Deriving intercity human flow pattern and mechanism based on cell phone location data: case study of Guangdong Province, China

Li Zhuo^{1,2,3}, Zhuo Chen^{1,3}, Chengzhuo Wu^{1,3}, Qingli Shi^{1,3}, Zhihui Gu^{4*}, Haiyan Tao^{1,3} and Qiuping Li^{1,3}

Abstract

The spatial pattern and mechanism of human flow are of great significance for urban planning, economic development, transportation planning and so on. In this study, we used cell phone location data to represent the human flow network in Guangdong Province, China, using the 21 cities in Guangdong as “nodes” and the human flow intensity among them as “edges”. Then we explored macro and micro features of the human flow network, by using the index of degree distribution, alter-based centrality and alter-based power, respectively. Finally, we proposed a human flow estimation model which integrates individual urban characteristics, intercity links, and differences to further analyze the affecting factors of human flow. We found that the human flow network in this region is significantly scale-free, with Guangzhou, Shenzhen, Foshan, and Dongguan being the most important cities. We also found that the newly proposed model can explain the human flow in the study area, with an R^2 of 0.914. Analysis results show that the factors of employment in tertiary sector, intercity internet attention, intercity differences in the number of tertiary workers, differences in population size, and distance have significant impacts on the human flow. This study may provide insights into human activity mechanisms that can contribute to urban planning and management.

Keywords: Cell phone location data, Network analysis, Human flow, Human mobility

1 Introduction

Spatio-temporal patterns of human flow can reflect the spatial structure, functional structure, economic vitality and development prospects of an area to a certain extent (Holland, 1996). Meanwhile, these patterns may also affect the spread of infectious diseases, transportation, pollution exposure (Dong et al., 2021; Tokey, 2021; Yu et al., 2020; Zhou et al., 2020). Exploring human flow patterns and their causes has, therefore, been an important research topic in human geography studies (Kou et al., 2021; Wang et al., 2021; Yang et al., 2019), and

has received more and more attentions in recent years with the development and applications of various spatio-temporal big data.

Human flow estimation is the basis of human flow pattern analysis. In the early stage, due to the lack of data which can intuitively represent the flow of people, scholars mainly used theoretical models to estimate the human flow. The “push-pull” theory, proposed by Ravenstein (1885) and refined by Heberle (Heberle, 1938), holds that human migration is affected by the repulsive force of the origin and the attractive force of the destination. Based on the “push-pull” theory and the law of universal gravitation, Taylor P. J. proposed the gravity model (Taylor et al., 2001). In this model, the human flow intensity between two places can be determined by

* Correspondence: gzh@szu.edu.cn

⁴School of Architecture and Urban Planning, Shenzhen University, Shenzhen 510861, China

Full list of author information is available at the end of the article

their residential population and physical distance. The gravity model explains human mobility intuitively with easily accessible data. It has, therefore, been widely adopted and improved by scholars in the quantitative researches of human flow (Belyi et al., 2017; Wang, Dong, et al., 2019). However, due to lack of data which can intuitively represent the flow of data, it is usually difficult to calibrate the unknown parameters in traditional models, as well as to verify their accuracy. Meanwhile, population size might not be able to accurately explain human flow (Liu et al., 2014).

With the increased accessibility of various spatio-temporal location big data, such as traffic data and user location data, scholars have started using various spatio-temporal big data to estimate the intercity human flow strength (Ahas et al., 2015; Jin et al., 2016; Lai & Pan, 2019). Among them, cell phone location data show greater potential of application. As a common device in modern society, cell phone has diverse user groups and large penetration rate. The location data of cell phone users are capable of representing close-to-real population distribution in some cases (Alexander et al., 2015; Shi et al., 2015; Zhang & Thill, 2019). More and more studies have, therefore, begun to use cell phone location data to estimate human flow and identify its patterns (Cao et al., 2021; Tokey, 2021; Zhang & Ng, 2021).

Network analysis is a commonly used method for pattern analysis, which regards human flows as “edges” and cities as “nodes”. There are mainly two types of patterns for human flow networks, namely the micro and macro patterns (Zhang & Thill, 2019). The macro patterns refer to the overall network structure characteristic of a region, which can be represented by the ranking and distribution of nodes, while the micro patterns refer to the importance of cities represented by the centrality of the nodes. Centrality measures are important indicators which can reflect the significance of a city in a given network. A variety of measures have been proposed. The intermediate centrality considers the number of times nodes appear on the shortest path in the network (Freeman, 1978). The degree centrality and weighted degree centrality consider the number and weight of edges associated with nodes (Brands, 2001). The proximity centrality considers the distances between one node and the other nodes (Neal, 2011). The alter-based centrality and power considers both the direct and indirect interaction and distinguish the centrality and power of nodes (Neal, 2012; Ziyu et al., 2017). Zhao et al. further improved the indicators of alter-based centrality and alter-based power to make them more suitable for China (Zhao et al., 2017).

Existing studies have not only described characteristics of human flow, but also analyzed factors that affect it. The most considered factors include individual characteristics of the origin and destination cities, such as

population size, GDP, and the geographic distance between the two cities (Wang, Dong, et al., 2019). Many scholars believe that cities with larger populations, larger GDP, and closer distance between them are more likely to generate massive human flow (Cao et al., 2018). Some scholars found that similarities in the intrinsic characteristics of the origin and destination cities, such as geomorphic condition, cultural background, administrative hierarchy, may also affect the human flow strength (CNNIC, 2015; Guo et al., 2018). There might be other factors influencing the human flow, for example, information flow, traffic flow, differences in the population and GDP, etc. Few studies, however, have explored influence of these factors on human flow.

