Granger C W J, 1969 "Spatial data and time series analysis" in *London Papers in Regional Science 1. Studies in Regional Science* Ed. A J Scott (Pion, London) pp 1-24
Granger C W J, 1975 "Aspects of the analysis and interpretation of temporal and spatial data" *The Statistician* 23 197-210
Haggett P, 1975 *Geography: A Modern Synthesis* (Harper and Row, New York)
Haining R P, 1976 "The moving average model of dependence for a rectangular plane lattice" abstract in *Advances in Applied Probability* 8 654-655
Haining R P, 1977a "Model specification in stationary random fields" *Geographical Analysis* 9 107-129
Haining R P, 1977b "Markov processes and slope series: the scale problem, a comment" *Geographical Analysis* 9 94-99
Hodder I, Orton C R, 1976 *Spatial Analysis in Archaeology* (Cambridge University Press, London)
Hodge D, Gatrell A C, 1976 "Spatial constraint and the location of urban public facilities" *Environment and Planning A* 8 215-230
Hudson J C, 1969 "A location theory for rural settlement" *Annals of the Association of American Geographers* 59 365-381
Huijbregts C J, 1975 "Regionalised variables and quantitative analysis of spatial data" in *Display and Analysis of Spatial Data* Eds J C Davis, M J McCullagh (John Wiley, New York) pp 38-53
Kendall M G, 1976 *Time Series* (Charles Griffin, London)
Kolars J, Nystuen J, 1974 *Human Geography: Spatial Design in World Society* (McGraw-Hill, New York)
Lébart L, 1969 "Analyse statistique de la contiguité" *Publications de l'Institut de Statistique de l'Universite de Paris* 18 81-112
Martin R L, 1974 "On autocorrelation, bias and the use of first spatial differences in regression analysis" *Area* 6 185-194
Matheron G, 1971 *The Theory of Regionalised Variables* (Cahiers du Centre de Morphologie Mathematique de Fontainebleau No.5, Ecole Nationale Superieure des Mines de Paris, Paris)
Naidu P, 1970 "Statistical structure of aeromagnetic field" *Geophysics* 35 279-292
Nelson C R, 1973 *Applied Time Series Analysis for Managerial Forecasting* (Holden-Day, San Francisco)
Nordbeck S, Rystedt B, 1972 *Computer Cartography* (Studentlitteratur, Lund)
Rayner J N, Golledge R G, 1972 "Spectral analysis of settlement patterns in diverse physical and economic environments" *Environment and Planning A* 4 347-371
Sayn-Wittgenstein L, 1970 "Patterns of spatial variation in forests and other natural populations" *Pattern Recognition* 2 245-253
Sibert J, 1975 *Spatial Autocorrelation and the Optimal Prediction of Assessed Values* (Michigan Geographical Publications 14, University of Michigan, Ann Arbor, Mich.)
Slutsky E, 1937 "The summation of random causes as the source of cyclic processes" *Econometrica* 5 105-146
Smith T R, 1976 "Set-determined process and the growth of spatial structure" *Geographical Analysis* 8 354-375
Tobler W, 1969 "Geographical filters and their inverses" *Geographical Analysis* 1 234-253
Tobler W, 1975 "Linear operators applied to areal data" in *Display and Analysis of Spatial Data* Eds J C Davis, M J McCullagh (John Wiley, New York) pp 14-37

5

A million or so correlation coefficients: three experiments on the modifiable areal unit problem

S Openshaw, P J Taylor

5.1 Introduction

Although geography and statistics shared a common intellectual heritage of local societies in the nineteenth century (Berry and Marble, 1968), they have diverged somewhat during the twentieth century (Taylor and Goddard, 1974). This divergence has not been reversed as much as might have been expected by the recent rise of a quantitative geography with its emphasis upon statistical analysis. The reason seems to be that the discipline of statistics have moved away from the original interest in the study of published data sources—archival data—so that the modern statistician typically researches in the theoretical mathematics of probabilities with carefully controlled experimental data. Hence where empirical data is employed, it tends to be specially collected and avoids many of the pitfalls that confront the user of 'dirty' archival data. In contrast the majority of geographical studies use archival sources for their data (Haggett, 1965, chapter 7).

The fact that geographical data is typically far removed either from the ideals of classical statistical inference or the assumptions of stochastic modelling has been often noted (Cliff and Ord, 1975). This has led to particular geographical interest in nonparametric statistics (French, 1971) and problems of spatial dependence (Cliff and Ord, 1973). However, one topic which has been relatively neglected has been the arbitrary nature of areal units. Since the area over which data is collected is continuous, it follows that there will be numerous alternative ways in which it can be partitioned to form areal units for reporting the data. There will, in theory, be an infinite number of ways in which a study area can be areally divided, although data will normally be presented for only one particular set of areal units. These units themselves, however, may be combined to form new larger units at a new scale in a large number of alternative ways. This is the modifiable areal unit problem, identified by Yule and Kendall (1950) who point out that "we must emphasise the necessity, in this type of work, of not losing sight of the fact that our results depend on our units" (page 313). Despite this warning there has been relatively little concern for this fundamental problem in spatial analysis. This chapter follows several previous considerations of the problem in geography (Robinson, 1956; Thomas and Anderson, 1965; Curry, 1966, Clark and Avery, 1976) but provides a far more comprehensive empirical study of the topic than has hitherto been attempted.

