Taking household data as ancillary information in areal interpolation

Wen Zeng[*1], Alexis Comber[†1]

[1] School of Geography, University of Leeds

January, 2019

Summary

This research explores the use of different areal interpolation methods to generate a spatially distributed surface of population from high level population data to lower-level units at different scales. Leeds in the UK and Qingdao in China have been taken as the case study areas. Seven classic methods and two augmented methods were applied. The results were compared with interpolation across the same scales based on household percentage method using household data from the population census (UK) and POI data from house sales (China). Correlation test was used to determine the best areal interpolation method.

KEYWORDS: Areal interpolation, interpolation method, population distribution, household data,

1. Introduction

Areal interpolation transfers attribute information from source zones with known values to smaller target zones with unknown values (Goodchild and Lam 1980). The areal interpolation problem is more common associated with geographical analysis than other fields (Lam 1983). Over the last three decades, various approaches for areal interpolation have been developed based on different assumptions about the underlying distributions of variables with the emerging of new data and methods (Comber et al., 2008; Langford, 2013; Bakillah et al., 2014; Lin and Cromley, 2015; Mennis, 2016). The aim of this research is to develop a local distribution of population in order to support health care planning and supply and demand modelling. Census data is not reported over small areas in China, but proxies for households and therefore population in China can be generated from housing rental information. On the other hand, household number has a well-known strong relationship with population as many previous studies have found. This paper uses the cases of the UK and China where lower level Geographies are known, to explore approaches for estimating population at lower scales in situations where they are unknown. This research firstly examines correlation between households and population over different Eds, and then compares well known methods of interpolation with estimations based on estimates of household number.

2. Methods

Study areas and data

This study uses Leeds in the UK and Qingdao in China as the case study areas. Leeds is the UK's third largest Metropolitan District with a population of 751,500 people according to the 2011 Census. Its boundaries cover some 552 square kilometres with a built-up area to the centre and south surrounded by a number of separate small towns and villages in a polycentric pattern. Qingdao is a typical large city in east coast of China, with the population of 3,779,000 in the central city area of this study according to the 2010 Census. The area is about 1407 square kilometres. The UK data is from the 2011 UK Census at scales from ward level (~6,500 people) to output area level (~300 people) and includes the population and household number for each enumeration district (ED) as well as geographical

[*] geowz@leeds.ac.uk
[†] a.comber@leeds.ac.uk

boundary information. The Chinese data is mainly from two sources. One derives from the 2010 Census and includes population and household number at district and subdistrict level. The other is the POI data, which is from Lianjia housing sales data. Lianjia is a property agency in China and crawler software was used to capture the data from the website. This data contains the location, household number, housing price, building numbers and other information about the residential quarter. The data has more than 2,800 records. The geographical administrative data of Qingdao was obtained from the website of National Geomatic Centre of China (NGCC).

Analysis

The objective of this research is to compare classic population interpolation methods with those based on household number information at multi-scales. It first examines the relationships between population and household number over different enumeration district (ED) scales to identify the correlation between households and population. Although a correlation between population and household number is expected, it is important to establish this relationship. Then, having determined relationships between different population measures over different ED scales, the study applies classic areal interpolation methods to the UK and China case studies to generate population estimates from high level source zone EDs to low level target zone EDs (Ward to LSOA in the UK, District to sub-District in China). **Binary dasymetric method and six classic methods without ancillary information were adopted**: simple area weighted, Areal weighted, Nearest Neighbour Interpolation, Pycnophylactic, Kriging and Inverse Distance Weighted method) at multiple scales. The two methods without ancillary information but with the highest correlation coefficients were chosen to be modified into weighted interpolation methods using the target zone household number as an auxiliary information. The correlation coefficients of the predicted population and real population in all methods were calculated to reflect their accuracy. The results of all above methods are compared with interpolation at the same scales using on household number (the percentage of the higher-level household total in each lower level area) using data from the population census (UK) and POI data from house sales (China) to identify which method is the best.

