Environment and Planning A 1995, volume 27, pages 211-224

Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation

P F Fisher, M Langford
Midlands Regional Research Laboratory, Department of Geography, University of Leicester, Leicester LE1 7RH, England
Received 4 October 1993; in revised form 21 February 1994

Abstract. Areal interpolation involves the transfer of data (often socioeconomic statistics and especially population data) from one zonation of a region to another, where the two zonations are geographically incompatible. This process is inevitably imprecise and is subject to a number of possible errors depending on the assumptions inherent in the methods used. Previous analysts have had only limited information with which to compare the results of interpolation and so assess the errors. In this paper a Monte Carlo simulation method based on modifiable areal units is employed. This allows multiple interpolations of population to be conducted from a single set of source zones to numerous sets of target zones. The properties of the full error distribution associated with a particular interpolation model can then be examined. The method based on dasymetric mapping consistently gave the highest accuracy of those tested, whereas the areal weighting method gave the lowest. More important than the results presented is the potential for future testing of other methods in increasingly complex situations.

Introduction

Socioeconomic data are generally published at some level of spatial aggregation in an endeavour to reduce the volume of published data and to preserve the confidentiality of the subjects. In the British census, for example, the smallest unit of spatial aggregation is called the enumeration district (ED): a zone that contains about 150 households (Rhind, 1991). The same term is used in rural areas of the USA, although the smallest unit employed within metropolitan areas is called the block (USBC, 1982). Whatever they are called, the increasing demand for sophisticated statistical and demographic analysis often requires the information they carry to be integrated with data derived from other sources. However, it is often found that the geographical boundaries used in census are incompatible with those employed during the collection of other data sets (Flowerdew and Openshaw, 1987). Alternative zonal systems can be derived both from the built environment (for example, postcodes: see Raper et al, 1992) and from the natural environment (for example, watersheds; see Goodchild et al, 1993). The ability to integrate data that have been collected under incompatible zonal systems is listed by Fotheringham and Rogerson (1993) as one of the pressing needs in spatial analysis. 'Areal interpolation' is the term given to the process of taking statistical information for one zonation of an area and converting it to give an estimate of that statistic for another incompatible zonation of the same area (Goodchild and Lam, 1980).

A large number of statistical methods have been suggested to achieve this aim, as reviewed by Goodchild et al (1993; see also Flowerdew and Green, 1989; 1991; Flowerdew et al, 1991; Goodchild and Lam, 1980; Lam, 1983; Langford et al, 1991; Tobler, 1979). No method is truly satisfactory, in the sense that it is impossible to derive perfect results, and thus some amount of error is inevitably introduced. In some of these studies only a method is proposed (for example, Tobler, 1979), whereas in others the method is also applied in a trial situation, and the results

compared either with known values (Flowerdew and Green, 1989; 1991; Flowerdew et al, 1991; Goodchild et al, 1993) or with other estimates (Langford et al, 1991). Testing against a single set of known values can only ever give a measure of the accuracy obtained in one particular geographical situation, and it says very little about the global applicability of the method. One of the most pressing problems in areal interpolation is to establish a methodology for evaluating errors in a systematic study and by repeated analysis.

In this paper we advocate the advantageous use of the modifiable areal unit, which is otherwise often regarded as a 'problem' generated by the spatial zonation of socioeconomic space. This allows the generation of multiple test situations for areal interpolation methods for the same geographical space and the same data. It then becomes possible to derive a full error distribution from which precise statements on error levels can be made. The basic methodology adopted is that of Monte Carlo simulation, and a first set of results is reported. The work relates solely to the spatial interpolation of count data from one set of spatial units to another ('type 11b'; see Flowerdew and Openshaw, 1987). The extension to further types of areal interpolation is a matter for future research. Parts of the work reported here are discussed in another paper (Langford et al, 1993).

Some of the past work in areal interpolation is reviewed in the next section of this paper, and we identify the methods that have been tested in the work reported here. The Monte Carlo simulation procedures are then outlined, along with the error measures that have been derived. In the penultimate section the results from the first application of the simulation procedure are given. These are based on data derived from a part of Leicestershire, England. There follows a discussion, which includes some possible avenues for further work.

Review of past work

The recent and widespread use of geographical information systems (GISs) has spurred research in areal interpolation. The fundamental problem of areal interpolation is perhaps most comprehensively discussed by Flowerdew and Openshaw (1987). Earlier reviews are presented by Goodchild and Lam (1980) and by Lam (1983); Goodchild et al (1993) give a more recent examination of the issue. Three major types of interpolation are identified here: cartographic, regression, and surface methods.

Cartographic methods

The simplest method of areal interpolation is undoubtedly the areal weighting method. Here, the population in each source unit is assumed to be evenly distributed across the area of that unit, and so the population density is estimated simply by dividing the population of the unit by its area. Areal interpolation is achieved by first overlaying the target units on the source units and determining the areas of intersection. Last, the populations of the target units are derived from the sum of the component portions of the source unit population:

$$\hat{P}_t = \sum \frac{A_{ts} P_s}{A_s}, \tag{1}$$

where $\hat{P}_t$ is the estimated population of the target zone, t; A_{ts} is the area of overlap of a source zone, s, with the target zone, t; P_s is the population of the source zone, s; and A_s is the area of the source zone, s. This method is simple to implement. It has no data requirements beyond the boundaries of the source and target units, and the populations in the source. Furthermore, the functionality it requires is present in almost every GIS on the market.

