Simulation of the spatial distribution of urban populations based on first-aid call data

Yinan Zhou,[1] Qin Zhong Zhu,[2] Li Luo[1]

[1]*School of Public Health, Fudan University;* [2]*Shanghai Medical Emergency Center, Shanghai, China*

Abstract

We examined the feasibility of estimating the spatial distribution of urban populations based on first-aid calls based on one high-density place, the Shanghai urban area and one low-density place, the Nanhai District of Foshan City in Guangdong Province. We aggregated the population and the total number of first-aid calls on digital maps divided by grids based on a Geographic Information System (GIS). Geographically weighted regression was applied to test the correlation between the population distribution simulated by first-aid call data and the actual residency. The impact of different population densities, different grid cell sizes and different types of first-aid calls on simulation correlation were tested. We found that the use of first-aid call data could explain 60-95% of the actual population distribution in Shanghai using a grid with 1000*1000 m cell size, while the Nanhai experience was that first-aid calls could only explain 4-76% of the actual population distribution using a grid with 2000*2000 m cell size. Thus, the higher the population density, the better the simulation effect. For a high-population density area, the overall accuracy of simulation can reach as high as 0.878 at the 1-km^2 resolution. However, there are different kinds of first-aid calls and for the best estimation of the population distribution in densely populated areas, we suggest using first-aid calls from people requiring acute medical care rather than all first-aid call data.

Introduction

High-resolution population distribution data are crucial to successfully address important issues ranging from socio-environmental research to public health (Hay *et al.*, 2005; Tatem *et al.*, 2011) and resource utilization since scientific analyses, operational activities and policy decisions are significantly influenced by this issue (Bhaduri *et al.,* 2007). So far, population data are generally acquired through a census. However, depending on the time for which these data are needed, this method can be relatively poor (Wardrop *et al.,* 2018) as censuses are usually held only every ten years, while populations are constantly changing due to growth and migration. An improved method is required to provide real-time population information. In addition, the census is costly both in terms of the manpower and financial resources needed, *e.g.*, as pointed out in the press by The Ledger (2010) and The Economist (2011), the 2010 United States Census cost $13 billion, approximately $42 per capita, and 635,000 temporary enumerators were hired. Finally, from a spatial perspective, census data are limited to administrative units, which often results in a different spatial dimension when dealing with the population and another when dealing with separate study data, which limits effective integration of the information (Yang *et al.*, 2002).

In recent years, the study of the spatial distribution of populations has been developing rapidly. When establishing a model considering population and impact factors, the simulating of the spatial distribution of a population and the process of dynamic change has been useful (Cai *et al.,* 2006; Linard *et al.*, 2012). For example, remote sensing and Geographical Information Systems (GIS) can be used to establish mathematical relationships between populations and impact factors (Tobler, 1979; Martin, 1989; Dobson *et al.*, 2000), which allows the establishment of a model including data regarding both land use and spatial population distribution (Qi *et al.*, 2015). The most statistically advanced models for gauging population distributions are classified as smart interpolation, in which a wide range of auxiliary inputs, such as night-time lights, land cover and topography can provide a weighting scheme to realign population counts in proportion to weights at the grid-cell level (Reed *et al.*, 2018).

This study aimed to analyze whether there is a correlation between the spatial distribution of first-aid calls and the spatial distribution of a population. Our hypothesis was based on the relatively stable level of common diseases in a certain region within a certain time, which means that the number of people being affected by one (or more) diseases is mainly related to the number

of people in that region. Due to the wide application of GPS real-time systems of emergency centres (Mintsis *et al.*, 2004; Athavan *et al.*, 2012), it is generally easy to get the precise coordinates of people making the first-aid calls. GPS real-time tracking means utilization of a vehicle-operated GPS instrument or a handset to send positioning messages (expressed as geographical coordinates) to the surveillance centres by the wireless Global System for Mobile communications (GSM), (Yao *et al.*, 2011). The distribution data provided by first-aid calls is a ready-made system which is completely accurate with respect to the individual making the call, which means that this method is not subject to administrative divisions. We analyzed the relationship between the distribution of first-aid calls and the population distribution as well as the impact of different types of first-aid calls, different population densities and different grid sizes with regard to correlation.