In summary, current human flow pattern analysis requires accurate data to construct the network and select the appropriate indicators for the study area. When analyzing the factors of human flow, only the inherent characteristics and geographical distance of individual cities are considered, while the differences formed in the dynamic development of cities and the various links between cities are usually ignored. In this study, we selected Guangdong Province, China as our study area and used cell phone location data and complex network analysis to analyze the macro and micro structure of human flow. After that, we analyzed the main factors influencing the human flow strength in this area. This paper can provide basic information and decision-making support for this region.

2 Study area and dataset

2.1 Study area

In this study, we select Guangdong province as our study area, which locates in Southern China (Fig. 1). The total population of the study area is 115.21 million, with an area of 179,725 km^2 . There are altogether 21 cities in the study area, with most populated cities, Shenzhen and Guangzhou, respectively. Guangdong Province governs 21 municipal cities, which can be further divided into four regions, namely the Pearl River Delta region, Western Guangdong, Eastern Guangdong, and Northern Guangdong. Adjacent to Hong Kong SAR (SAR, Special Administrative Regions) and Macao SAR, the PRD (PRD, Pearl River Delta) region is the most developed region in the province, with per capital GDP 130,182 Yuan. There are two first-tier cities, namely Guangzhou and Shenzhen in this region. Western Guangdong is the second economically developed region in the province, with per capital GDP 43,922 Yuan. The Northern and Eastern Guangdong is less developed, with per capital GDP 33,039 Yuan to 35,844 Yuan. Fig. 1 shows location of the study area, locations of the 21 cities in the study area.

Fig. 1 Location of Guangdong province

2.2 Dataset

Five types of data are included in this study, namely cell phone location data, Baidu index, intercity vehicle volume data, intercity road distance, and statistical data from yearbook. Specifically, Cell phone location data is the basis for human flow pattern analysis. The five datasets are used as variables in human flow estimation model. The specific descriptions of the datasets are as follows.

- (1) Cell phone location data. Nowadays, cell phone has become the most commonly used mobile device in

China. The penetration rate of cell phone had reached 112.2 cell phones per 100 people by the end of 2018 (MIIT, 2019). As for the study area, we calculated the cell phone penetration rates for all the 21 cities in Guangdong Province, using data from the 2019 Guangdong Provincial Statistical Yearbook. The penetration rate was calculated by dividing the number of cell phone subscribers in a city with its total population. The results show that the penetration rate of cell phones ranges from 0.8 to 2.4 in Guangdong, with an average value of 1.5 (Fig. 2). It should be noted that the subscribers

Fig. 2 Penetration rate of cities in Guangdong Province

include non-cellular devices which use mobile SIM cards. Although we do not have the specific number, most of the subscribers are personal cell phone users. The cell phone location data can, therefore, represent the human flow in each city well.

Each record in this data represents the number of cell phone users who appeared in both city i and city j during 30 days before day t . It can represent the human flow between cities over the 30-day period. The dataset was collected from June 13, 2018, to September 10, 2018, with a total of 61 days.

- (2) Baidu Index. The Baidu index characterizes the internet attention between cities. The data was obtained from Baidu Index platform, which is a data analysis platform based on Baidu user search behavior data and currently provides large number of keywords search indices from 2011 onwards. By setting the keyword as the target city, one can find users' attention in each city to the target city. For example, we can get the internet attention of Shenzhen to Guangzhou by setting the keyword as "Guangzhou" and the query area as "Shenzhen". The observation period was set the same as the mobile location data.
- (3) Traffic carrying capacity data. Traffic carrying capacity refers to the number of passengers that can be carried between cities by trains and buses. In this study, we used the Selenium Python package to crawl the train and bus schedules in one single day from the booking page on Ctrip.com, and then estimated traffic carrying capacity between cities, assuming that there are 50 passengers per bus, 600 passengers per high-speed train, and 1200 passengers per normal-speed train (Wang, Dong, et al., 2019).
- (4) Inter-city road distance data. Inter-city road distances are obtained through Baidu Maps. We use Baidu Map's route search function to set the two cities as the starting and ending points, and query to get the distance.
- (5) Statistical yearbook data. The statistical yearbook contains socio-economic data. The data obtained from the statistical yearbook include GDP, average salary, residential population, total employed population, employed populations in primary, secondary and tertiary sectors. The data in this paper were obtained from the 2018 Guangdong Statistical Yearbook.

3 Method

Our study includes three main steps. First, we built a human flow network based on cell phone location data;

second, we selected reasonable indicators of centrality and power for the human flow network and then analyzed its macro and micro characteristics; finally, we constructed an improved human flow estimation model and used it to analyze the factors affecting human flow.

3.1 Constructing human flow network

The key of constructing human flow network is to measure the weight of edges reasonably. In this study, we used cell phone location data as the weight of edges and constructed the intercity human flow network of the study area.

The 21 cities in the study area were defined as the nodes in the network. An edge is set for every two city nodes i and j ($1 \leq i, j \leq 21$), while the weight of edge W_{ij} is defined as the human flow volume between city i and j during the observation period, represented as, $t = 1, 2, \dots, 61$. The urban human flow network in the study area can then be constructed based on the nodes and edges defined above.

