Christopher D. Lloyd and Behnam Firoozi Nejad (2014) Surface models and the spatial structure of population variables: exploring smoothing effects using Northern Ireland grid square data. *Computers, Environment and Urban Systems*, 48, 64–72.

doi: http://dx.doi.org/10.1016/j.compenvurbsys.2014.07.001

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

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) to generate population grids with 100m square cells. The accuracy of the predictions is then systematically assessed using counts released for 100m 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.

Keywords: Population surface modelling; Spatial variation; Census

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 et al., 2002) and across the spatial sciences (Gotway and 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 et

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

al (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, 1996). In that study, total population counts were reallocated from irregular zones (output areas; OAs) to a 100m cell grid and the accuracy of the predictions was assessed using counts on a 100m grid which were an additional output from the Census (see Shuttleworth and 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 1km 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 1km 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 1km cells and 0.436 for OAs. These values were the smallest for 1km 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

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

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 100m cells are thus 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 and 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 Openshaw and Taylor 1979, Openshaw 1984, Wong 2009; Fotheringham et al., 2000 provide 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 et al (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).

[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 and Oliver 2007 for more on variogram models and nugget:sill ratios).

Some common approaches to areal interpolation are summarised here. The variable of interest is given by z. Data on z are extracted from a set of source zones s; they are, however, required for a set of target zones t, where both t and s belong to the same geographical area. The known value of z for zone s is indicated by z_s and the unknown value of z for zone t is represented with z_t (Flowerdew and Green 1994). The s zone will often be divided by the t zone boundaries into several intersection zones. The intersections between zones s and t are indicated by st. Therefore, the problem of calculating the values for the t zones can be reduced to the problem of calculating the intersection zone values. When the variable of interest is extensive (it does depend on the volume of the system; population counts are an example), and with an even distribution of the variable within the source zone (as is the accepted assumption), estimates (that is, reallocations to target zones) are given by:

$$\hat{z}_t = \sum_{s=1}^{n} \frac{A_{st}}{A_s} z_s \tag{1}$$

Where A_s is the area of the source zone, and A_{st} is the area of the intersecting parts of s and t. When the variable of interest is intensive (it does not depend on the volume of the system; rates of some population group are an example), the term A_s is replaced with A_t. This kind of approach is referred to as areal weighting – each target zone is assigned a population which is proportional to the size of the overlapping portion of the source zone. For example, if (part of) the target zone overlaps with 25% of the source zone then it is assigned 25% of the source zone population and the target zone population is the sum of all such population proportions for all overlapping source zones. Gregory and Ell (2005) assess the application of standard areal weighting and adaptations, which make use of ancillary data, in the transfer of counts from historic censuses between different zonal systems. Two approaches which apply ancillary data in reallocation of data values are dasymetric mapping (following Wright 1936), and the use of control zones (Gregory and Ell 2005). With dasymetric mapping, 'limiting variables' are used which provide information on the spatial distribution of the variable of interest within the source zone. As an example, areas of water or exclusively industrial areas have no resident population (see, for example, Fisher and Langford 1995). In the case of control zones, information within a set of zones other than the source zones (target zones or another set of zones) are used to inform reallocation of data values (Goodchild et al. 1993, Langford et al. 1991). In the present study, as detailed below, ancillary information for a set of zones other than the source zones is used and thus this corresponds to a control zone approach.

Population surface modelling is targeted at creating a population map on a regular grid system, where each grid cell represents an estimate of the number of people for that individual cell (Mennis 2003; Yue et al. 2003). Grid based data offer some advantages. It is not difficult to re-aggregate regular grid data to any areal arrangement required. Population data on a gridded base could be more compatible with other heterogeneous datasets (Yue et al. 2003). Also, it will help to avoid some of issues encountered by artificial (for example, administrative) boundaries (Martin and Bracken, 1991). Martin (1989) introduced a method for generating population surface from zone centroids[4]. This approach redistributes population-weighted zone centroid data using a smoothing kernel to produce a continuous population surface model (see Lloyd 2011 for an introduction

[4] http://www.public.geog.soton.ac.uk/users/martindj/davehome/software.htm

and example). The Pycnophylactic interpolation approach of Tobler (1979) provides another means of generating gridded population values from counts on an irregular zonal system whereby the population of grid cells overlapping a zone is constrained to sum to the population of that zone – the approach is mass preserving. With these methods, there is an assumption of smooth variation across zones. Within a zone, there may be considerable variations in population sub-group density. Therefore, combining a smoothing approach with use of secondary data which describe with-zone heterogeneity is likely to be optimal (Yue et al. 2003). Secondary data may take many forms and could include land use data, postcodes or any other source containing features which may relate to population structure.

