A GENERAL FRAMEWORK FOR THE SPATIAL INTERPOLATION OF SOCIOECONOMIC DATA

Uwe Deichmann[1], Michael F. Goodchild[1] and Luc Anselin[1]

Abstract

Spatial data are collected and represented as attributes of spatial objects embedded in the plane. We define basis change as the transfer of attributes from one set of objects to another. Methods of basis change for socioeconomic data are reviewed and seen to differ in the assumptions they make about underlying density surfaces. We extend these methods to more general cases and provide an illustration using California data. The implementation of this framework within a GIS is discussed.

INTRODUCTION

Geographic Information Systems are being increasingly used to support a variety of forms of spatial and regional analysis, using social, economic and demographic data. The databases which these systems access are discrete digital representations of real spatial variation using one of a number of standard discretizations or **data models** (for review see Burrough 1986). Variation can be described by attaching attributes to a set of spatial **objects**, which may be points, lines or areas, embedded in the 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 own 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

[1] National Center for Geographic Information and Analysis, University of California, Santa Barbara, Calif. 93106, USA

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 socio-economic reporting zones leads immediately to the difficulties known as the Modifiable Areal Unit problem (Openshaw and Taylor, 1981; Openshaw, 1984), or 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. 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 is available (Bennett *et al.*, 1984; Lam, 1983). A different case might be where data is available for a set of points but, is needed for a system of areas or zones (Bracken and Martin, 1989). Arbia (1989) has described the general problem of basis change and outlined an approach based on a space partitioned into discrete units. In this paper we assume space to be continuous, and feel that our approach is therefore more general.

The commonest 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 (**source** and **target** zones, respectively). This has been referred to as areal interpolation (Goodchild and Lam 1980; Lam 1983; Flowerdew and Openshaw 1988), and arises because of the general lack of coordination between agencies. Different departments, or even different offices in the same department, publish data based on a number of zonal arrangements that are often incompatible (e.g., counties, Standard Metropolitan Statistical Areas, labor market areas, hydrologic study areas, etc.). Reporting 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 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. The purpose of 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 $\{x,y,z_1,z_2,...,z_n\}$ where $z_1,z_2,...,z_n$ represent the values of n variables at location (x,y). Since both x and y are measured on continuous scales, 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 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:

$$P_i \;=\; {}_{Ai}\, p \; dA \tag{1}$$

where p(x,y) is the unobservable population density, A_i is the area of zone i and P_i is the zone's population. The minimum size of A depends on the variable. While 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 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 p 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 variable 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 using a radially symmetric function such as:

$$p(r) = p_0 e^{-br} \tag{2}$$

where p_0 is the central density and b is a constant. The literature records abundant tests of this and similar functions (Clark, 1951; Parr, 1985). However these monocentric models seem to be becoming less accurate with the transition to polycentric cities, particularly in North America. 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, but not have additional information on variation of density within it. Nevertheless, Honeycutt and Wojcik (1989) have recently used a circular Gaussian function to estimate a continuous surface of population density for the coterminous United States 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 numbers of the smaller units make this impractical. The geographical locations of the Enumeration District or Enumeration Area level (100-200 households) in the United States, United Kingdom and Canada are available in digital form only as "centroids" or representative points, usually placed in the area of highest density within the zone. Bracken and Martin (1989) describe a technique for estimating a continuous density surface from this data, as an essential step in basis change. A **kernel** function is defined to distribute the population at each centroid uniformly over a disc whose radius 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) proposes the term **pycnophylactic** 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, $i=1,...,n$, each with known area A_i and known population P_i. An estimation method for $p(x,y)$ is pycnophylactic if equation (1) above holds for every zone. Clearly this condition is not sufficient to estimate the surface, and many methods will have the pycnophylactic property. Tobler's method establishes additional boundary conditions on the surface p, for example that $p=0$ at the edge of the partitioned area, or that the gradient of p is zero at the edge. An iterative procedure finds 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

These methods partition the plane into zones of relative homogeneity, and then model the variation of density within each zone 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.

Goodchild and Lam (1980) examined the case where one can reasonably assume that the **source** zone density is constant. The derivation of 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 by:

$$V_t = \Sigma_s U_s (a_{st} / \Sigma_t a_{st}) \tag{3}$$

Note that as is appropriate for a spatially extensive variable such as population, the elements of the areal weight matrix are standardized over 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.

