4. The Effect of Aggregation On Univariate Statistics[1]

4.1. Summary

The resistance of the Modifiable Area Unit Problem to analytical solution requires that it be investigated by numerical and empirical studies that have the potential to lay the foundations for analytical approaches. The use of synthetic spatial datasets, whose spatial autocorrelation, mean, and variance of individual variables, and Pearson correlation between variables, can be controlled greatly enhances the ability of the analyst to study the MAUP in this manner. This chapter explores the effects of spatial aggregation on the variance and three univariate spatial autocorrelation statistics using a synthetic 400-region dataset. The relationship between the relative change in variance and a modified version of the G statistic that was first proposed by Amrhein and Reynolds (1996, 1997) is explored in more detail. These results compare favourably with results generated from the Lancashire dataset of Amrhein and Reynolds (1996).

4.2. Introduction

The Modifiable Area Unit Problem (MAUP) has been the focus of research interest for many years, with the current resurgence in interest being initiated by Openshaw and Taylor (1979) and fueled by the rapidly increasing computing power available to analysts. It is well known that the application of statistical results derived from one level of spatial resolution to a higher resolution (such as census tract data being used to predict individual household information) can result in serious errors; this all too common error has been named the *ecological fallacy*. An ancillary effect of the enhanced computing power is the proliferation of Geographical Information Systems (GIS) and other spatial analysis tools. As the MAUP has been either ignored or written off as intractable in many research results, it can be expected to get short shrift by users of this software who are unaware of the subtleties of spatial data analysis. The importance of gaining an understanding of the MAUP and how it can be taken into account in GIS software to reduce the numbers of flawed analyses and their possibly expensive repercussions cannot be understated.

---

[1] This is a modified version of the paper Reynolds and Amrhein (1998): Using a spatial dataset generator in an empricial analysis of aggregation effects on univariate statistics. *Geog. and Env. Modelling*, **1(2)**, 199-219.

Theoretical work, such as that by Arbia (1989), has shown that an analytical solution is possible, but under restrictive conditions that would seldom be found in real life situations. As a result, research into the MAUP has been primarily empirical, focusing on the effects of aggregation on various statistics computed from a specific dataset. For example, Openshaw and Taylor (1979) examine correlation coefficients using an Iowa electoral dataset, Fotheringham and Wong (1991) study multiple regression parameters using Buffalo census data, Amrhein and Reynolds (1996), one of the papers in the special issue of *Geographical Systems* that focuses on the MAUP, and Amrhein and Reynolds (1997) study the effects of aggregation on univariate statistics and make a tentative link between a spatial statistic and the relative change in variance. Recognition of spatial patterns is a fundamental requirement for landscape ecology, and various spatial autocorrelation statistics, such as the Moran Coefficient, are often employed as a tool for this task (Jelinski and Wu, 1996; Qi and Wu, 1996); hence it is important to know how spatial statistics are affected by aggregation as well.

The use of synthetic spatial datasets overcomes the difficulties inherent in publicly available sets, with census data being the prime example. Possible errors in the data notwithstanding, the greatest frustration for researchers into the MAUP is that one has no control over the values of spatial autocorrelation, means, variances, or Pearson correlations between variables; one must work with the data at hand. Amrhein (1995) is the first to use synthetic datasets in the study of the MAUP by locating points randomly within a unit square, assigning them random values, imposing various sized square grids, and aggregating the points within each square. This chapter extends this approach by employing more sophisticated synthetic datasets to explore the effects of spatial aggregation on the weighted variance and on three commonly-used spatial autocorrelation statistics, the Moran Coefficient, the Geary Ratio, and the Getis (G) statistic. The following sections discuss the method of analysis, the results, and the conclusions.

4.3. Method

The dataset generator, aggregation algorithm, and method for interpretation of the diagrams are described in detail in Chapter 3. The frequency distributions of values tend to be mound-shaped and unimodal, but are not usually normal (see Figure 4.1 for examples). The spa-

tial connectivity matrix is created from either a rectangular grid or a tessellation of randomly-generated Voronoi polygons, depending on the experiment.

