Applying population-flow-based spatial weight matrix in spatial econometric models: conceptual framework and application to COVID-19 transmission analysis

Abstract

This paper proposes a novel method for constructing an asymmetric spatial weight matrix and applies it to improve spatial econometric modeling. As opposed to traditional spatial weight matrices that simply consider geographic or economic proximity, the spatial weight matrix proposed in this study is based on large-volume daily population flow data. It can more accurately reflect the socio-economic interactions between cities over any given period. To empirically test the validity and accuracy of this proposed spatial weight matrix, we apply it to a spatial econometric model that analyzes COVID-19 transmission in Mainland China. Specifically, this matrix is used to address spatial dependence in outcome and explanatory variables, and to calculate the direct and indirect effects of all predictors. We also propose a practical framework that combines Instrumental Variable regressions and a Hausman test to validate the exogeneity of this matrix. The test result confirms its exogeneity, hence it can produce consistent estimates in our spatial econometric models. Moreover, we find that spatial econometric models using our proposed population-flow-based spatial weight matrix significantly outperform those using the traditional inverse distance weight matrix in terms of goodness-of-fit and model interpretation, thus providing more reliable results. Our methodology not only has implications for national epidemic control and prevention policies but can also be applied to a wide range of research to better address spatial autocorrelation issues.

Keywords: *Population flow, spatial weight matrix, endogeneity, spatial dependence (autocorrelation), COVID-19 transmission*

1. Introduction

Many phenomena in the real-world are spatially dependent. For example, economic development, meteorological conditions, air pollution and the spread of diseases between neighboring regions may correlate or interact with each other to a greater or lesser extent. Traditional econometric models assume the spatial dependence or spatial spillovers between research units to be zero, which may generate biased estimates of regression coefficients (Vega and Elhorst 2013). To empirically assess the magnitude and significance of the spatial dependence or spatial spillover effects, spatial econometric models have been widely used (LeSage 2008; ; Vega and Elhorst 2013). The interactions between spatial units in spatial econometric models are reflected through a spatial weight matrix (SWM), a square matrix of size N*N, with N being the number of research units being modeled. SWMs have played a vital role in deriving accurate models and estimates (Chen 2021). However, it is often challenging to select and construct an appropriate SWM that accurately reflects the spatial correlation and interactions between research units (Seya, Yamagata, and Tsutsumi 2013; Lam, and Souza 2020).

Traditional forms of spatial weight matrices include contiguity-based SWM (Cliff and Ord 1975), inverse-distance-based SWM (Anselin 2001), economically-based SWM (Conley and Ligon 2002) and nested SWM that combined both geographic and economic distances (Fingleton and Le Gallo 2008). The contiguity-based SWM remains one of the most popular spatial weight matrices (Getis 2009). Although the contiguity-based SWM is relatively easy to build, it ignores the varying degrees of interactions among neighboring units. The

inverse-distance-based SWM assumes that the intensity of interactions depends on geographic distance (Getis 2009). However, in many contexts, geographic proximity may not fully reflect the relevant connections between spatial units. For instance, after Wuhan's COVID-19 lockdown in January 2020, a SWM based on geographic proximity is inappropriate to capture the city's actual connections with other cities. Furthermore, new transportation technologies such as high-speed rail have changed the effective distance and the frequency and intensity of interactions between two places in real space, altering the strength and scope of spillover effects (Yu, Chen and Zhu,. 2012; Zhu et al. 2015; Cao and Zhu, 2017; Ahlfeldt and Feddersen, 2018; Zhu, 2021). Another major limitation of these traditional SWMs is the symmetric assumption of spatial spillover effects, which may deviate from the actual spatial interaction processes. Therefore, more appropriate spatial weight matrices that account for the real socio-economic interactions induced and amplified by modern technological innovations (e.g., transportation, communication) are needed to better reflect the spillovers between research units.

While most previous empirical studies focus on the application of spatial econometric models, only a few studies have introduced new ways to construct spatial weight matrices to better capture the underlying interaction process. For example, Case and Hines (1993) used income and racial composition to describe the associations between states. Zhang et al. (2009) proposed a co-movement SWM by accounting for the similarity of economic factors between regions. Getis and Aldstadt (2004) constructed a SWM based on the G_i^* local statistic. Emch et al. (2012) applied spatial proximity and social relationships separately to construct a SWM

and compare the spatial clustering pattern of disease transmission based on the two types of matrices. Taking into consideration the actual connections between spatial units, some studies have introduced actual population flow (e.g., Kordi and Fotheringham 2016) or flow intensity calculated based on GDP per capita and the number of people employed in different industries in connected cities (Li 2017) into the SWM. However, the above studies have mainly focused on the spatial interaction process itself. They do not apply their updated SWMs to improve spatial econometric models and illustrate the spillover effects generated by the spatial interaction processes.

The major contributions of this research, therefore, lie not only in constructing an asymmetric SWM based on large-volume daily population flows between geographic units, but more importantly, in applying the new SWM to improve spatial econometric modeling and to provide more reliable estimates of the (direct and spillover) effects of key variables. In the modeling of various social, economic, and health outcomes, a population-flow-based SWM may better approximate the actual process of spatial interactions than traditional spatial weight matrices using inverse distance or contiguity. This improved SWM also accounts for potential asymmetric spatial spillover effects (i.e., spillover from A to B not equal to spillover from B to A). This is consistent with the fact that the flows of people, goods, and information between spatial units are likely to be asymmetrical due to geographical constraints (e.g., Xu et al. 2016) or the uneven level of development between spatial units (e.g., Parent and LeSage 2006). In the empirical analysis of this paper, we construct the proposed SWM using real-time population flow data and apply it in spatial econometric models to analyze the

transmission of COVID-19 in Mainland China between January 11 and February 25, 2020 (excluding the two special administrative regions of Hong Kong and Macau and Taiwan region). By comparing the results of our model with an Ordinary Least Squares (OLS) model and spatial econometric model using inverse distance SWM, we find that the proposed model provides more reliable results that may better inform policy for epidemic control and prevention. In addition to application in infectious disease control, this methodology may also be applicable in investigating a variety of spatial outcomes that depend on human contact (e.g., economic, social, transportation, or public health studies).

The remainder of this article is structured as follows. Section 2 provides a brief review of the literature on spatial econometrics as well as the specification of the SWM. Section 3 introduces the construction process of the population-flow-based SWM. An example of the application of the proposed SWM in spatial regression is presented in Section 4 and conclusions are presented in the final section.

2. Literature review

Spatial econometrics was first proposed by Paelinck and Klaassen (1979) to improve traditional econometric approaches by capturing the spatial dependence between observations (Elhorst 2014). It has been widely used in fields such as transportation (e.g., Szabó, and Török 2020), environmental studies (e.g., Lv, Chen, and Cheng 2019), economics (e.g., J. Li, and S. Li 2020), and public health (e.g., Ispriyanti, Prahutama, and Taryono 2018).

The construction of spatial weight matrices is the key to spatial econometric models. SWMs are designed to reflect the interactions between spatial units (Kostov 2010) and may take different forms depending on the rationales behind the scenes. The earliest form of the SWM was the contiguity-based SWM (Getis 2009). This category of SWMs can be further divided into first-order contiguity matrices and high-order contiguity matrices. First-order contiguity matrices assume that spatial interactions only occur between spatial units sharing a common border (first-order neighbors) and the strength of the interactions between all pairs of first-order neighbors is the same (Getis 2009). High-order contiguity matrices are constructed in a similar way. For example, if one unit is given, its second-order neighbors are defined as the neighbors of its first-order neighbors, and so on. Although easy to implement, the basic assumption of the contiguity-based approach that no variations exist in the degree of interactions among neighbors of the same order is only a simplified measure of spatial interactions in the real world. Another type of SWM is the inverse distance matrix, which constructs spatial weights using the distance between pairs of observations (Perret 2011). Following Tobler's first law of geography (Tobler 1970), the underlying assumption of this approach is that the intensity of spatial relations among observations decreases as the distance between them increases. Most studies use the Euclidean distance between two spatial units to calculate spatial weights (e.g., Lu and Zhang 2011; Lv, Chen, and Cheng 2019), while other studies use alternative measures such as travel time (e.g., Conley and Topa 2002) or railway network distance (e.g., Lv, Chen, and Cheng 2019) to capture the physical proximity or travel costs between two units in the real world.

With high residential mobility and the development of advanced communication technologies in the modern world, geographic constraints have become weaker (Webber 1964; Wellman and Leighton 1979; Snyder 1995), and the influence of physical distance on the interactions among nodes in networks has declined (Conley and Topa 2002). To better capture the diverse spatial interaction processes that are not constrained by physical distance, some studies have used non-physical distance measures such as economic distance (Case et al. 1994; Pietrzak 2010), trade volumes (Aten 1997; Cohen and Paul 2004), industrial structure proximity (Zhang, Chen and Wang 2009), social contacts (Conley and Topa 2002) and social network/relationships (Emch et al. 2012; Leenders 2002) to substitute for physical distance. The nested weights matrix that combines the inverse geographic distance and non-physical distance matrices is another way to account for various spatial, economic, technological and transportation proximity factors influencing the spatial interaction processes (Parent and LeSage, 2008). However, one significant limitation of all the aforementioned SWM is the assumption of symmetric spatial spillover effects (i.e., the impact from observation *i* to observation *j* is equal to the impact from *j* to *i*), which does not accurately describe the spatial and socio-economic interaction processes in the real world. To deal with this limitation, a number of studies have constructed asymmetrical nested weights matrices (Li et al. 2010; Zheng et al. 2019). This type of matrix considers both geographic and economic proximity, but the indicator selection and matrix definition are relatively subjective and cannot incorporate population flow into measures of spatial associations. Focusing on infectious disease transmission, neighborhood relationships, hydrologic

connectivity of villages (Gu and Spear 2006), and kinship relationships (Emch et al. 2012) have also been used to construct asymmetric SWMs.

