THE MODIFIABLE AREA UNIT PROBLEM: EMPIRICAL ANALYSIS BY STATISTICAL SIMULATION

by

Harold David Reynolds

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Geography
University of Toronto

THE MODIFIABLE AREA UNIT PROBLEM:
EMPIRICAL ANALYSIS BY STATISTICAL SIMULATION

By Harold David Reynolds
Doctor of Philosophy
Graduate Department of Geography
University of Toronto
1998

Abstract

The Modifiable Area Unit Problem (MAUP) has been discussed in the spatial analysis literature since the 1930's, but it is the recent surge in the availability of desktop computing power and Geographical Information Systems software that have caused both a resurgence of interest in the problem and a greater need to learn more about it. Many spatial datasets are collected on a fine resolution (i.e. a large number of small spatial units) but, for the sake of privacy and/or size concerns, are released only after being spatially aggregated to a coarser resolution (i.e. a smaller number of larger spatial units). The chief example of this process is census data which are collected from every household, but released only at the Enumeration Area or Census Tract level of spatial resolution. When values are averaged over the process of aggregation, variability in the dataset is lost and values of statistics computed at the different resolutions will be different; this change is called the *scale effect*. One also gets different values of statistics depending on how the spatial aggregation occurs; this variability is called the *zoning effect*. The purpose of studying the MAUP is to try to estimate the true values of the statistics at the original level of spatial resolution. Knowing these would allow researchers to attempt to make estimates of the data values using either synthetic spatial data generators like the one described in this thesis or by other techniques.

Many studies of the MAUP have been made using specific datasets and examining various statistics, such as correlations. Although interesting properties have been documented, this approach is ultimately unsatisfactory because researchers have had no control over the various properties of the datasets, all of which could potentially affect the MAUP. This research has focused on the creation of a synthetic spatial dataset generator that can systematically vary means, variances, correlations, spatial autocorrelations and spatial connectivity matrices of variables in order to study their effects on univariate, bivariate, and multivariate statistics.

Even though the MAUP has traditionally been written off as an intractable problem, results from the various experiments described in this thesis indicate that there is a degree of regularity in the behaviour of aggregated statistics that depends on the spatial autocorrelation and configuration of the variable values. If the MAUP can be solved, however, it is clear that it will likely be a complex procedure.

Acknowledgments

The program that created the Voronoi tessellations was written by Jonathan Richard Shewchuk as part of the Archimedes project (parallel Finite Element Methods) and was made available over the World Wide Web. Financial support from SSHRC research grants #SSH 410-94-1736 and #410-97-0274 and doctoral fellowship #SSH 752-97-2107 were crucial in getting the research to where it is today.

I also could not have completed this work without the assistance and guidance of my supervisor, Professor Carl Amrhein. The efforts of Dan Griffith, who took the time to make the numerous suggestions for improvements in the final version of this thesis, are also gratefully acknowledged.

Finally, I wish to dedicate this work to my wife Jeannine, whose love and support made the long hours shorter, and to my son Joshua Morgan, who makes the dark days brighter, and whose appearance on August 22, 1997 opened my eyes to a whole new world.

Table of Contents

1. Introduction....1

2. Literature Review....3
   2.1 Univariate Statistics....3
   2.2 Bivariate and Multivariate Statistics....4
   2.3 Theoretical Work....6
   2.4 References....8

3. Technical Details....10
   3.1 The Spatial Dataset Generator....10
      3.1.1 Introduction....10
      3.1.2 Some Symbols Used in the Derivation....12
      3.1.3 The Dataset Generator....12
      3.1.4 Worked Example....13
   3.2 The Aggregation Model....17
   3.3 Interpretation of the Diagrams....17
   3.4 References....18

