5. The Effect of Aggregation on Bivariate Statistics

5.1. Summary

The synthetic spatial dataset generator described in Chapter 3 was used to seek a relationship between the behaviour of aggregated bivariate statistics and the spatial autocorrelation of the variables. It is found that a degree of dependence is visible, especially when their Moran Coefficients (MCs) are the same or when the initial correlation is zero. When the two variables have different MCs, the use of spatial autocorrelation is insufficient to completely describe the behaviour of the statistics, especially that of the correlation and MC of regression residuals. Correlation coefficients from a synthetic spatial dataset built on the Iowa connectivity matrix behave in a similar manner to those derived from the data used in Openshaw and Taylor (1979), helping to confirm the utility of the synthetic data generator as a tool for analysis of the MAUP. A numerical measure of spatial pattern is recognized as a requirement for more precise measurement of the MAUP as it affects the more complex univariate, bivariate, and multivariate statistics.

5.2. Introduction

The dependence of bivariate statistics, primarily correlation, on spatial resolution is what initially drew researchers' attention to what would be called the Modifiable Area Unit Problem (MAUP) (for example, Gehlke and Biehl, 1934; Robinson, 1950). Studies using specific datasets have appeared sporadically in the literature since then (e.g. Clark and Avery, 1976), but the daunting computational requirements for even the most basic study meant that systematic studies have been unfeasible until recently with the increasing availability of cheap, fast computers. Furthermore, studying bivariate statistics is complicated because they depend on the behaviour of two variables that are aggregated independently.

Openshaw and Taylor's (1979) examination of the effects of spatial aggregation on correlation coefficients has been widely recognized as the inspiration of an increasing body of research (see the 1996 special issue of *Geographical Systems*). Reynolds and Amrhein (1998) and Chapter 3 point out that the use of specific datasets greatly restricts the ability of researchers to study the Modifiable Area Unit Problem because the various spatial and aspatial parameters of the variables cannot be altered at will. The synthetic spatial dataset generator and random aggregation model described in detail in Chapter 3 are employed here to extend the work of Reynolds and Amrhein

(1998) to the bivariate statistics of covariance, correlation, regression slope parameters, and the Moran Coefficient of the regression residuals (MC_{RR}). Results from the analyses will be compared to results from Openshaw and Taylor (1979). The third section describes the rationale and method behind the experiments, the fourth and fifth present the results of the first and second experiments, the sixth section discusses the results, and the seventh presents conclusions of the chapter.

5.3. Method

Reynolds and Amrhein (1998) clearly demonstrate that the relative change in variance, defined on page 23, is clearly affected by both spatial autocorrelation and arrangement of the unaggregated variable and the number of aggregate cells. A similar formula cannot be used to express the change in covariance, unfortunately, because the covariance can be zero. Similar to the variance, the unaggregated covariance can be written as the sum of the covariance *between* the aggregated cells and the sum of weighted covariances *within* each cell as follows:

$$\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}\right)\left(y_{ij}-\bar{y}\right)=\frac{1}{N}\sum_{i=1}^{M}n_i\left(x_{i\bullet}-\bar{x}\right)\left(y_{i\bullet}-\bar{y}\right)+\frac{1}{N}\sum_{i=1}^{M}n_i\mathrm{Cov}_i(X,Y) \qquad (1)$$

where x_{ij} and y_{ij} are the observations of the "independent" and "dependent" variables in the j-th region in the i-th cell, M is the number of aggregated cells, n_i is the number of regions in cell i, $x_{i\bullet}=\frac{1}{n_i}\sum_{j=1}^{n_i}x_{ij}$ is the aggregated value of X in cell j, $\bar{x}=\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}x_{ij}=\frac{1}{N}\sum_{i=1}^{M}n_i x_{i\bullet}$ is the overall mean, and $\mathrm{Cov}_i(X,Y)$ is the covariance of the variables X and Y *within* aggregate cell i. The process of aggregation removes the weighted variances of variable X (and Y) within each aggregate cell and it removes the weighted covariances between X and Y. Unlike the variance, which is always positive, the covariance can be either positive or negative, so it is difficult to predict whether the net change for a given aggregation will be positive or negative. Intuitively, knowing the behaviour of variance, one would expect that covariance would tend to decrease in absolute value with aggregation (except of course when it is initially zero) due to a decrease in the variability of both variables, with this tendency becoming more likely as the initial correlation between the variables increases.