In the network analysis literature, network autocorrelations have been noticed and investigated by a growing body of studies. Black (1992) proposed a new method based on Moran's *I* statistics to measure network autocorrelations and argued that weight matrices are essential to reflect network structures. Leenders (2002) used social distance to construct spatial weights and applied the SWM in spatial modeling of the dependence embedded in electoral behavior. Chun (2008) applied eigenvector spatial filtering to build a network link matrix that combines the influences of competing destinations and intervening opportunities on travel behavior and used the matrix in a spatial filtering interaction model to analyze interstate migration behavior in the U.S. Ermagun and Levinson (2018) used the properties of networks to construct a SWM for examining spatial dependence in traffic network analysis. These studies recognize the mechanism and sources of network autocorrelations and emphasize the importance of incorporating network autocorrelations in understanding these networks. But none of them apply their updated SWMs in spatial econometric models to explicitly examine the direct and spillover effects induced by these network autocorrelations.

Other recent studies have incorporated actual population flows into the construction of spatial weight matrices. The flow of population is not only an important spatial interaction process but also facilitates other spatial spillover effects such as the transmission of ideas and beliefs (Leenders 2002; Homans 2013) as well as infectious diseases (Anderson 2013; Jia et al. 2020; Wei and Wang 2020;). Kordi and Fotheringham (2016) proposed a family of

localized spatially weighted interaction models (SWIM) including origin/destination-focused SWIM and flow-focused SWIM to address the spatial heterogeneity in spatial interactions using a geographical weighting approach; they proved that this method performed better in spatially nonstationary processes analysis. Moreover, the population flows between spatial units are usually asymmetric in nature and those areas with large net inflows or outflows may exert stronger spillover effects. For example, cities with more move-out population flow are more influential (Wei et al. 2018), serving as the critical nodes in the population flow network and influencing key properties of the network like clustering and transitivity (Xu et al. 2010; Alstott et al. 2014). These cities are highly associated with upsurges in epidemic transmission (Zhong and Bian 2016). Thus, a population-flow-based matrix not only better models the spatial interaction processes, but also better reflects the asymmetric nature of the interactions between spatial units. However, previous studies have mainly applied population-flow-based matrices to examine the spatial interaction process itself but have not incorporated these matrices into spatial regressions to model the actual channels of spatial spillover (i.e., human interactions) and further examine how the incorporation of population-flow-based matrices will improve spatial model specifications. The population flow data used in previous studies are usually yearly or monthly averages, while real-time data on daily population movement are seldom applied in spatial modeling.

3. Conceptual framework for constructing the SWM using population flow

To fill in the research gaps, this paper proposes a method of constructing a SWM based on the volume of population flow between two cities and applies the matrix in spatial regressions to better capture the actual spatial interaction processes. Using the COVID-19 outbreak in Mainland China as an example, we construct a population-flow-based SWM based on Baidu population flow data obtained from Baidu Huiyan platform (https://qianxi.baidu.com/2020/) and apply it in spatial regressions to explore the influencing factors and the real transmission mechanisms in the spread of COVID-19.

3.1 Theoretical framework

Spatial weight matrices represent the interaction processes between geographic units such as cities, regions, and provinces. The definition of spatial weights is the key element of matrix construction (Leenders 2002). In this study, spatial weights based on population movement are defined as follows:

$$w_{ij} = \frac{\sum_1^k Daily\ PM_{ij}}{k} \quad (1)$$

where the w_{ij} is the spatial weight of unit i towards unit j, PM_{ij} reflects the volume of population movement from city i to city j, k is the number of days in the period. Note that this is an asymmetrical weight matrix as PM_{ij} is not equal to PM_{ji}. This overcomes a major shortcoming of traditional symmetrical spatial weight matrices that ignores the direction of population flow between city pairs.

3.2 The construction of the SWM

Based on the findings of existing studies, human mobility significantly contributes to the transmission of COVID-19 (Fang, Wang, and Yang 2020; Qiu, Chen, and Shi 2020; Wei et al. 2021; Zhao et al. 2020). Hence, we use Baidu population flow data to construct the SWM in our spatial econometric model.

The daily inter-city population flow indices were collected from the Baidu Migration Platform developed by Baidu, Inc. This database applies a location-based service (LBS) technology to record and visualize the population movement trajectories of all mobile internet users throughout Mainland China. The database encompasses around 80 percent of the total number of mobile phone users in Mainland China, thus providing a strong approximation of the actual population flow between and within cities (Wei et al. 2018). This dataset has been used in several geospatial analyses of COVID-19 transmission (e.g., Fang, Wang, and Yang 2020; Qiu, Chen, and Shi 2020; Liu et al. 2020; Zhu and Guo 2021; Zhu and Tan 2021). For each pair of cities, the Baidu Migration Platform provides the daily population flow indices between them (including both the moving-out indices and moving-in indices[1]). For each city, the precise shares of outflows to the top 100 destinations are available, which on average cover over 97 percent outflows (Fang, Wang, and Yang 2020). This suggests that the data

[1] The moving-out indices and moving-in indices are based on the travel intensity between specific city-pairs. For example, the moving-out index of Beijing to Tianjin is referred to the volume of population flow traveling from Beijing to Tianjin. According to Baidu's meta-data, population flow from city i to city j is considered as the move-out index for city *i* and the move-in index for city *j*.

should be able to accurately capture the real volume of population flows among cities during the research periods. For the remaining 3 percent of outflows to other destinations where precise values are not given, we follow the first law of geography (Tobler 1970) and adopt inverse distance weighted interpolation[2] to estimate the outflow values for each remaining city. The specific process of population-flow weight matrix construction in this study is shown in Figure 1. Note that we also test the robustness of our results by alternatively using a gravity model to impute these missing values for the population-flow-based SWM, as shown in section 4.6.

[2] The inverse distance weighted interpolation follows the equation:

$$Corresponding\ Percentage_{ij} = \frac{\frac{1}{Distance_{ij}^2} * (100 - \sum_1^A Accurate\ Daily\ \ Moving\_Out\ Index_{ia})}{\sum_1^V \frac{1}{Distance_{iv}^2}}$$

where $Distance_{ij}$ is the Euclidean distance between city i and city j; A is the number of destination cities with accurate move-out indices for origin city i; V is the number of destination cities without accurate move-out indices for origin city i.

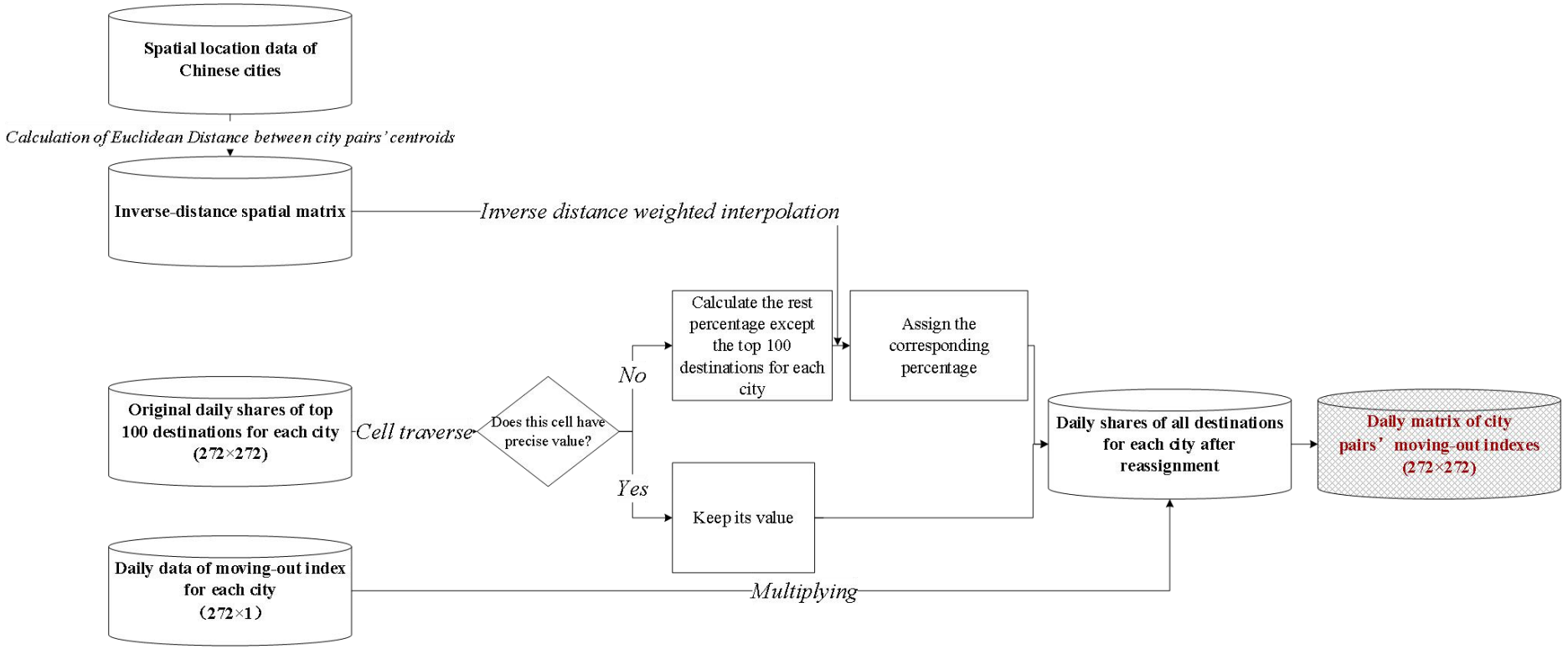

Figure 1 The process of population-flow based SWM construction

We take the averages of each element in the daily matrix of city pairs' moving-out indices by periods as $Index_{ij}^{A} = \frac{1}{N_A}\sum_{d=1}^{N_A} Index_{ijd}$, in which N_A denotes the number of the days in period A. After min-max normalization, the matrix of city pairs' moving-out indices was used as the SWM in the spatial models.

