PREAMBLE

A man should first direct himself in the way he should go.
Only then should he instruct others.

Buddha

This maxim is being exercised here by Upton, who reports on results stemming from problems encountered in studying the geographic distribution of voting changes over time. Statistical analysis of data can tell us a great deal about the nature of reality, although as yet it does not fully disclose a deep understanding of geo-referenced data. To this end, the future of spatial statistics in our age of sophisticated computing often is hotly debated, as can be seen in the earlier commentary made by Ripley on Ord, and by Ord on Griffith, for example, in this volume. Upton's paper follows the emerging tradition of computer-intensive statistical analysis that is so characteristic these days of a computer-rich scientific research environment. The purpose of this paper is to outline the problem of interpolating regional values, graphically representing regional data, and using regional data to make inferences about individual data (known as the Fallacy of Division in Logic, and more popularly known as the ecological fallacy). Upton continually gives instructions based upon his empirical experiences. Griffith's reaction to this paper is that Upton addresses two questions that deal with unresolved issues in spatial statistics, while addressing a third that has received considerable attention in the cartography literature. He further notes that again, as both Martin and Richardson, in her commentary on Martin's paper, point out, a fuller dialogue is needed between quantitative geographers and professional statisticians in order to erase such unawareness gaps.

The Editor

Information from Regional Data

Graham J. G. Upton

Department of Mathematics, University of Essex, Colchester, Essex, CO4 3SQ, England

Overview: This paper is concerned with three problems relating to regional data, namely, interpolation of regional values, pictorial representation of regional data, and the use of regional data to make inferences about individual data. In Section 2, regional values are treated as point values, with their regionality being recognized by giving particular weight to the values for neighbouring regions. A general weight function of the form $w = \text{(Population)}^i\,\text{(Area)}^j / \text{(Distance)}^k$ is used, where the distance concerned is that between the point representing the region of interest and the point representing the target region. Optimal values appear to be $i = 0$, $j = 1$, and $k = 3$, for first and second neighbours, with zero weights for more distant regions. A region may be small in size but have great importance—for example, Greater London in the context of population. This familiarly leads to the production of cartograms. Section 3 introduces two computer–based approaches to the production of cartograms, one based on the representation of regions by an appropriate number of points, and the other, due to my colleague, Dr. D. Fremlin, based on treating regions as polygons. Both methods lead to revised maps in which regions have areas proportional to importance. Section 4 is concerned with deducing individual values from regional values, and outlines a novel approach due to the Danish political scientist, Dr. S. Thomsen. Thomsen's method relies on the assumption that a cross–tabulation of two variables of interest may be regarded as a discretization of a bivariate normal distribution, with the underlying variables having values dependent on a common set of unmeasured (latent) variables. In the example considered, the method proves outstandingly accurate in retrieving information at the individual level from aggregate data.

1. Introduction

Motivation for the present work stems from problems experienced in carrying out a study of voting changes, which occurred between 1983 and 1987, in English constituencies (Upton, 1989). This study was concerned with disaggregating these voting changes into a number of components. Components of especial interest to political scientists concern the effect on the fortunes of a party of fielding a female candidate as opposed to a male candidate, and the extent of the personal allegiance that a sitting Member of Parliament might gain. The study suggested that, in a constituency in which 50,000 voted, the effect of fielding a female candidate resulted in a loss of between 125 and 250 votes, and that the advantage of incumbency was worth between 400 and 500 votes. Despite the small size of these effects, there were four constituencies (out of 512) in which the outcome might have been different if the candidates had been of the opposite sex.

In attempting to estimate such small effects, it is evident that one must reduce the "noise" in the data so far as is practicable by controlling for the variation between the characteristics of the constituencies. Some control would be possible using covariates such as the social class breakdown of the constituencies, or a breakdown of housing types. However, in the earlier paper I took the view that these should be regarded as genuine local effects, and instead

attempted to control for both tactical voting (using the 1987 voting profile) and geographical variation.

In the 1960s and 1970s, the changes between elections in allegiance towards the competing parties were remarkably homogeneous; a national swing towards the Conservative party, say, would be as apparent in the changes experienced in a Northumbrian constituency as in a Cornish constituency. However, in the 1980s, marked regional variations had become apparent, with the South of England moving away from the Labour party, so that it became commonplace to talk of a "North–South divide." Cozens and Swaddle (1987) have remarked that "[the election of] 1987 witnesses the emergence of fully–fledged regional patterns in Britain" (see also Johnston, 1985). Of course, while there was no sharp line separating North and South, the presence of this "divide" did imply that a change in the allegiance of a Northumbrian constituency between 1983 and 1987 was no longer a good guide to changes in the South.

In order to correct for geographical variations across the country, in my *Electoral Studies* paper I considered each constituency as being represented by a single point on the two–dimensional plane. The value of the voting change to be expected in a constituency then was taken to be that estimated by a weighted average of the voting changes experienced in all the remaining English constituencies, and this estimated change was subtracted from that actually experienced, leaving behind a residual in which the gender and incumbency effects would play a more prominent role. Clearly, different weighting procedures will lead to different estimated changes, and I chose weights based upon an inverse power of the distances between the notional point positions of the constituencies. The inverse power chosen was that which produced estimates most closely resembling the changes actually experienced, resulting in the use of an inverse square law.

Having estimated the residual local effects, a geographical element was still visible, implying that the weighting procedure adopted was sub–optimal. Therefore, a major theme of this present contribution is the comparison of alternative weighting schemes. For the most part we shall use data referring to the states of the United States, since these areal units form a more tractable data set. We shall consider several different sets of data, and will provide some general recommendations concerning a likely optimal procedure for interpolating regional data.

Having estimated the magnitudes of local effects, by whatever method, it is natural to consider presenting these estimates in the form of a map. However, the majority of the population live in conurbations, and this is reflected in a clustering of the positions of administrative groupings (*e. g.*, the English parliamentary constituencies). Using standard geographical co–ordinates to represent these areal units results in a map that is difficult to interpret, since the details of the fluctuations in local effects within the conurbations are almost invisible, unless the entire map is made unacceptably large.

Consequently, a second theme is the construction of scaling procedures that give equal prominence to each constituency. With constituencies represented as points, this problem implies a rearrangement of the points so that they are approximately uniformly distributed. With constituencies represented as polygons, achieving this goal implies a rescaling of the polygons, preserving their connectivities between regions, so that regions of equal population occupy the same area on the revised map. This latter approach, which results in the production of a cartogram, is not confined to a treatment of population, and can be easily

extended to give a "fair" representation of other aspects of regions. The work reported here is that of my colleague, Dr. Fremlin, and I am grateful to him for allowing me to report it in advance of publication.

The final theme of this paper is the so–called "ecological fallacy." This concept is concerned with the extent to which the actions of individuals can be deduced from knowledge of their aggregate behaviour. Here I report on the work of the Danish political scientist, Professor Thomsen, who has devised a method for estimating the body of a transition table from knowledge of its margins.

2. Spatial interpolation

2.1. Alternative approaches

Assume that we have n irregularly shaped regions whose positions are known, and that we wish to estimate the value of some variable, y, for a particular region, R, using the known value of y for at least one other region.

The motivation here is that the value of y for region R is either unknown or is suspected of having been influenced by some local factors. In the first case we have a genuine missing data problem, while in the second case (which prompted this study) it is the residual local effect that we are effectively estimating.

A number of general approaches to this problem could be considered. If we have information for all (or most) of the regions, on a vector of other relevant variables, say x, then we probably would do best to model the variation in y across the regions using standard multiple regression techniques for some function of x. In order to account for spatial effects, we might include location as a parameter in the model, or we might allow for a spatially correlated error structure (cf., Upton and Fingleton, 1985, Ch. 5). Two cases now arise, depending upon whether or not the x values for region R are known.

If the x values for region R are known, then substitution of these values into the regression model, using the values of the parameters estimated from the data from the other regions, will result in an estimate of the y value for region R whose reliability can be gauged from the associated standard errors.

If the x values for region R are not known, then a possible procedure would be as follows. First, use some interpolation procedure to estimate the x values for region R, then estimate the y value for region R using the fitted relation between y and x as described above. This approach appears to be novel, its properties are unknown, and it remains a topic for further research.

If there is no information on any relevant x–variables, then we must base our estimate on the available y–values, and the problem becomes one of interpolation in two dimensions.

2.2. Areal interpolation

Although there is a very large amount of literature on spatial interpolation, very little of it refers to the interpolation of data of the regional form being considered here. The special characteristic of the present data is that the y values refer to aggregate (or average) values for mapped regions whose boundaries are known. From the map we can obtain several pieces of information, which include the area of each region, the proportions of the edge of each region that border every other region, and the contiguities that exist between regions.

Interpolation could be based on any of these discernible properties of the regions on the map. To these we add population (if that information is available), since, if the variable y has any relation to human attributes, then it seems plausible to suppose that if R has two equi–sized neighbours (see Figure 1*a*), then that region having the greater population would have the greater bearing on the y value for R.

Kennedy and Tobler (1983) suggested that interpolation should be based solely on the information provided by contiguities between neighbouring regions, with weights being proportional to the total lengths of contiguous edge. However, this procedure requires the storage of a non–trivial amount of information, and also can lead to difficulties, as illustrated in Figure 1*b*. This figure shows Region R bordered by two regions of equal size, S and T. Logic suggests that each should be equally weighted, yet the Kennedy–Tobler procedure would result in region S being given much greater weight as a consequence of the jagged boundary between R and S.

"Jagged" boundaries result naturally when two regions are bounded by a natural feature, such as a river, while "straight" boundaries result from the use of lines of longitude or latitude. Accurate measurement will be difficult in the case of river boundaries. Further, if the river broadens into a lake, then a decision has to be made as to whether or not the two regions really have a common border at that point (after all, regions separated by the sea—the ultimate lake—would not be regarded as bordering one another).

Tobler and Kennedy (1985) describe the application of their procedure to the interpolation of values on a regular pixel mesh, and to interpolation on an irregular resel network (the states of the United States). They also extend their procedure to include second neighbours, giving the necessary formulae for the pixel mesh, and an illustration of the results for the resel case. One should note that their Figure 4 is a fine example of misleading statistics! This figure shows a scatter diagram of fitted $(\hat{y})$ and actual (y) values together with the regression line of $\hat{y}$ on y. This line is $Y = 39.3 + 0.52y$ (there is a typographical error on the figure) and leads to a multiple correlation R^2 value of 0.72. However, the true predictor is $\hat{y}$, not Y, and the relevant line is the 45 degree line $\hat{y} = y$, which has a much smaller R^2 value. The authors' claim that the value of 0.72 represents "an average success rate of 72%" is meaningless.

Judging by the recent review of Lam (1983), the Kennedy–Tobler procedures are the only ones that have been suggested for the direct interpolation of one regional y value from its companions. With one exception, all other methods discussed by Lam in her review paper are of the "overlay" type in which the y values from one geographical subdivision, the "target" zones, are computed from those for some alternative geographical subdivision, the "source" zones.

The exception is a procedure due to Tobler (1979) that allows one to interpolate sub–regional values from aggregated regional values. A brief description follows, and a simple worked example is provided by Lam (1983).

When a y value is quoted for a region and a different y value is quoted for its neighbour, then, if we assume homogeneity within the region, this implies that there is a sharp discontinuity along the edges of the regions. However, when one crosses the border from one region to its neighbour, one does not necessarily expect to immediately perceive that fact—only rarely is there a clear demarcation line in terms of ground cover, population density, social

Figure 1.

The effects of boundary and population on regional interpolation.

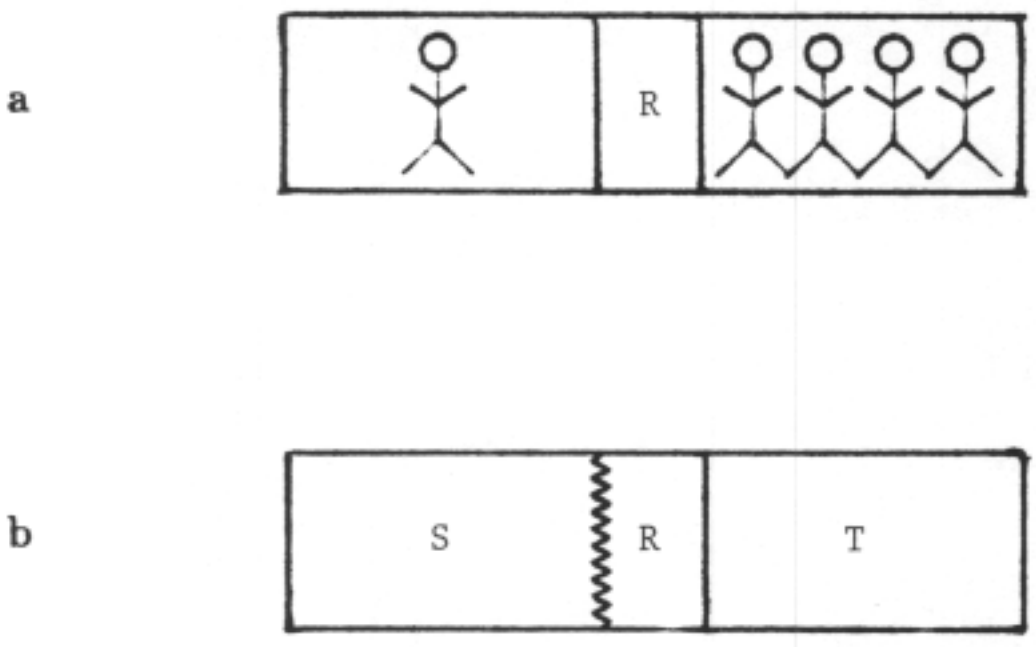
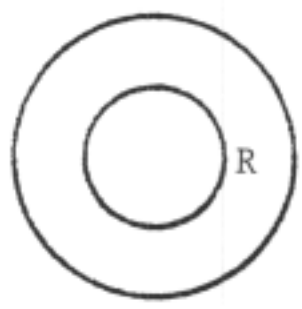


Figure 2.

A pathological example of a problem with the centre of gravity.



class split, or whatever. Tobler suggested a possible procedure for introducing a smooth lack of homogeneity into regional data so that these edge discontinuities are removed.

In order to illustrate Tobler's idea, consider the example of population density. For a region R, of area A and population density ρ, the total population is clearly $A\rho$. Suppose that the point (x, y) lies within R. Under the homogeneity assumption, the density at (x, y) is ρ for all points within R, thus leading to a plateau of density with edge–discontinuities. Tobler outlines a procedure for replacing the constant density ρ with a position–dependent density $D(x, y)$, which is such that

(i) $\iint D(x, y)dx\,dy = A\rho$,

(ii) there is no discontinuity along the edges of neighbouring regions.

Because of the double integration involved, this procedure is said to be volume preserving, or *pycnophylactic*.

2.3. Point interpolation

We now move from these area–based procedures to alternatives in which the y value for a region is assigned to some central point. Tobler and Kennedy (1985) assert that their procedure is "far superior to the use of arbitrary points ... to represent geographic areas," but they cited no evidence to support that statement. Upton (1985) showed that the estimates of state population density obtained using points were at least as good as those obtained by Kennedy and Tobler (1983), and further evidence of this is presented later in this section. The Tobler–Kennedy statement probably derives from the critique by Porter (1958) of the use of arbitrary points in the construction of isopleths.

A major problem that arises when representing an area by a "central" point is the choice of location for that point. As Figure 2 illustrates for a pathological case, there may be no point lying within the region that could reasonably be described as its centre! Moreover, even for a circular region the choice of location for the representative point is not clear cut. Suppose, for example, that this circular region consists of an uninhabited upper semi–circle containing forest, and a lower semi–circle representing a densely populated urban area. If our interpolation is concerned with some variable that is independent of population and afforestation, such as rainfall, then the centre of the circle would be a suitable choice as the representative point. If we are concerned with some property of forests, then we should choose a location in the upper semi–circle, while if we are concerned with people, or people's attitudes, then a location in the lower semi–circle would be appropriate.

In this study, these niceties have been ignored and the same central points have been used throughout. In the case of the contiguous regions of the United States, the "centre" has been taken to be the co–ordinates reported for that region in a standard atlas. In the case of the English constituencies, the "centre" has been taken to be the approximate co–ordinates of the population centre of gravity.