3. Data and methods

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; OAs nest within wards). The OAs were generated using an automated zone design approach; a measure of intra-area correlation was used to maximise social homogeneity within areas with the constraint that the total population and household numbers were above a predefined threshold and also close to the target size (Martin et al, 2001). The internal homogeneity of OAs makes them particularly suitable as a basis for population surface generation. The accuracy of the estimates is assessed using counts of the same variables taken from the Northern Ireland Census grid square dataset[5]. These are counts on a 100m cell grid with values only for populated cells (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 of population sub-groups. 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. Figure 1 shows the number of Catholics by 100m cells while Figure 2 shows the same counts, but for the Belfast region alone. As detailed below, the total population 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 (reallocations are constrained to populated cells). 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.

FIGURE 1 ABOUT HERE

FIGURE 2 ABOUT HERE

The basic surface modelling approach used in the present analysis is based on several steps. Firstly, a grid of points 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.

[5] http://cdu.mimas.ac.uk/2001/ni/grid/

The reallocation of counts from OAs or wards to 100m cells is based on this basic approach (the first case below) and three adaptations of the approach:

- Simple reallocation to cells within source zones – thus, if 10 cell centres overlap a source zone then each cell is assigned 1/10 of the zone population (case 1)
- Smoothing of outputs from case 1 (case 2)
- Reallocations constrained to populated cells within source zones (sub-group population is reallocated only to cells which are populated; case 3)
- Smoothing of outputs from case 3 (case 4)

Results for each of the four approaches are assessed 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 errors are also explored using linear regression. 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.

4. Results and discussion

Figure 3 shows estimates of the number of Catholics with smoothing using a 3 by 3 cell filter. Figure 4 shows, for the Belfast region, the estimates from Figure 3 minus the 'True' values. It is possible to discern some patterns from visual comparison of Figures 2 ('true' counts of Catholics for the Belfast region) and 4 (errors for the Belfast region). For example, there are few large errors in the east of Belfast. But, this is not particularly informative given the relatively small number of Catholics in east Belfast. 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'. There are many cases of zones which contain only one or two cells which the grid square resource shows to be populated and thus most estimates (that is, for cells which 'should' have a population of zero) are small over-estimates. This issue is explored further below. The errors are strongly spatially dependent; variograms estimated from all cells with non-zero errors for Catholic and LLTI counts show clear spatial structure, and this is expected given the likely patterns of under- and over-estimation suggested above. In other words, negative errors are likely to be found close to other negative errors, and similarly positive errors will tend to cluster with other positive errors.

FIGURE 3 ABOUT HERE

FIGURE 4 ABOUT HERE

4.1 Regression analysis of prediction errors

The characteristics of large negative and positive errors can be explored using regression. Cockings et al (1997) explore areal interpolation (areal weighting and a dasymetric approach) errors using sets of randomly generated source and target zones (aggregations of zones called enumeration districts; EDs). The paper assesses the relationships between the mean areal interpolation errors for target zones and parameters including target zone area, perimeter, compactness ratio (relating a polygon's area to its perimeter), total population, population density and the range of population densities within each zone (given the EDs within the zones). For most of the models considered, coefficients for all of the parameters were shown to be significant. Backward stepwise multiple regression was used to identify the most significant parameters in the model. In the case of areal weighting, perimeter and (to a lesser degree) total population were shown to be most significant, while for the dasymetric model trends were less clear with results differing according to the number of target zones (50 and 100 were used). In the present study, the focus is instead on population surface generation and 'true' gridded counts are used to assess performance of the approaches used. Here, errors are the dependent variable while the independent variables are (i) population count, (ii) the difference between a cell and the average of its immediate neighbours (DiffMean; cell(*i,j*) – mean of neighbouring cell values) and (iii) the difference between the cell value and the mean population for cells in the overlapping zone (zone population / number of cells in the zone) (DiffZ; cell(*i,j*) – zone population/number of cells in zone). This allows for the exploration of the relationship between errors and population size (i above), locally 'extreme' values (i.e., cells with much smaller or larger populations than neighbouring cells) (ii above) and cells which are 'extreme' relative to the overlapping zone (iii above). Table 1 shows the OLS regression model coefficients for errors against these three independent variables. Note that there is no evidence of multicollinearity for any model, judging by variance inflation factors and condition indexes (see Belsley et al., 1980). The residuals from the models were only moderately spatially autocorrelated, as judged by the variograms estimated from them. Spatially lagged dependent variable and spatial error models were also run (using a neighbourhood of 1000m) with the same variables[6] using the GeoDa™ software (Anselin et al. 2006) and the model coefficients were very similar to those obtained using standard OLS regression. The Population and DiffMean coefficients were, as for the OLS model, close to zero and the DiffZ coefficient values were similar, as were the R^2 values. Therefore, the OLS results were considered robust. The coefficients for DiffZ suggest that if the (true) population of a cell is larger than the mean of cells in the same zone then under-estimation of the cell population is likely; in contrast over-estimation is likely if a cell's population is smaller than the mean of cell populations in the same zone. For both Catholics and counts by LLTI, in the case where populated cells are *not* used to constrain reallocations, the model (no. 3) including *only* DiffZ explains as much variation as a model with *all three* independent variables. In short, Population and DiffMean do not explain variation in errors while DiffZ explains 85% (Catholics) and 86% (LLTI) of the variation in errors. 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. Where populated cells are used to constrain reallocations, both Population and DiffMean explain more of the variation in errors, although less than DiffZ. In this case, the

