PAPERS IN REGIONAL SCIENCE: The Journal of the RSAI 70, 3: 303–15

USING AREAL INTERPOLATION METHODS IN GEOGRAPHIC INFORMATION SYSTEMS

Robin Flowerdew
Mick Green
Evangelos Kehris
North West Regional Research Laboratory
Lancaster University
Lancaster LA1 4YB
UK

ABSTRACT This paper reports on a research project concerned with the areal interpolation problem — the problem of comparing different data sets when they have been made available for different zonal systems. Our approach is based on using additional information to guide the interpolation process. This paper emphasizes recent work applying the method to Poisson and binomially distributed data. There is also discussion of how the method can best be implemented in a geographic information system.

1. INTRODUCTION

A common problem in geographical or other regional research is the fact that the areal units for which data are available are not necessarily the ones that the analyst wants to study. This is a problem first of all for people interested in one particular type of unit, such as an administrative district, an electoral division, or a sales territory. It becomes more of a problem for people wishing to compare data over time when boundaries of data collection zones are subject to change. For example, if electoral reapportionment has taken place between elections, it is difficult to compare the performance of candidates between the two elections; in general, the monitoring of progress (or the reverse) becomes problematic if boundaries have changed. Even national censuses may be hard to compare over time because data collection units are subject to change (Norris and Mounsey 1983).

Situations frequently occur where a researcher wants to compare one variable that is available for one set of zones with another variable only obtainable for a different incompatible set. Common examples in the British context include comparisons between wards and parishes and between districts, constituencies, and travel-to-work areas (reviewed by Walford, Lane, and Shearman 1989). In many situations, it is important to compare data collected by private or commercial organizations with data made available officially. For example, client or customer addresses or survey results may be compared with officially provided demographic or socioeconomic data. This requires comparisons between zones such as postal-code areas or customer-service areas with officially designated zones. A further commonly encountered problem is the comparison of data for officially constituted areas with data for distance bands around a city, store, or pollution source. Sometimes the comparison may involve environmentally defined regions, such as geological, vegetation, or elevation zones or stream catchments, whose boundaries will seldom match those for which official data are available.

The comparisons mentioned above may be necessary for evaluating change, for assessing the performance of facilities or stores, or for producing effective maps to show two or more phenomena in relation to each other. The most important type of use, however, is statistical testing for relationships between the variables concerned.

This problem is a fundamental one in geographic information systems, where it is sometimes referred to as the polygon overlay problem. A preferable term, coined by Goodchild and Lam (1980), is "areal interpolation," which has the advantage of suggesting similarities with other types of interpolation problems. Reviews are available from Lam (1983) and Flowerdew and Openshaw (1987).

2. TERMS AND NOTATION

First of all, some terminology should be established. The problem of comparing data for two zonal systems can be restated as one of deriving data for one set of zones given the relevant data for another set. The variable of interest will be denoted Y; data on Y are available for a set of source zones S but are needed for a set of target zones T where both S and T cover the same geographical area. Figure 1 illustrates these terms. The value of Y for zone s is denoted y_s (which is known); the value for zone t is denoted y_t (which is not known).

If zone s intersects zone t, their boundaries form a zone of intersection st.