Studying the behaviour of the change in correlation, defined by $r_{xy} = \frac{Cov(X, Y)}{s_x s_y}$, where s_j is the standard deviation of variable j, is complicated by the fact that the covariance and variances of X and Y are all independent, and so vary independently under aggregation (s_x and s_y will both decrease, but the covariance can either decrease or increase). Openshaw and Taylor (1979) compare the aggregated correlation to the relative change in variance of the dependent variable, which, although not incorrect, is not anywhere nearly enough to gain an understanding of how it varies either due to spatial properties of the variables or to aspatial properties, such as the original correlation between the variables. Since the behaviour of the variance (and hence standard deviation) is already known, the behaviour of the covariance needs to be examined along with that of the correlation. To this end, the experiment is divided into two sections, the first in which both X and Y have the same level of spatial autocorrelation, as measured by the MC, and the second in which their MCs differ. The behaviour of the linear regression slope parameter $b_{xy} = Cov(X, Y) / s_x^2$ is also of interest, as it only depends on two independent, yet mathematically similar, factors. Finally, if the regression residuals are spatially autocorrelated, then the requirement of independent residuals is violated and the validity of the linear regression analysis is compromised because the sampling distributions of the parameters, and hence the probabilities of Type I and Type II errors, are changed (Griffith, 1988, pp. 82-83). Cliff and Ord (1981, p. 191) show that the least squares estimator of β has a variance that is higher when the residuals are spatially autocorrelated, and Dutilleul (1993) and Clifford et al. (1989) note that spatial autocorrelation in the variables requires a modified version of the t-test for the significance of the correlation coefficient. It is therefore of interest to analyze the spatial behaviour of the residual under aggregation to see if the process improves or worsens this problem.

The spatial dataset generator described in Reynolds and Amrhein (1998) (and in more detail in Chapter 3) allows the creation of datasets with variables that have specified means, variances, Moran Coefficients (MC) of spatial autocorrelation, and also of the matrix of Pearson correlations between the variables. The incompatibility of certain combinations of MC and correlation and the requirement of positive definiteness of correlation matrices both act to hamper investigations of the behaviour of bivariate statistics, especially for negative correlations. The datasets, generated on the irregular tessellation of 400 regions posited by Reynolds and Amrhein (1998)

(and Chapter 4), attempt to observe the widest possible range of combinations of MCs and correlations. The first experiment involves setting the MC of each of five variables to the same value (ranging between -0.4 and 1.0) and having the correlations between them set to values between -0.8 and 0.8. The second experiment requires that as many correlations as possible be fixed at a specific value while the MCs of the variables be varied within the limits imposed by the desired covariance matrix. In both experiments, the variances of the variables are set to 6.0 and the means to 20.0 in order to have non-zero values to better simulate real data. Each dataset is run through the random aggregation model of Reynolds and Amrhein (1998) (described in detail in Chapter 3) 1000 times, with the desired aggregated statistics computed and stored after each run, and the overall distributions of the statistics tested for normality using the Kolmogorov-Smirnov test.

5.4. Results for fixed Moran Coefficients, varying correlations

Figures 5.1, 5.2, and 5.3 illustrate the changes in covariance, correlation, and the upper triangle of the regression slope parameters matrix, when both variables have the same MC and different correlations, for MCs of a) -0.4 and b) +0.8. The lower triangle slopes behave in a similar manner and are not shown. These figures are generated by running the model on a dataset with five variables, and hence with a possibility of ten different correlations. Nine of the correlations are labeled on the plots and range from -0.8 to 0.8; the tenth is set to a value that makes the covariance matrix positive definite. Since this value is between -0.8 and 0.8, it is felt that including its results would not be necessary for the analysis. As explained in Chapter 3, each group of lines represents one statistic of interest, in this case a particular initial correlation, and each line in a group represents the range of values of the aggregated statistic for a particular level of aggregation. The heavy dot represents the mean of the distribution, and the tic marks above and below it are one standard deviation away from it, to give an idea of the shape of the distribution. As it turns out, nearly all of the frequency distributions of all of the statistics generated by these experiments are normal, according to the Kolmogorov-Smirnov test, and those that are not too different from normal, so this will not be further discussed. One of the features of all three figures is the symmetric behaviour of the statistics, which is not unexpected since greater organization is represented by values further away from zero in either direction.

