Cities 109 (2021) 103041

Contents lists available at ScienceDirect

Cities

journal homepage: www.elsevier.com/locate/cities

Measuring interaction among cities in China: A geographical awareness approach with social media data

Xinyue Ye [a], Shengwen Li [b], Qiong Peng [c,*]

[a] Department of Landscape Architecture and Urban Planning, Texas A&M University, USA
[b] School of Geography and Information Engineering, China University of Geosciences, China
[c] Department of Urban Studies and Planning, National Center for Smart Growth, University of Maryland, College Park, 1219D Preinkert Field House, College Park, MD 20742, USA

ARTICLE INFO

Keywords:
Geographical awareness
Spatial out-awareness rate index
Spatial in-awareness index
Social media
China
COVID-19

ABSTRACT

Unlike the large body of research on investigating interactions among cities using survey data, the social media-based city interaction study has received much less exploration. Based on geographical studies of social media content in China, we develop a few indices quantifying various levels of geographical awareness among cities. (1) We find that the geographical awareness proxy by the social media-based indices can measure interactions among cities. Specifically, the geographical awareness among cities follows gravitational law and is highly correlated with mobility flows. (2) The spatial in-awareness index (SIAI) is an appropriate index indicating a city's ranking in the urban hierarchy (3) the spatial out-awareness rate (SOAR) can indicate the interactions from a focal city to other cities. Our findings also show that SOAR can predict the number of people infected during a pandemic in a city system. Once the origin city or hotspots of the outbreak and the number of infected persons within those cities are known, we can use the social media-based SOAR index to predict number of cases for other else cities in the urban system. With this information, governments can properly and efficiently deliver medical equipment and staff to cities where large populations are infected.

1. Introduction

Big data provides enormous opportunities for studying place awareness and spatial interactions between cities. As defined in existing literature (Han et al., 2015, Meijers & Peris, 2019), geographic awareness refers to the comprehensive interactions between cities, based on references made to the other city via social media data. Place name information – referred to as a toponym – and users' geographical information included in social media posts can be used to quantify geographical awareness. Geographical awareness among cities reflects overall interactions. A greater understanding of geographical awareness will have direct and important city policy implications. For example, the spatial in-awareness index (SIAI) can serve as an indicator of a city's influence and ranking in city hierarchy. City rankings based solely on GDP or population miss other important factors, such as if a city is a regional rail transit center or if a city is a national historical city that draws a large number of tourists. The SIAI of a city indicates the overall influence of the city within its own hierarchical city system. The SIAI takes into account observed connections (e.g. traffic flows) and unobserved connections (e.g. cultural influences), which make SIAI an appropriate index of a city's overall influence within a city system. Another example is the spatial out-awareness rate (SOAR). SOAR measures outward interactions from a city to other cities, and these interactions may include observed connections (e.g. highway, rail, and flight) and unobserved connections. As such, this index can be used to predict the number of people who may become infected with an epidemic disease, once the origin city and number of infected persons within that city have been identified, and taken into account along with city infrastructure/health response/emergency response insights.

In the case of the 2019 novel coronavirus (COVID-19), SOAR can be applied to predict how many city residents may become infected. By examining SOAR between hotspot cities of the pandemic, we can predict the number of persons who may become infected outside the hotspot cities if conditions within these cities remain unchanged. These predictions could be invaluable, as both government and non-governmental organizations can properly and efficiently deliver medical equipment and staffs to cities where many may become infected with the disease.

Compared with survey data, big data (such as social media data or

* Corresponding author.
E-mail addresses: xinyue.ye@tamu.edu (X. Ye), swli@cug.edu.cn (S. Li), xqpeng@umd.edu (Q. Peng).

https://doi.org/10.1016/j.cities.2020.103041
Received 15 November 2019; Received in revised form 15 October 2020; Accepted 6 November 2020

cell phone positioning data) presents advantages in terms of quantifying geographical awareness. For example, social media data covers a large sample size, offers a high temporal resolution, and is accessible at a low cost. Such data can prove useful in quantifying geographical awareness, even despite the fact that the data may suffer from representative bias; that is, younger adults are significantly more likely to use social media platforms than the elderly. Cell phone positioning data could present an alternative if the cost of collecting the data were affordable and if the public did not have privacy concerns regarding the use of this data.

Despite the advantages of social media data and the significance of geographical awareness research, few studies of geographical awareness using social media data have been conducted. In particular, there are few studies that develop indices to proxy geographical awareness and test if the geographical awareness among cities can measure interactions among cities. This study attempts to fill these research gaps.

In this study, a couple of social media-based geographical awareness indices are first quantified and tested using an econometric model to determine whether the geographical awareness among cities follows a gravity model. We verify that the social media based indices can measure interactions among cities using mobility flow data. Next, we show that the spatial in-awareness index (SIAI), can indicate a city's ranking in its own city system. Finally, once the origin city or hotspots of the outbreak and the number of infected persons within those cities are known, the social media-based SOAR index can predict number of cases for other else cities in the urban system.

This paper proceeds as follows. We first review previous studies that define geographical awareness, measure geographical awareness, and explore interactions among cites using various data sources. We describe the data employed and follow this with an introduction of a framework of proposed social media-based geographical awareness indices. We introduce tests, report the results, and verify the results with another data-mobility flow data. Finally, we conclude with policy implications.

2. Literature review

2.1. Definition of geographical awareness

Geographical awareness – also referred to as geographic awareness or, roughly, spatial contextual awareness – is defined in terms of cognitive processes. It permits a unique, user-centered perspective upon which "conceptualizations imbue spatial structures with meaning" (Freksa et al., 2007). A narrow definition of geographical awareness omits the individual cognitive and computational functions involved in a complex geographic system but renders it as the specification of a point location on Earth. Since it is not easy to implement the definition of geographical awareness in practice, in some studies, researchers use more generalized versions of the definition. Such versions of the definition include the computation of the frequency of a place as mentioned by an individual or a group. Referencing a place by name on Twitter implies a certain level of awareness of that place (e.g. Han et al., 2015). However, it is imperative to note that geographical awareness does not equate geographical knowledge. For example, social media users may reference a given location name – such as New York City – in a post, but they might not know the exact geographic location of that city, nor might they know other geographic details about the city. Instead, users may just know the city's name or have a general awareness of the city.

