115          | 0.089          | 0.076          | 0.076          | 0.887          | 0.989          |
| $\gamma$1      | 2.489          | 1.454          | 1.343          | 1.223          | 1.441          | 1.898          |
| $\delta$       | 0.289          | 0.188          | 0.156          | 0.176          | 0.189          | 0.198          |
| $\rho$         | 0.167          | 0.143          | 0.122          | 0.156          | 0.166          | 0.172          |
infected people when different stages of measures are adopted. In stage 1, China did not adopt many prevention and control measures but medical intervention was deployed. At this stage, the $R_0$ was close to 3.1. Stage 2 began on 23 January 2020 and public transportation to and within Wuhan was stopped and large gatherings strictly prohibited. Strict quarantine and prevention and control measures were gradually implemented in Wuhan. In stage 2, the $R_0$ decreased to 2.6. Stage 3 began on 2 February and new infectious disease hospitals and mobile hospitals were activated. A large number of medical and public resources were used to combat disease transmission. Medical teams from other provinces in China also arrived in Wuhan. The community quarantine level was increased in quarantined areas. Testing capacity for COVID-19 also increased. Close contacts and suspected cases were also better identified and controlled. By mid-February, $R_0$ had decreased to 0.9–0.5. We employ the SIR model to simulate the graphs on the proportion of infected people and disease progression over the next 6 months in the whole of China (Figure 3) and

![Figure 3](https://example.com/figure3.png)

Figure 3  Epidemic prevention measures and number of infected people (nationwide): the y-axis presents the percentage of infected population, and the x-axis presents the predictive horizon (days). The first prevention stage is from 11 January 2020 to 23 January 2020. The second prevention stage is from 24 January 2020 to 1 February. The third prevention stage is from 2 February to 24 February. [Colour figure can be viewed at wileyonlinelibrary.com]
Wuhan (Figure 4) in the absence of any disease prevention, control and medical measures, and the various stages of the prevention and control measures.

The top left of Figure 3 shows the disease progression in China if there are no disease prevention, control and medical measures (i.e., a graph of the proportion of infected people). The graph shows the exponential growth of infected people over time. The model predicts that 110 days is required to infect the entire population of China, which is 1.3 billion people. The top right of Figure 3 shows that the disease progression curve is greatly flattened when medical intervention and prevention and control measures were carried out in the first stage. The peak of the number of infected people is close to 40 percent. The bottom left of Figure 3 shows the predicted disease progression curve

Figure 4  Epidemic prevention measures and number of infected people (Wuhan): the y-axis presents the percentage of infected population, and the x-axis presents the predictive horizon (days). The first prevention stage is from 11 January 2020 to 23 January 2020. The second prevention stage is from 24 January 2020 to 1 February. The third prevention stage is from 2 February to 24 February. [Colour figure can be viewed at wileyonlinelibrary.com]
based on $R_0$ and the initial proportions of infected and recovered people at the start of stage 2. We find that the peak number of infected people decreases to close to 30 percent when prevention and control and quarantine measures are strengthened by the government at this stage; the entire disease progression curve is flatter. At the start of stage 3, China greatly strengthened its testing capacity, which is combined with strict prevention and control and quarantine measures. This causes the disease progression curve to pass the inflection point and enter the decline stage. At this stage, the pandemic is successfully controlled. Figure 4 shows the disease progression in Wuhan and the effects of prevention and control and testing. When no intervention measure is employed, it takes around 80 days for the entire population of Wuhan to become infected, which is 10 million people. At stages 1 and 2, disease progression is well controlled and the peak number of infected people is 40–30 percent, and it begins to decline in stage 3. These results demonstrate and that the strict prevention and control measures and large area testing by the Chinese government during the COVID-19 pandemic successfully curbed the rapid spread of the disease during our period of study. At the same time, the implementation of these measures reduced COVID-19 transmission and enabled the economy to rapidly recover.

5. Conclusion

In this study, we combine epidemiological and spatio-temporal point process theories and propose a spatio-temporal point process model that we employ to examine the daily new confirmed cases of COVID-19 in Wuhan and five other major cities in China. Big data and machine learning are employed. Our model is used to describe temporal self-excitation of the daily number of new confirmed cases in the same region and spatial mutual-excitation between different regions. In this model, we incorporate the effects of some external factors on the disease transmission. We find that this model accurately predicts the spatio-temporal points of new confirmed cases at stage 1 of the pandemic, as well as the predicted values of new confirmed cases. The results show that our model is an appropriate tool to provide an early signal of disease transmission so that suitable and effective disease prevention, control and medical measures can be implemented in a timely manner.

We compared the fit of our new model using data for Wuhan and five other major cities in China during the three stages of the pandemic. The results show that our spatio-temporal point process model can more accurately predict the daily number of new COVID-19 cases, particularly at the outbreak phase when the daily number of new cases is large. Our model has some limitations. Firstly, it does not consider the transmission route of the disease like in the SIR model. Secondly, our model focuses on the total number of new confirmed cases in a region, not the population. Lastly, developments on the Internet and emergence of many social network data can greatly increase the prediction.

© 2021 Accounting and Finance Association of Australia and New Zealand
capabilities for human behaviour and emotional variables. However, our model did not use such data, which can be used as a future study direction. We also did not analyse the effects of quarantine measures and testing capacity on the pandemic. The results demonstrate the measures we examine are effective in successfully controlling the pandemic and play a large role in the subsequent resumption of production, work and daily life.