The major problem with this method is that it is incorrect to assume that the density of population within the source units is uniform. It is the same conceptual problem that is met with when using the choropleth map in cartography, and Wright (1936) proposed a cartographic alternative. The so-called dasymetric map uses knowledge of the locality to identify areas within zones that have different population densities, and so allows refinement of the assumption of an even distribution. Monmonier and Schnell (1984) show how the identification of areas of residential occupation in classified satellite imagery (Landsat) may be used to apply the method and thus allow a more realistic mapping of population density in, for instance, Pennsylvania. Flowerdew and Openshaw (1987) identify the possibility of using the dasymetric method, but as far as we are aware it has not been applied in any published research on areal interpolation.

Regression methods

Various workers have used a number of variants of regression models. These take the following general form:

$$\hat{P}_t = \mathrm{f}(x_1, x_2, ..., x_n)\,, \tag{2}$$

where $x_1, ..., x_n$ are control variables related to zone t. The appropriateness of the areas of different land-cover types as the control variables has been examined (Langford et al, 1991). Three different linear regression equations were developed, all based on land-use areas within polygons, with the regression line being made to pass through the origin of the graph (a zone with zero area has zero population). The area of land use within the source areas was collected, and then regression equations relating this to population were solved. It should be noted that no corrections were made to ensure that the populations reported for target zones were constrained to match the overall sum of the source units (the pycnophylactic property).

In the first so-called 'shotgun' model the areas of five different land-cover classes were used as independent variables to predict the population. The 'focused' model examined only the areas of high-density and low-density residential land use within the source units, and in the 'simple' model all residential land was regarded as one category. These methods were applied to the western part of Leicestershire, interpolating onto a 1 km grid.

Flowerdew (1988), Flowerdew and Green (1989; 1991), and Flowerdew et al (1991) have suggested that the appropriate regression model for the areal interpolation of population is not linear, but that a Poisson error distribution should be used. They suggested that other demographic variables may act as surrogates of population, and so be control variables for equation (2). Specifically, they examined the areal interpolation of the population of local government districts (census data) to parliamentary constituencies in Lancashire, using a variety of variables in the targets as indicators, including voting outcome, number of voters, cars per household, and people originating from the New Commonwealth and Pakistan. Using an iterative model-fitting algorithm in which the pycnophylactic criterion is maintained, they showed that simple models perform better than complex models. It was found that the proportion of people living in a situation with more than one person per room provided the best basis for estimation. Other experiments (Langford et al, 1991) suggest that the Poisson error model makes very little difference to the parameters of the simple regression models used, although with more complex models the differences between the coefficients increased markedly.

Goodchild et al (1993) suggest a number of further regression-based methods of areal interpolation in estimating the populations of drainage basins in California.

Five different methods were used as well as the areal weighting method and pycnophylactic surface (see next section). The primary conclusion from the work was the power of control zones in improving the interpolation process, where control zones were taken to be areas of even population distribution. They divided California into four zones known to have different population densities, being the central valley, the two major metropolitan areas, and the rest. When control zones were used a considerable improvement in estimates was reported. Flowerdew et al (1991) also used control zones, by taking a two-way split of Lancashire, making the split above and below the 400′ contour. This method showed no improvement over the Poisson regression method, however, but that is hardly surprising because of the simplistic use of a single-contour threshold as a controlling factor in population distribution. The value of control zones is assumed to be the same as in the cartographic dasymetric method discussed above.

Surface methods

A final group of methods has been proposed which are based on the mathematical assertion that, essentially, population density should be viewed as a continuously varying probability distribution. The major step with these methods is to define the distribution surface and, particularly, to use area-based statistics to approximate the surface. Once the distribution has been defined, integrating the volume under the surface gives one the population within any target zone. Tobler's (1979) widely quoted smooth pycnophylactic interpolation is the foremost such method. It minimises curvature on the surface and constrains the surface to zero population at the zone edges. Another method is based on use of the source zone centroids and a spreading function (Bracken, 1994; Bracken and Martin, 1989; Martin, 1989). From the published information, however, it is impossible to be precise about the properties of the surface generated. The method creates a dramatic visualisation of the population density distribution, but it is not clear that it will give a good areal interpolation and could benefit from explicit testing. It should be noted that the cartographic methods discussed above are really special cases of surface methods, where the surface is not seen to be continuous, but to have abrupt changes of population density.

Error analysis

Among studies of areal interpolation very few authors have addressed the reliability of the methods. Flowerdew (1988), Flowerdew and Green (1989; 1991), Flowerdew et al (1991), and Goodchild et al (1993) have compared the interpolated values for target zones with the actual values that were known from independent data sources. In so doing they make a single comparison for any particular interpolation. There has also been an attempt to validate their methods by examining the overall population patterns (Langford et al, 1991). The results of interpolating 1981 Census data onto a 1 km grid were compared with the values reported in the 1971 Census (as, unfortunately, grid cell values were not available in the 1981 Census). In summary, numerous methods have been proposed for areal interpolation, but very few analyses of the errors inherent in the process have been conducted. Those analyses which have been executed are all based on single interpolations and are therefore very limited statements as to the reliability of the methods concerned.

Monte Carlo simulation

The principles of Monte Carlo simulation have been widely exploited in the literature. Studies of point patterns by Besag and Diggle (1977) and migration routes by Hope (1968) are now well known, and Hagerstrand's (1965; Sechrist, 1992) diffusion

model may be implemented by Monte Carlo simulation. More recently, Openshaw et al (1987) used it as a fundamental element of their geographical analysis machine in identifying clusters of Leukaemia victims, and Openshaw et al (1991) examined the sensitivity of route selection for nuclear waste transport, again by Monte Carlo simulation. In the area of natural resources, Fisher (1991a) estimated by Monte Carlo simulation the effects of soil map errors on land valuation for taxation. Lee et al (1992) have examined the errors incurred in extracting flood plains. Of immediate relevance to the research reported here are the studies by Openshaw (1984), Openshaw and Taylor (1979), Fotheringham (1989), Fotheringham and Wong (1991), and Amrhein and Flowerdew (1992) on the modifiable area unit problem and the effects of aggregation on area-based demographic statistics.