5.2 Modifiable areal unit problems

The modifiable areal unit problem is in reality two separate but interrelated problems. Statisticians (Gehlke and Biehl, 1934; Neprash, 1934; Yule and Kendall, 1950) have generally discussed what we shall term the *scale problem*. This is defined simply as the variation in results that may be obtained when the same areal data are combined into sets of increasingly larger areal units of analysis. For instance, analysis of the same census data at scales ranging from enumeration districts, wards, local authorities, up to standard regions will almost certainly provide alternative results and possibly interpretations. It should be noted at this point that while this problem may be closely related to making inferences about individuals from aggregate data—in economics the micro/macroanalysis problem (Cramer, 1964) and in sociology the ecological fallacy problem (Robinson, 1950)—in this paper we shall remain interested only in combinations of the data that have already been spatially aggregated. Notice that this is where this study differs in purpose from the recent researches of Williams (1976; 1977a; 1977b).

Although scale differences are the most obvious manifestation of the modifiable areal unit problem there is also the problem of alternative combinations of base units at equal or similar scales. For instance Tooze (1976) and Kirby and Taylor (1976) have both divided up Britain at almost the same scale but have created very different units for analysis. Any variations in results due to alternative units of analysis where n, the number of units, is constant will be termed the *aggregation problem*.

In addition we can distinguish between two different forms of areal arrangement. In the regional taxonomy literature, contiguous areal arrangements are termed regions and noncontiguous areal arrangements are referred to as regional types (Spence and Taylor, 1970). Here we shall use the simpler terminology of *zoning system* for contiguous arrangement and *grouping system* for the noncontiguous case. Alternatively a zoning system may be considered a special case of a grouping system which incorporates a contiguity constraint in its formation. Since most spatial analyses use zoning systems we concentrate on them in what follows although simple groupings are considered for comparative purposes.

Finally Yule and Kendall's (1950) discussion of the problem was originally illustrated by using correlation coefficients and other researchers have followed suit. We continue this tradition because this statistic is easily interpretable and hence both scale and aggregation effects may be simply identified and described. This paper reports the results from three closely related experiments on variations in the correlation coefficient under different spatial and statistical conditions.

5.3 The first experiment: areal associations in Iowa

The first experiments on the correlation coefficient are carried out with the use of a set of data describing Iowa, USA. The data has been chosen partly because of its availability and manageability. The basic areal units are the ninety-nine counties of Iowa. For each unit we have two measures—the dependent variable is the percentage vote for Republican candidates in the congressional election of 1968 and the independent variable is the percentage of population over sixty-years old recorded in the 1970 US census[1]. Notice that we have no need to imply any individual-level correlations here. The percentage old people variable is interpreted as an index of demographic history, largely differential in- and out-migration, reflecting an economic environment which we hypothesise to be positively related to Republican voting. In fact the variables are correlated at +0·3466 over the set of ninety-nine counties. These data are regarded as our population of base units.

From Yule and Kendall's (1950) discussion we know that the correlation ($r = +0{\cdot}3466$) is specific to the particular set of areal units used. In table 5.1 alternative areal arrangements have been employed to compute the correlation to produce five sets of results for the same relationship. In each example the ninety-nine counties have been combined into just six new areal units[2]. Hence differences between these correlations and the original 'county' correlation reflect the scale problem, whereas differences in correlations between the five alternative aggregations illustrate the aggregation problem. The only 'official' aggregation, the congressional districts, produces a correlation below the original r value but the other correlations are, as is generally expected, higher than the base unit value.

Table 5.1. Some effects on the correlation coefficient of different areal arrangements of the Iowa counties into six zones.

Alternative combinations of counties	r
6 Republican-proposed congressional districts	0·4823
6 Democrat-proposed congressional districts	0·6274
6 Congressional districts	0·2651
6 Urban/rural regional types	0·8624
6 Functional regions	0·7128
99 Iowa counties	0·3466

[1] Problems caused by the fact that these measures define closed number sets are not critical for our subsequent analyses since no observations approach the limits of 0% and 100%.

[2] It should be noted that the variables were each aggregated by summing the county percentages and dividing by the number of counties in the aggregation. This does not produce the 'true' percentage values for the new aggregated units since counties vary in population size. This approach was adopted to facilitate computation in the subsequent experiments and will have no systematic effect on the experimental results.

If the researcher is interested in finding high correlations between his variables he will lament the final choice of congressional districts and will wish the Democrat proposal had won the political battle. Still higher correlations can be produced, however, by using specially designed 'geographical' arrangements—either urban/rural types based upon the largest urban area in a county or of which the county is part, or functional regions based upon the counties being allocated to the six largest urban zones by distance. Clearly the correlation between Republican voting and percentage old people is not an easy thing to measure.

5.3.1 Identification of the limits of the scale and aggregation problem

The dimensions of the modifiable areal unit problem for these Iowa data can be determined by finding the limits of the aggregation effects at different scales by applying an automatic zoning algorithm. This method has been developed to identify zonings or groupings of data that approximately optimise any general function defined in terms of the aggregated data (Openshaw, 1977a; 1977b; 1977c). Thus we can identify zoning or grouping systems of Iowa counties which approximately represent the limits of negative and positive correlation. These have been derived for various scales and are presented in table 5.2.

At the aggregation scale both for six zones and six groups the limits show that it is possible to produce the whole range of correlations. As the aggregation scale decreases the range of possible correlations declines especially for zoning systems where contiguity constraints are involved. We can conclude, however, that for all scales there is a relatively wide range of correlations which would certainly involve alternative substantive interpretations. Table 5.2, therefore, specifies the universe of alternatives

Table 5.2. Maximum and minimum values of the correlation coefficient.