Materials and methods

Databases

In order to test the applicability of this method, we selected two Chinese sample places with different population densities, *i.e.* the urban area of Shanghai on the east coast of China and Nanhai District, an urban area of Foshan City in Guangdong Province. With 7,258,043 people on 328.3 km^2 (22,107/km^2), the former has one of the highest population densities in the world, while Nanhai District of Foshan City, one of China's second-tier cities located in the south-eastern coastal area, has a population of 2,800,646 and the land area is 1,107.8 km^2, which translates into a population density of 2,528 people/km^2. Table 1 displays the acquired data used.

Since first-aid call data is the key data for research, detailed information including the addresses of the people making the calls (expressed in coordinates), time of the call, reason for the call (first-aid calls for acute medical care, transfer between hospitals, going home from hospital, etc.). Detailed information of first-aid call data in the two places is described in Tables 2 and 3.

Simulating the county-level administrative borders of Nanhai District

Lacking access to a ready-made, county-level digital map of Nanhai District, we simulated it by a GIS approach applying the "Thiessen Polygon" tool in ArcGIS v10.4.1 by ESRI (Redlands, CA, USA). After obtaining the population data at the county-level of Nanhai District from the Population Office of the Nanhai Public Security Bureau, we located each county and converted it into coordinate data using the location of the county administrative building as the centre. After this, we applied the "Thiessen Polygon" tool to simulate the county-level administrative division (left picture in Figure 1).
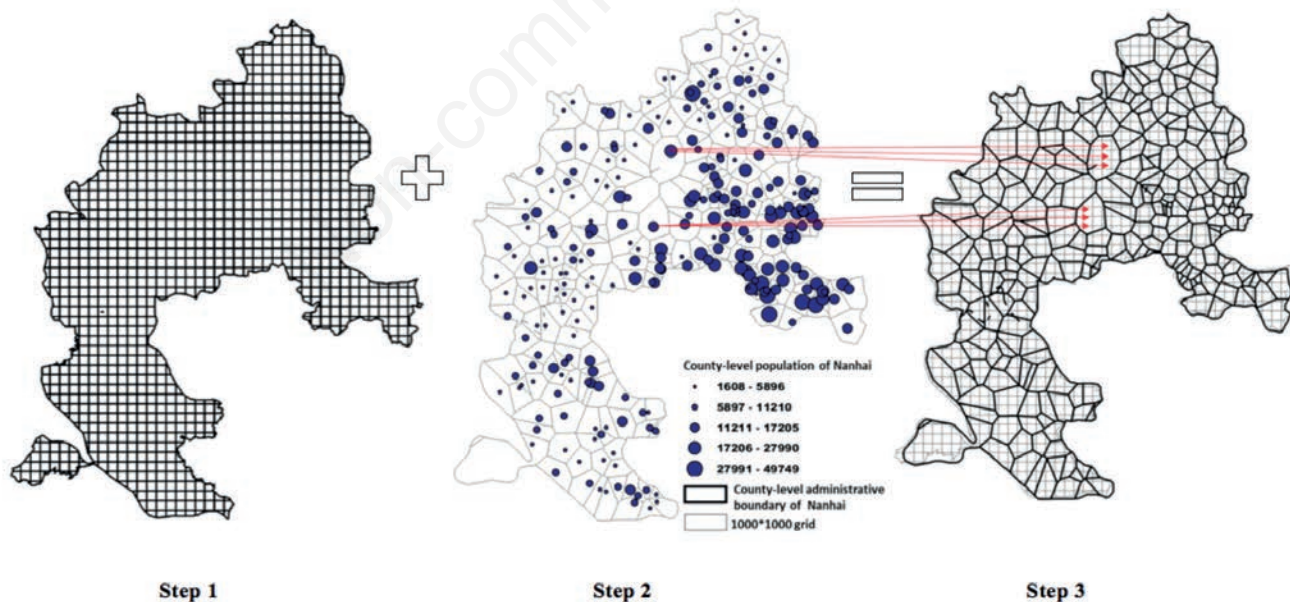