In essence, Monte Carlo simulation is based on the idea that, if there is insufficient knowledge of the processes operating in a particular situation such that it is impossible to develop a process error to predict the outcome, then given that some set of outcomes is known it is possible to use the statistical summary of those outcomes to determine random new values that conform to the distribution of observed values. One of the principles of Monte Carlo simulation is that a single realisation of the simulation process, yielding as it does a single value or outcome, is a statistical quirk, and that conclusions can only be reached from repeated realisations.

For example, Fisher (1991b; 1992) used the root mean squared error reported for a digital elevation model (DEM) to create alternative realisations of the elevation model, and so derive alternative realisations of the viewshed that may be determined mathematically from the DEM. In doing so he was able to establish the effects of elevation error on the viewshed for which no process-based formula exists. Similarly, the effects of the modifiable areal unit problem and spatial aggregation have been widely studied by Monte Carlo simulation.

It is quite possible to formulate the areal interpolation problem as a Monte Carlo process amenable to extensive sensitivity analysis. Given that users outside the census organisation are never going to have the point information of population to work with, we must start with the existing zonation. We will call this the elemental zonation of n zones. A purely fictional example is illustrated in figure 1(a). It does not matter at what actual geographical level in a hierarchical zonation system it exists; for the present discussion it is simply taken as being elemental. By some randomisation algorithm such as that given by Openshaw (1977) the elemental zones may be fused into k sets, l_k, of m aggregated zones [figures 1 (b)–(d)]. As a whole suite of demographic variables is known for the elemental zones, they are also known for all aggregate zones, because they are merely the sum of the values for the elemental zones within each aggregate zone.

The two zonations l_1 [for example, figure 1(b)] and l_2 [for example, figure 1(c)] may now form the source and target zones, respectively, of an areal interpolation. The actual population of all elemental zones is known, and therefore it is also known for all aggregated zones. Thus, the estimated population in the aggregated zones of zonation l_2 may be compared with their actual populations, and a precise error term can be established. By changing either the source or target zones to zonation l_3 [for example, figure 1(d)], and repeating the interpolation, it is possible to start to build up a picture of the distribution of errors in the estimation process. By increasing the number of times the zonation is changed, very precise statements can be made of the error characteristics of a particular estimation process in a particular geographical situation. By increasing the number of test geographical locations, it may also be possible to determine the relative performance of any number of different areal interpolation methods.

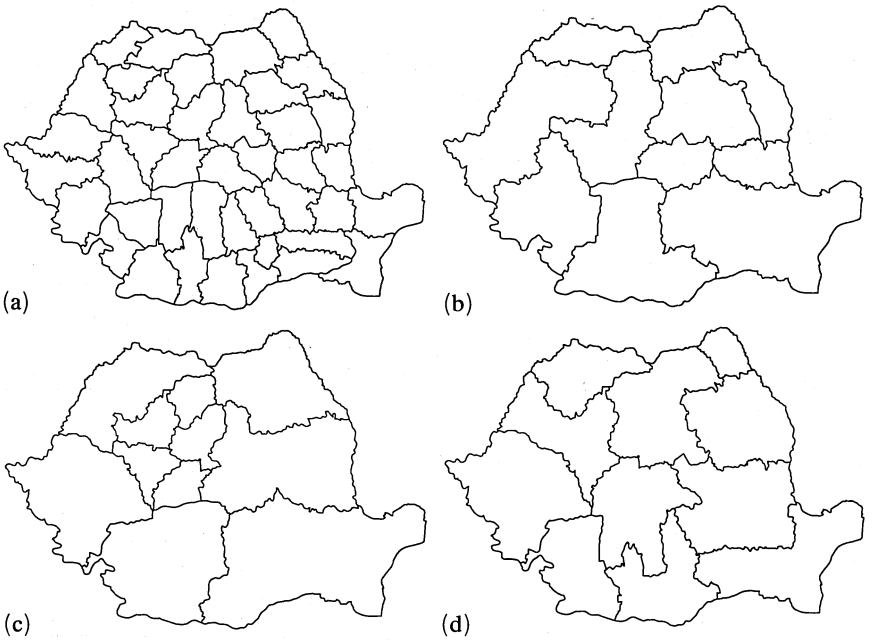

Figure 1. (a) The elemental zones in an imaginary region, and three different realisations of the random aggregation process: (b) set l_1, (c) set l_2, and (d) set l_3.

An experiment

Methods

Areal interpolation is usually a concern within countries and most often within small areas, although there is no actual reason why it should not be conducted across international boundaries, for example. Therefore, the elemental zones used in a Monte Carlo simulation process could be taken to be any subset of the whole globe. For experimental purposes here, and to confine the problem, an implementation of the method is proved on a small area of central England, taking the EDs of the 1981 Census as the elemental zones and agglomerating these to approximately the ward level. Wards are the official term for the next hierarchical agglomeration of EDs, and as such are just one realisation of the modifiable areal unit problem. In the experiment that follows the wards are taken as a fixed set of source zones. Monte Carlo simulation was used to generate random zonations of the EDs by using the algorithm of Openshaw (1977) which ensures that only neighbouring zones are merged.