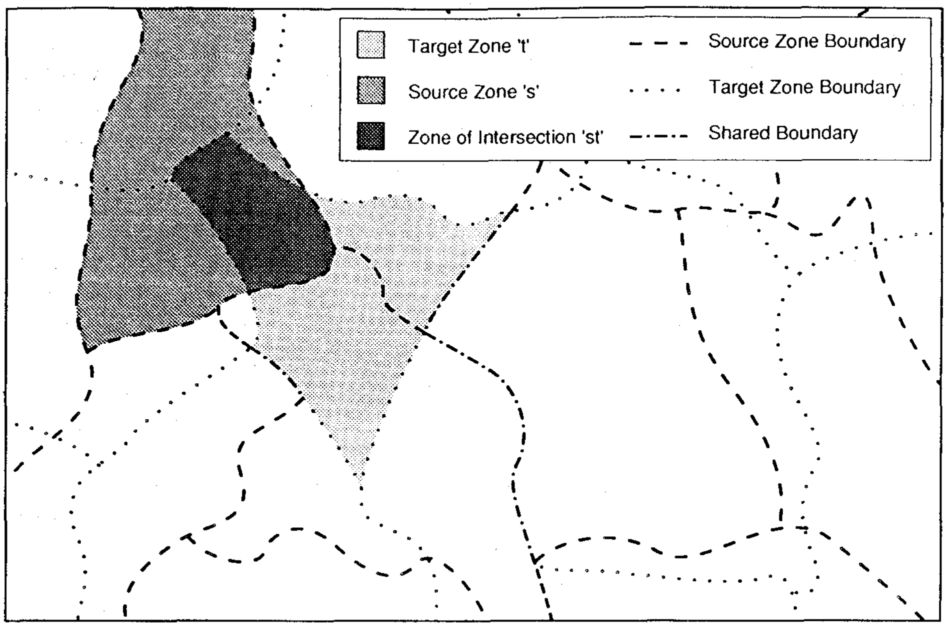

FIGURE 1. Source Zones, Target Zones, and Zones of Intersection

Sometimes a source zone will lie entirely within a target zone (or vice versa) but usually each source zone will be split by the target zone boundaries into several intersection zones, and each target zone will be similarly split. Usually there will be a simple relationship between the value y_t for target zone t and the values y_{st} for the intersection zones into which it is split. The problem of finding the values for the target zones then reduces to the problem of finding the intersection zone values.

The size of source, target, and intersection zones is highly relevant to the areal interpolation problem. The areas of these zones will be denoted as A_s, A_t, and A_{st}, respectively.

The variable of interest Y may be of several different measurement scales. Three important cases are where Y is categorical, where Y is extensive, and where Y is intensive. The terms extensive and intensive are derived from Goodchild and Lam (1980). The variable Y is extensive if its value for a target zone t is equal to the sum of its values for each zone st that intersects zone t:

$$y_t = \sum_s y_{st}.$$

Y will usually be extensive if it is a count or if it is a total of some sort (e.g., biomass, customer expenditure, wheat production).

The variable Y is intensive if its value for a target zone is a weighted average of its value for the intersection zones. Often the weights involved will be the areas involved:

$$y_t = \sum_s y_{st} A_{st} / \sum_s A_{st}.$$

Proportions and rates are examples of types of intensive variables. Most interval-scale variables are either intensive or extensive, but some are not: for example, relative relief (the difference between the highest and lowest points in a region) is neither extensive nor intensive.

The approach to areal interpolation to be developed in this paper involves the use of additional information about the target zones to help in the interpolation process. This requires the use of at least one ancillary variable available for the target zones; this will be denoted X_t. In some cases, X_t may be a vector of ancillary variables, not just a single variable.

3. AREAL WEIGHTING

The standard method for areal interpolation was stated by Markoff and Shapiro (1973) and discussed in detail by Goodchild and Lam (1980). Because it is based on combining source-zone values weighted according to the area of the target zone they make up, it will be referred to here as the areal weighting method. Different forms of the method are applicable for extensive and intensive data.

In the case of an extensive variable, it is assumed that the variable concerned is evenly distributed within the source zone. A subzone constituting half of a source zone will therefore have a value equal to half that of the source zone. In general, the ratio of the value for a subzone to the value of the source zone is assumed equal to the ratio of the area of the subzone to the area of the source zone:

$$\hat{y}_t = \sum_s A_{st}\, y_s / A_s.$$

For intensive data, an even distribution within source zones is again assumed. This means that the value for a subzone is the same as the value for the source zone. However, when subzone totals are combined to produce an estimate for the target zone, the result is a weighted average of the subzone values, with weights equal to the ratio of subzone area to target zone area:

$$\hat{y}_t = \sum_s A_{st}\, y_s / A_t.$$

In both cases, these methods assume that the variable of interest is evenly distributed within the source zones. This may be the most reasonable assumption if nothing is known about their distribution within the source zones. Frequently, of course, we do have additional information about distribution within the source zone, either directly or through knowledge of the distribution of other variables that we would expect to bear some relationship to the variable of interest. It is highly likely, for example, that knowledge of important variables like topography or population distribution would affect our expectations of how other variables might be distributed within source zones. It is the object of our research project to develop more sophisticated versions of the areal weighting method to take account of such additional information.