Flowerdew and Green (1989) describe the case where auxiliary information is available to assist in estimating target zone densities. In their example, the target zones are political voting districts, and a binary variable (Conservative or Labour Party representation) is available for each target zone. Letting ${}_1$ and ${}_2$ represent the population densities of Conservative- and Labour-held districts respectively, the expected population of source zone s is given by ${}_1A_1 + {}_2A_2$, where A_1 and A_2 are the areas of overlap between the source zone and Conservative- and Labour-held target zones respectively. The parameters ${}_1$ and ${}_2$ are estimated by regressing the known source zone population on the known overlap areas, using Poisson regression since the observed source zone population is integer and presumed to be sampled under a Poisson distribution.

Generalization I: uniform target zone densities

In this and the subsequent section we generalize the Goodchild and Lam (1980) and Flowerdew and Green (1989) approaches. Suppose that for some reason or other the population distribution can be assumed to be constant within the target zones for which data has to be derived. This might be the case where the source zones are drawn as arbitrary political boundaries, whereas the target zones follow clearly distinguishable physical features (e.g., steep slopes versus grass land) in which the homogeneity assumption is reasonable. Provided that the number of source zones n_s is larger than the number of target zones n_t, population densities for the target zones, d_t, can be estimated using the observed source zone populations, U_s, and the areas of overlap of source zone s with each target zone t:

$$U_s \;=\; \Sigma_t\, d_t\, a_{st} \qquad (4)$$

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. Since U_s will be integer in the population case, we should use Poisson regression. However, in other cases such as gross income, the dependent variable will be the sum of continuous variables, and the Gaussian error model implicit in OLS will be more appropriate.

If $n_t>n_s$, the d_t's will be indeterminate unless further assumptions can be made. Flowerdew and Green (1989) assume that target zones fall into a smaller number of classes. 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. Since 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 II: control zones

Suppose the analyst has access to a third set of zones, here referred to as control zones, which are believed to have constant densities. 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 by a_{cs} and between control zone c and target zone t by b_{ct}. Assuming that there are fewer control zones than source zones ($n_c<n_s$), the control zone densities can be estimated as before using the equations:

$$U_s = \Sigma_c \, d_c \, a_{cs} \qquad (5)$$

using 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's over the areas of each target zone:

$$V_t = \Sigma_c \, d_c \, b_{ct} \qquad (6)$$

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

Since the control zones can be independent of both source and 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 describe an illustration of the use of this method to a simple estimation problem.

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 United States, most socioeconomic data for California are published on the level of the 58 major civil divisions or counties. Data related to water issues, however, are gathered and distributed according to 12 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.

[Figure 1 about here]

The information published for the 12 basins consists not only of hydrologic data, but also data that is of direct relevance for socioeconomic modeling and that is available only on this spatial basis, e.g., water use by major industry, or residential water consumption. Recently, the California Department of Water Resources (1980) developed a multiregional input-output model for the 12 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 assuming uniform source zone densities. In the following paragraphs it will be demonstrated how such estimates can be improved by introducing a third set of control zones.

The performance of the method will be evaluated using population figures for 5377 census tract and enumeration districts in California. The data provide population counts for representative points within each zone, using census tracts (average population about 4,000) in major cities and enumeration areas (average population about 500) elsewhere. This information is available from the State Census Center of the California Department of Finance (1982) for the population census years, in this case 1980. When included in a GIS, this coverage allows for easy reaggregation to any set of zones. While this approach will still have an inherent unknown error because of incompatibility between census tracts, enumeration districts and hydrological basins, for the purpose of most analysis the accuracy thus attained should be sufficient.

The control zones in this example application 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, and 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 uniform source zone density assumption of Goodchild and Lam (1980) and the earier input-output study are clearly highly inaccurate. The desert counties of Southern California (Riverside and San Bernardino) contain both 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 by eye on a screen display of the entire state, using a screen cursor. We estimate a positional accuracy of 10km for this step, compared with much higher positional accuracies (better than 1km) for the source and target zone boundaries and for the reference point dataset. Equation (5) was used to estimate the four unknown control zone densities by OLS, and equation (6) was then used to estimate target zone populations. The accuracy of the technique was determined by comparing target zone populations with the values obtained by aggregating the point dataset. The latter were assumed to be true: no attempt was made to evaluate the error inherent in the aggregation.