3.2 Analyzing the macro characteristics of human flow

Based on the above-mentioned human flow network, we analyzed the macro characteristics of human flow from three aspect. The spatial distribution of human flow intensity between cities was firstly analyzed. Then, the cities were ranked according to the weighted degree centrality (Introduction) as Eq. (1), which can represent the importance of nodes well. Finally, linear correlation was performed for the logarithmic values of the rank order and the weighted degree centrality. If there is a significant linear correlation, the weighted degree centrality of nodes in the network is power-law distributed and the network is a scale-free network (Barabási & Albert, 1999).

$$C_{D_i} = \sum_{j=1}^n W_{ij} \quad (1)$$

where n represents the number of edges.

3.3 Analyzing the micro characteristics of human flow

In this part, we further analyzed the micro characteristics of the human flow network, including the status and role of cities in the network. These characteristics of nodes are usually described by centrality and power in network analysis. Among different indicators of centrality and power, the improved version proposed by Zhao et al. considers both direct and indirect connections between nodes (Zhao et al., 2017) and distinguish the centrality and power, reducing the overestimation of satellite cities caused by the impact of their connected core cities. They proposed the dependency parameter d_{ij} ,

which uses the ratio of the edge weight W_{ij} of city i and j to the degree centrality C_{D_j} of city j to characterize the dependence of the city j on the city i (Zhao et al., 2015). Their method is more suitable for portraying the importance and position of different cities in a region. In this study, we adopted this idea and constructed the alter-based centrality and power. The indexes can be expressed as Eqs. (2), (3) and (4).

$$AC_i = \sum_{j=1}^{n-1} (1-d_{ij}) W_{ij} C_{D_j} \quad (2)$$

$$AP_i = \sum_{j=1}^{n-1} (1-d_{ij}) \frac{W_{ij}}{C_{D_j}} \quad (3)$$

$$d_{ij} = \frac{W_{ij}}{C_{D_j}} \quad (4)$$

where AC_i is the alter-based centrality of city i , AP_i is the alter-based power of city i , W_{ij} is the weight of the edge of city i and city j , i.e., the actual human flow volume evaluated by cell phone location data between the

two cities, C_{D_j} is the weighted degree centrality of city j , i.e., the number of cities that city j connects with other cities, while d_{ij} is the dependency parameter, which portrays the dependency between cities.

Higher alter-based centrality of a city indicates stronger ability to collect and disperse resources and more likely to become a hub in the network. Whereas higher alter-based power means stronger ability to control resources and more potential to attract resources.

3.4 Building an improved human flow estimation model and analyze the key influencing factors of human flow

Human flow estimation model can explain human flow and estimate real human flow when lacking real human flow data. Therefore, based on the classical gravity model, we proposed an improved human flow estimation model, as well as analyze the factors affecting human flow.

Model construction requires appropriate dependent and independent variables. The independent variables include individual characteristics, links, and differences.

Table 1 Introduction to the data for modeling

Variables	Meaning	Data source
W_{ij}	Human mobility intensity between city i and j	Cell phone location data
P_{ij}^t	The number of mobile phone users who appeared in cities i and j during 30 days before the statistical day t	Cell phone location data
$bindex_{ij}$	Internet attention between city i and j	Baidu Index (e.g. the internet attention between Guangzhou and Shenzhen is the sum of daily Baidu Index with Guangzhou as the starting point and Shenzhen as the keyword plus the sum of daily Baidu Index with Shenzhen as the starting point and Guangzhou as the keyword)
$workpop_3i$	The employment population of the tertiary sector in city i	Statistical yearbook of Guangdong Province in 2018
$workpop_1i$	The employment population of the primary sector in city i	Statistical yearbook of Guangdong Province in 2018
$workpop_i$	The employment population in city i	Statistical yearbook of Guangdong Province in 2018
pop_i	The residential population in city i	Statistical yearbook of Guangdong Province in 2018
GDP_i	The GDP of city i	Statistical yearbook of Guangdong Province in 2018
$Salary_i$	The average wage of city i	Statistical yearbook of Guangdong Province in 2018
$dist_{ij}$	The network distant between city i and j	Baidu Map
$d_{Salary_{ij}}$	The difference of the average wage between city i and j	Statistical yearbook of Guangdong Province in 2018
$d_{workpop1_{ij}}$	The difference of the employment population of the primary sector between city i and j	Statistical yearbook of Guangdong Province in 2018
$d_{workpop2_{ij}}$	The difference of the employment population of the secondary sector between city i and j	Statistical yearbook of Guangdong Province in 2018
$d_{workpop3_{ij}}$	The difference of the employment population of the tertiary sector between city i and j	Statistical yearbook of Guangdong Province in 2018
$d_{pop_{ij}}$	The difference of the resident population between city i and j	Statistical yearbook of Guangdong Province in 2018
$d_{workpop_{ij}}$	The difference of the employment population between city i and j	Statistical yearbook of Guangdong Province in 2018
$d_{gdp_{ij}}$	The difference of GDP between city i and j	Statistical yearbook of Guangdong Province in 2018
ε	The error term	

Meanwhile, the dependent variable is the human flow obtained from cell phone location data. The reasons why we select these variables are in detail below. The variables are listed in Table 1.