Figure 1. Transforming administrative level population data to the cells of Nanhai District. Step 1: Divide the area into 1km*1km cell; Step 2: County-level population data and the administrative boundary of Nanhai was acquired. (each dot represents a county, the size of the circle represents the size of the population, i.e. larger circles represent a higher population); Step 3: Step 1 and Step 2 were combined, then the total number of cells in each county were counted. The population in each cell were calculated as follows: Population per cell = population of the county divided by the number of cells of the county (assuming that the population in the cells in each county is evenly distributed).

Table 1. Overview of acquired data

Category	Details	Source	Time covered
Shanghai urban area population data based on residential buildings	All buildings (point data)	Population Office Bureau of Shanghai Public Security	Data of December 2016
Population data of Nanhai District's 267 counties	County level data	Population Office of Nanhai Public Security Bureau	Data of December 2015
Digital map of the Shanghai urban area (geographic data)	County level data	Shanghai Urban Planning and Design Bureau	Data of December 2016
Digital map of geographic data of Nanhai District	County level data	Simulated by Thiessen Polygon	Data of December 2016
Shanghai first-aid call data	Point data	Shanghai Medical Emergency Center.	01/01/2016- 31/12/2016
Nanhai first-aid call data	Point data	Nanhai Medical Emergency Center	01/01/2015- 31/12/2015

The grid method

Aggregating population by applying a grid system makes it easier to integrate different kinds of spatial data for interdisciplinary comprehensive research (Deichmann *et al.,* 2001). Many approaches to the production of specialized grid-formatted population maps exist (Jia, 2012; Reed *et al.*, 2018). In this study, Shanghai's urban area was divided into four different grid sizes: 250*250 m, 500*500 m, 750*750 m and 1000*1000 m. The resolution used for Nanhai District is lower because of the lower population density, so the grid sizes there were: 1000*1000 m, 1500*1500 m, 2000*2000 m and 2500*2500 m. To do this we used the "Create Fishnet" tool in ArcGIS.

Aggregating first-aid calls and population data

Using ArcGIS, we mapped the first-aid call data and resident data in the form of point data onto digital maps of the two cities divided by grids into cells. These point data can be aggregated by cells of any resolution according to the study needs using the "Spatial Join" tool.

The population data of Shanghai in 2016 were aggregated by residential buildings, which means that the coordinates and population of each residential building were included. Whereas population data are often aggregated by administrative division, such as in Nanhai District, the way of transforming population data from county to grid cells is to distribute the population evenly within the county. First, we had to divide the Nanhai District by higher resolution than the research needed and the county (here we used grid cells of 200*200 m). Then we summed the number of cells in each county in the ArcGIS "Attribute Table" and distributed the population of the whole county equally into each grid cell (the average population of each grid cell of the county then equals the population of the county divided by the number of cells of the county). The method is illustrated in Figures 1 and 2 (The grid size of Figure 2 is 1000*1000 m for clearer display).

Simulating the distribution of population data by first-aid call data

After aggregating the number of first-aid calls and population data within each cell, the population in each cell can be simulated by the distribution of first-aid calls. The simulated population in each cell equals the proportion of first-aid calls in the cell to the total number of first-aid calls in the sample area times the number of people in this area, *i.e.* the total population is the same as the

Table 2. Types of first-aid call in the Shanghai urban area.

Type of call	Call count
First-aid calls requiring acute medical care	161,988 (51.9%)
Going home from hospitals	49,774 (15.9%)
Other reasons than medical treatment	37,311 (11.9%)
Transfer between hospitals	63,332 (20.3%)

Table 3. Types of first-aid call in Nanhai District

Type of call	Call count
First-aid calls requiring acute medical care	18,240 (57.0%)
Other reasons than medical treatment	13,757 (43.0%)

actual population. However, the population distribution is simulated by the distribution of first-aid calls (Figure 2).