[Figure 2 about here]

Technical Issues

A technical problem was encountered in the OLS estimation of the population densities of the regions. For Zone 1 (the rest of California), an area that has a very low density, the regression resulted in a negative coefficient. This can be avoided by constrained regression (e.g., Stein rule estimation, Judge and Yancey 1986) or by setting the problem coefficient to zero and reestimating the remaining coefficients, or by using some prior (again possibly subjective) estimate of density. Table 1 shows the population densities for each of the regions as derived from the point coverage, and by estimation where the density for Region 1 is constrained to zero, and by estimation constraining Zone 1 to its true value as obtained from the point data.

[Table 1 about here]

The table shows the very sharp contrast between the high densities of the major urban areas (691.4/km^2 in Zone 4 according to the point data) and the much lower densities of Zones 1 and 2 (11.5/km^2 in Zone 1). Note that the method overestimates 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 estimated populations for the hydrologic basins, the percentage error for the individual estimates, and the mean absolute percentage errors (MAPE). For both approaches that employ the control zone densities, the MAPE is considerably smaller than for the direct area weighted estimates (59.6 and 86.5 versus 124.2). Table 3 presents a number of summary statistics of the agreement between known and estimated hydrologic basin population.

[Table 2 about here]

[Table 3 about here]

Because equation (5) estimates the d_c's by forcing through the origin, the total of the estimates of the U_s's does not in general equal the known total state population. Nor is there an accounting constraint in equation (6), as Table 2 shows. The use of OLS ensures that R^2 is maximized in the estimation of the U_s's, but does not ensure the maximization of R^2 or any of the other measures of goodness of fit between the known and estimated V_t's shown in Table 3. In general, we do not know whether the goodness of fit measures in Table 3 would improve or deteriorate if an accounting constraint were imposed by proportionally adjusting the estimates.

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, there is a direct correlation between the information provided by the analyst and the accuracy of the results. In this case, one would expect continued improvement 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.

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 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, as follows:

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.

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.

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 data model is an obvious choice for any method of basis change that uses the areas of intersection between zones, as the latter 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 polygonal boundary; 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), which says in essence that one can learn nothing about the value of p at a point by looking across 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, or 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. Basis change using methods 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

Using the methods outlined above, the spatial basis can be changed from one set of reporting zones to another provided that the analyst is willing to assume that either source zones, target zones or some third set of control zones have uniform density. In the case of target and control zones, densities are estimated by OLS or Poisson regression. It is also possible for the analyst to introduce additional prior knowledge of certain control or target zone densities. Let the control zones be divided into two sets, A and B, where the densities of set A are known and those of set B are unknown. Then equation (5) becomes:

$$U_s - \Sigma_A d_c a_{cs} = \Sigma_B d_c a_{cs} \tag{7}$$

where the parameters on the left hand side are all known. For OLS we now require $n_B<n_s$. Moreover the structure of the matrix **A** in (5) (and appropriate submatrix in (7)) 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 B intersect the same source zone, and no other zone, and becomes increasingly likely as more control zones move to set A (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.

CONCLUSION

The integration of large amounts of spatially referenced socioeconomic data from heterogeneous sources in geographic information systems 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 has to be interfaced with environmental information, which is likely 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. We have described a technique for making use of knowledge of source zone populations, and control zones whose densities are assumed to be constant, and 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 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.

REFERENCES

Arbia, G., 1989, *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems.* Kluwer, Dordrecht.

Bennett, R.J., R.P. Haining and D.A. Griffith, 1984, "The problem of missing data on spatial surfaces." *Annals of the Association of American Geographers* 74:138-56.

Bracken, I. and D. Martin, 1989, "The generation of spatial population distributions from census centroid data". *Environment and Planning A* 21:537-43.

Burrough, P.A., 1986, *Principles of Geographical Information Systems for Land Resources Assessment.* Clarendon, Oxford.

California Department of Finance, 1982, "Population by census tract and enumeration district for 1980." State Census Center, Population Research Unit, Sacramento.

California Department of Water Resources, 1980, "Measuring economic impacts. The application of input-output analysis to California water resources problems." Bulletin 210, Sacramento.