4. AREAL INTERPOLATION USING ANCILLARY DATA

As stated above, our aim is to develop methods of areal interpolation that are more "intelligent" than areal weighting in the sense that they can take into account other relevant knowledge we may have about the source zones. One form of this, dasymetric mapping, has been known for some time (Wright 1936) although, as Langford, Unwin, and Maguire (1990) attest, it has not been widely used. It is applicable where it is known that part of a zone necessarily has a zero value for some variable. If the variable of interest is population, for example, it may (usually) be assumed to be zero for those parts of a zone covered in water, or even for those parts known to have nonresidential land use. Accordingly, it is assumed to be evenly distributed within those parts of the source zone not ruled out. This can be regarded as a step toward "intelligence" but a somewhat limited one; ancillary information is used to assign a subzone to one of only two categories — either it has the value of the whole zone (or an appropriate proportion thereof) or it has the value zero. The methods to be described here are more flexible in that they allow further information about a zone to be used to make quantitative estimates about the distribution of the variable of interest rather than a crude binary distinction.

The simplest example is where we have ancillary information about a target zone that allows us to assign it to one of two categories. In Figure 2, source zone s has a known value for variable Y, y_s, but in order to estimate the value y_t for target zone t we need to know the distribution of y_s between the two subzones created by the intersection of source- and target-zone boundaries. For

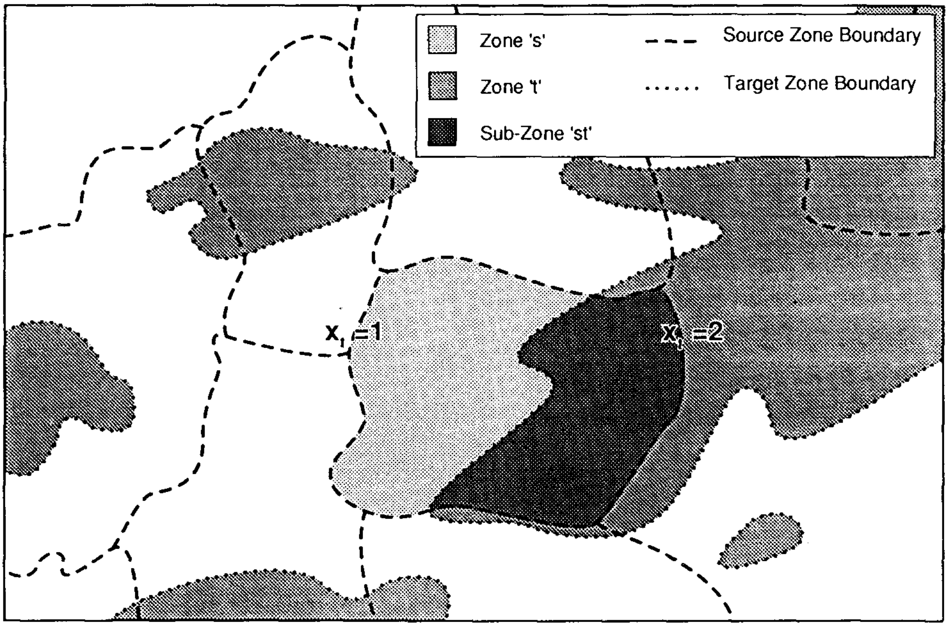

FIGURE 2. Use of Ancillary Data in Areal Interpolation

example, source zones might be administrative areas for which Y values (e.g., population) are readily available, whereas target zones might be defined by an environmental feature (e.g., rock outcrop), which might, for historical, economic, or other reasons, be expected to support population at a different density. In southern England, this might be true of limestone and clay belts. If it is desired to know how many people live in a limestone area (where $x_t = 2$ in Figure 2), the areal weighting method might give misleading answers if, in source zones like s, the majority of the population is concentrated in the clay portion of the zone.

If it is the case that population density is likely to be higher in a clay area than a limestone area, we would expect that the population of s would be predominantly located in the left part of the zone (dasymetric mapping techniques would not be applicable unless we were sure that nobody lived in the limestone area). Using the same principle for all source zones that include land of both types would lead to the conclusion that the population of a limestone target zone will be less than the areal weighting method would suggest. To estimate how much less, it is necessary to estimate the expected population for a given area of clay (y_1) and for a given area of limestone (y_2). This can be done using the whole data set, as described by Flowerdew and Green (1989), and estimates derived for y_{st} by scaling the results to fit observed y_s values. An essentially similar method has also been applied by Langford, Unwin, and Maguire (1990).