The huge literature on spatial interpolation using point values is spread over many disciplines, but almost exclusively deals with the case where a y value refers to an actual point location. There are two broad approaches to the interpolation of data of this type. One approach is to postulate the existence of some underlying continuum and fit a so–called trend surface. The principal problem with this approach is that the order of this surface is not known. An underlying first–order surface (a plane) is unlikely to be the case,

and once we move to higher–order surfaces we may achieve rather improbable estimated surfaces, leading to implausible estimates that lie outside the feasible range for y (Crain, 1970). One alternative here is to fit "local" surfaces subject to edge constraints that obviate discontinuities; this approach can be very expensive in terms of computational effort (see, for example, Haining *et al.*, 1984).

The second approach is to use, as an estimate, a simple weighted average of the available y–values. One major disadvantage here is that there are a very large number of plausible alternative weighting schemes. A second disadvantage is that this approach is innately conservative, in the sense that every estimated value will lie within the range of the observed y–values so that maxima will always be under–estimated and minima will always be over–estimated. Nevertheless, this approach has the great virtue of simplicity, and it is for that reason that we choose this approach for further study. In the context of genuine point values, Lam (1981) notes that this approach almost always leads to reasonable results.

Upton (1985) used inter–regional distances alone as the basis for weights. However, this is not really plausible as Figure 3*a* demonstrates. Regions S and T have centres that are equi–distant from the centre of R, but it would seem strange not to give more weight to S than to T.

Further evidence of the need to take area into account is provided by Figure 3*b*, which shows region S divided into two sub–regions, S_1 and S_2, that have centres equi–distant from the centre of R. Suppose that the value of y is uniform throughout S. Then before subdivision we have one y value at a distance d from the centre of R, and after subdivision we have two such values—yet nothing has changed but our imposition of a dividing line through S. If we weight by the area of S, then a single y–value is replaced by two half–weight y–values, and sanity is preserved. Note that the Kennedy–Tobler procedure also deals effectively with this problem.

Clearly area cannot be the sole ingredient of the weight function, since we can expect positive spatial correlation to prevail such that nearby y–values will have more relevance than more distant y–values. We can measure distance in many different ways (for example travel times), but the two methods considered here are in terms either of "flat–earth" distance or of contiguities.

In the case of the English constituencies, the co–ordinates used were based upon the standard National Grid, while in the case of the United States the distances were calculated in the following way. Let (x_r, y_r) and (x_s, y_s) be the co–ordinates (latitude, longitude) of the centres of regions R and S; then d_{rs}, the distance between these two centres, is defined as

$$d_{rs}^2 = (1.6) \times (x_r - x_s)^2 + (y_r - y_s)^2, \tag{1}$$

where the quantity 1.6 is a rough correction factor adjusting for the fact that one degree of longitude is not the same length as a degree of latitude. The reference to a flat–earth distance is a reflection that the distances calculated are not great–circle distances, and in view of this restriction nothing more sophisticated than equation (1) (which also ignores variations in distance with longitude) seems worthwhile.

The most obvious way to include distance in the weight function is to use some inverse power of distance, d^{-k}, although other functions of distance such as $\exp(-ad)$, $\exp(-ad^2)$ and $\exp(-ad^2)/(b + d^2)$ (see Ripley, 1981, Ch. 4) have been suggested. In these latter

Figure 3.

The relevance of area to spatial interpolation.

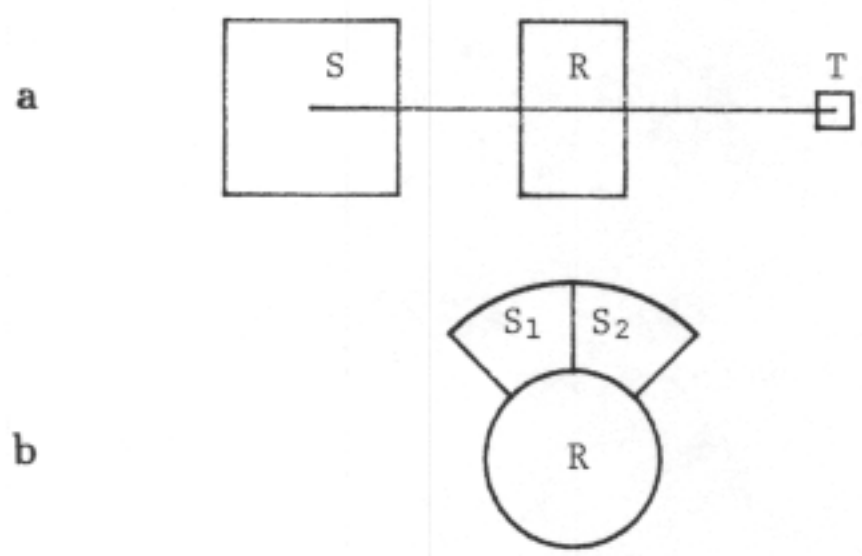
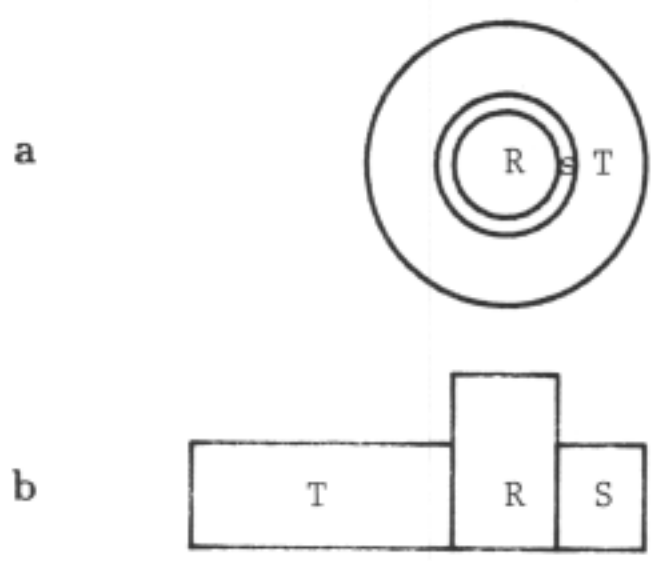


Figure 4.

Problems with weights that use edge contact only.



functions the choice of values for the arbitrary constants a and b will be dependent upon the unit of measurement (Stetzer, 1982). Therefore, in this present study the simple d^{-k} weighting is used with the sensitivity of the estimates to the choice of k being investigated, and recommendations being made concerning the optimal value of k.

The method suggested by Kennedy and Tobler (1983) implies that remoteness of one region from another could be measured using contiguity rather than distance *per se*, implying estimation using a weighted average of the y values exhibited by first neighbours. Clearly one should not expect to perceive sharply differing local characteristics simply by crossing over a border—any changes are sure to be gradual.

First neighbours alone, however, could be misleading, as the admittedly far–fetched Figure 4*a* illustrates. The region R has only one first neighbour, S, but an analyst surely would want to take rather more notice of region T than of region S.

Figure 4*b* illustrates another situation. Here regions S and T border R and account for the same proportions of the border of R. Region T is very much bigger than region S, but in this case one would not really want to give it that much more prominence because its own centre of gravity is so distant from R.

Taken together, Figures 4*a* and 4*b* suggest that it would probably be unwise to confine attention only to first neighbours, and that distance still has a role to play. Accordingly, attention now will turn to investigating a number of alternative weight functions of the form

$$w = (\text{Population})^{i} (\text{Area})^{j} / (\text{Distance})^{k}, \tag{2}$$

with $i = 0$ or 1, $j = 0$ or 1, and k taking on values in the range 0 to 6 inclusive. These weight functions are used in conjunction with all of the available data, with just the first neighbours, or with both first and second neighbours.

2.4. The United States data sets

Although this study was motivated by data from English elections, most attention will be paid to selected, very much smaller data sets relating to the contiguous states of the United States. These data sets are listed in Appendix A, together with the areas of the states and their co–ordinates as given in the gazetteer of the 1973 edition of *The Mitchell Beazley Concise Atlas of the Earth*. The first data set (I) is that used by Kennedy and Tobler (1983), and subsequently by Upton (1985), in their earlier works on spatial interpolation. This data set refers to the population densities of the forty–eight contiguous states (at an unspecified date). The second data set (II) has been the subject of some discussion in the context of the plotting of geographical information. These data were first illustrated using an unclassed choropleth map by Gale and Halperin (1982), and subsequently using rectangle charts by Cleveland and McGill (1984) and Dunn (1987). These data refer to the murder rates in 1978, expressed as murders per 100,000 inhabitants. Both of these data sets omit any specific information concerning the District of Columbia.

The remaining five data sets include information on the District of Columbia, and were chosen so as to provide a spectrum of applications. They were gleaned from the Statistical Abstract of the United States 1986 (table given in brackets) and are as follows:

(IIIa,b) the 1982 mortality rates of infants aged less than 1 (numbers per 1000) (Appendix V),

(IVa,b) the proportions of whites in the 1980 resident population (Table 32),

(Va,b) the population change between 1970 and 1980 expressed as a percentage of the 1970 population (Table 13),

(VIa,b) the percentage of voters supporting the Republican candidate for president in 1984 (Table 412), and

(VII) the numbers of food stamp recipients per 1000 population in 1984 (Appendix V).

The first four of these data sets contain one extreme outlier, the District of Columbia (D. C.), which is a very small densely populated area in which the vast majority of the residents are black. This areal unit has very high infant mortality and has experienced a net exodus in recent years. Because of these features we can expect that all the interpolation procedures will fare much worse when the Washington, D. C., data are included. For this reason these first four *Statistical Abstract* data sets were analysed twice, once (a) excluding the D. C. data, and once (b) including it.

2.5. Results for the United States data sets

In order to assess the merits of the various interpolation procedures, the true value, v_r for each region was compared with the interpolated value, v_r^*. A summary of the overall effectiveness of a procedure is given by evaluating

$$D = \sum (v_r - v_r^*)^2, \tag{3}$$

with the sum being over either the 48 contiguous states or over the 49 regions including the District of Columbia, as appropriate.

Since the actual magnitudes of the items in the various data sets differ considerably from each other, it is necessary to place each set of rival D values on a comparable footing. To this end, the smallest D value in each collection was set equal to 100, and the remaining D values in each collection were scaled accordingly. The resulting values are summarised in Table 1.

The first four columns of Table 1 refer to the particular form of weight function, namely equation (2), that has been used. The upper half of the table refers to cases where $k > 0$, implying that an inverse power of distance is included in the weight function. In the lower half of the table $k = 0$, implying that no explicit account is taken of inter–regional distances. Within each half of the table, three variants of the weight function are applied to all the data, to first and second neighbours only, or to first neighbours only. The variants have $(i, j) = (0, 0)$, $(1, 0)$ or $(0, 1)$. A fourth variant $(i, j) = (1, 1)$ gave values intermediate between the last two cases.

Thus Row 10 of Table 1 uses the grand mean of all the remaining values as an estimate of v_r, with no account being taken of either their distance from region R or their area or population size. This naturally leads to very poor estimates, which is evident in the table, where one can see, for example, that for the population density data set (Set I) the value of D was 4.2 times greater than the best value that was obtained. The scaled D values given in this row are (approximately) scaled versions of the variances of the various y values. A

TABLE 1
COMPARISON OF EFFICIENCIES OF 18 INTERPOLATION PROCEDURES ON 11 DATA SETS
(Each data set indexed at 100 for the overall optimal procedure)

Distance	Population	Area	Neighbours	Data Set I	II	IIIa	IIIb	IVa	IVb	Va	Vb	VIa	VIb	VII	Average
*			All	114	102	109	138	105	156	124	126	105	123	110	119
*	*		All	123	108	110	131	128	151	144	140	120	116	100	125
*		*	All	102	102	103	109	110	103	129	121	106	100	100	108
*			1st + 2nd	112	100	101	100	105	135	110	110	105	122	116	111
*	*		1st + 2nd	118	107	106	113	127	132	134	125	121	112	104	116
*		*	1st + 2nd	100	101	100	105	109	102	118	113	107	100	108	106
*			1st	120	108	120	115	100	105	100	100	107	118	123	111
*	*		1st	119	109	114	113	120	114	130	125	125	114	110	117
*		*	1st	111	108	112	108	106	100	109	109	100	101	114	107
			All	420	287	221	169	264	202	217	205	136	140	175	222
	*		All	435	325	224	166	293	204	224	208	147	138	174	231
		*	All	459	302	229	176	264	199	242	232	141	145	177	232
			1st + 2nd	221	133	102	100	130	142	119	116	118	123	129	130
	*		1st + 2nd	253	207	128	114	204	168	134	127	128	123	148	158
		*	1st + 2nd	221	155	109	105	142	145	124	121	115	122	148	128
			1st	165	131	122	115	111	106	100	100	107	119	129	119
	*		1st	211	153	129	113	155	128	130	126	125	125	139	139
		*	1st	153	154	125	108	130	122	109	109	100	122	143	125

NOTE. * denotes information that has been used.

large value in this row indicates that the best procedure for that data set was particularly effective in dealing with the observed spatial variations in the data values.

In the case of the procedures that use a distance component, d^{-k}, the actual power chosen varies from weight function to weight function and from data set to data set. All values of k between 0 and 6, in steps of 0.25, were considered, and the reported (scaled) D values are those that correspond to the optimal choice of k. The variability of D with changing k within a data set, and the variability in the optimal values of k from data set to data set are discussed later.

Table 1 shows clearly that procedures that incorporate a distance component almost always do very much better than those from which this variable is omitted. Generally speaking, when there is no distance component in the weight function, the smaller D values result from confining attention to the immediate neighbours of a region, while taking averages over all the remaining regions is markedly disastrous in the case of the population density and murder data sets.

When there is a distance component included, the picture is less clear as to which regions should be given non–zero weights. In one case (food stamp recipients—set VII) it is preferable to use information from all the regions, in four cases using both first and second neighbours is best (Sets I, II, IIIa, IIIb), in five cases (Sets IVa, IVb, Va, Vb, VIa) using only the first neighbours is best, and in the remaining case (Set VIb) using only the first neighbours is marginally inferior to the other alternatives. In contrast, using population as a component of the weighting procedure does not work well, even in the case of Sets VIa and VIb, which relate to opinion data (proportions supporting the Republicans).

Using a simple inverse distance weighting in order to combine information from all other constituencies (Row 1 of Table 1), which was the method suggested by Upton (1985) and used in Upton (1989), never provided optimal results with these data sets and occasionally led to seriously incorrect estimates. This technique performed least well for the data sets containing information on the District of Columbia (the sets marked with a suffix b), and the contrast with those excluding that information highlights the sensitivity of this procedure to the presence of an outlier.

The Kennedy–Tobler estimates for the population density data (Set I), which were based on edge–weighted first neighbours, with no distance effect, lead to a scaled index of 143, substantially less than the unweighted first neighbour estimate (165), but clearly far from optimal.

Although there is no especial reason for expecting the same weighting procedure to be optimal for every data set, from Table 1 there do appear to be some generalisations that can be made. The index having the lowest overall average is that involving a weighting function of the form Area $/d^k$. There is little to choose between the use of this weighting function applied to first neighbours only (average index 107), to first and second neighbours (average index 106), or to all regions (average index 108). Each of these produced the optimal index for two of the eleven data sets. However, figures given for the distance weighted index values are those for the optimal choice of k. Since this optimal value varies from data set to data set, some investigation of the sensitivity of these estimates to the choice of k is required.

Table 2 illustrates the dependence on k of interpolation procedures that use weights given by $w =$ Area $/d^k$. This table shows that the average indices using first neighbours are much less dependent on the choice of k than are the averages based on greater numbers of regions. However, there is a trade–off between remoteness of the observations from the target locality and quantity of observations. For the 49 contiguous regions of the United States, the number of first neighbours varies between 1 and 8, with a mean of 4.5, while the total number of first and second neighbours varies between 3 and 24 with a mean of 11.9.

Combining findings from Tables 1 and 2 suggests that the following five procedures give very comparable results:

(i) the unweighted average of first neighbours,

(ii) the weighted average of first neighbours, with weights given by $w =$ Area $/d^2$,

(iii) the weighted average of first neighbours, with weights given by $w =$ Area$/d^3$,

(iv) the weighted average of first and second neighbours, with weights given by $w =$ Area$/d^3$, and

(v) the weighted average over all regions, using weights given by $w =$ Area $/d^3$.