Relevance and effectual factors

Geographically weighted regression (Gwr) in GIS was applied to test the correlation between the distribution of population simulated by first-aid calls and actual residents, and how much of the population distribution can be explained by the distribution of first-aid calls. R^2 is a measure of goodness of fit whose value varies from 0.0 to 1.0, with higher values being preferable. It may be interpreted as the proportion of dependent variable variance accounted for by the regression model. The dependent variable is the actual population, and the explanatory variable is the simulated population.

First, the impact of different population densities on relevance was tested. We chose a high population density area and a low population density area to test the relevance of the method.

Second, the effect of different grid sizes on correlation was tested. Theoretically, a larger scale of the selected area would have a relatively large amount of data, so that the impact of extreme values would be relatively small and the stability better. Therefore, we explored the impact of different regional sizes on the simulation accuracy.

Third, the effect of the different types of first-aid call on correlation was tested. Theoretically, only first-aid calls requiring acute

medical care are related to the occurrence of accidents or acute illnesses and may have an indirect correlation with the population distribution. Therefore, we analyzed the correlation between the actual population distribution and all calls as well as the correlation between the actual population distribution and acute first-aid calls only.

Software

Primary data organization was performed using SAS 9.4. Mapping, calculation of spatial analysis, Gwr analysis were performed using ArcGIS v10.4.1 by ESRI (Redlands, CA, USA).

Results

Visualization of the distribution of actual and simulated populations

We visualized the population data simulated by the first-aid calls as well as the actual population both in Shanghai and Nanhai by color. As seen in Figure 2, the two distributions are very similar.