5. THE EM ALGORITHM

This method can be generalized using the EM (expectation/maximization) algorithm (Dempster, Laird, and Rubin 1977), a general statistical technique designed primarily to cope with problems of missing data. The EM algorithm consists of two iterated steps: the E step, in which the conditional *expectation* of the missing data is computed, given the model and the observed data, and the M step, in which the model is fitted by *maximum* likelihood to the "complete" data set (including the estimates made in the E step). These steps are then repeated until the algorithm converges.

If the areal interpolation problem is considered as a problem of estimating the values of y_{st} for each subzone formed by the intersection of the source-zone and target-zone boundaries, it can be regarded as a missing data problem. In general, y_{st} will be missing data but the row totals y_s are known. If the y_{st} values can be modeled on the basis of the areas of the subzones A_{st} and ancillary information x_t about the target zones, the EM algorithm can be applied to derive estimates for y_{st}. In practice, of course, y_{st} should only be nonzero for those combinations of source and target zone that actually intersect.

Extending the example in the previous section, where Y represented population, it may be reasonable to assume that Y_{st} has a Poisson distribution with parameter μ_{st}, and that this parameter is a function of x_t, A_{st}, and unknown parameters β. The E step then involves computing the conditional expectation $\hat{y}_{st}$ of y_{st} given the data y_s and the current model of μ_{st}:

$$\begin{aligned}\hat{y}_{st} &= \mathrm{E}[y_{st} \mid \hat{\mu}_{st}, y_s] \\ &= \hat{\mu}_{st}\, y_s / \sum_k \hat{\mu}_{sk}.\end{aligned}$$

The M step consists of fitting the model $\hat{\mu}_{st} = \mu(\beta, x_t, A_{st})$ by maximum likelihood as if the estimates $\hat{y}_{st}$ were independent Poisson data. This will yield values for the coefficients that give information about how the Poisson parameters μ_{st} are linked to the ancillary data and the subzone areas. The values of $\hat{\mu}_{st}$ derived in the M step can then be fed back into the E step, to derive better estimates of y_{st}. In turn, these can be used in the M step to provide better estimates of μ_{st}, and so on until convergence is reached. The areal weighting method can be used to give starting values for the y_{st} estimates.

If the ancillary information for target zones is geology, and each target zone is either clay ($x_t = 1$) or limestone ($x_t = 2$), population may be regarded as a Poisson-distributed variable whose parameter is either λ_1 (for clay zones) or λ_2 (for limestone zones). A suitable model for subzone population might be $\hat{\mu}_{st} = \lambda_t A_{st}$, where $t = 1$ if zone t is clay and 2 if it is limestone. The E step might then take the form

$$\hat{y}_{st} = \lambda_t A_{st}\, y_s / \sum_k \lambda_k \mathrm{A}_{sk}$$

for all target zones k intersecting a source zone s. The M step then consists of fitting the model $\mu_{st} = \lambda_t A_{st}$, using $\hat{y}_{st}$ as data. After iteration of the E and M steps until convergence, the target-zone totals can be estimated by summing the relevant subzone totals:

$$\hat{y}_t = \sum_s \hat{y}_{st}.$$

This method can be generalized easily for different types of ancillary data within the EM framework. Thus the ancillary variable X could have more than two possible cases: it could be a continuous variable, or there could be several different categorical or continuous variables used in combination to produce improved estimates of y_{st}. Further information is given for these cases in Green (1989).

The models discussed so far are based on the assumption that the variable of interest has a Poisson distribution. This is often a reasonable assumption when dealing with social and demographic data that may be presented as counts of people or events in each zone. In many situations, however, the variable of interest is not the total count, but the proportion of people or events who come into a particular category. We might be interested, for example, in the proportion of workers in a zone who are unemployed, the proportion of reported crimes using motor vehicles, or the proportion of diseased trees identified in a survey. For problems of this type, a binomial distribution may be appropriate, and Green (1990) has developed methods of areal interpolation applicable to this context.

Again, we assume that data on the proportion of a population sharing some characteristic are available for a set of source zones, but estimates are needed for a set of target zones. In this situation, it is not the area of the subzones produced by the boundary intersections of the source and target zones that is important but their total populations N_{st} (as intersection zones are whole wards, this information can be obtained from the census). We are concerned to estimate y_{st}, the number of "successes" (i.e., the cases whose proportion we are interested in), which is regarded as having a binomial distribution with parameters N_{st} and p_{st}, where p_{st} is the probability of a member of the population under study being a "success." Ancillary information x_t about the target zones is available, and p_{st} can be estimated as a function of x_t and a set of parameters β to be estimated.