Figure 5.1 illustrates a clear trend towards zero covariance as the number of aggregated cells decreases. Table 5.1 illustrates these observations numerically, with the top row being the value of the MC of both variables, the next row being the original correlation, the third being the original covariance values, and the entries being the mean values from 1000 runs of the aggregation model. Clearly the covariance tends to behave like the variance, at least when the MCs of X and Y are the same, even though the weighted sum of internal covariances from Equation (1) can be either positive or negative. The change in the concavity of a line formed by the heavy dots, which are the means of the distributions in each group of lines, as the MC of the two variables becomes more positive is also worthy of note, as it mimics that of the variance as shown in Figure 4.2. The range of values increases with decreasing number of aggregate cells for highly autocorrelated variables, while the range decreases with decreasing number of cells for negatively correlated variables, a pattern that shows up again in Figure 5.5a.

The table and figure show that more covariation is lost (in the sense that the covariance is brought closer to zero) when the variables are negatively autocorrelated (about 96% between 400 regions and 40 cells) or weakly positively autocorrelated than when strongly autocorrelated (about 58%), and these losses are approximately the same for all levels of initial correlation. When X and Y are both strongly positively autocorrelated, the juxtaposition of similar values means that the spatial arrangement of aggregated values will be similar to that of the unaggregated values, and thus the change in covariance will not likely be as great as it will be for less spatially organized variables. The covariance will tend to decrease (if initially non-zero) during aggregation because the change in spatial arrangements of both variables is more likely to make their association more random than it is to make it more related. When both variables are highly autocorrelated, their covariance, like their individual variances, will tend to vary more as the number of aggregate cells decreases because it becomes more likely that the larger cells will contain greatly differing values and so increasing the (co)variance lost.

Figure 5.2 illustrates the aggregation effect on the correlation for pairs of variables with the same MC, while Table 5.2 presents numerical values from selected original correlations, whose values are the means of the 1000 runs of the aggregation model and are represented in the figure by the heavy dots. In general, the means of the distributions remain close to the original values of the correlation coefficients and do not change significantly with the level of aggregation,

while the range of values increases markedly as the MC decreases. As the number of aggregate cells decreases, the mean correlation tends to decrease in magnitude when the variable MCs are positive, but tends to increase slightly as the MCs decrease. Since a change in correlation is the result of a combination of decreases in magnitude of three factors, the standard deviations of X and Y in the denominator and their covariance in the numerator, a net decrease is caused by the covariance decreasing more than the standard deviations, while a net increase is caused by the standard deviations decreasing more than the covariance. When X and Y are strongly positively autocorrelated, neither their individual variances nor the covariance between them are much affected by aggregation, hence the correlation coefficients tend to not be greatly affected by aggregation either. As the MCs of the variables decrease, X and Y become more likely to vary differently from each other under aggregation because of the increasing tendency for dissimilar values to be located next to each other, resulting in a greater variation of aggregated results.

Figure 5.3 shows the behaviour of the upper triangle of the matrix of regression slope parameters for the MCs of -0.4 and 0.8. It can be seen that these slope parameters, along with those in the lower triangle (not shown), behave very similarly to the correlations, which is reasonable since the two statistics have similar forms and since the denominator terms $s_x s_y$ for correlation and s_x^2 for the regression slope both represent the products of two variables with the same MC.

Figure 5.4 shows the behaviour of the upper triangle of the matrix of Moran Coefficients of the regression residuals (MC_{RR}) when the MCs of the variables are -0.4 and 0.8; those from the lower triangle behave similarly and are not shown. Since the linear regression procedure ignores the spatial locations of the variables, it is expected that the regression residuals should have a similar level of spatial autocorrelation as the original variables when they both have the same MC. As Chapter 4 shows, variables with the same MC will not necessarily have the same spatial arrangement and hence their statistics will behave differently under aggregation, with the MC itself being the most unpredictable. All of the plots show a tendency for the residuals to become more randomly autocorrelated as the number of aggregated zones decreases, with this becoming more defined as the MCs of the variables increase. This finding reflects the behaviour of the aggregated MCs as discussed in Chapter 4. It can also be seen that the behaviour of the MC_{RR} is almost independent of the initial correlation of the two variables for these two MCs, although there is a

slight downward trend with increasing correlation visible when the variables have intermediate values of the MC (not shown).