2.2. Measuring geographical awareness using social media data

Big data (e.g. social media data) provides us with invaluable opportunities to explore human behavior and social phenomena from the individual perspective to the population level. For example, social media data were broadly used to predict the potential impact of natural disasters – such as forest fires – and examine social network dynamics (Wang et al., 2019; Wang & Ye, 2019; Yue et al., 2019). The extensive use of social media data also offers us myriad opportunities to quantify geographical awareness. Traditionally, researchers use a mental map approach to evaluate one's level of geographical literacy or awareness (Beatty & Tröster, 1987; Chiodo, 1993). This approach is time-consuming and expensive, and the information it yields is quite arbitrary. Recently, researchers started using social media data to measure geographical awareness. While there has been a growing number of studies that quantify geographical awareness using social media data, such studies are limited. In some cases, scholars study the frequency of references to a place by name to indicate the level of geographical awareness a citizen has of that place. For example, Han et al. (2015) estimated the level of geographical awareness of cities for U.S.-based Twitter users. Xu et al. (2013) used tweets to assess geographical awareness characteristics of a biased sample population. Still, an in-depth assessment of individuals' geographical awareness would require more information than their references of place names (Xu et al., 2013). To use social media data in research, we should keep in mind that social media data – such as data derived from Twitter and Facebook – are criticized for being misrepresentative of residents and the data including a large amount of noise.

2.3. Evaluating geographical awareness from city perspective

There are two main perspectives from which to examine geographical awareness using social media data: the individual perspective and the city perspective. Based on Twitter data, Xu et al. use three indicators (mean center, median center, and standard deviation ellipse) to assess geographical awareness characteristics by individual (Xu et al., 2013). Examining geographical awareness from the individual perspective could offer very valuable insights; however, as Xu et al. (2013) mention, "Assessing the levels of geographical awareness may not have immediate practical values from an individual perspective as compared to the monitoring of social events or hazardous incidences." Furthermore, the value of geographical awareness at the individual level could be unstable as it varies significantly from individual to individual. Another perspective from which to examine geographical awareness is the city perspective. For example, Han et al. (2015) introduce Knowledge Discovery in Cyberspace for Geographical Awareness (KDCGA) to estimate the level of geographical awareness of Twitter users across 50 U.S. cities. Greater understanding of geographical awareness at the city level will have direct and important urban policy implications. For example, collective geographical awareness between cities can serve as an indicator of overall interactions between these cities, and overall, more studies that examine geographical awareness from the city perspective are needed.

2.4. Spatial interactions between cities and geographical awareness

Scholars of regional science, urban studies, and geography have long been interested in the spatial interactions between cities and regions as they convey the spatial structure of a region. Spatial interaction in this context refers to physical interactions, such as movements of people and cargo between places. The spatial interaction can also refer to virtual connections, such as information communication between places. We label the combination of physical interaction and virtual interaction between places as comprehensive spatial interactions. Yet, measuring the comprehensive spatial interactions between cities and regions is difficult. One of the barriers to studying spatial interactions is the challenge of using data to measure the strength of interactions between cities. Conventionally, interaction strength has been measured by volumes of passengers between two cities (Xiao et al., 2013), migration flows (Flowerdew & Lovett, 1988), trade flow (Hesse, 2010), and telecommunications (Guldmann, 1999). But more recently, social media check-in data (Liu, Sui, et al., 2014) and toponym co-occurrences (Liu, Wang, et al., 2014) have been used to measure interaction strength as well.

Researchers say that there is a close relationship between geographical awareness (roughly referred to as geographical

information) and individual spatial behavior (Richards, 1974). To start, geographical awareness levels may be affected by individual spatial activity and behavior (Golledge & Stimson, 1997; Horton & Reynolds, 1971). Thus, geographical awareness between cities, in some sense, can be used as an indicator of spatial interactions between cities. Evaluating geographical awareness based on social media offers a new opportunity to examine spatial interactions between cities. This new perspective is worth particular attention given the fact that the social media data upon which geographical awareness is assessed are more accessible than other indicators, such as data on the volumes of passengers between cities, migration flows, trade flow, etc.

Toponyms (place names) in social media are used to investigate urban spatial patterns and relationships (Hu, Ye, & Shaw, 2017; Meijers & Peris, 2019). Liu, Wang, et al. (2014) used this data to examine the connective strength and differences between geographical entities while Lin and Li (2015) used it to simulate urban growth in metropolitan areas. Also, toponym data have been used to determine social media users' geographical awareness of U.S. cities (Han et al., 2015). A greater value of awareness between two cities indicates a high level of interaction between the two cities.

3. Data

Analog to Twitter, Sina microblog (also labeled Sina Weibo) is a leading social networking service in China that allows users to publicly post very short messages and receive messages. Sina microblog is the leading social media platform in mainland China. It must be remembered that neither Twitter nor Facebook have an official presence in China. Researchers have conducted many studies about social networks using Sina microblog data, observed individual behaviors, and carried out sentiment analysis (Chen & She, 2012; Guo et al., 2011; Wang et al., 2013; Yan et al., 2013).

This paper adopts two types of spatial information from Sina microblog: toponym information from microblog messages and location information included in microblog messages. A post with text that contains the toponym of city j, posted from city i represents an instance of interaction from city i to city j. We refer to the total number of posts containing the toponym of city j posted from city i as geographical awareness from city i to city j (row 1 of Table 1).