4. Application of the SWM to COVID-19 transmission analysis

COVID-19 was first identified in Wuhan, the capital city of Hubei Province. It rapidly spread outward across China and to other countries, posing a severe threat to human health. COVID-19 is transmitted through human-to-human contacts, including airborne and fomite transmissions (World Health Organization 2020). Therefore, human interactions are critical to the transmission of COVID-19.

In this section, we use city pairs' moving-out indices as the spatial weights in the construction of SWM and estimate the direct, indirect, and total impacts of different independent variables on the number of cumulative confirmed cases of COVID-19 in each city through spatial econometric models. Baidu started to publicly report daily population flow data on January 11, 2020; hence the period of our analyses can only start on this date. The first imported case in Mainland China was confirmed on February 26, 2020, when a traveler from Iran was reported positive in the Ningxia Hui Autonomous Region. From then on, more imported cases were confirmed among travelers from foreign countries, contributing to a new wave of outbreaks in Mainland China. Since the impacts of the

imported cases on pandemic transmission cannot be distinguished from those of other independent variables in this study, we only look at the period before February 26, 2020 to avoid interference from these imported cases. In addition, population (out) flows from Wuhan, the epicenter of the COVID-19 outbreak, slumped after the city's lockdown at 10 am on January 23 and the implementation of strict nationwide prevention and control measures. Therefore, we divide the sample into two subperiods: 1) the period before the Wuhan lockdown (January 11–23, 2020), during which inter-city travel in Mainland China was normal; and 2) the post-lockdown period (January 24 – February 25, 2020). We implement spatial econometric analysis for these two periods separately to examine the different impacts of explanatory variables during the two periods.

4.1 Spatial econometric models

The transmission of COVID-19 is influenced by a wide variety of factors. Some studies have argued that coronavirus transmission is affected by geographical proximity, socioeconomic interactions, and the similarity of meteorological conditions across neighboring spatial units (Andersen et al. 2021; Sannigrahi et al. 2020). Therefore, spatial autocorrelation of virus transmission as well as of the independent variables need to be incorporated in spatial econometric models (e.g., spatial lag model [SLM], spatial error model [SEM] and spatial Durbin model [SDM]). This allows for more accurate estimation compared to an ordinary

least square (OLS) model. To correctly specify the model, the existence of spatial autocorrelation in the COVID-19 data must first be tested.

The global Moran's *I* index is widely used to detect spatial dependence (Moran 1950):

$$I = \frac{N}{W} \cdot \frac{\sum_{i=1}^{N} w_{ij} \cdot (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \quad (2)$$

where N is the number of cities; w_{ij} represents an element of the SWM, which defines the spatial relationships between cities; W is the sum of all w_{ij}; $x_i(x_j)$ and $\bar{x}$ is the specific variable in city i (j), and the $\bar{x}$ denotes the mean of x. The value of Moran's *I* does not range exactly from -1 to 1 but depends on the spatial weight matrix of the study area (De Jong, Sprenger and van Veen 1984). In general, a negative Moran's *I* value indicates spatial dispersion, while a positive value indicates spatial clustering.

Commonly used spatial econometric models include SLM, SEM, SDM, and the spatial Durbin error model (SDEM)[3]. Among these models, SLM accounts for spatial dependence in the dependent variable, while SEM accounts for spatial dependence in the error term (Gujarati 2021). The SDM is specifically designed to capture the spatial spillover effects of both the explanatory variables and the explained variable. It can also be treated as an unrestricted model that can be simplified into SLM and SEM by coefficient setting (LeSage 2008). The SDM can be denoted as:

$$Y = \rho W_P Y + \beta X + \vartheta W_G X + \alpha + \varepsilon \quad (3)$$

[3] According to Vega and Elhorst (2013), there are three other types of spatial econometric models: 1) the SLX (spatial lag of X) model that includes spatial interactions of explanatory variables; 2) the SAC model that includes a spatially lagged dependent and a spatially correlated error term; 3) the general nesting spatial (GNS) model that includes all three types of spatial interaction effects.

where Y is the total number of confirmed cases during each period (Period I or II, as previously defined) and X is a series of explanatory variables that may affect virus transmission. W_P and W_G are two spatial matrices constructed based on the population flow volume[4] and the inverse of geographical distance between city pairs, respectively. Because COVID-19 has clear evidence and characteristics of human-to-human transmission, it is more appropriate to use the population-flow-based matrix W_P to capture the spatial interactions of this (outcome) variable. For spatially lagged explanatory variables, we use geographical distance W_G as the weighting matrix because geographical proximity can better capture the spatial spillovers of our explanatory variables, e.g., socioeconomic factors and meteorological factors. α denotes the constant and ε is the error term. When $\vartheta = 0$, no spatial lagged explanatory variables are embedded and the SDM is transformed into the SLM.

To examine the potential endogeneity of the spatial weight matrix W_P, we follow the method in Cheng and Lee (2017) and propose a linear regression model with an endogenous variable *WY*:

$$Y = \alpha + \beta(WY) + \gamma X + \varepsilon \qquad (4)$$

where *W* is a N*N spatial weight matrix, *WY* is a N*1 column vector, and *X* is a N*k matrix. The spatial autoregressive term *WY* is an endogenous variable as it is affected by *Y,* regardless of whether *W* is endogenous or exogenous. Potentially we can adopt an

[4] During Period I, population flow between cities in China was not disrupted because it was only announced on January 20 that COVID-19 can be transmitted human-to-human and no further warnings on travel risks were announced until January 23. Moreover, Baidu did not publicly release daily population flow data until January 11. Therefore, in the models for Period I, the spatial weight matrix is built based on the population flow volume during Period I. In the models for Period II, we still use the population flow volume during Period I to account for the 14-day incubation period of the virus.

Instrumental Variable (IV) model and use *WX* as a series of instrumental variables to address the endogenous *WY*, because *WX* is clearly correlated with *WY*. The key point affecting their validity as instruments lies in whether *WX* is also correlated with ε (Greene 2012; Zhu 2011). If *W* is exogenous, then *WX* is uncorrelated with ε; hence *WX* is a valid instrument. If *W* is endogenous, then *WX* is correlated with ε; hence *WX* is no longer a valid instrument. Arguably, the inverse distance-based *W* is an exogenous SWM, which will lead to consistent 2SLS estimates if used in the above model setting. This lays the foundation for a Hausman test which can help us determine whether there are systematic differences between the model using population-flow-based W_P (potentially endogenous) and the model using inverse distance-based W_G (known exogenous); that is, whether the population-flow-based W_P is also exogenous. We follow a classic Hausman test specification where the null hypothesis is that W_P is exogenous (i.e., no systematic differences between the two models), and the alternative hypothesis is that W_P is endogenous. We first run the IV model using W_GX as instrumental variables and store the estimation results, which should give us a consistent estimator because W_G is known exogenous. We then rerun the model using W_PX as instrumental variables and compare the estimation results to the previous model via the Hausman test. The statistic of our Hausman test is found to be negative (-15.69), which suggests that we cannot reject the null hypothesis (Baltagi 2008; Hsiao 2014; StataCorp 2017). In other words, no systematic differences are found in the estimated coefficients

between the two models. Therefore, our population-flow-based W_P proves to also be exogenous and hence should produce consistent estimates in our spatial econometric models.[5]

In addition to the Hausman test above serving as technical evidence for the exogeneity of our population-flow-based SWM, we also believe it is conceptually convincing based on our research design. During Period I, inter-city travel in Mainland China was not disrupted at all, because no warnings about travel risks were announced until January 23, 2020 when the lockdown of Wuhan happened. Although virus transmission occurred during this period, population flow between cities was arguably exogeneous to the number of COVID-19 confirmed cases as life was normal at the time. In our Period II model, to account for the incubation period of the virus, we construct the SWM still based on population flow volume during Period I and use it for the spatial lagged dependent variable (i.e., total confirmed cases during Period II). This design further eliminates potential endogeneity of the weight matrix.

[5] Note that the exogeneity of our population-flow-based W depends on what outcome variable is used in the model. In our models, daily COVID-19 case number is the outcome variable. In other models using socioeconomic variables as outcome variables, the exogeneity of our population-flow-based W may be affected.