Correlation between the actual resident and the simulated population

We further tested the relationship between the actual population distribution and the simulated population distribution through the ArcGIS Gwr tool. In Shanghai, we can see the adjusted R^2 range from 0.544 to 0.878 as the grid size changes from 250*250 m to 1000*1000 m (Table 4). By visualizing the map of Local R^2 and Std.residual (Figure 3) using the 1000*1000 m grid size, we can see that the R^2 ranges from 0.60 to 0.95 in Shanghai, which means the distribution of first-aid calls can explain 60% to 95% of the actual population distribution in Shanghai using the 1000*1000 m grid size. We later obtained the Shanghai population and first-aid call data for the year of 2017, and the simulation using this method again showed a high correlation (87.3%) between the two by Gwr. In Nanhai, the adjusted R^2 ranges from 0.606 to 0.818 as the grid size changes from 1000*1000 m to 2500*2500 m (Table 5) and the distribution of first-aid calls can explain 4-76% of the actual population distribution using the 2000*2000 m grid size (Figure 3).
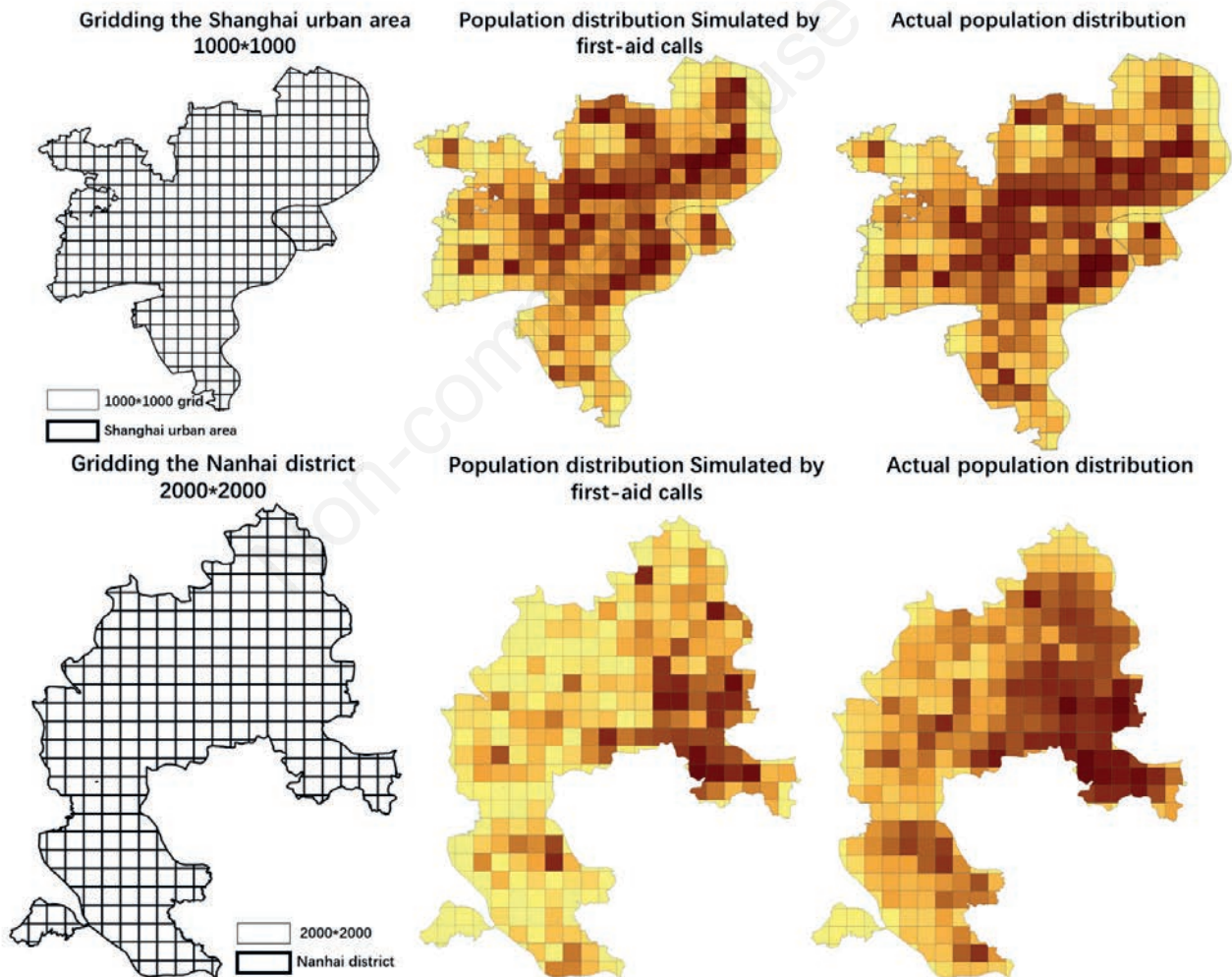

Figure 2. Population distribution simulated by first-aid call data of the Shanghai urban area in 2016 and Nanhai District in 2015. First-aid call data in the middle and the actual population distribution to the right. The deeper the colour, the larger the number of calls, and number of residents.

Influencing factors

First, the higher the population density, the better the simulation effect. With the same resolution of 1000*1000 m grid, the simulation effect is better in Shanghai than in Nanhai.

Second, the larger the size of the grid division, the higher the simulation accuracy (Table 4; Table 5). Both these results suggest that the more the average number of each grid cell, the better the stability of the model.

Third, first-aid call data for treatment only simulated better than the total number of first-aid call data in Shanghai (Table 4), and the lower the resolution, the more obvious the difference, whereas in Nanhai the difference is not obvious. This result suggests that the amount of data has a greater impact on the correlation. Since the density in Nanhai is relatively low, screening the acute first-aid call data only would result in fewer data per grid cell.