For this study, three steps were taken to collect, clean, and compile social media data related to Sina microblog text posts. First, text messages posted between the dates of January 1, 2015 and December 31, 2015 were collected using Application Programming Interfaces (APIs). Second, we omitted any text posts by users who did not utilize geographical tags on their accounts. Instead, we used only posts by users with geographical tags from any of the 32 capital cities collected. Afterwards, we counted the number of text posts by microblog user cities and by cities referenced in posted text messages. Unlike the United States, the most influential cities in China are overwhelmingly provincial capital cities. This is because these province capitals represent a large share of China's population as well as GDP within their respective province. These cities also attract a large number of businesses and immigrants, and many are experiencing rapid urbanization. Among them, Beijing, Shanghai, Tianjin, and Chongqing are federal municipalities that play significant roles in the Chinese economic and political systems. Fig. 1 shows the location of these 32 capital cities. We also collected information regarding Chinese capital city GDP, population, and average employee salary using Chinese City Survey 2016 data as issued by the Chinese Census Bureau.

4. A framework

To quantify geographical awareness for cities, Gong et al. (2020) list a few indices. Based on these indices, we add awareness of city i to city j to the list of indices and propose the formulas for spatial in-awareness rate (SIAR) and spatial out-awareness rate (SOAR), as shown in Table 1. The basic index is an awareness between two cities, which is calculated by the total number of posts in a city and the total number of posts that contain the toponym of another city (see row 1 of Table 1). A greater value of awareness between two cities suggests that there is a high level of interaction between the two cities. Besides the basic index of geographical awareness, Table 1 lists other spatial dimensions, such as the city relatedness degree, city relatedness rate, city relatedness index, spatial relatedness rate, and spatial relatedness index (see Table 1). In this paper, we focus on the index of geographical awareness between two cities, SOAR, and SIAI. In the next section, we develop equations to quantify these three indices and apply them in the empirical study.

Table 1
The indices quantifying geographical awareness.

Spatial dimension	Indices	Description
Relatedness degree (between two cities)	Awareness of city i to city j	The total number of posts in city i containing the toponym of city j
City relatedness degree	In-awareness	The total number of awareness for a city received from other cities.
	Out-awareness	The total number of awareness for a city send to other cities.
	Local-awareness	The total number of awareness for a city received from itself.
City relatedness rate	In-awareness rate (IAR)	The in-awareness of a city from another city relative to its local-awareness.
	Out-awareness rate (OAR)	The out-awareness of a city to another city relative to its local-awareness.
City relatedness index	In-awareness index (IAI)	The sum of the in-awareness rate for a city
	Out-awareness index (OAI)	The sum of the out-awareness rate for a city
Spatial relatedness rate	Spatial in-awareness rate (SIAR)	The in-awareness rate to spatial distance between two cities.
	Spatial out-awareness rate (SOAR)	The out-awareness rate to spatial distance between two cities.
Spatial relatedness index	Spatial in-awareness index (SIAI)	The sum of the spatial in-awareness rate for a city
	Spatial out-awareness index (SOAI)	The sum of the spatial out-awareness rate for a city

5. Methods and results

We first tested whether geographical awareness between cities follows gravitational law, estimated the decay function parameter, developed a spatial in-awareness index (SIAI), and calculated values of SIAI for all Chinese province capital cities using social media data and the estimated decay function parameter. Next, we tested whether SIAI could be used to indicate a city's ranking within its own city hierarchy. In addition, we tested that the social media-based SOAR index can predict number of cases for other else cities in the urban system, once the origin city or hotspots of the outbreak and the number of infected persons within those cities are known.

5.1. Fitting the gravity model

Spatial interaction systems can be generally expressed using the gravity model. Gravity models with friction are derived by Newton's law of gravitation, which take the following form:

$$X_{ij} = G\frac{Y_i^{\beta_1} Y_j^{\beta_2}}{\left(D_{ij} * exp\left(Z_{ij}\right)\right)^{\beta_3}} \quad (1)$$

where X_{ij} indicates spatial interaction strength from i to j; G is a constant

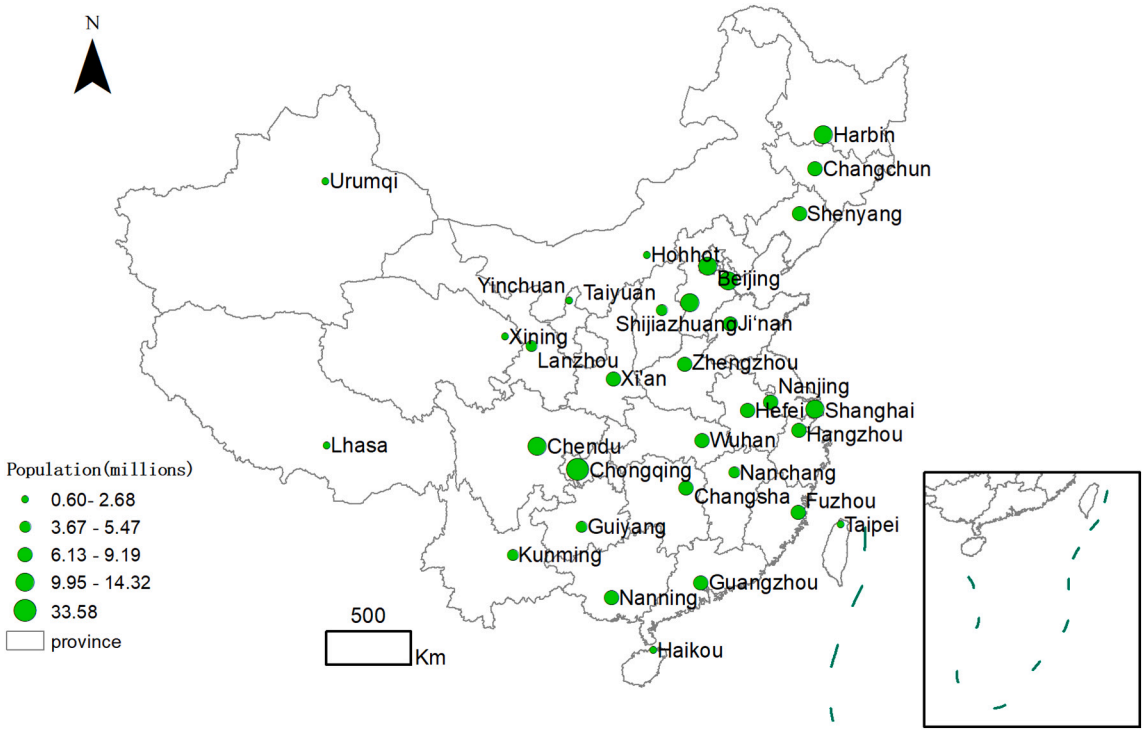

Fig. 1. Location of 32 capital cities in China.

