Population estimation in small areas: combining dasymetric mapping with pycnophylactic interpolation

Jega Idris Mohammed[1], Alexis Comber[2], Chris Brunsdon[3]

[1&2]Department of Geography, University of Leicester, Leicester, LE1 7RH, UK
Telephone: +44(0)116 252 3823, Fax: +44(0)116 252 3854
E-mail: ijm14@le.ac.uk[1] , ajc36@le.ac.uk[2]

[3]Department of Geography, University of Liverpool, Liverpool, L69 3BX, UK
Telephone: +44(0)151 794 2000
E-mail: Christopher.Brunsdon@liverpool.ac.uk[3]

ABSTRACT: Population censuses at fine levels of spatial detail provide potential demand information for effective health care planning and policy formulation. Previous studies have used different methods of areal interpolation to disaggregate population data to small areas. This study demonstrates the utility of combining dasymetric mapping with pycnophylactic interpolation to estimate population in small areas. The results were evaluated by comparing them with actual census data and measured using Root Mean Square Error (RMSE) and adjusted Root Mean Square Error (Adj-RMSE). The results show that the interpolated populations are reliable and suitable for use with location-allocation analyses of health facilities.

KEYWORDS: Population estimation, Areal interpolation, Health care planning, Dasymetric, Pycnophylactic

1. Introduction

Population estimates for small areas contribute significantly in analyses of spatial data. In analysing accessibility to public facilities (e.g. health centres), policy makers and planners need to have detailed information on population size to be capable of estimating facility demand. Geographic Information of an area at finer scale provides specific information based on local population characteristics which assist in coordinating, monitoring and evaluating service delivery (Curtis and Taket, 1989). This must be organised for effective planning and evaluation of health services (World Health Organisation, 1987). Population data from census in small areas is essential for the analysis of access in relation to demand and for supply for health service resources.

Population census data in some countries (e.g. Nigeria) are published only as spatially aggregate data for States and Local Government Areas. Health plans are made based on these larger estimates of the population and not use more detailed population data relating to small areas. There is a need for such data to be disaggregated to small areas to facilitate more robust spatial analysis. Areal interpolation is the process of estimating population distributions from aggregated census level to small areas within the aggregated boundary (Mennis, 2003). In order to overcome this problem, previous studies have used different techniques for areal interpolation to estimate population census data based on different assumptions about the original allocation of the known data and its dimensions (Hawley and Moellering, 2005). The two classes of techniques are: dasymetric techniques that use ancillary data (e.g. remote sensing, road network data) and those that do not use ancillary data. Regardless of approach, the major difficulty in applying interpolation techniques is that the estimation of data over small areas changes the aggregated boundary and effects the results of spatial analysis (Openshaw, 1984).

This study addresses the problem of estimating aggregated population census data to small areas by combining dasymetric mapping and pycnophylactic interpolation. The objective is to disaggregate population census data to small areas within the study area.

2. Areal interpolation

Areal interpolation is the transformation of aggregated population census data to where data is needed.

2.1 Areal interpolation methods using ancillary data

Two well-known techniques using ancillary data are dasymetric methods using remote sensing data and the road network method using road data. The dasymetric technique is volume preserving and residential land use types are represented using a two-dimensional zone system. The technique makes it easier to mask out known non-residential areas and gives better information about the distribution of population (Cai *et al.,* 2006). This technique was used by Wright (1936) to produce a density map and also estimates population distribution of Cape Cod using topographic sheet as ancillary data. Eicher and Brewer (2001) enhanced their analysis of socio-economic variables using urban land use data.

The road network technique estimates original values using one-dimensional street networks as ancillary data (Reibel and Bufalino, 2005). The technique assumes distribution of housing units, which identify areas of high population density, correlates with road networks (Brinegar and Popick, 2010). This is important where population is the variable of interest because most residential homes are located on road network. Xie (1995) used road network as ancillary data and developed three algorithms based around road classification, road length and internal node counts. Reibel and Bufalino (2005) interpolate 2000 census data in Los Angeles from 1990 census data using network length method with street network data (TIGER files from the U.S. census).