6. OPERATIONALIZING THE METHOD IN GIS

Areal interpolation can be regarded as a fundamental operation in the use of geographic information systems, because one of the central methods in GIS is the creation of new areal units through operations like buffering and overlay. Ascribing appropriate attribute values to these new units, as argued above, is highly problematic for certain kinds of variables, and hence it is important for a GIS to be linked to methods such as those described above. Ideally, it should be possible for a GIS user to derive appropriate values for new areal units without leaving the GIS; in practice, all that can reasonably be expected is an interface between the GIS and a statistical package that can fit the models described. The statistical package used is GLIM (Generalised Linear Interactive Modelling system) (Payne 1986), which has facilities for interactive modeling for a wide range of statistical distributions, and for adding additional procedures (such as the EM algorithm) in a macro language. In addition to cartographic display, the essential contribution of the GIS is in calculating the areas of the subzones formed by the intersections of the source and target zones. The GIS used in this project was ARC/INFO.

As Kehris (1989) explains, there are problems in linking data between ARC/

INFO and GLIM, but an effective method can be developed using the PASS facility in GLIM, provided that the ARC/INFO object code is available. PASS allows GLIM users to import and export data from other software, and its use in this context requires a user-written subroutine to read in an INFO data file. In the example below, the subroutine has been assigned the number 6, which is used to pick out the right code from the PASS facility. The following represents code that can be run from within GLIM to read an item called POPULATION from a polygon attribute table in INFO (in this case, WARDS.PAT in the directory shown) and to put it into a variate called POP in GLIM. The first three lines are GLIM commands and the remainder are generated by the user subroutine (prompts are in italics):

? $UNITS 80$
? $DATA POP$
? $PASS 6 POP$
ENTER THE DIRECTORY PATH
"/SCRATCH/COV/INFO"
ENTER THE FILENAME
"WARDS.PAT"
ENTER THE USER NAME
"ARC"
ENTER THE NAME OF THE ITEM
"POPULATION"

A similar subroutine may be added to the PASS facility to move data in the reverse direction. Thus source-zone data for the variable of interest, target-zone data for the ancillary variables, and area data for the intersection subzones can all be read from INFO to GLIM. The appropriate version of the areal interpolation algorithm can then be run within GLIM, and the estimated target-zone values can be passed back to INFO using the second PASS subroutine for subsequent graphic display and further analysis.

7. CASE STUDY

The methods described above can be illustrated using 1981 census information for North West England. Local government districts are the source zones and parliamentary constituencies (using 1983 boundaries) are the target zones. One advantage of using this particular data set is that the data are available for both sets of units; this makes it possible to estimate values for the constituencies from the district data, and then to check the results against the real data. Both districts and constituencies are aggregations of wards. In some cases, districts and constituencies are exactly coextensive (so areal interpolation is unnecessary); even when this is not so, they often have boundaries that are coextensive for part of their length. Figure 3 shows the boundaries for both sets of units.

There are thousands of count variables available in the Small Area Statistics from the 1981 census for which areal interpolation methods based on the Poisson distribution may be appropriate. The method is illustrated here using the population born in the New Commonwealth or Pakistan as the variable of

