*Environment and Planning A* 2011, volume 43, pages 1965 – 1980

doi:10.1068/a43485

# Evaluation of gridded population models using 2001 Northern Ireland Census data

**David Martin**
Geography and Environment, University of Southampton, Southampton SO17 1BJ, England;
e-mail: D.J.Martin@soton.ac.uk

**Chris Lloyd, Ian Shuttleworth**
School of Geography, Archaeology and Palaeoecology, Queen's University Belfast,
Belfast BT7 1NN, Northern Ireland; e-mail: c.lloyd@qub.ac.uk, i.shuttleworth@qub.ac.uk
Received 10 October 2010; in revised form 31 March 2011

**Abstract. There is growing interest in the use of gridded population models which potentially offer advantages of stability through time and ease of integration with nonpopulation data sources. This paper assesses the accuracy of models of the type introduced by Martin in 1989. Population counts for census output areas (OAs) are reallocated to a 100 m grid and then compared with true 100 m cell population counts uniquely available from the 2001 Northern Ireland Census. This analysis is novel, being the first large-scale assessment of gridded population models against true gridded population counts. We find evidence that kernel width and cell size are more important than the distance-decay parameter; that local mass preservation approaches are more appropriate in urban areas; but that the spatial scale of input data is more important than model parameters. It is suggested that more attention needs to be given to the varying spatial structures of population between places and that incorporating this information through geostatistical approaches could yield further insights.**

## 1 Introduction

There is a long history of modelling and mapping population data on regular geographical grids instead of, or in addition to, spatially irregular units derived from censuses and administrative systems. Important advantages of gridded population models include stability through time and ease of integration with other georeferenced data sources from environmental, physical, or social applications. In some countries, such as Finland (Rusanen et al, 2001), gridded population data are a standard product. However, in many others such as England and Wales, where gridded data are not standard outputs, researchers must use some form of spatial reallocation to create grid-based estimates from available point-referenced or area-referenced data. Many different approaches can be used to this end. Some estimate population distributions using remotely sensed data (eg, Bhaduri et al, 2007; Mesev, 2003). Major initiatives have focused on the production of global-scale reference models such as that described by Balk et al (2006) and Tobler et al (1997), which are available at relatively coarse spatial resolutions. The assessment of these techniques falls outside the scope of this paper which, instead, concentrates on the evaluation of gridded population models using high-resolution ground-based data sources. We focus specifically on the approach presented by Martin (1989) which is generally available for use by researchers to create their own models from local data, thus increasing the need for understanding of model performance. A key challenge for the assessment of all gridded models is that population counts are not usually available in the same study region for both irregular zones and a regular grid. This has presented long-term difficulties for those aiming to assess the usefulness of gridded models. However, uniquely in the UK, census counts for irregular zones and regular grids are available in Northern Ireland (NI).

The paper therefore addresses this challenge by comparing gridded population models with true 2001 Census population counts for 100 m and 1 km cells from the

NI Grid Square Product. It evaluates model parameters, assesses the effects of geographical context (eg, urban or rural), and considers the importance of the underlying zonal geography. Previous studies (Martin, 1996) demonstrate the theoretical advantages of gridded data over representations which seek, at one extreme, to allocate populations to centroid points or, at the other, to distribute populations across irregular zones. Model evaluation was restricted by the absence of true gridded population counts. Martin et al (2000) attempted more systematic evaluation using NI gridded population data but were limited because only a small sample of grid-square counts were available at that time. The present paper overcomes these deficiencies by generating point-referenced, gridded, and zonal (choropleth) representations from 2001 NI Census data and assessing performance against the complete set of gridded population counts. The analysis therefore provides empirical evidence to inform judgments of the accuracy of gridded population models. This has relevance to places where no true population count data are available for comparative purposes, with the caveat that the detailed conclusions are restricted to the specific type of modelling algorithm used here. The work is timely as there is an increasing interest in using gridded data to overcome the problem of continually changing small-area census geographies, with a new round of censuses internationally (Valente, 2010). Gallego (2010), for instance, uses land-use data to inform the interpolation of population totals to grid squares across Europe, one specific advantage being the continuity of gridded geographies across borders and differing statistical reporting zones.

The remainder of the paper comprises five sections. In section 2 we review the rationale and applications of gridded population modelling and in section 3 the particular algorithm which is being evaluated here. Section 4 describes the NI context and the data that are used for the analysis. Section 5 presents the implementation and comparison between model outputs and true population counts. Finally, we conclude by assessing the relative importance on model performance of the model parameters, geographical context, and input data. These wider implications have relevance beyond the specific NI context.

## 2 Applications of gridded population models

Numerous researchers choose to use gridded population models in order to overcome some of the more severe weaknesses of conventional population representations using irregular zonal boundaries. These are generally applications in which the spatial distribution of population has an important impact on the analysis, independent of the zonal geography used. The literature sometimes refers to such methods as producing population 'surfaces', although as few involve strictly continuous mathematical surface functions, the term 'gridded' is adopted here. Tate (2000) summarizes their general advantage as being a more realistic model of settlement pattern, reflecting density changes that are hard to represent using area-based representations. Example applications based specifically on the technique evaluated here include investigation of social class inequalities in the risk factors associated with flooding (Fielding, 2007); studies of environmental equity and risk assessment (Brainard et al, 2002); population exposure to transport of hazardous waste (Brainard et al, 1996; Lovett et al, 1997); more general transportation cost modelling (Brainard et al, 1997; Martin et al, 2002), and as the basis for cellular automata modelling of urban expansion (Wu and Martin, 2002). Mesev et al (1995) adopt a hybrid approach in which the census-based gridded population estimates are used to enhance the classification of urban areas from remotely sensed data.

One benefit of gridded population models common across these applications is that, at an appropriate spatial resolution, they are better able to convey the geography of settlement than zonal representations. Under most settlement patterns, particularly in rural areas, many cells of a gridded model will have zero population, reflecting the discontinuous nature of the population distribution. Moreover, many social and environmental processes are continuous over space and so suitable for representation using a regular grid. Remotely sensed data, for example, can readily be resampled onto a regular gridded geography. A second benefit is comparability over time. The UK, in particular, is subject to continual revision of census and administrative boundaries, primarily due to policies related to electoral representation. Considerable effort is being devoted to the maintenance of 2001 Census output areas (OAs) in 2011 to minimise change, but there will still inevitably be mergers and splits necessitated by substantial population changes (Cockings et al, 2009). It is therefore not possible to compare the results of the last four UK censuses on a consistent small-area geography. By contrast, gridded representations may be directly overlaid and reveal not only the numerical change in population but also the changing spatial extents of populated areas. In all these applications the comparability provided by gridded models is a key benefit, both between different data sources and within sources over time. An interesting further application, not considered here, is the modelling of more detailed temporal changes in population such as day/night cycles (Bhaduri et al, 2007). Most of the applications listed above also benefit from the abstract nature of the gridded structure which allows the output of multiple data modelling exercises to be combined using a common spatial framework. Others have identified the potential advantages of such models, but were unable to make use of them due to specific data limitations (eg, see Rosero-Bixby and Palloni, 1998; Verter and Kara, 2001).