2.2 Areal interpolation methods that do not use ancillary data

Two common interpolation techniques are pycnophylactic and areal weighting methods. The pycnophylactic approach (Tobler, 1979) predicts target zone estimates as volumes within each zone. It preserves the total volume and generates a 2 ½ dimensional continuously smooth surface (Cai *et al.*, 2006). This technique has been widely applied in different research areas. Some of the applications include triangulated Irregular Networks, TIN (Rase, 2001), point in polygon (Okabe and Sadahiro, 1997) geostatistical method of kriging (Kyriakidis, 2004) and modelling malaria in Kenya (Hay *et al*., 2005). Comber *et al.* (2008) used it to spatially disaggregate UK agricultural census data.

The areal weighting technique is a two-dimensional polygon overlay method that maintains volume and assume population is uniformly spread within the source zones (Lam, 1983). The disadvantage of this technique is the assumption of uniform distribution of population (Kim and Yao, 2010). Cromley *et al.* (2009) used areal weighting technique to correct changes in boundaries that occur between censuses in China. The result shows the methodology is applicable to areas with repeated change in unit boundaries.

3. Method

The methodology describes the use of a combination of dasymetric mapping and pycnophylactic interpolation to disaggregate population census data to small areas. The data used include:

- Land Cover/ Land use map of Leicester, UK
- 2001 population census of Leicester at Lower Super Output Areas (LSOAs) as the source zones.

- 2001 population census of Leicester at Output Areas(OAs) as the target zones

The methodology was carried out in two stages:

Stage 1: Binary dasymetric mapping

Binary dasymetric was chosen because previous research has shown no improvement in the accuracy by selecting multi-class (Langford, 2007). The technique was used to assign population density values over all the pixels in the study area with no values assigned to known large parks (green space areas), thereby creating a new polygon of the study area with the total population at LSOAs. A flow chart is shown in Figure 1 below.

Stage 2: Pycnophylactic interpolation

The pycnophylactic interpolation smoothes values assigned to each residential pixel. A 'Pycno' function written in R using a 30m grid was applied to the polygon created in Stage 1. The total population at LSOAs were disaggregated within the study area with the total source volume preserved as in Figure 1. The output, population density surface was converted to a points' file for further analysis. Allocation of census data from LSOAs to OAs in Leicester, UK was used to illustrate the method.