Three spatial datasets of 400 Voronoi polygons and 8 variables are created using the data-set generator. In order to test the effect of spatial autocorrelation on spatial aggregation, the first two sets are created with variables that are mutually uncorrelated, have variances of 6.0 and means of 20.0, and have Moran Coefficients of -0.4, -0.2, 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0. The non-zero mean is required so that all values are greater than zero in order for the Getis statistic to be valid, as well as to match most real datasets. To see if the variance of the variable affects the aggregated values, another set is created with variables that are mutually uncorrelated and have means of 20.0, but have the same Moran Coefficient values of 0.0 and variances of 5.0, 10.0, 20.0, 30.0, 40.0, 50.0, 60.0, and 70.0. The random aggregation model of Amrhein and Reynolds (1996, 1997) and Reynolds and Amrhein (1998)[2] was run 1000 times on each dataset and the relative change in variance, Moran Coefficient, Geary Ratio, and G statistic were saved for each of 8 levels of aggregation. Also saved were the following non-standard statistics:

$$MC_1 = \left[\frac{m}{S_C}\right]\left[\sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)\right] \tag{1}$$

$$GR_1 = \left[\frac{m-1}{2S_C}\right]\left[\sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij}\left(x_i - x_j\right)^2\right] \tag{2}$$

$$G = \left[\frac{m}{Sc}\right]\left[\sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij}x_i x_j\right]\left[2\sum_{i=1}^{m}\sum_{j=i+1}^{m} x_i x_j\right]^{-1}\left[\frac{1}{m}\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2\right]^{-1} \tag{3}$$

where $S_C = \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij}$ and m is the number of aggregate cells. MC_1 and GR_1 are just modified versions of the Moran Coefficient and Geary Ratio, while G is the G statistic (Getis and Ord, 1992; Ord and Getis, 1995) modified by dividing by the aggregate unweighted variance. These statistics are computed as part of the testing of possible correlation between equation (3) and the relative change in variance in Section 4.4. Equation (3) is slightly different from the modified G used in Amrhein and Reynolds (1996, 1997), who divided by the sum of squares of deviations, rather than

---

[2] Described in detail in Chapter 3.

the variance. To test the effectiveness of the new dataset generator at simulating a real dataset, the Lancashire dataset of Amrhein and Reynolds (1996) and a synthetic replication were run through the aggregation model and the results are compared. It is impractical to attempt to replicate large datasets such as the Toronto set of Amrhein and Reynolds (1997), since the time and effort required to compute the eigensystem of a matrix with 5370 rows and columns is enormous.

4.4. Results

4.4.1. The effects of aggregation on the variance

Figure 4.2a illustrates the aggregation behaviour of the relative change in variance (RCV), $\left(\sigma_o^2 - \sigma_{agg}^2\right) / \sigma_o^2$, where σ_o^2 is the variance of the N regions, and $\sigma_{agg}^2 = \frac{1}{N}\sum_{i=1}^{M} n_i \left(x_i - \overline{x}\right)^2$ is the aggregated variance that is weighted by the number of regions n_i in the M aggregated cells. A value of RCV near one (as in the first group of lines in Figure 4.2a) means that the aggregated weighted variance is much closer to zero than the original variance, while a value near zero (as in the last group of lines in Figure 4.2a) means that the new variance is very similar to the original. The diagrams are explained in detail in Section 3.3.

It can be shown that the variance of a spatially located variable can be partitioned into the sum of variances within various sub-regions and the variance of the average values of all the subregions (see Section 5.3 and Moellering and Tobler, 1973). The process of aggregation removes the former, so the more spatially homogeneous (i.e. positively autocorrelated) a variable is, the smaller the variance within each cell will be (on the average) and hence the less variance is lost. As the number of aggregate cells decreases (i.e. fewer, larger regions), the loss in variance obviously increases, since a greater number of values are being lost. Both of these patterns are well demonstrated in Figure 4.2a. As the number of aggregate cells decreases, the number of regions per cell increases on average, since the aggregation algorithm attempts to have similar numbers of regions per cell, but does not strictly enforce this ideal. When significantly positively autocorrelated variables are aggregated, increasing the number of regions per cell increases the likelihood that more widely differing values will be included in each cell, so one would expect the variability of possible aggregate variance values to increase with a decrease in the numbers of cells. With negatively or near-randomly autocorrelated variables, however, the tendency towards the

