Samantha Cockings, Peter F. Fisher and Mitchel Langford

Parameterization and Visualization of the Errors in Areal Interpolation

Areal interpolation involves the transfer of data from one zonation of a region to another, where the two zonations of space are geographically incompatible. By its very nature this process is fraught with errors. However, only recently have there been specific attempts to quantify these errors. Fisher and Langford (1995) employed Monte Carlo simulation methods, based on modifiable areal units, to compare the errors resulting from selected areal interpolation techniques. This paper builds on their work by parameterizing and visualizing the errors resulting from the areal weighting and dasymetric methods of areal interpolation. It provides the basis for further research by developing the methodology to produce predictive models of the errors in areal interpolation. Random aggregation techniques are employed to generate multiple sets of source zones and interpolation takes place from these units onto a fixed set of randomly generated target zones. Analysis takes place at the polygon, or target zone level, which enables detailed analysis of the error distributions, basic visualization of the spatial nature of the errors and predictive modeling of the errors based on parameters of the target zones. Correlation and regression analysis revealed that errors from the areal weighting technique were related to the geometric parameters of the target zones. The dasymetric errors, however, demonstrated more association with the population or attribute characteristics of the zones. The perimeter, total population, and population density of the target zones were shown to be the strongest predictive parameters.

The work presented in this paper focusses on the transfer of data from one zonation of a region to another, where the two zonations of space are geographically incompatible. This is known as *areal interpolation*. Although not new, areal interpolation has received renewed attention over recent years as rapid developments in Geographical Information Systems have enabled the integration of an overwhelming array of geographical data sets. These data sets are frequently collected, aggregated, and reported on unique spatial units, often of

Samantha Cockings is a research assistant in the Department of Geography at the University of Durham. Peter F. Fisher is senior lecturer and Mitchel Langford is lecturer at the Midlands Regional Research Laboratory, Department of Geography, University of Leicester.

Geographical Analysis, Vol. 29, No. 4 (October 1997)
Submitted: 12/17/96. Revised version accepted: 4/15/97.

convenience only to the collector, supplier, or user of the original data (Langford, Fisher, and Troughear 1993). Transferring data between these geographically incompatible spatial units is an inherently uncertain process. A huge literature exists concerning alternative methods of areal interpolation, with many representing an improvement on the simplest method of weighted overlay (Flowerdew and Green 1991; Lam 1983; Martin and Bracken 1991). However, few attempts have been made to measure the comparative accuracy of these methods, or to parameterize the errors resulting from them. Previous work (Fisher and Langford 1995) has employed Monte Carlo simulation to determine the comparative accuracy of areal interpolation techniques. The present paper builds on that work by attempting to produce predictive models of these errors.

1. APPROACH

Fisher and Langford's (1995) work employed Monte Carlo simulation based on modifiable areal units. This allowed multiple interpolations of population to be conducted from a single set of source zones onto numerous sets of target zones, the source zones being the spatial units for which data are available and the target zones those for which data are required. The properties of the full error distribution associated with a particular interpolation model could then be examined. The dasymetric method was found consistently to perform the best whilst the traditional areal weighting method was the least accurate of the five methods tested.

The primary aim of the work reported here is to produce the framework for deriving predictive models of the errors resulting from areal interpolation techniques. This requires three fundamental changes to the methodology previously employed by Fisher and Langford (1995). First, Monte Carlo simulation is employed to randomly aggregate the dataset into multiple sets of source zones (rather than a fixed set) and the population figures from these spatial units are then interpolated onto a constant set of target zones (as opposed to multiple sets). Second, there is a change of emphasis in the recording and analysis of errors. As well as recording summary error measures for the two methods of areal interpolation, the results are recorded at polygon (or target zone) level. This allows more in-depth analysis of the errors and basic visualization in the form of choropleth maps. Third, parameters likely to influence the degree of error in the interpolation are recorded for the target zones. This forms the basis of an attempt to derive predictive error models.

2. DATA AND METHODOLOGY

The area selected for the study comprised three census districts in Leicestershire, central England (Figure 1). The districts, namely Charnwood, Leicester, and Oadby and Wigston, were the same as those employed by Langford, Fisher, and Troughear (1993) and Fisher and Langford (1995). They were chosen because they represent three distinct landscapes. Charnwood is predominantly a rural district with one small city, Loughborough; Leicester district comprises the city of Leicester; and Oadby and Wigston is a suburban area to the southeast of the city. In addition, the availability of previously corrected and classified remotely sensed imagery of the area (from Langford, Maguire, and Unwin 1991), for use in the dasymetric method, made the area a logical choice.

The analysis was run on a 90 MHz Pentium, with most of the processing being carried out within the *Idrisi* GIS. The whole process was set up to be fully automatic through the use of a series of DOS batch files, employing a nesting