Despite their evident utility in a range of applications, a fundamental challenge in the calibration and evaluation of gridded population models from any source has been difficulty in testing the outputs due to the absence of sufficiently detailed true counts. Robinson and Zubrow (1997) used synthetic surfaces to evaluate the performance of four algorithms, but production of synthetic surfaces which display the range of characteristics encountered in a real-world settlement distribution is problematic. The use of actual census population counts, as here, overcomes this difficulty.

## 3 Population modelling method

This paper employs a method, first presented by Martin (1989), for the construction of gridded population estimates from zone centroids. Centroids are $(x, y)$ point locations identified as local centres of population: for example, the population-weighted centre of a census reporting zone. This is by no means the only method for generating gridded representations of socioeconomic data and various alternatives, driven by a similar desire to overcome weaknesses of zone-based representations, may be found in Goodchild et al, (1993), Langford and Unwin (1994), Tobler (1979), and Thurstain-Goodwin and Unwin (2000). These all utilise some form of interpolation or dasymetric allocation (Mennis, 2003) from centroid or area-referenced sources into grid cells. Kyriakidis (2004) proposes a geostatistical framework for generation of population surfaces by area-to-point Kriging and would prefer not to associate populations with centroid locations; Yoo et al (2010) compare the Kyriakidis and Tobler methods. We focus here on a method employing centroid locations because these are precisely the type of geographical reference locations generated by contemporary population data systems and they potentially offer greater geographical detail about population distributions than zone boundaries alone. The impact of using centroids at different spatial scales is one of the aspects to be investigated.

A more extensive review of the basic algorithm used here, including the concepts of count redistribution and adaptive kernel width, together with their strengths and weaknesses as compared to conventional zonal mapping, is presented in Martin (1996). That paper also discusses an extension in which population redistribution is constrained within known zone boundaries. Tobler's (1999) commentary on this approach emphasises the redistribution of population to all cells in the output grid, whereas our concern is with the creation of a spatially discontinuous representation, in which the retention of empty (unpopulated) cells is an important feature of the resulting models. We assume the existence of population count data for irregularly sized and shaped zones, which are represented by population-weighted centroid locations. The basic approach is not to interpolate or smooth the centroid populations but rather to redistribute the fixed total count of population from the point locations into the surrounding grid cells. To determine the weighting of each cell with respect to a centroid, we employ a locally adaptive kernel weighting function as illustrated conceptually in figure 1, showing a distance decay function centred on a centroid $j$. A review of the general principles of such approaches is provided by Lloyd (2011). The notation used in the figure relates to the following explanation.

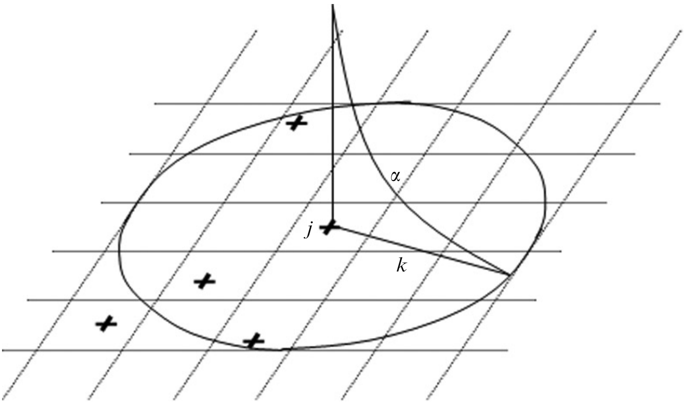


**Figure 1.** Redistribution of population counts from centroid $j$ to cells within kernel width $k$, using distance-decay parameter $\alpha$.

The method was first applied to enumeration districts (EDs) from the 1991 UK Census and a series of national gridded models assembled by Bracken and Martin (1995). Each centroid is treated as a local summary point for the more detailed, but unknown, actual population distribution and a zone may be represented by more than one centroid if such data are available. Modelling proceeds by processing each centroid in turn and estimating a local weighting function for this redistribution. The process is governed by an initial user-defined kernel width, within which the mean intercentroid distance is determined. The kernel width is then locally adapted to equal this distance and weights assigned to local cells according to a distance-decay function. In the construction of the 1991 UK models, $w_{ij}$, the weighting of cell $i$ with respect to centroid $j$, is determined by:

$$w_{ij} = \left( \frac{k^2 - d_{ij}^2}{k^2 + d_{ij}^2} \right)^{\alpha}, \qquad (1)$$

where $k$ is the kernel width and $d_{ij}$ is the distance between the centre of cell $i$ and centroid $j$. Beyond the adjusted kernel width $d > k$, all weights are zero: $w_{ij} = 0$. All that is required is a function that reduces the weighting from a peak at the centroid location to zero at the edge of the kernel, and many such functions exist. That given here is based on Cressman (1959). Thiebaux and Pedder (1987) add the exponent $\alpha$ in order to offer control over the shape of the distance-decay function within the extent of the spatial kernel. This function is convenient for experimentation in that it permits variation in the steepness of the distance decay within the kernel simply by varying $\alpha$. Values of 1 result in an approximately uniform decline from centroid to kernel edge, while values less than 1 produce flatter kernels with sharper decline at the edge and values greater than 1 produce more peaked functions with rapid decay away from the centroid. The total population recorded at each centroid is then redistributed in proportion to the weights assigned to cells within the kernel, and all other centroids are processed in the same way. The total population received by cell $i$ is thus the sum of its weighted population from all centroids:

$$\hat{P}_i = \sum_{j=1}^{N} P_j w_{ij} \ , \qquad (2)$$

where $N$ is the total number of centroids and $P_j$ is the population at centroid $j$. Edge effects are avoided by processing an area of at least one kernel width greater than the output region to be modelled, allowing population from centroids just beyond the region to be included, and for some of the population of centroids within $k$ distance of 4the edge to be lost.