4. The Effect of Aggregation on Univariate Statistics....20
   4.1 Summary....20
   4.2 Introduction....20
   4.3 Method....21
   4.4 Results....23
      4.4.1 The effects of aggregation on the variance....23
      4.4.2 The effects of aggregation on the Moran Coefficient....24
      4.4.3 Frequency distributions....26
   4.5 Correlating the change in variance with a spatial statistic....26
   4.6 Comparison of synthetic data to a real dataset....28
   4.7 Conclusions....29
   4.8 Tables....31
   4.9 References....32
   4.10 Figures for Chapter 4....33

5. The Effect of Aggregation on Bivariate Statistics....46
   5.1 Summary....46
   5.2 Introduction....46
   5.3 Method....47
   5.4 Results for fixed Moran Coefficients, varying correlations....49
   5.5 Results for fixed correlations, varying Moran Coefficients....52
   5.6 Discussion....55
   5.7 Conclusions....56
   5.8 References....58
   5.9 Tables....59
   5.10 Figures for Chapter 5....60

6. The Effects of Aggregation on Multivariate Regression Parameters....70
6.1 Summary....70
6.2 Introduction....70
6.3 The synthetic spatial dataset generator....71
6.4 The experiments....72
6.5 Results....73
6.6 Conclusions....78
6.7 References....80
6.8 Figures for Chapter 6....81

7. Summary of conclusions....89

8. Topics for future research....91

List of Tables

4.1a Selected K-S Test Statistics: Variable with Original MC of -0.4 31
4.1b Selected K-S Test Statistics: Variable with Original MC of 1.0 31
4.2a Selected Shapiro-Wilk Statistics: Variable with Original MC of -0.4 31
4.2b Selected Shapiro-Wilk Statistics: Variable with Original MC of 1.0 31
5.1 Variation of the covariance with original MC of the variables and correlations 59
5.2 Variation of the correlation with original MC of the variables and correlations 59
5.3 Summary information for the thirteenth group of distributions in Figure 5.8a 59
6.1 Total number of statistically significant correlations between the variables created by the aggregation process. 80

List of Figures

Figure 4.1: Frequency distributions of three variables generated by the new synthetic dataset 33
Figure 4.2: Variation of relative change in variance RCV (top) and MC with initial MC and 34
Figure 4.3a: Examples of variables with Moran Coefficients of 0.8 (top) and the variograms.....35
Figure 4.3b: Four more variables with MCs of 0.8 with length scales longer than those of.....36
Figure 4.3c: Four more variables with MCs of 0.8, all with longer length scales. Note the lack.....37
Figure 4.3d: The final four variables with MCs of 0.8, all with long length scales. On average, ...38
Figure 4.4a: Variation of the MCs of the variables in Figures 3a to 3d. It can be seen that the 39
Figure 4.4b: Variation of the variances of the variables in figures 3a to 3d. Results here 40
Figure 4.5: Relative change in variance (RCV) as a function of the aggregated MC without the...41
Figure 4.6: Relative change in variance (RCV) as a function of $\log_{10}$(G) (top) and $\log_{10}$(mod 42
Figure 4.7: Behaviour of the Relative Change in Variance with aggregation for the actual 43
Figure 4.8: Behaviour of the aggregated Moran Coefficients for the actual Lancashire dataset.....44
Figure 5.1: Variation of aggregated covariance with initial correlation where dependent and 60
Figure 5.2: Variation of aggregated correlation with initial correlation where dependent and 61
Figure 5.3: Variation of aggregated upper triangle (row is independent, column dependent) of....62
Figure 5.4: Variation of the MC of regression residuals with the original correlation, where 63
Figure 5.5: Variation of covariances (top) and correlations with the (MC independent, MC.....64
Figure 5.5, con't: Variation of upper triangle of regression slope parameters (top) and change.....65
Figure 5.6: Variation of covariance (top) and upper triangle of the matrix of regression slope.....66
Figure 5.6 (con't): The change in the MC_{RR} with the (MC independent, MC dependent).....67
Figure 5.7: Variation of correlation with the (MC independent, MC dependent) variables for 68
Figure 5.8: Variation of correlation for several combinations of variables whose MCs and.....69
Figure 6.1: The synthetic region used in all of the experiments, with its 400 cells (a) and a 81
Figure 6.2: Variation of R^2 (top) and the change of Moran Coefficient of the multivariate 82
Figure 6.3: Variation of the multivariate regression parameter β_0 and its standard error over 83
Figure 6.4: Variation of the multivariate regression parameter β_1 and its standard error over 84
Figure 6.5: Variation of the multivariate R^2 (top) and the change of the Moran Coefficient of.....85
Figure 6.6: Variation of the multivariate regression parameter β_0 (top) and its standard error, ...86
Figure 6.7: Variation of the multivariate regression parameter β_1 (top) and its standard error, ...87
Figure 6.8: Variation of correlation with aggregation for the datasets of experiment 1 in which ...88