Number of zones or groups	Zoning systems [a]		Grouping without contiguity	
	minimum r	maximum r	minimum r	maximum r
6	−0·999	0·999	−0·999	0·999
12	−0·984	0·999	−0·999	0·999
18	−0·936	0·996	−0·977	0·999
24	−0·811	0·979	−0·994	0·999
30	−0·770	0·968	−0·989	0·999
36	−0·745	0·949	−0·987	0·998
42	−0·613	0·891	−0·980	0·996
48	−0·548	0·886	−0·967	0·995
54	−0·405	0·823	−0·892	0·983
60	−0·379	0·777	−0·787	0·983
66	−0·180	0·709	−0·698	0·953
72	−0·059	0·703	−0·579	0·927

[a] Best from fifteen different random zoning systems used as starting points for the automatic zoning algorithm.

within which our experiments on the correlation coefficient are to be carried out.

5.3.2 Zoning and grouping distributions of correlation coefficients

The maximum and minimum correlation coefficients illustrated in table 5.2 may be interpreted as representing the known limits of the distribution of correlation coefficients for alternative aggregations at each scale. To produce the intervening distributions it is necessary to use a random zoning and grouping system generator (Openshaw, 1977d). By using this algorithm, random samples of ten-thousand alternative arrangements into six, twelve, eighteen, twenty-four, thirty, thirty-six, forty-two, forty-eight, fifty-four, sixty, sixty-six, and seventy-two zoning and grouping systems have been produced. Figure 5.1 shows the frequency distributions of correlation coefficients for each sample at class intervals of 0·1. In figure 5.1(a) zoning system correlations are shown and we may term these distributions the *zoning distributions* of correlation coefficients for aggregates of Iowa counties at different scales. Figure 5.1(b) shows the equivalent *grouping distributions* of correlation coefficients, which are sometimes considered as analogous to the sampling distribution of correlation coefficients in standard statistical theory (Williams, 1977a).

Interpretation of figure 5.1 is relatively straightforward. Both sets of distributions show a decrease in spread as the scale decreases (that is, more zones or groups) as would be expected. Furthermore the degree of bias in the estimators compared with the county-level correlation also declines with scale. In terms of our previous discussion this can be interpreted as the aggregation effect being illustrated by the horizontal spread and the scale effect, tending to increase the correlation, which is summarised by the vertical changes in the modes of the distributions. Differences between the two sets of distributions are more interesting. The zoning distributions visually exhibit less spread and less bias than the grouping distributions. The most obvious explanation for this phenomenon would seem to be the interaction between the contiguity constraint of the zoning systems and the positive spatial autocorrelation of the variables. In fact both variables are positively autocorrelated, on using Moran's I statistic $I_x = +0{\cdot}37$ and

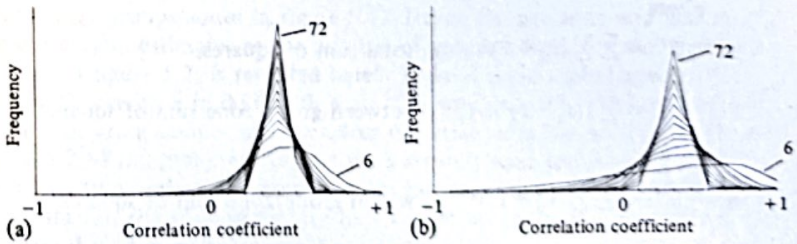

Figure 5.1. Frequencies of correlation coefficients at different scales. (a) Zoning distributions, (b) grouping distributions.

$I_y = +0{\cdot}43$ (Cliff and Ord, 1973). This hypothesis is formally tested in the next section and in this discussion we shall briefly consider other effects of this interaction between pattern and boundary.

The above visual interpretation is confirmed as far as the precision of the correlation estimates are concerned in table 5.3. This shows the spread of four zoning and four grouping distributions as measured by their standard deviations. The other feature that this table shows is that as the number of zones/groups increases, precision also increases. This brings us back to the analogy with sampling theory referred to above. This analogy is explicitly illustrated in table 5.3 by inclusion of the standard deviations for equivalent sampling distributions of the correlation coefficient. It should be noted that the analogy holds for grouping distributions but is highly unsatisfactory for zoning distributions.

Table 5.3. Precision of alternative correlation estimates for Iowa.

Number of zones	Standard deviation [a]	Number of groups	Standard deviation [a]	Size of sample [b]	Standard deviation [a]
6	0·218	6	0·408	6	0·429
12	0·161	12	0·273	12	0·273
24	0·122	24	0·189	24	0·172
72	0·051	72	0·066	72	0·057

[a] All standard deviations are based upon 10000 correlation coefficients derived from zones, groups, or samples.

[b] Samples generated by randomly selecting samples of Iowa counties without replacement, 10000 times.

5.3.3 Correlation coefficients and loss of variation

A random zoning system placed upon a positively autocorrelated variable will tend to have the effect of locating similar base units together. This will have a direct effect upon the inevitable loss in variation in the variable as a result of the aggregation of the base units. The loss can be described in terms of the sums-of-squares equality

$$S^a = S^B + S^W , \tag{5.1}$$

where

$$S^a = \sum_i \sum_j (x_{ij} - \bar{x})^2 , \quad \text{total sum of squares,}$$

$$S^B = \sum_j [(\bar{x}_j - \bar{x})^2 n_j] , \quad \text{between group/zone sum of squares,}$$

and

$$S^W = \sum_i \sum_j (x_{ij} - \bar{x}_j)^2 , \quad \text{within group/zone sum of squares,}$$

where x_{ij} is the value for the ith unit in the jth group/zone which has n_j base units. Since a sum-of-squares expression cannot be less than zero it follows that $S^B \leqslant S^a$. For any particular group/zone the degree of loss of variation in a variable can therefore be measured by

$$\Delta S^B = \frac{S^a - S^B}{S^a} . \tag{5.2}$$

At the ninety-nine-zone level each county is its own zone so that $S^a = S^B$ and $\Delta S^B = 0$. When the counties are aggregated into groups/zones, variation is lost until we are left with one zone (Iowa) when $S^B = 0$ so that $\Delta S^B = 1$. ΔS^B is therefore a simple measure of loss of variation ranging from zero, or no loss, to unity, or total loss of variation. [Although equation (5.2) actually defines the proportion of within group/zone variation, the interpretation in terms of loss of between group/zone variation is more useful for the present discussion.]