In previously published work, it has been noted that the pattern of populated cells in the output surface does not appear to be greatly sensitive to the value of $\alpha$ but that the initial kernel width, $k$, plays a crucial role. Figure 1 shows these parameters. Martin et al (2000) also examine various potential enhancements to the earlier techniques, and particularly note the sensitivity of these approaches to the precision and accuracy of the centroid locations. In general, the most promising results are obtained with the most detailed centroid locations, effectively moving the model closer to the 'ideal' in which each centroid represents the location of a single individual or household in the population, and no redistribution modelling would be required. This paper therefore considers the impact of the parameters $\alpha$ and $k$ and cell size on the estimated gridded population, and the impact of using centroids at varying spatial scales (namely OA centroids and unit postcodes). The paper also goes further by looking at model performance in urban versus rural areas, contrasting Belfast District Council (DC) with the rest of NI.

## 4 Data and analysis

Our analysis employs census data for the whole of NI in 2001. OA population counts were used as input to the modelling procedure to estimate populations at 100 m and 1 km cells. There were 5022 OAs in NI with a mean population size of 336 and a minimum threshold population of 100 persons and 40 households. OAs had a mean area of 2.7 km$^2$, a median size of 0.147 km$^2$, and varied from a minimum of 0.001 km$^2$ to a maximum of 101.32 km$^2$. OAs in NI were designed using the same algorithm as in England and Wales (Martin, 2002). This produced a tradeoff between population size, social homogeneity, and shape but as can be seen above there is still considerable variation in population size and area. For each OA a digital boundary and address-weighted centroid location were available. OAs were selected as the starting point because they are the most commonly used small-area geography for census counts. Smaller unit postcodes were also introduced to the analysis to enable consideration of the impact of spatial scale—or the support—on the accuracy of gridded population models. There were 35 320 unit postcodes in NI, representing the lowest level of the postal delivery system. Their mean population was 48, with a minimum of 3 and maximum of 2582. A look-up table was obtained from the Northern Ireland Statistics and Research Agency (NISRA) which showed population counts at postcode level and their allocation to census OAs.

True population counts for 2001 for 100 m and 1 km cells were obtained from the NI Grid Square Product (Shuttleworth and Lloyd, 2009; Shuttleworth et al, 2006). This resource is unique to NI within the UK and allows direct assessment of gridded models. In 1971, in addition to standard census output geographies, grid square counts were provided in Great Britain and NI. However, in NI only, these counts have continued to be provided from subsequent censuses. In 1991, as in previous census years, 1 km grid squares were available for the whole of NI while 100 m grid squares were provided only for urban areas. In 2001 both 1 km grid squares and 100 m grid squares were made available for the whole of NI. For grid squares containing fewer than 25 persons and fewer than 8 households (it was necessary that both criteria were satisfied) outputs are restricted to total males, total females, total persons, and total households. Figure 2 shows the population of NI in 2001 for 100 m grid cells, which is here used as the main benchmark for the assessment of gridded model performance. A successful gridded population model will be a close approximation to this distribution and, at this scale, visually very similar.

In assessing the performance of these gridded models it is useful to understand something of the population geography of NI. The city of Belfast and its surrounding urban area is the largest population concentration in NI and is clearly visible in the east of figure 2. Belfast DC (2001 Census population 277 391) forms the core of the Belfast urban area which extends geographically beyond this purely formal urban definition, and is comparable in size to many UK regional cities. The majority of NI

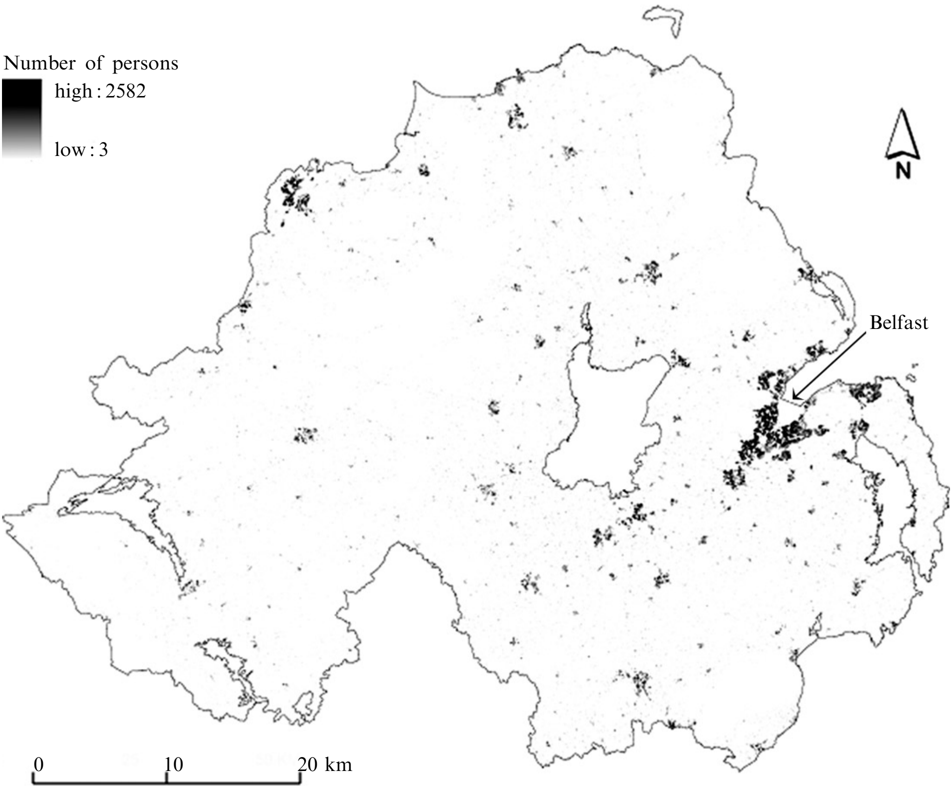


**Figure 2.** 2001 Census population of Northern Ireland in 100 m grid cells.

outside Belfast is rural and sparsely populated, especially the rural areas in the west and south. This translates into a gridded population distribution that is highly skewed—a majority of populated 100 m cells have fewer than 10 people in them and a majority of all the NI surface area is empty of people. Highly populated cells with 80 or more people are comparatively rare. NI thus offers a contrasting and complex environment in which to evaluate the population models; it has a major city but also sparsely populated areas which are typical of many regions elsewhere in the UK, Ireland, and beyond. These contrasts are very evident in figure 2. The true population and modelled results for Belfast are shown in more detail in section 5 below.

Our analysis uses a Visual Basic implementation of the gridded population algorithm called SurfaceBuilder which interprets a variety of ASCII input file layouts containing X, Y (coordinate) and Z (count value) fields. The definition of a particular model run, in terms of all the selected model parameters, may be stored and retrieved for reuse. The user is able to export the completed model for mapping, as has here been conducted using ArcGIS. The compiled SurfaceBuilder program may be downloaded from the software link at http://www.public.geog.soton.ac.uk/users/martindj/.