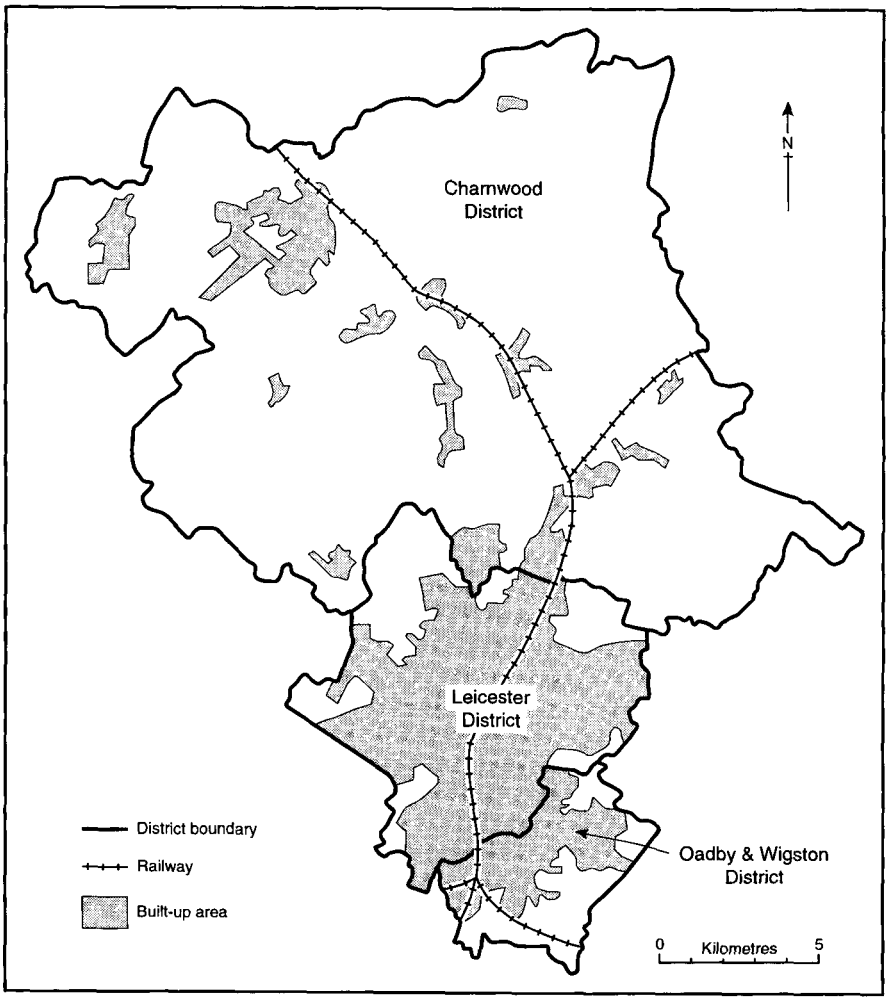

FIG. 1. Map of the Study Area Showing Major Urban Areas (Fisher and Langford 1995, p. 217)

structure to call the various programs and commands as many times as necessary. The elemental spatial units (1991 enumeration districts) were randomly aggregated to give two fixed sets of target zones of fifty and one hundred polygons, and 250 sets of fifty source zones. Figure 2 shows the fixed set of fifty target zones. The random aggregation program works by randomly selecting m (elemental) zones, where m is the number of aggregated zones required. These zones act as core zones for the aggregation and neighboring (elemental) zones are randomly selected to merge with these core zones (Openshaw 1977). The process is then repeated until all elemental zones are allocated.

A series of geometric and attribute parameters of the target zones was recorded to aid the parameterization of errors. The area, perimeter, and compactness ratio of each target zone were calculated and recorded. The compactness ratio compares a polygon's area to its perimeter and is therefore a measure of the polygon's shape. The attribute parameters recorded were the total population and population density of each target zone as a whole, and the range of population density within each zone. Consideration was given to including land