4.2 Data source and variable selection

4.2.1 City-level COVID-19 epidemiological data

Daily COVID-19 infection data for the period from January 11 to February 25, 2020 in Mainland China was retrieved from the China Data Lab of Harvard Dataverse[6]. This data was scraped from the daily COVID-19 infection data on DXY.cn, one of the earliest open datasets developed to track the COVID-19 outbreak[7]. Some cities such as the Special Administrative Regions of Hong Kong and Macao are excluded from our analysis due to the lack of socio-economic data, Baidu population flow indices or meteorological variables. Note that Wuhan is also excluded because it is regarded as an outlier for the purposes of this study and may lead to biased results. The final number of cities contained in this study is 272. The number of cumulative confirmed cases is separately calculated for each of the two study periods (i.e., the pre-lockdown and post-lockdown periods) for each city and used as the dependent variable. For the Period II model, the cumulative confirmed cases of each city at the end of period I were included as an explanatory variable to represent their initial infection levels at the time of the Wuhan lockdown.

[6] https://doi.org/10.7910/DVN/MR5IJN

[7] According to DXY.cn, the COVID-19 infection data they published was reported by 32 provincial-level Health Commissions in China

4.2.2 Baidu population flow Data

As mentioned in Section 3.2, the population flow indices obtained from the Baidu population flow dataset reflect the daily population movements between cities in Mainland China. Meanwhile, the dataset also provides a daily within-city population flow index for each city. These two indices were used in this study as proxies for the intensity of inter-city and within-city population flow. Specifically, considering the significant level of infection risks due to population outflow from Wuhan (Qiu, Chen, and Shi 2020), the average population outflow originating from Wuhan towards each destination city[8] was incorporated into the model as an important explanatory variable. Moreover, transportation research often relates residents' travel demand to social and economic interactions (Zhou, Zhang and Zhu 2019; Zhu et. al 2020), hence within-city population flow should also be included as an explanatory variable.

4.2.3 Socio-economic data

Previous research on the development of epidemics has suggested that it is necessary to take socio-economic factors such as population, economic development, and medical resources

[8] To convert the two indices into the actual volume of person-movements in and out of each city, we use the daily number of people traveling into and out of Hong Kong provided by the Hong Kong Immigration Department to calibrate and calculate the number of people that each moving-in index and moving-out index unit corresponds to. Using this data, we estimate that one index unit in the move-in index and move-out index corresponds to 71,121 person-movements. This estimated converting factor is constant across all cities and is used to calculate the actual daily volume of population inflow and outflow of each city.

into consideration since they can greatly affect social interactions, residents' behavior, and pandemic diagnosis effectiveness, thereby influencing the transmission of COVID-19 (Oyedotun and Moonsammy 2021; Qiu et al. 2020; Zhai et al. 2021). The socio-economic variables selected include the total population of urban areas, GDP per capita and licensed doctors per capita for each city collected from the latest version of the China City Statistic Yearbook (2018).

4.2.4 Natural meteorological data

As meteorological conditions potentially play a role in the transmission of contagious diseases (Li et al. 2019; Shi et al. 2020), this study also considers natural meteorological factors including average daily temperature, average daily wind speed, and average daily air quality index (AQI). Meteorological data were acquired from the China Meteorological Data Service Centre[9], which includes hourly records of meteorological elements of each meteorology observation. We first calculated the daily data by averaging the hourly data of each day for each variable. For each city where meteorological data were not available, the value of each meteorological variable was imputed using the Empirical Bayesian Kriging inverse distance weighted interpolation in ArcGIS (Krivoruchko and Gribov 2019)[10]. We

[9] http://data.cma.cn/

[10] Kriging is a type of statistical technique for optimal spatial prediction, which has been used widely in meteorological applications, agriculture, geosciences and many other disciplines due to its minimized prediction error. Compared with classical Kriging methods, the Empirical Bayesian Kriging is more robust by accounting for the errors introduced by the estimation of the Semivariogram model (Krivoruchko and Gribov 2019).

then further aggregate the daily meteorological data into the two defined periods for each city using the average values of each meteorological element.

AQI is based on the level of six atmospheric pollutants (SO_2, NO_2, PM2.5, PM10, CO, O_3) measured at all monitoring stations throughout every city. Each record includes information on the daily average, maximum, minimum, and standard deviation values of AQI. In this study, we will use the daily average AQI to represent the air quality conditions of each research unit.

The descriptive statistics of all the variables are given in Table 1. Because of their skewed distributions, the dependent variables and some of the explanatory variables are transformed into logarithms in the models (e.g., total population, GDP per capita, doctors per capita, total number of confirmed cases in Period I).

[Table 1 Descriptive Statistics of the Dependent and Independent Variables]

4.3 Spatial autocorrelation tests and model selection

The spatial autocorrelation test results are illustrated in Table 2. The significantly positive values of Moran's *I* indices suggest the existence of strong spatial autocorrelation for the dependent variable during both periods. These results imply that the OLS estimates are invalid and justify the use of spatial models to address the spatial dependence of the number of cumulative confirmed cases of COVID-19.

The SWM constructed with Baidu population flow data is applied to capture the spatial interaction of the dependent variable in our model, while the inverse distance weight matrix is used to capture that of our explanatory variables such as socioeconomic factors and

natural meteorological conditions. Following the general-to-specific rule for spatial econometric modeling (Elhorst 2014), we apply a two-stage testing procedure to select the most appropriate model, with results reported in Table 3.

First, the Lagrange Multiplier (LM) tests (LM-lag and LM-error) and the robust LM tests are applied to spatial models for Period I and Period II, respectively. For Period I, the LM test results are all significant at the 99 percent percent confidence level when the population-flow-based weight matrix (W_P) is applied in the spatial model but are all insignificant when the inverse geographic distance matrix (W_G) is used. This indicates that when the inverse geographic distance matrix (W_G) is used for the spatial models, neither spatial error model (SEM) nor spatial lag model (SLM) performs better than the non-spatial model (i.e., OLS model) for Period I. Meanwhile, because the Moran's *I* indices suggest the existence of strong spatial autocorrelation for the dependent variable, these insignificant LM test results prove that the inverse geographic distance matrix fails to capture the spatial autocorrelation identified in Period I. On the other hand, when the population-flow-based weight matrix (W_P) is used, both SEM and SLM outperform the non-spatial model for Period I, suggesting W_P is able to accurately capture the spatial autocorrelation.

For Period II, the LM-lag and LM-error tests and the robust LM tests using different weight matrices (i.e., W_P and W_G) are all significant. They suggest that both W_P and W_G capture the spatial autocorrelation in this Period and that both SLM and SEM outperform the non-spatial OLS model.

Second, the Likelihood Ratio (LR) test for the spatial common factors between SLM and SDM is conducted to verify if the SDM should be simplified to an SLM (Elhorst 2010; Seldadyo, Elhorst and Haan 2010). Based on the LR test results in Table 3, SDM can be degenerated to SLM for Period I, while SDM is preferred to SLM for Period II.

[Table 2 Results of Global Moran's *I* tests of spatial autocorrelation for dependent variables based on Population-flow weight matrix]

[Table 3 Results of LM tests, Robust LM tests and LR tests]

4.4 Results

Table 4 reports the empirical results of the two study periods. The spatial models outperform the OLS model in terms of the Goodness-of-Fit (R^2) values in both periods. Additionally, the coefficients of spatially lagged cumulative confirmed cases are significant in models for both periods (before and after Wuhan's lockdown). This indicates that the COVID-19 outbreak in a city significantly affected the number of cases in its adjacent cities due to population flow, and inversely, was affected by the COVID-19 incidence in its surrounding areas due to the spillover effects. Therefore, spatial autocorrelation exists in the number of COVID-19 confirmed cases across cities and spatial econometric models (SLM and SDM) are more effective and reliable than the non-spatial model (OLS model) for analyzing the potential determinants of COVID-19 transmission. Additionally, the global Moran's *I* test for residuals of the spatial econometric models indicates that no significant spatial autocorrelation in residuals exists (see Appendix 1).

[Table 4 Empirical results for OLS and spatial econometric models]

According to LeSage and Pace (2010), using point estimates of spatial regression models to estimate the spatial spillover effects may lead to incorrect conclusions because of feedback loop effects. They also pointed out that partial derivative interpretation of the impacts from a dependent variable to an independent variable can provide a more valid basis to interpret spatial effects. Hence, to reveal the exact spatial spillover effects, we estimate each independent variable's direct, indirect, and total effects in the spatial econometric models separately. The decomposed results are illustrated in Table 5.

[Table 5 Estimation results for decomposition of the spatial effects]

The results suggest that the total urban population had a positive direct effect on the number of cumulative confirmed cases during both periods, indicating that a larger population increases the locally confirmed cases in that city after controlling for the effects of other factors. In other words, epidemic control is more challenging in larger cities than in smaller cities. Specifically, this positive direct effect is stronger after Wuhan's lockdown. One possible reason is that COVID-19 transmission within cities gradually became severe during this period and a larger population base implies higher chances of virus transmission. Positive indirect impacts of a city's total population on adjacent cities' cumulative confirmed cases are only found in Period I, which may indicate the effectiveness of nationwide prevention and control measures swiftly adopted by other cities after the Wuhan lockdown[11].