3. Results

In order to adopt household number as the auxiliary information, we need to identify the relationship between household number and population. Thus, the correlation coefficients of the two variables were calculated at multi-scales. The relationship between population and household number gets weaker the finer the scale of the geographic unit in the UK case. While in the Chinese case, the relationship is rather robust on both scales. The correlation coefficients between real household number and population are above 0.97, and the correlation between household number estimated by POI data and population at subdistrict scale is also strong (0.785). Although the correlation of population and household number estimated by POI is not very significant at district level, the number of districts in Chinese case of this study is only 6 and it can not prove that the relationship at district level is not strong.

Based on the established relationship between household number and population, we can examine different interpolation methods and household-weighted interpolation methods. From the results, Pycnophylactic and NNI are the relatively best methods with highest correlation coefficients in all methods without ancillary information. Although the correlation coefficients of areal weighted method are higher than that of NNI at some scales in the UK case, the Chinese case cannot use areal weighted method. Despite simple weighted method performs better in the UK case, it fails in Chinese case. Thus, in order to compare classic and weighted interpolation methods, this study chose Pycnophylactic and NNI to be modified into weighted methods. However, the correlation between real population and population predicted by all areal interpolation methods without ancillary information are generally poor in both cases, the coefficients of which are all below 0.6. Especially for Kriging and IDW methods, the correlations between predicted population and real population are very poor. Thus, it seems to be needed to modify existing interpolation methods to obtain better results.