[6] Note that the spatial lag models were run on a random sample of the prediction locations comprising 5% of the original grid cells, as it proved problematic to fit models to the full dataset (1,357,708 points) using the GeoDa™ software. Testing indicated that the sample was representative and OLS results for the full data set and the 5% sample were very similar.

DiffMean coefficients suggest that where a cell's population is larger than the (mean of the) populations of the neighbouring cells, over-estimation is the likely outcome. Thus, the local configuration of cells has a bigger impact on errors where populated cells are used to constrain reallocations than where they are not.

TABLE 1 ABOUT HERE

4.2 Summaries of prediction errors

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 not constrained to populated cells), the errors for each approach and both sub-groups (Catholics and persons by LLTI) are summarised in Table 2 using the RMSE. 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. Where estimates are not constrained to populated cells, for OAs as source zones the smallest RMSE for Catholics and LLTI is for a 3 by 3 window. In the case where estimates are not constrained to populated cells, and with wards as source zones, 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. Where estimates are constrained to populated cells, for OAs as source zones the smallest RMSE for Catholics and LLTI is for a 3 by 3 window. With estimates constrained to populated cells with wards as source zones, the smallest RMSE for Catholics and for LLTI is for a 3 by 3 cell window. Using populated cells to constrain estimates clearly reduces the difference between the RMSE values for estimates derived from OA and ward-level counts for both Catholics and LLTI. In the case of reallocations constrained to populated cells, the gains are proportionately greater for wards as source zones than for OAs. As suggested by Table 1, for ward source zones (with estimates not constrained to populated cells), smoothing does not make a great deal of difference and the RMSE values are quite similar for all window sizes considered. However, for OA source zones, smoothing makes quite a large proportional difference for both Catholics (the RMSE for 3 by 3 cells is 3.6% smaller than for the model with no smoothing) and LLTI (2.9%).

TABLE 2 ABOUT HERE

Another useful summary is the ratio of RMSE by ward to RMSE by OA values for counts of Catholics and persons with a LLTI with no smoothing. In the case of reallocations not constrained by total population the ratios are 1.147 (Catholics) and 1.143 (LLTI). With reallocations constrained to total population the ratios are 1.143 (Catholics) and 1.109 (LLTI). The ratios are, therefore, smaller where estimates are constrained to populated cells. Thus, using total population by grid cells to constrain estimates reduces the differences between OA- and ward-derived surfaces with no smoothing (that is, counts in all cells within a zone) relative to the case where estimates are not constrained to populated cells. In other words, if populated cells are used to constrain reallocations, OA and ward results are more similar then when populated cells are not used in this way. This makes intuitive sense — constraining reallocations to populated cells reduces the benefits of geographically fine-grained OA data relative to ward data as the population data help to represent fine-scale variation in population sub-groups. In other words, using populated cells to constrain estimates means that those estimates are likely to more closely mirror the underlying population structure than when estimates are not constrained to populated cells. Estimates made using different source zones are likely to be more similar in the former case than in the latter.

Table 3 shows the values in Table 1 expressed as ratios (RMSE for reallocations constrained to population cells/RMSE for reallocations not constrained to populated cells). In this table, small ratios indicate that the use of population data has a larger impact on the RMSE, while larger values indicate that population has a smaller impact. Thus, the increase in accuracy associated with constraining estimates to populated cells is proportionally greatest for 3 by 3 pixels for OAs and wards for both counts of Catholics and persons with a LLTI. As suggested previously, the proportional gain is greater for wards than for OAs and supports the assertion that use of secondary data will tend to result in larger gains where the source zones used are larger.

TABLE 3 ABOUT HERE

Another way to summarise the values in Table 2 is to express them as a percentage of RMSE for reallocations not constrained to populated cells and no smoothing – so, the RMSEs are expressed as a percentage of the RMSEs for the most basic approach to reallocation. As an example: the RMSE for LLTI OAs, with estimates constrained to populated cells with window size 3 (1.336) / RMSE estimates not constrained to populated cells and no smoothing (1.495) = 0.894, thus 89.4%. Computing ratios in this way indicates that the largest increase in accuracy over a standard overlay approach (that is, estimates not constrained to populated cells and no smoothing) is gained by using total population data to constrain reallocation of counts rather than by using smoothing, but that constraining estimates to populated cells and smoothing *in combination* result in the greatest increase in accuracy. To clarify, constraining estimates to populated cells and no smoothing results in a RMSE which is more than 10% smaller than for the base model (no total population constraint and no smoothing) for both OAs and wards for Catholics, and for wards for LLTI; for OAs for LLTI the reduction is 7.6%. In the case of constraining estimates to populated cells and smoothing together, the maximum decreases in RMSE with respect to the base model are (in all cases for a window size of 3 by 3 cells) 14.6% (OAs, Catholics), 10.6% (OAs, LLTI), 16.9% (wards, Catholics) and 15.6% (wards, LLTI). Thus, there are potentially considerable gains if secondary data are used in conjunction with smoothing. The potential gain is greater for wards than for OAs, for the reasons elucidated previously.

