

A framework for the areal interpolation of socioeconomic data

M F Goodchild, L Anselin, U Deichmann

National Center for Geographic Information and Analysis, Department of Geography, University of California, Santa Barbara, CA 93106, USA

Received 4 March 1992; in revised form 27 July 1992

Abstract. Spatial data are collected and represented as attributes of spatial objects embedded in a plane. Basis change is defined as the transfer of attributes from one set of objects to another. Methods of basis change for socioeconomic data are reviewed and are seen to differ in the assumptions made in each about underlying density surfaces. These methods are extended to more general cases, and an illustration is provided by using Californian data. The implementation of this framework within a geographical information system is discussed.

Introduction

Geographical information systems (GIS) are being increasingly used to support a variety of forms of spatial and regional analysis, which use social, economic, and demographic data. The databases which these systems access are discrete digital representations of real spatial variation which use one of a number of standard discretizations or data models (for a review, see Burrough, 1986; Goodchild, 1992). Variation can be described by attaching attributes to a set of spatial objects, which may be points, lines, or areas, embedded in a plane. Thus the behavior of a variable such as population density is often described by attaching the density attribute to each of a set of reporting zones which collectively partition the plane. But, equally, we might represent the same spatial variation by attaching attributes either to a set of contours (line objects) or to a set of zone centroids (sample points). Each of these options has its advantages and disadvantages in terms of accuracy of representation, storage, and processing efficiency. We define the set of discrete objects used to describe the spatial variation of a parameter as its spatial basis.

The basis of social and economic statistics is usually defined independently of the variable itself, through the use of a set of reporting zones. In this sense, social and economic data models are different from those commonly used for environmental data, where the basis is derived from, and therefore unique to, the spatial variation itself (Mark and Csillag, 1989). The arbitrary nature of such socioeconomic reporting zones leads immediately to the difficulties known as the modifiable areal unit problem (Openshaw, 1984; Openshaw and Taylor, 1981), that is, the dependence of the results of spatial analysis on the arbitrary spatial basis of the data used.

A change of the spatial basis of regional analysis can be defined as the process of transferring attributes from one set of spatial objects to another within a defined portion of geographic space. An example of such basis change is the point interpolation problem, where the variable z is known for a number of points (x, y) but is needed at another point for which no data are available (Bennett et al, 1984; Griffith et al, 1989; Lam, 1983). A different case might be where data are available for a set of points but are needed for a system of areas or zones (Bracken and Martin, 1989; Martin and Bracken, 1991). Arbia (1989) has described the general problem of basis change and has outlined an approach based on a space partitioned into discrete units. In this paper we assume space to be continuous, and we will propose an approach that could be considered more general.

The most common form of spatial basis change for socioeconomic data occurs when data must be transferred from one set of reporting zones or areas to a second, independent set. These are termed *source* and *target* zones, respectively, where the zones are generally not nested hierarchically and the boundaries between zones generally do not coincide. This has been referred to as areal interpolation (Flowerdew and Openshaw, 1987; Flowerdew et al, 1991; Goodchild and Lam, 1980; Lam, 1983). It arises because of the general lack of coordination and commonality of objectives between data-collecting agencies. Different departments, or even different offices in the same department, publish data based on a number of zonal arrangements that are often incompatible (for example, counties, metropolitan statistical areas, labor-market areas, hydrologic study areas, etc). The zones often change through time, making longitudinal analysis difficult. Recently, marketing applications have begun to create a demand for statistics based on postal zones, which have been defined solely for the efficient delivery of mail and are thus largely incompatible with any other zone system.

These issues create significant problems for regional and multiregional analysis and modeling, which require both cross-sectional and longitudinal data and must often combine data from different sources. The data models and geometric algorithms of GIS provide a suitable environment for the development of a set of general procedures for spatial basis change for socioeconomic data. Our purpose in this paper is to lay the groundwork for such a set of procedures by presenting a general framework and by reviewing and extending existing methods.

General framework

The fundamental unit of spatial data is the tuple $\{x, y, z_1, z_2, \dots, z_n\}$, where $z_1, z_2, \dots, z_n$ represent the values of n variables at location (x, y) . Both x and y are measured on continuous scales, and in the absence of strong constraints on the degree of variability (homogeneity, stationarity, isotropy), the number of tuples required to describe the complete variation of z_1 through z_n is infinite. However, the presence of strong positive spatial autocorrelation in most spatially distributed variables ensures that a number of forms of discrete approximation will be acceptably accurate.

Unfortunately, most socioeconomic variables are not directly observable in this form. For example, it is not possible to visit location (x, y) and determine the population density or average personal income as an attribute of that point. Instead, most variables must be observed in the form of integrals or averages over finite areas and reported as attributes of a discrete set of objects, each of finite spatial extent. For example,

$$P_i = \int_x \int_y p(x, y) dx dy, \quad (1)$$

where $p(x, y)$ is the unobservable population density function at location (x, y) and P_i is the population of zone i . The minimum size of each zone i depends on the variable. Although a single block might yield a meaningful population count, statistics based on firms clearly require larger areas. We can place various interpretations on the unobservability of $p(x, y)$. Population is discrete, so it becomes impossible to approximate its distribution by a continuous function below a certain level of spatial resolution, which itself depends on density. In the limit $p(x, y)$ takes only two values, finite at points occupied by people and zero elsewhere. However, each individual's location is fundamentally uncertain because of data-collection errors, migration, diurnal movement, and geodetic imprecision, so $p(x, y)$ could be interpreted as the aggregate of a set of probability distributions. Similar arguments apply at appropriate scales to other socioeconomic variables.