and, as such, of no significant concern; Y_i and Y_j are the masses of the original and destination cities (e.g. the GDP) and D_{ij} is the distance between the two cities. Here, we use X_{ij} to denote unilateral geographical awareness from city i to city j, which indicates the frequency of users in city i who reference the name of city j. Researchers have already shown that the higher the level of individual geographical awareness of persons within a city, the more likely it is that the individuals have established connections with the city. In our case, Y_i and Y_j are the population of the original and destination cities. D_{ij} is the direct distance between city i and city j. Z_{ij} is the friction factor, also labeled as the control variable. Note that Z_{ij} can be more than one independent variable. After taking the logarithm for both sides of the equation, we have:

$$log(X_{ij}) = log(G) + \beta_1 * Y_i + \beta_2 * Y_j - \beta_3 * log(D_{ij}) - \beta_3 * Z_{ij} \quad (2)$$

We then rewrote the above equation in an econometric regression form and applied it to our case:

$$log(awareness_{ij}) = \alpha_0 + \alpha_1 * log(pop_i) + \alpha_2 * log(pop_j) + \alpha_3 * log(Distance_{ij}) + \alpha_4 * log(salary_j) + \alpha_5 * HistoricalCity_j + \alpha_6 * RailHub_j + \varepsilon_{ij} \quad (3)$$

where $awareness_{ij}$ indicates the number of social media posts in city i that name city j; pop_i is the population in city i; pop_j is population in city j; Variable $Distance_{ij}$ indicates the distance between city i and city j; $salary_j$ indicates average employee salary in city j; Variable $HistoricalCity_j$ is a dummy variable indicating if city j is a national historical city as of 2015; Variable $RailHub_j$ is a dummy variable indicating if city j is a railway hub; and ε_{ij} is the error term. α_0, α_1, α_2, α_3, α_4, α_5, and α_6 are corresponding coefficients. α_3 is the coefficient of variable $Distance_{ij}$ and is expected to be negative. For outside visitors, destination city j is attractive if city j offers more historic scenery, or more profitable income opportunities, or if it offers highly accessible transportation. Thus, we include variables $salary_j$, $HistoricalCity_j$, and $RailHub_j$ in the model.

Table 2 lists the variables that are included in the analysis. Each unilateral interaction from the original city to the destination city is treated as one observation. For example, an event that users in Beijing mention X number of times in reference to Shanghai is treated as an observation. A second event that users in Shanghai mention in reference to Beijing is treated as another observation. After data cleaning, we take into account 917 observations.[1] For each observation, we have attributes: $awareness_{ij}$ indicates the frequency at which destination city j is mentioned in microblogs of users in an original city I; Pop_i indicates the population of original city i; Pop_j indicates the population of destination city j; $Distance_{ij}$ indicates the direct distance between original city i and destination city j; $HistoricalCity_j$, is a dummy variable that indicates if the destination city j is a national historical city; $RailHub_j$, is a dummy variable that indicates if a destination city j is a national railway hub; $Salary_j$ indicates if a destination city j is a national railway hub.

Table 2
Definition of variables.

Variable name	Definition
$awareness_{ij}$	Frequency of mentioning a destination city j in microblogs of users in an original city i in 2015.
Pop_i	Population of original city i (2015)
Pop_j	Population of destination city j (2015)
$Distance_{ij}$	Direct distance between original city i and destination city j.
$HistoricalCity_j$	Dummy variable. If a destination city j is labeled as national historical city, assign 1. Otherwise, 0.
$RailHub_j$	Dummy variable. If a destination city j is a national railway hub, assign 1. Otherwise, 0.
$Salary_j$	Average employee salary in destination city j (2015)

The characteristics of the sample are shown in Table 3. The mean observation has a frequency of 112,689 count, 5.009 million population for original city, 5.024 million population for destination city, interaction distance of 1360 km, 74.4% probability of being an historical city, 22.8% probability of being a railway hub, and an average salary of 70,429 Yuan for the destination city.

We estimated Eq. (3) using the ordinary least square (OLS) and

[1] A couple of observations is missing counts of post numbers. We drop those observations when we run the regression.

Table 3
Descriptive statistics.

Variable name	Mean	Standard deviation	Minimum	Maximum
$awareness_{ij}$	112,689	1,737,958	5	40,184,256
Pop_i (10,000 persons)	500.9	441.2	22.3	2126.7
Pop_j (10,000 persons)	502.4	441.9	22.3	2126.7
$Distance_{ij}$ (km)	1360	719.4	96	3554
$HistoricalCity_j$	0.744	0.4368	0	1
$RailHub_j$	0.228	0.4197	0	1
$Salary_j$ (Yuan)	70,429	13,968	54,200	113,073

Note: the dataset has 917 observations.

reported the results in Table 4. All signs of coefficients aligned with our expectations. The coefficient of log (Distance) – also labeled as the distance decay parameter – is −0.3085 and significant at the 0.001 level. This parameter reveals that the distance between cities impacts on their interaction behaviors. We can compare different interaction behaviors using their distance decay parameter values. A greater distance decay parameter value implies a faster decay effect and spatial interactions are influenced by distance. The absolute value of the distance decay parameter in our case is smaller than that in Liu, Wang, et al. (2014) and Lu and Liu (2012). In their studies, they use air passenger flow data and social media check-in data. The coefficients of the *population* of the original city and the destination city are positive and significant, which means that spatial interaction increases as the population of cities increase. The coefficient of variable *Historical* is positive and significant, which means that the spatial interaction increases if the destination city is a national historical city. If the destination city is a historical city, it is apt to attract more activities, such as tourism and commercial investment. The coefficient of *railway hub* is positive and significant, which implies that railway hub cities generate more spatial interactions between itself and other cities. The coefficient of *salary* variable is positive and significant, which means that cities with higher average wages would trigger more spatial interactions from other cities. Our model has a respectful R^2 with the magnitude of 33.7%, given the limited explanatory variables we have. In summary, the regression results showed that social media-based geographical awareness indices follow the gravity law, which implies that to the indices may indicate interaction between cities; in addition, inter-city interactions are governed by the gravity model. We further verify that the indices can measure interactions among cities using mobility flow data in Section 5.4.