[11] After January 24, different levels of prevention and control measures (i.e., the shutdown of public transport and public places, the lock down of residential buildings/neighborhoods, and the set-up of checkpoints to control the population

GDP per capita shows no direct impact on the local COVID-19 situation during Period I, but illustrates significant and positive spillover effects during Period II. This result suggests that proximity to economically advanced cities may increase the chances of epidemic transmission for surrounding cities. Significant and positive direct effects of licensed doctors per capita are identified in the pre-lockdown period (Period I) but not the after-lockdown period (Period II). The results for Period I indicates that a 1 percent percent rise in licensed doctors per capita was directly associated with a 99.96 percent increase in cumulative confirmed cases. While this result may sound counterintuitive, it is likely that cities with more medical resources were more effective in terms of diagnosing patients with symptoms of COVID-19. Conversely, some patients may not have been diagnosed effectively in areas with low levels of healthcare. In Period II, with progress in identification techniques and enhancement of nationwide publicity about COVID-19, the accuracy of diagnosis in most places in Mainland China had improved, and most cities had the resources needed to make diagnoses efficiently and correctly. This may explain why the number of licensed doctors per capita does not show significant direct impacts in Period II. Additionally, its significant and positive indirect impacts during Period I but negative indirect impacts during Period II imply that the medical resources of a city have significant spillover effects that help nearby cities diagnose and control COVID-19.

As for the three natural meteorological variables and AQI, only average humidity shows significant positive direct effects on the number of cumulative confirmed cases in both

entering the city) were implemented by most of the cities, especially for cities in Hubei province and some cities with relatively more confirmed cases like Wenzhou, Hangzhou, and Harbin (Fang, Wang, and Yang 2020).

periods. This is consistent with the findings in studies by Doğan et al. (2020), and Chien and Chen (2020). Meanwhile, average temperature shows positive direct effects and negative indirect effects in Period II. Note that the specific correlation between meteorological conditions and the COVID-19 pandemic is still controversial (Harmooshi, Shirbandi, and Rahim 2020; McClymont and Hu 2021) and needs to be further explored. Even so, this study provides evidence that meteorological variables are significant contributing factors in COVID-19 transmission and exhibit spillover effects on the COVID situation in surrounding regions.

One point worth noting is that the index of travel intensity within cities shows no significant effects in the Period I model but is found to exert significant and negative direct, indirect, and total impacts in the Period II model. The insignificant impact of within-city travel flow in the first period may be explained by the lag in COVID-19 transmission from Wuhan. That is, in the early stages of the pandemic (i.e., prior to Wuhan's lockdown), most cities only had a small number of confirmed cases so within-city travel flows would have a relatively weak impact on the spread of the disease. However, with the increasingly serious situation of the COVID-19 pandemic in Period II, local governments began to take different control and prevention measures, including suspending public transportation, closing public places and factories, and locking down communities. Cities with more severe local outbreaks generally implemented more strict control and preventative measures, resulting in lower intra-city travel intensity. In addition, the significant spillover effects of within-city travel may be explained by policy imitation and referencing between neighboring cities.

Population outflow from Wuhan is positively associated with local COVID-19 cumulative confirmed cases as well as cases in surrounding cities during both periods. This result indicates that population outflow from Wuhan, the outbreak source region, posed high risks to destination cities and adjacent cities. This also suggests the lockdown policy implemented in the region effectively prevented the further spread of COVID-19, which is consistent with the conclusions of studies by Liu et al. (2020), , Qiu, Chen and Shi (2020) and Yang et al. (2020).

Unsurprisingly, the direct and indirect impacts of cumulative confirmed cases in Period I are significantly positive in Period II, suggesting that the more cumulative confirmed cases were present in a city before the lockdown of Wuhan, the more serious the outbreak situation in the city would be during Period II. A higher initial infection level generally means a greater possibility for human-to-human transmissions in the later stage. The strong indirect effects further emphasize the necessity of control measures in neighboring cities near those hotspots.

4.5 SWM performance comparison

In this section, we further compare the performance of the spatial econometric models using different SWMs for each period.

1) The Period I model

Regardless of whether the population-flow-based SWM or inverse distance SWM was used, the global Moran's *I* tests for Period I consistently exhibit significantly positive values (see section 4.3), confirming the existence of spatial autocorrelation in the dependent variable. However, as mentioned in section 4.3, LM tests and robust LM tests indicate that spatial models (e.g., SLM, SEM) based on inverse distance SWM fail to capture such spatial autocorrelation in the dependent variable and these models perform no better than the OLS model. On the other hand, spatial models using population-flow-based SWM for the spatial lagged dependent variable successfully address the spatial autocorrelation issue, with the model residuals exhibiting no systematic spatial pattern at a 95 percent confidence level (see Appendix 1).

2) The Period II model

As discussed earlier, our Period II spatial econometric model uses population-flow-based SWM for the spatially lagged dependent variable and inverse distance SWM for the spatially lagged independent variable. To validate the model results, we further estimate the Period II model using only the inverse distance SWM for both spatially lagged dependent and independent variables and compare the performance of these two models (results provided in Appendix 2).

Compared with the model using only inverse distance SWM, the model employing mixed SWMs (i.e., inverse distance SWM for the independent variables and population-flow-based SWM for the dependent variable) has a higher R-square. Although

there is no major difference across the two models in terms of the significance level of explanatory variables, the decomposition of direct and indirect effects does suggest that the model using only inverse distance SWM fails to capture the indirect effects of most explanatory variables, whereas the model employing mixed SWMs has done well. These comparisons thus confirm the robustness of our results.

4.6 Additional robustness checks

Finally, we further conduct two robustness checks using: 1) symmetric population-flow-based SWM[12] derived from the average of population move-in and move-out within each city pair; 2) (asymmetric) population-flow-based SWM with missing values imputed by a gravity model (Wilson 1974)[13], rather than the inverse distance weighted interpolation. Compared with the original asymmetric population-flow-based SWM

---

[12] Based on our original population-flow-based SWM (before normalization), we calculate the value of element in symmetric population-flow-based SWM as:

$$w_{ij} = \frac{1}{2}\left(Moving\_Out\ Index_{ij} + Moving\_Out\ Index_{ji}\right) = w_{ji}$$

where the w_{ij} is the spatial weight of unit *i* towards unit *j*, and w_{ji} is the spatial weight of unit *j* towards unit *i*; $Moving\_Out\ Index_{ij}$ is the index that reflects the volume of population traveling from city *i* to city *j* ; $Moving\_Out\ Index_{ji}$ is the index that reflects the volume of population traveling from city *j* to city *i* .

[13] The gravity model used in this paper can be expressed as follows:

$$w_{ij} = \frac{GDP_i^{\alpha} GDP_j^{\beta} Pop_i^{\varphi} Pop_j^{\omega}}{Distance_{ij}{}^{\tau}}$$

where the w_{ij} is the spatial weight of unit *i* towards unit *j*; GDP_i (GDP_j) is the GDP of city *i* (city *j*); Pop_i (Pop_j) is the total population of city *i* (city *j*); $Distance_{ij}\ is\ the$ geographic distance between city *i* and city j.

We first select city-pairs with accurate move-out indices in our dataset to estimate $\alpha, \beta, \varphi, \omega, and\ \tau$ based on the equation above. We then use the estimation results to impute move-out indices for city-pairs without accurate move-out indices.

elaborated in previous sections, the estimation results using these two alternative SWMs show no major differences in terms of the statistical significance and the coefficient estimates of key explanatory variables but have slightly lower model R-squares (around 5 percent lower) (see Appendices 3 and 4). These comparisons suggest that spatial econometric models using the original asymmetric population-flow-based SWM produce robust results with better goodness of fit, and thus better capture the role of spatial interactions in COVID-19 transmission.

5. Conclusions

This paper proposes an innovative method of constructing SWM based on the real-time population flow data into spatial econometric models. Unlike traditional SWMs which only consider geographic or economic proximity, the matrix in this study introduces travel volume as the measure of connection between city pairs, aiming to better approximate the impact of social interactions. City pairs with more intensive travel interactions are defined as closely connected even if they are geographically far from each other, such as in the case of Sanya and Beijing. Incorporating the population-flow-based SWM into spatial econometric modeling improves classical spatial economic theory by more accurately approximating the spatial interaction processes underlying the spillover effects of spatial outcomes. The proposed method can be applied to a wide range of economic and social research, including environment, public health, demography, and social welfare studies.

In the application part of this study, we adopt spatial econometric models based on the population-flow-based SWM and traditional geographical distance-based SWM to examine the potential transmission determinants of COVID-19 in Mainland China. Our sample is from January 11 to February 25, 2020, covering the critical episodes of both the initial spread and the peak of infections before the first imported case being reported in Mainland China. Considering the changes in population flow from the epicenter of the outbreak as well as the widespread use of prevention and control measures, we divide the whole research period into two subperiods using the date of Wuhan lockdown (January 23, 2020). The results reveal that the advanced health care system played an essential role in the early diagnosis and control of the epidemic. We also find that spatial autocorrelation should be considered when exploring the correlation between meteorological conditions and diseases. Furthermore, the results confirm the significant time-lagged effects of traveler outflows from the outbreak source region on pandemic transmission (Qiu, Chen and Shi, 2020). Interestingly, according to the results of the Period II model, we find that the population outflow from Wuhan is significantly associated with a higher number of local cumulative confirmed cases as well as the number of cases in surrounding cities due to positive spillover effects. These results emphasize the importance of checking the entire records of travel routes for local governments.

There are two limitations in this study. First, due to data availability, we use only Baidu population flow data as our data source of the population flow between and within cities. Although it has been demonstrated that Baidu population flow data captures the real-time

actual population flow through Baidu LBS with relative accuracy, this still excludes the movement of those people who do not use electronic devices, such as elderly people and young children. Second, even though the data of over 200 cities can be obtained from the Baidu Migration Platform, the precise moving-out indices of each source-city are only available for the top 100 destinations. As a result, the data in this paper may not accurately reflect smaller flows of population between cities. Hence, the analytical approach of this paper can still be extended when more reliable data resources become available for future research.