Three districts (the hierarchical agglomeration of wards in the UK census) of Leicestershire, namely Charnwood, Leicester, and Oadby and Wigston were used for the experiment. Each is listed with some summary statistical information in table 1 and illustrated in figure 2 where the main urban areas are also shown. Charnwood is primarily a rural landscape with one small city, Loughborough, and a number of small towns and villages; Leicester district is almost entirely coincident with the city of Leicester; and Oadby and Wigston is a suburban area to the immediate south and east of Leicester city (see table 2). For experimental purposes the three districts are treated separately and together to give four different test situations.

The wards of the census were used for the source zones. The number of target zones in each set of aggregated zones was varied to give m = 5, 10, and 15 zones in the single-district examples, and m = 10, 50, and 100 zones in the combined case.

In each test, 250 different sets of aggregated zones of the elemental zones were found (that is, $k = 250$).

The ED outlines were digitised (with permission from the Ordnance Survey), and georegistered to a classified Landsat image. The image and classification used were the same as reported in an earlier paper (Langford et al, 1991), and the occurrence of the five different cover types recognised in each of the three districts are shown in table 2.

Table 1. Summary statistics of the three districts of Leicestershire used in the tests.

District	Number of wards (source zones)	Number of EDs	Area (km^2)	Total population	Maximum ED population density (persons per km^2)[a]
Oadby and Wigston	10	92	24	50673	25550
Leicester	16	586	73	276228	52500
Charnwood	23	307	280	132871	15550
Combined	49	985	817	459772	52500

[a] The minimum population densities with enumeration districts (EDs) in all districts is zero; such low densities are caused by institutional buildings, such as universities, being separate EDs with no permanent population.

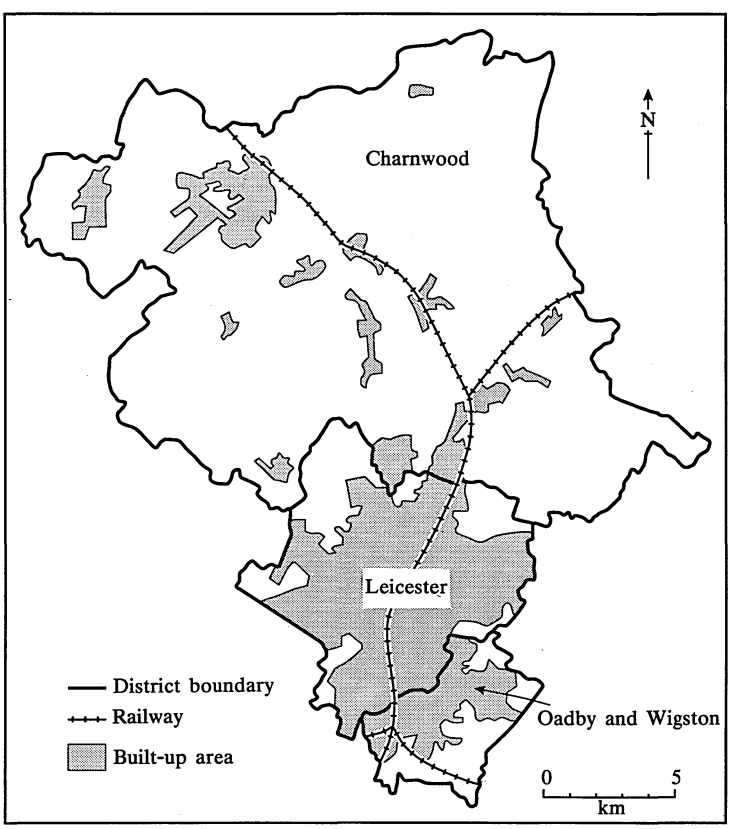

Figure 2. The study area, showing the three census districts, and the major population centres.

Five different areal interpolation methods were used: the three regression models briefly described above—shotgun, focused, and simple (Langford et al, 1991)—and the two cartographic methods—areal weighting and dasymetric mapping. The dasymetric and regression methods all relied on the classified Landsat image. In the shotgun multiple regression model the areas of five land-cover types within a source zone were independent variables for the dependent variable, the population. The focused model used only the two residential land-use classes as independent variables, and for the simple model the two residential covers were combined to give a binary map of residential and nonresidential areas. Populations were found for the target areas by solving the regression equations with areas of different use types within the different target zones. In the dasymetric method, the two residential land-use types were combined as a single populated land use, and the remaining land uses were grouped as unpopulated areas.

The major differences in the methods examined are that the regression methods are fitted globally, whereas the cartographic methods are local. Furthermore, the dasymetric method is constructed by using control zones, whereas the areal weighting method uses only the source zone.

All analysis was executed on a 486 personal computer using the raster-based Idrisi GIS package (Eastman, 1989) with some additional programming for this project. The approach could be implemented within a vector GIS, but it is easier to execute in raster mode, and the use of raster processing of population-related data is espoused by Martin and Bracken (1991). The selection of a raster cell size of 30 m × 30 m made it simple to integrate the satellite image of land use. Also, it was sufficiently small, relative to the size of EDs, to avoid complications arising from the rasterisation process.

Table 2. Land-cover types in each of the three districts.

Land-cover type	Oadby and Wigston		Leicester		Charnwood	
	percentage	area (km^2)	percentage	area (km^2)	percentage	area (km^2)
Industrial	8.75	2.0	17.77	13.0	3.29	9.2
Dense residential	0.44	0.1	8.20	6.0	0.29	0.8
Residential	39.37	9.0	43.75	32.0	10.41	29.1
Agriculture	49.87	11.4	28.71	21.0	81.87	228.9
No population	1.57	0.36	1.56	1.14	4.15	11.6

Error measures

The root mean square error, E^{RMS}, for one realisation of the aggregation process to m target aggregations is given as follows:

$$E^{\mathrm{RMS}} = \left[\frac{1}{m} \sum_{j=1}^{m} (P_t - \hat{P}_t)^2 \right]^{1/2} . \tag{3}$$

This value is highly dependent upon the magnitude of the mean population in the target zones, itself a reflection of the number of target zones. Therefore standardisation against that mean [see equation (4) below] gives a basis for comparison between experimental situations and an equivalent of the coefficient of variation, V:

$$V = \frac{1}{k\bar{P}_t} \sum_{i=1}^{k} E_i^{\mathrm{RMS}} , \tag{4}$$

where $\bar{P}_t$ is the mean of actual target populations in any experimental situation, and with k individual aggregated zones.