Discussion

The hypothesis was that the incidence of common diseases in a certain region within a certain time period is relatively stable, so the number of people affected by these diseases should mainly be related to the population number. Testing the methodology in a high- and low-density area, we learnt the following. First, there is a correlation between the population simulated by the distribution of first-aid call data and the actual population distribution. Second, the higher the population density, the better the simulation effect, for a high-population density area, the distribution of first-aid calls can explain 87.8% of the actual population distribution at the 1-km^2 grid resolution. In relatively low-population density areas, however, the simulation accuracy was lower. Third, the larger the size of the grid divisions (the cells), the higher the simulation accuracy. Fourth, in a high-density area, for the best estimation of the population distribution, we suggest using the data that corresponds to first-aid calls for acute medical care rather than all first-aid call data. It is worth noting that Shanghai's population data is accurate as it relies on the point data of each building, so it can be accurately aggregated at any dimension, which means that the result should be relatively reliable. However, we did not obtain a similarly accurate actual population distribution data in Nanhai, so it had to be estimated, which partly explains its much lower correlation.

Existing spatial demographic datasets for low-income countries are generally based on outdated and low-resolution data (Tatem *et al.,* 2011; Wardrop *et al.*, 2018). The strength of the methodology described here depends on its accuracy and spatially detailed population distribution data, such as first-aid call data. This can make accessibility measures considerably more valuable than those obtained through aggregated population data. The main problem of most previous GIS-based approaches is the lack of sufficiently detailed population and settlement distribution, which impacts the availability and reliability of spatially-detailed accessibility information (Linard *et al*., 2012).

In this study, the input data for simulation of the population distribution was first-aid call data and the total population of the region. The latter is easy to obtain through annual statistics reports, while the ambulance GPS system provides a ready-made first-aid call data expressed as latitude and longitude automatically provided by the driver when the emergency vehicle arrives at the point of call, so there is no need for additional geolocation. This approach allows a relatively accurate estimate of the population distribution in areas where it is not easy to obtain population information. In addition, first-aid calls data provide an estimate of the population distribution without administrative boundaries, so as to aggregate the population data at any dimension according to the research needs.

Population data from the Census Bureau are usually counted and aggregated in accordance with administrative boundaries. In most places, the county level is the highest resolution (Qi *et al*., 2015). First-aid call data is a point dataset that is not subject to administrative division. We can therefore divide the area as needed according to the purpose of the study. For example, World Health Organization (WHO) recommends that Community Health Centres (CHCs) be located so that all people can reach one within

Table 4. Gwr-adjusted R^2 applied for grid cell size and type of first-aid calls in Shanghai.

Grid cell size	Number of grid cells	Gwr-adjusted R^2 (acute calls)	Gwr-adjusted R^2 (all calls)
250*250	4,876	0.544	0.542
500*500	1,284	0.678	0.655
750*750	602	0.802	0.707
1000*1000	355	0.878	0.705

Table 5. Gwr-adjusted R2 applied for grid cell size and type of first-aid calls in Nanhai.

Grid cell size	Number of grid cells	Gwr-adjusted R^2 (acute calls)	Gwr-adjusted R^2 (all calls)
1000*1000	1,091	0.606	0.685
1500*1500	522	0.535	0.533
2000*2000	307	0.692	0.693
2500*2500	204	0.818	0.819

15 minutes. It is clear that the use of data sorted by administrative divisions is not suitable for this kind of planning, but the use of first-aid calls makes it possible to divide the areas in question by grids into cells of the required 15-minute walk size. Thus, this approach should be helpful for social, economic, and scientific research (Zhou *et al.,* 2018).