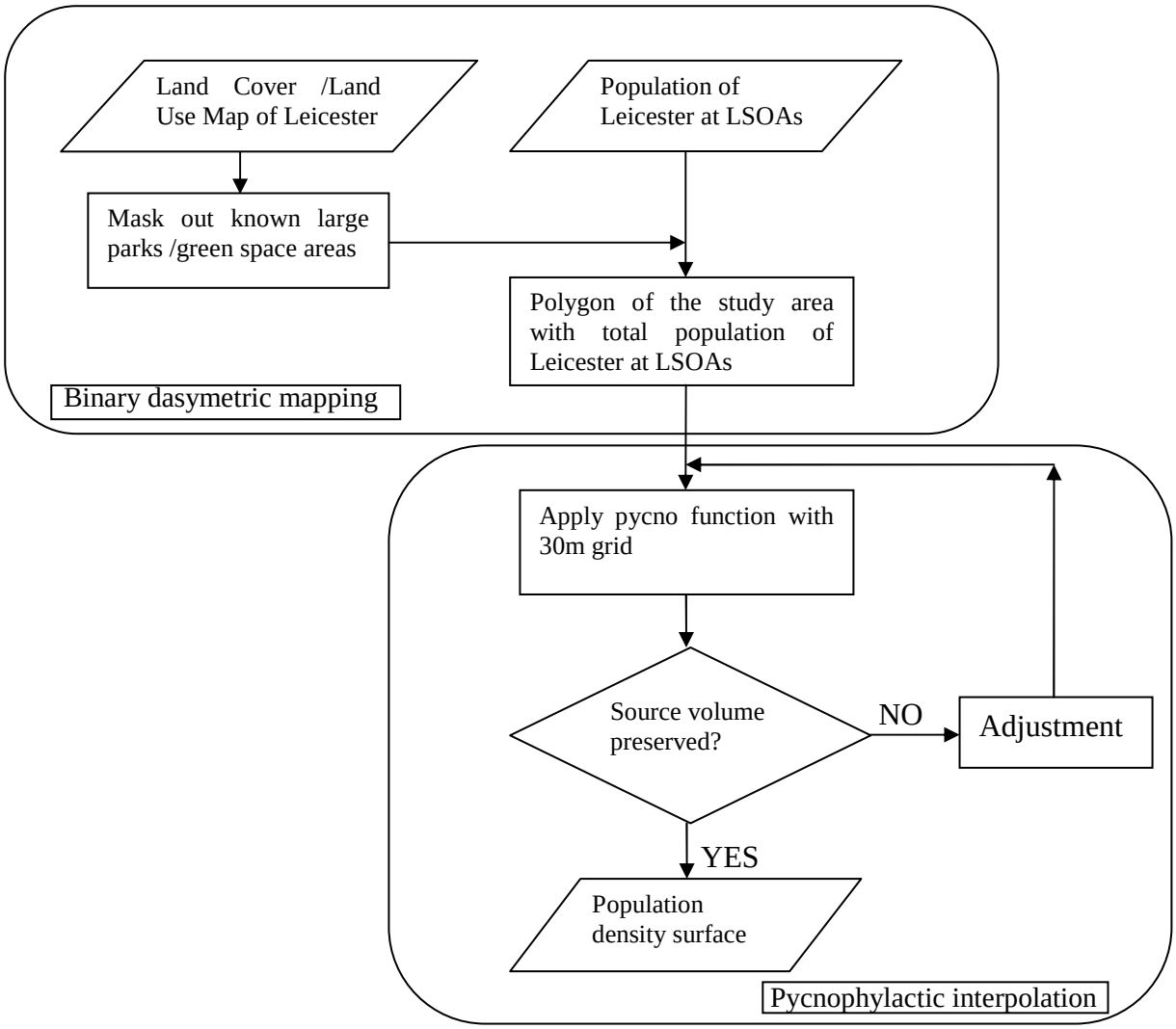

Figure 1. Flowchart combining dasymetric mapping with pycnophylactic interpolation

4. Results

The population of LSOAs as in Figure 2 shows LSOAs with low population in red and those with high population in yellow. The legend shows range of values in three (3) classes with colour changing from red to yellow as the population increases. The predicted density surface from a combination of dasymetric mapping with pycnophylactic interpolation as in Figure 3 shows the population density values as a continuous surface ranging from low population density (shown in red) to high population density (shown in yellow). The map of OAs with population density (Figure 4) shows OAs with low populations in red and those with high population in yellow. The residuals, the difference between predicted and actual population in each OA are shown in Figure 5. This is important in visualising errors spatially. The red colour indicates a negative residual while yellow colour indicates positive residual. Although the actual population of OAs at the city centre have high population, the technique predicts high population values in and around the city centre.