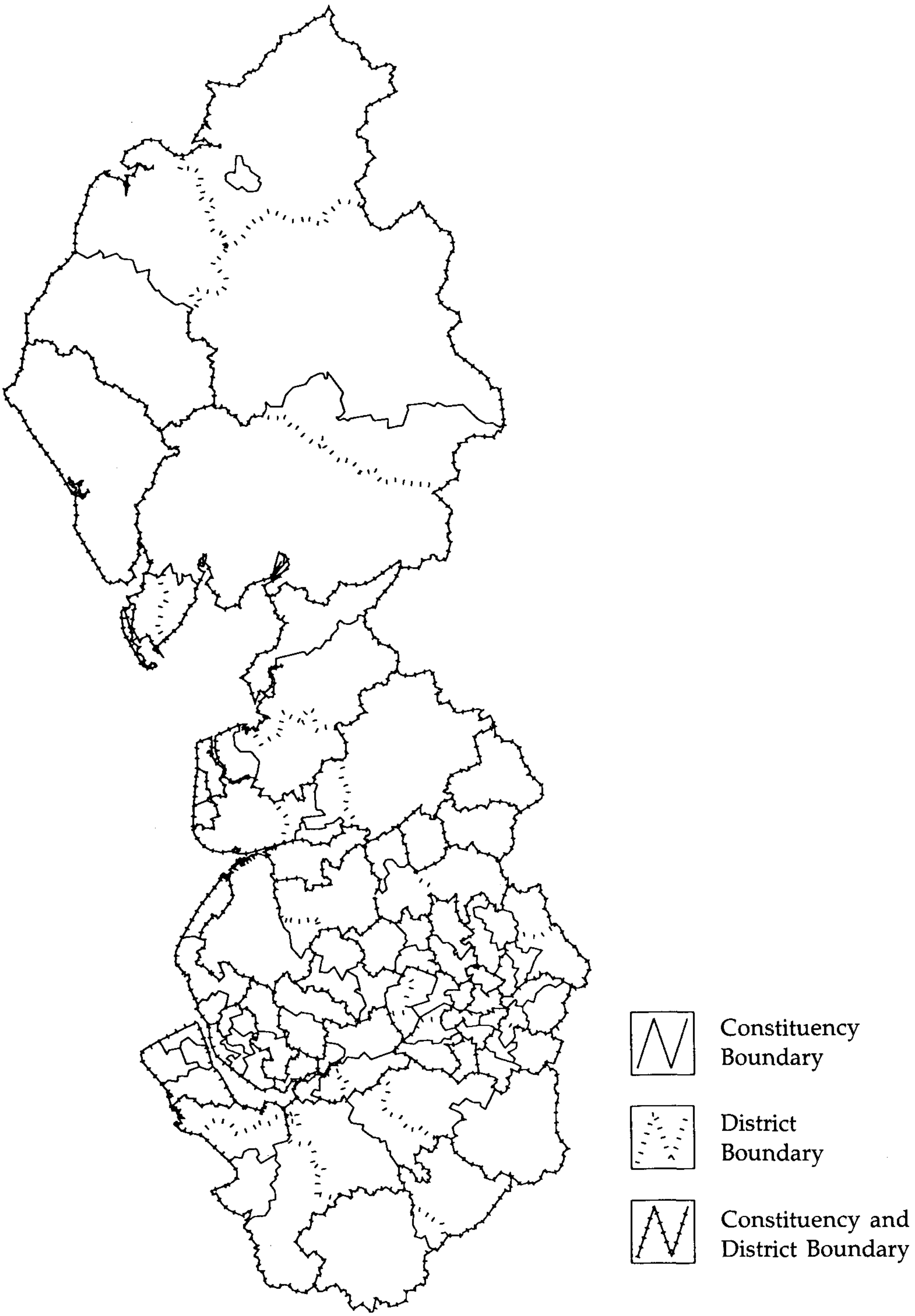

FIGURE 3. Boundaries of Districts and Parliamentary Constituencies in North West England

interest. Using districts as source zones, the technique described above was used with several different target-zone variables used as ancillary data, some from the census and some from other sources.

Some idea of the effectiveness of the method may be gained from the goodness of fit of the model fitted in the M step of the EM algorithm. This is assessed using the deviance (likelihood ratio); lower values represent better model fit. Table 1 shows the goodness of fit for five areal interpolation models ("source deviance"). The areal weighting model is the conventional method described by Goodchild and Lam (1980) in which the variable of interest is assumed to be evenly distributed within the source zones. Bearing in mind that the target zones are constituencies, a simple ancillary variable is the political party winning the constituency at the 1983 election (effective use of an ancillary variable in areal interpolation does not require there to be a clear cause-effect relationship with the variable of interest). Car ownership, taken from the 1981 census, is used as an ancillary variable in two different forms: first as a three-level categorical variable and second as a ratio (cars per household). Lastly, overcrowding (proportion of people living at more than 1 person per room) is used. The figures in the source deviance column are difficult to interpret individually, but a comparison between them suggests that all the other models are considerably better than the areal weighting method, even the rather simplistic political party model. The most successful is the categorical version of the car ownership variable, which reduces the deviance to about one-third of its value for the areal weighting model.

Normally this source deviance statistic would be the only way of evaluating the success of the areal interpolation exercise. In this case, however, we know the true target-zone values of y_t as well as the source zone values y_s. The estimated values can then be compared with the true values for the target zones; the results of this comparison are given in the "target deviance" column, which is based on the fit of the estimated $\hat{y}_t$ values to the known y_t values. The values in this column are comparable with each other but are not comparable with the source deviance values.