This method is applicable to the study of urban areas, especially areas with a high population density. As mentioned earlier, the correlation between simulated population distribution and an actual population distribution was significantly reduced for areas with low population density. Thus, it is not applicable to use first-aid call data to simulate the population distribution in rural areas where the population is sparse. An additional limitation is the fact that ready-made first-aid call data with coordinates are only available in an ambulance equipped with a GPS system is a limitation. Although such data are readily available in some areas, researchers
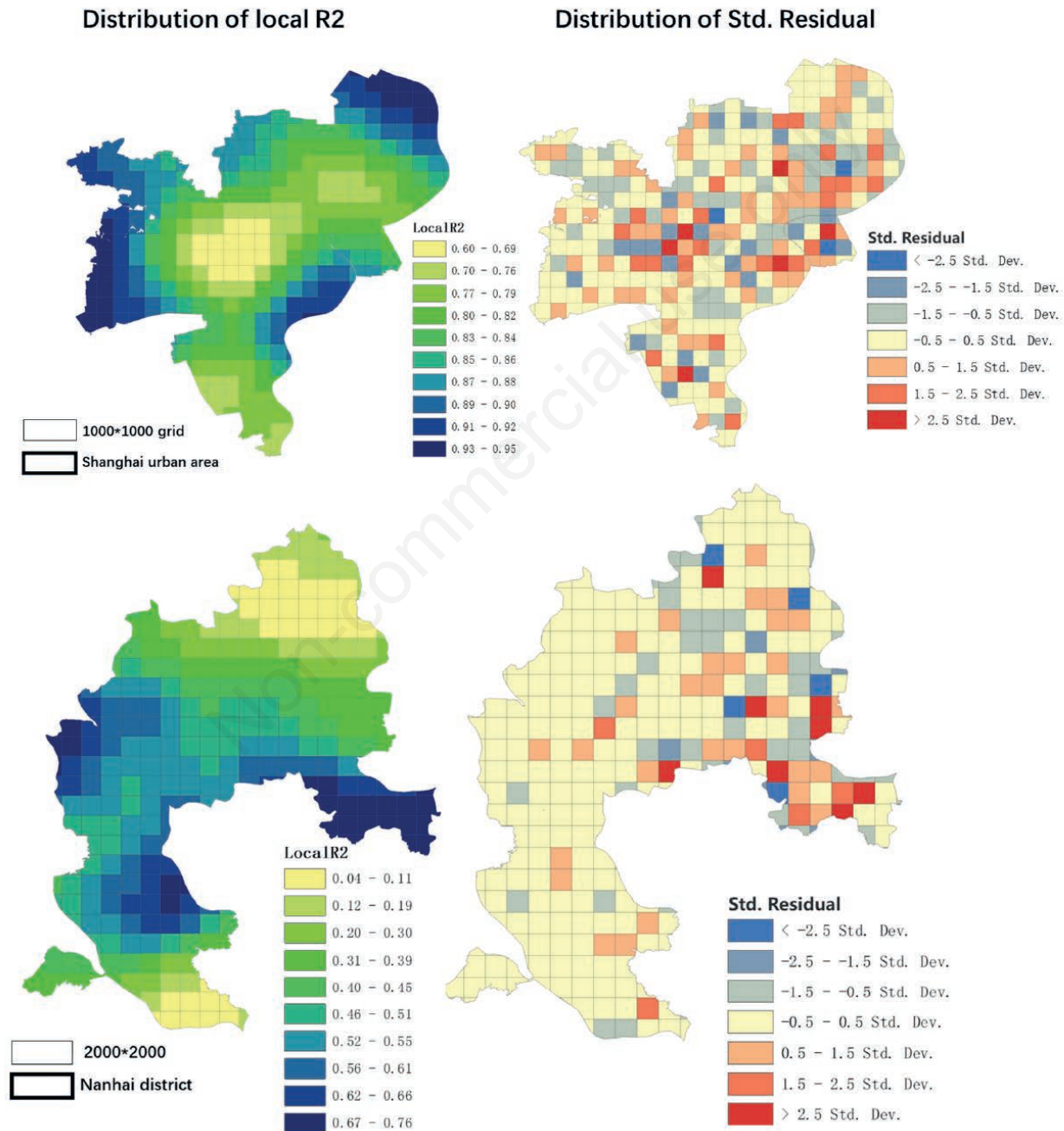

Figure 3. Distribution of local R^2 and Std.residual by the Gwr method.

still have difficulty in obtaining the data because that are not made public.

The main function of this article is to stimulate further studies. We maintain that the method presented here has certain practical prospective applications and may prove to be a good foundation for the development of a general model for the estimation of population distribution. However, in this work, we only used data from two cities, are considering only three influencing factors. Thus, more studies need to be done to further test applicability and improve feasibility of this method.