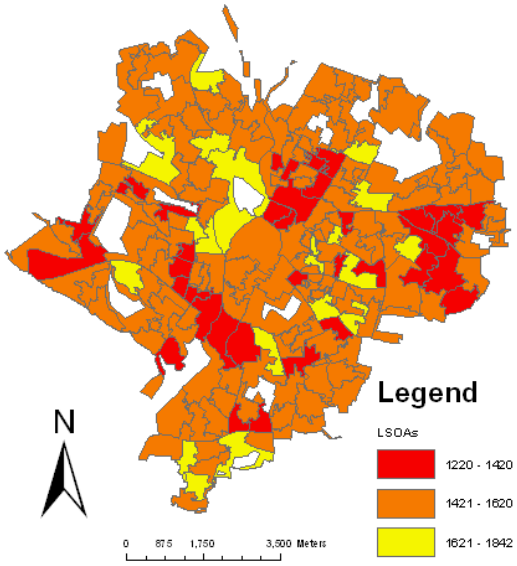

Figure 2 Map of LSOAs with population density

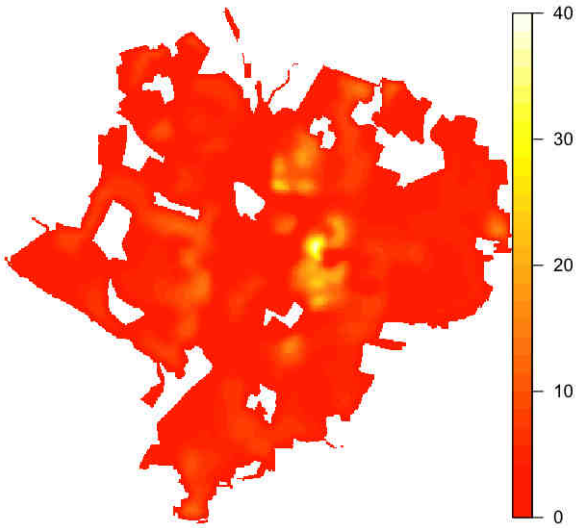

Figure 3. Predicted density map

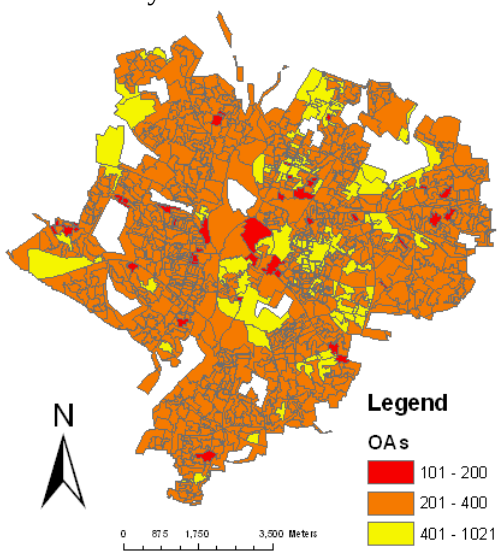

Figure 4. Map of OAs with population density

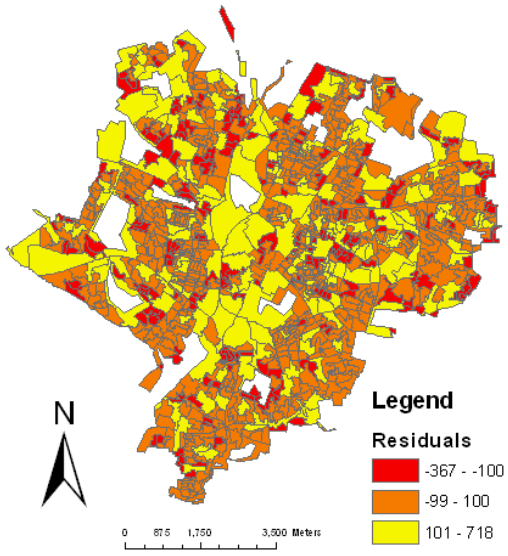

Figure 5. Map of residuals

The scatterplot of predicted densities versus OAs as in Figure 6 presents a general description of the data structure. It shows a strong clustering of points with relatively little scatter. This indicates positive relationship between predicted densities and actual population at OAs.