Results

The results of analysis are summarised in table 3(a)-3(d), and some illustrative histograms of the three-district case are shown in figures 3-5. In table 3, in every case the accuracy of the areal interpolation decreases as the number of target zones increases, with the number of source zones held constant. This is to be expected, and essentially says that the finer the spatial resolution, the less accurate the estimate.

The general trend in accuracy between the interpolation methods is very consistent. The least accurate method is usually the areal weighting method. One exception is in the Leicester district when estimating to 5 and 10 target zones. In these cases the simple regression method gives the worst result [table 3(b)]. Another exception is in the estimates for the three combined districts for 5 target zones, where all three regression methods gave worse results [table 3(d)]. It is believed that this says more about the problems with the regression models than the advantages of the areal weighting method. By contrast, the dasymetric method is always the most accurate.

Table 3. Results from (a) Oadby and Wigston district, (b) Leicester district, (c) Charnwood district, and (d) the three districts combined.

Number of target zones	Average population	Method				
		areal weighting	shotgun	focused	simple	dasymetric
(a) Oadby and Wigston district						
5	10101.2	2393.8	2228.3	1179.7	1173.7	422.88
		0.24	0.22	0.11	0.11	0.04
10	5106.4	2125.5	1185.4	836.6	834.8	400.4
		0.41	0.23	0.16	0.16	0.08
15	3412.5	1816.1	838.2	667.2	665.8	383.6
		0.53	0.24	0.20	0.19	0.11
(b) Leicester district						
5	55245.0	4933.3	4917.4	5109.4	6807.9	2880.0
		0.09	0.09	0.09	0.12	0.05
10	27622.8	4355.7	4093.5	4115.2	4794.9	2722.7
		0.16	0.15	0.15	0.17	0.09
15	18415.2	4501.2	3454.9	3430.0	3852.4	2497.3
		0.24	0.19	0.19	0.21	0.13
(c) Charnwood district						
5	26573.7	4007.3	2488.0	2396.3	2457.7	1206.0
		0.15	0.09	0.09	0.09	0.04
10	13287.1	3884.5	1673.2	1888.4	1928.8	1178.9
		0.29	0.13	0.14	0.14	0.09
15	8858.1	3613.9	1375.3	1597.4	1621.5	1110.4
		0.41	0.16	0.18	0.18	0.13
(d) The three districts combined						
10	45977.9	5758.6	6342.9	8605.5	11015.4	2639.3
		0.13	0.14	0.19	0.24	0.06
50	9198.6	3575.5	2411.6	2789.4	3201.2	1629.9
		0.39	0.26	0.30	0.35	0.18
100	4601.2	2707.8	1558.6	1737.7	1912.3	1233.8
		0.59	0.34	0.38	0.42	0.27

Note: the number of realisations of the randomisation process is 250; in each cell in the table the upper value is the mean root mean square [equation (3)], and the lower value is standardised against the average population [equation (4)].

The regression models are fitted globally and so they are bound to give a poorer estimate than the locally fitted (that is, dasymetric) method. Within the regression methods, the pattern is less clear than among the cartographic methods. In all areas, except Oadby and Wigston district [table 3(a)], the simple regression method gives the worst result, and the shotgun method commonly gives the best. In several cases, however, the improvement from the focused method to the shotgun method is extremely small, meaning that the complex regression models are unnecessary. On the other hand, in Charnwood district [table 3(c)] when interpolating to the smallest number of target zones, all regression methods yield equally accurate results, and in Leicester district [table 3(b)] the focused and shotgun methods are equally accurate for all numbers of target zones. Therefore, there seems to be little support for Flowerdew and Green's (1989; 1991) conclusion that simple regression models are better, although complex ones do not always yield useful improvements in the accuracy of the interpolation.

Error distributions

In figures 3 - 5 histograms of the three-district case are shown. They can be considered representative of all those examined for other cases. Each diagram is based on the root mean squared errors for each of the 250 different realisations of the aggregation process. The dasymetric and regression methods yield limited ranges of values with near-normal distributions and single modes or twin modes close together. The distribution of the errors associated with the areal weighting method