Much of the following discussion will be concerned with population density as an example of a spatially distributed variable. Its integral P is observable over reporting zones above some minimal size, and other densities, such as gross income density, have similarly observable integrals. Ratios or averages, such as average personal income, are interpreted as the ratio of two densities or integrals, in this case gross personal income divided by population. Goodchild and Lam (1980) refer to spatially integrated densities as *spatially extensive*, and to the ratios of densities as *spatially intensive*.

The success of any basis change depends on the accuracy with which the underlying continuous function p can be estimated or modeled, and thus on the accuracy of hypotheses about the form of p . In the following sections, we review a number of alternative hypotheses from previous work on the basis change problem:

Radially symmetric kernel functions

It is possible to model the variation in population density over a city by using a radially symmetric function such as

$$p(r) = p_0 \exp(-Br), \quad (2)$$

where p_0 is the central density, B is a constant, and r is the radial distance. The literature records abundant tests of this and similar functions (Clark, 1951; Nordbeck, 1971; Parr, 1985). However, these monocentric models seem to be becoming less accurate with the transition to polycentric cities, particularly in North America (Gordon et al, 1986). Moreover, from the perspective of basis change it seems unlikely that one would know the central location and central density (or total population) of a city, or have additional information on variation of density within it. Nevertheless, Honeycutt and Wojcik (1990) have recently used a circular Gaussian function to estimate a continuous surface of population density for the conterminous USA from point data on city locations and populations.

In many countries it is possible to obtain digital boundary files for the larger census reporting zones, but the sheer number of the smaller units make this impractical. The geographical locations of the enumeration district or enumeration area level (100–200 households) in the USA, United Kingdom, and Canada are, at this time, available in digital form as ‘centroids’ or representative points only, usually placed in the area of highest density within the zone. In the near future, this may become less of a constraint because of the increasing availability of detailed digital boundary data for small socioeconomic areas. Bracken and Martin (1989; Martin and Bracken, 1991) describe a technique for estimating a continuous density surface from these data, as an essential step in basis change. The centroids are seen as summary measures of population distribution, and a kernel function is defined to distribute the population at each centroid uniformly over a window whose size is determined in each case by an analysis of the locations of nearby centroids, in an adaptive form of density estimation (Silverman, 1986).

Maximally smooth estimation

Tobler (1979) proposed the term ‘pynophylactic’ to describe a method of estimating a density surface which is density preserving, in the following sense. Suppose the plane is partitioned into a set of zones indexed by i ($i = 1, \dots, n$), each with known area A_i and known population P_i . An estimation method for $p(x, y)$ is pynophylactic if equation (1) holds for every zone. Clearly, this condition is not sufficient to estimate the surface, and many methods will have the pynophylactic property. Tobler’s method established additional boundary conditions on the surface $p(\cdot)$, for example that $p(\cdot) = 0$ at the edge of the partitioned area, or that the gradient of $p(\cdot)$ is zero at the edge. An iterative procedure is utilized to find a solution which satisfies

these constraints and minimizes the curvature of the surface. However, the minimum curvature objective derives more from a desire for simplicity and a lack of alternative hypotheses than from any observations about real population density surfaces.

Piecewise approximation

In these methods the plane is partitioned into zones of relative homogeneity, and then the variation of density within each zone is modeled by using a simple function. The possibility of using a constant value of density within each zone is suggested by the tract form of much housing development in recent years, and by the element of homogeneity in the definition of many census reporting zones, particularly census tracts. In general, however, homogeneous densities within socioeconomic reporting zones are exceptions rather than the rule.

Goodchild and Lam (1980) examined the case where one can reasonably assume that the source-zone density is constant. The derivation of the target-zone estimates is straightforward. If we denote the area of overlap between source zone s and target zone t by a_{st} , and the known source-zone population by U_s , the target-zone population, V_t , can be estimated from

$$V_t = \sum_s U_s \left(a_{st} / \sum_t a_{st} \right). \quad (3)$$

Note that, as is appropriate for a spatially extensive variable such as population, the individual areas of overlap, a_{st} , are standardized by using the total target-zone area. The resulting coefficients thus represent the share of the area of source zone s that lies within target zone t , and the source-zone statistic is allocated accordingly over target zones. In actual computation these coefficients are usually arranged in a matrix, A , which is referred to as the areal weights matrix or the intersection matrix.

Flowerdew and Green (1989) and Flowerdew et al (1991) describe the case where auxiliary information is available to assist in the estimation of target-zone densities. In one example, the target zones are political voting districts, and a binary variable (Conservative or Labour Party representation) is available for each target zone. By letting λ_1 and λ_2 represent the population densities of Conservative-held and Labour-held districts, respectively, we can express the expected population of source zones by $\lambda_1 A_1 + \lambda_2 A_2$, where A_1 and A_2 are the areas of overlap between the source zone and Conservative-held and Labour-held target zones, respectively. The parameters λ_1 and λ_2 are estimated by regressing the known source-zone population on the known overlap areas by means of Poisson regression because the observed source-zone population is integral and is presumed to follow a Poisson distribution.

Auxiliary information that could aid the areal interpolation process is often not consistently available for large areas. One data source that can be utilized, however, is remotely sensed data, such as real photography or satellite imagery. Langford et al (1990) statistically related land-cover information derived from a classified Landsat Thematic Mapper (TM) image to observed population totals in a set of administrative units. This relationship was subsequently used to derive population estimates for other arbitrary regional subdivisions.