5.5. Results for fixed correlation, varying Moran Coefficients

When the MCs of X and Y are allowed to vary independently, the number of potential combinations of MC and correlation increases dramatically. Some of them can be ruled out as impossible to create, if not theoretically then at least with the dataset generator, these being sets with variables that have high correlations and greatly differing MCs. This is not unreasonable, since highly correlated variables need to have similar spatial arrangements and this is simply not possible with variables that have very different spatial autocorrelations. Setting all of the correlations to the same value and varying the MC can be done for any value of the correlation that exceeds -0.2; for correlations less than -0.2 only the top row (and leftmost column from symmetry) of the matrix were set to the desired value and the remainder were adjusted until the covariance matrix became positive definite. Several different datasets are required for the larger correlations (especially large negative ones) in order to examine as many combinations as possible, which has the unfortunate effect of introducing pairs of variables with the same MCs and different spatial arrangements, whose aggregated statistics behave differently from each other and make it harder to derive general conclusions.

Interpretation of the results becomes more complex with this experiment as well. All of the remaining diagrams are similar to Figures 5.1 to 5.4, except that the initial correlation of the two variables is held constant while their respective MCs vary. Hence, the groups of lines are labeled (MC_x, MC_y), representing the Moran Coefficients of the independent and dependent variables. Figure 5.5 shows the behaviour of the covariance, correlation, upper triangle of the matrix of regression slope parameters, and the upper triangle of the MC_{RR} for an initial correlation of 0.0, for which only one data file was required to be generated. The first three statistics have initial values of zero and are equally likely to be positive or negative on aggregation, as the symmetry of the diagrams confirms. The most interesting feature of Figure 5.5a is the transition from the covariance increasing with decreasing number of aggregate cells for two highly autocorrelated variables (left hand group of lines) to it decreasing with decreasing number of cells for two negatively

autocorrelated variables. This can also be seen in Figures 5.1a and 5.1b for all the initial correlations, and is explained in the previous section.

Figure 5.5b shows that the range of aggregate correlations increases with decreasing number of cells for all combinations of variable MCs. As the MC of either variable decreases, the range of correlations for all levels of aggregation increases. Since the variability of the covariance does not appear to be much affected by the spatial autocorrelations of the two variables, as Figure 5.5a shows, this behaviour is due to the increasing variability of the variance (and hence standard deviation) of a variable as its MC decreases. The variability of the regression slope parameters increases as the difference between the MCs of the two variables increases, as shown in Figure 5.5c, and as with correlations it can be attributed to the variability of the variance of the independent variable increasing with decreasing MC. Finally, since the original slope parameter is zero for the uncorrelated data, the regression residual will be just the deviation of the dependent variable from its mean and hence the MC_{RR} is the MC of the dependent variable. Figure 5.5d shows that indeed the variation does not depend on the independent variable's MC.

As the original level of correlation between the two variables increases, similar patterns appear in the aggregated data as in the zero correlation example, albeit usually with less symmetry. As one would expect, the patterns for initially negative correlations are similar to those of their corresponding positive correlations, but reflected in the x-axis. Figure 5.6a, the change in covariance for an initial correlation of 0.4, illustrates the tendency for covariance to decrease in absolute value as the number of aggregate cells decreases, and as the MC of either variable decreases. As with the zero correlation case, the size of the range does not usually change significantly with the number of cells, except for cases of two highly autocorrelated variables, when the range increases with decreasing number of cells, and two negatively autocorrelated variables when the range decreases with decreasing number of cells.

The behaviour of the regression slope parameter b_1, is more regular than that of the other two statistics. Figure 5.6b shows the upper triangle of the matrix of b_1 for an initial correlation of 0.4 and was created by merging the results from two different files. The pattern with the zero initial correlation is repeated here, with the range showing a tendency to increase for all levels of aggregation as the independent variable decreases in MC, but with only a slight dependence on

the dependent variable's MC, which is reasonable given that the only influence the dependent variable can exert on the regression slope is through the covariance.