To deconstruct these findings further, the example of estimation of Catholics using OAs as source zones is considered. For smoothing using a 3 by 3 cell window, the range of errors where estimates are not constrained to populated cells is -456.0 to 319 while with the populated cells constraint it is -343.5 to 319. Thus, there is no change in the maximum over-estimate, but the maximum under-estimate decreases by 113. In terms of large under-estimations, in the case of no populated cell constraint, this includes cases where several cell centroids fall within a source zone, but only one is populated in reality (or at least that is what the grid square dataset indicates) – thus the populations are too spread out if the total population constraints are not used. In some cases, there are large under-estimates both without populated cell constraint and with populated cell constraint approaches – one example is the case of an OA which contains ten 100m grid cell centroids, but most of the population is concentrated in one discrete area – thus, again, the population is more spread out than it should be. The case of the maximum over-estimate corresponds to a case where only one cell centroid falls within an OA, and so the entire OA population is assigned to that cell. In that case, the source OA is elongated with a majority of its area covered by 100m cells whose centroid falls (in some cases only just) into other OAs. Thus, no variant of the approach employed here could reduce the error in this case. In general, the cell size should be much smaller than the areal unit and where there are few cells within a zone the population may be wrongly forced to one or two cells. The cell size of 100m is used here to enable accuracy assessment with 'true' 100m cell data; one possible improvement is to initially reallocate to cells smaller than 100m square (e.g., 25 m cells) and then aggregate upwards (see Martin et al. 2011 for an example of this kind of approach). In this case, the mean error with no total population constraint is 0.18 while with the populated cell constraint it is 2.6. Thus, the tendency to over-estimate increases as might be expected if the estimates are constrained to populated cells (in some cases, the estimates will be falsely pushed upwards). But, in addition, large negative estimates are reduced when the populated cell constraint is used.

4.3 Spatial variation in counts and prediction errors

These findings conform to expectation in that community background, when expressed as percentages of Catholics (or, for example, as log-ratios), is more spatially continuous than the percentages (or log-ratios derived from these) of persons with a LLTI. The same is true of counts of persons by community background and LLTI, with the former more spatially structured than the latter. 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 as neighbouring count of Catholics are more likely to be similar than neighbouring counts of persons by LLTI. Using gridded total population counts as analogous to land use data makes a greater difference than smoothing, but smoothing in isolation does make a difference and populated cell constraints and smoothing in combination produce the smallest errors. The findings also suggest that the amount of smoothing and the spatial variation in rates derived from the counts are related to the accuracy of derived population surfaces.

Clearly, if populations in source zones are homogenous then smoothing will not increase the accuracy of estimates. Conversely, in locations where there are abrupt within-zone changes in population density (e.g., there is a tower block), without external information on the location and nature of housing or on population counts, all approaches are likely to result in large errors. Such errors may be minimised when estimates are constrained to populated cells. In this study, the total counts by 100m grid cells are used to constrain estimates – estimates are only made to cells which the grid square data indicate are populated. Where total counts in this form are generally available obviously they could be used to proportionately assign counts of sub-groups to cells. But, here the total count data are used as analogous to land use data to make the approach generalisable to cases where perhaps land use data can be used to determine areas with or without populations, but only in a binary sense (that is, cells are populated or they are not). Differences in results obtained with no populated cell constraint and with a populated cell constraint are a function of internal variability of zones. In the extreme case, the entire population of a zone may be concentrated into a small area in a tower block and using populated cells to constrain estimates (or land use data, for example) is likely to make a much larger difference than when the population is evenly spread across a zone. Differences in results obtained with no smoothing or with smoothing are also partly a function of internal variability but also the degree to which neighbouring zones are similar.