SurfaceBuilder was used to construct a series of gridded population models for NI, using grid cell sizes of 25 m and 50 m, which have subsequently been aggregated to match the cells for which true counts are available. These have been created separately from OA centroids and unit postcodes. A raster mask has been used to ensure that, subject to the cell resolution, population counts from a centroid may be redistributed only to cells falling within the appropriate OA. The use of a mask in this way enforces local mass preservation, whereby the total population allocated to each zone is constrained to its true total. The alternative, which does not use any masking values, will preserve the total population of the study area (and of any settlement surrounded by unpopulated cells) and is termed global mass preservation; both options are evaluated. Modelling scenarios were run using a range of kernel widths and distance-decay functions so as to better understand their impact on the accuracy of the gridded population estimates. We have considered a large range of parameter combinations and selected models for presentation here which span typical values used in previous studies using comparable UK datasets based on this method. A more limited analysis based on unit postcodes was undertaken to explore the implications of changing the spatial scale of the input centroids.

Two reference distributions were also generated. Firstly, population counts were allocated entirely to the cells containing their OA centroids, which equates to a representation using only centroid locations with no redistribution. Secondly, a zonal model was created which allocates the population of an OA evenly across all the raster cells assigned to that OA. The latter has the equivalent representational characteristics to a conventional shaded area map.

The question of the appropriate metric to use to evaluate the utility of the gridded population models is difficult because it depends on the aim of the analysis and the type of area in which it is being undertaken. In rural areas it might be appropriate to know, for example, simply if an area is populated or not. In urban areas a user may be more interested in the size of the population. A variety of measures were therefore employed for assessment and table 1 presents a summary of the model variants which are tabulated in the following section. One measure was the extent to which the pattern of populated and unpopulated cells was accurately produced by the algorithm. However, for areas that were populated, and for Belfast DC in particular, numeric indicators such as mean error, standard deviation, and root mean square error (RMSE) were selected as being more appropriate. For a range of models different $\alpha$ and $k$ parameters were assessed, as were local and global mass preservation.

**Table 1.** Summary of model runs reported in tables 2–7.

| Table | Coverage[a] | Input data[b] | Parameters held constant | Parameters varied | Outcome measures |
|---|---|---|---|---|---|
| 2 | NI | OA | none | mass preservation, search radius, distance decay, cell size | presence/absence of population |
| 3 | NI | postcode | | | |
| 4 | NI | OA | cell size (50 m) | mass preservation, search radius, distance decay | mean error, standard deviation, root mean square error |
| 5 | Belfast | OA | cell size (50 m) | | |
| 6 | Belfast | OA | cell size (25 m) | | |
| 7 | NI | postcode | cell size (50 m) | | |

Note. In tables 2–7 local mass preservation refers to the constraint of population redistribution within the same zone; in global mass preservation, totals are constrained only within the entire study area.

[a] NI—Northern Ireland.
[b] OA—output area.

We have selected parameter ranges which reflect the typical applications and data reviewed in section 2 and which cover the principal interactions between parameters. These considerations inform the presentation of results set out in more detail below.

## 5 Results

Here, we first examine the extent to which the centroid, zonal, and gridded models correctly predict populated and unpopulated 100m cells for NI as a whole (tables 2 and 3, section 5.1). We then explore mean error, standard deviation, and RMSE of the output from the gridded models using varying parameters and input data, for populated cells for NI as a whole and for Belfast separately (tables 4 to 7, section 5.2). The tabulation of results that follows is necessarily selective and further tables, presenting a wider range of parameter combinations, are presented in the online appendix (http://dx.doi.org/10.1068/a43485).

### 5.1 100 m cells—presence or absence of population

Tables 2 and 3 present information on the extent to which different population models predict which cells in NI are populated or unpopulated using OA and unit postcodes centroids as inputs, respectively. If 0.5 or more people were identified in a 100 m cell,

**Table 2.** Presence or absence of population in 100 m cells from output area (OA) centroids.

| Model | Mass preservation | Search radius, $k$ (m) | Distance decay, $\alpha$ | Cell size (m) | Percentage correctly | |
|---|---|---|---|---|---|---|
| | | | | | populated | unpopulated |
| 1 | OA centroids—global | na | na | 100 | 3.70 | 99.9 |
| 2 | OA zones—global | na | na | na | 56.8 | 79.1 |
| 3 | global | 250 | 2.00 | 25 | 30.8 | 97.3 |
| 4 | local | 250 | 2.00 | 25 | 30.2 | 97.4 |
| 5 | global | 500 | 1.00 | 50 | 42.2 | 91.5 |
| 6 | global | 500 | 2.00 | 50 | 39.9 | 93.0 |
| 7 | local | 500 | 1.00 | 50 | 40.8 | 91.8 |
| 8 | local | 500 | 2.00 | 50 | 38.9 | 93.2 |

Note. na—not applicable

**Table 3.** Presence or absence of population in 100 m cells from postcode centroids.

| Model | Mass preservation | Search radius, $k$ (m) | Distance decay, $\alpha$ | Cell size (m) | Percentage correctly populated | Percentage correctly unpopulated |
|---|---|---|---|---|---|---|
| 1 | global | 500 | 0.25 | 50 | 66.3 | 79.9 |
| 2 | global | 500 | 1.00 | 50 | 67.3 | 81.5 |
| 3 | global | 500 | 2.00 | 50 | 65.8 | 83.7 |
| 4 | local | 500 | 0.25 | 50 | 65.8 | 80.4 |
| 5 | local | 500 | 1.00 | 50 | 66.3 | 82.1 |
| 6 | local | 500 | 2.00 | 50 | 64.9 | 84.4 |

then this was rounded to 1 and the cell was deemed to be populated. This criterion was also followed for the transfer of zone-based counts to cells and, for some zones, the count of persons per 100 m square was less than 0.5. This explains why, in table 2, OA zones (model 2) has a positive value for percentage correctly unpopulated. Modelling was undertaken on 50 m cells that were then aggregated to 100 m cells. The $\alpha$ and $k$ parameters were allowed to vary as well as the cell size at which modelling was performed. The metrics used were the percentage of populated 100 m cells that were correctly modelled as being populated and the percentage of unpopulated 100 m cells that were also correctly modelled as being unpopulated.

