Assessing the Accuracy of Population Sub-Group Surfaces

Behnam Firoozi Nejad[2] and Christopher D. Lloyd[1]

[1]Department of Geography and Planning, School of Environmental Sciences,
University of Liverpool, Roxby Building, Chatham Street, Liverpool, L69 7ZT, UK
Tel. (+44) 0151 794 2857; c.d.lloyd@liverpool.ac.uk;
http://www.liv.ac.uk/environmental-sciences/staff/christopher-lloyd/
[2]School of Geography, Archaeology and Palaeoecology, Queen's University Belfast,
Belfast, BT7 1NN, UK

1. Introduction

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 (see Martin 1996 and Lloyd 2014 for summaries). 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 study applies four different approaches. Each has as its first stage the overlap of grid cell centroids with source zones and proportional allocation of the source zone population to the overlapping cells. So, where 10 grid cell centres overlap a source zone then each cell is assigned 1/10 of the zone population. The four approaches comprise: (1) Simple reallocation to gridded cells within source zones as described previously, (2) Smoothing of outputs from 1, (3) an approach the same as 1, but where reallocation to cells within source zones is only made to cells which external data show to be populated (so, reallocations cannot be made to non-residential areas) and (4) Smoothing of outputs from 3.

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 100m square cells. The accuracy of the predictions is then assessed using the 100m grid counts released 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, over common areal weighting approaches.

2. Data and methods

2.1 Data

The analysis is based on two sets of counts from the 2001 Census of Northern Ireland. The source data from which counts are reallocated are from Tables KS007b (Community Background: Religion or Religion Brought Up in) and KS008 (Health and Provision of Unpaid Care) for wards (*n* = 582) and output areas (OAs, *n* = 5022). The accuracy of the estimates is assessed using counts of the same variables taken from the Northern Ireland Census grid square dataset[1]. These are counts on a 100m cell grid with values only for populated cells, released as an additional output from the Northern Ireland Census. Counts for grid squares have been released as outputs from all Northern Ireland Censuses since 1971 (see Shuttleworth and Lloyd, 2009, for more on the grid square resource). Note that only total persons and households were reported for cells with less than 25 persons or 8 households. Therefore, only cells exceeding those thresholds are included in this analysis. The lack of small counts means that comparison of estimates and grid cell counts is not ‘like with like’; but the grid-based counts do provide a representation of sub-group population structure and thus this comparison is considered appropriate. The total counts for 100m cells are also used as a mask whereby estimates are only made (that is, OA or ward counts are only reallocated) if a corresponding cell is populated. Obviously, many users will not have access to such data but they are used here as a proxy for land use data which indicate areas which are likely to be populated or not populated.

2.2 Methods

The basic surface modelling approach used in the present analysis is based on several steps. Firstly, a grid of points (cell centres) with a 100m spacing is overlaid on the source zones (OAs or wards). Secondly, the two are spatially joined so that each 100m grid point is assigned the population of the source zone it falls within. Thirdly, the number of 100m points within each zone is computed and the population assigned to each 100m point is divided by the number of 100m points within each zone. At the end of this process, the total counts assigned to the 100m points sum to the total population in the source zones.

As summarised in the introduction, the reallocation of counts from OAs or wards to 100m cells is based on this basic approach (number 1 below) and three adaptations of the approach:

1. Simple reallocation to gridded cells within source zones
2. Smoothing of outputs from 1
3. Reallocation constrained to populated cells within source zones
4. Smoothing of outputs from 3

After smoothing, the cell values are readjusted to ensure that the populations of the cells sum to the population value of the zone in which they fall. The rationale behind a smoothing approach is that neighbouring areas may be expected to be similar and thus adapting estimates at the common edges of neighbouring zones to make them more similar to one

[1] These data simply comprise counts of persons, as derived from the Census, allocated to square cells. See http://borders.edina.ac.uk/html/easy_download/NIGRID.html

another may be expected to increase the accuracy of estimates. The population counts are strongly spatially structured – large counts are often found next to other large counts while small counts are often found next to other small counts. Thus, a smoothing approach is intuitively sensible.

Results for each of the four approaches are examined and the accuracy of the estimates is assessed by computing the root mean squared error (RMSE) given the difference between the estimates and the population sub-group values as represented in the grid square count data described above. The smoothing approach used here is similar, although not identical, to the pycnophylactic interpolation approach of Tobler (1979). Specifically, once the counts have been distributed to cells within zones (with or without constraining to populated cells) they are smoothed using a square filter window. The application of window sizes from 3 by 3 to 15 by 15 cells is assessed. With the present approach, the smoothing is conducted in one step, whereas Tobler's approach is iterative with successive smoothing until a convergence criterion is met.