juxtaposition of widely differing values means that as the number of regions per cell increases, the opportunity for variation in the aggregate variance values will tend to remain the same or decrease. Both of these patterns are demonstrated in Figure 4.2a. When variables of the same MC but different variances were aggregated, it was found that the variance of the original variable had no discernible impact upon the distributions of the RCV (not shown). Only the spatial organization of the variable plays a major role in the new variance.

4.4.2. The effects of aggregation on the Moran Coefficient

Explanation for the changes in spatial autocorrelation, as explained by the aggregated Moran Coefficient, is more difficult. Figure 4.2b was created by running the model on the same dataset as Figure 4.2a. Unfortunately, the nice clear pattern seen in the figure for variances is not present here. There is an upward trend in the ranges as the MC increases for the first three and last three variables, but the variables whose MCs are 0.2 and 0.4 behave very similarly to the one with MC of -0.2. Clearly the behaviour of the MC is much more complex than the variance and further exploration is required.

Figures 4.3a to 4.3d illustrate 16 variables, 8 on the irregular tessellation used in the other experiments and 8 on a 20×20 square grid, each of which has a MC of 0.8. Each figure has four variables illustrated at the top and their estimated variograms (Cressie, 1993, p. 69) below. The variograms are isotropic (i.e. a function of distance only, not of direction) and computed using the standard method of moments estimator $2\hat{\gamma}(h) = \frac{1}{N(h)} \sum_{N(h)} \left(Z(\mathbf{s}_i) - Z(\mathbf{s}_j) \right)^2$ (Cressie, 1993, p. 69), where h is the Euclidean distance between the points $\mathbf{s}_i=(x_i, y_i)$ and $\mathbf{s}_j=(x_j, y_j)$ and $Z(\mathbf{s})$ is the variable value at point $\mathbf{s}$. Because the data locations are regions, their centroids are used for the values of $\mathbf{s}$. This formula states that the value of the variogram at a distance h (plotted as the x coordinate of the diagram) is the sum of all the values of $\left(Z(\mathbf{s}_i) - Z(\mathbf{s}_j) \right)^2$ where the Euclidean distance between $\mathbf{s}_i$ and $\mathbf{s}_j$ is less than or equal to h divided by the number of pairs of points that meet this criterion. The variogram "acts as a quantified summary of all the available [spatial] structural information, which is then channeled into the various procedures of resource and reserve evaluation" (Journel and Huijbregts, 1978, p. 12).

Figures 4.3a to 4.3d clearly show that variables with the same MC can have very different spatial structures, although the possibilities decrease as the MC approaches the maximum allowed by the spatial structure. The location of the maximum of the variogram can be used as a crude approximation of the length scale of the spatial structure. Variables with a short length scale, such as those in Figures 4.3a and 4.3b, also have variograms that oscillate about the asymptotic value. The downward component of the oscillation occurs when the distances are great enough to reach from one cluster to another similar one, allowing more differences between similar values to be included in the sum, and the upward component occurs when the distances allow more dissimilar pairs of values to be included in the sum.

Figures 4.4a and 4.4b illustrate the effect of the spatial arrangement on the aggregated MC and RCV respectively. Each set of lines has a label that corresponds to the respective variable in Figures 4.3a to 4.3d, and the diagrams are divided into four sections to indicate in which figure each variable is located. As expected, the behaviour of both of the statistics is related to the arrangement of the values. As long as the aggregate cells are, on average, of a similar or smaller size than the length scale of the variable, then similar values will tend to be aggregated and hence the variance will not be greatly affected. With the aggregate cells having similar values to the unaggregated cells, similar values will still tend to be next to each other and so the spatial autocorrelation will not be much affected either and in fact may even increase somewhat (Figure 4.4a, Variables 11 to 15). As the number of cells decreases and size increases to reach and exceed the length scale, then more and more dissimilar values will be included within an aggregate cell and the loss in variance will be greater. Increasing variability of the values within the aggregate cells makes it more likely that dissimilar values will be located next to each other in the aggregated region, hence lowering the spatial autocorrelation, sometimes dramatically, creating a strongly negatively aggregated variable where it was strongly positive before. A more detailed analysis of spatial pattern's effect on aggregation will be a topic for future research.