Clark, C., 1951, "Urban population densities." *Journal of the Royal Statistical Society A* 114:490-6.

Flowerdew, R. and M. Green, 1989, "Statistical methods for inference between incompatible zonal systems." In M.F. Goodchild and S. Gopal, editors, *Accuracy of Spatial Databases*, pp. 239-48. Taylor and Francis, London.

Flowerdew, R. and S. Openshaw, 1988, "A review of the problem of transferring data from one set of areal units to another incompatible set." Northern Regional Research Laboratory, Research Report No. 4.

Goodchild, M.F., 1978, "Statistical aspects of the polygon overlay problem." In *Harvard Papers on Geographic Information Systems* 6. Addison-Wesley, Reading, MA.

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

Honeycutt, D. and J. Wojcik, 1989, "Population density surface calculation for the coterminous United States." Presented at ESRI Users Conference '89, Palm Springs, Calif.

Judge, G.G. and T.A. Yancey, 1986, *Improved Methods of Inference in Econometrics*. North Holland, Amsterdam.

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

Mark, D.M. and F. Csillag, 1989, "The nature of boundaries on area-class maps." *Cartographica* 26:65-78.

Openshaw, S., 1984, *The Modifiable Areal Unit Problem*. Concepts and Techniques in Modern Geography (CATMOG) No. 38. Geo Books, Norwich, UK.

Openshaw, S. and P.J. Taylor, 1981, "The modifiable areal unit problem." In N. Wrigley and R.J. Bennett, editors, *Quantitative Geography: A British View*. Routledge, London.

Parr, J.B., 1985, "The form of the regional density function." *Regional Studies* 19:535-46.

Silverman, B.W., 1986, *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Tobler, W.R., 1979, "Smooth pycnophylactic interpolation for geographical regions." *Journal of the American Statistical Association* 74:519-30.

FIGURE CAPTIONS

1. a) The 58 counties of California (source zones)

   b) The 12 hydrologic basins (target zones)

2. The four control zones

TABLE 1
Estimated Population Densities for Control Zones
Population / sq km

	Point Based	Dens 1 = 0	Dens 1 set a priori
Zone 1 (Rest of Cal)	11.5	0.0	11.5
Zone 2 (Central Valley)	59.7	58.5	46.1
Zone 3 (San Francisco)	446.0	476.7	467.3
Zone 4 (Los Angeles)	691.4	859.4	838.5

TABLE 2
Population Estimates for Hydrological Study Areas

	Point Based Estimate	Estimated Population Area Weights	Perc. Error	Estimated Population, Dens 1 = 0	Perc. Error	Estimated Population Dens 1 set a priori	Perc Error
NC	434229	492404	13.40	138999	-67.99	714822	64.62
SF	4775448	3954032	-17.20	3254029	-31.86	3232440	-32.31
CC	1005486	1436609	43.88	1183665	17.72	1468564	46.06
LA	8047510	5627519	-30.07	6945949	-13.69	6810930	-15.37
SA	2811837	1417893	-50.57	2804334	-0.27	2780204	-1.13
SD	2002046	1908370	-5.68	4401910	119.87	4351458	117.35
SC	1651086	1499991	-9.15	1065513	-35.47	1492151	-9.63
SJ	1018370	1652891	62.31	1102597	8.27	1200062	17.84
TL	1144648	997241	-13.88	1222762	6.82	1218579	6.46
NL	60649	70254	16.84	0	-100.00	180935	198.33
SL	316634	3033075	858.91	993063	213.63	1710843	440.32
CR	314849	1492515	374.04	0	-100.00	591852	87.98
	23582792	23582792	124.16*	23112820	59.63*	25752841	86.45*

* Mean *Absolute* Percentage Error

Hydrological Study Areas: 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

TABLE 3
Summary Statistics for the Estimation of Basin Population

	Area Weights	Dens 1 = 0	Dens 1 set a priori
RMSE	1337363.3	1016017.7	1087321.1
MAE	837921.7	608633.0	675863.3
R^2	0.44	0.80	0.66
MAPE	124.16	59.63	86.45

RMSE:	Root Mean Squared Error
MAE:	Mean Absolute Error
R^2:	Correlation With Actual Population
MAPE:	Mean Absolute Percentage Error