The results from the present analysis support the key findings of Kim and Yao (2010). The latter study was based on the Atlanta metropolitan statistical area and showed that the smallest RMSE without the use of ancillary information was for the pycnophylactic approach (the RMSE for standard areal weighting was larger). The smallest errors of any approach were for a combined approach which made use of ancillary information and smoothing. Kim and Yao (2010) showed that there was a greater proportional decrease in RMSE with increase in smoothing window size for the combined (hybrid) approach than for the simple pycnophylactic approach. In the present study, it was shown that use of smoothing (but only over a small window), as opposed to no smoothing, reduces the RMSE proportionately more in the case when ancillary data are used than when they are not. The present study extends the findings of Kim and Yao to another geographical context but also assesses alternative source zones and considers two separate population sub-groups. The analysis shows that increases in estimation accuracy associated with smoothing are proportionately greater in the community background case than in the LLTI case. The optimal filter size (corresponding to the smallest RMSE) is 3 by 3 cells in most cases (only results for wards as source zones with no populated cell constraint differ from this), but smoothing brings greater gains for the more spatially structured Catholic counts than for the more spatially variable LLTI counts. Thus, while the optimal approach may be the same for different population sub-groups, the impact on accuracy is likely to vary between population sub-groups. More work is needed on the importance of choice of source zones, ancillary data, and smoothing given particular population sub-groups in different countries and regions, where the nature of available data may be highly variable.

5. Conclusions and future work

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 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 impact when OAs are source zones than when wards are the source zones. When used in combination, populated cell 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. Catholic percentages (log-ratios) are more spatially continuous than LLTI percentages (log-ratios). Given that the population base is the same in both cases, counts of Catholics are likely to be more continuous than LLTI counts. Using secondary data (here total population counts for grid cells) 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.

An obvious addition to this analysis would be the use of geostatistical approaches to population surface modelling. The present analysis provides some context to future work which will fully assess when it is useful to make use of information on the spatial structure of individual variables through variogram estimation and modelling, and deconvolution to estimate the point support variogram followed by area-to-point kriging (see Goovaerts 2008). In cases where rich (in terms of geographical and attribute detail) population data are available in conjunction with spatially-detailed secondary data sources, simple areal reallocation approaches may be sufficient. However, in more sparsely populated rural areas in particular, it may be that the spatial correlation structure of variables has a more important role in terms of the accuracy of population counts reallocated to grid cells. This research will be extended methodologically but also in terms of time period and geography. The present paper has used data for Northern Ireland to assess differences in approaches, and, after further testing, the analysis will be extended to the rest of the UK.

Acknowledgments

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

References

Anselin, L., Syabri, I. and Kho, Y. (2006) GeoDa: an introduction to spatial data analysis. *Geographical Analysis*, 38, 5–22.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Cockings, S., Fisher, P. F. and Langford, M. (1997) Parameterization and visualization of the errors in areal interpolation. *Geographical Analysis*, *29*, 314–28.

Fisher, P. F. and Langford, M. (1995) Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A*, *27*, 211–224.

Flowerdew, R. and Green, M. (1994) Areal interpolation and types of data. In S. Fotheringham and P. Rogerson (Eds.), *Spatial analysis and GIS* (pp. 121–45). London: Taylor and Francis.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2000) *Quantitative Geography: Perspectives on Spatial Data Analysis*. London: SAGE Publications.

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

Goodchild, M. F., Anselin, L. and Deichmann, U. (1993) A framework for the areal interpolation of socioeconomic data. *Environment and Planning A*, *25*, 383–397.

Gotway, C. A. and Young, L. J. (2002) Combining incompatible spatial data. *Journal of the American Statistical Association*, *97*, 632–48.

Goovaerts, P. (2008) Kriging and semivariogram deconvolution in the presence of irregular geographical units. *Mathematical Geosciences*, *40*, 101–28.

Gregory, I. and Ell, P. (2005) Breaking the boundaries: geographical approaches to integrating 200 years of the census. *Journal of the Royal Statistical Society, Series A*, *168*, 419–437.

Kim, H. and Yao, X. (2010) Pycnophylactic interpolation revisited: integration with the dasymetric-mapping method. *International Journal of Remote Sensing*, *31*, 5657–5671.

Lam, N.S-N. (1983) Spatial interpolation methods: a review. *American Cartographer, 10***,** 129–149.

Langford, M., Maguire, D. J. and Unwin, D. J. (1991) The areal interpolation problem: estimating population using remote sensing in a GIS framework. In I. Masser and M. Blakemore (Eds.) *Handling Geographical Information: Methodology and Potential Applications* (pp. 55–77). Harlow: Longman Scientific and Technical.

Lloyd, C. D. (2010) Exploring population spatial concentrations in Northern Ireland by community background and other characteristics: an application of geographically weighted spatial statistics. *International Journal of Geographical Information Science, 24,* 1193–1221.

Lloyd, C. D. (2011) *Local Models for Spatial Analysis*. Second Edition. Boca Raton: CRC Press.

Lloyd, C. D. (2012) Analysing the spatial scale of population concentrations by religion in Northern Ireland using global and local variograms. *International Journal of Geographical Information Science, 26,* 57–73.

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

Martin, D. (1989) Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers, New Series, 14*, 90 – 97

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

Martin, D. and Bracken, I. (1991). Techniques for modelling population-related raster databases. Environment and Planning, A 23, 1069–1075.