The simplest of these five procedures is clearly (i), while procedure (iv) has the smallest average index, and also the smallest maximum index for the eleven data sets so far considered. However, all of the above testing has been done on data obeying the contiguities of the United States, and it is clearly sensible to look at data relating to some other region for confirmation of these findings.

TABLE 2
THE EFFECT OF THE PARAMETER k IN d^{-k} WEIGHTED AREA-BASED ESTIMATORS
(Each data set indexed at 100 for the overall optimal procedure)

		Data Set											
Procedure	Value of k	I	II	IIIa	IIIb	IVa	IVb	Va	Vb	VIa	VIb	VII	Average
All	0	459	302	229	176	264	199	242	232	141	145	177	233
Other	1	352	228	158	136	204	167	158	151	119	126	136	176
Regions	2	200	144	113	112	140	122	129	121	107	104	108	127
	3	132	108	103	112	112	103	132	124	107	101	100	112
	4	109	102	108	143	112	127	138	133	111	126	108	120
	5	103	104	119	172	120	159	143	142	115	151	119	132
		22	16	12	10	14	12	9	9	10	11	12	Optimal value of 4k
First	0	221	155	109	105	142	145	124	121	115	122	148	137
and	1	186	131	103	105	127	130	118	113	110	112	127	124
Second	2	146	111	100	107	114	110	123	115	107	102	113	113
Neighbours	3	117	102	102	112	109	102	131	123	108	101	108	110
	4	103	102	109	144	113	128	138	133	112	126	113	120
	5	100	105	119	173	122	159	144	142	116	152	121	132
		20	14	9	2	10	12	4	5	9	11	12	Optimal value of 4k
First	0	153	154	125	108	130	122	109	109	100	122	143	125
Neighbours	1	146	133	117	110	116	110	116	112	102	108	125	117
Only	2	134	117	112	112	107	103	126	118	107	101	115	114
	3	121	110	113	118	107	101	135	126	112	104	114	115
	4	113	109	119	150	113	129	141	136	116	130	121	125
	5	111	110	128	177	122	160	146	144	120	154	128	136
		20	15	10	0	10	11	0	0	0	9	11	Optimal value of 4k

2.6. The English constituency voting data

Consider data concerning voting profiles for the 523 English constituencies for the General Elections of 1979, 1983 and 1987. Further, consider only the shares of the vote obtained by the three main parties, with these shares being scaled so as to sum to one for each constituency at each election.

These election data are used in two different ways. First, the change in the percentage of the vote obtained by a party between two successive elections is analyzed. Three parties and two inter–election periods gives rise to six possible data sets, and the values reported refer to the aggregate fit over all six sets. The second treatment is of the actual percentages themselves, and here results refer to the aggregate fit over all nine party/election combinations.

Results reported in Table 3 for the two data set groups are divided into two sections, one incorporating an area component in the weight function and one omitting this component. A convenient source of information about constituency areas is a publication of the Office of Population Censuses and Surveys (1981). Results are given for weights involving inverse distance powers between 0 and 4 inclusive. These results are indexed so that 100 represents the best fit found for that data set.

The findings reported here confirm those for the United States data. If there is no knowledge about the constituency areas, then the simple mean of first neighbours works

very well for both data set groups. If area measures are used in the weighting function, then the minimax choice is $w = \text{(Area)}/d^3$ for first and second neighbours.

A noticeable feature of the calculations that does not show up in Table 3 is the difference in the computational effort between using all of the data and a restricted set of neighbours. For these 523 English constituencies the number of first neighbours varies between 1 and 15, with an average of 5.5, while the combined total of first and second neighbours varies between 2 and 45 with an average of 18.2. The English constituencies have larger average numbers of neighbours than the states of the United States because a smaller proportion of the regions are situated on a boundary.

TABLE 3
ALTERNATIVE d^{-k} INTERPOLATION PROCEDURES
APPLIED TO VOTING DATA FOR 523 ENGLISH CONSTITUENCIES

Data Set	Procedure	With Area Component					Without Area Component				
	k:	0	1	2	3	4	0	1	2	3	4
Inter-	All	136	124	106	104	114	123	107	101	110	120
election	1st + 2nd	120	110	104	108	117	102	100	105	115	123
change	1st	116	111	112	119	125	104	107	115	124	131
Party	All	298	245	153	114	115	301	182	144	138	140
percentages	1st + 2nd	179	142	116	108	110	115	107	108	115	124
	1st	132	118	109	109	112	100	100	106	114	121

2.7. Stetzer's results

Stetzer (1982) was interested in a somewhat different problem, namely the optimal choice of a weights matrix for use with STARIMAR models. Stetzer was concerned with point values and simulated data having a variety of autocorrelation structures. Stetzer considered both the accuracy of parameter estimates and the accuracy of forecasts using various weight matrices. For irregular meshes, which correspond to the present situation, Stetzer considered both 0/1 weights based on the connectivities in minimum spanning trees (roughly equivalent to the use of first neighbours with no distance, population or areal weighting), and two alternative distance–decay weightings. These distance weightings took either the form d^{-1} or $\exp(-cd)$, where c was given some appropriate value; but, the weights were set to zero if $d > D_{\max}$, the cut–off distance. Stetzer reports results for three cut–off distances, with the smaller $D_{\max}$ value restricting attention to nearby points and the largest $D_{\max}$ value allowing information from all areal unit values to be used. Therefore, these three values correspond reasonably well to the three levels of neighbour considered in the present study.

Confining attention to the results that Stetzer obtained for the errors in forecasts, three generalizations emerged:

(i) using distance–decay weights, the largest of the $D_{\max}$ values was worst and the smallest was preferable,

(ii) the d^{-1} weights were preferable to the $\exp(-cd)$ weights, for Stetzer's choice of c, and

(iii) the simple connectivity weights were often much inferior to the distance–decay

weights.

Although the contexts are distinctly different, these findings are reassuringly similar to those being reported in the present analysis.

2.8. Recommendations

If information concerning the areas of the various regions is not available, then apparently the best procedure is to take *the simple average of the immediate (first) neighbours*. However, disregarding the area of a region seems somewhat illogical, since if a homogeneous first–neighbour of a region R were to suddenly disintegrate into m pieces, all bordering the region R, then the weight of that region would increase by a factor of m. Hence, despite the undoubted success of the simple first–neighbour estimate, an area–based estimate seems more reliable. With large numbers of regions there is a huge computational gain from confining attention to just a few constituencies, though one is reluctant to ever use just one constituency as an estimator of another. For these reasons the preferable procedure would seem to be to use *first and second neighbours, with weighting (Area)* $/d^3$.

Figure 5 displays results for this procedure applied to the population density data. Since population densities have a highly skewed distribution, with most being small (smallest is Wyoming at 3.4) and a few being very large (largest is New Jersey at 953.1), the data have been plotted on a log–log scale, so that all the states receive comparable prominence. This figure may be contrasted with Figure 4 of Tobler and Kennedy (1985).

A perfect estimation procedure would result in all of the points in the diagram lying on the 45 degree line shown. Given the logarithmic scaling, the greatest vertical deviations from the line correspond to the greatest *multiplicative* discrepancies. These are for Nevada (observed 4.4, predicted 102.7) and California (observed 127.6, predicted 6.4). Although neighbours, these states are separated by mountains, and evidently any interpolation scheme that fails to take proper account of barriers of this kind is doomed to failure. The greatest absolute error is that for the state of New Jersey (observed 953.1, predicted 346.4), though as noted before, every interpolation scheme based on weighted averages is certain to underestimate the maximum and overestimate the minimum.

3. Redrawing the map

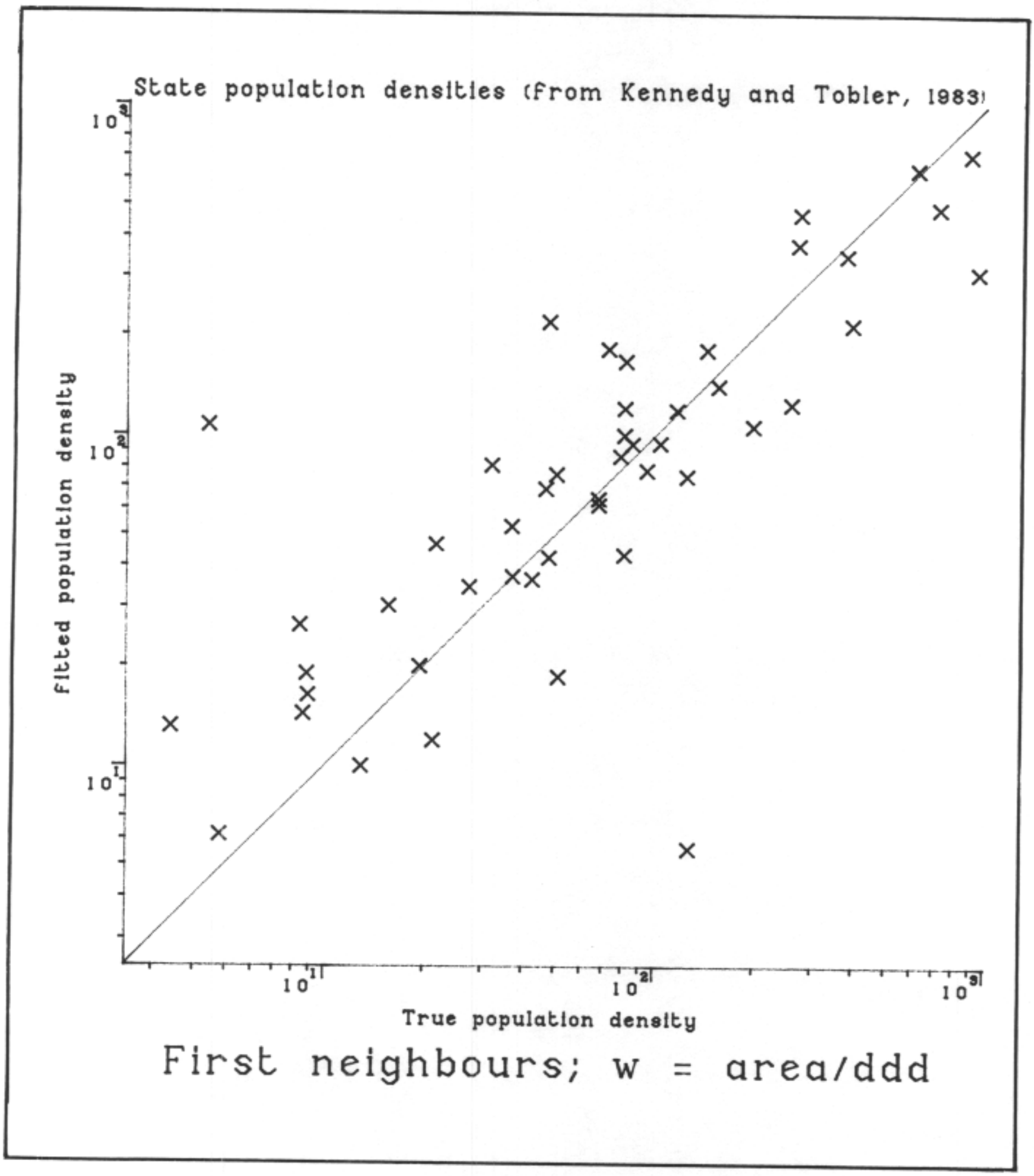
As mentioned in the introduction, the motivation for this section came from a need to coherently display information concerning English constituencies on a single map. One standard procedure for illustrating quantitative geographic information is the use of a choropleth map, described as the "lazy man's map" by Evans and Jones (1981), who discuss various alternatives to it. More recent alternatives involve the use of rectangle charts (Cleveland and McGill, 1984; Dunn, 1987), though these do not address the problems of the visibility of small but important locations.

In this section, two computational approaches will be considered, one resulting in a redistribution of point locations, and the other, due to my colleague, Dr. Fremlin, resulting in a direct redrafting of a map to produce a cartogram. An alternative approach, when there are only small numbers of locations, involves multidimensional scaling using contiguities to specify structure (see, *e. g.*, Gatrell, 1981); this technique is not feasible with present programs when the numbers of locations are much in excess of 100.

The point relocation procedure will be illustrated in the context of the British parlia-

Figure 5.

Scatter diagram, using a log–log scale, of observed and fitted population density values.



mentary constituencies that motivated this study; Fremlin's procedure will be illustrated in the context of the United States.

3.1. Point locations

Representing each of the 631 contiguous British constituencies (including the Western Isles constituency, but omitting the Northern Ireland constituencies and the remote Orkney and Shetland constituency) by a point located at the approximate centre of gravity of its population, leads to the map displayed in Figure 6. The choice of the point location was subjective and based upon a scrutiny of the relevant constituency map.

In Figure 6, the constituencies have been subdivided into 11 groups following the classification employed by Waller (1983). A breakdown of the allocation of counties to groups is given in Appendix B. The most obvious features of this map are the concentrations of constituencies in Greater London, Birmingham, Liverpool and Manchester, Tyneside, Edinburgh and Glasgow, and south Wales. If one wants to display information about these urban constituencies, then this map will have to be enormous if Central London is to be visible! The more usual solution is to have magnified maps of the conurbations displayed adjacent to a main map, with the conurbation information on these inset maps and the rural information on the main map. This procedure works well unless one is interested in the urban–rural interface, or in national trends rather than local details. It was the latter requirement that motivated this study.

The objective now is to rearrange the points in Figure 6 in such a way that all are equally visible, and so that both the relative and absolute geographic positionings of the constituencies are preserved (*e. g.*, London constituencies remain adjacent to each other and continue to be located towards the South–East corner of the revised map).

Consider, for simplicity, a rectangle of land within England. If the point density within the rectangle is uniform, then it follows that the density of the x–coordinates will be uniform, and that the same will be true for the y–coordinates. This suggests that if a rectangle contains a non–uniform arrangement of points, with non–uniform spreads of coordinates, then a simple rescaling of the axes might suffice to correct matters. However, Figure 7 demonstrates that this is not the case.

Figure 7 shows a rectangle containing eight points whose x–coordinates are equi–spaced, and whose y–coordinates are equi–spaced, yet two of these points have considerably greater visibility than the rest. Study of the marginal coordinate distributions provides no information concerning the clusters present in the pattern. Although this is an extreme case, it will be observed that similar features appear in the real data of Figure 6, with London and Glasgow being in opposite corners of the "British rectangle."

Nevertheless, the concept of coordinate scaling underlies the method actually adopted, which involves a scaling procedure in which x–coordinates and y–coordinates are alternately and iteratively adjusted (one should note that the resulting solution is not unique). Ideally the number of points is arranged to be a power of 2 and, in this case, at each iteration all the data have either their x–coordinate or their y–coordinate adjusted, with the number of data points being scaled in precisely the same way and being halved each time.

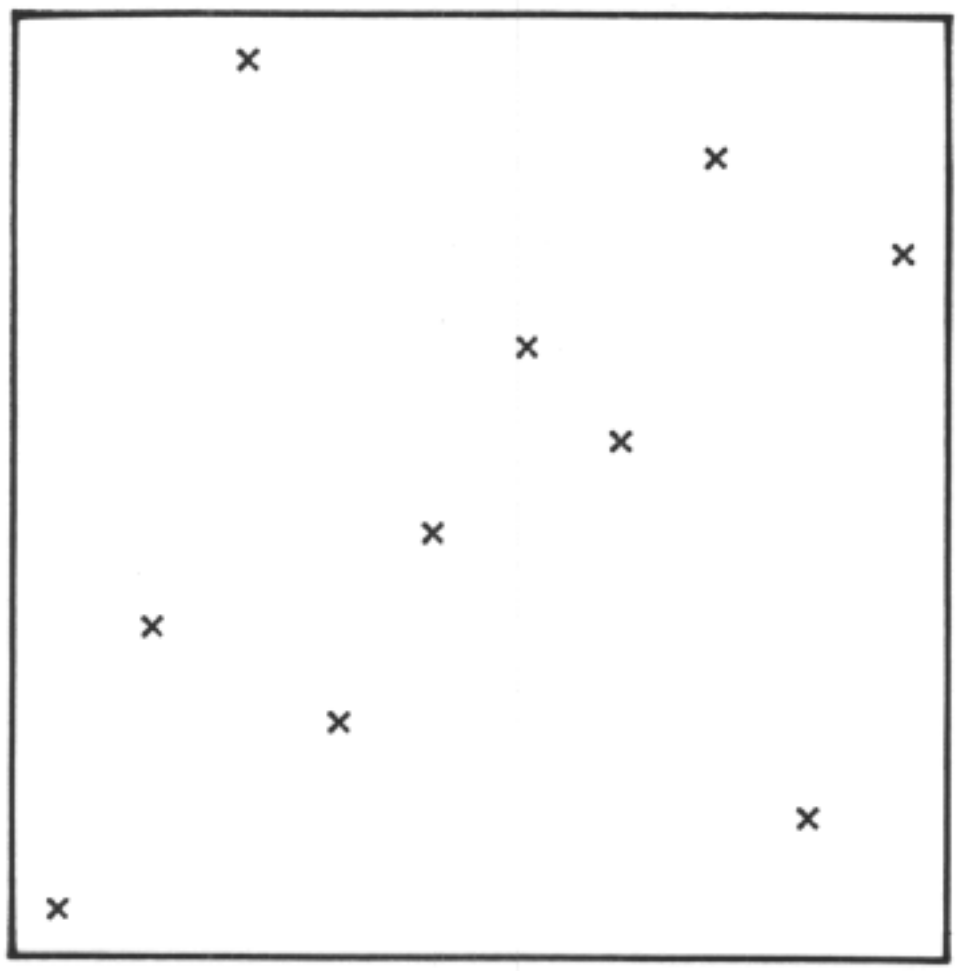
In Figure 8*a* the study area is a rectangle with width w and height h. For convenience suppose the origin of the axes is at the bottom left corner. Let M_x be the median x–

Figure 6.