References

Abbott, S., J. Hellewell, J. Munday, CMMID nCoV working group, & S. Funk, 2020, The transmissibility of novel Coronavirus in the early stages of the 2019–20 outbreak in Wuhan: exploring initial point-source exposure sizes and durations using scenario analysis. *Wellcome Open Research*, 5, 17.

Ahmed, F., N. Zviedrite, and A. Uzicanin, 2018, Effectiveness of workplace social distancing measures in reducing influenza transmission: a systematic review, *BMC Public Health* 18(1), 518.

Backer, J. A., D. Klinkenberg, and J. Wallinga, 2020, Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020, *Eurosurveillance* 25(5), 2000062.

Brooks, S. K., R. K. Webster, L. E. Smith, *et al.*, 2020, The psychological impact of quarantine and how to reduce it: rapid review of the evidence, *The Lancet* 395(10227), 912–920.

Chan, J. F.-W., S. Yuan, K.-H. Kok, *et al.*, 2020, A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster, *The Lancet* 395(10223), 514–523.

Ferguson, N. M., M. J. Keeling, W. J. Edmunds, *et al.*, 2003, Planning for smallpox outbreaks, *Nature* 425(6959), 681–685.

Fong, M. W., H. Gao, J. Y. Wong *et al.*, 2020, Nonpharmaceutical measures for pandemic influenza in nonhealthcare settings – social distancing measures, *Emerging Infectious Diseases* 26(5), 976–984.

Hawkes, A. G., 1971, Spectra of some self-exciting and mutually exciting point processes, *Biometrika* 58(1), 83–90.

Kermack, W. O., and A. G. McKendrick, 1927, A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society of London A* 115(772), 700–721.

Klepac, P., L. W. Pomeroy, O. N. Bjørnstad, T. Kuiken, A. D. M. E. Osterhaus, and J. M. Rijks, 2009, Stage-structured transmission of phocine distemper virus in the Dutch 2002 outbreak, *Proceedings. Biological Sciences* 276(1666), 2469–2476.

Koelle, K., and M. Pascual, 2004, Disentangling extrinsic from intrinsic factors in disease dynamics: a nonlinear time series approach with an application to cholera, *The American Naturalist* 163(6), 901–913.

Koelle, K., X. Rodó, M. Pascual, M. Yunus, and G. Mostafa, 2005, Refractory periods and climate forcing in cholera dynamics, *Nature* 436(7051), 696–700.

Kraemer, M. U. G., C.-H. Yang, B. Gutierrez, *et al.*, 2020, The effect of human mobility and control measures on the COVID-19 epidemic in China, *Science* 368(6490), 493–497.

Kucharski, A. J., T. W. Russell, C. Diamond, *et al.*, 2020, Early dynamics of transmission and control of COVID-19: a mathematical modelling study, *The Lancet Infectious Diseases* 20(5), 553–558.

Kulldorff, M., 1997, A spatial scan statistic, *Communications in Statistics – Theory and Methods* 26, 1481–1496.

© 2021 Accounting and Finance Association of Australia and New Zealand
Kulldorff, M., R. Heffernan, J. Hartman, R. Assunção, and F. Mostashari, 2005, A space–time permutation scan statistic for disease outbreak detection, *PLoS Medicine* 2(3), e59.

Mossong, J., N. Hens, M. Jit *et al.*, 2008, Social contacts and mixing patterns relevant to the spread of infectious diseases, *PLoS Medicine* 5(3), e74.

O’Hara, R. B., and D. J. Kotze, 2010, Do not log-transform count data, *Methods in Ecology and Evolution* 1(2), 118–122.

Prem, K., A. R. Cook, and M. Jit, 2017, Projecting social contact matrices in 152 countries using contact surveys and demographic data, *PLOS Computational Biology* 13(9), e1005697.

Quilty, B. J., S. Clifford, S. Flasche, and R. M. Eggo, 2020, Effectiveness of airport screening at detecting travellers infected with novel coronavirus (2019-nCoV), *Eurosurveillance* 25(5), 2000080.

Riley, S., C. Fraser, C. Donnelly *et al.*, 2003, Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions, *Science* 300(5627), 1961–1966.

Riou, J., and C. L. Althaus, 2020, Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020, *Eurosurveillance* 25(4), 2000058.

Tian, H., Y. Liu, Y. Li, *et al.*, 2020, An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China, *Science* 368(6491), 638–642.

Unkel, S., C. P. Farrington, P. H. Garthwaite, C. Robertson, and N. Andrews, 2012, Statistical methods for the prospective detection of infectious disease outbreaks: a review, *Journal of the Royal Statistical Society: Series A* 175(1), 49–82.

Wang, H., Z. Wang, Y. Dong *et al.*, 2020, Phase-adjusted estimation of the number of Coronavirus Disease 2019 cases in Wuhan, China, *Cell Discovery* 6(1), 10.

Wu, J. T., K. Leung, and G. M. Leung, 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.

Zhang, J., P. Klepac, J. M. Read *et al.*, 2019, Patterns of human social contact and contact with animals in Shanghai, China, *Scientific Reports* 9(1), 15141.