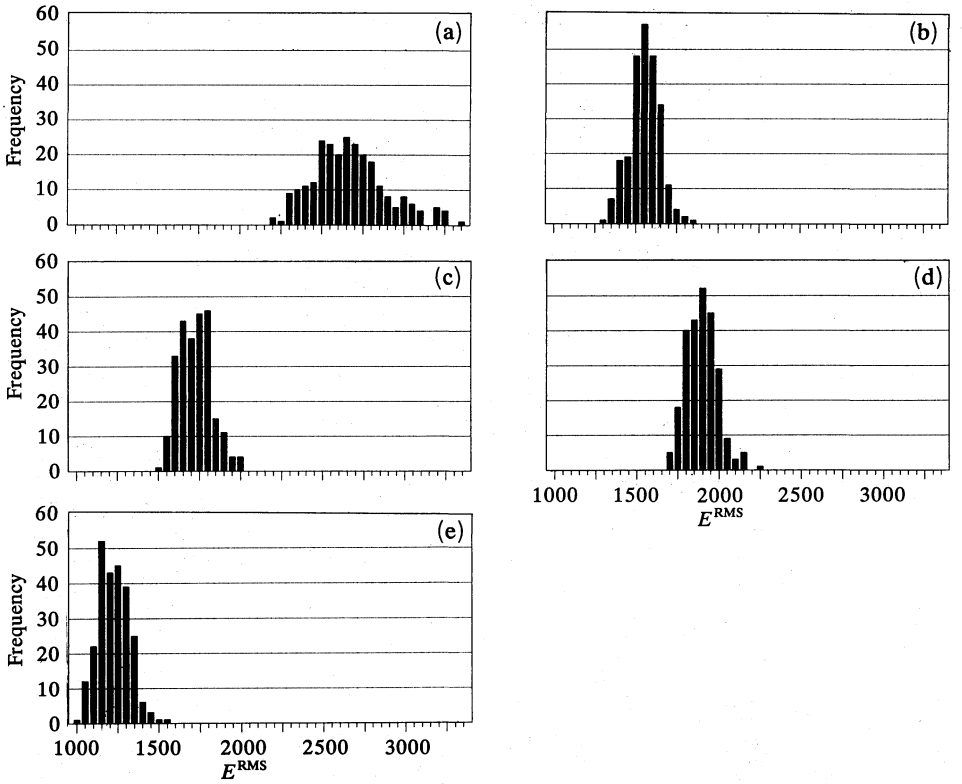

Figure 3. Histograms of the root mean squared errors, E^{RMS}, for (a) the areal weighting method, (b) the shotgun method, (c) the focused method, (d) the simple method, and (e) the dasymetric method, for the three districts combined when 100 target zones are used; 250 values of E^{RMS} are used to construct each diagram.

by contrast is rather dispersed and multimodal. In the 10-target case (figure 5) all histograms are more dispersed than in others, and less peaked.

The normality and general compactness of the regression methods are perhaps to be expected because of the mathematical process of regression analysis. The complexity of the curves of the areal weighting method is likewise to be expected. The compact and normal form of the dasymetric method is very reassuring, especially in the 10-target-zone case where even the regression methods are less well behaved, and confirms the advantages of this method over the others.

Summary

Above all, the experimental results presented here show that, of the methods examined, the dasymetric method gives the best estimate of the population in the target zones. This is the case in every situation tested. The method used here utilises a binary division of the land covers, and so is a special case of the full dasymetric mapping method suggested by Wright (1936). Commonly, the areal weighting method for interpolation is the worst, showing the lack of precision in the choroplethic assumptions in this method. This is not surprising because of the lack of local information other than the zone boundaries. The simple regression method is the worst of the second class of methods. Increasing the number of land-cover parameters improves the accuracy of the estimation by regression methods, but not greatly.

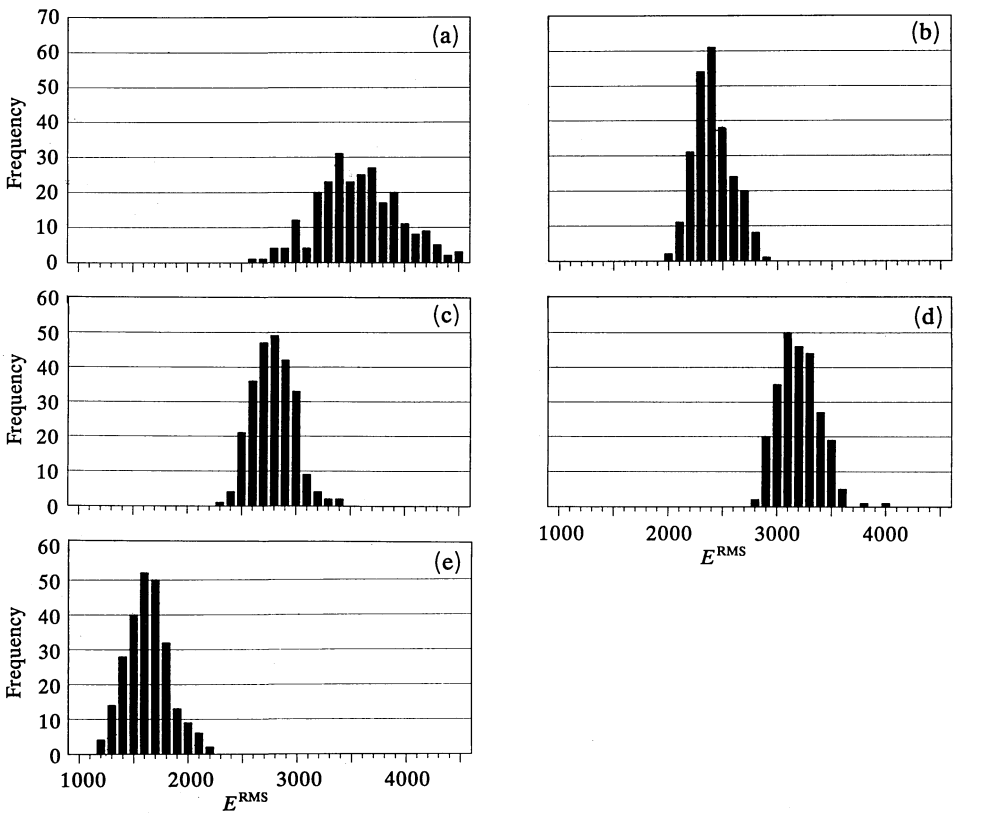

Figure 4. Histograms of the root mean squared errors, E^{RMS}, for (a) the areal weighting method, (b) the shotgun method, (c) the focused method, (d) the simple method, and (e) the dasymetric method, for the three districts combined when 50 target zones are used; 250 values of E^{RMS} are used to construct each diagram.