ΔS^B can be computed for either variable, of course, although theoretical work in related areas suggests that loss of variation in the independent variable is the more interesting. Cramer (1964), for instance, has shown that the most efficient grouping of individuals in a linear regression equation for aggregate data is where the loss of variation in the independent variable is minimised. He thus employs income classes based upon individual income returns.

Since positive autocorrelation may be regarded as the 'norm' in most spatial analyses, it follows that aggregation by zoning will have a similar effect on the independent variable as 'efficient grouping' has in econometrics. This would account for the more precise correlation estimates given by the zoning systems as indicated by the smaller spread in figure 5.1(a). A similar result has been found by Williams (1976) in a more closely related study. He contrasted random and homogeneously grouped data and found that the latter provided much more precise estimators in regression and correlation analyses.

Given the above studies, we have chosen to characterise all of our zoning and grouping systems in terms of ΔS^B for the independent variable. In figure 5.2 one-thousand correlation coefficients are arranged against ΔS^B for zoning and grouping systems at each of the aggregations into six, twelve, and twenty-four units. Notice, first of all, that the vertical spread of observations is the same as the horizontal spread for six, twelve, and twenty-four groups/zones in figure 5.1. Hence the precision and bias in the correlation estimates as the number of groups/zones increases from left to right in figure 5.2, is repeated here. These diagrams also show a systematic variation in ΔS^B with scale. As would be expected, as the number of groups/zones increases, loss of variation is less noticeable. In figure 5.2(b) the grouping systems for a six-unit scale lost nearly all of their variation, only a few arrangements lose less than 90% ($\Delta S^B = 0{\cdot}9$). In contrast at the scale of twenty-four almost all grouping systems have lost less than 90% of the original sums of squares and typically about three-quarters of the variation is lost ($\Delta S^B = 0{\cdot}75$).

Similar patterns are found in the zoning systems on figure 5.2(a). The contrasts with figure 5.2(b) are interesting, however. Brought over from figure 5.1 is the smaller spread and bias on the vertical (correlation) scale. Differences along the ΔS^B scale are even more clear-cut, however. As our previous discussion has implied, there is far less loss of variation in the zoning systems at each of the three scales. In fact the level of ΔS^B for six zones is approximately equivalent to ΔS^B for twenty-four groups. Only about half of the original variation is lost when zoning systems of twenty-four are produced. Figure 5.2 clearly shows that choice of a zoning system has a very profound effect on the type of aggregation produced when the data are autocorrelated. Very similar results have been produced for arrays of r against ΔS^B for the dependent variable.

Before leaving figure 5.2 it is worth noting where the five aggregations of six from table 5.1 occur among these randomly generated systems. The four zoning systems occur within the distribution of random zoning systems but this is not true of the one grouping system in table 5.1, the urban/rural types. This grouping preserves much more variation than any random grouping and is similar in fact to the zoning systems. The particular criteria for the grouping seem to be equivalent to zoning constraints in terms of ΔS^B. In contrast the criteria employed for the four zoning systems seem to have had no additional effect beyond contiguity constrained random zoning.

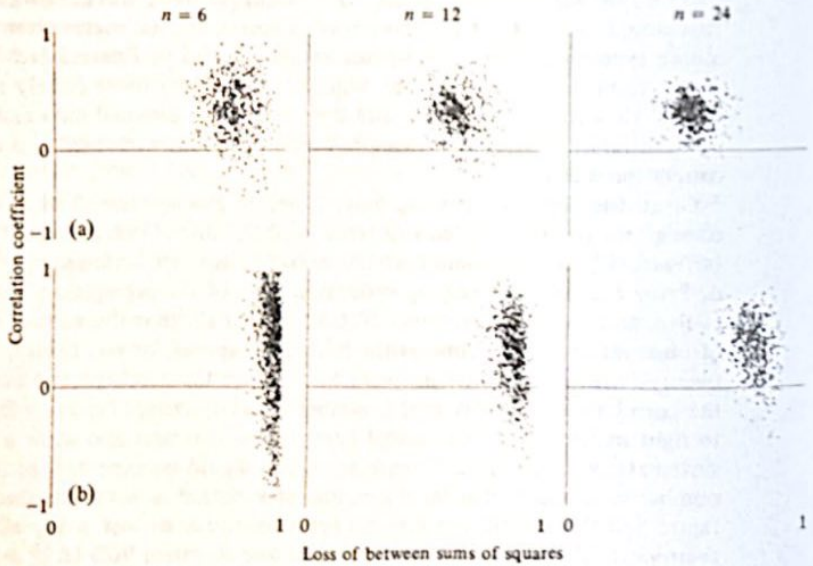

Figure 5.2. Graph plots of one-thousand correlation coefficients against loss of between sums of squares for the independent variable. (a) Zoning systems, (b) grouping systems.