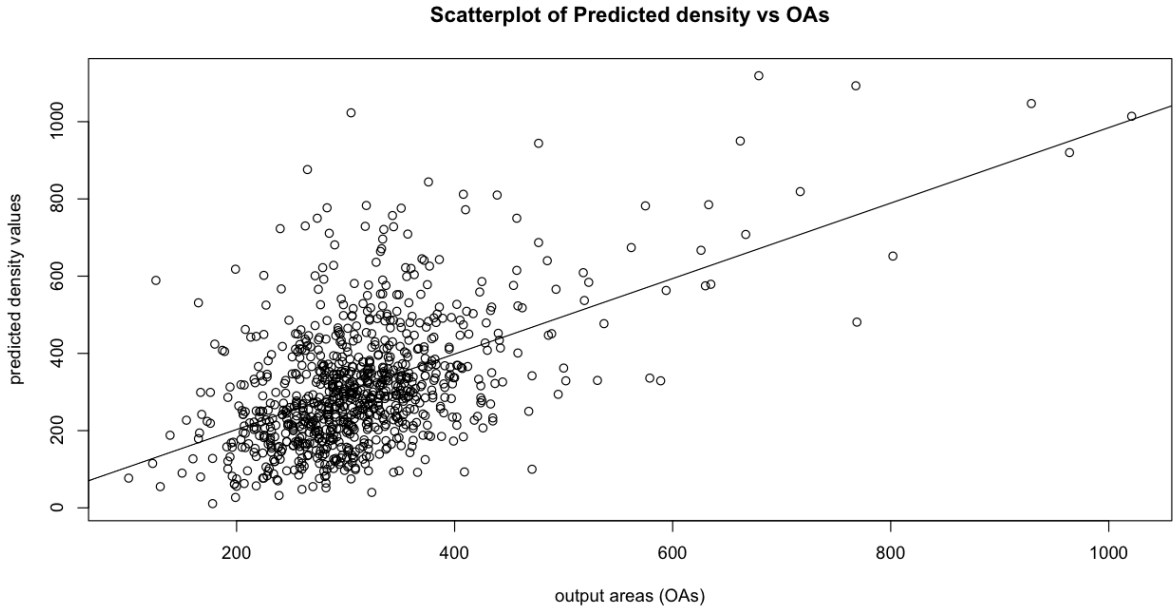

Figure 6. Scatterplot with regression line of Leicester population: Predicted density values vs. OAs

The histogram of residual values shows a visual presentation of value of residuals and their frequency of occurrence as in Figure 7. The spacing and number of bins were selected based on size and the distribution of the data, this could sometimes construct a misleading histogram (Simonoff, 1996).