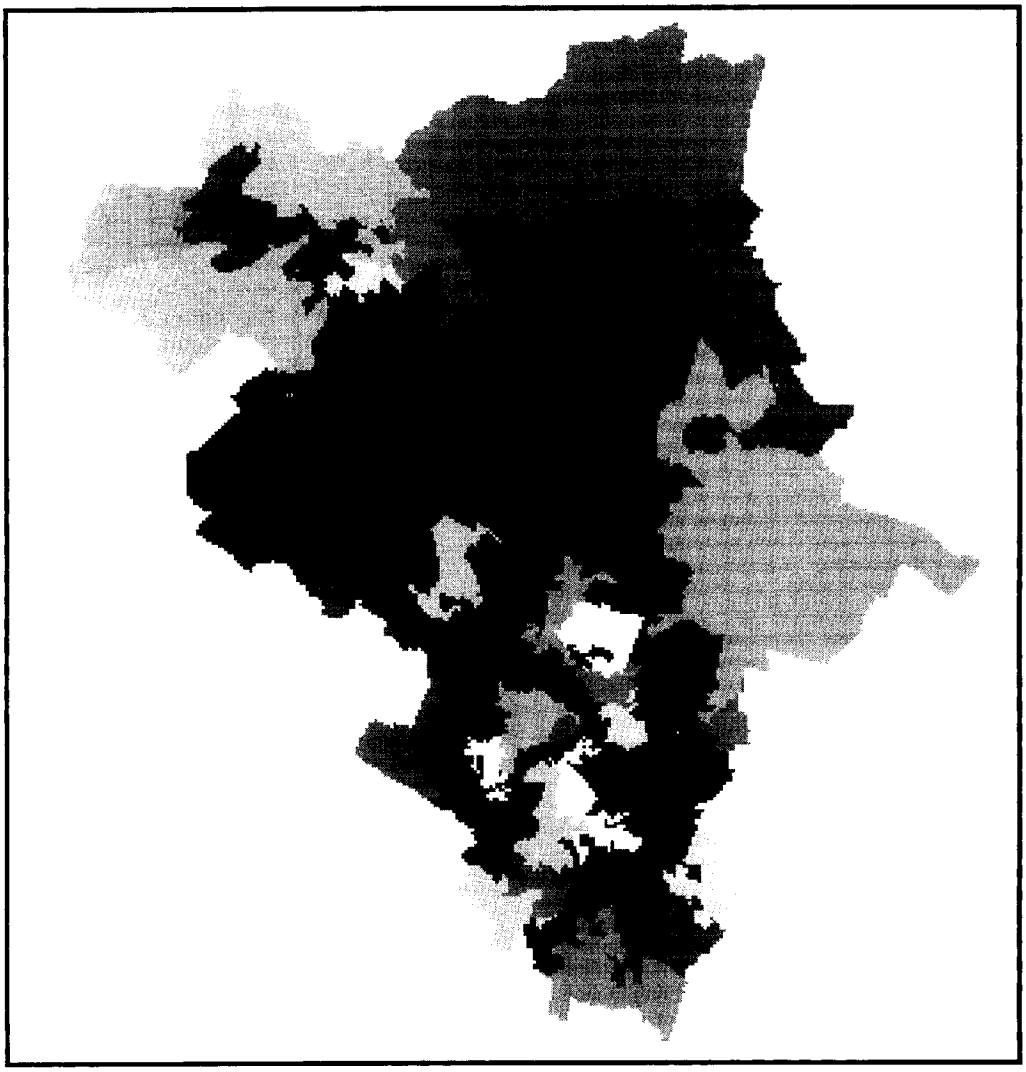

FIG. 2. Fixed Set of Fifty Target Zones

cover as a parameter; however, this was rejected because land cover is employed as a determining variable in the dasymetric process.

The population totals for the multiple source zones were then interpolated onto the fixed target zones using the areal weighting and dasymetric methods of areal interpolation, and the resulting population estimates were recorded. These methods represent the "best" and "worst" case scenarios according to previous work carried out by Langford, Fisher, and Troughear (1993) and Fisher and Langford (1995). The areal weighting method assumes a homogenous distribution of population across the spatial unit. Interpolation takes place by overlaying the target zones onto the source zones, finding the areas of intersection, and then summing the populations of the component parts of the source zones contained within each target zone (Lam 1983; Fisher and Langford 1995). The dasymetric method makes use of additional geographical infor-

mation about the distribution of population to inform the estimation (Wright 1936; Langford and Unwin 1994; and Langford, Maguire, and Unwin 1990). In this case, the extra information was derived from a classified Landsat TM image of the study area.

Various measures of the errors from interpolation were calculated. The difference between the estimated and real population value (calculated by summing the population of the elemental zones contained within each target zone) for each polygon in each run was recorded and the mean error was determined for each polygon. The standard deviation of the estimates for each polygon was recorded to give an indication of their accuracy or reliability. Correlation, simple regression, and multiple regression analyses were undertaken to identify any relationships among the parameters and error measures. Basic visualization of the error measures at target zone level, in the form of choropleth maps, was employed to indicate the spatial characteristics of the error distributions.

3. RESULTS

3.1 Mean Errors

Figure 3 shows the mean error for each target zone in the fifty-target-zone experiment, based on 250 runs. The mean error is calculated by using equation (1).

$$ME_t = \frac{\sum_{i=1}^{k} (\hat{P}_{ti} - P_t)}{k} \tag{1}$$

where ME_t is the mean error of target zone t; k is the number of runs; $\hat{P}_{ti}$ is the population estimate for target zone t for run i; and P_t is the real population value of target zone t. A positive mean error therefore represents an overestimate of population for the zone, whereas a negative value indicates an underestimate.

In terms of the magnitude of errors, as would be expected, the dasymetric estimate is consistently closer to the real population value for the zone than is the areal weighting estimate. However, Figure 3 shows that for some zones the mean errors for the two methods of interpolation are very different. This sort of zone-level analysis allows detailed comparisons to be made. For example, for zone 25 in Figure 3 (zone numbers are plotted on the x-axis), the areal weighting technique gives a large underestimate as opposed to a small overestimate by the dasymetric method. Further analysis shows this zone to have a large perimeter and low compactness ratio, indicating an irregularly shaped polygon. It is also in a sparsely populated region. These characteristics produce exactly the sort of conditions in which the areal weighting technique is believed to perform poorly.

In the one-hundred-target-zone experiment, the areal weighting estimate is less accurate than the dasymetric equivalent for 84 percent of the zones. Zones with especially large overestimates by both methods tend to have large areas, perimeters, and high total population. However, their population density tends to be well below average, indicating sparsely populated regions.

3.2 Standard Deviation of the Estimates

The standard deviation of the estimates for each target zone was recorded to investigate the consistency or reliability of estimation by each method of interpolation. The standard deviation of the areal weighting estimates is greater than