- (1) Individual characteristics of cities. The GDP and population size of a city are often mentioned in the analysis of human flow factors (Shen & Liu, 2016; Windzio, 2018; Zhang et al., 2020). We believe that the number of workers in different sectors is also a key factor affecting human flow. We, therefore, chose GDP (*GDP*) and average wage (*Salary*) as the measurement of economic development level, resident population (*pop*), and total employed population (*workpop*) as the indicators of population size, and employed population in primary, secondary and tertiary sectors (*workpop1*, *workpop2*, *workpop3*) as the measurement of each sector's scale. This data was obtained from the statistical yearbook of Guangdong Province in 2018.
- (2) Links between cities. We considered two primary links between cities, i.e., the transportation accessibility and internet attention between cities.
 - (a) Inter-city transportation accessibility index. The higher the traffic accessibility, the more convenient it is to travel across cities, and the more likely to generate close urban human flow. In this paper, we adopted geographic distance $dist_{ij}$ and traffic carrying capacity $traffic_{ij}$ to characterize inter-city traffic accessibility, where traffic carrying capacity refers to the number of passengers that can be carried between cities by high-speed/normal-speed trains and long-distance buses.
 - (b) Internet attention between cities. In the information age, using the internet to search for information has become a major activity of internet users (Wang & Loo, 2019). Internet attention has, therefore, become an important inter-city link. The more web searches a city has, the more likely it is to become a potential destination. We characterized $bdindex_{ij}$, the internet attention of city i to city j , by counting the sum of Baidu index of search cities to the target city during the observed time.
- (3) Differences between cities. In physics, gradient refers to the rate of change with respect to distance of a variable quantity. It can also describe the uneven distribution of geographical elements. The generalized gradient theory suggests that gradients are embodied in natural, economic, social, human resource, ecological, and institutional differences. A gradient in any sense is both a gradient nudging party and a gradient receiving party. It exists

throughout human economic activities (Li, 2004). At present, little attention has been paid to the impact of gradient effect on human mobility. This paper aims to analyze the influence of the difference of GDP (d_{gdp}), the difference of average wage (d_{salary}), the difference of permanent resident population (d_{pop}), the difference of total employment population ($d_{workpop}$), the difference of employment population in various industries ($d_{workpop1}$, $d_{workpop2}$, $d_{workpop3}$) on the human flow between cities.

Adding more explanatory variables can make the description of the human flow more complete. To find its unknown parameters, the logarithm is often taken for both sides of the model. Then we performed stepwise linear regression (Hair et al., 1998) for the data from June 13 to August 11, 2018, and selected the model with maximum R^2 value as the optimal model, as shown in Eq. (5).

Since population contributes a lot to human flow estimation and may be closely related to the employed population and internet attention, we added population as a confounding variable to control the population effect (Eq. (5)). The parameters description is in Table 2.

$$\ln W_{ij} = \sum_{i=0}^8 \beta_i \ln(X_i) + \alpha_1 \ln(pop_i) + \alpha_2 \ln(pop_j) + \varepsilon \quad (5)$$

In order to validate the above-mentioned models, we used the data from August 12, 2018 to September 10, 2018 as test data, and used the index of R^2 and $RMSE$ to evaluate the prediction accuracy. The formula of $RMSE$ is shown in Eq. (6), where W_{ij}^{ref} is the real value, W_{ij}^{pred} is the predicted value, and n is the total number of edges.

Table 2 Explanation of variables in Eq. (5)

Factors	Variables
$bdindex_{ij}$	X_0
$workpop3_i$	X_1
$workpop3_j$	X_2
$workpop_i$	X_3
$workpop_j$	X_4
$dist_{ij}$	X_5
$d_{workpop3_{ij}}$	X_6
$d_{pop_{ij}}$	X_7
$d_{gdp_{ij}}$	X_8

$$RMSE = \sqrt{\frac{\sum (W_{ij}^{ref} - W_{ij}^{pre})^2}{n}} \quad (6)$$

Based on the above constructed model and the coefficients determined by regression analysis, we further analyzed the correlations between the variables and human flow, as well as their influences.

4 Results

4.1 Macro characteristics of human flow network

According to the human flow network construction method described in 3.1, the human flow network map of the study area can be obtained, see Fig. 2. The local human flow patterns in the study area can be divided into four levels according to the linkage strength from high to low (In Fig. 2, the strength is listed from smallest to largest in order from top to bottom). The first level is the strong human flows between Guangzhou, Dongguan and Shenzhen, which are located in the center of the PRD. The second level is mainly the sub-strength links among Guangzhou, Dongguan and Foshan. The third level is the moderate links, which includes two types, namely, internal PRD links and PRD-peripheral cities links. The fourth level is the weak connection among the peripheral cities.

The weighted degree centrality is calculated according to Eq. (1). The results shown in Fig. 3 and Fig. 4(a) can represent the human flow passing through the node.

The human flow of each city node has great spatial heterogeneity (Fig. 4(a)). Among them, Shenzhen has much more human flow than other cities, Guangzhou and Dongguan are next, which are basically at the same level, and Foshan is slightly less than the above three cities. They have much more human flow than other cities and occupy 78.4% of the human flow in the network. It can indicate that a few cities dominate the human flow in the network.

We further analyzed distribution of the weighted degree centrality of nodes in the human flow network. By performing linear regression analysis on the logarithm of weighted degree centrality and its rank, as shown in Fig. 4(b), the two shows significant linear correlation with R^2 reaching 0.8376. This indicates that the weighted degree centrality of the cities conforms to a power law distribution, and the regional urban human flow network is a scale-free network with great heterogeneity (Analyzing the macro characteristics of human flow).