5.2. Spatial in-awareness index (SIAI)

The spatial in-awareness index (SIAI) is designed to measure the collective spatial inward-awareness rate for a city. Using the index, it is worthwhile to compare the differences in the levels of SIAI in different cities. The index is seen as an indicator of a city's ranking in its own city hierarchy. Here we suggest that the index can be calculated using the form.

$$SIAI_i = \frac{\sum_{j=1}^{n}\left\{X_{ij} * \left(\frac{D_{ij}}{max\left(D_{ij}\right)}\right)^{-\alpha}\right\}}{X_{ii}} \text{ for } i \neq j \quad (4)$$

where $SIAI_{ij}$ is spatial in-awareness index; X_{ij} indicates geographical awareness of city i with regard to city j; n is the number of cities in the analysis; X_{ii} indicates the total number of awareness for a city i received from itself; D_{ij} indicates the actual distance between city i and city j; max (D_{ij}) indicates the largest distance from city i to other cities. α is a parameter used to indicate distance decay. α is set to the value that we estimated in gravity equation model, which equals to 0.3085. $AIGA_i$ indicates spatial in-awareness index with regard to city i.

Using the above Eq. (4), we calculated SIAI for 32 capital cities in China (see Table 5). Table 5 and Fig. 2 show city ranks based on SIAI. These city rankings aligned with our expectations. Shanghai, Beijing, Tianjin, and Guangzhou – all of which were considered national central cities by the Chinese federal government in 2010 – represented the top four on the list. These cities were followed by Zhengzhou, Chengdu, Chongqing, and Xi'an, which are regional urban centers in Central China, Southwestern China, and Northwestern China, respectively. Capital cities in the developing areas – such as Guiyang, Haikou, Yinchuan, Hohhot, and Urumqi – rounded out the bottom of the ranking list.

5.3. Out-awareness rate (SOAR) and the spread of 2019-coronavirus in China

Out-awareness rate (OAR) is the outward-awareness from a city to another city, relative to the original city's local-awareness. Spatial out-

Table 4
Results of regression.

	Estimate	Std. error
Intercept	−7.9460***	1.9603
Log(*Pop*) (original city)	0.4482***	0.0336
Log(*Pop*) (destination city)	0.2254***	0.0375
Log(*Distance*)	−0.3085***	0.0479
HistoricalCity	0.2944***	0.0700
RailHub	0.1555ˆ	0.0812
Log(*Salary*) (destination city)	1.0810***	0.1760
Sample size	917	
R-squared	33.7%	
Adjusted R-squared	33.3%	

The dependent variable is log ($Frequency_{ij}$). The diagnostic of error term shows that residuals are normally distributed and equal variance (homoscedasticity).
*** $p = 0.001$ (two-tailed).
ˆ $p = 0.1$(two-tailed).

Table 5
SIAI of 32 capital cities in China.

City name	SIAI	
	Value	Rank
Shanghai	4,022,493	1
Beijing	2,072,568	2
Tianjin	1,686,948	3
Guangzhou	1,590,888	4
Zhengzhou	417,497.7	5
Chengdu	8265.303	6
Chongqing	8124.824	7
Xi'an	7406.047	8
Nanjing	7239.427	9
Wuhan	6420.44	10
Lanzhou	5979.48	11
Hangzhou	5877.02	12
Changsha	5652.886	13
Harbin	4353.62	14
Kunming	4320.249	15
Shenyang	4105.439	16
Taipei	4081.644	17
Taiyuan	3874.083	18
Jinan	3854.372	19
Nanchang	3712.317	20
Changchun	3541.766	21
Fuzhou	3513.15	22
Hefei	3462.144	23
Shijiazhuang	3450.072	24
Xining	3414.347	25
Lhasa	3133.61	26
Nanning	2968.343	27
Guiyang	2923.827	28
Haikou	2469.157	29
Yinchuan	2429.685	30
Hohhot	1390.172	31
Urumqi	502.9634	32