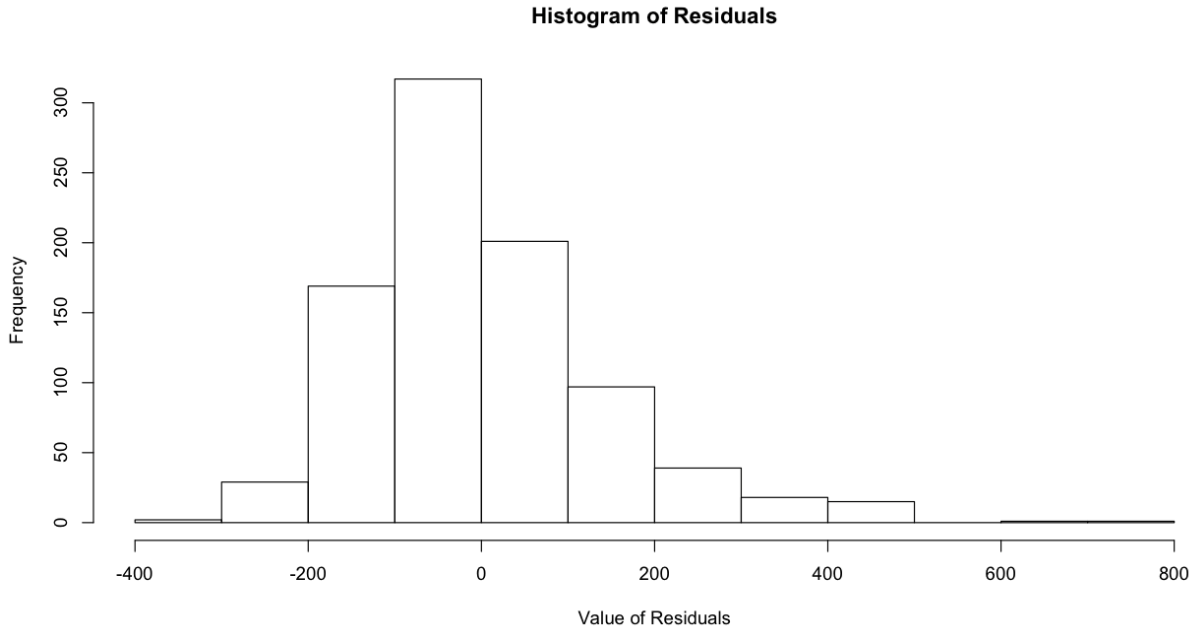

Figure 7. Histogram of residual values

A smoothed version of the histogram, the Kernel density estimates, as in Figure 8, shows a bell-shaped distribution for the residual values centred around zero which signifies positive relationship.

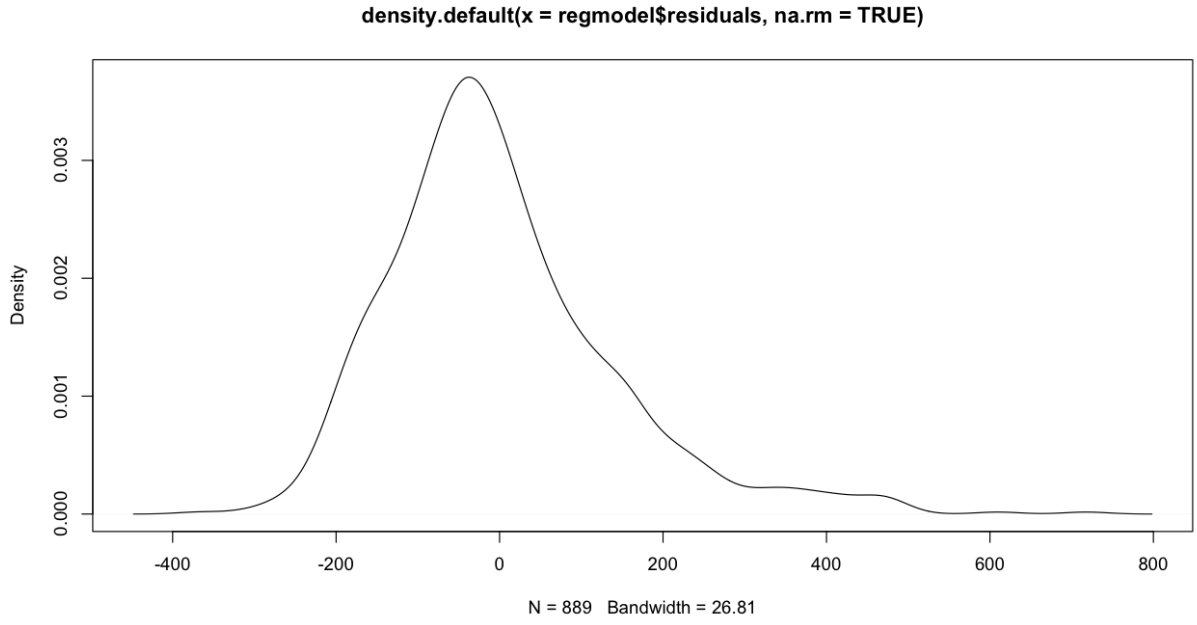

Figure 8. Kernel density of the residuals at default bandwidth

The residual values were sorted to check for serial correlation. This is useful as it shows the distribution of all data points and possible outliers. The sorted data was plotted against its index as in Figure 9.

