Computers, Environment and Urban Systems 48 (2014) 64–72

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Surface models and the spatial structure of population variables: Exploring smoothing effects using Northern Ireland grid square data

Christopher D. Lloyd [a,*], Behnam Firoozi Nejad [b]

[a] Department of Geography and Planning, School of Environmental Sciences, University of Liverpool, Roxby Building, Chatham Street, Liverpool L69 7ZT, UK
[b] School of Geography, Archaeology and Palaeoecology, Queen's University Belfast, Belfast BT7 1NN, UK

ARTICLE INFO

Article history:
Received 9 October 2013
Received in revised form 30 June 2014
Accepted 2 July 2014
Available online 5 August 2014

Keywords:
Population surface modelling
Spatial variation
Census

ABSTRACT

Where areal units used to report population counts from Censuses and other sources are incompatible, direct comparison of counts is not possible. To enable such comparisons, a wide variety of areal interpolation and surface modelling approaches have been developed to reallocate counts from one zonal system to another or to a regular grid. The particular characteristics of individual variables, representing population sub-groups, mean that the most accurate results for each sub-group may be obtained using quite different approaches, or different model parameters. This paper seeks to assess how the degree of smoothing associated with population surface modelling relates to the accuracy of predictions made using two variables in Northern Ireland – the number of Catholics and persons with a limiting long term illness (LLTI). The study makes use of counts for 2001 released for output areas (OAs) and wards to generate population grids with 100 m square cells. The accuracy of the predictions is then systematically assessed using counts released for 100 m grid cells as an additional output from the 2001 Census. The results show that the amount of smoothing and the spatial structure of the variables are related to the prediction errors and this suggests that use of information on the spatial structure of variables is likely to provide benefits, in terms of accuracy of population reallocations, over common areal weighting approaches.

1. Introduction

The reallocation of counts from one set of zones to another (areal interpolation) is a common objective in Census research (e.g., Martin, Dorling, & Mitchell, 2002) and across the spatial sciences (Gotway & Young, 2002). Examples include accounting for boundary changes between Censuses, or transfer from higher to lower level geographies. Possible approaches include kernel smoothing, to create a population grid, regression based on land use data, and areal weighting (using overlay operators). Variables representing different population sub-groups may exhibit very different spatial patterns – for example, there may be more variation in employment or educational status across an area than there is with respect to car ownership. Thus, the approach used to create a population grid in each case should be adapted to account for the heterogeneity of a variable. Areal interpolation approaches are needed since any analysis based on area data, such as counts for wards provided as outputs from the UK Census of Population, is partly a function of the size and shape of those zones and the capacity to reallocate counts to different zonal systems may facilitate an enhanced analysis of the variable by removing the dependence on the zonal system. In addition, where zones change between Censuses, direct comparisons for different Census dates are not possible. Furthermore, taking the case of the UK, if the 2011 Census is to be the last (the Beyond 2011 Programme[1] considers possible future alternatives for the provision of national-level population statistics), then there will be an even greater need for flexible approaches to mapping populations using diverse data sources.

The focus in this paper is on the construction of population surfaces. Most current attempts to create population surfaces are based on (i) kernel smoothing type approaches used in isolation (see, for example, Martin, 1989) or (ii) areal reallocation informed by external datasets such as land use data. This study builds on previous work by Martin, Lloyd, and Shuttleworth (2011), which was based on data from the 2001 Census of Population of Northern Ireland and made use of the kernel smoothing method of Martin (1989 and 1996). In that study, total population counts were real-

* Corresponding author.
E-mail address: c.d.lloyd@liverpool.ac.uk (C.D. Lloyd).

[1] http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/index.html.

http://dx.doi.org/10.1016/j.compenvurbsys.2014.07.001
0198-9715/

located from irregular zones (output areas; OAs) to a 100 m cell grid and the accuracy of the predictions was assessed using counts on a 100 m grid which were an additional output from the Census (see Shuttleworth & Lloyd, 2009, for a summary of the Northern Ireland Census grid square resource). The present study applies data from the same sources, but instead uses counts for two population sub-groups, rather than the total population. The counts used are the number of Catholics by community background (as defined in Section 3) and persons with a limiting long term illness (LLTI). The two sets of counts were selected as previous research shows that they have (when expressed as percentages or transforms of percentages) distinct spatial structures (see Lloyd, 2010, and also see Lloyd, 2012 for an analysis of spatial scales of variation in religion and community background). In an analysis based on log-ratios[2] derived from 2001 Census data for Northern Ireland, Lloyd (2010) showed that community background (Catholics/non-Catholics) log-ratios were more strongly spatially structured (that is, spatially dependent) than a host of socioeconomic and demographic variables. The Moran's I spatial autocorrelation coefficient (using queen contiguity whereby similarity in values for adjacent zones is measured) was computed for log-ratios given three zonal systems: OAs, wards and 1 km grid cells. The values of I indicate how spatially dependent are individual variables, with large positive values indicating greater spatial dependence (i.e., neighbouring values tend to be more similar). For the community background log-ratio, I was 0.752 for 1 km cells and it was 0.826 for OAs. These were the largest values for all variables considered. For the LLTI/non-LLTI log-ratios, the values of I were 0.060 for 1 km cells and 0.436 for OAs. These values were the smallest for 1 km grid cells and the third smallest out of 14 variables for OAs. Thus, the Catholics/non-Catholics log-ratio was relatively homogenous over quite large areas while the LLTI/non-LLTI log-ratio varied more over small distances. Thus, the optimal approaches to population surface modelling in the case of counts of Catholics and persons with a LLTI might be expected to be quite different.