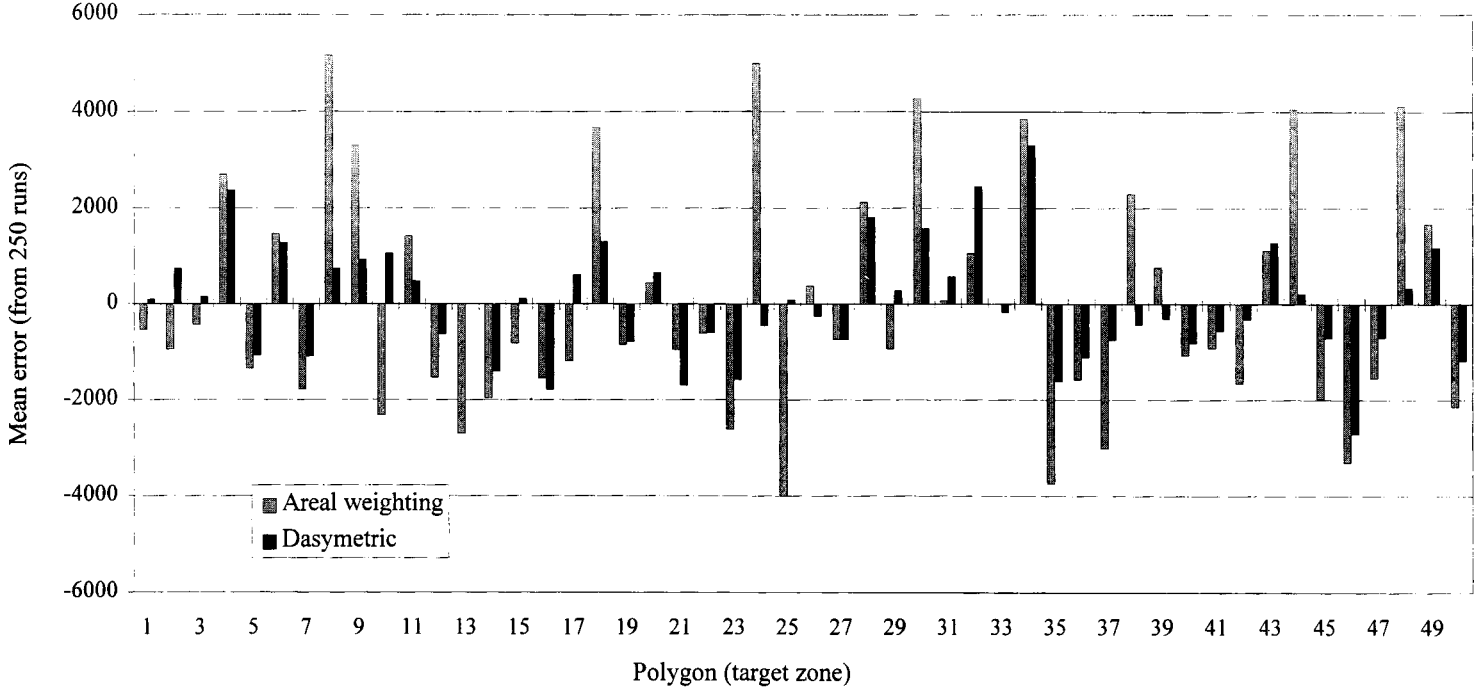

FIG. 3. Mean Errors (fifty target zones)

TABLE 1
Correlation Coefficients and Significance Values for Mean Error and Parameters

	50 target zones				100 target zones			
	Areal weighting		Dasymetric		Areal weighting		Dasymetric	
Area	0.595	*<0.001*	0.225	*0.116*	0.631	*<0.001*	0.405	*<0.001*
Perimeter	0.686	*<0.001*	0.360	*0.010*	0.621	*<0.001*	0.439	*<0.001*
Population density	−0.603	*<0.001*	−0.530	*<0.001*	−0.381	*<0.001*	−0.471	*<0.001*
Minimum Population density	−0.555	*<0.001*	−0.540	*<0.001*	−0.286	*0.004*	−0.361	*<0.001*

NOTE: All values to three decimal places; values in plain type are product moment correlation coefficients; values in italics are significance values of the *t*-statistics resulting from regression analysis.

the dasymetric equivalent for every target zone in the fifty-zone experiment, indicating a less reliable result. In many cases the difference between the two methods' estimates is quite considerable.

Similar trends are observed in the one-hundred-target-zone experiment, where the dasymetric method produces a lower standard deviation of estimates than does the areal weighting in all but two of the zones. The estimates for these two zones have low standard deviations by both methods of interpolation. Investigation of the parameters of these zones shows them to be very small but densely populated areas. By way of contrast, the largest standard deviations from the areal weighting technique are recorded for zones with large, sparse populations. These zones are the same as those with the largest mean error.

3.3 Correlation and Regression Analysis

Correlation coefficients were calculated between the parameters and mean error for each target zone, and between the parameters and standard deviation of estimates. Scatterplots were constructed to visualize the distribution of the errors, and backward stepwise multiple regression was employed in an attempt to develop predictive equations of the rates of error.

Mean Error. The correlation coefficients and significance of the *t*-statistics from regression analysis are presented in Table 1. A *p*-value of less than 0.05 was considered to be a significant result. Figure 4 presents a sample scatterplot of population density against mean error for the fifty target zones.

Table 1 suggests that all of the parameters shown (except for area in the fifty-zone dasymetric experiment) play a significant role in determining the mean error. The number of zones in the experiment also appears to have an effect on the strength of correlation. For the areal weighting technique, area and perimeter are strongly correlated with the mean error in both the fifty- and one-hundred-zone experiments. Population density is also significant in both, but the degree of correlation is lower for the one-hundred-target-zone experiment. For the dasymetric technique, in both the fifty- and one-hundred-zone experiments, population density exhibits strong correlation with the mean error, as does minimum population density in the fifty-zone case. In every case, the area and perimeter of the target zones show positive correlation with the mean difference, whilst population density and minimum population density have negative correlation.

It is unlikely that the parameters are mutually exclusive in their effect on the errors in areal interpolation. In most cases, the error for a target zone will be dependent on a range of parameters. The relative influence of the most important parameters was therefore evaluated by using backward stepwise multiple regression analysis. In this case, the dependent variable is the level of error