The results are complex and need careful interpretation. Table 2 shows how gridded models offer a tradeoff between the characteristics of zonal and centroid representations. Using the indicator of the percentage of cells correctly identified as being unpopulated, model 1 (centroids) might be seen as being the most successful, with its 99.87% accuracy. However, it is the weakest model in identifying populated 100 m cells. This is because it concentrates OA populations at OA centroids and overestimates empty cells. At the other extreme model 2 (zonal) allocates OA populations across all OA zones. In spreading the population like this, it correctly populates nearly 57% of 100 m cells—a higher figure than the gridded population models in models 3–8—but at the expense of incorrectly populating many empty cells. All the gridded models lie between these two extremes. They have a poorer performance in correctly predicting populated cells than model 2 but are more accurate than model 1 in this respect; and they are not as accurate as model 1 in predicting unpopulated cells but outperform model 2 in this respect. In these aspects, the gridded population models are more accurate overall than models 1 and 2. There are clear differences between models 3–8. In correctly predicting populated cells, models 5–8 ($k = 500$ m, cell size $= 50$ m) perform better than models 3 and 4 ($k = 250$ m, cell size $= 25$ m), indicating that cell size and the search radius $k$ parameter are important. Varying the distance-decay parameter $\alpha$ in models 5–8 while $k$ and cell size are held constant, produces differences between the models of around 3 percentage points, suggesting that cell size and $k$ (either singly or combined) are more important than $\alpha$. There is no evidence to suggest that local mass preservation is superior, although differences are small.

Table 3 represents a change of spatial scale, being based on unit postcodes. Models 2 and 3 (table 3) can be directly compared with models 5 and 6 (table 2) and models 5 and 6 (table 3) can be directly compared with models 7 and 8 (table 2). The models in table 3 perform worse than those in table 2 in predicting unpopulated cells. However, the table 3 models are more accurate in predicting populated cells. This can be attributed primarily to the smaller size of unit postcodes, which are closer in size to the 100 m cells on which population is being modelled than the larger OAs.

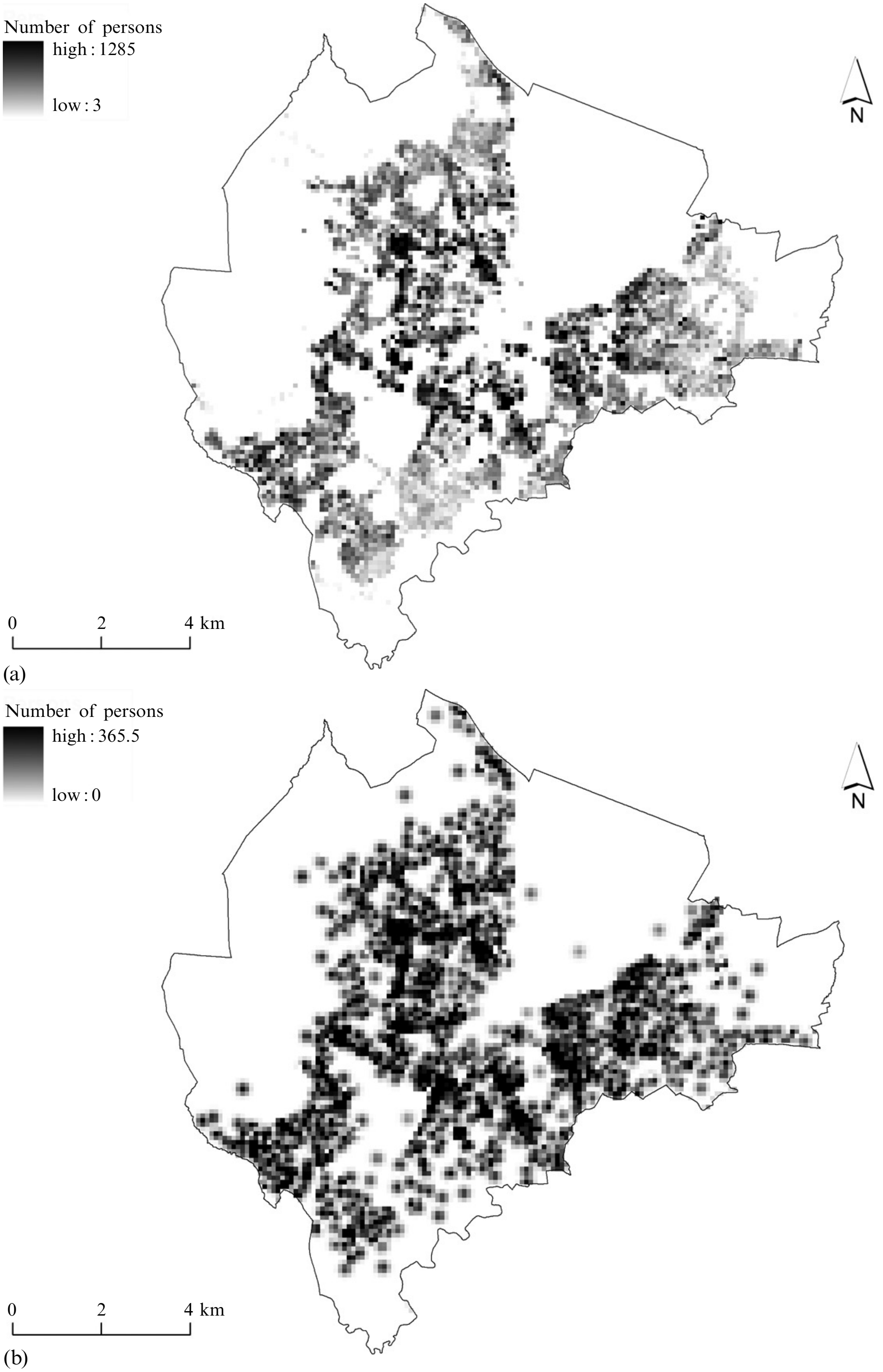


**Figure 3.** Belfast District Council: (a) 'true' census population counts and (b) modelled counts for 100 m cells from output area counts using local mass preservation: search radius 500 m, distance decay 2, 50 m grid cells.

However, the greater number of centroid locations means that overall there is more spreading of population into areas that are truly unpopulated.

Figure 3 shows (a) 'true' counts and (b) modelled counts for 100 m cells for Belfast DC. Model 12 from table 4 is used here as representative of the appearance of the gridded models. The two maps display similar spatial patterns although, as expected, the maximum modelled counts are much smaller than the maximum true counts. Inevitably, the modelling procedure does not capture well high-intensity populations in individual 100m cells which are not captured by OA data.

### 5.2 Population counts

We now turn from the presence or absence of population in grid cells to the estimation of population counts. OA zone counts per 100 m cell were obtained by overlaying vector OA boundaries and vector 100 m cells. The centroid model achieves an overall mean error of −258.35 and RMSE of 271.78, compared with the zonal model values of −8.78 and 156.85, respectively. As might be expected, the choropleth zonal model produces lower RMSE values overall than the centroid model, essentially because the centroid model allocates population into far too few cells. The zonal model, although both overestimating and underestimating population densities in different places provides a better overall coverage than the centroids.