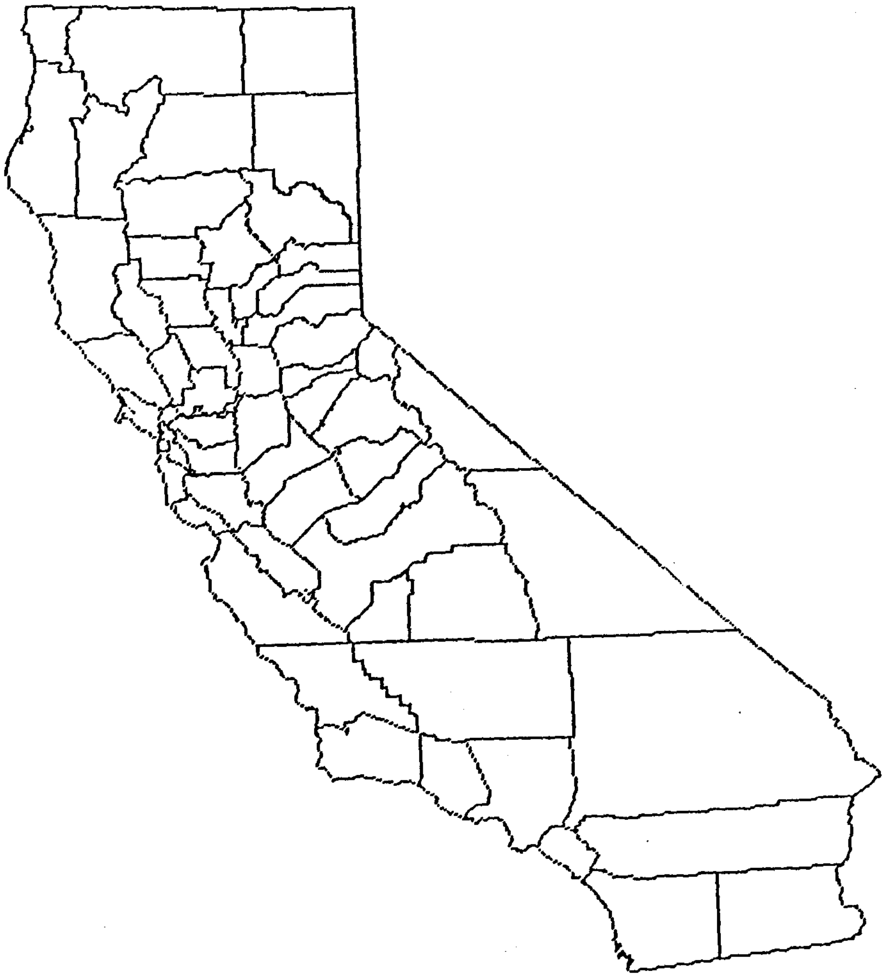

Figure 1a: The 58 counties of California (source zones).