Location of 631 British constituencies. (Symbols refer to different regions—see Appendix B.)

Figure 7.

Ten points, non–uniformly spread, but with equi–spaced $x-$ and $y-$coordinates.



coordinate of the N points. Ideally, as remarked above, N is arranged to be a power of 2; but this may prove impractical, and so a procedure for general N will be described. If N is even, with $N = 2n$, then M_x is the average of the nth and $(n+1)$–th ordered x–coordinates. If N is odd, with $N = 2n + 1$, then M_x is equal to the $(n + 1)$–th ordered x–coordinate. Let $R_x = 2M_x/w$. If the points were distributed evenly over the rectangle then R_x would be equal to 1. If R_x is less than 1 then this indicates clustering toward the left of the rectangle, while R_x greater than 1 indicates clustering toward the right. Denote the group of points having the n smallest x–coordinates as group A, and the remainder as group B. Let x_A denote the x–coordinate of a point in group A and x_B denote the x–coordinate of a point in group B. Revised coordinates x^*_A and x^*_B may be calculated using the formulæ

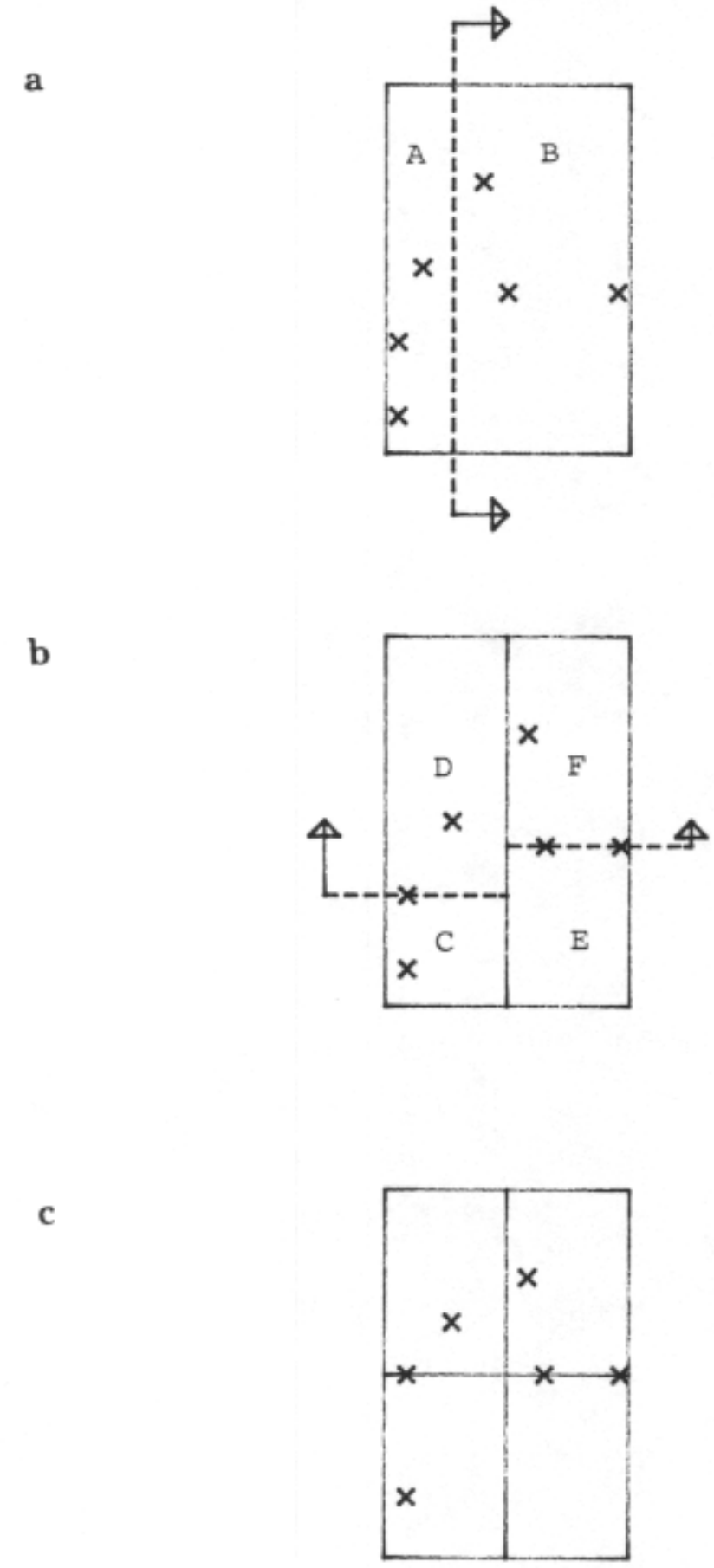
$$x^*_A = x_A/R_x, \quad x^*_B = [w + (2x_B - wR_x)/(2 - R_x)]/2. \tag{4}$$

These computations result in a set of revised x–coordinates having median $w/2$, so that in general there are equal numbers of points in the left– and right–hand sides of the rectangle. In the case of N being odd there also will be one point having a revised x–coordinate equal to $w/2$, and this x–coordinate will remain unchanged throughout subsequent iterations. The adjustments are illustrated in Figure 8*a*, and lead to the revised x–coordinates shown in Figure 8*b*.

This first iteration concludes with a scaling of the y–coordinates. In this scaling the

Figure 8.

One complete iteration of the scaling procedure.



two groups A and B are scaled *separately*. If N is even, then each of A and B contains $N/2 = n$ points. If N is odd, then the point that has the median x–coordinate is assigned to either A or B, the assignment being based on the location of its nearest neighbour. If its nearest neighbour belongs to group A, then this point also belongs to that group; otherwise this point is assigned to group B.

Now the y–coordinates of group A may be scaled by dividing the points in that group into subgroups, say C and D, with the y–coordinates of subgroup C being smaller than those of subgroup D. The median y–coordinate, M_y, is calculated in the manner described earlier and, taking $R_y = M_y/h$, the revised y–coordinates are given by

$$y_C^\star = y_C/R_y, \quad y_D^\star = [h + (2y_D - hR_y)/(2 - R_y)]/2, \tag{5}$$

which are an exact equivalent of the previous x–coordinate equations given by (4) above.

Next, turning attention to group B, one may form subgroups E and F, calculate the values of M_y and R_y for group B (which will, in general, be different than those for group A), and revise the y–coordinates for group B using equation (5), with the new R_y value and with the labels C and D being replaced by E and F. This procedure is illustrated in Figure 8*b*, and leads to the revised positions shown in Figure 8*c*. This step completes one cycle of the iterative procedure. If any point has been assigned the median y–coordinate value $(h/2)$, then this coordinate will be unchanged through subsequent iterations.

At the start of the next cycle there are four separate groups, namely C, D, E and F, each of which may be treated as a separate entity. The scaling formulæ (4) and (5) will require adjustment in order to take account of which quarter of the original square is being treated; nevertheless, the general procedure remains the same. At the end of this cycle there will be eight groups to consider: each successive cycle will result in a doubling of the number of groups to be considered, while the number of points in these groups will be roughly halved on each occasion. The reduction is not exactly a halving because of the presence of an increasing number of points whose coordinates have been fixed by virtue of their occupying a median position. The iterations continue until each subregion contains only a single point that is relocated at the centre of that subregion. One should note that in the case where N is a power of 2, only at the last stage will the final coordinates of a point be determined.

This foregoing description of the procedure assumes that one is working with a rectangular study region. However, if one simply encloses the point locations of Figure 6 within a rectangle and then applies the procedure, the outcome would be an unrecognizable "rectangularized" Britain! In order to avoid this problem the artificial rectangle must be filled up around the existing land constituencies with a lattice of "sea constituencies." These synthetic areal units are positioned on a grid having approximately the same density of points as the land constituencies, so that the general shape of the land is preserved. Small random increments may be added to the lattice point positions so as to avoid problems with multiple tied coordinates; and, these fictitious constituencies may be given identifiers that will prevent them from being plotted when the map finally is redrawn. The revised map given in Figure 9 involved a supplement of 764 sea constituencies to the 631 original constituencies (the supplementary number is large because the land mass of Britain is inclined relative to the National grid, resulting in a wide rectangle—a different coordinate system could result in a closer fitting encompassing rectangle).

That the algorithm has been reasonably successful in giving each constituency approximately equal prominence is immediately apparent. Studies of the patterns of regional symbols confirm that this procedure does not scatter constituencies unduly, and preserves their general relative positions. However, a curious feature of the procedure becomes apparent when Figure 9 is held at eye level and is viewed from the bottom of the page. Apparently there are a number of nearly empty "channels" running through the point pattern. These channels are a consequence of the splitting procedure and arise in the following manner. Consider the initial split of the 1395 (land and sea) constituencies. Precisely one has its x–coordinate assigned to the median value $w/2$. At the next iteration there are four groups, two of size 348 and two of size 349. The latter two cause precisely two constituencies to have their x–coordinates fixed (at $w/4$ or $3w/4$). At the next iteration there are 16 groups each of size 87, so that four constituencies have their x–coordinates fixed at each of $w/8$, $3w/8$, $5w/8$ and $7w/8$. Successive iterations lead to steadily increasing numbers of x–coordinates being fixed, with, in general, each subdivision of w being allocated to a greater number of constituencies. However, none of the subsequent allocations are to $w/2, w/4, 3w/4, w/8, \ldots$, and so, inevitably, these channels appear in the data.

Therefore, the channels apparent in Figure 9 are a direct consequence of the fact that the number of points being rearranged (1395) is not a power of 2. With, for example, 1024 points, there will be no premature allocations to any of $w/2, w/4, 3w/4, \ldots$. Instead, in the final stage the points will be allocated to $w/2048, 3w/2048, 5w/2048, \ldots$, and thus all of them will lie on an evenly spaced grid.

Figures 10*a* and 10*b* show the results of confining attention to the 523 English constituencies and surrounding them by 501 "sea constituencies." The figures differ in that in Figure 10*a*, in each iteration, the first adjustment was made to the x–coordinates, and the second to the y–coordinates, whereas in Figure 10*b*, in each iteration, the first adjustment was made to the y–coordinates, and the second to the x–coordinates. Examination of the figures reveals that while most of the regions remain as compact entities, there are a few instances in which some constituencies have lost contact with the bulk of their region. Figure 10*b* is noticeably worse in this respect.

In addition to the arbitrary order of coordinate adjustment, there are many other arbitrary elements, such as the precise position of the enveloping rectangle and the choice of positions for the phantom sea constituencies; there is no unique end product.

3.2. From points to areas

After all the point positions have been adjusted to their final positions, one then could redraw the constituency boundaries. However, in view of the arbitrary nature of the transformations employed, no attempt will be made here to restore the contiguities present in the original data. But, by connecting each constituency to all its original neighbours, using the revised coordinates, an idea of the distortions that have been induced may be attained. Figure 11 shows results for the English constituencies, using the revised coordinates illustrated in Figure 10*a*.

If all the constituencies were in their same relative positions, then one would see a network of short lines with very few crossovers. Figure 11 shows that this holds true for Greater London and the North West, but, not unexpectedly, the vast expansions of these regions have resulted in considerable stress around their peripheries. The initial binary

Figure 9.

Map of Britain showing constituencies in revised positions (symbols correspond to regions).

Figure 10a.

Revised English constituency positions with first revision performed on x co–ordinates.