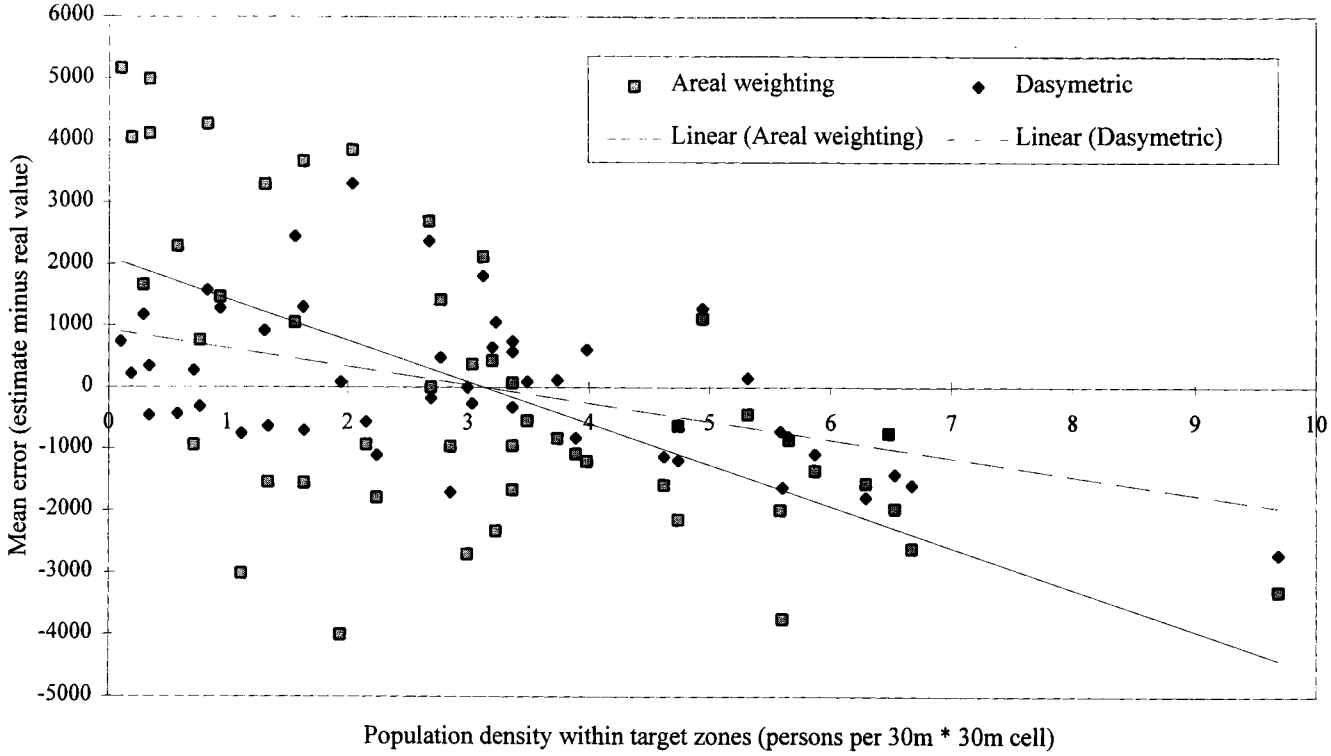

FIG. 4. Population Density against Mean Error (fifty target zones)

TABLE 2
Stepwise Regression for Mean Errors from Areal Weighting Technique

Zonation	Multiple *R*	*R*-squared	Coefficients	*P*-value
50 zones	0.703	0.494	Intercept	0.057
			Perimeter	<0.001
100 zones	0.720	0.519	Intercept	0.320
			Perimeter	<0.001
			Total population	<0.001
			Density range	0.006
			Minimum population density	0.007
			Maximum population density	0.006

TABLE 3
Stepwise Regression for Mean Errors from Dasymetric Technique

Zonation	Multiple *R*	*R*-squared	Coefficients	*P*-value
50 zones	0.530	0.281	Intercept	0.001
			Population density	<0.001
100 zones	0.581	0.338	Intercept	0.489
			Perimeter	0.011
			Population density	0.035
			Density range	0.010
			Minimum population density	0.010
			Maximum population density	0.010

from areal interpolation and the explanatory variables are the parameters recorded for the target zones. In this way, the level of error for a specific target polygon can be predicted from its parameters. Initially, the full model is fitted using all of the variables. Insignificant variables are then removed one at a time until those remaining contribute significantly to the model. In this case, the criterion $p < 0.05$ was used to determine significance. At each step, the variable with the smallest contribution to the model (or the largest p-value) was removed, as long as that p-value was greater than the chosen 0.05 level.

Tables 2 and 3 present the significant parameters following stepwise regression between the mean error and the various parameters, for the areal weighting and dasymetric techniques, respectively. For the areal weighting technique the perimeter and, to a lesser extent, total population are the most significant parameters in the model. The trends are less clear for the dasymetric technique, with population density being the most significant parameter for the fifty-zone case, but with other attribute parameters playing a major part in the one-hundred-zone experiment. In general, as the number of zones in the experiment increases, so the number of significant parameters also increases. The perimeter, however, plays an important part in almost all cases.

Standard Deviation of Estimates. The correlation coefficients and significance of the t-statistics from regression analysis for the standard deviation and various parameters are presented in Table 4. The parameters producing the most significant prediction of the standard deviation for the areal weighting technique, are the area, perimeter, total population, and population density of the target zones. Area, perimeter, and total population all display positive correlation with the standard deviation, compared to the negative relationship for population density. For the dasymetric technique, the most significant parameters are

TABLE 4
Correlation Coefficients and Significance Values for Standard Deviation and Parameters

	50 target zones				100 target zones			
	Areal weighting		Dasymetric		Areal weighting		Dasymetric	
Area	0.710	*<0.001*	0.015	*0.917*	0.789	*<0.001*	0.265	*0.008*
Perimeter	0.783	*<0.001*	0.027	*0.852*	0.860	*<0.001*	0.349	*<0.001*
Compactness ratio	−0.310	*0.029*	−0.113	*0.433*	−0.422	*<0.001*	−0.412	*<0.001*
Total population	0.601	*<0.001*	0.535	*<0.001*	0.709	*<0.001*	0.739	*<0.001*
Population density	−0.532	*<0.001*	0.343	0.015	−0.506	*<0.001*	−0.052	*0.609*
Population density range	−0.225	*0.117*	0.624	*<0.001*	0.152	*0.131*	0.614	*<0.001*
Minimum Population density	−0.430	*0.002*	0.025	*0.175*	−0.521	*<0.001*	−0.282	*0.004*
Maximum Population density	−0.296	*0.049*	0.610	*<0.001*	−0.117	*0.246*	0.434	*<0.001*

NOTE: All values to three decimal places; values in plain type are product moment correlation coefficients; values in italics are significance values of the *t*-statistics resulting from regression analysis.