4.4.3. Frequency distributions

As it is of interest, and potentially useful, to learn about the frequency distributions of the aggregated statistics, the distribution of statistic values for each statistic at each level of aggregation is tested for normality using both the Kolmogorov-Smirnov (K-S) and Shapiro-Wilk tests. In

order to see if having more points is beneficial, the tests are conducted cumulatively on the first 100 runs, the first 200 runs, and so on until all 1000 points are included. Tables 1a and 1b (at the end of the chapter) present a summary of the K-S test results for selected statistics, aggregation levels, and numbers of runs for variables with initial MCs of -0.4 and 1.0 respectively. The second column lists the critical value of the K-S test; if the computed statistic is less than it (for example, the RCV for 180 cells at 100 runs is 0.0431 and the corresponding critical value is 0.1360) then the frequency distribution is normal. All of the distributions are either normal or close to normal, including the ones not shown. As a general rule, the distribution deviates more from a bell-shaped curve as the number of aggregate cells decreases. As the number of runs increases, the K-S statistics indicate a trend towards a less normal distribution, but this is probably at least partly an artifact of the $n^{-1/2}$ dependence of the critical value. This sort of problem is common among simulation analyses in which one must decide the optimum number of experiments based on an increase in accuracy due to more runs versus a shrinking confidence interval. For the most part, the values of the K-S statistic decrease slightly or remain about the same with increasing MC of the unaggregated variable, meaning that the values become more normally distributed. Curiously, the RCV of the 180 cell aggregation is a glaring exception to this observation; why this is so requires further investigation. Tables 4.2a and 4.2b on page 31 present selected results for the Shapiro-Wilk tests for the same variables as above, and the values corroborate the conclusions drawn from the first two tables.

4.5. Correlating the change in variance with a spatial statistic

Amrhein and Reynolds (1996, 1997) and Reynolds and Amrhein (1998) have indicated that a relationship could exist between the relative change in variance (RCV) and the aggregated G statistic, defined as G by Equation (3), which is the classic G statistic (Getis and Ord, 1992) modified by dividing it by the unweighted variance σ_u^2 of the aggregated values. The primary challenge is to prove that this relationship is not simply due to the presence of similar terms on both sides of the equation: the weighted variance in the numerator of the Relative Change in Variance (RCV) and the unweighted variance in the denominator of the modified G.

Figure 4.5a illustrates the RCV as a function of the aggregated variable MC_1, defined by Equation (2), for the variable whose initial MC is -0.4, while Figure 4.5b illustrates that of RCV

and the aggregated regular MC. Plots for the modified and regular Geary Ratio are very similar and so are not shown. These plots and those of Figure 4.6 are created using the statistic values from every tenth model run, and each level of aggregation has its own symbol. It is immediately obvious that the inclusion of the sum of squares of deviations term turns a fairly strong non-linear relationship into a very weak one. Figure 4.5 and the equivalent Geary Ratio plots serve as a counterexample to the argument that the relationship between the modified G statistic and the RCV is caused by the inclusion of this term.

Figure 4.6a shows the relationship between the RCV and $\log_{10}(G)$ for the variable with MC of -0.4, while 4.6b illustrates that between RCV and $\log_{10}$(modified G). The logarithm is required for clarity because the G and modified G values occur over two orders of magnitude. It is clear that inclusion of the aggregated variance (with its sum of squares of deviations) creates a very good non-linear relationship where there was none before. Note that the initial MCs of -0.4 are used in Figures 4.5 and 4.6 because they best illustrate the argument. With a little work it can be shown that the Moran Coefficient and modified G statistic can be written in terms of the Geary Ratio (for the former, see Griffith, 1987, p. 44), and it is this relationship, coupled with the evidence in Figure 4.5, that suggests that the relationship between the RCV and the modified G statistic is a real one, and not one created by the presence of similar terms on both sides of the equation.