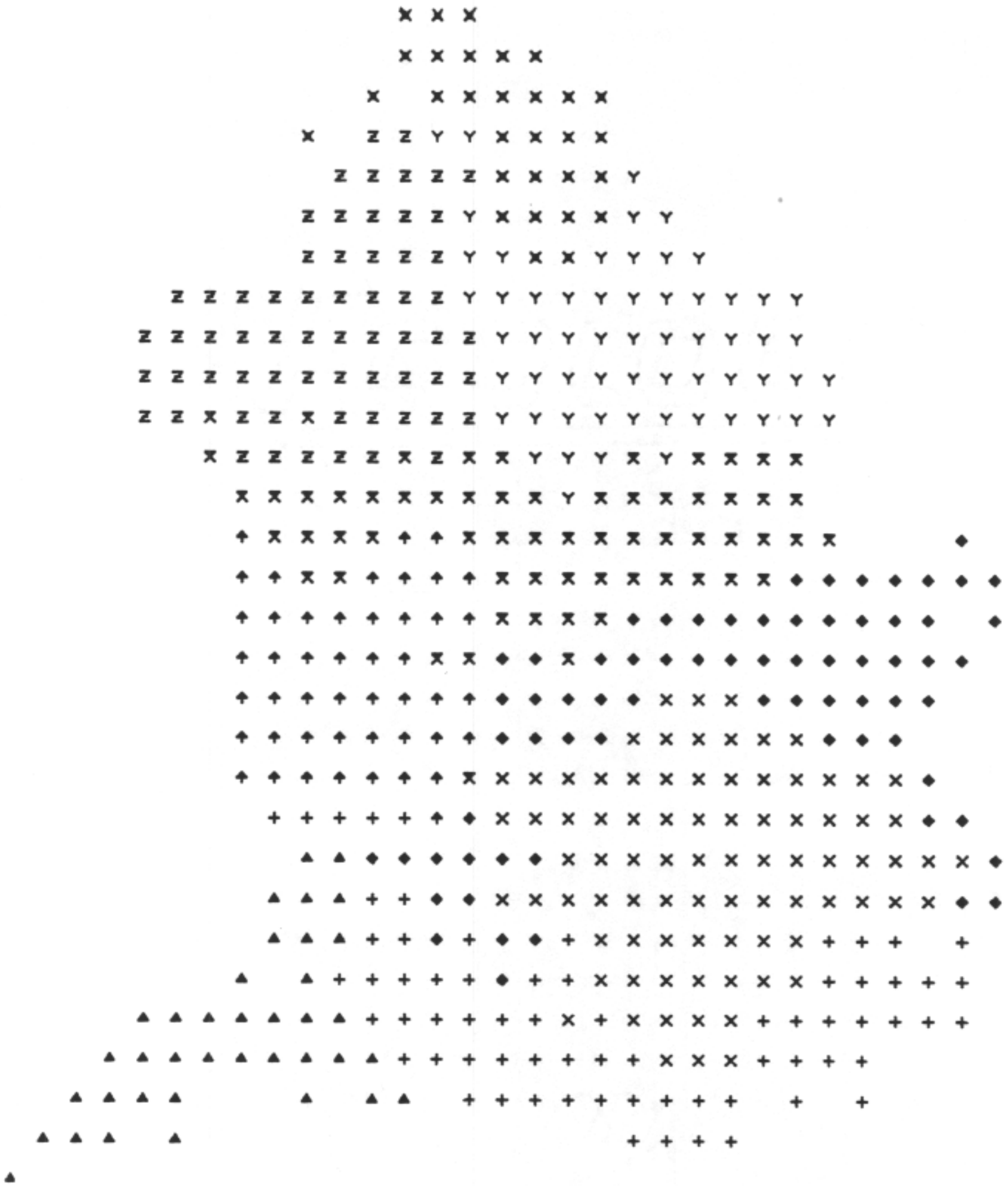
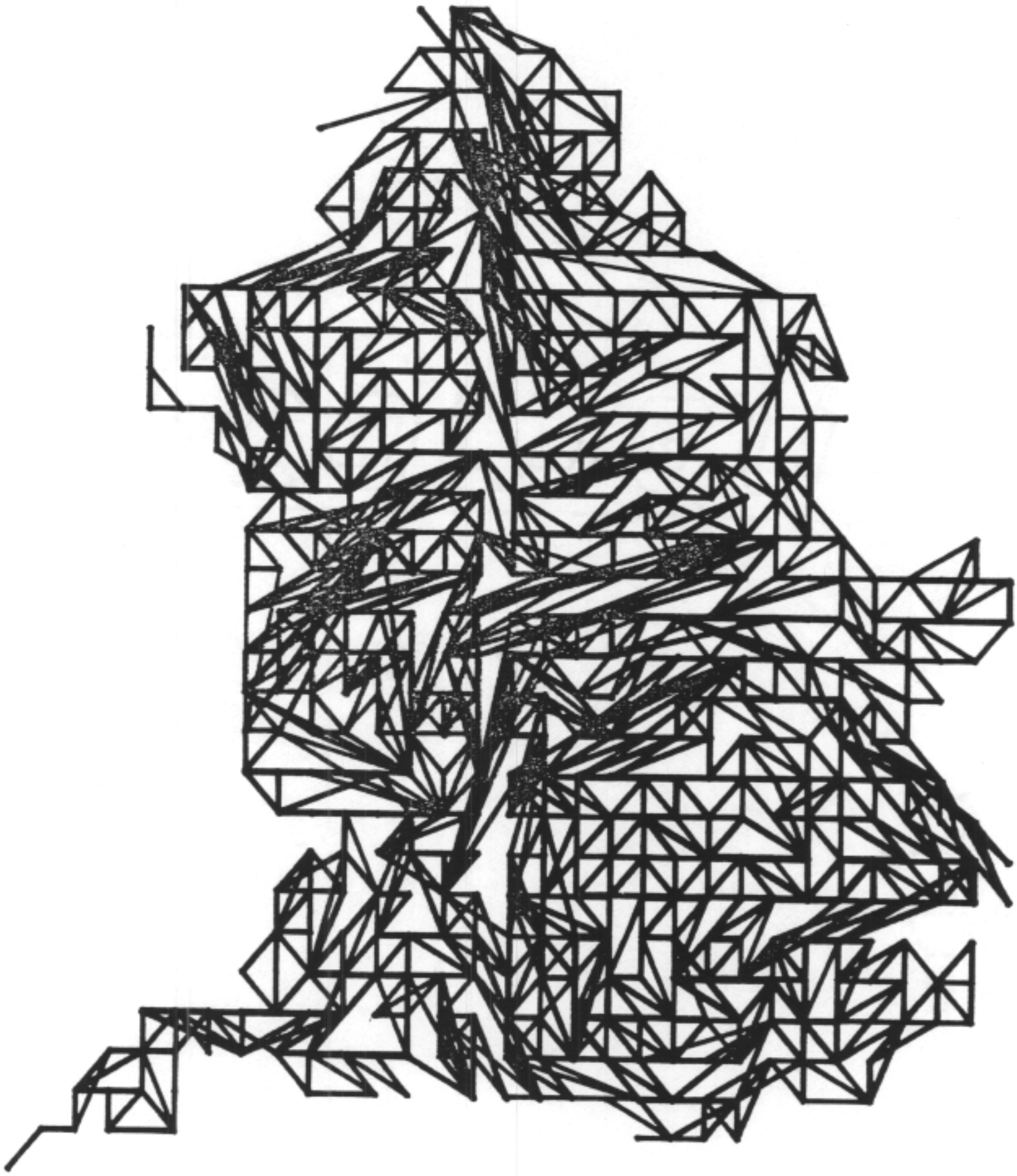


Figure 10b.

Revised English constituency positions with first revision performed on y co–ordinates.

Figure 11.

Map of revised nearest neighbour connectivities showing the distortions induced by scaling.



splits impart particularly noticeable displacements, which are evidenced not only by the long diagonal lines, but also by the lack of horizontal lines in the central column, as well as the lack of vertical lines in the central row. Those constituencies most affected fall to the east of an expanding Birmingham and lie in the west of Yorkshire, which are being squeezed by the simultaneous expansions of Tyneside in the North region and Merseyside in the North West.

In order to quantify the stress, a list was made, for each English constituency, of its ten nearest neighbours. After each iteration a count was made of the number of these nearest neighbours that remained among the ten nearest constituencies according to the revised positions. The results are summarised in Table 4.

Table 4 shows that in the final map transformation arrangement (Figure 10*a*) all constituencies retained at least one of their original ten nearest neighbours in their revised ten, with only two retaining all ten original neighbours. Not unexpectedly, the greatest changes are experienced during the first few iterations. Constituencies affected most adversely are listed in this table, and interestingly these are not persistent, which indicates that subsequent iterations can "repair the damage" done by previous iterations. Bosworth, Sutton Coldfield, and Bromsgrove are constituencies in central England in the neighbourhood of Birmingham that become displaced because of the simultaneous expansions of the Birmingham and Greater London conurbations. Skipton and Ripon, Bishop Auckland, and Keighley become displaced because of the simultaneous expansions of the Tyneside and west Yorkshire conurbations. In terms of actual first neighbours, Bishop Auckland retains 3 out of 5, Bromsgrove 4 out of 10, and Keighley 2 out of 4; the figures in Table 4 are pessimistic with respect to these constituencies, since each is physically close to a multi–constituency urban centre.

TABLE 4
DISPLACEMENT OF THE TEN NEAREST NEIGHBOURS OF 523 ENGLISH CONSTITUENCIES AFTER EACH ITERATION OF THE ADJUSTMENT PROCEDURE

After	Number Still in Nearest 10											Mean	Worst
Iteration	0	1	2	3	4	5	6	7	8	9	10	Number	Cases
0	0	0	0	0	0	0	0	0	0	0	523	10.0	*
1	0	0	1	3	2	6	9	14	58	210	220	9.1	*
2	0	0	1	7	10	19	31	68	124	179	84	8.2	*
3	1	0	2	15	18	51	71	118	121	93	33	7.2	Skipton, Ripon
4	0	1	5	14	23	55	115	120	101	69	20	6.8	Bosworth
5	0	2	8	16	37	81	127	115	81	47	9	6.4	Bosworth, Sutton Coldfield
6	0	3	8	21	68	116	114	103	68	20	2	5.9	Bishop Auckland, Bromsgrove, Keighley

Figure 11 shows the stress induced when an arbitrarily oriented rectangle including arbitrarily placed "sea constituencies" was used to surround the genuine constituencies. The placement of the "sea constituencies" and the size of the rectangle used are unlikely to make much difference to the outcome; but an improved rearrangement undoubtedly could be obtained by choosing the orientation of the rectangle so as to minimize the stress induced. An untested hypothesis is that the rectangle of minimum area would be near optimal.

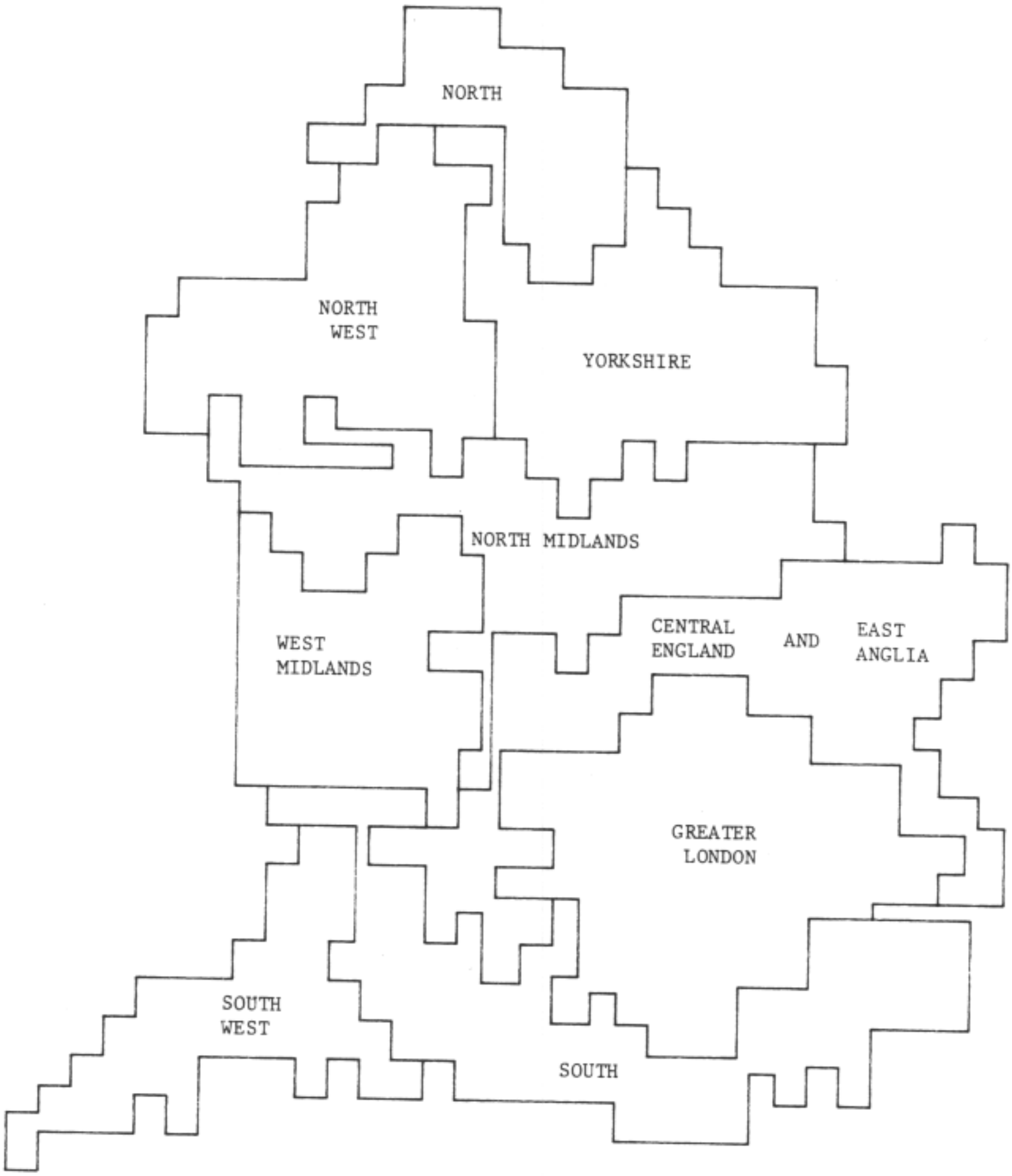
The array of symbols in Figure 10*a* is potentially distracting, so Figures 12*a* and 12*b* show the effects of rescaling in areal terms. Figure 12*a* shows the regions of England as defined by Waller (1983), while Figure 12*b* is a rendering of Figure 10*a* into a comparable

Figure 12a.

Map of England showing regions before constituency rearrangement.

Figure 12b.

Regions of England after rescaling (from Figure 10a).



form. This figure suggests that the North Midlands and Yorkshire regions have become noticeably misshapen, and have developed a number of "pseudopodia" corresponding to the few severe constituency displacements noted earlier.

Before leaving the English voting data, Figure 13 shows the application of this procedure to the study of voting behaviour within the context of the 1987 general election. The symbols C, L and A have been used to denote constituencies in which the Conservative, Labour and Alliance parties had majorities of more than 10% over their nearest rival. Marginal seats (less than a 10% lead) are indicated by one of the integers $1, 2, \ldots, 6$, depending upon the voting outcome. The main features apparent from this figure are the North–South divide, with Labour seats primarily in the North and the conurbation centres, and the clusters of marginal constituencies on the urban–rural interfaces of the West Midlands and the South-East of London. The boundaries marked are those from Figure 12*a*.

One should note that the type of data portrayed in Figure 13 can be much better displayed in colour. With colour one replaces the C, L and A symbols by different coloured dots, and the six integers by the letters C, L and A in an appropriate colour. The colour represents the party that won the seat and the letter identifies the runner–up. More quantitative displays also are possible.

3.3. Rescaling areas

This section reviews a very flexible procedure devised by my colleague Dr. Fremlin for the computer production of cartograms. This procedure is still under development, but seems to me to provide exciting possibilities, since it works directly with areas, and hence guarantees that existing geographical contiguities are preserved. Figure 14 shows the final output of the current version of the procedure. This figure shows all the states of the USA (together with the District of Columbia) scaled in such a way that the area of each is proportional to its population. The figure is plainly recognizable, and while the scaling obviously distorts the states, the undesirable "pseudopodia" of the previous method (see Figure 12*b*) are not present here; recognition of individual states is a simple matter, except for a few in the mountains of the west (and except for Alaska!).

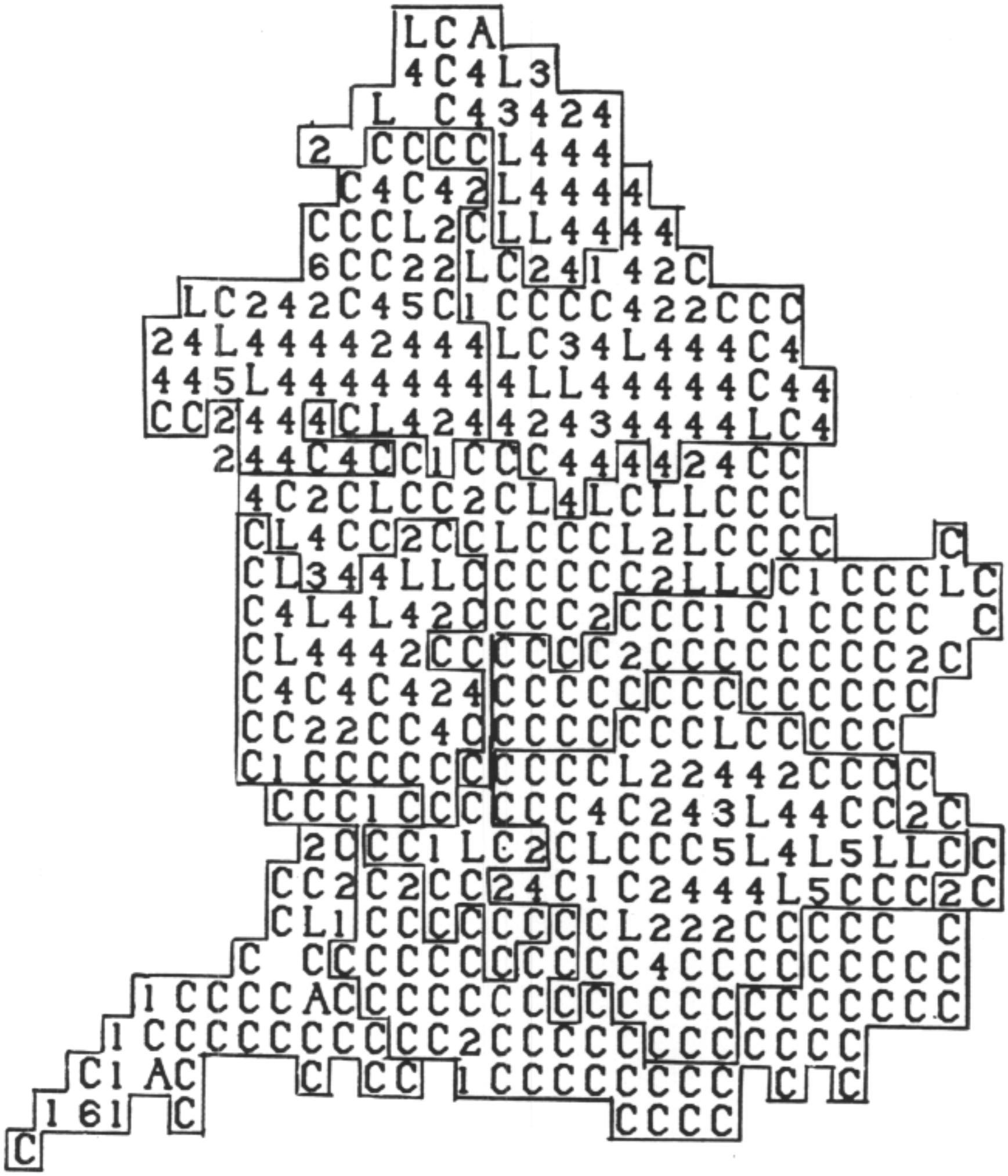
The idea of harnessing the computer to produce cartograms is not novel (*e. g.*, Tobler, 1973), but has received rather little attention in the literature. In part this may be because of the natural reluctance of the cartographer to accept the potentially inartistic output of the computer. An example of an unacceptable computer–produced cartogram is provided by Sen (1976), and the warning of Griffin (1980) is relevant: "the novelty of an automated approach may lead to intemperate haste in its utilization ...." It should be noted that the computer–produced cartogram of 1970 United States population reproduced on page 119 of Muehrcke and Muehrcke (1986) is in fact grossly inaccurate.