Again all the models fitted are far better than the areal weighting method. However, there seems considerably less difference among the other models than the source deviance figures would suggest. The political party model actually

TABLE 1. Goodness of Fit of Areal Interpolation Models — Population Born in the New Commonwealth or Pakistan, North West England 1981

Model	Source Deviance	Target Deviance
Poisson		
Areal Weighting	347,834	115,638
Political Party	283,515	54,499
Car Ownership (Categorical)	117,394	62,030
Cars per Household	127,868	53,726
Overcrowding	183,240	40,667
Binomial		
Areal Weighting	90,651	44,622
Political Party	84,933	34,436
Car Ownership (Categorical)	69,398	34,329
Cars per Household	69,912	32,748
Overcrowding	65,145	26,621

gives better estimates than the car ownership models, and the overcrowding model is the most successful. The implications of this are, first, that "intelligent" areal interpolation does provide considerably better estimates than areal weighting and, second, that the best model in terms of source deviance is not necessarily the best in terms of target deviance. In addition to the models presented in Table 1, further models were fitted including combinations of the ancillary variables described already. Although these models had slightly lower source deviances than the single-variable models, the target deviances did not show similar improvement. Indeed, our experience suggests that a simple model is likely to be more successful than a more complicated one.

The work described so far models the New Commonwealth- and Pakistani-born population as if its distribution were independent of the population as a whole. It may be more sensible instead to model its distribution with respect to the total population, using the binomial model described above. In such a model, the proportion of New Commonwealth- and Pakistani-born people in the total population is a function of one or more ancillary variables. The lower half of Table 1 shows the results of fitting binomial models based on the same ancillary variables, again interpolating from districts to constituencies. The target deviance values are considerably less than for the Poisson models — perhaps not surprisingly since the overall distribution of population is taken into account in these models. Again, the source deviances do not give a perfect indication of the success of the interpolation as measured by target deviance, but the results are more consistent for the binomial models than for the Poisson.

8. CONTROL ZONES

The methods described so far have been intended to yield improved estimates of variable values for a set of target zones, given ancillary information about the same target zones. There may however be situations where target-zone estimates may be improved by ancillary information available for a third set of spatial units, which will be referred to here as control zones. This approach was first suggested by Goodchild, Anselin, and Deichmann (1989).

The North West England study described above was extended by using altitude as a control variable, defining control zones according to whether they were above or below the 400-foot (122-meter) contour, which roughly divides lowland and higher land in the region. Superimposition of source-zone, target-zone, and control-zone boundaries produces three-way intersection zones — for example, the zone of intersection between Lancaster district, Morecambe constituency, and land over 400 feet. The techniques described in previous sections can be generalized for this problem. In practice, altitude made little difference to goodness of fit. The approach might be fruitful, however, with other data or with different control zones.

9. DIFFICULTIES IN USING INTELLIGENT AREAL INTERPOLATION METHODS

The methods developed in this project represent an improved approach to the areal interpolation problem. They enable additional information to be taken into account when values of a source-zone variable are needed for a target zone, and they provide better estimates than the standard method of areal weighting. Nevertheless, there are a number of difficulties in using them.

First, there are practical difficulties in finding the appropriate data and areal unit boundaries and in putting them in the GIS. In the case of this project, there were considerable lengths of boundary that were identical for both district and constituency coverages. However, the two coverages were separately digitized. When one was overlaid on the other, a large number of very small and spurious zones of intersection emerged, created because of small differences in the way the lines had been digitized.

Second, there are problems in linking the GIS used to the analytical routines used in areal interpolation. In our case, we were able to operationalize a link between ARC/INFO and GLIM, but in doing this we were dependent on access to the object code for both packages. It is possible that the link would not be feasible for all hardware platforms, and it may or may not be practical to construct similar links for other GIS packages or for other appropriate statistical software.

Third, the methods are dependent on finding suitable ancillary variables. The analyst may not have access to suitable target-zone data, and the relationship of the variable of interest to available ancillary variables may be tenuous. Our experience has been that some improvement in estimates can often be gained even when the relationship is flimsy, but it is clearly sensible to use ancillary information that is strongly related to the variable of interest, when such information is available.