Generalization 1: uniform target-zone densities

In this and the subsequent section we generalise the approaches taken by Goodchild and Lam (1980) and Flowerdew and Green (1989). Suppose that for some reason or other the population distribution can be assumed to be constant within the target zone for which data have to be derived. This might be the case if the source zones are drawn as arbitrary political boundaries, whereas the target zones follow clearly

distinguishable physical features (for example, built-up urban area versus agricultural land) in which the homogeneity assumption is reasonable. Provided that the number of source zones, n_s , is at least as large as the number of target zones, n_t , population densities for the target zones, d_t , can be estimated by using the observed source-zone populations, U_s , and the areas of overlap of source zone s with each target zone t :

$$U_s = \sum_t d_t a_{st}. \quad (4)$$

If the number of source zones is equal to the number of target zones, $n_t = n_s$, one can solve for the densities by inverting the area weight matrix A , which will be square in this case, provided A has an inverse. If $n_t < n_s$, then densities can be estimated as coefficients in a linear regression forced through the origin. One might argue that, as U_s will be integral in the population case, we should use Poisson regression. However, in other cases, such as for gross income, the dependent variable will be the sum of continuous variables and the Gaussian error model implicit in ordinary least squares (OLS) will be more appropriate.

If $n_t > n_s$, the values of the d_t will be indeterminate unless further assumptions can be made. Flowerdew and Green (1989) assume that target zones fall into a smaller number of classes than the number of source zones. In effect, they use an aggregation of target zones ($n_t = 2$) for the estimation of source-zone densities, and a subsequent disaggregation for the estimation of target-zone populations. As spatial arrangements are represented in these methods only through the areas of overlap, zones can be aggregated whether or not they are actually contiguous in space.

Generalization 2: control zones

Suppose neither source nor target zones have homogeneous densities. If the analyst has access to a third set of zones, here referred to as control zones, which are believed to have constant densities, these can be employed in an intermediate step to derive target-zone estimates. We will denote these control zones by the index c , and the unknown density of these zones by d_c . The areas of overlap between source zone s and control zone c will be denoted as a_{cs} and those between control zone c and target zone t by b_{ct} . If we assume that there are fewer control zones than source zones ($n_c < n_s$), the control-zone densities can be estimated as before from the equation:

$$U_s = \sum_c d_c a_{cs}, \quad (5)$$

with the known source-zone populations and areas of overlap. The target-zone populations can then be derived by integrating the piecewise population density surface represented by the d_c over the areas of each target zone:

$$V_t = \sum_c d_c b_{ct}. \quad (6)$$

Note that there are no restrictions on the number of target zones in this case. Flowerdew and Green (1989) created control zones by using surrogate information presumed to be helpful in estimating densities, but this requires the control zones to be congruent with target zones (or aggregates of target zones).

As the control zones do not need to be congruent either with source zones or with target zones, we see a major advantage of this approach in the freedom which it gives the analyst to introduce additional knowledge into the estimation process. The boundaries of the control zones need not be defined very carefully, but may simply express the analyst's personal experience of the area in relatively imprecise and general terms. In the next section, we discuss database and estimation issues

related to the approach outlined above. This will be followed by an illustration of the use of this method in a simple estimation problem.

Implementation issues

We have argued that the problem of basis change for socioeconomic data can be seen as one of estimating one or more underlying density surfaces. The different methods reviewed above amount to alternative sets of assumptions about the nature of the underlying surface(s). In this section, we first present a spatial database perspective on the alternatives and then discuss a number of issues concerned with the implementation of a general set of procedures for basis change.

Database issues

Various methods have been devised for the digital representation of the spatial variation of a parameter such as p (Burrough, 1986). The alternatives fall generally into three classes.

- (1) Point sampling: the value of the variable is obtained at a finite set of points, often arrayed as a grid or raster. Tobler (1979) uses this representation for estimates of p in his pycnophylactic method.
- (2) Contours: if p is continuous, a sampling of the domain of p defines a number of linear point sets or contours, which can be approximated by a finite number of points connected by straight line segments. However, this representation is inefficient as a method of sampling the spatial variation of p ; it provides high accuracy in the neighborhood of each contour, but much lower accuracy elsewhere.
- (3) Piecewise approximation: as noted earlier, the plane can be partitioned into zones within which variation is assumed to follow some simple mathematical function. Digitally, the zones are approximated by polygonal boundaries (points connected by straight lines) with associated attributes defining the chosen function.

This last data model is an obvious choice for any method of basis change in which the areas of intersection between zones are used, as these areas can be obtained by invoking the GIS procedure of polygon overlay. If p is constant within zones, then the entire surface can be described by attaching a single value to each polygon; discontinuities in the surface occur in general across all zone boundaries. Tobler's objective of maximum smoothness in the density surface in the pycnophylactic method stands in sharp contrast to this model (Goodchild and Lam, 1980). In essence, in this model, one can learn nothing about the value of p at a point by looking cross a zone boundary at adjacent points, however close they may be.

If the variable of interest is topographical elevation, it is common to use a variant of piecewise approximation known as the TIN (triangulated irregular network) (Burrough, 1986), in which the variation within each zone is assumed to be planar. To avoid discontinuities in the value of p across zone boundaries, it is necessary for the zones to be triangles. However, this model seems inappropriate for socioeconomic data, where there are no parallels to the propensity of geomorphological processes to create areas of constant slope.