5.3.4 Correlation coefficients and differential loss of variation

One feature of figure 5.2 is that there is no apparent relationship between r and ΔS^B. Whatever degree of loss of variation that occurs, a wide range of correlations can result. There is nothing in the literature to suggest otherwise. Blalock (1964), however, does propose that differential loss of variation between independent and dependent variables will have a systematic effect on the correlation coefficient. His argument is simply that if the dependent variable, y, is caused by several independent variables, then if the variation of one independent variable, x, is particularly maintained in an aggregation, this will enable this particular variable to account for more of the remaining variation in y, that is, increase the correlation. This is because, of the total variation of y that is lost in the aggregation, less of that part of the variation of y caused by x will be lost than that part caused by the other independent variables whose own variation is not being maintained. This proposed effect may be measured by

$$\Delta S^D = \Delta S^B_x - \Delta S^B_y , \qquad (5.3)$$

where the subscripts x and y refer to independent and dependent variable sums of squares respectively. ΔS^D may be termed the differential loss of the sums-of-squares term. This measure ranges from -1 when there has been no loss in the variation of x and total loss of the variation of y to $+1$ when total loss of variation of x coincides with no loss in variation of y. These two limiting cases afford a theoretical addition to Blalock's hypothesis despite the fact that the correlation coefficient itself is indeterminate in each case owing to the zero variation in one of the variables. If we briefly revert to a regression format, however, we can see that the case where $\Delta S^D = -1$ corresponds to a horizontal line and $\Delta S^D = +1$ to a vertical line. Maximum loss in the sum of squares of y with respect to the sum of squares of x ($\Delta S^D = -1$) corresponds to perfect prediction of y from x, whereas maximum loss of the sum of squares of x with respect to the sum of squares of y ($\Delta S^D = +1$) corresponds to no prediction of y from x and variations in y depend solely on other variables. These are limiting cases, however. In general we expect that as ΔS^D gets larger, r will decline since variables other than x are maintaining that part of the variability of y not dependent on x at the expense of that part dependent on x.

Figure 5.3 shows scatters of one-thousand zoning and grouping systems at scales of six, twelve, and twenty-four units with r plotted against ΔS^D. The two sequences of scatters are very different from one another with respect to the changes in the spread of ΔS^D. With aggregations into just six units the spread of ΔS^D is far greater for zoning systems than for simple random groupings. However, the difference is eliminated by the twenty-four-unit scale. In effect we find two different scale effects—in zoning systems the spread of ΔS^D declines with scale whereas with grouping systems the range increases. The explanation for this phenomenon is not readily apparent. Presumably the particular scale of six units

interacts with the particular spatial autocorrelation pattern in these variables to counteract the expected decline in ΔS^D as illustrated in the grouping systems and is even strong enough to slightly increase the range of ΔS^D. We present an informal test of this idea in the next section.

A second difference between the zoning and grouping systems in figure 5.3 concerns the modal values of ΔS^D. In the grouping scatters, the mode is at zero, which is expected, that is, random grouping is as likely to cause loss of variation in x as in y. In the zoning systems, however, the mode of ΔS^D is positive for all three data plots. This suggests that zones tend to cause more loss of variation in x than y. The explanation of this may be presumed to lie in the interaction of the contiguity constraint and the particular autocorrelation patterns of the two variables. Once again we informally test this assertion in the next section.

Having commented on the nature of ΔS^D in the data plots in figure 5.3 we can now turn to the hypothesised relationship between it and the correlations. A cursory glance at the six diagrams indicates no tendency towards $r = 0$ for increasing ΔS^D as we had previously suggested.

Finally before we leave these plots we can note once again where the five six-unit aggregations from table 5.1 fit onto the diagrams. In this case all five grouping/zoning systems lie within the scatter of aggregations. Although they all lie away from the ΔS^D modes on the two diagrams, there is no particular pattern to this divergence.

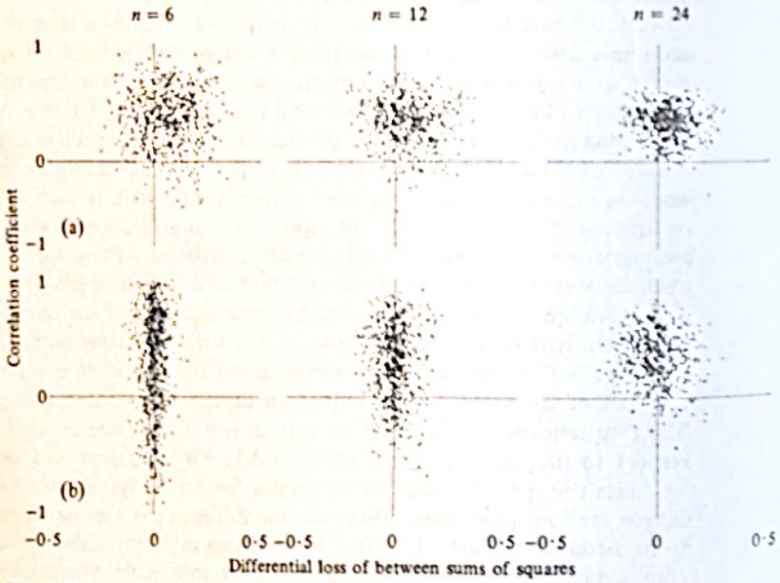

Figure 5.3. Graph plots of one-thousand correlation coefficients against differential loss of between sums of squares. (a) Zoning systems, (b) grouping systems.

We can summarise these experiments on the Iowa data in two general statements.

(1) There seem to be very distinct differences between zoning and grouping systems in many situations and these seem to be caused by the interaction of the contiguity in the zoning with the spatial autocorrelation in the data.

(2) There seem to be no systematic relationships between the sums-of-squares terms and the correlation coefficients.

We have designed two further experiments to explore both statements more fully.

5.4 The second experiment: spatial autocorrelation

We investigate the effects of spatial autocorrelation on the correlation coefficient and sums-of-squares terms by employing specially constructed artificial data. The experiments carried out upon the Iowa data are then repeated on the artificial data.