Table 4 shows the prediction errors (predicted – observed) where modelling was undertaken on 50 m cells which were then aggregated to 100 m cells for NI as a whole. The mean error closest to zero is in model 9 (local mass preservation, $k = 250$ m, $\alpha = 2$). The smallest RMSE was for model 6 (global mass preservation, $k = 500$ m, $\alpha = 2$). There are not major differences between the outputs from the models in table 4, indicating that the outcomes are not very sensitive to the choice of $\alpha$ and $k$, nor indeed to the choice of global versus local mass preservation. Most errors in the table are small. There is only one error greater than −2000, and only five in total greater than −1000 : for model 6, for example. When these are excluded, the RSME drops to 28.086. The largest negative error occurs in a small OA near Belfast International Airport and is believed to be a single large communal establishment that is not being modelled effectively. Other specifications were explored for prediction to 25 m cells and then aggregation to 100 m (detailed in the online appendix), but these

**Table 4.** Prediction errors (prediction – observed) for Northern Ireland: prediction to 50 m cells and aggregation to 100 m cells.

| Model | Mass preservation | Search radius, $k$ (m) | Distance decay, $\alpha$ | Cell size (m) | Mean error | Standard deviation | RMSE[a] |
|---|---|---|---|---|---|---|---|
| 1 | global | 250 | 0.25 | 50 | −7.12 | 36.77 | 37.46 |
| 2 | global | 250 | 1.00 | 50 | −6.20 | 38.07 | 38.57 |
| 3 | global | 250 | 2.00 | 50 | −5.49 | 41.52 | 41.88 |
| 4 | global | 500 | 0.25 | 50 | −10.58 | 32.96 | 34.61 |
| 5 | global | 500 | 1.00 | 50 | −9.60 | 32.34 | 33.74 |
| 6 | global | 500 | 2.00 | 50 | −8.76 | 32.27 | 33.43 |
| 7 | local | 250 | 0.25 | 50 | −6.82 | 35.71 | 36.36 |
| 8 | local | 250 | 1.00 | 50 | −5.81 | 37.18 | 37.63 |
| 9 | local | 250 | 2.00 | 50 | −5.06 | 40.41 | 40.73 |
| 10 | local | 500 | 0.25 | 50 | −10.26 | 32.56 | 34.13 |
| 11 | local | 500 | 1.00 | 50 | −9.54 | 32.29 | 33.67 |
| 12 | local | 500 | 2.00 | 50 | −9.89 | 32.36 | 33.56 |

[a] RMSE—root mean square error.

were no more accurate than the results as tabulated in table 4. The models presented in table 4 were an advance on the zonal and centroid models in terms of RMSE, although the picture was more mixed when evaluating the mean error. Further modelling was undertaken on 50 m cells which were then aggregated to 1km cells and for prediction to 25 m cells and aggregation to 1 km cells using a 250 m search radius in both instances. In these cases (again detailed in the online appendix), the RMSE values are clearly smaller for large search windows and more gradual distance decay functions. The smallest RMSE values were obtained with global, rather than local, mass preservation.

There are several trends apparent from table 4. Each variant of the gridded model produces RMSE values much lower than the centroid and zonal representations. The RMSE values for global and local mass preservation where $k = 250$ m are larger for larger $\alpha$ values. Conversely, the RMSE values for global and local mass preservation where $k = 500$ m are smaller for larger $\alpha$ values. These figures demonstrate the interactions between (i) spatial variation in the population distribution and (ii) homogeneity of zones (here, OAs). A small search radius $k$ with a small distance-decay parameter $\alpha$ (giving more weight to distant observations than a large $\alpha$) spreads the population more evenly than would a small $k$ with a large $\alpha$—this latter combination would concentrate population more densely around the grid cell at the centre of the kernel. The results suggest that the most accurate predictions are obtained when the population is spread evenly, but not *too* evenly. With a small $k$ and large $\alpha$ parameter, the population is too densely clustered around centroid locations, whereas with a large $k$ and small $\alpha$, the population is too evenly spread out. If $k$ and $\alpha$ parameters are both large or both small, an acceptable balance appears to be achieved.

Table 5 gives the prediction errors just for Belfast for prediction to 50 m cells and aggregation to 100 m cells, while table 6 gives the equivalent figures for prediction to 25 m cells and aggregation to 100 m cells, the latter only for a 250 m search radius. These tables highlight the ways in which the models perform differently when looking only at an urban area. The mean error closest to zero across tables 5 and 6 is for model 3 in table 5 (global mass preservation, $k = 250$ m, $\alpha = 2$) for prediction to 50 m cells. However, the smallest RMSEs are for local mass preservation models. The smallest RMSE, for example, is for local mass preservation using $k = 500$ m, $\alpha = 2$ also for

**Table 5.** Predication errors (prediction – observed) for Belfast District Council: prediction to 50 m cells and aggregation to 100 m cells.

| Model | Mass preservation | Search radius, $k$ (m) | Distance decay, $\alpha$ | Cell size (m) | Mean error | Standard deviation | RMSE[a] |
|---|---|---|---|---|---|---|---|
| 1 | global | 250 | 0.25 | 50 | −7.51 | 40.95 | 41.64 |
| 2 | global | 250 | 1.00 | 50 | −6.50 | 41.38 | 41.89 |
| 3 | global | 250 | 2.00 | 50 | −5.73 | 45.17 | 45.53 |
| 4 | global | 500 | 0.25 | 50 | −11.46 | 42.51 | 44.03 |
| 5 | global | 500 | 1.00 | 50 | −9.98 | 40.81 | 42.01 |
| 6 | global | 500 | 2.00 | 50 | −8.71 | 39.50 | 40.45 |
| 7 | local | 250 | 0.25 | 50 | −7.64 | 39.74 | 40.46 |
| 8 | local | 250 | 1.00 | 50 | −6.67 | 40.41 | 40.96 |
| 9 | local | 250 | 2.00 | 50 | −5.93 | 43.30 | 43.70 |
| 10 | local | 500 | 0.25 | 50 | −9.71 | 39.33 | 40.51 |
| 11 | local | 500 | 1.00 | 50 | −9.08 | 38.89 | 39.94 |
| 12 | local | 500 | 2.00 | 50 | −8.47 | 38.74 | 39.66 |

[a] RMSE—root mean square error.