From a database perspective, the polygonal model with zones of constant density is compatible both with the likely form of data input, as population values for irregular zones, and with the likely required form of output. Unlike the pycnophylactic method, it avoids the need for a further internal form of representation of the density surface. Methods in which basis change is used and which are based on intersection areas can be carried out by straightforward extensions of the polygon overlay operation, by processing the polygonal representation of various sets of zones

Estimation issues

By using the methods outlined above, one can change the spatial basis from one set of reporting zones to another, provided that one is willing to assume that source zones, target zones, or some third set of control zones have uniform density. For many socioeconomic variables (such as population, income, or employment) the data do not take on negative values. In the approach outlined above, this may lead to a problem when the predicted values for the target zones are negative. Indeed, a simple application of OLS regression does not necessarily yield positive predicted values. In order to avoid this problem, a number of alternative approaches can be taken. In the context of areal interpolation, the most common suggestion is to use Poisson regression, which is ideally suited to deal with count variables and ensures positive predicted values. This is the approach taken in the work of Flowerdew and associates (Flowerdew and Green, 1989; Flowerdew et al, 1991). However, a Poisson regression is based on a very limiting assumption that the mean of the process equals the variance. In other words, a number of commonly encountered problems, such as spatial dependence, are 'not allowed'.

An alternative to the Poisson approach is to impose constraints on the coefficients in an OLS regression. This is a special case of incorporating additional information in the estimation procedure (for an overview, see Judge and Yancey, 1986). Two sets of constraints could be considered. In the first, the estimated coefficients are forced to satisfy, for example, the total for population in the region. Clearly, the total population (say P) equals the sum of the values for the source zones:

$$P = \sum_s U_s. \quad (7)$$

This condition is implemented by imposing a linear constraint on the coefficients of the form:

$$P = \sum_c d_c \left(\sum_s a_{cs} \right), \quad (8)$$

where the last term, in parentheses, in equation (8) is nothing but the total for each control zone. Such a linear constraint can be easily incorporated into most commercial regression software. Although this approach makes sure that the sum over the target zones corresponds to the total for the source zones, it does not guarantee that the estimated coefficients are positive. The latter condition can be enforced in two ways in an ordinary regression. One approach would be to set the negative coefficient(s) to zero or to some other value by using some prior (again possibly subjective) estimate of density. Of course, the other coefficients would need to be reestimated, and the nature of the statistical inference in the model changes substantially (see Judge and Yancey, 1986). Clearly, these prior values can be imposed in combination with the constraint expressed in equation (8) on the total. Another approach is to take the Bayesian perspective, suggested by Geweke (1986; 1988), where a prior distribution takes the form of a product of a conventional uninformative distribution with an indicator function that represents the inequality constraint. Inference is based on a posterior distribution that is computed by means of Monte Carlo numerical integration (for technical details, see Geweke, 1986; 1988). The Geweke approach is implemented in some specialized econometric software, such as Shazam (White, 1988).

An alternative is to move away from the standard regression model and to implement a lognormal regression. This type of regression imposes the positive nature of the dependent variable but results in a particular form of heteroscedasticity

In a lognormal regression, the dependent variable, y , is assumed to follow a lognormal distribution, in contrast to the standard assumption of normality. Formally, a lognormal regression can be expressed in the form

$$y = X\beta + \varepsilon, \quad (9)$$

where the error term, ε , follows a lognormal distribution. As a result, the expected value of y , $E(y)$, is given by $E(y) = X\beta$, but its variance, $V(y)$, is given by $V(y) = \sigma^2(\beta'x)^2$, where σ^2 is the variance of the lognormal distribution and $\beta'x$ is the predicted value for an observation. In the lognormal regression, the variance is thus proportional to the mean, which necessitates a maximum likelihood approach towards estimation (for details, see Amemiya, 1973). The lognormal regression technique is implemented in specialized econometric software, such as Limdep (Greene, 1991).

The lognormal model might be considered the most appropriate estimation method in terms of the preferred error structures (errors due to the multiplication of a large number of unknown effects) compared with OLS (additive errors), or the Poisson model, in which only Poisson sampling error is assumed—in reality the assumption of uniform densities is a larger source of error. Moreover, the lognormal model has a variance-stabilizing effect, which is commonly exploited in econometric practice (Greene, 1991; Spanos, 1986).

It is also possible for the analyst to introduce additional prior knowledge of certain control-zone or target-zone densities. This may occur in cases where more information about a particular subarea of the study region is available than about the rest of the region. Let the control zones be divided into two sets, X and Y , where the densities of set X are known and those of set Y are unknown. Then equation (5) becomes

$$U_s - \sum_X d_c a_{cs} = \sum_Y d_c a_{cs}, \quad (10)$$

where the parameters on the left-hand side are all known. For least-squares estimation we now require $n_Y < n_s$. Moreover, the structure of the matrix A in equation (5) [and appropriate submatrix in equation (10)] must be such that no column (variable, in the regression sense) is a linear function of any other column or columns. This condition will arise, for example, when two control zones in set Y intersect the same source zone, and no other zone, and becomes increasingly likely as more control zones move to set X (have known densities). An implemented package would need to assist the analyst by identifying such conditions.