With the above conclusion reached, the points for all levels of aggregation and the various MCs of the original variables were fitted, using least squares, to an equation of the form $RCV = A*G + B*\log_{10}(G) + C*M + D*\log_{10}(M+\alpha) + E$, where G is the aggregated modified G statistic, M is the Moran Coefficient of the unaggregated variable, and α is a number large enough to ensure that the logarithm is defined. In this case, α=0.5 since the lowest MC used is -0.4, but values in the 0.4 to 0.6 range produce fits with similar values of R^2. The original MC is included in this equation because of the obvious dependence of RCV on it that is displayed in Figure 4.1a. Fits generated from various datasets with variables of varying MC consistently generated R-squared values in the 0.9 range and have very significant F-test results. Unfortunately, initial attempts to exploit this relationship to predict the variance of an unaggregated variable have not been successful, and work on this continues.

4.6. Comparison of synthetic data to a real dataset

The use of synthetic spatial datasets to systematically examine the MAUP is essential, as real datasets do not offer the flexibility of spatial and aspatial parameter control that can be defined by an appropriate experimental design. In any sort of empirical experiment, one must be able to identify any factors, such as the spatial autocorrelation and pattern, variance, and correlation of the variables or the level of aggregation, that might have an impact on the results. After these factors are identified, the experiments must be designed in such a way as to allow each factor to be systematically varied over its feasible or practical range in order to judge its influence on the outcome. When a single dataset is used, such as in Openshaw and Taylor (1979) to study correlations, or in Fotheringham and Wong (1991) to study multivariate statistics, the researcher is limited to whatever means, variances, correlations, MCs, and other properties that the variables have. Conclusions that are drawn cannot be tested for the effects of a different MC or correlation coefficient, resulting in what is effectively one tree in the forest of the behaviour of the MAUP.

It is important, however, to see how well the behaviour of a real dataset is mimicked by that of a synthetic counterpart, i.e. a dataset created to have the same MCs, variances, correlations, and means (so long as none of the synthetic variable values are negative). A good correspondence will increase confidence in the validity of applying conclusions about the MAUP based upon synthetic data to real world situations. Two weaknesses of this dataset generator became apparent during the experimentation that led to this paper. The first, an inability to control the frequency distribution of the values, often manifested itself in a need to shift the mean of a variable so that the lowest value was zero, but was otherwise not of much consequence. The second, an inability to control the spatial pattern of the values, poses a greater potential problem to dataset simulation, as the behaviour of the spatial characteristics like MC depends on the spatial arrangement (section 4.5.2) as well as the level of spatial autocorrelation inherent in it.

To this end, we employ the Lancashire dataset previously used in Amrhein and Reynolds (1996). Figure 4.7 compares the behaviour of the RCV of all eight variables in this dataset to a set of synthetic counterparts whose parameters match the originals. Generally speaking, there is a good correspondence between the locations of the means of the distributions from the two datasets, though it can be seen that the values from the synthetic set generally occupy wider ranges. This difference may be caused at least in part by differences between the spatial arrangements of

the original and synthetic variable values (such as in Figure 4.9), and needs further investigation. Figure 4.8 compares the behaviour under aggregation of the Moran Coefficients of the variables in the two datasets. It can be seen that the last four variables of the sets behave similarly, while the first four have often dramatic differences, the greatest of which occurs with the first variable, MTDEP. Figure 4.9 compares the spatial distributions of the original and synthetic values of this variable, with the distribution ranges divided up such that each encloses an equal number of the 304 wards to facilitate visual comparison. The dramatic differences between the two, which both have an MC of 0.36, are more than likely to be the cause of the differences in the behaviour under aggregation of their MCs, as is mentioned above.