4.2 Micro characteristics of human flow network

Alter-based centrality measures the ability of cities to collect and disperse human flow in the network. The alter-based centrality of each city is calculated according to Eq. (2). The results are shown in Fig. 5(a) and Table 3, where the natural breakpoint method was used to show its gradation. The first grade includes Dongguan (1.722), Guangzhou (1.891), Shenzhen (2.521), with alter-based centrality higher than 0.895. Located in the central area of the PRD, they serve as the political, economic and cultural

Fig. 3 Human flow network in Guangdong Province

Fig. 4 Correlation between rank and weighted degree: (a) Scatter Plot of rank and weighted degree; (b) Scatter Plot of rank and $\ln(\text{weighted degree})$

centers and important connecting hubs in the region. The second level includes Foshan, with alter-based centrality of 0.895. The third class includes the rest cities with alter-based centrality below 0.25, which differs greatly from the rest of the classes. This indicates that the distribution capacity of the PRD region (especially Guangzhou, Dongguan, Shenzhen and Foshan) has a dominant position.

The alter-based power reflects the node's ability to dominate the human flow in the network. The alter-based power of each city node is calculated according to Eq. (3), and the results are shown in Fig. 5(b) and Table 3. Alter-based power also shows a hierarchical structure. The first level includes Guangzhou, Dongguan and Shenzhen, whose alter-based power reaches more than 1.021. It is worth mentioning that the alter-based power of Shenzhen reaches 3.957, which is 1.47 times higher than that of Guangzhou and locates at an absolute leading position in the city network. This may be explained by the fact that Shenzhen is the most economically developed and high-tech city in the region, and also an important window connecting the region with home and abroad. The second level includes Foshan, with alter-based power of 1.021. The third level includes other cities, with alter-based power of 0.521 or less.

It is worth mentioning that Zhanjiang City, which is located in the westernmost Guangdong with the fewest bordering cities, is ranked fifth in terms of both alter-based centrality (0.25) and alter-based power (0.521). It shows strong centrality and power comparing to the other cities in western Guangdong, like Maoming, Yangjiang, and Yunfu. This may be due to that Zhanjiang is closely connected to the core of the PRD (Fig. 3) where the human mobility is active, thus having a strong centrality relative to other peripheral

cities. Zhanjiang city itself not only has a certain ability to collect and disperse resources, but also has the potential to drive other cities in Western Guangdong to accelerate their development.

4.3 Evaluation of the improved human flow estimation model

Following the method described in 3.2, we constructed the improved logarithm model of human flow. The model passed the F-test with the goodness of fit reached 0.914 and the p -value of each variable less than 0.01, indicating that the improved model was more suitable for describing the human flow in the study area. By analyzing the regression parameters shown in Table 4, we can see that the independent variables all pass the significance test and there is no collinearity. This indicates that the added variables can largely explain the amount of intercity human flow and the selection of independent variables is reasonable.

Furthermore, the proposed model was validated using the test data set. The result shows that there is a significant linear correlation between the prediction results and the reference data (Fig. 6), with an R^2 value of 0.9066. The high prediction accuracy further proves that the constructed model is reasonable.

4.4 Analysis on key influencing factors of human flow

Using the proposed model, we further analyzed the influences of different factors on human flow.

The coefficient and standardized coefficient for all independent variables are in Table 4. The standardized coefficient can eliminate the effect of different units of the independent variables to characterize the contribution of

Fig. 5 Altered Centrality and power of Cities: (a) alter-based centrality of cities; (b) alter-based power of cities

Table 3 Alter-based Centrality and power of cities

city	AC	AP	city	AC	AP
Shenzhen	2.521	3.957	Zhaoqing	0.075	0.331
Guangzhou	1.891	2.677	Meizhou	0.061	0.277
Dongguan	1.722	2.314	Shaoguan	0.056	0.267
Foshan	0.895	1.021	Shanwei	0.055	0.274
Zhanjiang	0.25	0.521	Chaozhou	0.054	0.299
Huizhou	0.202	0.453	Heyuan	0.046	0.282
Zhongshan	0.195	0.383	Yangjiang	0.039	0.281
Jiangmen	0.169	0.433	Yunfu	0.075	0.331
Shantou	0.133	0.327			
Maoming	0.118	0.342			
Jieyang	0.108	0.358			

each independent variable. The larger the standardized coefficient, the more constructive it is for the dependent variable.

Among the indicators of individual characteristics, two of them passed the significance test, including the total employed population and the employed population in tertiary sector of the two cities. Specifically, the employed population in the tertiary sector of the two cities is the most constructive in this model. The larger the employed population in the tertiary sector, the stronger the human flow. It can therefore be concluded that peripheral cities should develop the tertiary industry and create more tertiary employment opportunities, strengthening the human flow with other cities.

Among the indicators of links between cities, distance passed the significance test. The intercity human flow is negatively associated with the intercity distance. The

shorter the distance, the stronger the human flow. That indicates that physical distance still has great impact on human flow. Peripheral cities in this region should, therefore, improve the transportation system to form a close connection with the core cities, hence reducing the development hindrance caused by the distance effect.

In addition to distance, the indicator of internet attention between two cities also passed the significance test. Internet attention shows a positive correlation with human flow strength. The higher the intercity internet attention, the stronger the human flow. This indicates the dissemination of network information may affect users' intercity interaction behavior in the information era. Users' internet attention to a city may reflect their willingness to travel to that city to a certain extent. Meanwhile, intercity internet attention, as the only data that changes over time in the improved model, plays an important role in estimating the amount of monthly-scale human flow. Peripheral cities could provide diversified information in cyberspace to improve their visibility online, which may draw users' attention to them inspire travel intentions and attract human flow in turn.