In general, we expect alternative sets of reporting zones for a given area to share certain common boundary segments and for the matrix A to be sparse. However, A may be less sparse than expected because in practice the two digitized versions of a supposedly common line will not in fact coincide, and small sliver polygons with finite area will produce nonzero entries in certain cells of A (Goodchild, 1978). These small nonzero entries can propagate into significant spurious estimates of population if they occur in columns which would otherwise be collinear.

Illustration using control zones

The data

We now illustrate the use of control zones with an example application for regions in the state of California. As is typical of all states in the USA, most socioeconomic data for California are published on the level of the fifty-eight major civil divisions or counties. Data related to water issues, however, are gathered and distributed according to twelve hydrologic study areas that follow major watershed boundaries

As can be seen in figure 1, the boundaries of the two zonal systems are largely incompatible.

The information published for the twelve basins consists not only of hydrologic data, but also data that are of direct relevance for socioeconomic modeling and that are available only on this spatial basis (for example, water use by major industry, or residential water consumption). The California Department of Water Resources (CDWR, 1980) developed a multiregional input-output model for the twelve hydrologic basins for an economic impact study. For this purpose, all output, employment, income, and population data had to be transferred from the basis of counties to the hydrologic regions. The areal interpolation method applied in this study consisted of direct areal weighting from counties to basins, with uniform source-zone densities assumed.

In a similar study, Anselin et al (1990) evaluated the effect of using two different weighting schemes upon the results of a multiregional activity analysis model for the state of California, where the base regions were aggregates of counties. The model was used to assess how population dynamics translate into changes in final demand for industrial and agricultural sectors. These in turn lead to changes in water use by region and thus to a shift in the pattern of water flows among the hydrologic study areas. The areal weighting method was compared with a weighting scheme based on population figures for 5757 census tract and enumeration districts in California. The data provide population counts for representative points within each zone, with census tracts (average population about 4000) used in major cities and enumeration areas (average population about 500) used elsewhere. This information is available from the Status Census Center of the California Department of Finance (CDF, 1982) for the population census year, in this case 1980. When included in a GIS, this coverage allows for easy reaggregation to any set of zones. The model results showed that, depending on the choice of the weighting method, the optimal flows vary considerably in magnitude as well as in direction of flow. Consequently, the choice of an interpolation strategy has a strong influence on model results and thus on potentially far-reaching policy decisions.

In many actual applications of areal interpolation, ancillary data, such as the census tract centroids, are unavailable. In the following paragraphs it will be demonstrated how the areal interpolation can be improved in the absence of such

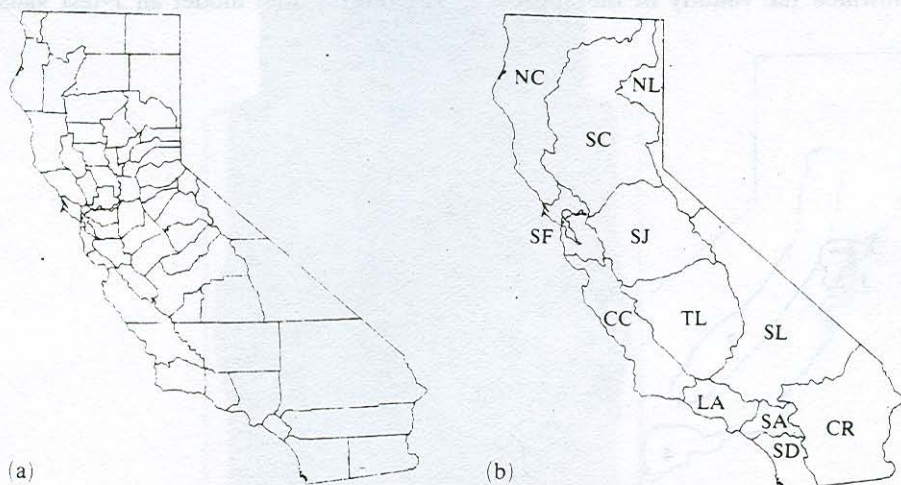


Figure 1. (a) The fifty-eight counties in California (source zones), and (b) the twelve hydrologic basins (target zones). Note: see table 2 for fuller description of basins.

data by introducing a third set of control zones. The 'truth' against which the performance of the method will be evaluated is the results of the reaggregation of the census tract centroids and associated population figures. Although this approach will still have an inherent unknown error because of incompatibility between census tracts, enumeration districts, and hydrologic basins, for the purpose of most analysis the accuracy thus attained should be sufficient.

The control zones in this example are based on four subjective divisions that are assumed to show a homogeneous or constant population density. These are the urbanized regions in the Bay area, the Southern California metropolitan area, the agricultural region of the Central Valley, with the rest of the state as a residual. By means of a GIS these regions can be designed interactively by a user who is familiar with the study area. The state of California displays extreme population density gradients even within counties, so the assumption of a uniform source-zone density made by Goodchild and Lam (1980) and in the earlier input-output study is clearly highly inaccurate. The large desert counties of Southern California (Riverside and San Bernardino) contain extensive unpopulated areas and parts of the heavily populated Los Angeles basin, and similar contrasts can be found elsewhere in the state. Our purpose is to show that a small amount of subjective information can produce a substantial improvement in the accuracy of the results when compared with the naïve assumption of uniform source-zone densities.