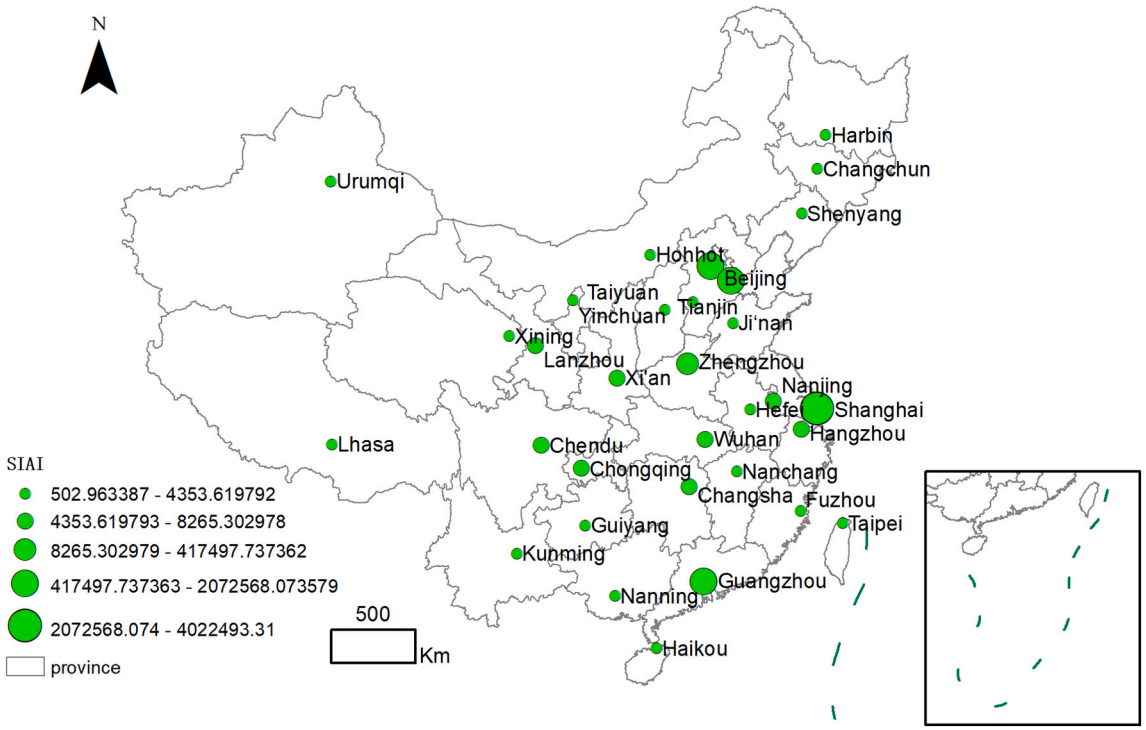

Fig. 2. SIAI level of 32 capital cities.

awareness rate (SOAR) is the representation of an outward-awareness rate to spatial distance between two cities. The mathematical expression for calculating $SOAR_{ij}$ is given by:

$$SOAR_{ij} = \left(\frac{X_{ij}}{X_{ii}}\right)\left(\frac{D_{ij}}{max(D_{ij})}\right)^{-\alpha} \text{ for } i \neq j \quad (5)$$

where $SOAR_{ij}$ indicates spatial out-awareness rate from city i to city j to the spatial distance of between two cities; X_{ij} indicates the awareness index, which is the number of posts by any posters located in city i, and whose text contains the name of city j. The larger X_{ij} denotes the stronger awareness from city i to city j. X_{ii} indicates city i's local awareness; D_{ij} indicates the distance between city i and city j; max (D_{ij}) indicates the largest distance from city i to other cities; α is a parameter used to indicate distance decay. We set α equaling to 0.3085, which we previously estimated.

The first case of the novel COVID-19 coronavirus was identified in early December 2019 in Wuhan, the capital of Hubei Province, China. The outbreak of the virus coincided with the most significant Chinese holiday – the Lunar New Year – and millions of residents moved from Wuhan to other cities via various transportation modes to observe the holiday. Some travelers were infected with the virus and thereby spread it to residents of other cities. The number of cases and deaths continues to rise as of this writing, making this a challenging health event both within China and across the world. In early response efforts, Chinese governments worked to identify and isolate persons infected with the virus throughout the country, and delivered medical equipment and staffs to cities where large populations were infected. Yet, the Chinese government had limited knowledge about which cities would suffer most and to which cities more medical equipment and staffs should be delivered rapidly.

The logic of the analysis in this section is as follows. COVID-19 is spread via human-to-human transmission, primarily through close contact and in-person interactions (Guan et al., 2020). Regular interactions between cities in which the virus first originated served as a critical factor in the rapid spread of the disease. The SOAR index can indicate interactions among cities; as such, we expect the SOAR to reflect a significant correlation with the number of COVID-19 cases in cities.

The calculation of SOAR values from Wuhan to other cities provides us with a critical opportunity to investigate the collective relatedness from Wuhan to other capital cities. Table 6 reports the SOAR value as well as the numbers of infected persons in other Chinese capital cities. Fig. 3 shows that the number of infected persons is positively correlated with the logarithm of SOAR at a significant level of 0.001with an R-squared value of 0.397. This means that the SOAR value can account for the logarithm of the number of infected persons in cities, with a 39.7% variation. Note that the number of infected persons is also attributed to other factors that we do not include in the regression, such as a city's infrastructure, health response, emergency response, etc. Given the fact that we do not include other explanatory variables in the regression, the R-squared value of 0.397 is a respectable value. COVID-19 is still spreading but epidemiologists face enormous challenges in trying to predict where the next cases of newly infection persons will occur. By estimating the SOAR values from Wuhan to other cities, we could ascertain the collective relatedness of a city with Wuhan and estimate the number of people who may become infected in a given city over a given period. Based on the SOAR values of cities, governments could more efficiently deliver the appropriate amounts of medical equipment and staff when public health emergencies occur.

5.4. Verification of geographical awareness index

To verify that the awareness indices that we developed above can indicate interactions among cities, we further investigated human interactions among cities using an additional dataset. This dataset illustrates human mobility flows among provincial capital cities in China. The data were retrieved from Baidu (Alias Chinese Google) (China Data Lab, 2020). It includes human mobility flows among cities on January 1st, 2020. In general, the human mobility flow from city i to city j is not same as the human mobility flow from city j to city i. Each unilateral mobility flow from the original city to the destination city is treated as one observation. By running a linear regression between the volumes of mobility flows and awareness (number of social media posts in an

Table 6
Ranking of spatial out-awareness rate and ranking of cities with confirmed cases of COVID-19.

City name	SOAR value	Number of infected persons
Chongqing	2.117644	525
Beijing	1.906248	366
Guangzhou	2.132203	327
Shanghai	1.991529	315
Changsha	2.79333	228
Nanchang	2.46833	210
Hangzhou	2.36014	162
Hefei	2.448262	161
Harbin	1.473183	159
Zhengzhou	2.103662	141
Chengdu	1.828899	131
Tianjin	2.13036	117
Xian	1.865542	114
Nanjing	2.728432	90
Fuzhou	2.081659	64
Kunming	1.594614	48
Jinan	2.210397	47
Nanning	1.920687	44
Changchun	1.24687	42
Lanzhou	1.82387	35
Guiyang	1.529272	33
Yinchuan	0.95829	32
Haikou	1.416319	29
Shenyang	1.332595	28
Shijiazhuang	1.561715	27
Urumqi	0.279489	21
Taipei	1.815492	18
Taiyuan	1.619474	18
Xining	1.240932	15
Hohhot	0.642387	7
Lhasa	1.373572	1

Note: the number of patients infected with COVID-19 was updated at 12:00 pm on February 13, 2020 (EST). Wuhan has 32,994 patients infected with COVID-19. We retrieved the number of infected persons from the Tencent (2020).

original city that name a destination city), we found that the volumes of mobility flow are highly correlated with awareness. Specifically, the awareness index can explain 17.5% of variation of the human mobility flows. The coefficient of the awareness index is 0.421 at a significant level of 0.0001, which means that a 1% increment of the awareness index indicates a 0.421% increase of human mobility flows. The verification test shows that the geographical awareness index can illustrate interactions among cities.