1. Introduction

The Modifiable Area Unit Problem (MAUP), a term introduced by Openshaw and Taylor's (1979) classic paper, has long been recognized as a potentially troublesome feature of aggregated data, such as census data. Aggregation of high resolution (i.e. a large number of small areas) data to a lower resolution (i.e. a smaller number of larger areas) is an almost unavoidable feature of large spatial datasets due to the requirements of privacy and/or data manageability. When the original data are aggregated, the values for the various univariate, bivariate, and multivariate parameters will change because of the loss of information. This phenomenon is called the *scale effect*. The M spatial units to which the higher-resolution data are aggregated, such as census enumeration areas or tracts, postal code districts, or political divisions of various levels, are arbitrarily created by some decision-making process and represent only one of an almost infinite number of possible partitionings of the region M ways. Each partitioning will result in different values for the aggregated statistics; this variation in values is known as the *zoning effect.* As will be shown in the following chapters, the statistic values form distributions that are normal or nearly so. The two effects are not independent, because the lower-resolution spatial structure may be built from contiguous higher-resolution units, such as census tracts from enumeration areas, and the resulting aggregate statistics will be different for each possible arrangement of the high-resolution units.

This research is timely and necessary. The increasing availability of powerful microcomputers, workstations, and Geographical Information Systems (GIS) software suggests that undertaking complex spatial analyses is no longer limited to those trained in the vagaries of spatial data. Large numbers of users are blissfully unaware that aggregation effects may cause widespread misuse of results. For example, Openshaw and Taylor (1979) demonstrate that the sign of the correlation between two variables can change, depending on the spatial resolution of the dataset that is used, which means that if the data were to be used to influence a decision in public policy a serious error could be made. The stubborn refusal of this problem to be solved analytically, except for some carefully defined and unrealistic problems (Arbia, 1989) means that, for the moment, the most useful information about the MAUP can only be gleaned through the use of statistical simulations. Ironically, it is the same increase in computing power that makes the extensive simulations performed for this research possible.

The purpose of this research is to shed some light on the behaviour of statistics that are computed with aggregated data by using a set of systematic empirical experiments. It is hoped that the results of these experiments will bring us one step closer to the ultimate goal of being able to accurately estimate the true statistical relationships within datasets that, for reasons of confidentiality, size, or other factors, are only available in aggregated form. Knowing the statistic values would allow researchers to attempt to make estimates of the data values using either synthetic spatial data generators like the one described in this thesis or by other techniques. Until Amrhein (1995), research into the MAUP has primarily consisted of examining the effects of aggregation on various statistics, usually correlations, computed from a single dataset. The primary drawback to this method is that the researcher is unable to vary the properties (such as means, variances, covariances, and spatial autocorrelations) of the particular dataset, somewhat akin to trying to determine the properties of a forest by studying a few trees here and there.