Martin, D., Dorling, D. and Mitchell, R. (2002) Linking censuses through time: problems and solutions. *Area*, *34*, 82–91.

Martin, D., Lloyd, C. and Shuttleworth, I. (2011) Evaluation of gridded population models using 2001 Northern Ireland Census data. *Environment and Planning A*, *43*, pp.1965–1980.

Martin, D., Nolan, A. and Tranmer, M. (2001) The application of zone-design methodology in the 2001 UK Census. *Environment and Planning A*, *33*, 1949–62

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

Openshaw, S. (1984) *The Modifiable Areal Unit Problem*. Concepts and Techniques in Modern Geography 38. Norwich: GeoBooks.

Openshaw, S. and Taylor, P. J. (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In N. Wrigley (Ed.) *Statistical Applications in the Spatial Sciences*. London: Pion, pp. 127–144.

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*, 213–229.

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

Webster, R. and Oliver, M. A. (2007) *Geostatistics for Environmental Scientists*. Second Edition. Chichester: John Wiley and Sons.

Wong, D. (2009) The modifiable areal unit problem (MAUP). In A. S. Fotheringham and and P. A. Rogerson (Eds.) *The SAGE Handbook of Spatial Analysis*. London: SAGE Publications, pp. 105–123.

Wright, J. K. (1936) A method of mapping densities of population: with Cape Cod as an example. *Geographical Review*, *26*, 103–110.

Yue, T. X., Wang, Y. A., Chen, S. P., Liu, J. Y., Qiu, D. S., Deng, X. Z., Liu, M. L. and Tian, Y. Z. (2003) Numerical simulation of population distribution in China. *Population and Environment*, *25*, 141–163.

Table 1. OLS regression model summary for errors (three-by-three cell smoothing) against Population, the difference between the cell value and the mean population for cells in the overlapping zone (zone population / number of cells in the zone) (DiffZ) and the difference between a cell and the average of its immediate neighbours (DiffMean). Without reallocations constrained by population (no Pop.) and with reallocations constrained (Pop.). All coefficients are significant to the 0.001 level.

Variable		No Pop.			Pop.		
		Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Catholics	Constant	0.264	0.265	0.272	0.138	0.160	0.249
	Population	0.004			0.102		
	DiffZ	-0.947	-0.942	-0.954	-0.688	-0.574	-0.720
	DiffMean	-0.022	-0.023		-0.263	-0.279	
	R squared	0.847	0.847	0.847	0.738	0.727	0.615
LLTI	Constant	0.108	0.108	0.112	0.059	0.070	0.106
	Population	0.001			0.098		
	DiffZ	-0.945	-0.943	-0.961	-0.736	-0.630	-0.793
	DiffMean	-0.028	-0.028		-0.238	-0.253	
	R squared	0.860	0.860	0.859	0.781	0.771	0.689

Table 2. Catholics and LLTI: RMSE by zone (OA or ward) and for different degrees of smoothing. Without reallocations constrained to populated cells and with reallocations constrained to populated cells (POP). The smallest values in each column are given in bold.

	OAs		Wards		OAs POP		Wards POP	
Window	Catholics	LLTI	Catholics	LLTI	Catholics	LLTI	Catholics	LLTI
0	3.846	1.495	4.411	1.709	3.456	1.382	3.949	1.532
3	**3.708**	**1.451**	4.363	1.693	**3.285**	**1.336**	**3.666**	**1.443**
5	3.729	1.459	4.351	1.688	3.391	1.367	3.752	1.467
7	3.758	1.468	4.345	1.686	3.439	1.380	3.814	1.485
9	3.780	1.476	**4.342**	1.685	3.457	1.386	3.850	1.496
11	3.795	1.481	**4.342**	1.684	3.462	1.387	3.876	1.504
13	3.807	1.484	4.344	**1.683**	3.464	1.388	3.896	1.509
15	3.814	1.486	4.346	**1.683**	3.467	1.388	3.908	1.514

Table 3. Values in Table 2 expressed as ratios (RMSE for reallocations constrained to populated cells/RMSE for reallocations not constrained to populated cells). Cath is Catholic.

Window	OA Cath Pop/Non	OA LLTI Pop/Non	Ward Cath Pop/Non	Ward LLTI Pop/Non
0	0.899	0.924	0.895	0.896
3	0.886	0.921	0.840	0.852
5	0.909	0.937	0.862	0.869
7	0.915	0.940	0.878	0.881
9	0.915	0.939	0.887	0.888
11	0.912	0.937	0.893	0.893
13	0.910	0.935	0.897	0.897
15	0.909	0.934	0.899	0.900