As for the index of difference between cities, three influencing factors, namely the differences of residential population, tertiary industry employment, and GDP between two cities, are found to be significant. Specifically, the difference of residential population shows a negative correlation with the human flow, which means that cities with similar population size are strongly connected, and weakly connected otherwise; The difference of employed people in the tertiary sector shows a positive correlation with the human flow. This means cities with greater differences in the number of people employed in the tertiary sector are relatively more connected, while cities with similar tertiary sector scales have weak human flows among them. Finally, the difference between the two cities' GDP shows a positive correlation with the human flow. However, the difference of GDP has little impact on the human flow between two cities. The government should encourage companies to settle in the peripheral cities, providing more work opportunities in the tertiary sector. Meanwhile, initiating attractive talent introduction programs and household registration policies may also help to improve the human flow.

5 Discussion

Currently, spatial interaction models are widely used in the study of human mobility. Among them, the classical gravity model is still used today and explains the human flow well in various studies. Although the gravity model has been confirmed in many scenarios, relatively few factors are considered, and whether it is applicable to portraying the human flow characteristics of the current study area remains to be explored.

Table 4 Results of Regression

variables	coefficient	Standardized coefficient
$\ln(\text{workpop}_i)$	-0.196***	-0.106***
$\ln(\text{workpop}_j)$	-0.173***	-0.092***
$\ln(\text{workpop3}_i)$	2.154***	0.605***
$\ln(\text{workpop3}_j)$	2.249***	0.621***
$\ln(\text{dist}_{ij})$	-0.486***	-0.138***
$\ln(\text{bdindex}_{ij})$	0.877***	0.217***
$\ln(d_{\text{pop}_{ij}})$	0.091***	-0.073***
$\ln(d_{\text{workpop3}_{ij}})$	-0.170***	0.101***
$\ln(d_{\text{gdp}_{ij}})$	0.216***	0.036***
$\ln(\text{pop}_i)$	-0.614***	-0.148***
$\ln(\text{pop}_j)$	-0.526***	-0.129***
ε	-14.468***	
R^2	0.914	
F	331.104***	

note: *** represents passing the significance test at the 0.01 level

Fig. 6 Scatter Plot of $\ln(\text{reference})$ and prediction

There are unknown damping parameters in the gravitational model, which need to be obtained by tuning the parameters. In this paper, the tuning parameter range is set to $[0, 0.01, 2]$, and the final parameter is determined to be 0.472. In addition, this paper compares the human flow calculated based on the gravity model with the human flow volume obtained from cell phone location data, calculates R^2 , and obtains the $R^2 = 0.77$, which means there are significant differences between the human flow estimated by the gravity model and the observed human flow. Compared with the classical gravity model, the improved model works better ($R^2 = 0.914$). So, using only three factors is not sufficient to explain the crowd interaction.

In the analysis of the factors influencing the human flow, the development of the tertiary industry plays an important role in it ([Analysis on key influencing factors of human flow](#)). The following is a speculation on the intrinsic reason of it: the tertiary industry is mostly labor-intensive. Unlike the primary and secondary industries, which are limited by the nature of work and have relatively fixed workplaces, the tertiary industry has high employment elasticity and absorbs more labor forces, which will form a strong human flow. In addition, the tertiary industry includes transportation and communication services. When a city has more people employed in the tertiary industry, the city's transportation and communication services may be more developed, which

can promote cross-city economic development and inter-city population movement.

After explaining the mechanism of abortion, we can try to give the following suggestion ([Analysis on key influencing factors of human flow](#)). The cities with developed tertiary industry in the region should create tertiary industry jobs that are needed by both core cities and peripheral cities, and increase the population of tertiary industry employment while expanding the urban population size, so that the connection between peripheral cities and core cities can be enhanced; in addition, the government should develop transportation infrastructure construction to reduce the cost of inter-city travel, so that the interaction of people between cities can increase and weaken the geographical limitation. At the same time, relevant departments should pay attention to the interaction of cities in the information network, improve the image of cities, and increase their internet attention, so as to strengthen the human flow. The government should exert its macro-control ability to promote close collaboration and linkage development in order to achieve regional integration.

There are still some limitations in this study. The model constructed in this paper performs well at the monthly scale, exploration of its estimation effect at a finer time scale is inadequate, due to the data limits. Besides, the parameters estimated in the model are based on the data obtained only in our study area, whether

they are applicable in other regions needs further verification.

In future research, the relationship between internet attention and human flow can be further explored to build a more accurate human mobility model.

6 Conclusions

In this paper, we estimated the amount of human flow among 21 cities of the Guangdong Province based on mobile phone positioning data, performed an almost unbiased and accurate analysis on the spatial structure of the human mobility network, and analyzed factors that may influence formation of this spatial structure. In addition to commonly used indicators, we further discussed the factors of internet attention and intercity level. The main conclusions are listed as follows:

First, cell phone location data shows intercity human flow in the Guangdong Province is a significant sale-free network, which means few nodes control most of the human flow. Cities in the PRD account for most of the human flow. The human flow network is, therefore, much denser within the PRD than in the peripheral regions; Second, the improved alter-based centrality and power can highlight important nodes and edges of the network. Zhanjiang is the only peripheral city that has a close relationship with the PRD. As a result of alter-based centrality and power, Zhanjiang can be distinguished from other peripheral cities; Besides, the improved human mobility model is better than the classical model. The model accuracy can be significantly improved by considering individual characteristics, links, and differences between cities. It is worth mentioning that daily intercity network attention enables the model to estimate the monthly human flow in the region with high accuracy; Finally, the combination of city employment, tertiary industry employment, distance, internet attention, inter-city population difference, tertiary industry employment difference, and GDP difference can explain 91% of inter-city human flow. The number of people employed in the tertiary sector and the internet attention are two important driving factors that have not been noticed in previous studies.