Amrhein's (1995) study, described in more detail in the next chapter, represents an initial, relatively simple, attempt to use synthetic data to study the MAUP by aggregating points into squares. My research required that I extend this process to the ability to control key parameters like means, variances, correlations, and Moran Coefficients of spatial autocorrelation, as well as the ability to generate connectivity matrices by subdividing a region with random Voronoi polygons (Okabe et al., 1992). Systematically varying these parameters permits examination of their influence on the MAUP, while creating synthetic datasets whose parameters are the same as those of a real dataset allows the researcher to ensure that the results obtained are realistic.

The second chapter of this thesis presents a literature review that will helps to define its context. The third chapter consists of a detailed description of the spatial dataset generator, the aggregation model, and instructions on the interpretation of the diagrams. Chapter 4 explores the effects of aggregation on the variance and the Moran Coefficient, and continues earlier efforts to correlate the change in variance to a spatial statistic. Chapter 5 continues this research with analysis of the bivariate statistics covariance, correlation, regression slopes, and the Moran Coefficient of the regression residuals, comparing results to those found in Openshaw and Taylor (1979). Chapter 6 presents the extension of the studies to multivariate regression parameters, comparing the results to those of Fotheringham and Wong (1991). Finally, chapter 7 contains a discussion and summary of the conclusions from the previous three chapters.

2. Literature Review

The Modifiable Area Unit Problem has been recognized in the literature since at least Gehlke and Biehl's (1935) work. Due to its inherent analytical intractability, it has been either downplayed or ignored in various studies using spatial data and in textbooks on spatial analysis. Only within the past 15 years or so with the advent of cheaper, faster, and more powerful computers, has an in-depth examination of the behaviour of the MAUP become possible. The extensive literature can be divided into two broad categories, empirical analyses and theoretical developments. I have not tried to make this literature survey complete, since good survey papers (Openshaw and Taylor, 1981; Dudley, 1991) exist already; rather it is intended to place my work in context of the main body of MAUP research.

2.1. Univariate Statistics

The behaviour of univariate statistics such as mean, variance, and Moran Coefficient (MC) under aggregation has received little attention in the literature, since it is inferences about relations between two or more variables that is the focus of most research involving spatial data. Spatial autocorrelation statistics, however, are often used to test for patterns in a satellite image by landscape ecologists. As these patterns influence ecological processes, such as population dynamics, biogeochemical cycling, and aspects of biodiversity (Qi and Wu, 1996), it is useful to know how the spatial scale of the analysis affects the spatial autocorrelation statistics. This is problematic because the various satellites have different spatial resolutions. Qi and Wu (1996) and Jelinski and Wu (1996) conclude that the Moran Coefficient, Geary Ratio, and Cliff-Ord statistic are scale dependent, showing an overall decline in spatial autocorrelation with scale, and are also dependent on the zoning system used in the aggregation.

Amrhein and Reynolds (1996, 1997) present results based on census datasets from Lancashire in England and from the Greater Toronto Area's enumeration areas respectively. The average variance of the 8 Lancashire variables (all of which were averaged during aggregation) and the 5 Toronto variables (the first three of which were summed and the last two averaged during aggregation) is found to vary systematically with the change in scale. The change in variance is also found to correlate well for all variables in both datasets with the G statistic (Getis and Ord, 1992), which was modified by dividing it by the global sum of squares of deviations of the aggre-

gated variable. The fit is not as good with the fifth variable of the Toronto dataset, which is likely due to the presence of a large number of suppressed (zero) values of the EA average income, but the overall results are good enough to indicate the potential of using a spatial statistic to predict the effect of the MAUP on an aggregated dataset.