This is not the place for a detailed mathematical description of Fremlin's procedure, but it is possible to give a general idea of his strategy. The essential input to the program is a list of the coordinate positions of the vertices of the polygons that are used to approximate the boundaries of the states, together with the values (here population sizes) for the states. The procedure can be used equally well to display the importance of the states in terms of features other than population—all that would be required is a change in the input values.

The first stage of the iterative procedure consists of dividing the present version of the country map into a rectangular array of cells, some of which may be completely or partly

Figure 13.

The results for the 523 English constituencies in the 1987 election.

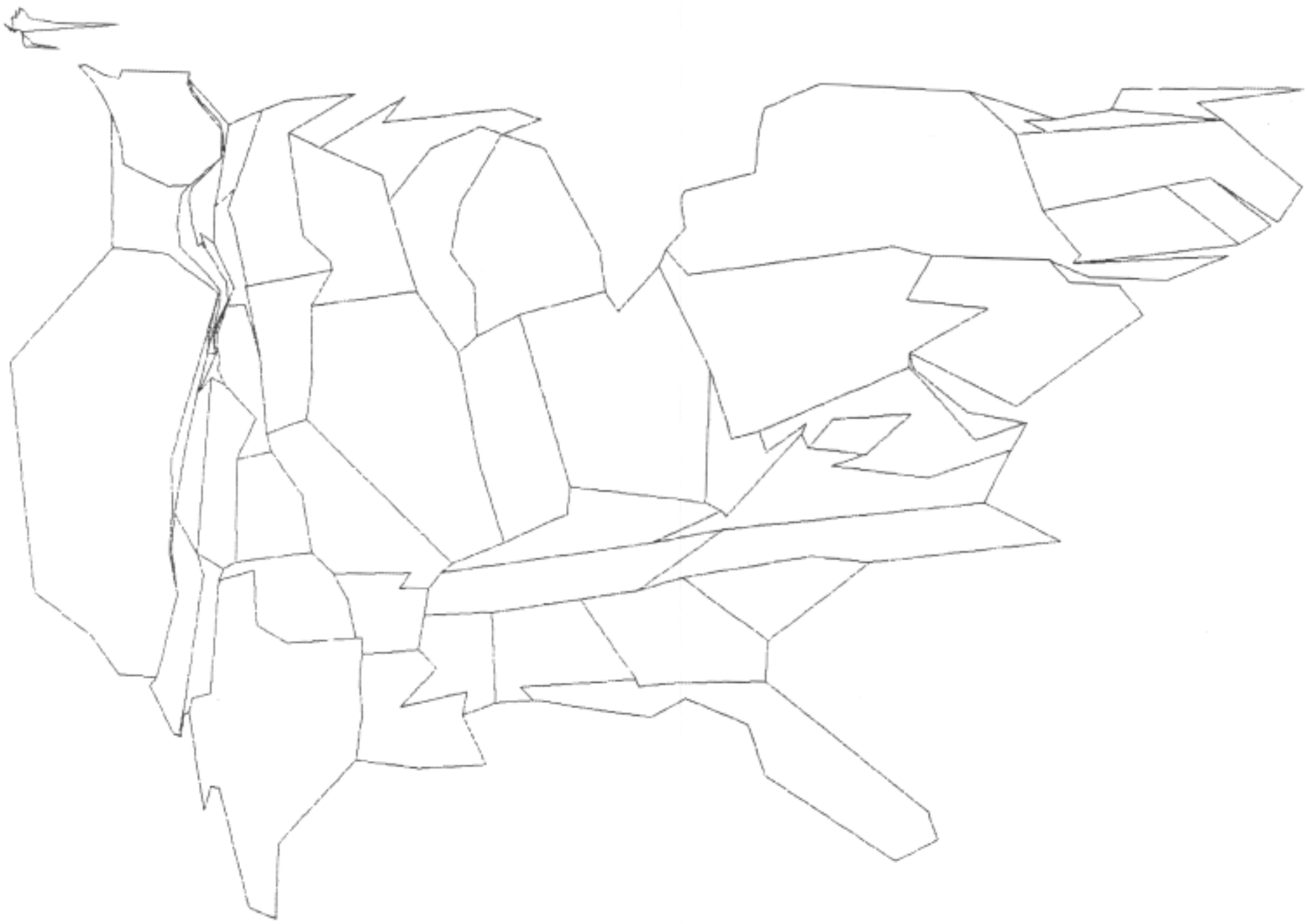


Key: C, L, A Seats having ≥10% majority for the Conservative (C), Labour (L) and Alliance (A) parties.

1-6 refer to other (marginal) seats in which the party orders are as follows: 1 CAL, 2 CLA, 3 LAC, 4 LCA, 5 ALC, 6 ACL.

Figure 14.

The United States (including Alaska) scaled using Fremlin's procedure and retaining contiguities.



filled with sea. For the first iteration the "present version" is the usual geographical map. The total land area and the estimated total population (assuming uniform density within a state) then is calculated separately for each column of the rectangle. From these results calculation of the average population density within that column is easy, and this average value is used to invent "sea people" to inhabit the areas of sea lying in this column, so that the overall density for the column matches the density for the land in that column.

Various improvements could be considered in the implementation of the first stage, which, in the form described above, suffers from the assumption of uniform population density within

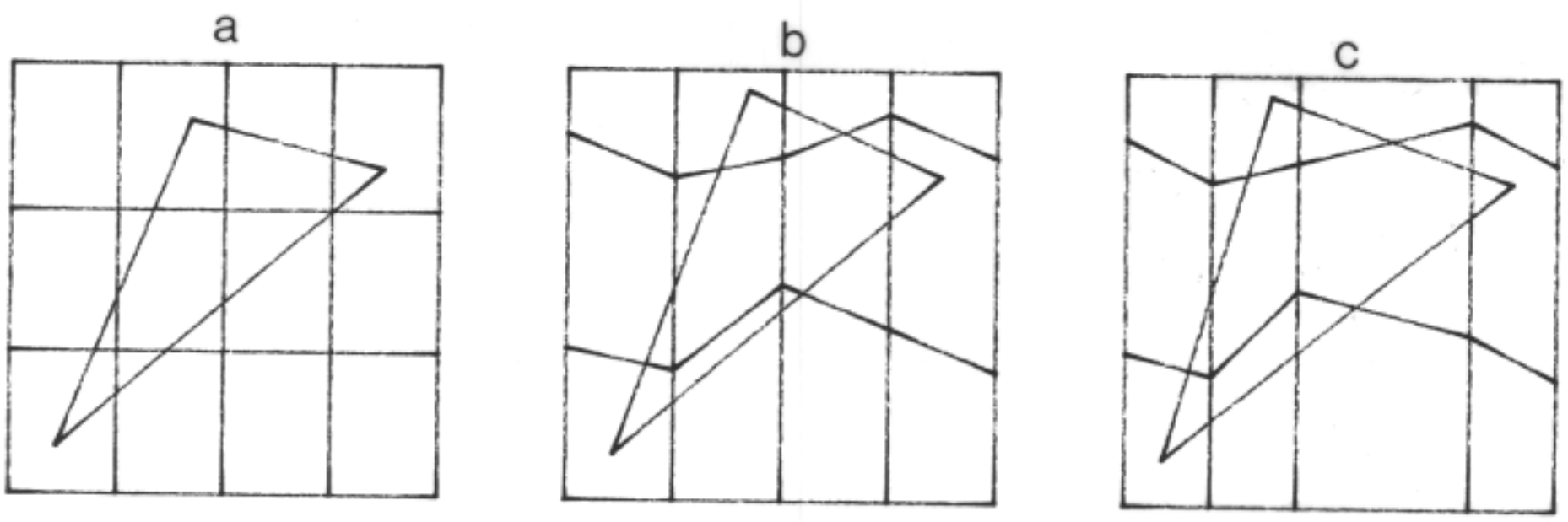
a state. If individual population figures and precise boundaries are available for some or all of the subregions of a state (*e. g.*, principal cities or counties), then this information could be used to provide improved estimates of the populations in the cells of the rectangular array. Alternatively, it might be thought appropriate to employ Tobler's (1979) pycnophylactic approach to produce a smoothed version of the underlying population density before implementing the first stage procedure.

The second stage also treats each column separately. The initially rectangular cells are now distorted into trapezia having vertical sides—in other words, the initially horizontal top and bottom of each cell are tilted in some appropriate manner. The nature of this tilting is a function of the population densities of the cells in neighbouring columns, with the aim being to make the fluctuations in density as gradual as possible.

After these adjustments have been made, the original rectangular mesh (Figure 15*a*) is changed into a mesh consisting of equi–spaced vertical lines and "wavy" horizontal lines (Figure 15*b*). A final adjustment now consists of an alteration in the column widths so as to reflect the variations in population density from column to column (Figure 15*c*). It is natural to consider altering the spacings so as to make each adjusted column have the same density. However, this adjustment turns out to create most undesirable distortions in the map, because of the large fluctuations in density that are present during the first iterations, and in practice a less extreme adjustment of the same character is used.

Figure 15.

The stages in an iteration of Fremlin's procedure.



The preceding step completes the first iteration, after which it is a comparatively easy matter to compute the revised locations of the vertices defining the regions. The map is now rotated through 90 degrees and the entire procedure is repeated. The iterations continue until the variation in population density from state to state has become acceptably small.

One necessary refinement to the procedure is required in order to avoid state boundaries crossing over each other, and hence resulting in figure–eight shapes! When computing the adjusted vertex locations, a check is made that this crossover has not happened. If it has happened, then this iteration is abandoned and is restarted with extra "vertex" coordinates being interpolated along the offending sides. This procedure may need to be implemented several times for any given iteration, but ultimately these undesired crossovers are eliminated.

It should be noted that, while the computer performs the necessary calculations, there are a number of variable parameters within the computer program. Moreover, the skill of the cartographer is translated into choosing values for these parameters that give the most acceptable cartogram for a given data set. One cannot expect optimal values for these parameters to exist that will apply to all data sets; in other words, there is still considerable scope for the blending of art with science.

4. Inferences from aggregate data

4.1. The "ecological fallacy"

The central problem of interest in this section is the following:

Given information concerning a number of *groups* of individuals, what can we conclude about the *individuals* separately?

All attempts to answer this type of question are coloured by knowledge of the results set out in a seminal paper by Robinson (1950), who used data on race and literacy taken from the 1930 U. S. census. These data are reproduced in two forms in Table 5.

TABLE 5
INFORMATION ON RACE AND LITERACY PROVIDED BY THE 1930 US CENSUS

(a) A cross–classification for the entire country (figures in thousands)

	Black	White	Total
Illiterate	1512	2406	3918
Literate	7780	85574	93354
Total	9292	87980	97272

(b) Percentage figures by region

	Percent Black	Percent Illiterate
New England	1.1	3.7
Mid Atlantic	4.0	3.5
East North Central	3.7	2.1
West North Central	2.6	1.4
South Atlantic	27.6	8.3
East South Central	27.2	9.6
West South Central	18.8	7.2
Mountain	0.9	4.2
Pacific	1.1	2.1

Robinson considered the question "To what extent are race and literacy interrelated?" Tables 5*a* and 5*b* both provide information about this relationship. Table 5*a* provides

information at the individual level, whereas Table 5*b* provides aggregated information. Multiplying the figures in Table 5*b* by the region populations and summing these products yields the marginal totals (9292, 87980, 3918, 93354) of Table 5*a*.

A correlation coefficient may be calculated from Table 5*a* using the four cell frequencies (1512, 2406, 7780 and 85574). There are various possible correlation formulæ depending upon how the two classifying variables are treated. If specific values (for example 0 and 1) are assigned to the two categories of each classifying variable, then the Pearson product–moment correlation coefficient is appropriate. This was the coefficient calculated by Robinson, and it proves to have the value 0.203, which suggests a rather slight connection between race and literacy.

Thomsen (1987) suggested calculating the *tetrachoric* correlation coefficient, which has value 0.747. The tetrachoric coefficient is appropriate when the two classifying variables are innately continuous variables (such as height and weight), but just happen to have been dichotomized (*e. g.*, into categories short and tall, light and heavy). When the joint distribution of the classifying variables is bivariate normal, this correlation coefficient is an unbiased estimate of the actual population correlation.

In the present case, neither of these correlation coefficients seems entirely appropriate, since although literacy can certainly be regarded as a continuous variable, it is difficult to regard color in that light. However, since the normal distribution is commonly used as an approximation for the binomial, the tetrachoric coefficient seems a better choice than the product moment coefficient used by Robinson.

For Table 5*b* Robinson calculated the correlation between percentage black and percentage illiterate, weighting these percentages by the populations of the regions concerned; this has been termed the *ecological* correlation, which may be denoted as ρ_E. Specifically, if region i has population n_i and the values of the two characteristics of interest are denoted by p_i and q_i, then ρ_E is calculated from the expression

$$\rho_E = \sum n_i(p_i - \bar{p})(q_i - \bar{q}) / \{\sum n_i(p_i - \bar{p})^2 \sum n_i(q_i - \bar{q})^2\}^{1/2}, \qquad (6)$$

where $\bar{p}$ is the mean of the p–values and $\bar{q}$ is the mean of the q–values. The nine pairs of regional values in Table 5*b* give the strikingly high correlation of 0.946, with Robinson having taken the discrepancy between 0.203 and 0.946 to imply that aggregated results might well give a highly misleading impression of individual–level relations between variables.

The high correlation obtained for Table 5*b* is clearly a consequence of the huge gap between the values for the three southern regions and those for the rest of the map. Because the gap is apparent in both columns of data, the correlation is very great. Firebaugh (1978) segregated the regions into two "super–regions" on this basis, and showed that, for explaining the variations in literacy rates, it is the interaction between race and super–region that is the dominant explanatory variable, rather than race itself. In a more detailed analysis, Hanushek *et al.* (1974) performed a multiple regression of literacy rates on race, available schooling and various other variables. They concluded that it was the schooling variable that played the dominant role: in 1930 in the southern states the proportions of those eligible for schooling that were actually enrolled in schools were lower than in other states. Since these same states contained the highest proportions of blacks, in the absence of schooling information it is race that appears to be the controlling variable.

Firebaugh, Hanushek *et al.*, and the many other scholars who have written about the problem posed by Robinson are in general agreement that, where there is a difference between the correlation in the table of individual–level data and the correlation apparent in a spatially aggregated form of that same data, this difference will be attributable to some unmeasured "latent" variable that has a marked spatial variation that matches the aggregation used. For Table 5 this would be the "available schooling" variable.

Robinson also considered the effect of a less extreme spatial aggregation, using state percentages, which resulted in an apparent correlation of 0.773 (very close to the tetrachoric value of 0.747). A detailed example of the way in which increasing spatial aggregation can lead to increasing apparent correlation is provided by Yule and Kendall (1950, Ch. 13). The fact that results vary according to the aggregation used is part of a more general problem entitled "the modifiable areal unit problem" by Openshaw and Taylor (1981). These authors regarded the general problem as follows:

> ... it would appear that geographers are doing their best to build what can only be described as intrinsically non–geographical models of geographical phenomena based on unreasonable assumptions regarding the nature of zonal data.

Closely related to these problems, and with the same essential cause—unmeasured latent variable(s)—are the so–called "nonsense" regressions and Simpson's paradox (Simpson, 1951). A typical example of the former is the strong positive relationship found between the numbers of clergy in Cuba at various decades of the century and the amount of rum consumed in Havana at those times (both increase as the population increases). Simpson's paradox is a discrete data analogue of the problems that arise when data from two distinct populations are combined—separate positive correlations then may appear to become a negative correlation.