Abbreviations

SAR: special administration region; GDP: gross domestic product; PRD: Pearl River Delta

Acknowledgements

Not applicable.

Code availability

Not applicable.

Authors' contributions

Conceptualization, Li Zhuo, Zhuo Chen; methodology, Li Zhuo, Zhuo Chen; formal analysis, Zhuo Chen; data, Zhihui Gu; writing—original draft preparation, Zhuo Chen; writing—review and editing, Li Zhuo, Chengzhuo

Wu, Qingli Shi; funding acquisition, Li Zhuo, Zhihui Gu, Qiuping Li. All authors have read and agreed to the published version of the manuscript.

Funding

This work was funded by the National Natural Science Foundation of China (Grant No. 41971372, Grant No. 41971345 and Grant No. 51778366), the National Key R&D Program of China (Grant No. 2019YFB2103103), in part by the Natural Science Foundation of Guangdong Province (Grant No. 2020A1515010680).

Availability of data and materials

Not applicable.

Declarations

Competing interests

The authors declare that they have no conflicts of interests.

Author details

¹Guangdong Provincial Engineering Research Center for Public Security and Disaster, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China. ²Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai 519000, China. ³Guangdong Provincial Key Laboratory of Urbanization and Geo-simulation, Sun Yat-sen University, Guangzhou 510275, China. ⁴School of Architecture and Urban Planning, Shenzhen University, Shenzhen 510861, China.

Received: 16 July 2021 Accepted: 25 October 2021

Published online: 21 January 2022

References

- Ahas, R., et al. (2015). Everyday space time geographies using mobile phone based sensor data to monitor urban activity in Harbin Paris and Tallinn. *International Journal of Geographical Information Science*, 29, 2017–2039.
- Alexander, L., Jiang, S., Murga, M., & González, M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data transportation research. *Part C, Emerging technologies*, 58, 240–250. <https://doi.org/10.1016/j.trc.2015.02.018>
- Barabási, A., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286, 509. <https://doi.org/10.1126/science.286.5439.509>
- Belyi, A., Bojic, I., & Sobolevsky, S. (2017). Global multi-layer network of human mobility. *International Journal of Geographical Information Science*, 31(7), 1381–1402. <https://doi.org/10.1080/13658816.2017.1301455>
- Brands, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2), 163–177. <https://doi.org/10.1080/0022250X.2001.9990249>
- Cao, J., Li, Q., Tu, W., Gao, Q., Cao, R., & Zhong, C. (2021). Resolving urban mobility networks from individual travel graphs using massive-scale mobile phone tracking data. *Cities*, 110, 103077. <https://doi.org/10.1016/j.cities.2020.103077>
- Cao, Z., Zheng, X., Liu, Y., Li, Y., & Chen, Y. (2018). Exploring the changing patterns of China's migration and its determinants using census data of 2000 and 2010. *Habitat International*, 82, 72–82. <https://doi.org/10.1016/j.habitatint.2018.09.006>
- CNNIC. (2015). Available. http://www.cnnic.net.cn/hlwfzyj/hlwxbzg/hlwjtjbg/202102/t20210203_71361.htm. Accessed 3 Feb 2021.
- Dong, W., Wang, S., & Liu, Y. (2021). Mapping relationships between mobile phone call activity and regional function using self-organizing map computers. *Environment and Urban Systems*, 87, 101624. <https://doi.org/10.1016/j.compenvurbysys.2021.101624>
- Freeman, L. C. (1978). Centrality in social networks conceptual clarification social. *Networks*, 1, 215–239. [https://doi.org/10.1016/0378-8733\(78\)90021-7](https://doi.org/10.1016/0378-8733(78)90021-7)
- Guo, D., Jin, H., Gao, P., & Zhu, X. (2018). Detecting spatial community structure in movements. *International Journal of Geographical Information Science*, 32(7), 1326–1347. <https://doi.org/10.1080/13658816.2018.1434889>
- Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate Data Analysis* (5th ed.). Prentice-Hall.
- Heberle, R. (1938). The causes of rural-urban migration a survey of German theories. *The American Journal of Sociology*, 43(6), 6–21. <https://doi.org/10.1086/217875>