References

Ahlfeldt, G., and A. Feddersen. 2017. From periphery to core: measuring agglomeration effects using high-speed rail. *Journal of Economic Geography* 18 (2):355-390.

Alstott, J., Panzarasa, P., Rubinov, M., Bullmore, E. T., and Vértes, P. E. 2014. A Unifying Framework for Measuring Weighted Rich Clubs. *Scientific Reports* 4(1):1-6.

Andersen, L. M., Harden, S. R., Sugg, M. M., Runkle, J. D., and Lundquist, T. E. 2021. Analyzing the spatial determinants of local Covid-19 transmission in the United States. *Science of The Total Environment* 754:142396.

Anderson, R. M. 2013. *The population dynamics of infectious diseases: theory and applications*. Springer.

Anselin, L. 2001. Spatial econometrics. *A companion to theoretical econometrics*, 310-330.

Aten, B. 1997. Does Space Matter? International Comparisons of the Prices of Tradables and Nontradables. *International Regional Science Review* 20 (1-2):35-52.

Baltagi, B. H. 2008. *Econometric analysis of panel data (Vol. 4)*. Chichester: John wiley & sons.

Black, W. R. 1992. Network autocorrelation in transport network and flow systems. *Geographical Analysis* 24(3): 207-222.

Cao, J., and Zhu, P. 2017. High-speed rail. *Transportation Letters* 9(4): 185-186.

Chen, Y. 2021. An analytical process of spatial autocorrelation functions based on Moran's index. *PLOS ONE* 16 (4): e0249589.

Cheng, W. and Lee, L. F. 2017. Testing endogeneity of spatial and social networks. *Regional science and urban economics* 64: s. 81–97.

Chien, L. C., and Chen, L. W. 2020. Meteorological impacts on the incidence of COVID-19 in the US. *Stochastic Environmental Research and Risk Assessment* 34(10): 1675-1680.

Chun, Y. 2008. Modeling network autocorrelation within migration flows by eigenvector spatial filtering. *Journal of Geographical Systems* 10(4): 317-344.

Cliff, A. D. and Ord, J. K. 1975. Model Building and the Analysis of Spatial Pattern in Human Geography. *Journal of the royal statistical society: series B (methodological)* 37 (3): s. 297–328.

Cohen, J. P., and Paul, C. J. M. 2004. Public infrastructure investment, interstate spatial spillovers, and manufacturing costs. *Review of Economics and Statistics* 86(2): 551-560.

Conley, T. G., and Ligon, E. 2002. Economic distance and cross-country spillovers. *Journal of Economic Growth* 7(2): 157-187.

Conley, T., and Topa, T. 2002. Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics* 17 (4):303-327.

De Jong, P., Sprenger, C., and Van Veen, F. 1984. On extreme values of Moran's *I* and Geary's c. *Geographical Analysis*, *16*(1): 17-24.

Doğan, B., Jebli, M. B., Shahzad, K., Farooq, T. H., and Shahzad, U. 2020. Investigating the effects of meteorological parameters on COVID-19: case study of New Jersey, United States. *Environmental Research* 191: 110148.

Elhorst, J. P. 2010. Applied Spatial Econometrics: Raising the Bar. *Spatial Economic Analysis* 5(1)*:* 9-28.

Elhorst, J. P. 2014. *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. 2014th ed. Berlin: Springer.

Emch, M., E. Root, S. Giebultowicz, M. Ali, C. Perez-Heydrich, and M. Yunus. 2012. Integration of Spatial and Social Network Analysis in Disease Transmission Studies. *Annals of the Association of American Geographers* 102 (5):1004-1015.

Ermagun, A., and Levinson, D. 2018. An introduction to the network weight matrix. *Geographical Analysis* 50(1): 76-96.

Fang, H., Wang, L., and Yang, Y. 2020. Human mobility restrictions and the spread of the novel coronavirus (2019-ncov) in China. *Journal of Public Economics* 191:104272.

Fingleton, B. and Le Gallo, J. 2008. Estimating spatial models with endogenous variables, a spatial lag and spatially dependent disturbances: Finite sample properties*. *Papers in regional science* 87 (3): s. 319–339.

Getis, A. 2009. Spatial Weights Matrices. *Geographical Analysis* 41 (4):404-410.

Getis, A., and Aldstadt, J. 2004. Constructing the Spatial Weights Matrix Using a Local Statistic. *Geographical Analysis* 36 (2):90-104.

Greene, W. 2012. *Econometric Analysis*. 7th Edition, Prentice Hall, Upper Saddle River.

Gujarati, D. N. 2011. *Econometrics by example* (Vol. 1). New York: Palgrave Macmillan.

Homans, G. 2017. *Human Group*. Taylor and Francis.

Hsiao, C. (2014). *Analysis of panel data (No. 54)*. Cambridge university press.

Ispriyanti, D., A. Prahutama, and A. Taryono. 2018. Modeling space of spread Dengue Hemorrhagic Fever (DHF) in Central Java use spatial Durbin model. *Journal of Physics: Conference Series* 1025:012112.

Jia, J., X. Lu, Y. Yuan, G. Xu, J. Jia, and N. Christakis. 2020. Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature* 582 (7812):389-394.

Kordi, M., and A. Fotheringham. 2016. Spatially Weighted Interaction Models (SWIM). *Annals of the American Association of Geographers* 106 (5):990-1012.

Kostov, P. 2010. Model Boosting for Spatial Weighting Matrix Selection in Spatial Lag Models. *Environment and Planning B: Planning and Design* 37 (3):533-549.

Krivoruchko, K., and A. Gribov. 2019. Evaluation of empirical Bayesian kriging. *Spatial Statistics* 32:100368.

Lai, S., I. Bogoch, A. Watts, K. Khan, Z. Li, and A. Tatem. 2020. *Preliminary risk analysis of 2019 novel coronavirus spread within and beyond China*. https://www.worldpop.org/resources/docs/china/WorldPop-coronavirus-spread-risk-analysis-v1-25Jan.pdf. (Accessed 18 March 2020)

Lam, C., and Souza, P. C. 2020. Estimation and selection of spatial weight matrix in a spatial lag model. *Journal of Business & Economic Statistics* 38(3): 693-710.

Leenders, R. 2002. Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks* 24 (1):21-47.

LeSage, J. 2008. An Introduction to Spatial Econometrics. *Revue d'économie industrielle* 123:19-44.

Li, J., and Li, S. 2020. Energy investment, economic growth and carbon emissions in China—Empirical analysis based on spatial Durbin model. *Energy Policy* 140:111425.

Li, S. 2017. Construction and Application of Spatial Weight Matrix Based on Urban Flows Intensity (in Chinese). *Statistics and Decision* 24: 70-73.

Liu, K., S. Ai, S. Song, G. Zhu, F. Tian, H. Li, Y. Gao, Y. Wu, S. Zhang, Z. Shao, et al. 2020. Population Movement, City Closure in Wuhan, and Geographical Expansion of the COVID-19 Infection in China in January 2020. *Clinical Infectious Diseases* 71 (16):2045-2051.

Lu, J., and Zhang, L. 2011. Modeling and Prediction of Tree Height–Diameter Relationships Using Spatial Autoregressive Models. *Forest Science* 57 (3):252-264.

Lv, Y., Chen, W., and Cheng, J. 2019. Direct and Indirect Effects of Urbanization on Energy Intensity in Chinese Cities: A Regional Heterogeneity Analysis. *Sustainability* 11 (11):3167.

McClymont, H., and Hu, W. 2021. Weather variability and COVID-19 transmission: A review of recent research. *International journal of environmental research and public health* 18(2): 396.

Moran, P. A. 1950. Notes on continuous stochastic phenomena. *Biometrika* 37(1/2): 17-23.

Harmooshi, N. N., Shirbandi, K., & Rahim, F. 2020. Environmental concern regarding the effect of humidity and temperature on 2019-nCoV survival: fact or fiction. *Environmental Science and Pollution Research* 27(29): 36027-36036.

Oyedotun, T. D. T., and Moonsammy, S. 2021. Spatiotemporal variation of COVID-19 and its spread in South America: A rapid assessment. A*nnals of the American Association of Geographers* 111(6): 1868-1879.

Paelinck, Jean H. P. and Klaassen, Leo H. 1979. *Spatial econometrics*. Farnborough, Eng: Saxon House

Parent, O., and J. LeSage. 2006. Using the Variance Structure of the Conditional Autoregressive Spatial Specification to Model Knowledge Spillovers. *Journal of applied Econometrics 23*(2): 235-256.

Perret, J. K. 2011. *A proposal for an alternative spatial weight matrix under consideration of the distribution of economic activity (No. 2011-002)*. Schumpeter Discussion Papers.

Pietrzak, M. 2010. Application of economic distance for the purposes of a spatial analysis of the unemployment rate for Poland. *Oeconomia Copernicana* 1 (1):79-98.

Qiu, Y., Chen, X., and Shi, W. 2020. Impacts of social and economic factors on the transmission of coronavirus disease 2019 (COVID-19) in China. *Journal of Population Economics* 33(4): 1127-1172.