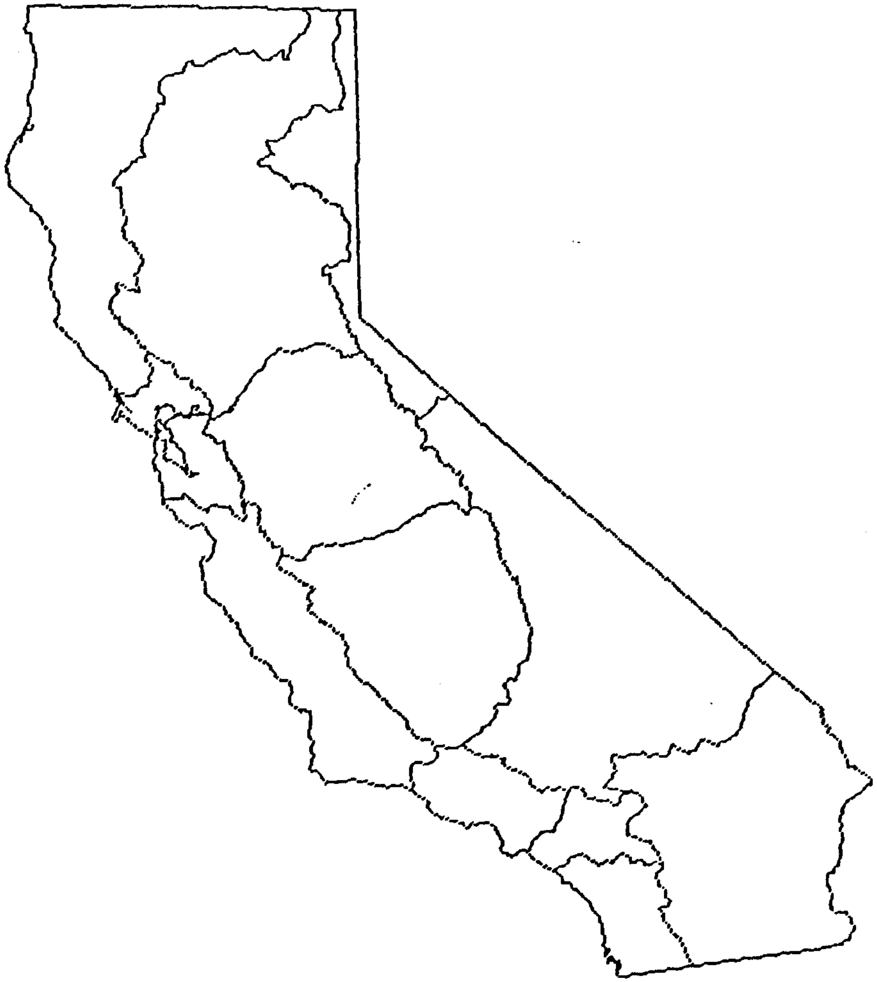

Figure 1b: The 12 hydrologic basins (target zones).

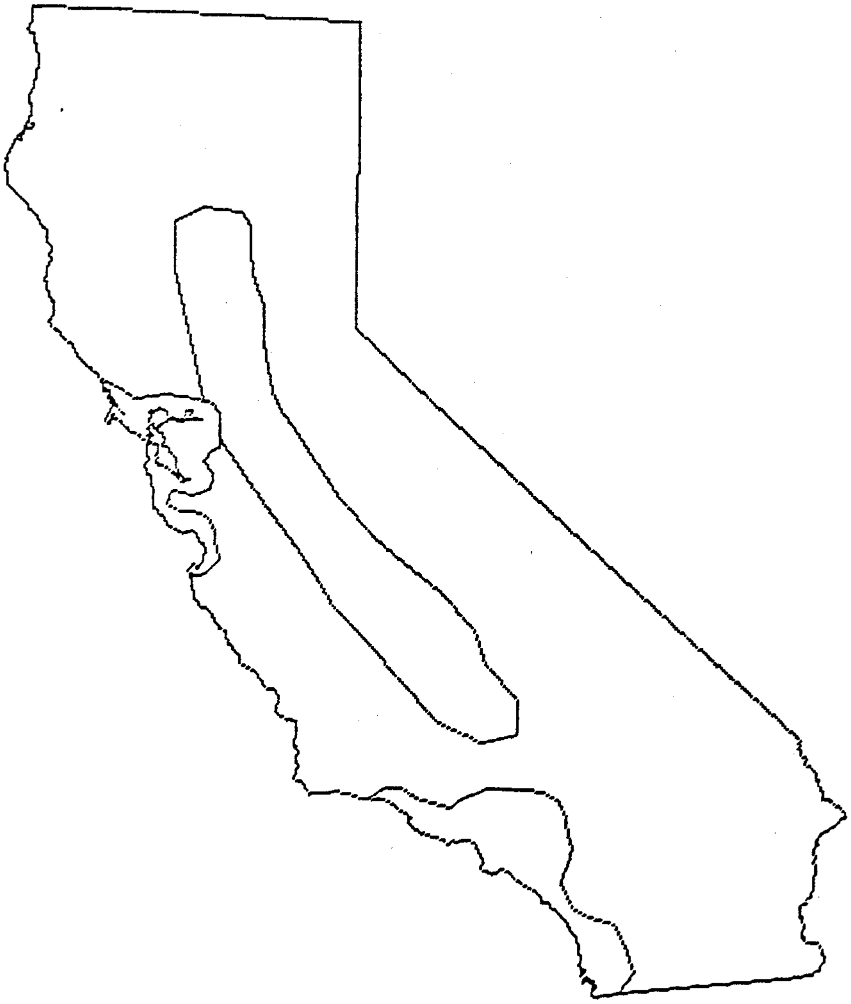

Figure 2: The four control zones