- Holland JH (1996) Hidden order: How adaptation builds complexity Vol. Addison Wesley Longman publishing co., Inc., 350 bridge pkwy suite 208 Redwood City, CA United states.
- Jin, F., Jiao, J., & Qi, Y. (2016). Evolution and geographic effects of high-speed rail in East Asia. *Acta Geographica Sinica*, 71, 576–590.
- Kou, L., Kwan, M., & Chai, Y. (2021). Living with urban sounds: Understanding the effects of human mobilities on individual sound exposure and psychological health. *Geoforum*, 126, 13–25. <https://doi.org/10.1016/j.geoforum.2021.07.011>
- Lai, J., & Pan, J. (2019). Spatial pattern of population flow among cities in China during the Spring Festival travel rush based on "tencent migration" data. *Human Geography*, 34, 108–117.
- Li, J. (2004). Generalized Gradient Theory: A New Perspective on Coordinated Regional Economic Development. *Social Science Research*, 4, 21–25.
- 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), e86026. <https://doi.org/10.1371/journal.pone.0086026>
- MIIT. (2019). 2018 communications industry statistical bulletin. Available: https://www.miit.gov.cn/gxsj/tjfx/xy/art/2020/art_f9b061284a1646498f135584d8f78757.html. Accessed 20 Aug 2021.
- Neal, Z. (2011). Differentiating Centrality and Power in the World City Network. *Urban Studies*, 48, 2733–2748. <https://doi.org/10.1177/0042098010388954>
- Neal, Z. (2012). Does World City network research need eigenvectors? *Urban Studies*, 50(8), 1648–1659. <https://doi.org/10.1177/0042098013477702>
- Ravenstein, E. G. (1885). The laws of migration. *Journal of the Statistical Society of London*, 48, 167–235.
- Shen, J., & Liu, Y. (2016). Skilled and less-skilled interregional migration in China: A comparative analysis of spatial patterns and the decision to migrate in 2000–2005. *Habitat International*, 57, 1–10. <https://doi.org/10.1016/j.habitatint.2016.06.007>
- Shi, L., Chi, G., Liu, X., & Liu, Y. (2015). Human mobility patterns in different communities: a mobile phone data-based social network approach. *Annals of GIS*, 21(1), 15–26. <https://doi.org/10.1080/19475683.2014.992372>
- Taylor, P. J., Catalano, G., & Walker, D. R. F. (2001). Measurement of the World City Network. *Urban Studies*, 13, 2367–2376. <https://doi.org/10.1080/004209802200002701>
- Tokey, A. I. (2021). Spatial association of mobility and COVID-19 infection rate in the USA: A county-level study using mobile phone location data. *Journal of Transport and Health*, 22, 101135. <https://doi.org/10.1016/j.jth.2021.101135>
- Wang B, Loo BPY (2019) The hierarchy of cities in Internet news media and Internet search: Some insights from China Cities 84:121–133. <https://doi.org/10.1016/j.cities.2018.07.013>.
- Wang, J., Dong, L., Cheng, X., Yang, W., & Liu, Y. (2019). An extended exploration and preferential return model for human mobility simulation at individual and collective levels. *Physica A: Statistical Mechanics and its Applications*, 534, 121921. <https://doi.org/10.1016/j.physa.2019.121921>
- Wang, R., Li, N., & Wang, Y. (2021). Does the returners and explorers dichotomy in urban human mobility depend on the observation duration? An empirical study in Guangzhou, China. *Sustainable Cities and Society*, 69, 102862. <https://doi.org/10.1016/j.scs.2021.102862>
- Windzio M (2018) The network of global migration 1990–2013 Soc networks 53: 20–29. <https://doi.org/10.1016/j.socnet.2017.08.006>.
- Yang, X., Fang, Z., Yin, L., Li, J., Lu, S., & Zhao, Z. (2019). Revealing the relationship of human convergence–divergence patterns and land use: A case study on Shenzhen City, China. *Cities*, 95, 102384. <https://doi.org/10.1016/j.cities.2019.06.015>
- Yu, X., et al. (2020). Quantifying the impact of daily mobility on errors in air pollution exposure estimation using mobile phone location data. *Environment International*, 141, 105772. <https://doi.org/10.1016/j.envint.2020.105772>
- Zhang, W., Chong, Z., Li, X., & Nie, G. (2020). Spatial patterns and determinant factors of population flow networks in China: Analysis on Tencent Location Big Data. *Cities*, 99, 102640. <https://doi.org/10.1016/j.cities.2020.102640>
- Zhang, W., & Thill, J. (2019). Mesoscale Structures in World City Networks. *Annals of the American Association of Geographers*, 109(3), 887–908. <https://doi.org/10.1080/24694452.2018.1484684>
- Zhang, Y., & Ng, S. T. (2021). Unveiling the rich-club phenomenon in urban mobility networks through the spatiotemporal characteristics of passenger flow. *Physica A: Statistical Mechanics and its Applications*, 584, 126377. <https://doi.org/10.1016/j.physa.2021.126377>
- Zhao, M., Liu, X., Derudder, B., Zhong, Y., & Shen, W. (2015). Mapping producer services networks in mainland Chinese cities. *Urban Studies*, 52(16), 3018–3034. <https://doi.org/10.1177/0042098014558541>
- Zhao, Z., Wei, Y., Pang, R., Wang, S., & Feng, Z. (2017). Alter-based centrality and power of Chinese city network using inter-provincial population flow. *Acta Geographica Sinica*, 72, 1032–1048. <https://doi.org/10.11821/dlxb201706007>
- Zhou, Y., Xu, R., Hu, D., Yue, Y., Li, Q., & Xia, J. (2020). Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: a modelling study using mobile phone data. *The Lancet Digital Health*, 2(8), e417–e424. [https://doi.org/10.1016/S2589-7500\(20\)30165-5](https://doi.org/10.1016/S2589-7500(20)30165-5)
- Ziyu, Z., Ye, W., Ruiqiu, P., Shijun, W., & Zhangxian, F. (2017). Alter-based centrality and power of Chinese city network using inter-provincial population flow. *Acta Geographica Sinica*, 72, 1032–1048. <https://doi.org/10.11821/dlxb201706007>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.