The four control zones shown in figure 2 were entered interactively by eye on a screen display of the entire state by using a screen cursor. We estimate a positional accuracy of 10 km for this step, compared with much higher positional accuracies (better than 1 km) for the source-zone and target-zone boundaries and for the reference point data set. Equation (5) was used to estimate the four unknown control-zone densities. We carried out this estimation by using four methods in addition to OLS, in order to guarantee that the total population for the state equals the sum of the interpolated parts and/or to deal with the positive nature of the variable (population). As is well known, standard unconstrained OLS does not necessarily satisfy either condition. We implemented two constrained least squares (LS) estimates, one in which the overall sum of population in the state is imposed (LS Sum), and one in which, in addition, the negative coefficient estimate was set to zero (imposed as a prior constraint; LS $b_1 = 0$ Sum). The test on the significance of the constraints confirmed the validity of the approach. For the LS Sum model an F -test value of



Figure 2. The four control zones.

0.167 (probability 0.685) was obtained, for the LS $b_1 = 0$ Sum model the value was 0.113 (probability 0.893), which means that the null hypothesis (that the constraints hold) cannot be rejected. Diagnostics showed no evidence of spatial error autocorrelation but did give a strong indication of heteroscedasticity.

In addition, we implemented two nontraditional estimation techniques: a Bayesian approach in which we imposed positivity on the coefficients (Bayes LS Ineq) and a lognormal regression (Lognormal). Either of these two methods necessarily satisfies the total population constraint without imposing this explicitly (as an additional constraint on the estimated coefficients).

By using the estimates for the coefficients in equation (5), we interpolated the target zone populations by means of equation (6). The accuracy of the technique was determined by comparing target-zone populations with the values obtained by aggregating the point data set. The latter were assumed to be true: no attempt was made to evaluate the error inherent in the aggregation. For comparison, the results for an interpolation in which area weights were used (where the source zones were assumed to have homogeneous densities) and for a pycnophylactic interpolation were included as well. The pycnophylactic interpolation was performed on a grid of 540 rows and 460 columns, corresponding to an approximate cell size of 2 km by 2 km. As the boundary constraint, the Neumann condition (the rate of change of the densities across the boundary was set to zero) was seen as being most appropriate given the high concentration of population along the California coast (see Tobler, 1979).

Results

In table 1 we show the coefficient estimates (densities) for each of the four control zones, arranged by estimation methods. The standard measure of population density (persons km^{-2}) is derived by multiplying the coefficients by one thousand. The very sharp contrast between the high densities of the major urban areas (691.4 persons km^{-2} in zone 4 according to the point data) and the much lower densities

Table 1. Coefficient estimates, by method.

Model	Zone 1	Zone 2	Zone 3	Zone 4
Actual	0.0115	0.0597	0.4460	0.6914
Unrestricted OLS	-0.0027	0.0615	0.4790	0.8645
Standard error	0.0065	0.0313	0.1372	0.0533
Robustness	0.0035	0.0104	0.1123	0.2005
LS Sum	-0.0013	0.0671	0.5046	0.8691
Standard error	0.0054	0.0281	0.1220	0.0521
LS $b_1 = 0$ Sum	0.0000	0.0628	0.4929	0.8638
Standard error	na	0.0219	0.1123	0.0474
Bayes LS Ineq	0.0043	0.0572	0.4722	0.8511
Standard error	0.0034	0.0283	0.1405	0.0543
Lognormal	0.0021	0.0289	0.3665	0.3998
Standard error	0.0004	0.0091	0.1265	0.2783

na—not applicable.

Note: Standard population densities (persons km^{-2}) are derived by multiplying the coefficient by one thousand.

Unrestricted OLS, ordinary least squares without constraints on the coefficients; LS Sum, least squares regression with total population constraint; LS $b_1 = 0$ Sum, least squares regression with negative coefficient set to zero a priori and with total population constraint; Bayes LS Ineq, Bayesian regression imposing positivity of coefficients; Lognormal, lognormal regression imposing positivity of coefficients.

of zones 1 and 2 (11.5 persons km^{-2} in zone 1) was evident. Note that all but the Lognormal model overestimate the densities of the two major urban areas. In reality the densities of the areas in zones 1 and 2 close to the boundaries of zones 3 and 4 are higher than elsewhere, which tends to bias the zone 3 and 4 density estimates upward.

Table 2 shows the interpolated populations for the hydrologic basins. The percentage error for the individual estimates, as well as the mean percentage error (MPE) and the mean absolute percentage errors (MAPEs) are displayed in table 3. In general, the standard errors of the coefficients as well as the actual interpolation errors are fairly large, which emphasizes the difficulties in areal interpolation for nonhomogeneous regions. For all approaches in which control zone densities were employed, the MPE and MAPE are considerably smaller than for the area weighted estimates and the pycnophylactic approach (for the MAPE, 46.8–67.4 versus 124.2 and 86.1, respectively).

The three LS estimates display a fairly low MPE, but the predicted negative or zero population for some of the hydrologic study areas are clearly unacceptable. The two nonstandard estimation methods score better on the MAPE. However, the Bayesian approach overpredicts the population total, whereas the lognormal, which has the lowest overall MAPE, systematically underpredicts population. These results show that the choice of the estimation method should not be taken lightly. In cases where the source and target zones are more homogeneous than in our example, constrained LS methods might very well produce acceptable results, thus making more complex estimation methods unnecessary.

As noted previously, the intent of this illustration is to show the versatility of the method, and the degree to which a small amount of subjective information, in the form of four crudely defined zones, can improve on the assumption of uniform source-zone density. We do not wish to suggest that four zones are adequate, or that the errors in the estimates shown in table 3 are acceptable for any practical purpose. In reality, it is expected that there would be a direct correlation between

Table 2. Interpolated hydrologic study area populations.