5.4.1 A data generator

We require to produce new variables, the properties of which we can carefully control, for our set of ninety-nine Iowa counties. A suitable data generator can be based upon the quadratic loss function $F(x^*, y^*)$, which tends to zero as data with the required properties are produced. Thus our objective function to be minimised is

$$\begin{aligned} F(x^*, y^*) = {} & w_1 (r^*_{yx} - r_{yx})^2 \\ & + w_2 (S^*_x - S_x)^2 + w_3 (S^*_y - S_y)^2 + w_4 (K^*_x - K_x)^2 \\ & + w_5 (K^*_y - K_y)^2 + w_6 (I^*_x - I_x)^2 + w_7 (I^*_y - I_y)^2 , \end{aligned} \qquad (5.4)$$

where x^* and y^* are 198 undetermined parameters (two for each county) which are estimated to minimise F; r^*_{yx}, S^*_x, S^*_y, K^*_x, K^*_y, I^*_x, and I^*_y are values of correlation, skewness, kurtosis, and spatial autocorrelation for any given vectors x^* and y^*; r_{yx}, S_x, S_y, K_x, K_y, I_x, and I_y are desired values which x^* and y^* should possess; and $w_1, \ldots, w_7$ are a set of weights that indicate the relative importance associated with each characteristic of the data. The spatial autocorrelation measure used is Moran's I statistic based on first-order contiguity relations (Cliff and Ord, 1973).

The function $F(x^*, y^*)$ is minimised by a quasi-Newtonian optimisation procedure available as a Harwell subroutine VA10AD (Fletcher, 1972). This data generator proved remarkably successful albeit expensive in computer time with a typical run requiring 920 seconds of central processing unit time on an IBM 370/168.

Two sets of data were generated for our experiment. In both sets our target r_{yx} was +0·3466 for comparability with the Iowa data. Furthermore we defined normal data so that $S_x = S_y = K_x = K_y = 0$ were targets. The data sets were designed to differ only in terms of spatial autocorrelation.

In data set A the targets were $I_x = I_y = 0$ to produce no spatial autocorrelation and in data set B the targets were $I_x = I_y = 1$ to produce maximum positive spatial autocorrelation. The targets are summarised in table 5.4 together with the properties of the generated data.

Table 5.4. Properties of generated data.

Data set	Targets				Actual						
	r	S	K	I	r	S_x^*	S_y^*	K_x^*	K_y^*	I_x^*	I_y^*
A	0·34	0	0	0	0·34	−0·00	0·00	−0·00	0·00	0·00	0·00
B	0·34	0	0	1	0·34	0·01	0·00	−0·00	−0·00	0·82	0·92

5.4.2 Zoning and grouping distributions of correlation coefficients

The zoning and grouping distributions for both artificial data sets are shown in figure 5.4. If we consider data set A first, we can see that with no autocorrelation there are no major differences between the zoning and grouping distributions. In contrast, for data set B, differences between the two sets of distributions are very clearly illustrated. Data set B replicates our findings from the Iowa data with the zoning distributions showing less spread and less bias than the grouping distributions. We may conclude that spatial autocorrelation and zoning does interact in the way suggested for the Iowa data.

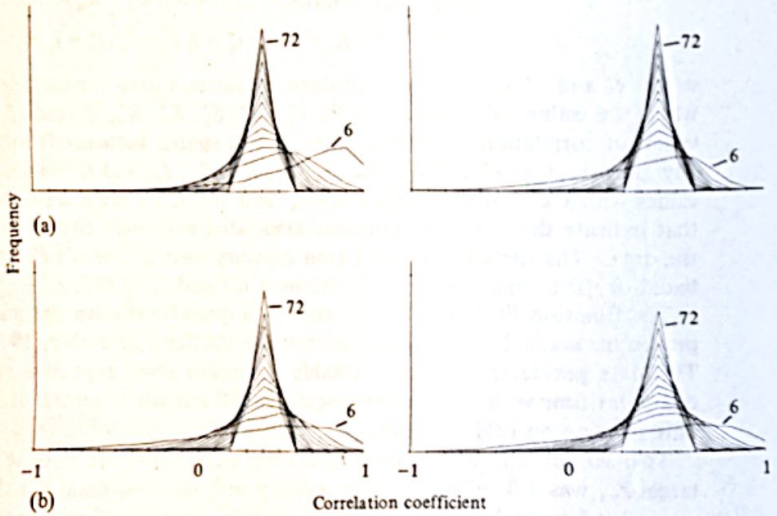

Figure 5.4. Frequencies of correlation coefficients at different scales for the artificial data. (a) Zoning distributions, (b) grouping distributions.

5.4.3 Correlation coefficients and the sums-of-squares terms

Let us turn to the sums-of-squares analysis. With one exception, the scale effects noted in the Iowa data are repeated for both sets of artificial data and so we shall just consider one scale here, that of twelve zones/groups. The exception will be briefly commented on below.

The scatter of results for the correlation coefficient against ΔS^B is shown in figure 5.5. Clearly the distinctive feature is the reduced loss of sums of squares for the zoning distributions in data set B. Hence this data set has again replicated the result found for the Iowa data and we therefore confirm the hypothesis of the interaction of contiguity constraint and autocorrelation, with reduction in loss of variation in a variable with aggregation.