Because the initial MC_{RR} is very different for each variable, the difference between it and the aggregated MC_{RR} is examined. It can be seen that, at least for the case of an original correlation of 0.4 shown in Figure 5.6c, the behaviour seems more related to the MC of the independent variable than that of the dependent variable, as was the case for the initial correlation of 0.0. A general trend toward decreasing MC_{RR} for highly autocorrelated variables and increasing MC_{RR} for negatively autocorrelated variables indicates a tendency toward more random autocorrelation of residuals being produced by aggregation, indicating again that aggregation may actually improve the statistical reliability of regression results. Unfortunately, the need to create and merge several files for the initial correlation of 0.8 case and the resulting influence of the initial spatial distributions make drawing conclusions for higher correlations difficult (not shown).

As the initial level of correlation increases, the behaviour of the aggregated correlation becomes more unpredictable. When the initial correlation is moderate, such as in Figure 5.7a where it is 0.4, there is a strong tendency for correlations to increase with aggregation for all but the least spatially autocorrelated pairs of variables. This agrees with the general conclusions of papers published prior to Clark and Avery (1976) that state that correlations tend to increase with aggregation (Clark and Avery, 1976), a conclusion somewhat discounted by Openshaw and Taylor's (1979) results which show the peaks of the various distributions at or near the original correlation value. Clark and Avery's (1976) results show a correlation coefficient that increases steadily with level of aggregation from its initial value near 0.4, except for the last level where it decreases slightly, a behaviour that they considered an anomaly. Robinson (1950) described a correlation coefficient that increased from 0.203 at the individual level to 0.773 at state level and 0.946 at the (U.S. Census) division level, and Gehlke and Biehl (1935) presented two, the first which increased in absolute value monotonically from -0.502 to -0.763 and the second which started from -0.563, decreased in absolute value and then increased to end at -0.621. No information on the spatial autocorrelations of the variables was available for either of these three papers, but it is reasonable to assume that they were moderately positive.

Figure 5.7b shows the change in correlation for an initial correlation of 0.8 and graphically illustrates that the tendency for correlations to increase with aggregation does not always hold, at

least not for highly correlated variables. Each group of lines in a dashed box represents the behaviour of the aggregated correlation between two variables with the same combination of MCs as the other group. It can be seen that pairs of variables with the same MCs can behave quite differently under aggregation, an effect that is likely caused by differences in the spatial arrangements of the dependent and independent variables. This behaviour is a good subject for future research.

5.6. Discussion

In order to facilitate comparison with Openshaw and Taylor's (1979) study of the aggregation effect on correlations, a dataset with 8 variables, whose MCs alternate between 0.37 and 0.43, and which are all mutually correlated at 0.3466, is created using the correlation matrix of the 99 counties of the state of Iowa. Unlike the MCs and correlation, the means and variances were not stated in the paper, so they were all arbitrarily set to 20.0 and 6.0 respectively, the same as in the other experiments. The aggregation model is only run 1000 times on this dataset, as compared to the 10,000 runs of Openshaw and Taylor (1979), but prior experience has shown that there is little to gain in going beyond 1000 runs. As the model automatically generates eight levels of aggregation, from 45% to 10% of the original number of cells, the counties were aggregated to 45, 40, 35, …, and 10 regions. Figure 5.8a shows the variation in correlation between the pairs of variables whose MCs were 0.37 and 0.43. Table 5.3 presents summary information for the thirteenth group of lines of Figure 5.8a, which was selected because it has among the greatest extremes in the 10 aggregate cells values.

The patterns of the figure and the table show behaviour similar to that in Openshaw and Taylor's (1979) Figure 5.1, with normally or near-normally distributed variables whose frequency distributions become wider and flatter as the number of aggregate cells decreases. Figure 5.8b provides a comparison to a synthetic dataset in which all variables have MCs of 0.4 and varying degrees of correlation, as in Figures 5.1 to 5.4, but generated on the Iowa connectivity matrix, and it can be seen that the third group of lines from the right, representing the original correlation of 0.4, is similar to the groups in Figure 5.8a. The wider ranges in Figure 5.8b, as compared to a similar diagram for the 400-zone connectivity matrix (not shown, but see Figure 5.2), is due to the smaller number of zones in the Iowa dataset because the smaller numbers of zones means that dissimilar values will be closer together and hence more likely to be included within aggregate cells.