6. Conclusion and discussion

This study develops social media-based geographical awareness indices to quantify geographical awareness for 32 capital cities in China. We utilized an econometric model to demonstrate that the geographical awareness indices follows the gravity law and then estimated the decay function parameter as 0.3085. This confirms the validity of Tobler's first law of geography in cyberspace: "Everything is related to everything else, but near things are more related than distant things (Tobler, 1970)." The explanation is that individuals may have more spatial interactions – such as by visiting friends, traveling, doing business, and shopping – in nearby cities than in distant cities. Hence, microblog users more frequently mention nearby city names rather than distant city names and they have higher awareness of nearby cities than distant cities. Using mobility flow data, we verify that the social media based geographical awareness indices can measure interactions among cities.

In addition, the proposed spatial in-awareness index (SIAI) can indicate a city's influence and attractiveness in its own city hierarchy. Finally, the proposed spatial out-awareness rate (SOAR) can predict number of cases for other else cities in the urban system, once the origin city or hotspots of the outbreak and the number of infected persons within those cities are known.

The study results should be interpreted with some caution because of two limitations. First, the analysis does not include data from non-

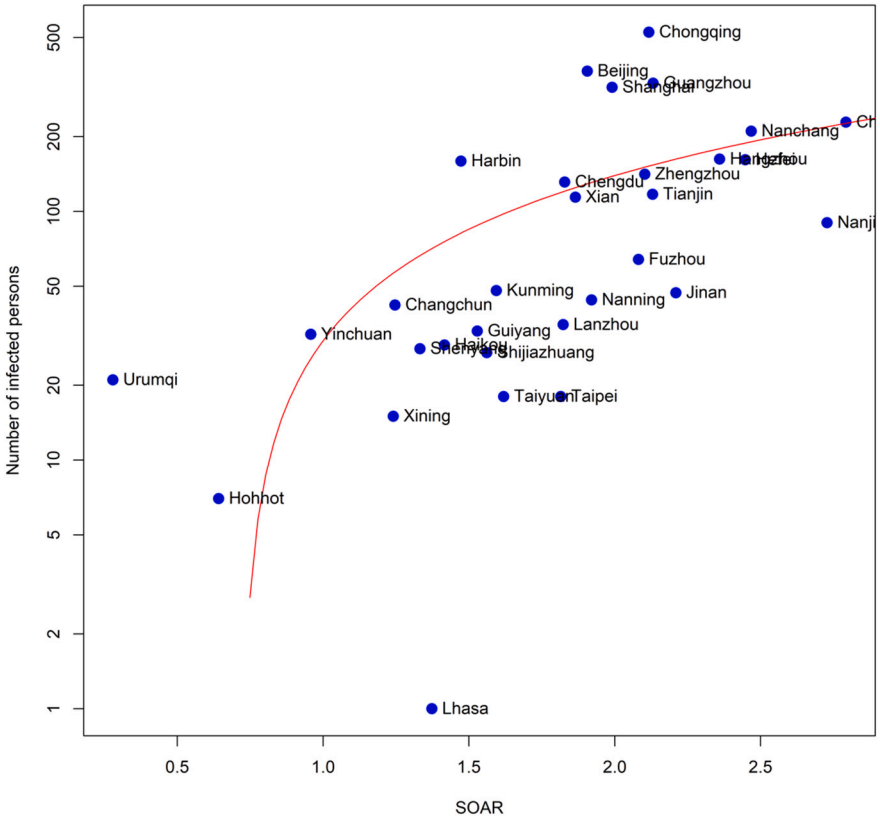

Fig. 3. Correlation between the logarithm of SOAR and the number of infected COVID-19 persons (R-squared 0.397).

capital megacities such as Qingdao, Shenzhen, or Suzhou. Those megacities have large populations and a significant impact on the regional economy. The reason for this omission is that we were unable to access the Sina Weibo data for these cities. Yet, provincial capital cities play a major role in the national urban system. Second, research using social media data has being criticized for being misrepresentative of residents. Young adults are more likely than their older counterparts to use social media platforms. To resolve this limitation, this study employs one more dataset – migration data among provincial capital cities – to validate the results. The high correlation between migration and place awareness demonstrates that the results are valid.

This study presents three important implications. First, the proposed social media-based geographical awareness indices – such as SIAI and SOAR – can be extended to other datasets and cases in other countries. Second, the SIAI can indicate the attractiveness of cities and demonstrate the influences that cities have in their own city hierarchy. As such, policymakers and urban planners could utilize the index to investigate the city hierarchy in terms of influence. Investors, such as tourism agencies, may adjust their investments based on the SIAI. The higher a city's level of SIAI, the more attractive the city is. Investors make rational decisions and tend to invest in cities with high inter-geographical awareness indices. Third, once the origin city or hotspots of the outbreak and the number of infected persons within these cities are known, we can use the social media-based SOAR index to predict number of cases for other else cities in the urban system.

The social media-based geographical awareness indices we developed can be easily applied in other countries, particularly those that have detailed social media data. For instance, social media data are quite accessible and detailed in the United States. Applying the indices we developed, researchers can investigate interactions among various cities. Even more, on a global scale, researchers can calculate the ranking of cities using social media-based geographical awareness indices. These indices could prove useful in predicting disease spread; with COVID-19, this data could have shaped our understanding of hotspot cities and how and why infection rates later spiked in other cities. Armed with this information, federal and local governments can properly and efficiently deliver medical equipment and staffs to cities where large populations are infected.

CRediT authorship contribution statement

Xinyue Ye: Conceptualization, Methodology, Writing- Reviewing and Editing.