4.7. Conclusions

The preceding experiments have demonstrated some interesting properties of statistics that are computed from spatially aggregated data. They were made possible by the creative control over the synthetic data provided by the new generator. All statistics, even the complex spatial ones, fall within well-defined distributions that are normal or nearly so, and whose parameters (mean and standard deviation) are determined by the level of aggregation. The RCV shows a strong dependence on the spatial autocorrelation of the original variable, as opposed to the spatial statistics like the MC and Geary Ratio whose dependence on the original spatial autocorrelation (as measured by the original MC) is unclear. The spatial arrangement of the data, especially for high levels of MC, also plays an important role for both the aggregated MC and variance. None of the statistics shows any discernible relationship with the variance of the unaggregated dataset, however, indicating that it is the spatial distribution of the values, rather than the values themselves, that largely determine the behaviour of the dataset under spatial aggregation. The RCV is also found to be highly correlated with a non-linear function of both the original MC and the modified G statistic, having an R^2 value of the order of 0.9. It is argued that the strength of this relationship is not due to the presence of similar terms on both sides of the equation (weighted variance in the LHS and unweighted in the RHS) but is in fact genuine. This represents a small step toward the ultimate goal of estimating the values of the various unaggregated statistics, but more work is required in order to effectively exploit this relationship. Various attempts to use it

to predict the original variance of an aggregated dataset have been unsuccessful, and research on this problem continues.

The new spatial dataset generator provides more flexibility in the creation of datasets than does the old one. The pair-swapping algorithm employed in the older generator does not allow for the creation of variables whose spatial patterns are representative of the entire range of possible patterns, and also only allows the first row of desired correlations to be computed. Unfortunately, it does not allow for control over the final spatial distribution of a variable, or the frequency distribution of its values. While this does not appear to seriously affect the ability of synthetic datasets to mimic the aspatial aggregation properties of their univariate statistics, the behaviour of spatial statistics like the Moran Coefficient can be dramatically different between the true variable and its synthetic counterpart due to differences in the spatial arrangements. It is clear that the dataset generator is still in need of some refinements.

Among the most interesting and potentially useful results include the fact that aggregate statistics, both spatial and non-spatial, form normal or near-normal sampling distributions whose bounds are relatively small compared to the range of possible values of the statistics. This is a strong indication that the results of aggregation are not chaotic, but behave in a well-defined manner. The normality of the distributions is interesting because of the complexity of the processes involved, especially for the spatial statistics. Since most statistical theory is built around assumptions of normally distributed data, a cynic would expect Murphy's Law to act to make the distributions something other than normal. Exploration of this feature is another topic for future research. Programs to estimate the effect of the MAUP such as the ones used here have the potential to be incorporated into routines in GIS software packages once sufficiently sophisticated algorithms, backed by a more thorough knowledge of the theory behind what is going on, become available. As this occurs, one of the most troublesome sources of error in the analysis of spatially referenced data may finally be rendered tractable to even the most inexperienced GIS users and the ultimate goal of being able to estimate the true statistical parameter values of a spatially aggregated dataset may finally be achieved.

4.8. Tables

Table 4.1a: Selected K-S Test Statistics: Variable with Original MC of -0.4

	Critical	RCV		Moran Coeff		Geary Ratio		Modified G	
RUNS	K-S	180	40	180	40	180	40	180	40
200	0.0962	0.0395	0.0807	0.0534	0.0508	0.0339	0.0405	0.0553	0.0673
400	0.0680	0.0215	0.0920	0.0322	0.0471	0.0251	0.0372	0.0393	0.0624
600	0.0555	0.0262	0.0770	0.0238	0.0446	0.0209	0.0305	0.0335	0.0624
800	0.0481	0.0249	0.0797	0.0138	0.0368	0.019	0.0288	0.0262	0.0655
1000	0.0430	0.0266	0.0719	0.0147	0.0375	0.0198	0.0246	0.0227	0.0728

Table 4.1b: Selected K-S Test Statistics: Variable with Original MC of 1.0

	Critical	RCV		Moran Coeff		Geary Ratio		Modified G	
RUNS	K-S	180	40	180	40	180	40	180	40
200	0.0962	0.0473	0.0363	0.0358	0.0324	0.0324	0.0453	0.0382	0.0426
400	0.0680	0.0355	0.0313	0.0302	0.0196	0.0323	0.0431	0.0322	0.0347
600	0.0555	0.0345	0.0263	0.0204	0.0211	0.0355	0.0329	0.0241	0.0399
800	0.0481	0.0348	0.0350	0.0193	0.0182	0.0278	0.0341	0.0233	0.0410
1000	0.0430	0.0304	0.0292	0.0175	0.0187	0.0261	0.0336	0.0226	0.0353