3. Analysis

Figure 1 shows estimates of the number of Catholics with smoothing using a 3 by 3 cell filter. Figure 2 shows, for the Belfast region, the estimates from Figure 1 minus the 'True' values. Detailed examination of the errors shows that the largest errors occur where there are large concentrations of people (specifically Catholics in this example) – a tower block would be an example. Thus, in the estimated grid of values, there would be likely to be too few people in the tower block location cell, but too many in surrounding cells over which the tower block residents have been 'spread'. While there do tend to be large over-estimates in locations with small populations and large under-estimates in places with large populations, population sizes and estimation errors are not strongly linearly related as mid-range errors may be associated with moderately small or moderately large populations. In other words, errors in estimates for places with moderately large populations are not biased. The RMSE values for Catholics are larger than those for LLTI as the counts of Catholics are larger than the counts of persons with a LLTI.

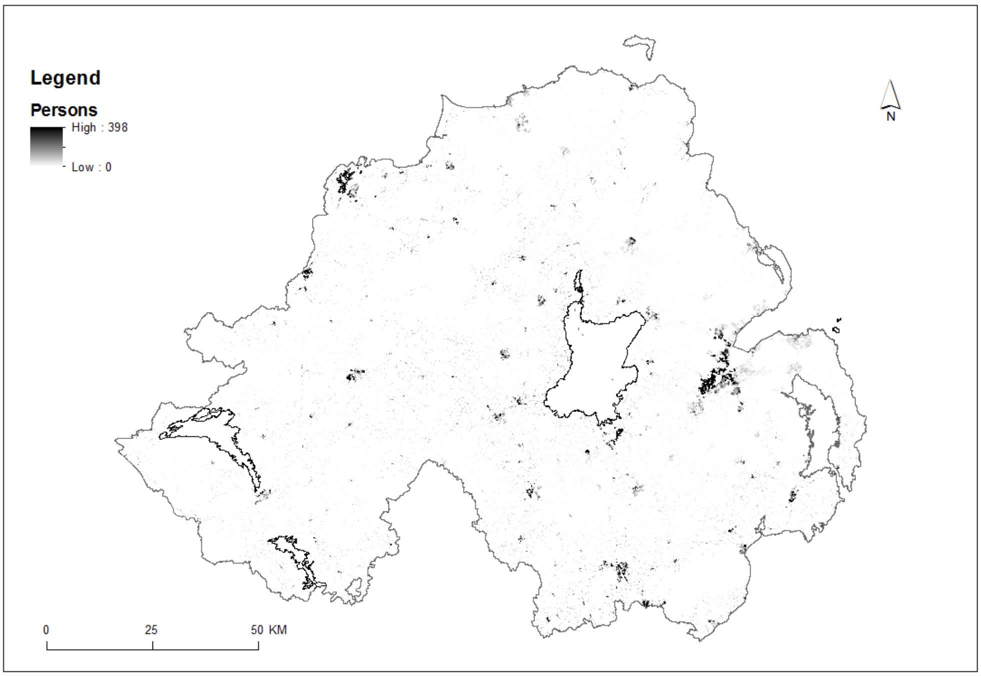

Figure 1. Estimates of Catholics using method 2 with 3 by 3 pixel smoothing

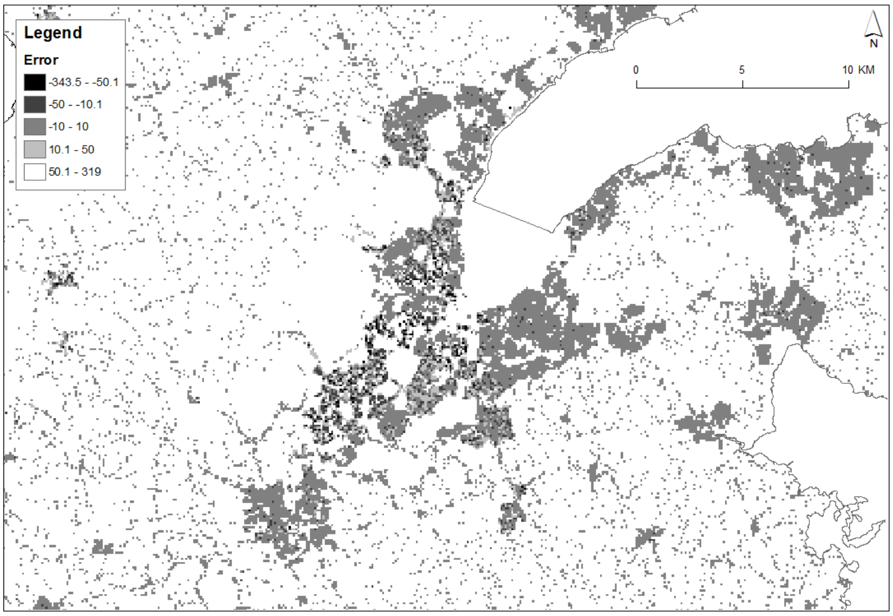

Figure 2. Errors of estimates (estimates minus 'True' values) of Catholics using method 2 with 3 by 3 pixel smoothing: Belfast region