The four scatter plots of correlation coefficients against ΔS^D are shown in figure 5.6. Once again it is the zoning system with data set B which has the most distinctive pattern. Here, however, replication of the Iowa data is not exact. Although autocorrelation and zoning interact to produce a wide spread of ΔS^D values, the bias in the values is the opposite to that of the Iowa data. This is inconsequential because specification of x and y as dependent and independent variables is wholly arbitrary in our artificial data. These results do suggest that the particular bias in the Iowa data set is specific to that data set. Furthermore the peculiar scale effect for the

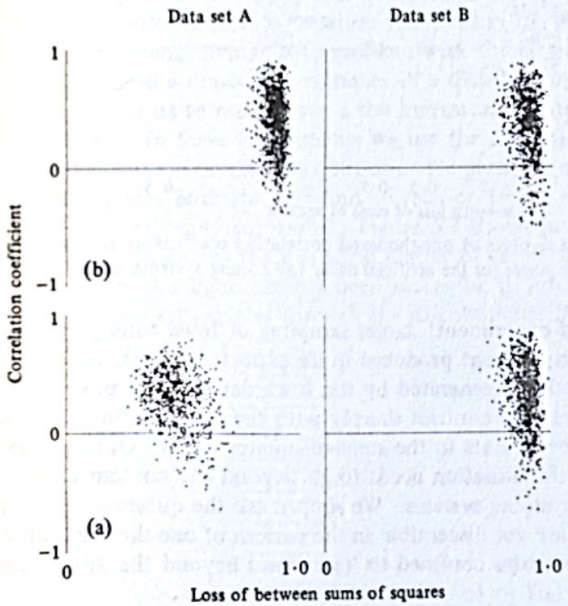

Figure 5.5. Graph plots of one-thousand correlation coefficients against loss of between sums of squares for the independent variable in the artificial data. (a) Zoning systems, (b) grouping systems.

zoning system with ΔS^D for Iowa that was found in figure 5.4 is not repeated for data set B, which further suggests an effect specific to the Iowa data.

The conclusions to be drawn from these second experiments are that zoning and spatial autocorrelation do interact in quite predictable ways and that this interaction explains much of the variety of results previously obtained from the original autocorrelated Iowa data.

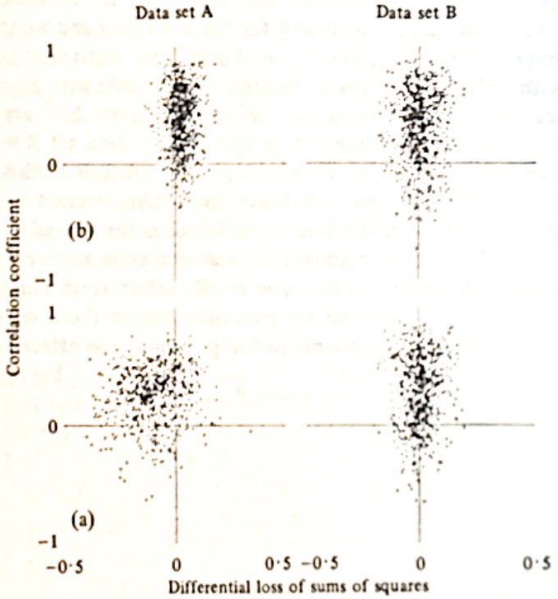

Figure 5.6. Graph plots of one-thousand correlation coefficients against differential loss of sums of squares for the artificial data. (a) Zoning systems, (b) grouping systems.

5.5 The third experiment: target sampling of Iowa zoning distributions

The second experiment produced quite expected results, which generally support hypotheses generated by the Iowa data. These positive findings from the Iowa data contrast sharply with the negative findings relating correlation coefficients to the sums-of-squares terms. Our strategy for investigating this situation needs to go beyond the random generation of zoning and grouping systems. We simply ask the question: if the proposed relationships are not discernible in the pattern of one-thousand observations, are the relationships confined to 'rare' cases beyond the distributions identified so far?

5.5.1 Target sampling

Target sampling involves defining a desired value of correlation or sums-of-squares term and producing a set of aggregations that have the desired value. For instance we may wish to know the pattern of correlation coefficients for zoning systems at a specific scale for which $\Delta S^D = +0{\cdot}5$ or $-0{\cdot}5$. Such aggregations can be produced by using the automatic zoning algorithm employed in section 5.3 to define maximum or minimum correlation coefficients. In this case we define an aggregation of x and y which minimises the quadratic loss function $Z(x^1, y^1)$

$$Z(x^1, y^1) = (\Delta S^D - \text{target})^2 . \tag{5.5}$$

For any one target we can often generate several different solutions by starting from different initial random zoning systems. This is possible both because of the large number of alternative solutions in terms of equation (5.5) and because of the suboptimal nature of the automatic zoning algorithm. In effect we are sampling vertical strips in the distribution represented by the data plots of one thousand previously described (figure 5.6). If in equation (5.5) we replace ΔS^D by r_{xy} we can sample horizontal strips in the same way.

5.5.2 Target sampling experiments

The purpose of this exercise is to obtain a more thorough understanding of the sum-of-squares/correlation relationship than can be obtained from random arrangements. The problem with the latter is that they fail to emphasise the limits and extremes of a distribution. Our target sampling will enable us to push beyond the known distribution patterns previously reported. In these experiments we use the original Iowa data and concentrate on zoning distributions. We generate one-hundred solutions of twelve zones both for ΔS^D and correlation targets, although we are not always completely successful. Figure 5.7 shows both sets of target samples and in both graphs there is some scatter of results around extreme targets. Elsewhere the algorithm has been successful in producing specific vertical and horizontal samples through the distribution. Both graphs should be