Table 4.2a: Selected Shapiro-Wilk Statistics: Variable with Original MC of -0.4

	RCV		Moran Coeff		Geary Ratio		Modified G	
RUNS	180	40	180	40	180	40	180	40
200	0.9824	0.9445	0.9838	0.9663	0.9720	0.9572	0.9557	0.9530
400	0.9795	0.9115	0.9770	0.9673	0.9735	0.9518	0.9636	0.9447
600	0.9772	0.9239	0.9781	0.9737	0.9685	0.9624	0.9662	0.9347
800	0.9782	0.9275	0.9744	0.9768	0.9685	0.9662	0.9700	0.9349
1000	0.9773	0.9295	0.9726	0.9754	0.9675	0.9664	0.9689	0.9137

Table 4.2b: Selected Shapiro-Wilk Statistics: Variable with Original MC of 1.0

	RCV		Moran Coeff		Geary Ratio		Modified G	
RUNS	180	40	180	40	180	40	180	40
200	0.9606	0.9658	0.9621	0.9754	0.9728	0.9623	0.9515	0.9662
400	0.9644	0.9679	0.9669	0.9746	0.9683	0.9629	0.9614	0.9651
600	0.9669	0.9707	0.9720	0.9756	0.9651	0.9669	0.9669	0.9648
800	0.9644	0.9702	0.9723	0.9734	0.9670	0.9660	0.9701	0.9651
1000	0.9640	0.9691	0.9746	0.9719	0.9680	0.9636	0.9697	0.9657

4.9. References

Amrhein, 1995: Searching for the elusive aggregation effect: Evidence from statistical simulations. *Env. and Planning A*, **27**, 259-274.

Amrhein, C. G., and H. Reynolds, 1996: Using spatial statistics to assess aggregation effects. *Geographical Systems*, **3**, 143-158.

Amrhein, C. G., and H. Reynolds, 1997: Using the Getis statistic to explore aggregation effects in Metropolitan Toronto census data. *The Canadian Geographer*, **41(2)**, 137-149.

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

Cressie, N. A. C., 1993: *Statistics for Spatial Data, Revised Edition*. (New York: Wiley)

Fotheringham, A. S., and D. W. S. Wong, 1991: The modifiable area unit problem in multivariate statistical analysis. *Env. and Planning A*, **23**, 1025-1044.

Getis, A., and K. Ord, 1992: The analysis of spatial information by use of a distance statistic. *Geographical Analysis*, **24**, 189-206.

Griffith, D. A., 1987: *Spatial Autocorrelation: A Primer*. (Washington, DC: American Association of Geographers).

Griffith, D. A., 1996: Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. *The Canadian Geographer*, **40(4)**, 351-367.

Jelinski, D. E., and J. Wu, 1996: The modifiable area unit problem and implications for landscape ecology. *Landscape Ecology*, **11(3)**, 129-140.

Journel, A. G., and C. J. Huijbregts, 1978: *Mining Geostatistics*. (London: Academic Press)

Moellering, H., and W. Tobler, 1973: Geographical Variances. *Geographical Analysis*, **4**, 34-50.

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.

Ord, J. K., and A. Getis, 1995: Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis*, **27(4)**, 286-306.

Qi, Y., and J. Wu, 1996: Effects of changing resolution on the results of landscape pattern analysis using spatial autocorrelation indices. *Landscape Ecology*, **11(1)**, 39-49.

Reynolds, H., and C. G. Amrhein, 1998: Some effects of spatial aggregation on multivariate regression parameters. *Econometric Advances in Spatial Modelling and Methodology: Essays in Honour of Jean Paelinck*, D. Griffith, C. Amrhein and J-M. Huriot (eds.). Dordrecht: Kluwer.

Tiefelsdorf, M., and B. Boots, 1995: The exact distribution of Moran's I. *Env. and Planning A*, **27**, 985-999.