Sannigrahi, S., Pilla, F., Basu, B., Basu, A. S., and Molter, A. 2020. Examining the association between socio-demographic composition and COVID-19 fatalities in the European region using spatial regression approach. *Sustainable Cities and Society* 62:102418.

Seldadyo, H., Elhorst, J. P., and Haan, J. D. 2010. Geography and governance: Does space matter?. *Papers in Regional Science* 89(3):625-640.

Seya, H., Y. Yamagata, and M. Tsutsumi. 2013. Automatic selection of a spatial weight matrix in spatial econometrics: Application to a spatial hedonic approach. *Regional Science and Urban Economics* 43 (3):429-444.

Shi, P., Y. Dong, H. Yan, C. Zhao, X. Li, W. Liu, M. He, S. Tang, and S. Xi. 2020. Impact of temperature on the dynamics of the COVID-19 outbreak in China. *Science of The Total Environment* 728:138890.

Snyder, H. 1995. The Network Nation: Human Communication Via Computer. *Information Processing and Management* 31 (2):265-266.

StataCorp. 2017. Stata 15 Base Reference Manual. College Station, TX: Stata Press.

Szabó, Z., and Á. Török. 2019. Spatial Econometrics – Usage in Transportation Sciences: A Review Article. *Periodica Polytechnica Transportation Engineering* 48 (2):143-149.

Tobler, W. 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 46:234.

Transmission of SARS-CoV-2: implications for infection prevention precautions: scientific brief. 2020. https://apps.who.int/iris/bitstream/handle/10665/333114/WHO-2019-nCoV-Sci_Brief-Transmission_modes-2020.3-rus.pdf (last accessed 6 July 2021).

Vega, S. H., and Elhorst, J. P. (2013, August). On spatial econometric models, spillover effects, and W. In *53rd ERSA Congress, Palermo, Italy*.

Webber, M. 1964. The Urban Place and the Nonplace Urban Realm. *Explorations into Urban Structure*.

Wei, S., and Wang, L. 2020. Examining the population flow network in China and its implications for epidemic control based on Baidu migration data. *Humanities and Social Sciences Communications* 7 (1): 1-10.

Wei, Y., Wang, J., Song, W., Xiu, C., Ma, L., and Pei, T. 2021. Spread of COVID-19 in China: analysis from a city-based epidemic and mobility model. *Cities* 110:103010.

Wei, Y., W. Song, C. Xiu, and Z. Zhao. 2018. The rich-club phenomenon of China's population flow network during the country's spring festival. *Applied Geography* 96:77-85.

Wellman, B., and Leighton, B. 1979. Networks, Neighborhoods, and Communities. *Urban Affairs Quarterly* 14 (3):363-390.

Wilson, A. G. 1974. *Urban and regional models in geography and planning*. London, UK: John Wiley & Sons Incorporated.

Xu, B., Gong, P., Seto, E., Liang, S., Yang, C., Wen, S., Qiu. D., Gu, X., and Spear, R. 2006. A Spatial-Temporal Model for Assessing the Effects of Intervillage Connectivity in Schistosomiasis Transmission. *Annals of the Association of American Geographers* 96 (1):31-46.

Yang, Z., Zeng, Z., Wang, K., Wong, S. S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao, Z. Mai, Z et al. 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of thoracic disease* 12(3): 165.

Yu, T., Z. Chen, and Zhu, P. 2012. Characteristics and mechanism of high speed rail-driven suburbanization in China: A case study of Beijing-Shanghai high-speed rail. *Scientia Geographica Sinica*, 32 (9), 1041-1046.

Zhai, W., Liu, M., Fu, X., & Peng, Z. R. 2021. American inequality meets COVID-19: Uneven spread of the disease across communities. *Annals of the American Association of Geographers* 111(7): 2023-2043.Zhang, J. W., Chen, X., and Wang, S. Y. 2009. New Spatial Weight Matrix and Its Application in China's Regional Foreign Trade. *Systems Engineering-Theory and Practice* 29: 84-92.

Zhao, S., Zhuang, Z., Ran, J., Lin, J., Yang, G., Yang, L., and He, D. 2020. The association between domestic train transportation and novel coronavirus (2019-nCoV) outbreak in China from 2019 to 2020: A data-driven correlational report. *Travel Medicine and Infectious Disease* 33:101568.

Zheng, D., Hao, S., Sun, C., and Lyu, L. 2019. Spatial Correlation and Convergence Analysis of Eco-Efficiency in China. *Sustainability* 11 (9):2490.

Zhong, S., and Bian, L. 2016. A Location-Centric Network Approach to Analyzing Epidemic Dynamics. *Annals of the American Association of Geographers*:1-9.

Zhou, J., Zhang, M., and Zhu, P. 2019. The equity and spatial implications of transit fare. *Transportation Research Part A: Policy and Practice*, 121 (2019), pp. 309-324.

Zhu, P. 2011. *Telecommuting, Travel Behavior and Residential Location Choice: Can Telecommuting be an Effective Policy to Reduce Travel Demand?* PhD thesis, University of Southern California.

Zhu, P., Yu, T., and Chen, Z. 2015. High-speed rail and urban decentralization in China. *Transportation Research Record: Journal of the Transportation Research Board* (2015), pp. 16-26

Zhu, P., Ho, S. N., Jiang, Y., and Tan, X. 2020. Built environment, commuting behaviour and job accessibility in a rail-based dense urban context. Transportation Research Part D: Transport and Environment, 87, 102438.

Zhu, P. 2021. Does high-speed rail stimulate urban land growth? Experience from China. *Transportation Research Part D: Transport and Environment* 98: 102974.

Zhu, P., and Guo, Y. 2021. The Role of High-speed Rail and Air Travel in the Spread of COVID-19 in China. *Travel medicine and infectious disease*: 102097.

Zhu, P., and Tan, X. 2021. Is compulsory home quarantine less effective than centralized quarantine in controlling the COVID-19 outbreak? Evidence from Hong Kong. *Sustainable Cities and Society*, 74, 103222.

Pengyu Zhu is an associate professor in the Division of Public Policy at the Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR. E-mail: pengyuzhu@ust.hk. His research interests include the big data and urban planning,

sustainable transportation, economic development policy, housing and land use policy, and migration and employment.

Jiarong Li is a PHD student in the Society Hub at the Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR. E-mail: jligj@connect.ust.hk. Her research interests include the spatial data analysis, urban transportation and spatial econometrics.

Yuting Hou is an assistant professor in the Department of Building and Real Estate at the Hong Kong Polytechnic University, Kowloon, Hong Kong SAR. E-mail: yuting.hou@polyu.edu.hk. Her research interests include the interactions between land use and transportation, urban and regional economics, and applied spatial analysis.

[Appendix 1 Global Moran's *I* of residuals based on Population-flow weight matrix]

Appendix 2 Period II SDM using only the inverse distance SWM for both spatially lagged dependent and independent variables

[Appendix 2.1 Coefficient Estimates]

[Appendix 2.2 Decomposition of the spatial effects]

[Appendix 3 Estimated spatial effects using symmetric population-flow-based SWM derived from the average of population move-in and move-out within each city pair]

[Appendix 4 Estimated spatial effects using asymmetric population-flow-based SWM with missing values imputed by a gravity model]

1

2 Table 1 Descriptive Statistics of the Dependent and Independent Variables

Variable	Note	Obs.	Mean	Std. Dev.	Min	Max
Total number of confirmed cases in Period I	Number of cumulative confirmed cases in Period I	272	3.165	30.139	0	495
Total population	Total population of urban area (ten thousand)	272	174.529	234.2	16	2465
GDP per capita	GDP per capita (RMB)	272	74992.32	37511.13	19212	217313
Doctors per capita	Licensed doctors per capita	272	0.004	0.002	0.001	0.010
Average temperature during Period I	Average daily temperature during Period I (℃)	272	2.02	8.92	-18.90	22.56
Average humidity during Period I	Average daily relative humidity during Period I (%)	272	75.18	9.499	32.08	90.19
Average wind speed during Period I	Average daily wind speed during Period I (m/s)	272	1.90	0.36	1.24	3.20
Average air quality index during Period I	Average air quality index during Period I	272	95.93	46.69	26.40	237.00
Within-city population flow index during Period I	Index of travel intensity within cities during Period I	272	5.33	0.66	2.89	7.18
Population from Wuhan during Period I	Population from Wuhan during Period I (thousand)	272	206.54	774.21	0	8024.92
Total number of confirmed cases in Period II	Number of cumulative confirmed cases in Period II	272	271.39	2859.74	0	46946
Average temperature during Period II	Average daily temperature during Period II (℃)	272	4.48	7.62	-15.17	20.79
Average humidity during Period II	Average daily relative humidity during Period II (%)	272	69.10	11.16	29.30	84.61
Average wind speed during Period II	Average daily wind speed during Period II (m/s)	272	2.21	0.425	1.24	3.91
Average air quality index during Period II	Average air quality index during Period II	272	68.42	26.42	24.65	140.28
Within-city population flow index in Period II	The average index of travel intensity within cities in Period II	272	2.65	0.70	0.65	4.96

Table 2 Results of Global Moran's *I* tests of spatial autocorrelation for dependent variables based on Population-flow weight matrix

	Global Moran's *I statistics*
Total number of confirmed cases during Period I	0.48 ***
Total number of confirmed cases during Period II	0.21***

Table 3 Results of LM tests, Robust LM tests and LR tests

Model specification	**Period I**	**Period II**
Weight matrix: *Population-flow weight matrix*		
LM-lag test	11.39***	8.94***
LM-error test	6.93***	7.79***
Robust LM-lag test	14.44***	5.95***
Robust LM-error test	3.2***	4.79**
Weight matrix: Inverse *geographic distance matrix*		
LM-lag test	0.01	54.92***
LM-error test	0.16	34.81***
Robust LM-lag test	0.18	23.36***
Robust LM-error test	0.34	3.24**
LR test statistics between SDM and SLM	12.99	126.24***

Note: 1. *** p<0.01, ** p<0.05, * p<0.1. This paper uses the traditional cutoff value of p<0.05 to determine statistical significance.