This study uses a population surface modelling approach based on (but not identical to) the method of Tobler (1979). In essence, the first stage of the method entails overlaying a grid on the input zones and assigning the zone counts to the overlapping cells. The counts for each zone are then divided by the number of grid nodes which fall within each zone. Thus, the population of the zone and the overlapping grid nodes are the same. The grid node counts are then smoothed using a filter of predetermined size; this has the effect of making neighbouring grid node counts in different zones more similar. The grid node counts are rescaled so that, again, the population of the zone and the overlapping grid nodes are the same. The optimal size of the filter window for a given population sub-group is likely to be related to the degree of spatial variation in that population sub-group, and this is the key issue explored in the paper. The previous discussion suggests that spatial autocorrelation analysis in *rates* (percentages, log-ratios, etc.) may provide a guide to the likely degree of spatial dependence in *counts*, and vice versa. That is, if a population sub-group is more spatially continuous then, all else being equal, smoothing over large areas is more likely to bring benefits as cells at the edges of two zones (such as OAs and wards, the two source zones used here) are more likely to be similar. Taking the example of Catholics in Northern Ireland, the strong spatial dependence in rates suggests that neighbouring zones (or at least the areas along their common edges) will often have similarly large populations of that group. In contrast, for persons with a LLTI, few such distinct areas of small and large counts are likely to exist and we might expect smoothing to be less beneficial. Indeed, variograms (see Lloyd, 2012 for an introduction to variogram estimation) estimated from counts of Catholics and persons by LLTI suggest that the former are much more spatially structured than the latter.[3] Thus, smoothing may be more likely to beneficial in the case of Catholic counts than in the case of counts of persons with a LLTI. Land use data and other ancillary data sources have been used to increase the accuracy of population reallocations between zonal systems. In this paper, counts of *all* persons per grid cell, as represented in the Northern Ireland grid square resource, are used as analogous to land use data such that population sub-group values are only reallocated to cells which are populated. In other words, the total counts for 100 m cells are used as a mask. In this case, the paper thus refers to estimates (or reallocations) constrained to populated cells.

This paper is the first to assess the performance of population surface generation methods using input zones at two spatial scales and two population sub-groups with 'true' gridded population counts used as a benchmark. There is relatively little existing research which assesses population surfaces using a 'true' population grid for comparison, little work on population sub-group surfaces and no systematic assessments of accuracy which, like the present paper, explore the errors and their relationship to other characteristics. The paper first considers some approaches to areal interpolation modelling, and population surface modelling specifically. Next, the data used in the analysis are detailed. The analysis assesses the accuracy of population surfaces for Catholics and persons with a LLTI and demonstrates that the optimal approach in the two cases differs if the source zone sizes differ.

2. Areal interpolation and population surface modelling

A common problem in regional analysis is that data are not in the spatial units (that is, zones such as Census tracts or wards) that the analyst requires. This may be even more problematic for researchers wanting to compare data such as national Censuses over time when boundaries of data collection or output areas are subject to change. The problem regularly arises where scientists want to compare a variable which is accessible for one set of zones with an additional variable only available for a different and incompatible zonal system (Flowerdew & Green, 1994). More generally, the results of any statistical analysis are a function of the size and shape of zones used to report values. The modifiable areal unit problem (MAUP) encapsulates the idea that zones are generally arbitrary and changes to zones will affect results of analyses (see Fotheringham, Brunsdon, & Charlton, 2000; Openshaw, 1984; Openshaw & Taylor, 1979; Wong, 2009 provides an overview and consider some possible avenues for research). As a solution to these problems, methods have been developed to reallocate counts from one set of zones to another or from irregular zones to regular grids so that data on different zonal systems can be compared and the dependence on a single zonal system removed. Methods to reallocate population counts can be divided into two groups (1) areal interpolation and (2) surface modelling (Yue et al., 2003), although the second is sometimes considered as a subset of the first. Goodchild and Lam (1980) and Langford, Maguire, and Unwin (1991) provide reviews of areal interpolation methods. A review of interpolation methods, including areal interpolation, is available from Lam (1983), while an overview is provided by Lloyd (2014).

Some common approaches to areal interpolation are summarised here. The variable of interest is given by z. Data on z are

[2] A transform of percentages which makes their analysis using standard statistical methods appropriate.

[3] This is indicated by the nugget:sill ratio (derived from models fitted to the variograms); the ratio for Catholics was smaller than that for LLTI, indicating more spatial structure in counts of Catholics than persons by LLTI (see Webster & Oliver, 2007 for more on variogram models and nugget:sill ratios).