This, plus the behaviour of the means of the distributions, which both increase, decrease, and remain approximately the same, emphasizes the above conclusion that the behaviour of the correlation under aggregation is very difficult to predict and will depend on the spatial configurations and number of observations of the two variables.

5.7. Conclusions

The synthetic spatial dataset generator of Reynolds and Amrhein (1998) is used to search for a relationship between the effects of aggregation on the covariance and correlation and the spatial autocorrelations of the two variables whose interaction is measured. Two experiments are performed, the first in which the Moran Coefficients of the variables are equal and the correlations varied, and the second in which the correlations of variables are held constant and their MCs are varied. In both experiments, it is observed that the magnitude of the ranges of the covariances decreases with the decreasing number of aggregate cells for low values of variables' MC, but this gradually changes as the MCs increase until the ranges increase with decreasing numbers of aggregate cells. Even though the covariance can either increase or decrease with aggregation, unlike the variance which always decreases, in the vast majority of cases it decreases in magnitude, showing that variability is lost both within each variable and between them. One common factor of all the statistics and levels of aggregation is that all of the frequency distributions are either normal or nearly normal, even for the very complex MC of regression residuals (MC_{RR}).

When both of the variables have the same Moran Coefficient, the behaviour of the covariance, correlation, and regression slope parameter β_1 is quite regular, with the ranges of the statistics tending to increase as the MCs decrease, increase as the number of aggregate cells decreases, and decrease as the original correlation increases in magnitude. The MC_{RR} shows little variation with initial correlation, but its behaviour changes as the MCs of the two variables increase, showing a marked tendency to decrease as the number of aggregate cells decreases. Since spatial autocorrelation of residuals is a violation of the desirable property of independent residuals, the decrease in MC indicates that the quality of results of linear regression will actually be improved by aggregation, although the loss of information through aggregation makes this improvement questionable.

When the variables' MCs differ and the initial correlation is zero, the behaviour of the bivariate statistics is still reasonably regular. The covariance has its properties discussed above, while the range of correlations shows a definite trend toward increasing as the MCs of the variables decrease. As expected, the greatest variability in the b_1 values occurs for the variables with the greatest differences in MCs, while again the ranges generally increase as the MCs of the variables decrease. The behaviour of the MC_{RR} depends on the MC of the dependent variable only, since an initially zero b_1 means the initial MC_{RR} is that of the deviation of y about its mean. When the variables' MCs differ but the initial correlation is non-zero, reliable prediction of the statistics becomes much more difficult, especially for MC_{RR} and correlation, as differences in results due to different spatial configurations of the variables can be dramatic. The unfortunate conclusion that must be drawn is that prediction of the unaggregated values of bivariate statistics will be, if possible at all, a very difficult process. Clark and Avery (1976) hypothesize that deviations in the behaviour of the coefficients are related directly to how the covariation is affected by aggregation and indirectly by the spatial autocorrelations of the variables, but do not agree with a hypothesis by Blalock (1964) that the deviations are caused by reduction in variation of the dependent or independent variable. My results indicate that both are partially correct – the behaviour is related to *all* of these causes, which is why they, using only a few real datasets without the benefit of being able to vary parameters at will, had difficulty drawing their conclusions.

In order to compare the results of the experiments to those of Openshaw and Taylor (1979), a synthetic dataset was generated on the connectivity matrix of the 99 counties of Iowa whose variables have MCs of 0.37 and 0.43 and correlations of 0.3466 to match the properties of the variables in that paper. The results appear to be in agreement, with the distributions becoming wider and flatter with aggregation, and the ranges becoming quite large as the number of zones becomes small. The ranges are larger with the smaller number of initial regions as compared to the 400 zones of the test datasets because dissimilar values are closer together, even for high MCs, increasing the chance of having aggregate cells with larger internal variations. The fact that some distribution means increase, while others decrease or stay roughly the same, highlights the dependence of the correlation on the spatial distribution of the variables, even though the correlation has no spatial component.