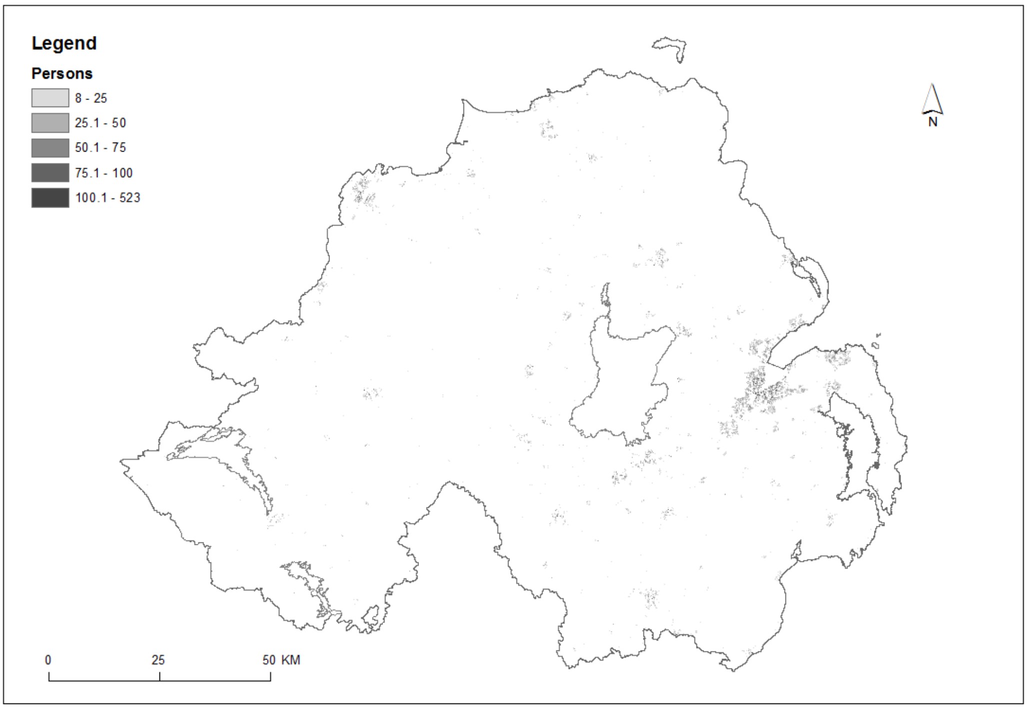

Figure 1. Number of Catholics by 100m cells

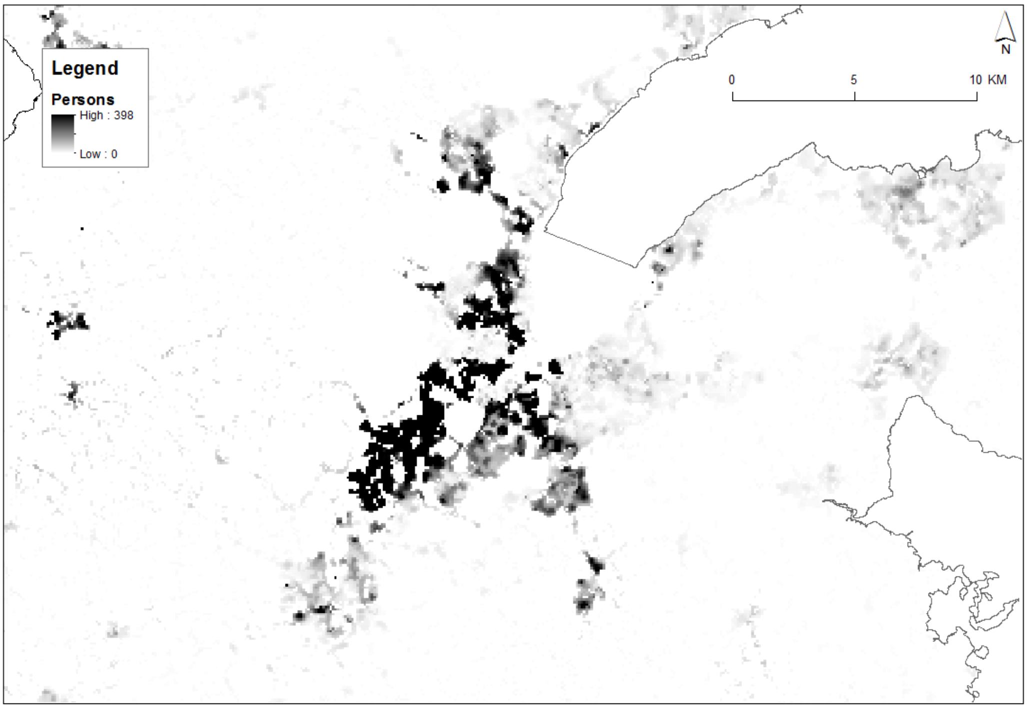

Figure 2. Number of Catholics by 100m cells: Belfast region