Given that the focus in this paper is on how errors vary for different population sub-groups and using different approaches (smoothing or non-smoothing, estimates constrained to populated cells or estimates with no constraint), the errors for each approach and both sub-groups (Catholics and persons by LLTI) are summarised in Table 1 using the RMSE. Where estimates are not constrained to populated cells, for OAs the smallest RMSE for Catholics and LLTI is for a 3 by 3 window. In the case of no population constraint for wards, the smallest RMSE for Catholics is for a 9 by 9 and an 11 by 11 window while for LLTI it is for a 13 by 13 and a 15 by 15 window. For OAs as source zones, where estimates are constrained to populated cells, the smallest RMSE is for Catholics and LLTI it is for a 3 by 3 window. For wards, where estimates are constrained to populated cells, the smallest RMSE for Catholics and for LLTI is for a 3 by 3 cell window. Using populated 100m grid cells to constrain estimates clearly reduces the differences between the RMSE values for estimates derived from OA and ward-level counts for both Catholics and LLTI. In other words, there is less to gain by using smaller source zones when reallocations are constrained to populated cells when they are not.

Table 1. Catholics and LLTI: RMSE by zone (OA or ward) and for different degrees of smoothing (window size). Without reallocations constrained to populated cells and with reallocations constrained to populated cells (POP).

	OAs		Wards		OAs POP		Wards POP	
Window	Catholics	LLTI	Catholics	LLTI	Catholics	LLTI	Catholics	LLTI
0	3.85	1.50	4.41	1.71	3.46	1.38	3.95	1.53
3	3.71	1.45	4.36	1.69	3.29	1.34	3.67	1.44
5	3.73	1.46	4.35	1.69	3.39	1.37	3.75	1.47
7	3.76	1.47	4.35	1.69	3.44	1.38	3.81	1.49
9	3.78	1.48	4.34	1.69	3.46	1.39	3.85	1.50
11	3.80	1.48	4.34	1.68	3.46	1.39	3.88	1.50
13	3.81	1.48	4.34	1.68	3.46	1.39	3.90	1.51
15	3.81	1.49	4.35	1.68	3.47	1.39	3.91	1.51

4. Discussion and conclusions

These findings conform to expectation in that community background, when expressed as percentage of Catholics, is more clustered than the percentage of persons with a LLTI. Thus, more is gained (in terms of accuracy) by using immediate neighbouring values via smoothing in the community background case than in the LLTI case. In other words, RMSE values decrease more with smoothing for Catholics than for LLTI. Using gridded total population counts as analogous to land use data (that is, constraining estimates to populated cells) makes a greater difference than smoothing, but smoothing in isolation does make a difference and population and smoothing in combination produce the smallest errors. The findings also suggest that the amount of smoothing and the spatial variation in counts are related to the accuracy of derived population surfaces.

The choice of method for redistributing counts (standard areal weighting, smoothing), and the potential benefits of using ancillary data to inform this process, depends on the spatial structure of the population sub-set (in this study, Catholics or persons with a LLTI) and the availability of secondary data which provide useful information on the distribution of the population sub-set. The analysis suggests that use of total population data to constrain total counts offers greater benefits than smoothing for both Catholics and LLTI and for both OA and ward source zones. In isolation, smoothing has a bigger impact on estimates of Catholics than for LLTI and has a greater proportional impact for OAs than for wards. When used in combination, total population constraints *and* smoothing have a bigger impact for Catholics than for LLTI and more for wards than for OAs. In short, smoothing is likely to provide greater gains when a variable is more spatially continuous. Variograms estimated from the counts show that counts of Catholics are more strongly spatially structured than counts by LLTI. Using secondary data (here total population counts) to help reallocate sub-group counts from source zones is likely to have a greater impact where zones are larger and thus likely to be more heterogeneous. While these results are intuitively sensible, few studies have provided empirical tests of this nature which could be used to refine surface mapping approaches.

5. Acknowledgements

The grid square data were provided by the Census Office (part of the Northern Ireland Statistics and Research Agency – NISRA). This research was supported by the British Academy/Leverhulme Trust Small Research Grant award SG121849, and this funding is acknowledged gratefully.

References

LLOYD, C. D., 2014. *Exploring Spatial Scale in Geography* (Chichester: Wiley-Blackwell)

MARTIN, D., 1996. An assessment of surface and zonal models of population. *International Journal of Geographical Information Systems*, **10**, pp. 973–989

SHUTTLEWORTH, I. G. and LLOYD, C. D., 2009. Are Northern Ireland’s communities dividing? Evidence from geographically consistent Census of Population data, 1971–2001. *Environment and Planning A*, **41**, pp. 213–229

TOBLER, W. R., 1979. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, **74**, pp. 519–530

Biography

Chris Lloyd is a Senior Lecturer in Geography. His research focuses on spatial data analysis, and in particular on local spatial statistics and the exploration of spatial scale. He has a particular interest in the conceptualisation and measurement of residential segregation.

Behnam Firoozi Nejad is a researcher in quantitative human geography and GIS. He is interested in population surface modelling and the application of spatial analysis methods to Census data.