Hydro- logic areas	Actual	Area weights	Model					
			Pycno- phylactic	OLS	LS Sum	LS b1 = 0 Sum	Bayes LS Ineq	Log- normal
NC	434229	492404	444156	-1251	80308	143734	357881	216402
SF	4775448	3954032	4272176	3259269	3439506	3364868	3239524	2510125
CC	1005486	1436609	1192638	1114267	1203311	1215497	1290239	879742
LA	8047510	5627519	6355566	6978795	7020068	6981399	6891388	3237608
SA	2811837	1417893	2267183	2810195	2830749	2818646	2793869	1312892
SD	2002046	1908370	2125383	4414174	4444827	4424376	4380697	2058442
SC	1651086	1499991	1545043	961604	1137687	1139079	1281857	686619
SJ	1018370	1652891	1533092	1078861	1210938	1175041	1190005	666873
TL	1144648	997241	787217	1223787	1374086	1314131	1293549	653959
NL	60649	70254	24851	-44069	-20898	0	68862	34251
SL	316634	3033075	2570096	818233	928881	1004403	1271160	608329
CR	314849	1492515	465391	-144153	-68358	0	225251	112039
Total	23582792	23582792	23582792	22469711	23582792	23582792	24285723	12978426

Note: OLS, ordinary least squares; NC, North Coast; SF, San Francisco; CC, Central Coast; LA, Los Angeles; SA, Santa Ana; SD, San Diego; SC, Sacramento; SJ, San Joaquin; TL, Tulare Lake; NL, North Lahontan; SL, South Lahontan, CR, Colorado River. For definitions of model types, see note to table 1.

the information provided by the analyst and the accuracy of the results. In this case, one would expect continued improvements in the quality of estimates as the number of control zones increases, up to the limit of the number of source zones. But the concept of the control zone seems to provide a relatively efficient way of introducing a limited amount of information into the estimation process.

Table 3. Percentage error of population estimates.

Hydro- logic areas	Area weights	Model					
		Pycno- pyclactic	OLS	LS Sum	LS b1 = 0 Sum	Bayes LS Ineq	Log- normal
NC	13.4	2.3	-100.3	-81.5	-66.9	-17.6	-50.2
SF	-17.2	-10.5	-31.7	-28.0	-29.5	-32.2	-47.4
CC	43.9	18.6	10.8	19.7	20.9	28.3	-12.5
LA	-30.1	-21.0	-13.3	-12.8	-13.2	-14.4	-59.8
SA	-50.6	-19.4	-0.1	0.7	0.2	-0.6	-53.3
SD	-4.7	1.1	120.5	122.0	121.0	118.8	2.8
SC	-9.2	-6.4	-41.8	-31.1	-31.0	-22.4	-58.4
SJ	62.3	50.5	5.9	18.9	15.4	16.9	-34.5
TL	-13.9	-31.2	6.9	20.0	14.8	13.0	-42.9
NL	16.8	-59.0	-172.7	-134.5	-100.0	13.5	-43.5
SL	858.9	711.7	158.4	193.4	217.2	301.5	92.1
CR	374.0	47.8	-145.8	-121.7	-100.0	-28.5	-64.4
MPE	103.6	57.0	-16.9	-2.9	4.07	31.4	-31.0
MAPE	124.2	81.6	67.4	65.4	60.9	50.6	46.8

Note: A negative sign for percentage error indicates an underprediction of hydrologic basin populations, a positive value indicates an overprediction. MPE, mean percentage error; MAPE, mean absolute percentage errors; see also tables 1 and 2 for definitions of models and hydrologic areas, respectively.

Conclusions

The integration of large amounts of spatially referenced socioeconomic data from heterogeneous sources in GIS leads to an increasing need for reliable methods of spatial basis change, particularly to overcome problems of incompatible reporting zones. This seems to be equally true if socioeconomic data have to be interfaced with environmental information, which is likely to be reported for entirely incompatible zones. In this paper, we have proposed that successful basis change requires the identification of one or more acceptable underlying continuous surfaces. Various existing methods have been classified according to the assumptions which they make about these surfaces. In addition, we have looked at the issues involved in representing such surfaces digitally. In essence, the method and the data model used for internal surface representation impose strong and limiting views on the form of spatial variation of the phenomenon.

The generalization of area-of-overlap methods to homogeneous control zones appears to offer substantial potential in allowing the user to input prior knowledge of the area. A package built around this approach and coupled with the functionality of a GIS could implement a flexible and general capability for basis change, provided it were able to interact intelligently with the analyst and give accurate and appropriate summaries of the options and limitations.

The example used as an illustration in this paper is clearly unrealistic, but serves as a simple indication of the improvement in accuracy that is possible with a package that can exploit various forms of prior knowledge. Although in practical application

aggregation error in its various manifestations remains an often intractable problem, the accuracy of target-zone estimates will be directly proportional to the amount and quality of control-zone data utilized. We have described a technique for making use of knowledge of source-zone populations and of control zones whose densities are assumed to be constant and to be either known or unknown. Many forms of prior knowledge seem to fall into this framework, with the notable exceptions of Tobler's (1979) objective of minimum curvature, and that described in the literature on urban density functions. However, Tobler's interpolation could be generalized to allow for discontinuities of density across linear features defined by the analyst.