Shengwen Li: Methodology, Data curation, Visualization, Investigation.

Qiong Peng: Methodology, Validation, Writing Original draft preparation, Editing.

Declaration of competing interest

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

Acknowledgements

This material is partially based upon work supported by the National Science Foundation under Grant No. 1416509. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Beatty, W. W., & Tröster, A. I. (1987). Gender differences in geographical knowledge. *Sex Roles, 16*(11–12), 565–590.

Chen, J., & She, J. (2012, June). An analysis of verifications in microblogging social networks—Sina Weibo. In *2012 32nd international conference on distributed computing systems workshops* (pp. 147–154). IEEE.

China Data Lab, 2020, "baidu_in_20200101.csv", Baidu Mobility Data, doi:10.7910/DVN/FAEZIO/6RWNXE, Harvard Dataverse, V16.

Chiodo, J. J. (1993). Mental maps: Preservice teachers' awareness of the world. *Journal of Geography, 92*(3), 110–117.

Flowerdew, R., & Lovett, A. (1988). Fitting constrained Poisson regression models to interurban migration flows. *Geographical Analysis, 20*, 297–307.

Freksa, C., Klippel, A., & Winter, S. (2007). A cognitive perspective on spatial context. In *Dagstuhl seminar proceedings*. Schloss Dagstuhl-Leibniz-Zentrum fr Informatik.

Golledge, R. G., & Stimson, R. J. (1997). Spatial cognition, cognitive mapping, and cognitive maps. In G. Golledge (Ed.), *Spatial behavior: A geographic perspective* (pp. 224–257).

Gong, J., Li, S., Ye, X., Andris, C., & Peng, Q. (2020). *Measuring the dynamic impact of high speed railways on urban interactions in China*. arXiv e-prints, arXiv-2010. https://www.researchgate.net/profile/Xinyue_Ye2/publication/344734261_Measuring_the_Dynamic_Impact_of_High-Speed_Railways_on_Urban_Interactions_in_China/links/5f983c0092851c14bceaffa4/Measuring-the-Dynamic-Impact-of-High-Speed-Railways-on-Urban-Interactions-in-China.pdf.

Guan, W. J., Ni, Z. Y., Hu, Y., Liang, W. H., Ou, C. Q., He, J. X., … Du, B. (2020). *Clinical characteristics of 2019 novel coronavirus infection in China* (MedRxiv).

Guldmann, J.-M. (1999). Competing destinations and intervening opportunities interaction models of inter-city telecommunication flows. *Papers in Regional Science, 78*, 179–194.

Guo, Z., Li, Z., & Tu, H. (2011, October). Sina microblog: An information-driven online social network. In *2011 international conference on cyberworlds* (pp. 160–167). IEEE.

Han, S. Y., Tsou, M. H., & Clarke, K. C. (2015). Do global cities enable global views? Using Twitter to quantify the level of geographical awareness of US cities. *PLoS One, 10*(7), Article e0132464.

Hesse, M. (2010). Cities, material flows and the geography of spatial interaction: Urban places in the system of chains. *Global Networks, 10*(1), 75–91.

Horton, F. E., & Reynolds, D. R. (1971). Effects of urban spatial structure on individual behavior. *Economic Geography, 47*(1), 36–48.

Hu, Y., Ye, X., & Shaw, S. L. (2017). Extracting and analyzing semantic relatedness between cities using news articles. *International Journal of Geographical Information Science, 31*(12), 2427–2451.

Lin, J., & Li, X. (2015). Simulating urban growth in a metropolitan area based on weighted urban flows by using web search engine. *International Journal of Geographical Information Science, 29*(10), 1721–1736.

Liu, Y., Sui, Z., Kang, C., & Gao, Y. (2014). Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PLoS One, 9*(1), Article e86026.

Liu, Y., Wang, F., Kang, C., Gao, Y., & Lu, Y. (2014). Analyzing relatedness by toponym co-occurrences on web pages. *Transactions in GIS, 18*(1), 89–107.

Lu, Y., & Liu, Y. (2012). Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies. *Computers, Environment and Urban Systems, 36*(2), 105–108.

Meijers, E., & Peris, A. (2019). Using toponym co-occurrences to measure relationships between places: Review, application and evaluation. *International Journal of Urban Sciences, 23*(2), 246–268.

Richards, P. (1974). Kant's geography and mental maps. *Transactions of the Institute of British Geographers*, 1–16.

Tencent. (2020). Tracking coronavirus. Retrieved from https://news.qq.com/zt2020/page/feiyan.htm.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. Economic Geography, 46(sup1), 234–240.

Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., & Bao, Z. (2013, April). A depression detection model based on sentiment analysis in micro-blog social network. In Pacific-Asia conference on knowledge discovery and data mining (pp. 201-213). Springer, Berlin, Heidelberg.

Wang, Z., Lam, N. S., Obradovich, N., & Ye, X. (2019). Are vulnerable communities digitally left behind in social responses to natural disasters? An evidence from Hurricane Sandy with Twitter data. *Applied Geography, 108*, 1–8.

Wang, Z., & Ye, X. (2019). Space, time, and situational awareness in natural hazards: A case study of Hurricane Sandy with social media data. *Cartography and Geographic Information Science, 46*(4), 334–346.

Xiao, Y., Wang, F., Liu, Y., & Wang, J. (2013). Reconstructing gravitational attractions of major cities in China from air passenger flow data 2001–2008: A particle swarm optimization approach. *The Professional Geographer, 65*, 265–282.

Xu, C., Wong, D. W., & Yang, C. (2013). Evaluating the "geographical awareness" of individuals: An exploratory analysis of Twitter data. *Cartography and Geographic Information Science, 40*(2), 103–115.

Yan, Q., Wu, L., & Zheng, L. (2013). Social network based microblog user behavior analysis. *Physica A: Statistical Mechanics and its Applications, 392*(7), 1712–1723.

Yue, Y., Dong, K., Zhao, X., & Ye, X. (2019). Assessing wild fire risk in the United States using social media data. Journal of Risk Research, 1-15.