1

2 Table 4 Empirical results for OLS and spatial econometric models

	Period I (before lockdown) Cumulative cases (January 11 - January 23)		**Period II (after lockdown) Cumulative cases (January 24 - February 25)**	
MODEL	OLS	SLM	OLS	SDM
Total population (log)	0.44***	0.34***	0.46***	0.42***
GDP per capita (log)	0.12	0.16*	-0.17	-0.11
Doctors per capita (log)	117.70***	99.64***	8.34	32.01
Average temperature	0.01	-0.01	0.02**	0.06***
Average wind speed	0.11	0.18	-0.12	0.12
Average humidity	0.01**	0.01**	0.04***	0.03***
Average air quality index	-0.01***	-0.01***	0.01**	0.01
Within-city population flow index	-0.05	-0.04	-0.68***	-0.56***
Population from Wuhan during Period I	0.01***	0.01***	0.01***	0.01***
Total number of confirmed cases during Period I (log)			0.41***	0.32***
Constant	-4.06***	-4.20***	1.60*	-22.85**
W*Total population (log)				-1.70*
W*GDP per capita (log)				3.78***
W*Doctors per capita (log)				-1,792.00***
W*Average temperature				-0.20**
W*Average wind speed				-0.84
W*Average humidity				-0.07
W*Average air quality index				0.05**
W*Within-city population flow index				-0.33
W*Population from Wuhan during Period I				0.01***
W_G*Total number of confirmed cases during Period I (log)				2.48**
W_P*Total number of confirmed cases during Period I (log)		0.74***		
W_P*Total number of confirmed cases during Period II (log)				0.30**
R-square	0.45	0.50	0.67	0.79

Observations	272	272	272	272

Note: *** p<0.01, ** p<0.05, * p<0.1. This paper uses the traditional cutoff value of p<0.05 to determine statistical significance.

Table 5 Estimation results for decomposition of the spatial effects

	Effect	**Period I (before lockdown) cumulative cases (January 11 -January 23)**	**Period II (after lockdown) cumulative cases (January 24 - February 25)**
Total population (log)	Direct	0.34***	0.42***
	Indirect	0.07**	-1.78*
	Total	0.41***	-1.36
GDP per capita (log)	Direct	0.16*	-0.11
	Indirect	0.03	4.02***
	Total	0.19*	3.92
Doctors per capita (log)	Direct	99.96***	30.15
	Indirect	20.41**	-1907.21***
	Total	120.37***	-1877.05***
Average temperature	Direct	-0.01	0.06***
	Indirect	-0.01	-0.21**
	Total	-0.02	-0.16**
Average wind speed	Direct	0.18	0.11
	Indirect	0.04	-0.89
	Total	0.21	-0.78*
Average humidity	Direct	0.01**	0.03***
	Indirect	0.01*	-0.07
	Total	0.02**	-0.04*
Average air quality index	Direct	-0.01***	0.01
	Indirect	-0.01***	0.05**
	Total	-0.02***	0.05**
Within-city population flow index	Direct	-0.04	-0.55***
	Indirect	-0.01	-0.39
	Total	-0.05	-0.94
Population from Wuhan during Period I	Direct	0.01***	0.01***
	Indirect	0.01**	0.01***
	Total	0.02***	0.02***
Total number of confirmed cases during Period I (log)	Direct		0.33***
	Indirect		2.66**

Total	2.98**

Note: *** p<0.01, ** p<0.05, * p<0.1. This paper uses the traditional cutoff value of p<0.05 to determine statistical significance.

Appendix 1 Global Moran's *I* of residuals based on Population-flow weight matrix

	Global Moran's I	**P-value**
Residuals of SLM for Period I (before lockdown)	-0.05	0.06
Residuals of SDM for Period II (after lockdown)	0.04	0.08

Appendix 2 Period II SDM using only the inverse distance SWM for both spatially lagged dependent and independent variables

Appendix 2.1 Coefficient Estimates

MODEL	SDM
Total population (log)	0.56***
GDP per capita (log)	-0.15
Doctors per capita (log)	53.58
Average temperature	0.06***
Average wind speed	0.05
Average humidity	0.03**
Average air quality index	0.01
Within-city population flow index	-0.53***
Population from Wuhan during Period I	0.01***
Total number of confirmed cases during Period I (log)	0.38***
Constant	-18.59*
W*Total population (log)	-1.75*
W*GDP per capita (log)	3.20***
W*Doctors per capita (log)	-1,730.00***
W*Average temperature	-0.23***
W*Average wind speed	-0.97
W*Average humidity	-0.07
W*Average air quality index	0.04*
W*Within-city population flow index	0.09
W*Population from Wuhan during the Period I	0.01***
W*Total number of confirmed cases during Period I (log)	2.51**
W*Total number of confirmed cases during Period II (log)	0.78***
R-square	0.75
Observations	272

Appendix 2.2 Decomposition of the spatial effects

	Effect	Coefficient
Total population (log)	Direct	0.53***
	Indirect	-5.91
	Total	-5.37
GDP per capita (log)	Direct	-0.09
	Indirect	13.85
	Total	13.76
Doctors per capita (log)	Direct	21.68
	Indirect	-7580.86
	Total	-7559.18
Average temperature	Direct	0.06***
	Indirect	-0.81
	Total	-0.75
Average wind speed	Direct	0.03
	Indirect	-4.16
	Total	-4.12
Average humidity	Direct	0.03**
	Indirect	-0.22
	Total	-0.20
Average air quality index	Direct	0.01
	Indirect	0.18
	Total	0.18
Within-city population flow index	Direct	-0.53***
	Indirect	-1.43
	Total	-1.96
Population from Wuhan during Period I	Direct	0.01***
	Indirect	0.01
	Total	0.01
Total number of confirmed cases during Period I (log)	Direct	0.43***
	Indirect	12.62
	Total	13.05

1

2

Appendix 3 Estimated spatial effects using symmetric population-flow-based SWM derived from the average of population move-in and move-out within each city pair

	Effect	Period I	Period II
Total population (log)	Direct	0.34***	0.49***
	Indirect	0.09**	-1.77*
	Total	0.40***	-1.28
GDP per capita (log)	Direct	0.13	-0.14
	Indirect	0.04	3.82***
	Total	0.17	3.68***
Doctors per capita (log)	Direct	85.98***	37.55
	Indirect	26.51**	- 2008.35***
	Total	112.49***	-1970.81***
Average temperature	Direct	-0.01	0.06***
	Indirect	-0.01	-0.18**
	Total	-0.02	-0.13**
Average wind speed	Direct	0.18	0. 06
	Indirect	0.05	-0.78
	Total	0.23	-0.71
Average humidity	Direct	0.01**	0.03***
	Indirect	0.01*	-0.06
	Total	0.02**	-0.03
Average air quality index	Direct	-0.01***	0.01
	Indirect	-0.01**	0.06***
	Total	-0.01***	0.06***
Within-city population flow index	Direct	0.01	-0.51***
	Indirect	0.01	-0.69
	Total	0.02	-1.20
Population from Wuhan during Period I	Direct	0.01***	0.01***
	Indirect	0.01**	0.01***
	Total	0.01***	0.02***
Total number of confirmed cases during Period I (log)	Direct		0.35***
	Indirect		2.77***
	Total		3.12**
R^2		0.47	0.76

Note: *** p<0.01, ** p<0.05, * p<0.1.

Appendix 4 Estimated spatial effects using asymmetric population-flow-based SWM with missing values imputed by a gravity model

	Effect	Period I	Period II
Total population (log)	Direct	0.31***	0.46***
	Indirect	0.09***	-1.62
	Total	0.40***	-1.16
GDP per capita (log)	Direct	0.14	-0.16
	Indirect	0.04	1.33**
	Total	0.18	1.17
Doctors per capita (log)	Direct	90.33***	41.24
	Indirect	27.61**	-1253.95**
	Total	117.94***	-1212.70***
Average temperature	Direct	-0.01	0.03**
	Indirect	-0.01	-0.22**
	Total	-0.01	-0.19**
Average wind speed	Direct	0.16	0. 09
	Indirect	0.05	-0.19
	Total	0.21	-0.11
Average humidity	Direct	0.01*	0.03***
	Indirect	0.01*	-0.02
	Total	0.02**	-0.01
Average air quality index	Direct	-0.01***	-0.01
	Indirect	-0.01**	0.02
	Total	-0.01***	0.01
Within-city population flow index	Direct	-0.01	-0.55***
	Indirect	-0.01	-0.45
	Total	-0.01	-0.99
Population from Wuhan during Period I	Direct	0.01***	0.01***
	Indirect	0.01**	0.01***
	Total	0.01***	0.02***
Total number of confirmed cases during Period I (log)	Direct		0.36***
	Indirect		2.60***
	Total		2.96***
R^2		0.48	0.76

Note: *** p<0.01, ** p<0.05, * p<0.1.