Acknowledgement. The National Center for Geographic Information and Analysis is supported by the National Science Foundation (grant SES 88-10917). A previous version of the paper was presented at the Third European Conference on Geographical Information Systems, Munich, 1992.

References

- Amemiya T, 1973, "Regression analysis when the variance of the dependent variable is proportional to the square of its expectation" *Journal of the American Statistical Association* **68** 928-934
- Anselin L, Rey S, Deichmann U, 1990, "The implementation of integrated models in a multi-regional system", in *New Directions in Regional Analysis: Multi-regional Approaches* Eds L Anselin, M Madden (Belhaven Press, London) pp 146-170
- Arbia G, 1989 *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems* (Kluwer, Dordrecht)
- Bennett R J, Haining R P, Griffith D A, 1984, "The problem of missing data on spatial surfaces" *Annals of the Association of American Geographers* **74** 138-156
- Bracken I, Martin D, 1989, "The generation of spatial population distributions from census centroid data" *Environment and Planning A* **21** 537-543
- Burrough P A, 1986 *Principles of Geographical Information Systems for Land Resources Assessment* (Clarendon Press, Oxford)
- CDF, 1982 *Population by Census Tract and Enumeration District for 1980 State Census Center*, Population Research Unit, California Department of Finance, Sacramento, CA 95814
- CDWR, 1980, "Measuring economic impacts. The application of input-output analysis to California water resources problems" bulletin 210, California Department of Water Resources, Sacramento, CA 95802
- Clark C, 1951, "Urban population densities" *Journal of the Royal Statistical Society, Series A* **114** 490-496
- Flowerdew R, Green M, 1989, "Statistical methods for inference between incompatible zonal systems", in *Accuracy of Spatial Databases* Eds M F Goodchild, S Gopal (Taylor and Francis, London) pp 239-248
- Flowerdew R, Openshaw S, 1987, "A review of the problem of transferring data from one set of areal units to another incompatible set", RR4, Northern Regional Research Laboratory, Universities of Lancaster and 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
- Geweke J, 1986, "Exact inference in the inequality constrained normal linear regression model" *Journal of Applied Econometrics* **1** 127-141
- Geweke J, 1988, "Antithetic acceleration of Monte Carlo integration in Bayesian inference" *Journal of Econometrics* **38** 73-89
- Goodchild M F, 1978, "Statistical aspects of the polygon overlay problem", in *Harvard Papers on Geographic Information Systems 6: Spatial Algorithms: Efficiency in Theory and Practice* Ed. G Dutton (Addison-Wesley, Reading, MA)
- Goodchild M F, 1992, "Geographical data modeling" *Computers and Geosciences* **18** 401-408
- Goodchild M F, Lam N S-N, 1980, "Areal interpolation: a variant of the traditional spatial problem" *Geo-Processing* **1** 297-312
- Gordon P, Richardson H W, Wong H L, 1986, "The distribution of population and employment in a polycentric city: the case of Los Angeles" *Environment and Planning A* **18** 161-173
- Greene W H, 1991 *Limdep, User's Manual and Reference Guide* (Economic Software Inc, Bellport, NY)

- Griffith D A, Bennett R J, Haining R P, 1989, "Statistical analysis of spatial data in the presence of missing observations: a methodological guide and an application to urban census data" *Environment and Planning A* **21** 1511-1523
- Honeycutt D, Wojcik J, 1990, "Development of a population density surface calculation for the conterminous United States", in *GIS/LIS 1990 Proceedings Volume 1* Anaheim, CA (American Congress on Surveying and Mapping, Washington, DC) pp 484-496
- Judge G G, Yancey T A, 1986 *Improved Methods of Inference in Econometrics* (North-Holland Amsterdam)
- Lam N S-N, 1983, "Spatial interpolation methods: a review" *American Cartographer* **10** 129-149
- Langford M, Unwin D J, Maguire D J, 1990, "Generating improved population density maps in an integrated GIS", in *Proceedings, First European Conference on Geographical Information Systems Volume 2* Eds J Harts, H F L Ottens, H J Scholten, EGIS Foundation, PO Box 80.115 3508 TC Utrecht, pp 651-660
- Mark D M, Csillag F, 1989, "The nature of boundaries on area-class maps" *Cartographica* **2** 65-78
- Martin D, Bracken I, 1991, "Techniques for modelling population-related raster databases" *Environment and Planning A* **23** 1069-1075
- Nordbeck S, 1971, "Urban allometric growth" *Geografiska Annaler B* **53** 54-67
- Openshaw S, 1984 *CATMOG 38. The Modifiable Areal Unit Problem* (Geo Books, Norwich)
- Openshaw S, Taylor P J, 1981, "The modifiable area unit problem", in *Quantitative Geography. A British View* Eds N Wrigley, R J Bennett (Routledge, Chapman and Hall, Andover, Hants) pp 60-70
- Parr J B, 1985, "The form of the regional density function" *Regional Studies* **19** 535-546
- Silverman B W, 1986 *Density Estimation for Statistics and Data Analysis* (Routledge, Chapman and Hall, Andover, Hants)
- Spanos A, 1986 *Statistical Foundations of Econometric Modelling* (Cambridge University Press, New York)
- Tobler W R, 1979, "Smooth pycnophylactic interpolation for geographical regions" *Journal of the American Statistical Association* **74** 519-530
- White K J, 1988 *Shazam Econometrics Computer Program* (McGraw-Hill, New York)