**Table 6.** Prediction errors (prediction – observed) for Belfast District Council: prediction to 25 m cells and aggregation to 100 m cells.

| Model | Mass preservation | Search radius, $k$ (m) | Distance decay, $\alpha$ | Cell size (m) | Mean error | Standard deviation | RMSE[a] |
|---|---|---|---|---|---|---|---|
| 1 | global | 250 | 0.25 | 25 | −7.70 | 40.68 | 41.40 |
| 2 | global | 250 | 1.00 | 25 | −6.72 | 41.08 | 41.62 |
| 3 | global | 250 | 2.00 | 25 | −5.97 | 44.47 | 44.86 |
| 4 | local | 250 | 0.25 | 25 | −8.21 | 39.08 | 39.93 |
| 5 | local | 250 | 1.00 | 25 | −7.23 | 39.72 | 40.38 |
| 6 | local | 250 | 2.00 | 25 | −6.48 | 42.29 | 42.78 |

[a] RMSE—root mean square error.

prediction to 50 m cells (model 11 in table 5). There was little benefit in predicting to 25 m cells. The advantages of local mass preservation are more clearly seen in Belfast than in NI as a whole (compare table 4). This is intuitively correct as urban areas such as Belfast have smaller OAs in closer proximity than rural areas and thus more local information is available. In contrast, in rural areas with large OAs where the population may be heterogeneously distributed, methods that rely too much on local information may result in reduced prediction accuracy.

Table 7 returns to consideration of the whole of NI, but this time using postcode centroids as input. The results can be compared with those in table 4. The smaller mean errors and RSMEs in table 7 illustrate the importance of the spatial scale of the input units since they are generally less than those in table 4 where the analysis starts with OAs. It is better to begin with centroids representing zones that are nearer in size to the regular grid for which estimates are being made.

**Table 7.** Postcode data: prediction errors (prediction – observed) for Northern Ireland—prediction to 50 m cells and aggregation to 100 m cells, local mass preservation to output areas.

| Model | Mass preservation | Search radius, $k$ (m) | Distance decay, $\alpha$ | Cell size (m) | Mean error | Standard deviation | RMSE[a] |
|---|---|---|---|---|---|---|---|
| 1 | global | 500 | 0.25 | 50 | −7.31 | 21.71 | 22.91 |
| 2 | global | 500 | 1.00 | 50 | −6.49 | 20.82 | 21.81 |
| 3 | global | 500 | 2.00 | 50 | −5.73 | 19.99 | 20.79 |
| 4 | local | 500 | 0.25 | 50 | −5.73 | 22.11 | 22.84 |
| 5 | local | 500 | 1.00 | 50 | −5.13 | 21.64 | 22.23 |
| 6 | local | 500 | 2.00 | 50 | −4.58 | 21.79 | 22.27 |

## 6 Conclusion

Certain properties of gridded models of population are inherent—for example, the independence of the arbitrary grid from boundary changes—aiding analysis of population change over time. In most countries, gridded counts are not available as direct aggregations from individual-level data, so the construction of gridded population models is necessary to achieve these benefits. Potential benefits such as the preservation of settlement pattern and the preservation of unpopulated areas are dependent to a significant extent on the input data and model parameters used, such as grid cell size and the shape and extent of some spatial redistribution function. Due to the absence of gridded counts, it is normally impossible to assess the performance of the resulting models in those situations where they may be of most value. The unique NI grid square product has provided us with a means of undertaking such an evaluation for the first time and these findings have relevance to analysis of spatial population distributions in any context.

Gridded population models offer some generic advantages over centroid and zonal (choropleth) representations of population, as was shown in table 2. Zonal models spread the population far too evenly over space and perform more poorly than gridded models in predicting unpopulated cells. This is a particular weakness for applications which involve distance to population, such as the transportation or hazard models considered in section 2, yet zonal representations continue to be the most pervasive form of population mapping. Representations based on centroids alone concentrate the population in only a few places and, whilst identifying the largest proportion of unpopulated cells correctly, perform very poorly in indicating all populated cells. The use of centroid locations remains common in accessibility modelling, for example (Langford and Higgs, 2006). Gridded population models lie between these two approaches, taking some of the strengths of each. Nevertheless, their accuracy in picking out populated cells is only between 30% and 40%. This does not seem particularly high, but is perhaps explained as a consequence of real population geographies which have large numbers of vacant cells or cells with very low population densities—particularly in rural areas, which fall below the spatial resolution of the available population data. One of our most important conclusions is to highlight the weaknesses of the conventional approaches, and this has broad international applicability.

Our analysis also indicates that the accuracy of the gridded models is not, in general, very sensitive to the model parameters (global versus local mass preservation, $\alpha$, $k$, and cell size) used, with only small variations in diagnostics such as RMSE. A wide range of parameters will have the effect of concentrating population into truly populated areas and leaving remote areas unpopulated. However, there are some subtle differences apparent. Search radius ($k$) and cell size seem to be more important than distance decay ($\alpha$), as the results appear to be more sensitive to the choices made about them than for $\alpha$ (Fotheringham et al, 2002). Identification of populated areas has more impact than the detailed distribution within these areas. Moreover, there are some indications that local mass preservation works better than global mass preservation in densely populated areas. This suggests that users of gridded population models would do well to consider use of different parameters in urban and rural areas. However, the spatial resolution of the input centroids appears to be more important overall than parameterisation of the models. Models based on unit postcodes produced more accurate results than those based on OAs, the postcodes being much closer in spatial extent to the cell size of the output model. Spatial scale—specifically the support in geostatistical terms—seems to be more important than the selection of model parameters.

In drawing wider conclusions it is also worth considering the importance of place and spatial population structures. It has already been seen that local mass preservation techniques seem better suited to urban areas with higher population densities. In urban areas, for instance, analysts might be more interested in the predicted size of populations in cells (rather than whether they are populated or not) whereas in rural areas with low-density populations, there may be greater interest in estimating whether a cell is populated or not. There might thus be different motives for gridded population modelling in different places. However, our analysis also suggests that there are different challenges. Given the comments about spatial scale above, it is likely that gridded population models would face significantly greater challenges in areas with geographically large zones and very sparse populations than in more densely populated places, depending on the spatial scale of the grid to which predictions are being made. This suggests that, instead of a formulaic rule set, analysts need to assess spatial scale and model parameters flexibly according to geographical context and to assess the sensitivity of results across a series of models.

Our conclusions also indicate a direction for future research. There is a need to assess gridded population models wherever data permit, in other more varied geographical contexts and using a wider range of modelling approaches. A second route, allied to this, would be incorporate measures of the spatial structure of populations through the use of geostatistical methods to include more information in the models and to allow greater sensitivity to the analysis of between-place variation.