4.2. Thomsen's approach

Political scientists are interested in the so–called "floating voters" whose political allegiance changes from election to election. Aggregate election results convey only the gross changes of support—if the Democrat vote increases by 5%, then (neglecting births, deaths and so forth) all one can say is that $(5+x)\%$ of the electorate moved from Republican to Democrat and $x\%$ moved from Democrat to Republican. Little is known about the size of x.

Expressed in terms of the layouts of Tables 5*a* and 5*b*, the constituency aggregated figures are comparable to the data of Table 5*b*, and the interest of the political scientist is in reconstructing the figures in a tabular form akin to Table 5*a*. There have been several attempts to achieve this goal, of which the most recent are those by Brown and Payne (1986), Johnston *et al.* (1982), and Thomsen (1987). These authors attended a recent workshop on ecological regression at Lund University, where their various approaches were discussed and where the consensus appeared to be that Thomsen's approach was the most promising.

Thomsen's description of his procedure is naturally couched specifically in terms of voting figures, but his rationale is not confined to voting and will be described here in general terms. In the case where each variable of interest has just two categories (as for the data of Table 5), the procedure leads to an explicit formula for the reconstruction of the individual–level data. When more than two categories are involved, an iterative computer program is required, with a version suitable for a PC being available from Professor Thomsen (Department of Political Science, University of Aarhus, Denmark).

In general terms, one is interested in quantifying the joint occurrence of the four category combinations of the two binary variables X and Y, both of which, without loss of generality, can be assumed to take on the values 0 or 1. An implicit assumption of the procedure is that the values of both variables are governed by the values of one or more unobserved latent variables; this is termed the latent structure approach. There is no need to specify the number of latent variables, nor to give them names. The theory extends to cover any number (providing certain assumptions are met). Solely for simplicity of presentation, suppose that there are just two latent variables, denoted by Z_1 and Z_2.

Since different regions have different characteristics, one can expect that the mean values of Z_1 and Z_2 for the inhabitants of region i, μ_{1i} and μ_{2i} say, will differ from those in region j (μ_{1j} and μ_{2j}). Since Z_1 and Z_2 are unmeasured quantities, a general scale should be set in some arbitrary way, which will be done here by assuming that for each Z–variable the various μ–values have mean 0 and variance 1. Within region i individuals will have their own specific Z–values, which will vary about their regional means μ_{1i} and μ_{2i}. Basic assumptions are that the Z–variables have a common variance, σ^2, that this variance is the same for each region, and that within a region their joint distribution is bivariate normal. Thus, within region i the joint probability density function of the Z–values may be written as $\phi(\mu_{1i}, \mu_{2i}, \sigma^2)$.

Combining the assumptions about the variability of the regional means and the variability of Z–values within a region, and assuming that the joint distribution of regional means also is bivariate normal, the overall joint distribution of the Z–values across all the regions has the bivariate normal density function $\phi(0, 0, 1 + \sigma^2)$.

Since the Z–variables influence X (and Y), it is reasonable to assume that, for a given individual, the probability that X takes on the value 0 is some function of the values of Z_1 and Z_2 for that individual. A convenient assumption is that $P(X = 0) = \Phi(a + bz_1 + cz_2)$, where the symbol Φ refers to the distribution function of the unit normal distribution. Likewise, $P(Y = 0) = \Phi(d + ez_1 + fz_2)$.

Let the overall proportion of people for whom X takes the value 0 be denoted by p. Given these various normality assumptions, the value of p is given by the double integration of the product of $\Phi(a + bz_1 + cz_2)$ and $\phi(0, 0, 1 + \sigma^2)$ over the ranges of Z_1 and Z_2. Similar double integrations render expressions for q, the overall proportion of people for whom Y takes the value 0, and for the joint probability $P(X = 0, Y = 0)$. Combining this pair of expressions yields the following one, for the overall correlation coefficient ρ_A:

$$\rho_A = (be + cf)(1 + \sigma^2)/\{[1 + (b^2 + c^2)(1 + \sigma^2)][1 + (e^2 + f^2)(1 + \sigma^2)]\}^{1/2}. \qquad (7)$$

Equation (7) is the quantity that is estimated by the tetrachoric coefficient from a complete 2–by–2 table.

At this point a recapitulation of which quantities are supposedly known and which are supposedly unknown is worthwhile. The problem posed at the beginning of this section was in effect "What can be deduced about individuals given regional values?" Given the regional values $(p_1, p_2, \ldots)$ for $P(X = 0)$, and the regional values $(q_1, q_2, \ldots)$ for $P(Y = 0)$, calculating the values of p and q is easy. Thus, in the race/literacy example these computations are given by the marginal totals of Table 5*a* as $3918/97272 = 0.0403$ and $9292/97272 = 0.0955$. These then are known values, as are the values $p_1, p_2, \ldots, q_1, q_2, \ldots$

given as percentages in Table 5*b*. However, the individual cell entries in Table 5*a* (*e. g.*, 1512) are supposedly unknown, and constitute what is to be estimated.

If the value of ρ_A were known, then, given p and q, one could reconstruct the cell entries. To do so exactly requires a computer program, but an approximation is provided by the formula

$$P(X=0,Y=0) = \{k - [k^2 - 8\rho_A(1+\rho_A)pq]^{1/2}\}/(4\rho_A), \tag{8}$$

where $k = 1-\rho_A+2\rho_A(p+q)$. This formula is a straightforward inversion of an approximation originally suggested by Yule (1897).

Although the value of ρ_A is unknown, ρ_E, the ecological correlation displayed by the set of individual regional p_i and q_i values, can be calculated using equation (6). One also could calculate the ecological correlation between a transformation of the p_i values and the same transformation of the q_i values. Thomsen shows that, if a particular transformation of the values is used, then this correlation is given by the expression

$$\rho_E = (be+cf)/[(b^2+c^2)(e^2+f^2)]. \tag{9}$$

Comparing expression (7) for ρ_A with expression (9) for ρ_E indicates that if $1/(1+\sigma^2)$ is negligible by comparison with the sums (b^2+c^2) and (e^2+f^2), then ρ_E will be approximately equal to ρ_A; otherwise it will be greater than ρ_A. In practice it appears that ρ_E often is very close to ρ_A, providing that ρ_E is computed from reasonably homogeneous areas—thus, for Robinson's data the value of ρ_E for state data exceeds ρ_A by only 0.026, whereas for the data aggregated over regions the excess correlation is appreciably greater. Thomsen himself feels that a state has too large an area to be regarded as a homogeneous unit, and hence advocates performing calculations using data at the county level.

The transformation used to obtain equation (9) was the so–called probit transformation, which is closely approximated by the simpler logit transformation, in which the proportion p_i is replaced by $\log_e\{p_i/(1-p_i)\}$. An example of the affiliated calculations is given below.

4.3. An example

Table 6*a* shows the numbers of households in the four regions of the United States in 1980. These data have been extracted from Tables 64 and 527 of the 1981 edition of the *Statistical Abstract of the United States*. The numbers of households classified as black and the numbers in receipt of food stamps also are shown, together with the corresponding p and q values (thus, for example, $1632/17447 = 0.09354$). These particular characteristics were chosen because Table 527 of the Abstract also conveys the information that a total of 2409 black families were in receipt of food stamps. The intention here is to use Thomsen's procedure to obtain an estimate of this total figure from ecological data, typified by those appearing in Table 6*b*. Knowing the true value furnishes a guide to how well the procedure has performed.

Table 6*b* summarizes, for the Northeast region, the information available in Tables 32 and 203 of the 1982 edition of the Abstract. Although this information refers to persons rather than households, this reference creates no difficulty providing that, for example, the mean size of black households does not vary appreciably from state to state within a region. The (false) assumption of equal family sizes for black and non–black families is not required—this simply generates a scaling factor that cancels out in the computation of ρ_E.

TABLE 6
INFORMATION CONCERNING HOUSEHOLDS OF THE UNITED STATES

(a) Numbers of households (figures in thousands) in various regions in March 1980

Region	Total	Black	Receiving food stamps	p	q
Northeast	17447	1632	1359	0.09354	0.07789
Midwest	20933	1808	1251	0.08637	0.05976
South	25523	4125	2401	0.16162	0.09407
West	15205	840	900	0.05524	0.05919

(b) Numbers of people (figures in thousands) in Northeast region at April 1, 1980

State	Total	Receiving food stamps	Others	Weighted logit	Black	Others	Weighted logit
ME	1125	139	986	− 65.7	3	1122	−198.7
NH	921	53	868	− 84.8	4	917	−164.9
VT	511	46	465	− 52.3	1	510	−140.9
MA	5737	446	5291	−187.3	221	5516	−243.7
RI	947	87	860	− 70.5	28	919	−107.4
CT	3108	174	2934	−157.5	217	2891	−144.4
NY	17558	1804	15754	−287.2	2402	15156	−244.1
NJ	7365	600	6765	−207.9	925	6440	−166.5
PA	11864	1030	10834	−256.3	1047	10817	−254.4

(c) Results

Region	Correlation Between Weighted Logits	Estimated number of black families receiving food stamps	
Northeast	0.7236	450	
Midwest	0.5649	289	
South	0.7883	1298	
West	0.9087	372	
		2409	Total
		2417	True number
		8	Error of estimation

One should note that if Table 6*a* had referred to people rather than households, then it could have been constructed from an extended version of Table 6*b*. In essence, therefore, one is hoping to estimate the overall number for the black/food stamp combination simply from aggregate state values. Thus this is a classic example of the ecological problem.

The calculation of the weighted logits in Table 6*b* requires explanation. Consider the 17558 households in New York state, of which 1804 received food stamps. Thus, $p_{NY} = 1804/17558$, and hence $U_{NY} = \log_e [p_{NY}/(1 - p_{NY})] = \log_e(1804/15754)$. Similarly, $q_{NY} =$

2402/17558, so that $V_{NY} = \log_e[q_{NY}/(1-q_{NY})] = \log_e(2402/15156)$. In the same way, U–values and V–values can be obtained for each of the other states in the region, and a straightforward (unweighted) correlation between the U–values and the V–values would involve sums such as $C = (U_{ME}V_{ME} + \cdots + U_{MA}V_{MA})$. However, sums calculated in this way fail to account for the precision with which the various logit values are known. Evidently more attention should be paid to states with large populations because these will give more precise estimates. It turns out that an appropriate set of weights is provided by the set of state populations $N_{ME}, N_{NH}, \ldots, N_{PA}$. Therefore the value of C is replaced by the weighted version $(N_{ME}U_{ME}V_{ME} + \cdots + N_{MA}U_{MA}V_{MA})$. In view of the other sums that need to be calculated, it is convenient to write the product $N_{ME}U_{ME}V_{ME}$ as $N_{ME}^{1/2}U_{ME} \cdot N_{ME}^{1/2}V_{ME}$, and the quantities $N^{1/2}U$ and $N^{1/2}V$ define the weighted logits in Table 6*b*.

Table 6*c* summarizes the final results for this procedure. The value for ρ_E for the Northeast region is 0.7236. Using this value for ρ_A in equation (8), with the values of p and q (0.09354 and 0.07789) given in Table 6*a*, leads to the estimate $P(X=0, Y=0) = 0.02579$. Multiplying this value by the regional population, 17447, one gets an estimated 450 thousand black families in the Northeast region that received food stamps. Summing the regional estimates leads to an estimated total of 2409 thousand black families receiving stamps in the nation as a whole. This is remarkably close (a 0.33% error) to the true value of 2417 thousand; but, it would be unwise to expect that the method always would perform that well.

5. Conclusions

Regions provide a natural and convenient framework for summarising geographical data. Invariably government statistics are summarised in this way, using various levels of aggregation, ranging from aggregations of states, states themselves, counties, cities, wards, and postal districts.

One might have the impression that data so frequently used and so commonly cited would be transparent to the user. However, as this paper has shown, this is not the case. It is not always easy to display regional data in a manner appropriate to its use, and the map used often is a matter of judgement by the cartographer. At present interpolation of univariate regional data essentially is an empirical task. Multivariate regional data provide a minefield of potential misunderstandings.

Given the difficulties inherent with this common type of geographical data, it is a surprise to find that it has been accorded rather little attention in either the geographical or the statistical literature. The purpose of this paper has been to highlight these difficulties, and hence to stimulate further research on the interesting problems that regional data present.

6. References

Brown, P., and C. Payne. (1986) Aggregate data, ecological regression and voting transitions, *Journal of the American Statistical Association.* **81**, 452-460.

Cleveland, W., and R. McGill. (1984) Graphical perception: Theory, experimentation, and application to the development of graphical methods, *Journal of the American Statistical Association.* **79**, 531-554.

Cozens, P., and K. Swaddle. (1987) The British General Election of 1987, *Electoral Studies.*

6, 263-266.

Crain, I. (1970) Computer interpolation and contouring of two–dimensional data: a review, *Geoexploration*. **8**, 71-86.

Dunn, R. (1987) Variable–width framed rectangle charts for statistical mapping, *The American Statistician*. **41**, 153-156.

Evans, I., and K. Jones. (1981) Ratios and closed number systems, in *Quantitative Geography: A British View*, edited by (N. Wrigley and R. Bennett, pp. 123-134. London: Routledge and Kegan Paul.

Firebaugh, G. (1978) A rule for inferring individual–level relationships from aggregate data, *American Sociological Review*. **43**, 557-572.

Gale, N., and W. Halperin. (1982) A case for better graphics: the unclassed choropleth map, *The American Statistician*. **36**, 330-336.

Gatrell, A. (1981) Multidimensional scaling, in *Quantitative Geography: A British View*, edited by N. Wrigley and R. Bennett, pp. 151-163. London: Routledge and Kegan Paul.

Griffin, T. (1980) Cartographic transformation of the thematic map base, *Cartography*, **11**, 163-174.

Haining, R., D. Griffith, and R. Bennett. (1984) A statistical approach to the problem of missing spatial data using a first–order Markov model, *The Professional Geographer*. **36**, 338-345.

Hanushek, E., J. Jackson, and J. Kain. (1974) Model specification, use of aggregate data, and the ecological correlation fallacy, *Political Methodology*. **1**, 89-107.

Johnston, R. (1985) *The Geography of English Politics*. London: Croom Helm.

Johnston, R., A. Hay, and P. Taylor. (1982) Estimating the sources of spatial change in election results: a multiproportional matrix approach, *Environment and Planning A*. **14**, 951-961.

Kennedy S., and W. Tobler. (1983) Geographic interpolation, *Geographical Analysis*. **15**, 151-156.

Lam, N. (1981) The reliability problem of spatial interpolation models, *Modeling and Simulation*. **12**, 869-876.

Lam, N. (1983) Spatial interpolation methods: a review, *The American Cartographer*. **10**, 129-149.

Muehrcke, P., and J. Muehrcke. (1986) *Map Use: Reading, Analysis and Interpretation*. Madison, WI: JP Publications.

Office of Population Census and Surveys. (1981) *Census 1981: Parliamentary Constituency Monitors—1983 Boundaries*. London: Government Statistical Office.

Openshaw, S., and P. Taylor. (1981) The modifiable areal unit problem, in *Quantitative Geography: A British View*, edited by N. Wrigley and R. Bennett, pp. 60-69. London: Routledge and Kegan Paul.

Porter, P. (1958) Putting the isopleth in its place, *Proceedings*, Minnesota Academy of Science. **26**: 372-384.

Ripley, B. (1981) *Spatial Statistics*. Chichester: Wiley.

Robinson, W. (1950) Ecological correlations and the behavior of individuals, *American Sociological Review*. **15**: 351-357.

Sen, A. (1976) On a class of map transformations, *Geographical Analysis*. **8**: 23-37.

Simpson, E. (1951) The interpretation of interaction in contingency tables, *Journal of the Royal Statistical Society*. **13B**: 238-241.

Statistical Abstract of the United States, 1981, 1982, 1986. Washington: U. S. Department of Commerce, Bureau of the Census.

Stetzer, F. (1982) Specifying weights in spatial forecasting models: the results of some experiments, *Environment and Planning A*. **14**: 571-584.

The Mitchell Beazley Concise Atlas of the Earth, 1973. London: Mitchell Beazley.