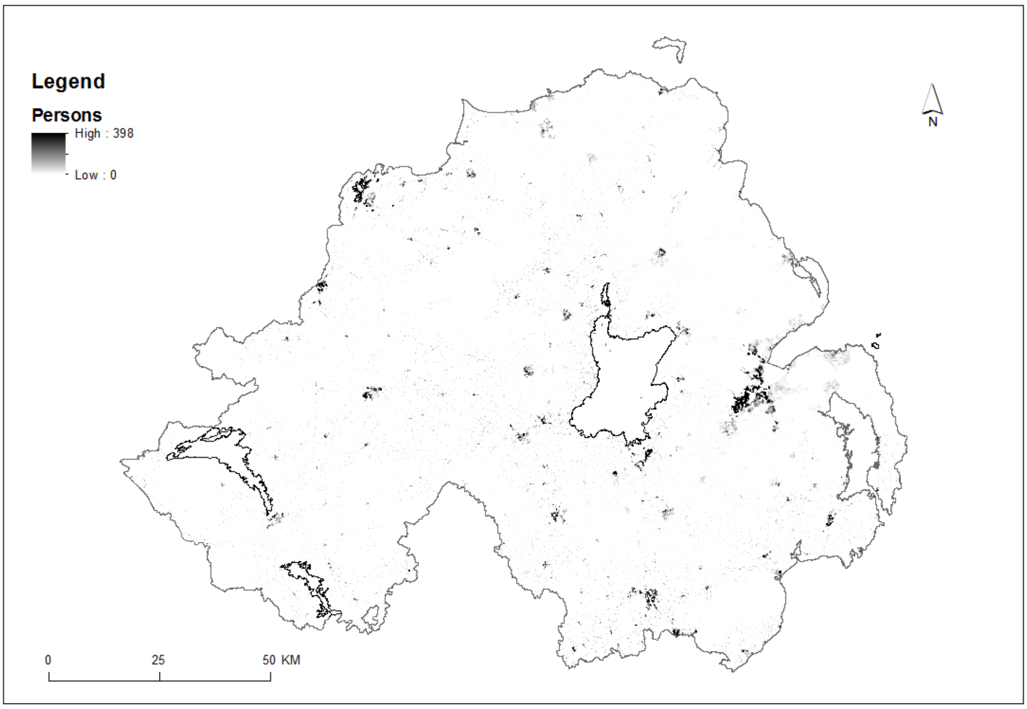

Figure 3. Estimates of Catholics: 3 by 3 pixel filter

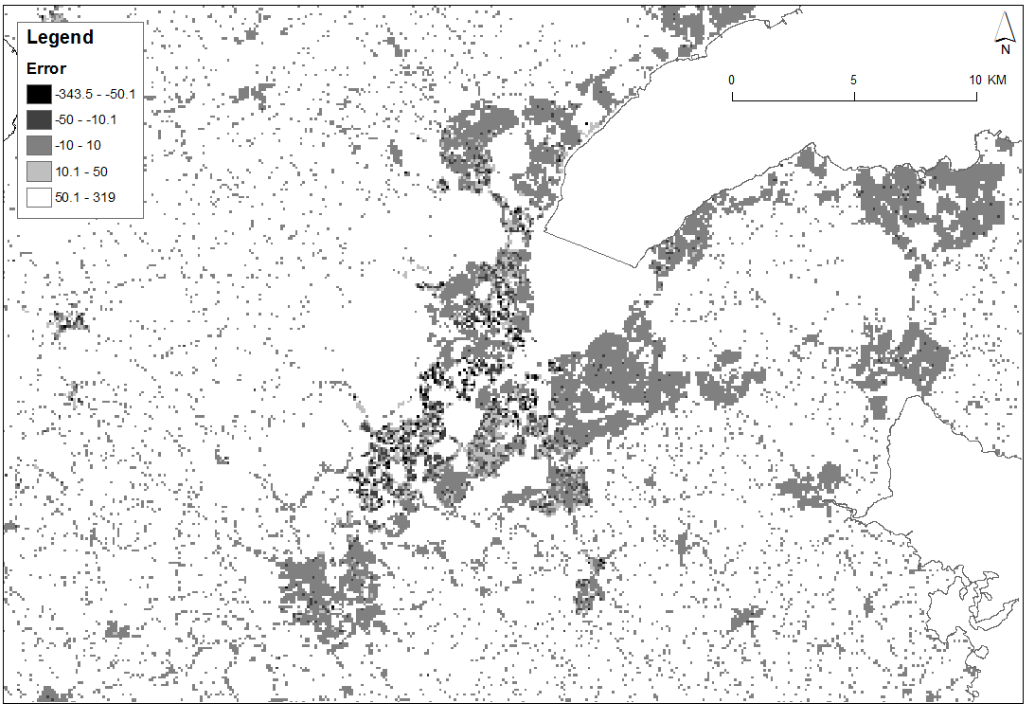

Figure 4. Estimates of Catholics, 3 by 3 pixel filter minus 'True' values: Belfast region