TABLE 5
Stepwise Regression for Standard Deviations from Areal Weighting Technique

Zonation	Multiple *R*	*R*-squared	Coefficients	*P*-value
50 zones	0.813	0.661	Intercept	0.009
			Perimeter	<0.001
			Total population	0.014
100 zones	0.900	0.809	Intercept	0.029
			Perimeter	<0.001
			Total population	<0.001

TABLE 6
Stepwise Regression for Standard Deviations from Dasymetric Technique

Zonation	Multiple *R*	*R*-squared	Coefficients	*P*-value
50 zones	0.695	0.484	Intercept	0.185
			Total population	<0.001
			Population density	<0.001
100 zones	0.760	0.578	Intercept	0.102
			Total population	<0.001
			Population density	0.008

total population, population density range and maximum population, all of which display positive correlation with standard deviation. This analysis suggests that, overall, total population is consistently the most significant parameter for predicting the standard deviation of estimates.

Tables 5 and 6 present the significant parameters following stepwise regression between the standard deviation of estimates and the various parameters, for the areal weighting and dasymetric techniques respectively. Multiple regression against the standard deviation of estimates produces consistent results. For the areal weighting technique, the perimeter, and, to a lesser extent, total population are the most significant parameters, compared to total population and population density for the dasymetric technique.

Overall then, correlation and regression analysis suggests that the errors from the areal weighting technique are strongly related to the spatial parameters of the target zones, whereas the dasymetric method shows more correlation with the attribute parameters of the zones. The parameters playing the most significant roles in regression models are the perimeter, total population, and population density.

3.4 Visualization of Errors

One important aim of the research was to incorporate a spatial component into analysis of the errors occurring in areal interpolation. One of the major advantages of recording error measures at the target zone, or polygon, level is that it allows the visualization of errors. Choropleth maps were therefore created to present and analyze the mean error and standard deviation of estimates for the target zones in each experimental situation.

Choropleth Maps of Mean Error. Figure 5 shows the mean error from both techniques for all target zones in the fifty-zone experiment. It confirms the more accurate estimations resulting from the dasymetric technique (b), as opposed to those from the areal weighting method (a). On the whole, the two methods tend to overestimate or underestimate the same zones. Analysis at the polygon level allows certain zones to be isolated where the interpolation methods are performing particularly well or poorly, and when there is inconsistency between the two techniques.

The choropleth maps seem to confirm that no specific individual parameter has an overriding influence on the accuracy of the results. The different parameters have a cumulative effect, making trends in the errors difficult to identify. For example, the maps show that a number of large and irregularly shaped zones have inaccurate estimates from the areal weighting technique. However, these are contradicted by smaller zones also having large errors. In addition, the distribution of errors does not seem to be determined by the distribution of population or land use, with high and low estimates in a mixture of suburban, rural, and city center locations. These factors may be playing a part in combination, but no one specific parameter has an overriding influence.

Choropleth Maps of Standard Deviations. Figure 6 presents the standard deviation of estimates by the areal weighting (a) and dasymetric methods (b) for fifty target zones. The areal weighting map clearly indicates a larger deviation of estimates compared to the dasymetric equivalent. On the whole, the values in the one-hundred-zone case are lower than those in the fifty-zone experiment for both methods. Again, there are difficulties in identifying spatial trends due to the different influences of the various parameters.

4. DISCUSSION

In general, the errors and reliability of estimation from the areal weighting technique were found to be strongly related to the *geometric* parameters of the target zones, whereas the dasymetric method showed more correlation with population or *attribute* parameters. These results are largely consistent with what we might expect given the information used by the two techniques. The areal weighting interpolation is based solely on the areal extent of the zones. Inaccuracies are likely to result if population is not distributed uniformly within the zones, for example, in rural communities where zones are frequently large but contain small clusters of populations. We would expect, therefore, that the shape, area, and distribution of population within the target zones would be the most important parameters in predicting errors from the