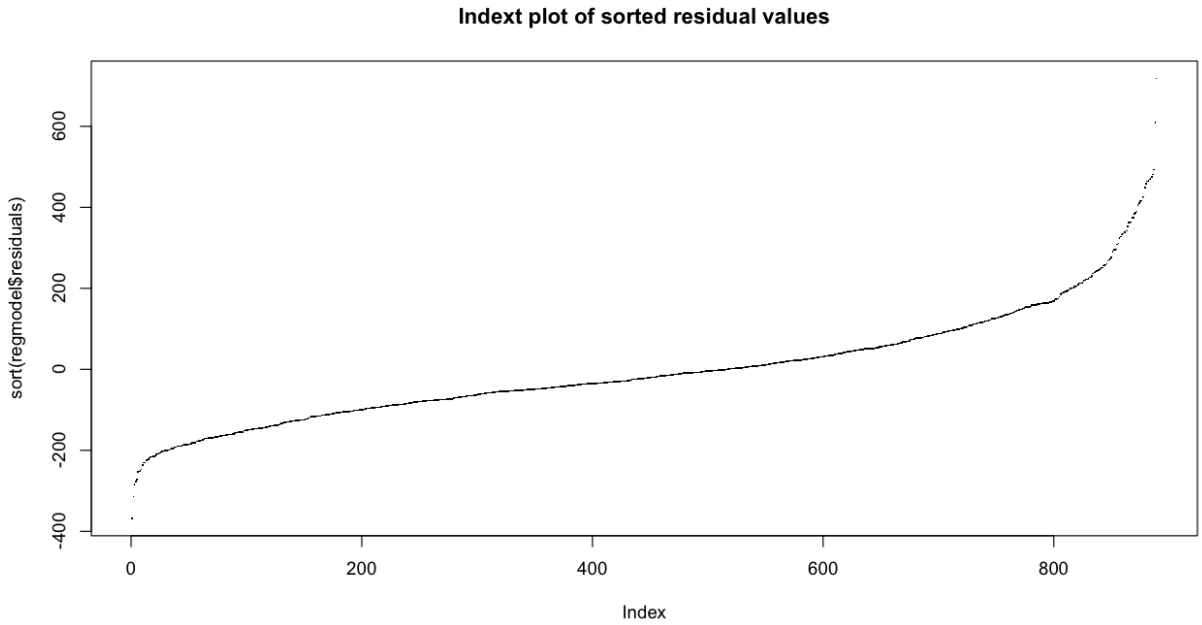

Figure 9. Index plot of sorted residual values

The results were measured for accuracy derived from Root Mean Square Error (RMSE) and adjusted Root Mean Square Error (Adj-RMSE) (Gregory, 2000). The RMSE uses absolute values of the difference between actual population and the predicted population within each of the target zones. A RMSE value of 1.442174 was obtained. The equation is represented as;

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2}$$

The Adj-RMSE is a measure of percentage error that normalises RMSE by the actual population within each target zone. Adj-RMSE value of 0.004607584 was obtained. The equation is represented as;

$$Adj - RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - y_i}{x_i}\right)^2}$$

Where;
x_i is the actual population of zone *i*
y_i is the predicted population of zone *i*
n is the number of zones

5 Conclusion

In this study, a combination of dasymetric mapping and pycnophylactic interpolation methods were used to estimate population of Leicester from LSOAs to OAs.

The results from this study has shown that the aggregated population censuses can be disaggregated to fine levels of spatial detail providing necessary information about the volume of demand in a geographic region. This must be obtained to adequately allocate demand to facilities within the region and to effectively evaluate accessibility to facilities. This can best be done when demand at output area level is analysed using origin and destination matrix.

Finally, an Adj-RMSE value of 0.004607584 means the technique has very little errors related to it. This shows the accuracy of the technique for population density estimation. This is important in addressing the problem of using larger estimates of the population in health care planning objectives and policy formulation, especially in areas where detailed census data (for example at the level of Output Areas) may not be available.