Statistical simulation is proving to be a useful tool in the continuing attempts to understand the workings of the MAUP, especially with the more complex bivariate and multivariate statistics. Unfortunately, it seems that a higher level of sophistication than the Moran Coefficient is required to numerically describe the spatial pattern if attempts to predict and hence exploit the behaviour of statistics under aggregation are to have any hope of success.

5.8. References

Blalock, H., 1964: *Causal Inferences in Nonexperimental Research*. (Chapel Hill: University of North Carolina Press).

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

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

Clifford, P., S. Richardson, and D. Hémon, 1989: Assessing the significance of the correlation between two spatial processes. *Biometrics*, **45**, 123-134.

Dutilleul, P, 1993: Modifying the t-test for assessing of the correlation between two spatial processes. *Biometrics*, **49**, 305-314.

Gehlke, C. E., and K. Biehl, 1934: Certain effects of grouping on upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, **29**, 169-170.

Griffith, D. A., 1988: *Advanced Spatial Statistics*. (Dordrecht: Kluwer).

Openshaw, S., and P. Taylor, 1979: A million or so correlation coefficients: Three experiments on the modifiable area unit problem. In *Statistical Applications in the Spatial Sciences*, Ed. N. Wrigley, (Pion, London), 127-144.

Robinson, W. S., 1950: Ecological correlations and the behavior of individuals. *American Sociological Review*, **15**, 351-57.

Reynolds, H., and C. Amrhein, 1998: Using a spatial dataset generator in an empirical analysis of aggregation effects on univariate statistics. *Geog. and Env. Modelling*, **1(2)**, 199-219.

5.9. Tables

Table 5.1: Variation of the covariance with original MC of the variables and correlations

	Original MC = -0.4			Original MC = 0.8		
Cells	r = -0.6	r = 0.4	r = 0.8	r = -0.6	r = 0.4	r = 0.8
400	-3.6000	2.4000	4.8000	-3.6000	2.4000	4.8000
180	-1.0733	0.6401	1.3696	-3.0226	2.0130	4.0241
160	-0.8969	0.5340	1.1428	-2.9355	1.9506	3.9038
140	-0.7287	0.4296	0.9260	-2.8314	1.8747	3.7574
120	-0.5844	0.3404	0.7388	-2.6993	1.7812	3.5717
100	-0.4299	0.2601	0.5497	-2.5401	1.6691	3.3467
80	-0.3204	0.1869	0.4054	-2.3294	1.5157	3.0468
60	-0.2095	0.1217	0.2640	-2.0166	1.3023	2.6201
40	-0.1151	0.0688	0.1457	-1.5468	0.9773	1.9725

Table 5.2: Variation of the correlation with original MC of the variables and correlations

Cells	Original MC = -0.4			Original MC = 0.8		
400	r = -0.6	r = 0.4	r = 0.8	r = -0.6	r = 0.4	r = 0.8
180	-0.6202	0.3899	0.8008	-0.6041	0.4008	0.8011
160	-0.6238	0.3911	0.8020	-0.6035	0.4002	0.8000
140	-0.6240	0.3874	0.8040	-0.6030	0.3994	0.7984
120	-0.6289	0.3881	0.8041	-0.6013	0.3983	0.7956
100	-0.6220	0.3927	0.8032	-0.5995	0.3979	0.7922
80	-0.6301	0.3895	0.8071	-0.5967	0.3957	0.7861
60	-0.6288	0.3898	0.8044	-0.5869	0.3905	0.7742
40	-0.6242	0.3928	0.8014	-0.5710	0.3815	0.7518

Table 5.3: Summary information for the thirteenth group of distributions in Figure 5.8a

Cells	Mean	Std Dev	Min	Max	Range
99	0.3466				
45	0.3193	0.0500	0.1497	0.4938	0.3440
40	0.3112	0.0557	0.0761	0.4500	0.3739
35	0.3048	0.0643	0.0898	0.5023	0.4125
30	0.2928	0.0767	0.0048	0.5254	0.5206
25	0.2813	0.0951	-0.1720	0.5309	0.7029
20	0.2692	0.1166	-0.2637	0.6245	0.8882
15	0.2483	0.1672	-0.5425	0.7013	1.2438
10	0.2212	0.2565	-0.7585	0.9003	1.6588