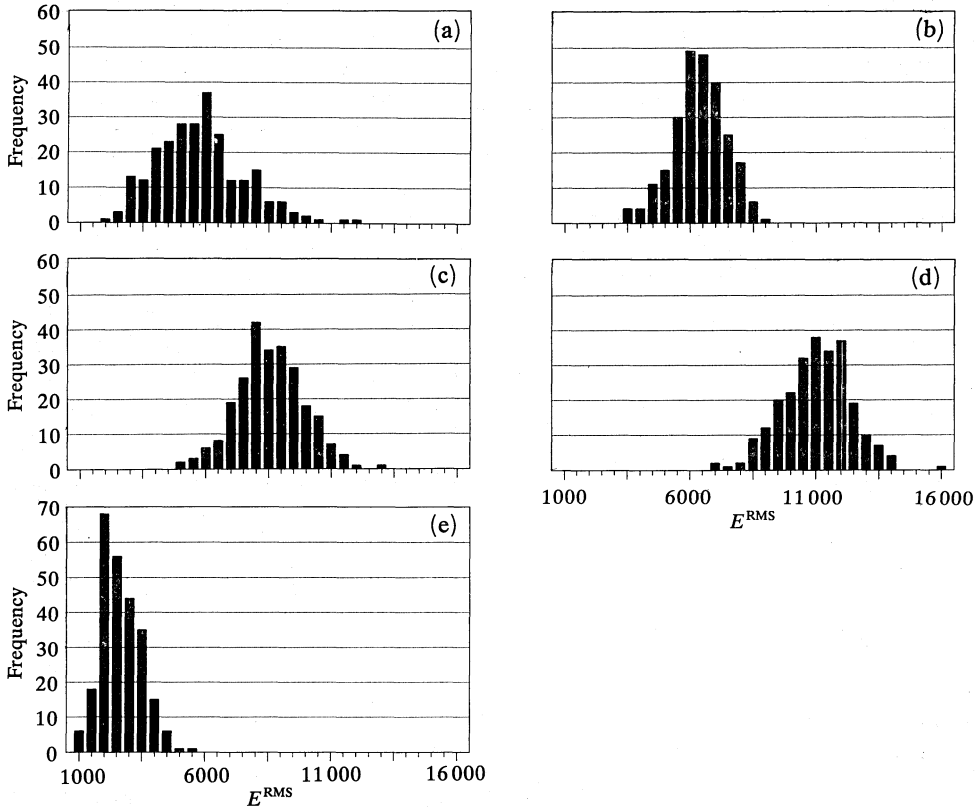

Figure 5. Histograms of the root mean squared errors, E^{RMS}, for (a) the areal weighting method, (b) the shotgun method, (c) the focused method, (d) the simple method, and (e) the dasymetric method, for the three districts combined when 10 target zones are used; 250 values of E^{RMS} are used to construct each diagram.

Conclusions

The work reported here is the first rigorous attempt to examine the errors inherent in the process of areal interpolation, and the results presented are very limited and subject to several shortcomings.

1. Only five methods are examined. At the least, the Poisson regression with the EM algorithm (Flowerdew and Green, 1989), the other regression models (Goodchild et al, 1993), and the surface methods (Bracken, 1994; Tobler 1979) should also be tested, so that not only statements of the relative accuracy may be made but also the relative computational costs may be measured.
2. In the work reported here, classified Landsat imagery has been used to construct both the regression model and the dasymetric model. The results of this analysis are impressive, but classified images are not without their own accuracy problems (for example, Campbell, 1987, chapter 12). There is a need to examine the effects of this accuracy on the estimation process, and to explore the potential of other data sources.
3. Last, the method has been applied only over a limited area. There is an obvious need to broaden the area of analysis so that valid estimates can be made of the error over large areas.

The basic approach of using modifiable area units generated by a Monte Carlo simulation to provide error estimates of areal interpolation is considered to be innovative, and it offers the chance to generate a systematic comparison of the interpolation methods that have been proposed. However, Monte Carlo simulation is a computationally intensive method for statistical analysis, which has severe limitations in the real-world analysis of errors (Openshaw et al, 1987). The method outlined would allow the parameterisation of a number of predictive formulae for different methods of areal interpolation, when sensitised to variations in a number of control variables. Thus application of this method is not the end in itself of the research. The method can enable the development of predictive equations for future use where Monte Carlo simulation would not be possible. Furthermore, it allows the development of a detailed set of recommendations as to when particular methods may be expected to perform best and permits the estimation of the relative accuracy of all methods of areal interpolation, including those which may be suggested in the future.

Acknowledgements. The authors would like to thank Doris Brencke, David Troughear, and Philip Jutson who all assisted with the work reported here at various stages and in different ways. The intellectual stimulation of the Midlands Regional Research Laboratory (MRRL) is acknowledged, and especially the comments of Alan Strachan, David Unwin, and David Walker. Carl Amrhein and Mike Goodchild both provided encouraging comments on the work. The work reported here has been conducted as a part of the Transition Funding awarded to the MRRL from the Economic and Social Research Council.