**Acknowledgements.** The authors acknowledge the following data sources: Northern Ireland Statistics and Research Agency, 2001 Census: Northern Ireland Grid Square Data and Standard Area Statistics (Northern Ireland) [computer file]. ESRC Census Programme, Census Dissemination Unit, MIMAS (University of Manchester); Northern Ireland Statistics and Research Agency, 2001 Census: Digitised Boundary Data (Northern Ireland) [computer file]. ESRC Census Programme, Census Geography Data Unit (UKBORDERS), EDINA (University of Edinburgh). Northern Ireland Statistics and Research Agency, 2001 Census: Directory of postcodes and output areas, NISRA. Ian Shuttleworth and Chris Lloyd received support from ESRC award RES-000-23-0478.

**References**

Balk D L, Deichmann U, Yetman G, Pozzi F, Hay S I, Nelson A, 2006, "Determining global population distribution: methods, applications and data" *Advances in Parasitology* **62** 119 – 156

Bhaduri N, Bright E, Coleman P, Urban M L, 2007, "LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics" *Geojournal* **69** 103 – 117

Bracken I, Martin D, 1995, "Linkage of the 1981 and 1991 UK Censuses using surface modelling concepts" *Environment and Planning A* **27** 379 – 390

Brainard J, Lovett A, Parfitt J, 1996, "Assessing hazardous-waste transport risks using a GIS" *International Journal of Geographical Information Systems* **10** 831 – 849

Brainard J S, Lovett A A, Bateman, 1997, "Using isochrone surfaces in travel-cost models" *Journal of Transport Geography* **5** 117 – 126

Brainard J S, Jones A P, Bateman J J, Lovett A A, Fallon P J, 2002, "Modelling environmental equity: access to air quality in Birmingham, England" *Environment and Planning A* **34** 695 – 716

Cockings S, Harfoot A, Hornby D, 2009, "Towards 2011 output geographies: exploring the need for, and challenges involved in, maintenance of the 2001 output geographies" *Population Trends* **138** 38 – 49

Cressman G P, 1959, "An operational objective analysis system" *Monthly Weather Review* **97** 367 – 374

Fielding J, 2007, "Environmental injustice or just the lie of the land: an investigation of the socio-economic class of those at risk from flooding in England and Wales" *Sociological Research Online* **12**(4), http://dx.doi.org/10.5153/sro.1570

Fotheringham A S, Brunsdon C, Charlton M, 2002 *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships* (John Wiley, Chichester, Sussex)

Gallego J, 2010, "A population density grid of the European Union" *Population and Environment* **31** 460 – 473

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

Kyriakidis P C, 2004, "A geostatistical framework for area-to-point spatial interpolation" *Geographical Analysis* **36** 259 – 289

Langford M, Higgs G, 2006, "Measuring potential access to primary healthcare services: the influence of alternative spatial representations of population" *The Professional Geographer* **58** 294 – 306

Langford M, Unwin D J, 1994, "Generating and mapping population density surfaces within a geographical information system" *The Cartographic Journal* **31** 21 – 26

Lloyd C D, 2011 *Local Models for Spatial Analysis* 2nd edition (CRC Press, Boca Raton, FL)

Lovett A A, Parfitt J P, Brainard J S, 1997, "Using GIS in risk analysis: a case study of hazardous waste transport" *Risk Analysis* **17** 625 – 633

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, 2002, "Geography for the 2001 Census in England and Wales" *Population Trends* **108** 7 – 15
Martin D, Tate N, Langford M, 2000, "Refining population surface models: experiments with Northern Ireland census data" *Transactions in GIS* **4** 343 – 360
Martin D, Wrigley H, Barnett S, Roderick P, 2002, "Increasing the sophistication of access measurement in a rural healthcare study" *Health and Place* **8** 3 – 13
Mennis J, 2003, "Generating surface models of population using dasymetric mapping" *The Professional Geographer* **55** 31 – 42
Mesev V (Ed.), 2003 *Remotely Sensed Cities* (Taylor and Francis, London)
Mesev V, Longley P, Batty M, Xie Y, 1995, "Morpohology from imagery: detecting and measuring the density of urban land use" *Environment and Planning A* **27** 759 – 780
Robinson J M, Zubrow E, 1997, "Restoring continuity: exploration of techniques for reconstructing the spatial distribution underlying polygonized data" *International Journal of Geographical Information Science* **11** 633 – 648
Rosero-Bixby L, Palloni A, 1998, "Population and deforestation in Costa Rica" *Population and Environment* **20** 149 – 185
Rusanen J, Muilu T, Colpaert A, Naukkarinen A, 2001, "Finnish socio-economic grid data, GIS and the hidden geography of unemployment" *Tijdschrift voor Economische en Sociale Geografi*e **92** 139 – 147
Shuttleworth I G, 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
Shuttleworth I, Lloyd C D, Anderson J, 2006, "Linking Northern Ireland Census of Population data 1971 – 2001", Final Report of RES-000-23-0478, School of Geography, Queen's University, Belfast
Tate N J, 2000, "Surfaces for GIScience" *Transactions in GIS* **4** 301 – 303
Thiebaux H J, Pedder M A, 1987 *Spatial Objective Analysis: With Applications in Atmospheric Science* (Academic Press, London)
Thurstain-Goodwin M, Unwin D J, 2000, "Defining and delineating the central areas of towns for statistical monitoring using continuous surface representations" *Transactions in GIS* **4** 305 – 318
Tobler W R, 1979, "Smooth pycnophylactic interpolation for geographical regions" *Journal of the American Statistical Association* **74** 519 – 530
Tobler W R, 1999, "Linear pycnophylactic interpolation: comment on a paper by D Martin" *International Journal of Geographical Information Science* **31** 85 – 90
Tobler W R, Deichmann U, Gottsegen J, Maloy K, 1997, "World population in a grid of spherical quadrilaterals" *International Journal of Population Geography* **3** 203 – 225
Valente P, 2010, "Census taking in Europe: how are populations counted in 2010?" *Population and Societies* **467** 1 – 4
Verter V, Kara B Y, 2001, "A GIS-based framework for hazardous materials transport risk assessment" *Risk Analysis* **21** 1109 – 1120
Wu F, Martin D, 2002, "Urban expansion simulation of Southeast England using population surface modelling and cellular automata" *Environment and Planning A* **34** 1855 – 1876
Yoo E H, Kyriakidis P C, Tobler W R, 2010, "Reconstructing population density surfaces from areal data: a comparison of Tobler's pycnophylactic interpolation method and area-to-point Kriging" *Geographical Analysis* **42** 78 – 98