6 Acknowledgement

I would like to acknowledge The Petroleum Technology Development Fund (PTDF), Nigeria for funding my PhD

7 References

Brinegar S J and Popick S J (2010). A comparative analysis of small area population estimation methods, *Cartography and Geographic Information Science.*

Cai Q Rushton G Bhaduri B Bright E and Coleman P (2006). Estimating Small-Area Populations by Age and Sex Using Spatial Interpolation and Statistical Inference Methods, *Transactions in GIS*, 10(4), pp.577-598.

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

Cromley R G Ebenstein A Y and Hanink D M (2009). Estimating Components of Population Change from Census Data for Incongruent Spatial/Temporal Units and Attributes, *Spatial Science*, 54(2), pp.89-100.

Curtis S E and Taket A R (1989). The development of Geographical Information Systems for Locality Planning in Health Care, *Royal Geographical Society with Institute of British Geographers*, Area 21(4), pp.391-99.

Eicher C and Brewer C (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science*, 28, pp.125-38.

Gregory I N (2000). An evaluation of the accuracy of the areal interpolation of data for the analysis of long-term change in England and Wales. *Paper at Geocomputation 2000.*

Hawley K and Moellering H (2005). A comparative analysis of areal interpolation methods, *Cartography and Geographic Information Science*, 32(4), pp.411-23.

Hay S I Noor A M Nelson A and Tatem A J (2005). The accuracy of human population maps for public health application, *Tropical Medicine and International Health*, 10, pp.1073–1086.

Kim H and Yao X (2010). Pycnophylactic interpolation revisited: integration with the dasymetric-mapping method, *International Journal of Remote Sensing*, 31(21), pp.5657-71.

Kyriakidis P C (2004). A geostatistical framework for area-to-point spatial interpolation, *Geographical Analysis*, 36, pp.259-89.

Lam N S (1983). Spatial Interpolation Methods: A Review, *The American Cartographer*, l0 (2), pp.129-49.

Mennis J (2003). Generating surface models of population using dasymetric mapping, *The Professional Geographer*, 55, pp. 31–42.

Okabe A and Sadahiro Y (1997). Variation in Count Data Transferred from a Set of Irregular Zones to a Set of Regular Zones through the Point-in-Polygon Method, *Geographical Information Science*, 11(1), pp.93-106.

Openshaw S (1984). The modifiable areal unit problem, *Concepts and Techniques in Modern Geography*, 28, pp.38-41.

Rase W D (2001). Volume-preserving Interpolation of a Smooth Surface from Polygon Related Data, *Journal of Geographical Systems*, 3, pp.199 – 203.

Reibel M and Bufalino M E (2005). Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems, *Environment and Planning A*, 37, pp. 127–139.

Simonoff J (1996). *Smoothing methods in Statistics*. New York: Springer.

Tobler W (1979). Smooth Pycnophylactic Interpolation for Geographical Regions, *Journal of the American Statistical Association*, 74, pp.519-36.

World Health Organisation (1987), *Indicators for Primary Health Care*, World Health Organisation Regional Office for Europe, Copenhagen

Xie Y (1995). The Overlaid Network Algorithms for the Areal Interpolation Problem, *Computer, Environment and Urban Systems*, 19(4), pp.287 – 306.

8 Biographies

Jega Idris Mohammed is a second year PhD student in Geographical Information Science at the department of Geography, University of Leicester, Leicester. My research interest is on population estimation in small areas and spatial analysis of health facility distribution, policy and planning.

Alexis comber is a senior lecturer in Geographic information at department of Geography, University of Leicester, Leicester. His research interest includes Accessibility, equity of access and optimisation, uncertainty and representation in spatial data and the use of spatial analyses to evaluate policy.

Chris Brunsdon is Professor of Geographical information at the department of Geography, University of Liverpool, Liverpool. His research interest includes; Methodologies underlying spatial statistical analysis and geographical information systems and their applications in a number of subject areas; exploratory data analysis; data visualisation; house price modelling and crime pattern analysis.