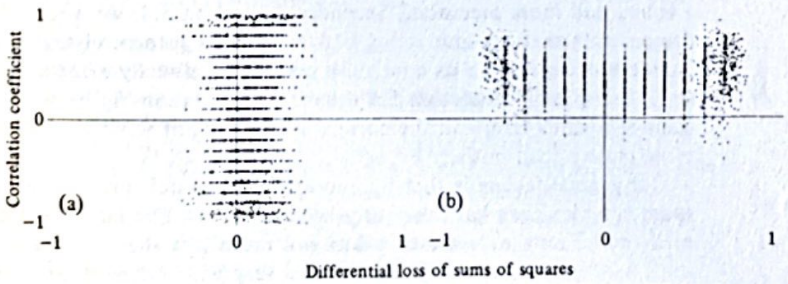

Figure 5.7. Target sampling of correlation coefficients against differential loss of sums of squares. (a) Correlation targets, (b) sums-of-squares targets.

compared with the random twelve-zone distribution for the Iowa data in figure 5.3(a). This comparison clearly illustrates how the target sampling provides information for a much wider range of r and ΔS^D.

If we consider the correlation target results first [figure 5.7(a)], we notice that although the whole range of possible r values have been generated, no systematic relationship emerges. Since we have previously argued that the size of correlation is likely to be dependent upon the sum-of-squares term then we can expect a higher probability of discerning a relationship when it is ΔS^D which is varied. This is achieved by the target sampling for ΔS^D illustrated in figure 5.7(b). In this graph there is again no systematic trend in the main central portion of the plot but the extreme ΔS^D samples do suggest some effects on the correlation coefficient. For high negative values of ΔS^D (that is, when much of the variation in x is preserved relative to y) at less than $-0{\cdot}4$, there is a slight tendency for a relatively large number of higher correlation coefficients to be produced. This is consistent with the previous arguments of section 5.3. At the other extreme, for high positive values of ΔS^D (that is, when most of the variation in x is lost relative to y) at greater than $+0{\cdot}4$, a relatively large quantity of high correlations are again produced suggesting a shallow parabola. In this case, however, the high correlations are accompanied by a set of weak negative correlations. This suggests that since high ΔS^D values eliminate most of the variation in x we are approaching the 'vertical' regression situation discussed in section 5.3, where the underlying relationship will be highly unstable.

5.6 Conclusions

The purpose of these experiments has been to increase our understanding of the modifiable areal unit problem both from statistical and geographical perspectives. Three particular sets of findings are likely to be of some general interest. In section 5.3 the relationship between zoning, grouping, and also sampling distributions was illustrated in terms of correlation coefficients. In particular it is clearly shown that the simple zoning/grouping/sampling analogy is misleading since zoning distributions exhibit less bias and more precision. Secondly, in section 5.4, we were able to demonstrate that the interaction between spatial autocorrelation and the zoning procedure with its contiguity constraints directly affected resulting statistics. Finally in section 5.5 the expected relationship between the sums-of-squares term and the correlation coefficient was found to be much more elusive than initially expected.

Our general feeling is that the modifiable areal unit problem is much more complex than has previously been believed. The fact that there seems to be no simple solution does not mean that the problem is any less important. We *have* been able to find a very wide range of correlations. We simply do not know why we have found them. Hence we can make no general statements about variations in correlation coefficients so that each areal unit problem must be treated individually for any specific piece of research.

In many ways the use of any particular zoning or grouping system may not be such a bad thing if it brings the problem of the geographical individual back into consideration in spatial analysis. Hannan (1971) has suggested that the substantive way to overcome scale problems is to develop cross-level theory. The aggregation problem of the geographer seems to be much simpler. No special theory linking scales is required; all that is needed is agreement upon what constitute the objects of the geographical enquiry. This question has long been a thorny one for geographers and seems to have been avoided more than it has been explicitly faced. *The question is simply what objects at what scales do we wish to investigate?* In contrast to our previous statistical perspective, this is an essentially geographical problem and answers to the question should form the basic geographical contribution to, and solution of, the modifiable areal unit problem. If all researchers in a field of geographical enquiry agree on their objects of interest, and these objects can be defined in a nonarbitrary manner, then this constitutes a unique set of units and the problem disappears. Of course the situation is not, and is never likely to be, that simple and is discussed further in Openshaw and Taylor (1978).

Furthermore, even if it is possible to agree on a unique set of areal units on geographical grounds, statistical variations, such as those illustrated in this paper, will continue to be of interest in order to place a set of results into a meaningful statistical and spatial perspective. For as Williams (1976, page 16) has so neatly put it "No self-respecting statistician would take just any selection of individuals as his sample in a study and give it no further thought. Likewise we would hope that the days are numbered for urban and regional scientists who produce zoning systems, as it were, out of a hat and proceed to use them, blissfully unaware of the effects the grouping might have on any subsequent empirical investigations they carry out".

References

Berry B J L, Marble D F, 1968 *Spatial Analysis* (Prentice Hall, Englewood Cliffs, NJ) pp 1-9

Blalock H M, 1964 *Causal Influences in Non Experimental Research* (University of North Carolina Press, Chapel Hill, NC)

Clark W A V, Avery K L, 1976 "The effects of data aggregation in statistical analysis" *Geographical Analysis* 8 428-438

Cliff A D, Ord J K, 1973 *Spatial Autocorrelation* (Pion, London)

Cliff A D, Ord J K, 1975 "Model building and the analysis of spatial patterns in human geography" *Journal of the Royal Statistical Society B* 37 297-348

Cramer J S, 1964 "Efficient grouping, regression and correlation in Engel curve analysis" *Journal of the American Statistical Association* 59 233-250

Curry L, 1966 "A note on spatial association" *Professional Geographer* 18 97-99

Fletcher R, 1972 "FORTRAN subroutines for minimization by quasiNewton methods" *AERE* R7125 (HMSO, London)