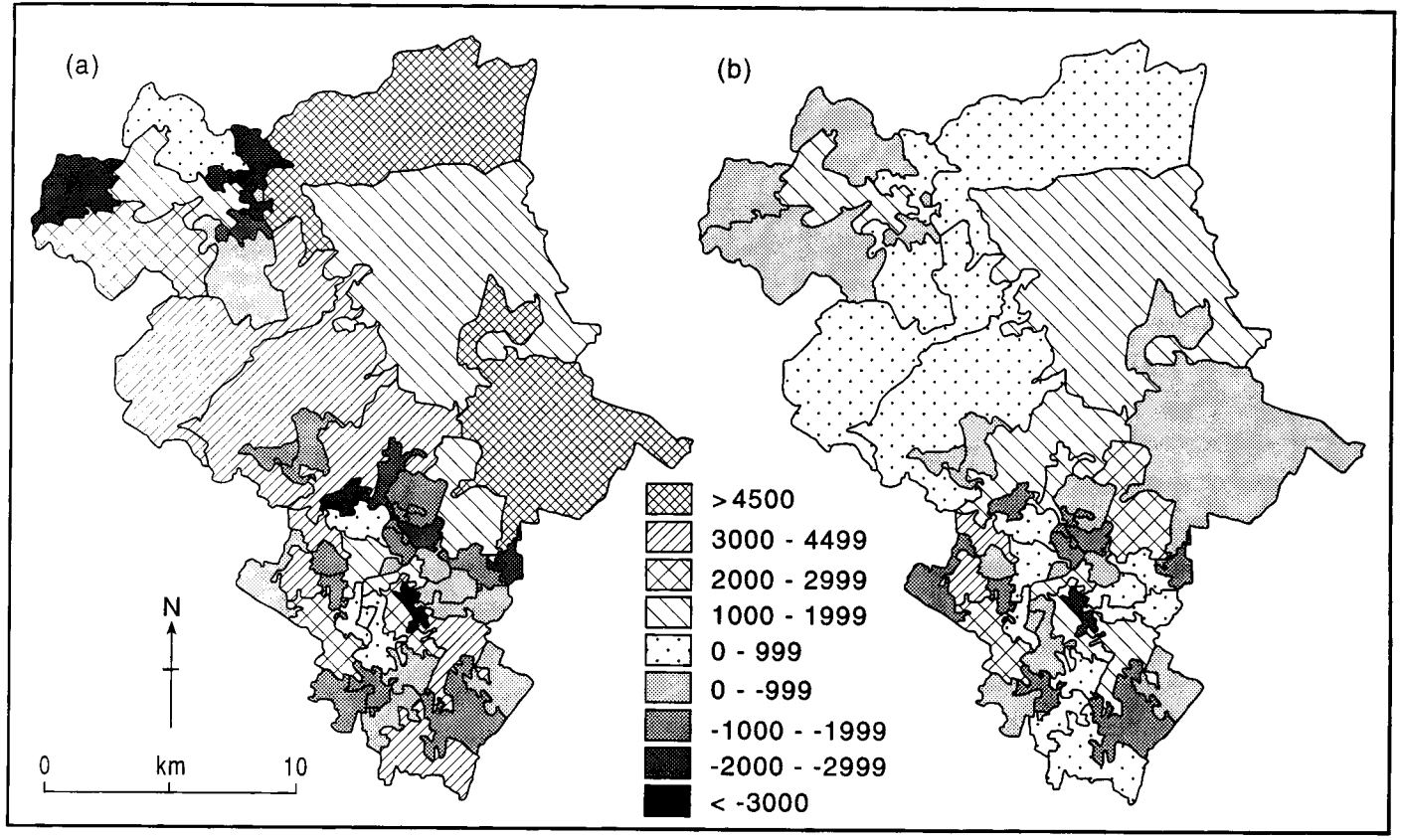

FIG. 5. Choropleth Map of Mean Errors (fifty target zones)

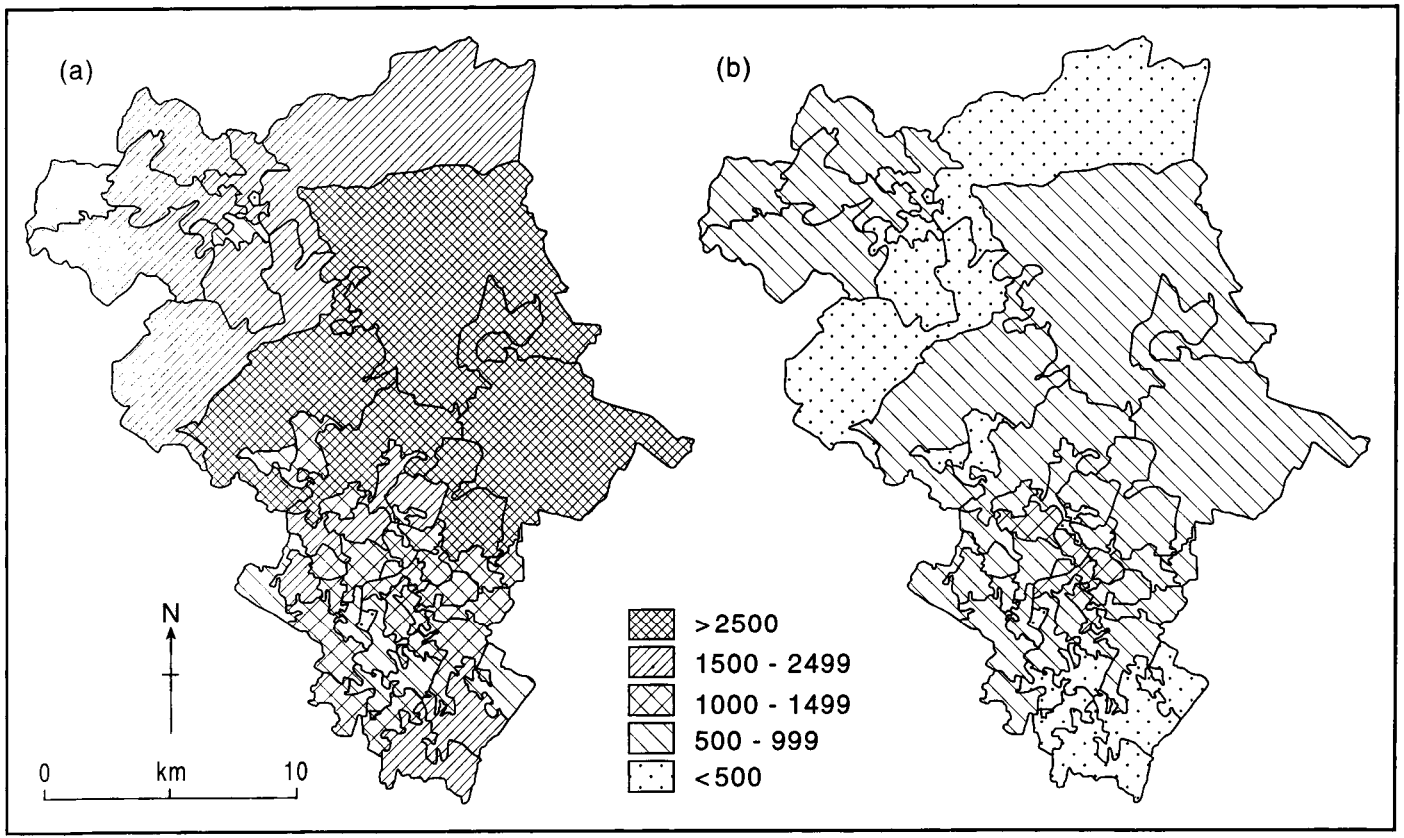

FIG. 6. Choropleth Map of Standard Deviation of Estimates (fifty target zones)