Thomsen, S. (1987) *Danish Elections 1920–79: A Logit Approach to Ecological Analysis and Inference*. Aarhus: Politica.

Tobler, W. (1973) A continuous transformation used for districting, *Annals of the New York Academy of Science*. **219**: 215-220.

Tobler, W. (1979) Smooth pycnophylactic interpolation for geographical regions, *Journal of the American Statistical Association*. **74**: 121-127.

Tobler, W., and S. Kennedy. (1985) Smooth multidimensional interpolation, *Geographical Analysis*. **17**: 251-257.

Upton, G. (1985) Distance–weighted geographic interpolation. *Environment and Planning A*. **17**: 667-671.

Upton, G. (1989) The components of voting change in England 1983–1987, *Electoral Studies*. **8**: 59-74.

Upton, G., and B. Fingleton. (1985) *Spatial Data Analysis by Example*, Volume 1: Point Pattern and Quantitative Data. Chichester: Wiley.

Waller, R. (1983) *The Atlas of British Politics*. London: Croom Helm.

Yule, G. (1897) On the theory of correlation, *Journal of the Royal Statistical Society*. **60**: 812-854.

Yule, G., and M. Kendall. (1950) *An Introduction to the Theory of Statistics* (14th ed.). London: Griffin.

APPENDIX A

Data for the United States

State	Area	Latitude	Longitude	Data Set I	II	III	IV	V	VI	VII
Alabama	131994	32.83	87.00	67.9	13.3	13.8	73.8	13.1	60.5	151.1
Arizona	295121	34.00	112.00	15.6	9.4	9.3	82.4	53.1	66.4	70.4
Arkansas	135403	34.83	93.67	37.0	9.1	10.1	82.7	18.9	60.5	112.4
California	406377	37.50	119.50	127.6	11.7	9.9	76.2	18.5	57.5	61.9
Colorado	269347	39.50	105.50	21.3	7.3	9.1	89.0	30.8	63.4	51.9
Connecticut	12667	41.75	72.75	623.6	4.1	11.1	90.1	2.5	60.7	46.6
Delaware	5023	39.17	75.50	276.5	6.7	14.1	82.1	8.4	59.8	63.6
D. C	164	38.90	77.02	999.9	999.9	21.2	26.9	−15.6	13.7	122.0
Florida	140798	28.00	82.00	125.5	11.0	12.8	84.0	43.5	65.3	55.8
Georgia	150946	32.83	83.25	79.0	14.4	12.7	72.2	19.1	60.2	96.5
Idaho	214271	45.00	115.00	8.6	5.4	9.9	95.6	32.4	72.4	55.9
Illinois	144677	40.00	89.00	199.4	9.9	13.6	80.8	2.8	56.2	95.9
Indiana	93423	40.00	86.25	143.9	6.2	11.4	91.1	5.7	61.7	72.9
Iowa	145509	42.25	93.25	50.5	2.6	10.2	97.4	3.1	53.3	64.6
Kansas	212623	38.75	98.25	27.5	5.7	10.4	91.7	5.1	66.3	48.4
Kentucky	103139	37.50	85.25	81.2	9.0	12.0	92.3	13.7	60.0	149.9
Louisiana	115755	31.25	92.25	81.0	15.8	13.0	69.2	15.4	60.8	140.1
Maine	80587	45.25	69.25	32.1	2.7	9.0	98.7	13.2	60.8	96.9
Maryland	25576	39.00	76.75	396.6	8.2	11.9	74.9	7.5	52.5	68.5
Massachusetts	20342	42.25	71.83	727.0	3.7	10.1	93.5	0.8	51.2	59.2
Michigan	148080	44.00	85.00	156.2	10.6	12.1	85.0	4.3	59.2	112.2
Minnesota	206825	46.00	94.25	48.0	2.0	9.5	96.6	7.1	49.6	54.3
Mississippi	122806	32.83	89.50	46.9	12.6	15.4	64.1	13.7	61.9	190.1
Missouri	179257	38.50	93.50	67.8	10.4	11.7	88.4	5.1	60.0	74.1
Montana	378009	47.00	110.00	4.8	4.8	10.1	94.1	13.3	60.5	66.7
Nebraska	199274	41.50	100.00	19.4	3.0	10.0	94.9	5.7	70.6	54.2
Nevada	285724	39.00	117.00	4.4	15.5	10.2	87.4	63.8	65.8	34.0
New Hampshire	23382	43.58	71.67	81.7	1.4	11.0	98.8	24.8	68.6	29.7
New Jersey	19417	40.25	74.50	953.1	5.4	11.7	83.2	2.7	60.1	63.6
New Mexico	315471	34.50	106.00	8.4	10.2	11.3	75.1	28.1	59.7	107.4
New York	123180	43.00	75.00	381.3	10.3	12.1	79.5	− 3.7	53.8	103.5
North Carolina	126992	35.50	80.00	104.1	10.8	13.7	75.8	15.7	61.9	75.3
North Dakota	180180	47.50	100.25	8.9	1.2	10.6	95.9	5.7	64.8	39.4
Ohio	106610	40.25	82.75	260.0	6.9	11.5	88.9	1.3	58.9	104.5
Oklahoma	178503	35.50	98.00	37.2	8.5	12.3	85.9	18.2	68.6	77.6
Oregon	250078	44.00	121.00	21.7	5.0	10.5	94.6	25.9	55.9	79.7
Pennsylvania	116709	40.75	77.50	272.3	6.2	11.6	89.8	0.5	53.3	89.9
Rhode Island	2743	41.67	71.50	902.5	4.0	10.0	94.7	− 0.3	51.7	73.8
South Carolina	78528	34.00	81.00	85.7	11.5	16.1	68.8	20.5	63.6	112.4
Tennessee	107003	35.83	85.50	94.9	9.4	12.0	83.5	16.9	57.8	110.2
South Dakota	197475	44.25	100.00	8.8	1.9	10.2	92.7	3.7	63.0	63.7
Texas	681244	31.50	99.00	42.7	14.2	10.9	78.7	27.1	63.6	74.6
Utah	213390	39.50	111.50	12.9	3.7	11.0	94.7	37.9	74.5	44.2
Vermont	24110	43.83	72.75	47.9	3.3	9.3	99.1	15.0	57.9	84.9
Virginia	103230	37.50	78.75	116.9	8.8	12.8	79.1	14.9	62.3	64.9
Washington	172929	47.50	120.50	51.2	4.6	10.6	91.5	21.1	55.8	60.7
West Virginia	62709	38.75	80.50	72.5	6.8	11.4	96.2	11.8	55.1	135.8
Wisconsin	141508	44.75	89.50	81.1	2.5	9.5	94.4	6.5	54.2	74.1
Wyoming	252171	43.00	107.50	3.4	7.1	9.8	95.0	41.3	70.5	43.1

Data Set I: Population density

Data Set II: Murders (per 100,000)

Data Set III: Infant mortality (per 1000)

Data Set IV: Percent white, 1980

Data Set V: Percent increase in population

Data Set VI: Percent voting Republican

Data Set VII: Food stamp recipients (per 1000)

APPENDIX B

The British Regions used by Waller (1983)

1. **The South West**:

Counties of Avon, Cornwall, Devon, Dorset, Somerset — 38 constituencies.

2. **The South of England**:

Counties of East Sussex, Gloucestershire, Hampshire, Isle of Wight, Kent, Surrey, West Sussex, Wiltshire — 68 constituencies.

3. **Greater London**:

84 constituencies.

4. **Central England and East Anglia**:

Counties of Bedfordshire, Buckinghamshire, Cambridgeshire, Essex, Hertfordshire, Norfolk, Oxfordshire, Suffolk — 69 constituencies.

5. **The West Midlands**:

Counties of Hereford and Worcester, Shropshire, Warwickshire, West Midlands — 47 constituencies.

6. **The North Midlands**:

Counties of Cheshire, Derbyshire, Leicestershire, Lincolnshire, Northamptonshire, Nottinghamshire, Staffordshire — 63 constituencies.

7. **The North West**:

Counties of Greater Manchester, Lancashire, Merseyside — 64 constituencies.

8. **Yorkshire**:

Counties of Cleveland, Humberside, North Yorkshire, South Yorkshire, West Yorkshire — 60 constituencies.

9. **The North of England**:

Counties of Cumbria, Durham, Northumberland, Tyne and Wear — 30 constituencies.

10. **Wales**:

38 constituencies

11. **Scotland**:

72 constituencies.

DISCUSSION

"Information from regional data"

by Graham J. G. Upton

Upton addresses three interesting questions concerning spatial statistics, within the context of several empirical examples. Two of these questions deal with unresolved issues in spatial statistics, while one illustrates the lack of awareness by many non–geography spatial statisticians of developments that have occurred in the geographical sciences. This latter view is reinforced by some of the graphics research that has been undertaken at the National Center for Supercomputer Applications (University of Illinois), too, and in part can be seen as a justification for the establishment of the National Center for Geographic Information and Analysis by the National Science Foundation.

Upton first discusses the determination of appropriate weights for numerically handling spatial autoregressive structures. He reports an all too common finding in this section, namely that those weights producing estimates most closely resembling the geographic pattern in question still leave a visible geographical element in residuals. Knowing how to statistically evaluate this residual situation is quite difficult, if presently not impossible, as Cliff and Ord (1981) have noted. The reader should recognize here that Upton is somewhat in error, in that the approach he suggests is not really all that novel. Work on the missing data problem for spatial surfaces has been studied in considerable detail by Martin (1984, forthcoming) as well as Haining, Griffith, and Bennett (1984, 1989). Flowerdew and Green (1988, 1989) have explored the problem of estimating disaggregated areal unit values from aggregate values. Further, two drawbacks should be acknowledged concerning Upton's analysis. First, as fractals and cartographic line generalization research indicate, boundary length measuring is both difficult and of questionable accuracy. Second, an overlooked and very serious problem with trend surface modelling is the underlying geographic distribution of points (see Unwin and Wrigley, 1987). As a supplement, Griffith (1988, 1989) has reported some interesting findings concerning outlier diagnostics for geo–referenced data that at least supplement Upton's work, here, too.

This first section raises several troublesome questions in this reader's mind, as well. First, does ignoring the "... variation in distance with longitude ..." introduce serious measurement error? Second, is the "first neighbors only" scheme equivalent to a conditional autoregressive (CAR) model, and is the "first plus second neighbors" scheme equivalent to a simultaneous autoregressive model (SAR)? Third, does the standardization of results for Table I really make the different data sets comparable? In other words, what is the source of and structure of error being evaluated? Is the aggregate fit of the model over all six data sets achieved with covariance analysis? Finally, while Ripley (1988) argues for a CAR model, Upton ends up arguing for an SAR model; why do these two scholars differ on this point?

In the next section Upton's preoccupation with computational constraints fails to take into account their considerable relaxation by supercomputing capabilities. A major portion of this section seems like an attempt to re–invent the wheel, for Upton actually seems to be talking about continuous cartograms (see Tobler, 1963; Monmonier, 1982). From a numerical point of view, the use of "sea constituencies" is bothersome, mostly because these imaginary areal units comprise such a large proportion of the final geo–referenced data set (763 imaginary versus 631 actual, in one case, and 523 imaginary versus 501 actual in an-

other case). Griffith (1982) has discussed the problem of shape and boundary buffer zones, in terms of percentages of internal and border areal units. Perhaps one needs to determine some threshold value beyond which an approach like Upton's becomes too artificial. Certainly the impact of having such a disproportionate number of manufactured areal units on the final statistical properties of an analysis merits very close scrutiny.

The final section of Upton's paper deals with the general problem of modifiable areal units and statistical inference complications introduced by aggregation. Again the Flowerdew and Green (1988, 1989) treatment of this topic is of relevance. Arbia's (1989) recent book addresses this issue, too, in a very imaginative and systematic way. Another important but often overlooked theme where the ecological fallacy emerges is in location–allocation modelling, where an areal unit centroid is used as the geo–referenced coordinate for all items located within that unit (introducing severe measurement error in many cases). Upton discusses the tetrachoric correlation coefficient, which is merely the square root of a chi square statistic that has been divided by the sample size. But this constrained estimation approach appears elsewhere in the spatial analysis literature, and so a comparison is in order. Certainly the volume of work that has been published on entropy maximizing models is relevant here, as it follows this same strategy. Upton's reflection on the question of "What is an appropriate areal unit?" is reminiscent of the familiar geographic question of "What is a region?"; perhaps some insights can be gained by reviewing the classical literature on this topic.

All in all, Upton's paper is a treatment of three interesting issues, with continual reference to selected empirical examples. His conceptual reflections upon these empirical examples is the major strength of this paper. His lack of acknowledgement of considerable relevant geographic literature supports the contention that a fuller dialogue is needed between quantitative geographers and professional statisticians. This, indeed, is the gap that this very book seeks to help fill!

References

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

Cliff, A., and J. Ord. (1981) *Spatial Processes*. London: Pion.

Flowerdew, R., and M. Green. (1988) "Statistical methods for inference between incompatible zonal systems," paper presented at Initiative #1, NCGIA, UC/Santa Barbara, December 12-16.

Flowerdew, R., and M. Green. (1989) "Inference between incompatible zonal systems using the EM algorithm," paper presented to the Sixth European Colloquium on Theoretical and Quantitative Geography, Chantilly, France, September 5-9.

Griffith, D. (1982) Geometry and spatial interaction, *Annals* of the Association of American Geographers, **72**, 332-346.

Griffith, D. (1988) "Interpretation of standard influential observations regression diagnostics in the presence of spatial dependence," paper presented to the 35th Regional Science Association, Toronto, November 11-13.

Griffith, D. (1989) Pure error and lack–of–fit regression diagnostics in the presence of spatial dependence. *Sistemi Urbani*, No. 2, in press.

Haining, R., D. Griffith, and R. Bennett. (1984) A statistical approach to the problem of missing spatial data using a first–order Markov model. *The Professional Geographer*, **36**, 338-345.

Haining, R., D. Griffith, and R. Bennett. (1989) Maximum likelihood estimation with missing spatial data and with an application to remotely sensed data. *Communications in Statistics: Theory and Methods*, **18**, 1875-1894.

Martin, R. (1984) Exact maximum likelihood for incomplete data from a correlated Gaussian process. *Communications in Statistics: Theory and Methods*, **13**, 1275-1288.

Martin, R. (forthcoming). Information loss due to incomplete data from a spatial Gaussian one–parameter first–order conditional process. *Communications in Statistics: Theory and Methods*.

Monmonier, M. (1982) *Computer Assisted Cartography*. Englewood Cliffs, N. J.: Prentice Hall, pp. 123-134.

Ripley, B. (1988) *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.

Tobler, W. (1963) Geographic area and map projections. *Geographical Review*, **53**, 59-78.

Unwin, D., and N. Wrigley. (1987) Control point distribution in trend surface modelling revisited: an application of the concept of leverage. *Transactions of the Institute of British Geographers*, N. S. 12: 147- 160.

Daniel A. Griffith, Syracuse University

A REJOINDER TO GRIFFITH'S DISCUSSION

by Graham J. G. Upton

This response is addressed to those who, like myself, decide whether a paper is worth reading on the basis of the ensuing discussion. To such readers I would comment that much of Griffith's discussion appears to be on a different paper to that which I thought I had written! In particular, I would stress that my interpolation procedures are not based on boundary lengths for precisely the reasons that Griffith indicates. Further, I do not advocate the fitting of trend surfaces—quite the reverse.

In commenting on the second section of my paper, Griffith correctly alludes to the difficulties posed by the presence of artificial "sea constituencies." However, the advent of supercomputers noted by Griffith means that this problem is conceptual, not computational. Indeed, if these artificial constituencies are excluded, then a far more compact representation will result—though a rectangular Britain would not be aesthetically pleasing. Whilst cartograms are not new, I am not persuaded that there are any existing computer–based procedures that lead to totally acceptable output.

Next, a more general observation, motivated by Griffith's discussion of CAR and SAR models. I wholeheartedly agree that the present dialogue between geographers and statisticians should be promoted. However, there can be few people as well qualified as Griffith to bridge the geography–statistics interface, and I worry that the current advances in "geometrics" may not be meaningful to the less quantitative geographer. "First neighbours" is an easy concept to grasp, whereas I cannot say the same for the "conditional autoregressive model"!

Finally (despite his misguided comments on my paper), a belated "thank you" to Dan Griffith for his generous hospitality!