References

Amrhein C G, Flowerdew R, 1992, "The effect of data aggregation on a Poisson regression model of Canadian migration" *Environment and Planning A* **24** 1381 - 1391

Besag J, Diggle P J, 1977, "Simple Monte Carlo tests for spatial patterns" *Applied Statistics* **26** 327 - 333

Bracken I, 1994, "A surface model approach to the representation of population-related social indicators", in *Spatial Analysis and GIS* Eds S Fotheringham, P Rogerson (Taylor and Francis, London) pp 247 - 260

Bracken I, Martin D, 1989, "The generation of spatial population distributions from census centroid data" *Environment and Planning A* **21** 537 - 543

Campbell J B, 1987 *Introduction to Remote Sensing* (Guildford Press, New York)

Eastman R, 1989 *Idrisi User's Guide, Version 4* Idrisi Project, Clark University, Worcester, MA

Fisher P F, 1991a, "Modelling soil map-unit inclusions by Monte Carlo simulation" *International Journal of Geographical Information Systems* **5** 193 - 208

Fisher P F, 1991b, "First experiments in viewshed uncertainty: the accuracy of the viewable area" *Photogrammetric Engineering and Remote Sensing* **57** 1321 - 1327

Fisher P F, 1992, "First experiments in viewshed uncertainty: simulating the fuzzy viewshed" *Photogrammetric Engineering and Remote Sensing* **58** 345 - 352

Flowerdew R, 1988, "Statistical methods for areal interpolation: predicting count data from a binary variable", RR-15, Northern Regional Research Laboratory, University of Lancaster, Lancaster, and University of Newcastle upon Tyne, Newcastle upon Tyne

Flowerdew R, Green M, 1989, "Statistical methods for inference between incompatible zonal systems", in *The Accuracy of Spatial Databases* Eds M F Goodchild, S Gopal (Taylor and Francis, London) pp 239 - 247

Flowerdew R, Green M, 1991, "Data integration: statistical methods for transferring data between zonal systems", in *Handling Geographical Information: Methodology and Potential Applications* Eds I Masser, M Blakemore (Longman, Harlow, Essex) pp 38 - 54

Flowerdew R, Openshaw S, 1987, "A review of the problems of transferring data from one set of areal units to another incompatible set", RR4, Northern Regional Research Laboratory, University of Lancaster, Lancaster, and University of Newcastle upon Tyne, Newcastle upon Tyne

Flowerdew R, Green M, Kehris E, 1991, "Using areal interpolation methods in geographic information systems" *Papers in Regional Science* **70** 303 - 315

Fotheringham A S, 1989, "Scale-independent spatial analysis", in *The Accuracy of Spatial Databases* Eds M F Goodchild, S Gopal (Taylor and Francis, London) pp 221-228

Fotheringham A S, Rogerson P A, 1993, "GIS and spatial analytical problems" *International Journal of Geographical Information Systems* **7** 3-19

Fotheringham A S, Wong D W S, 1991, "The modifiable areal unit problem in multivariate statistical analysis" *Environment and Planning A* **23** 1025-1044

Goodchild M F, Lam N S-N, 1980, "Areal interpolation: a variant of the traditional spatial problem" *Geoprocessing* **1** 297-312

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

Hagerstrand T, 1965, "A Monte Carlo approach to diffusion" *European Journal of Sociology* **6** 43-67

Hope A C A, 1968, "A simplified Monte Carlo significance test procedure" *Journal of the Royal Statistical Society B* **30** 582-598

Lam N S-N, 1983, "Spatial interpolation methods: a review" *American Cartographer* **10** 129-149

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

Langford M, Fisher P F, Troughear D, 1993, "Comparative accuracy measurements of the cross areal interpolation of population", in *Proceedings of EGIS '93* (EGIS Foundation, Utrecht) pp 663-674

Lee J, Snyder P K, Fisher P F, 1992, "Modelling the effect of data errors on feature extraction from digital elevation models" *Photogrammetric Engineering and Remote Sensing* **57** 1321-1327

Martin D, 1989, "Mapping population data from zone centroid locations" *Transactions of the Institute of British Geographers: New Series* **14** 90-97

Martin D, Bracken I, 1991, "Techniques for modelling population-related raster databases" *Environment and Planning A* **23** 1069-1075

Monmonier M, Schnell G, 1984, "Land use and land cover data and the mapping of population density" *International Yearbook of Cartography* **24** 115-121

Openshaw S, 1977, "Algorithm 3: a procedure to generate pseudo-random aggregations of N zones into M zones, where M is less than N" *Environment and Planning A* **9** 169-184

Openshaw S, 1984 *The Modifiable Areal Unit Problem* CATMOG 38 (Geobooks, Norwich)

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

Openshaw S, Charlton M, Craver S, 1991, "Error propagation: a Monte Carlo simulation", in *Handling Geographical Information: Methodology and Potential Applications* Eds I Masser, M Blakemore (Longman, Harlow, Essex) pp 78-101

Openshaw S, Charlton M, Wymer C, Craft A, 1987, "A mark 1 geographical analysis machine for the automated analysis of point data sets" *International Journal of Geographical Information Systems* **1** 335-358

Raper J, Rhind D W, Shepherd J, 1992 *Postcodes: The New Geography* (Longman, Harlow, Essex)

Rhind D W, 1991, "Counting the people: the role of GIS", in *Geographical Information Systems, Volume 2: Principles and Applications* Eds D J Maguire, M F Goodchild, D W Rhind (Longman, Harlow, Essex) pp 127-137

Sechrist R P, 1992, "Simulation of the spatial diffusion process" *Computers and Geosciences* **18** 965-974

Tobler W, 1979, "Smooth pycnophylactic interpolation for geographical regions" *Journal of the American Statistical Association* **74** 519-530

USBC, 1982 *Guide to the 1980 Census of Population and Housing* US Department of Commerce, US Bureau of the Census, Washington, DC

Wright J K, 1936, "A method of mapping densities of population with Cape Cod as an example" *Geographical Review* **26** 103-110