areal weighting technique. In contrast, the dasymetric technique is able to use information about the distribution of population within the target zones and is therefore less reliant on the geometric properties of the zones. We may expect inaccuracies to arise in sparsely populated areas, where the classification of population using remotely sensed data, is likely to be less accurate.

Visualization of the errors failed to identify spatial trends and links with the parameters. However, choropleth maps confirmed that prediction of the error values in areal interpolation is always going to be difficult due to the confounding influences of various parameters on the estimates. The maps showed the differences between the two methods of interpolation clearly and the basis for further research, discussed below, has been laid.

Further analysis and testing of parameters is required before full predictive models can be derived. The parameters employed in this study produced promising results, although there are many more parameters requiring investigation. In particular, a more effective measure of the population distribution within zones would be useful, effectively a measure of texture or pattern. The validity and feasibility of recording parameters of the source zones may also warrant further consideration; the characteristics of both the source and target zones may well be required to derive reliable predictive models.

Visualization of the errors in areal interpolation is a subject that as yet has received little attention. The most effective form of visualization would be software to enable interactive spatial analysis of the dataset. This would allow simple mapping of the errors, but would also create an environment whereby the user could interrogate the parameters of the zones and undertake spatial statistical analysis of the data set.

The methodology needs expanding to encompass further geographical areas, different socioeconomic data sets and other interpolation methods. In terms of geographical areas there is a clear need to investigate the effect of areas of high-rise buildings and areas of natural vegetation on the errors in the interpolation process. The method can also be expanded to incorporate other data sets, especially within the socioeconomic domain, such as health statistics, police records and other geo-referenced data.

There is also a clear need to incorporate other documented areal interpolation techniques into the analysis. As well as the regression models already investigated by Fisher and Langford (1995), obvious examples include pycnophylactic interpolation (Tobler 1979), density surface estimation (Martin and Bracken 1991), and the Poisson distribution of error (Flowerdew and Green 1991).

Despite the dasymetric method consistently producing the most accurate results it must not be forgotten that it is still producing errors. Further improvements in the areal interpolation methods must therefore always be sought. However, perhaps a more long-term solution is to address the problem of areal interpolation at source—that is, to eradicate the problem of geographically incompatible spatial units altogether. Efforts have been made to move toward data integration and to establish a standardized set of spatial units on which to collect and distribute data throughout the United Kingdom, but this can have a positive influence only on *future* data collection. For historical data, of course, it will only produce even more incompatible boundaries.

5. CONCLUSIONS

The issue of areal interpolation is not a new one; it is a problem that faces most users of spatial data at some time or another. It therefore needs attention if Geographical Information Systems are to be employed to their full potential.

This study represents an important step forward in investigating the errors in areal interpolation. The recording of error measures at *polygon* level is an innovative approach that has allowed in-depth analysis and basic visualization of the errors. Furthermore, it has enabled investigation into the links between a set of zonal parameters and the errors resulting from areal interpolation. The error resulting from the areal weighting technique is shown to be strongly influenced by the geometric parameters of the target zone, whilst the dasymetric technique is more strongly related to the population or attribute characteristics of the zone. This methodology is applicable to other geographical areas, other data sets and other methods of areal interpolation.

LITERATURE CITED

Fisher, P. F., and M. Langford (1995). "Modeling the Errors in Areal Interpolation between Zonal Systems by Monte Carlo Simulation." *Environment and Planning A* 27, 211–44.

Flowerdew, R., and M. Green (1991). "Data Integration: Statistical Methods for Transferring Data between Zonal Systems." In *Handling Geographic Information*, edited by I. Masser and M. Blakemore, pp. 38–54. Harlow: Longman.

Lam, N. S-N. (1983). "Spatial Interpolation Methods: A Review." *American Cartographer* 10(2), 129–49.

Langford, M., P. F. Fisher, and D. Troughear (1993). "Comparative Accuracy Measurements of the Cross-Areal Interpolation of Population." In *Proceedings of European Conference on Geographical Information Systems* (EGIS) 93. Utrecht: EGIS Foundation, 663–74.

Langford, M., D. Maguire, and D. Unwin (1990). "Mapping the Density of Population: Continuous Surface Representations as an Alternative to Choropleth and Dasymetric Maps." Midlands RRL Research Report No. 8.

——— (1991) "The Areal Interpolation Problem : Estimating Population Using Remote Sensing in a GIS Framework." In *Handling Geographic Information*, edited by I. Masser and M. Blakemore, 55–77. Harlow: Longman.

Langford, M., and D. Unwin (1994). "Generating and Mapping Population Density Surfaces within a GIS." *The Cartographic Journal* 31, 21–26.

Martin, D., and I. Bracken (1991). "Techniques for Modelling Population-related Raster Databases." *Environment and Planning A* 23, 1069–75.

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, 169–84.

Tobler, W. R. (1979). "Smooth Pycnophylactic Interpolation for Geographical Regions." *Journal of the American Statistical Association* 74 (367), 519–35.

Wright, J. K. (1936). "A Method of Mapping Densities of Population with Cape Cod as an Example." *Geographical Review* 26, 103–10.