Conclusions

The first-aid call approach represents a rapid and effective approach to simulate population distributions. It is also simple and inexpensive and not hindered by administrative divisions.

References

Athavan K, Balasubramanian G, Jagadeeshwaran S, Dinesh N, 2012. Automatic ambulance rescue system. In: 2012 Second International Conference on Advanced Computing & Communication Technologies. IEEE 190-95. doi:10.1109/ACCT.2012.34

Bhaduri B, Bright E, Coleman P, Urban M, 2007. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. GeoJ 69:103-17. Available from: https://link.springer.com/article/10.1007/s10708-007-9105-9

Cai Q, Rushton G, Bhaduri B, Bright E, Coleman P, 2006. Estimating small area populations by age and sex using spatial interpolation and statistical inference methods. Transactions GIS 10:577-98.

Deichmann U, Balk D, Yetman G, 2001. Transforming population data for interdisciplinary usages: from census to grid. Washington (DC): Center for International Earth Science Information Network, 2001. Available from: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.529.9943&rep=rep1&type=pdf

Dobson JE, Bright EA, Coleman PR, Durfee RC, Worley BA, 2000. LandScan: a global population database for estimating populations at risk. Photogrammetric Eng Remote Sens 66:849-57.

Hay SI, Noor AM, Nelson A, Tatem AJ, 2005. The accuracy of human population maps for public health application. Trop Med Int Health 10:1073-86.

Jia P, 2012. Modeling High-resolution Gridded Population Surface in Alachua County, Florida. University of Florida. Available from: https://ufdc.ufl.edu/UFE0044292/00001

Linard C, Gilbert M, Snow RW, Noor AM, Tatem AJ, 2012. Population distribution, settlement patterns and accessibility across Africa in 2010. PLoS One 7:e31743. doi:10.1371/journal.pone.0031743

Martin D, 1989. Mapping population data from zone centroid locations. Trans Inst Br Geogr 14:90-7.

Mintsis G, Basbas S, Papaioannou P, Taxiltaris C, Tziavos IN, 2004. Applications of GPS technology in the land transportation system. Eur J Oper Res 152:399-409.

Qi W, Liu S, Gao X, Zhao M, 2015. Modeling the spatial distribution of urban population during the daytime and at night based on land use: A case study in Beijing, China. J Geograph Sci 25:756-68. Available from: https://link.springer.com/article/10.1007/s11442-015-1200-0

Reed FJ, Gaughan AE, Stevens FR, Yetman G, Sorichetta A, Tatem AJ, 2018. Gridded population maps informed by different built settlement products. Data 3:33.

Tatem AJ, Campiz N, Gething PW, Snow RW, Linard C, 2011. The effects of spatial population dataset choice on estimates of population at risk of disease. Popul Health Metr 9:4.

The Economist, 2011. Censuses: Costing the count. Available from: https://www.economist.com/international/2011/06/02/costing-the-count

The Ledger, 2010. US Census Takers Attacked on the Job. National Ledger. Available from: https://nationalledger.com/2010/05/us-census-takers-attacked-on-the-job/

Tobler WR, 1979. Smooth pycnophylactic interpolation for geographical regions. J Am Stat Assoc 74:519-30.

Wardrop NA, Jochem WC, Bird TJ, Chamberlain HR, Clarke D, et al., 2018. Spatially disaggregated population estimates in the absence of national population and housing census data. PNAS 115:3529-37. doi:10.1073/pnas.1715305115

Yang X, Wang N, Jiang D, Xiong L, Liu H, 2002. Regionalization of Population Distribution Based on Spatial Analysis. Acta Geographica Sinica 57:76-81Available from: http://www.geog.com.cn/EN/abstract/abstract18914.shtml

Yao H, Dong W, Liang D, Rogner A, Lai J, 2011. Application of GIS on emergency rescue. Procedia Eng 11:185-8.

Zhou Y, Bai G, Luo L, 2018. Development of a hexagonal, mesh-based distribution method for community health centres. Geospat Health 13:648. doi:10.4081/gh.2018.648