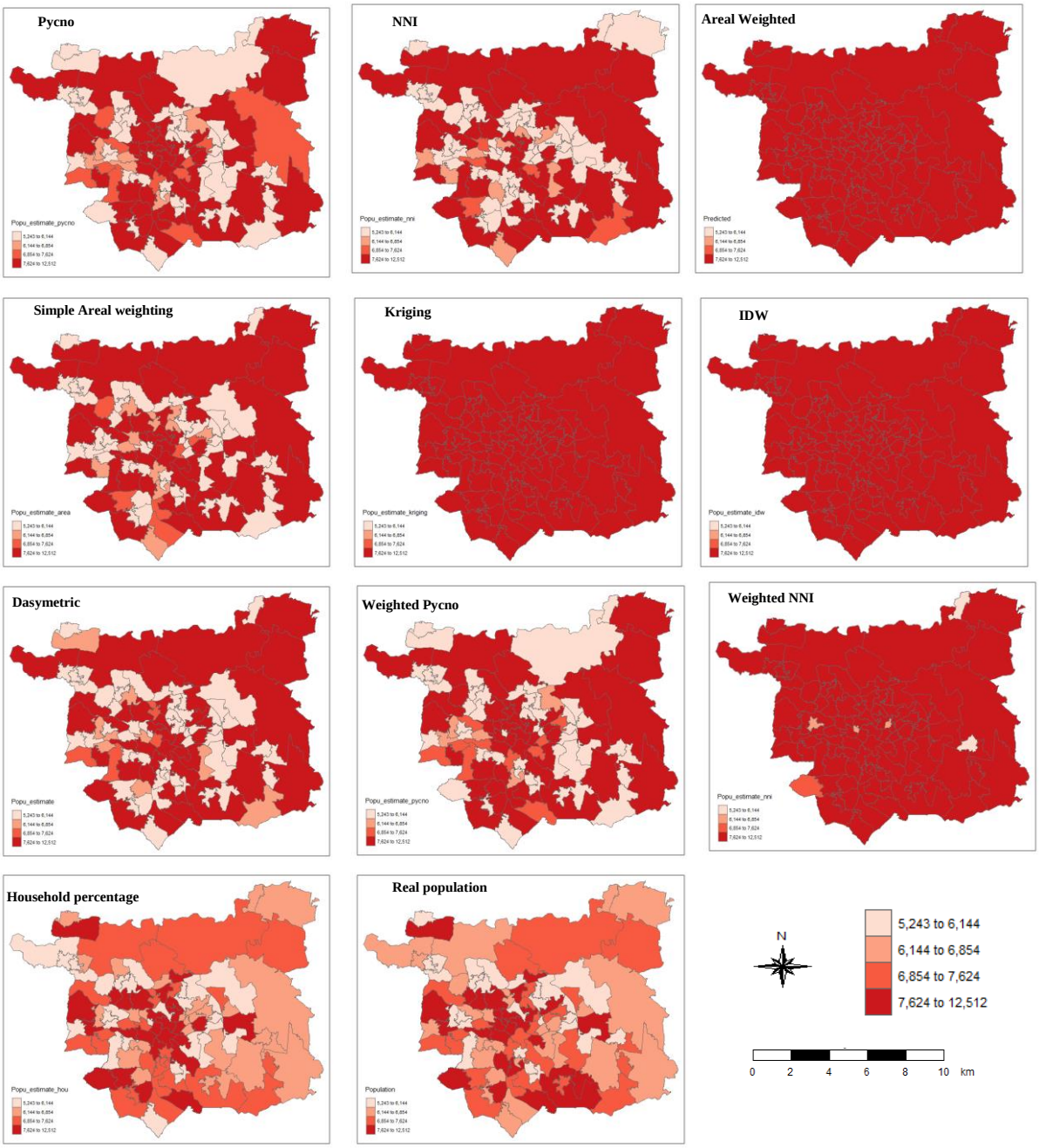

Figure 1 results of different areal interpolation methods at MSOA level predicted from Ward population in Leeds, UK

The pycno and NNI were modified into weighted interpolation methods with household number as weights. However, the correlation coefficients of both weighted methods in the UK case and weighted pycno method in Chinese case were just a little bit improved. The results are obviously unsatisfactory. Then, this study adopts the simple household number percentage method to re-distribute population into lower-level EDs. It is very surprising that the correlation coefficients at all scales are much higher than that of both classic and weighted interpolation methods. This means the simple household percentage method can estimate population very well from higher-level to lower-level enumeration districts.

4. Discussion and outlook

It can be seen that simple household percentage method is the best among all methods in this research. This method only uses household number as ancillary information, which can be obtained from many channels. This suggests that there may be opportunities from the many new forms of data arising from portals, open data initiatives, APIs, etc., much of which always have location attached, to augment and even supplant a number of current methods for manipulating spatial data. In developing countries where low-scale census data is rarely open to the public, open access data sets and volunteered geographic information provide new chances to geographical researchers. For instance, based on house sales data, this research generates the predicted population surface in Qingdao, which is used for healthcare facilities planning. This population surface can be used in health care planning and supply and demand modelling.

Besides, the results of this work suggest some key opportunities for methods for working over small scale geographies and with microsimulation data arising from the many new forms of data, particularly where these can be easily related to specific socio-economic variables, such as with household and resident population.

References

Bakillah, M., Liang, S., Mobasheri, A., Jokar Arsanjani, J., & Zipf, A. (2014). Fine-resolution population mapping using OpenStreetMap points-of-interest. International Journal of Geographical Information Science, 28(9), 1940-1963.

Comber, A., Proctor, C., & Anthony, S. (2008). The creation of a national agricultural land use dataset: combining pycnophylactic interpolation with dasymetric mapping techniques. Transactions in GIS, 12(6), 775-791.

Fisher, P. F., & Langford, M. (1996). Modeling sensitivity to accuracy in classified imagery: A study of areal interpolation by dasymetric mapping. The Professional Geographer, 48(3), 299-309.

Goodchild, M.F. and Lam, N.S.-N., 1980. Areal interpolation: a variant of the traditional spatial problem. Geo-Processing, 1, 297–312.

Lam, N. S. N. (1983). Spatial interpolation methods: a review. The American Cartographer, 10(2), 129-150.

Langford, M. (2013). An evaluation of small area population estimation techniques using open access ancillary data. Geographical Analysis, 45(3), 324-344.

Lin, J., & Cromley, R. G. (2015). Evaluating geo-located Twitter data as a control layer for areal interpolation of population. Applied Geography, 58, 41-47.

Mennis, J. (2016). Dasymetric spatiotemporal interpolation. The Professional Geographer, 68(1), 92-102.

Biographies

Dr. **Wen Zeng** is a post-doctoral researcher at the School of Geography, University of Leeds. His interests are quality of life, accessibility, urban and community development and social inequalities from geographical perspective. His current research is about population and health planning.

Professor **Lex Comber** holds a Chair in Spatial Data Analytics at the School of Geography. Lex is a leading international researcher in many areas of spatial science and geocomputation, with publications in accessibility, facility location optimisation, graph and network theory, spatial data uncertainty, citizen science, land use / land cover and remote sensing.