Fourth, as shown above, there are problems in evaluating the goodness of fit and in comparing models based on different ancillary variables. The source deviance values derived for the models are unreliable guides to the quality of estimates derived as evaluated by target deviance. All we can suggest is that the analyst be guided by a combination of source deviance values, intuitive reasonableness of the relationships postulated, and model simplicity. A complicated model with several parameters to be estimated may produce worse target-zone estimates than a simpler model based on one ancillary variable, so it may be sensible to use two or more ancillary variables only when there are good theoretical grounds to expect them to have strong and independent relationships to the variable of interest.

Fifth, areal interpolation cannot claim to give totally accurate target-zone figures, regardless of what method is used. If interpolated values are incorporated into the GIS, it is desirable that users be aware that they have been interpolated and hence may be inaccurate. Ideally they should be accompanied by some quantitative measure of reliability. This applies not just to the methods described in this paper, but also to the standard methods like areal weighting already used as a default in some geographic information systems. This concern is part of the wider topic of error in GIS and its representation, discussed from many perspectives in Goodchild and Gopal (1989).

Sixth, the areal interpolation methods discussed here are based on a limited set of statistical models, namely Poisson and binomial models. These are appropriate for count variables such as the census variables used in the case study, but they do not cover all the measurement scales that are likely to be encountered in a GIS. In particular, they do not cover continuous variables, and of course many different continuous distributions must be considered before intelligent areal interpolation methods are generally available for all types of data commonly used in GIS. Work is continuing on extensions of these methods for new types of data.

ACKNOWLEDGMENTS

The research on which this paper is based was conducted under research grant R000 231373 from the Economic and Social Research Council. We are also indebted to colleagues in the North West Regional Research Laboratory for ideas and assistance, to the ESRC and Lancaster University for access to hardware and software used in the project, and to the ESRC Data Archive and Manchester Computing Centre for the data.

REFERENCES

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 39: 1–38.

Flowerdew, R., and Openshaw, S. 1987. A review of the problems of transferring data from one set of areal units to another incompatible set. Lancaster and Newcastle, England: Northern Regional Research Laboratory, Research Report 4.

Goodchild, M. F., Anselin, L., and Deichmann, U. 1989. A general framework for the spatial interpolation of socioeconomic data. Paper presented at the Thirty-sixth North American Meetings of the Regional Science Association, Santa Barbara, California.

Goodchild, M. F., and Gopal, S., eds. 1989. *The accuracy of spatial databases.* London: Taylor and Francis.

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

Green, M. 1989. Statistical methods for areal interpolation: the EM algorithm for count data. Lancaster, England: North West Regional Research Laboratory, Research Report 3.

Green, M. 1990. Statistical models for areal interpolation. In *EGIS '90: Proceedings, First European Conference on Geographical Information Systems, 1,* eds. J. Harts, H. F. L. Ottens, and H. J. Scholten, pp. 392–99. Utrecht, The Netherlands: EGIS Foundation.

Kehris, E. 1989. Interfacing ARC/INFO with GLIM. Lancaster, England: North West Regional Research Laboratory, Research Report 5.

Lam, N. S.-N. 1983. Spatial interpolation methods: a review. *American Cartographer* 10: 129–49.

Langford, M., Unwin, D. J., and Maguire, D. J. 1990. Generating improved population density maps in an integrated GIS. In *EGIS '90: Proceedings, First European Conference on Geographical Information Systems, 2,* eds. J. Harts, H. F. L. Ottens, and H. J. Scholten, pp. 651–60. Utrecht, The Netherlands: EGIS Foundation.

Markoff, J., and Shapiro, G. 1973. The linkage of data describing overlapping geographical units. *Historical Methods Newsletter* 7: 34–46.

Norris, P., and Mounsey, H. M. 1983. Analysing change through time. In *A census user's handbook,* ed. D. Rhind, pp. 265–86. London: Methuen.

Payne, C. D., ed. 1986. *The Generalised Linear Interactive Modelling system, Release 3.77.* Oxford: Numerical Algorithms Group.

Walford, N. S., Lane, M., and Shearman, J. 1989. The Rural Areas Database: a geographical information and mapping system for the management and integration of data on rural Britain. *Transactions, Institute of British Geographers* 14: 221–30.

Wright, J. K. 1936. A method of mapping densities of population with Cape Cod as an example. *Geographical Review* 26: 103–10.