Amrhein (1995) is the first paper based solely on statistical simulation of the MAUP. The experiments are based on 10 000 points located randomly within a unit square region, each representing an individual. The x and y coordinates are generated first from a uniform distribution and then from a normal N(0,1) distribution. Each location is assigned two values representing observed variables, with the values again being drawn from first a uniform and then a normal distribution, thus creating four combinations in total. To examine the scale effects, the points are aggregated into 100, 49, and 9 square areal units, and to account for zoning effects, the process of aggregating the 10 000 points into the 100 region grid is repeated for 100 independent sets, and for 50 sets for the other two grids. Summary statistics for each aggregation are computed and stored for comparison purposes with the original "population" statistics. It is found that the weighted mean does not display any aggregation effects, which is to be expected since the aggregate weighted mean is mathematically identical to the population mean. The variance is not found to display scale effects beyond what could be expected from the decrease in observations, though it is noted that scale-specific variance values cannot be imputed to other scales without adjusting for the change in number of units. Populations with higher variances tend to display more pronounced zoning effects than those with a lower variance. The regression slope coefficient and the Pearson correlation coefficient both display scale effects that increase systematically with a decreasing number of zones. The standard deviation of the regression coefficient displays pronounced zoning effects, to the point where it fails to provide useful information. Sign changes of the regression coefficient are also noted. These results provided the starting point for Steel and Holt's (1996) theoretical results.

2.2. Bivariate and Multivariate Statistics

Gehlke and Biehl (1935) appears to be the first publication cited that describes an interesting phenomenon, the tendency for correlation coefficients to increase as areal regions are aggregated into fewer numbers of larger regions. When male juvenile delinquency was correlated with

median equivalent monthly rental, the correlation coefficient varied monotonically from-0.502 for 252 census tracts to -0.763 for 25 regions; delinquency rates varied non-monotonically from -0.516 to -0.621. Two other experiments were also performed that illustrated that the method of grouping also affected the aggregated correlation.

Robinson (1950) examined correlations between race and illiteracy at the U.S. Census Division (0.946), state (0.773) and individual (0.203) levels, and foreign birth and illiteracy at the Census Division (-0.619), state (-0.526) and individual (0.118), but it should be noted that he uses data that appear in contingency tables rather than the more usual x-y point data. He also describes a mathematical relationship between his "ecological" correlations and individual correlations and asserts (correctly) that one should not use conclusions derived from data at one level of spatial resolution to units at another resolution (primarily individuals). A possible solution to the contingency tables type problem is described in King (1997).

Clark and Avery (1976) looked at correlations derived from data collected from 1596 census tracts, and correlations from a survey of households, both from the Los Angeles area. They found a systematic increase in the correlation coefficients (and systematic changes in other bivariate statistics) as the number of aggregated units decreased, except for a slight decrease in the fifth level of aggregation from the value at the fourth level. They also conclude that their results do not agree with a hypothesis by Blalock (1964) that changes in the slope coefficient are explained by the reduction in variation of the independent or dependent variable, but instead could be related directly to how covariation changes with aggregation, and independently on the spatial autocorrelation of the micro- and macrolevel data.

Openshaw and Taylor (1979) are credited with introducing the term Modifiable Area Unit Problem. They use a dataset of percentage voters for Republicans in the 1968 congressional elections as a dependent variable and the percentage of population over sixty as recorded in the 1970 US census over the 99 counties of Iowa to examine the effect of the MAUP on bivariate correlation coefficients. Ten thousand aggregations are performed at each of twelve different spatial scales, ranging from six to 72 areal units, and the correlation coefficients are computed. These aggregations are performed with two separate algorithms, one that requires spatial contiguity and one that does not. As illustrated by their Table 5.2, they find that the range of correlation coefficients becomes broader as the number of zones decreases, to the point where all possible values

for the coefficient are computed for the six and twelve zone groups, and even for the 48 zones in the non-contiguous aggregations the range is from -0.967 to 0.995. No relation is found between the correlation coefficient and the relative loss of variation (original - aggregate variance)/(original variance) of the independent variable, though there is a systematic trend in of the loss of variation with scale. They also show that the interaction between spatial autocorrelation and the contiguous zoning procedure directly affects the resulting statistics.

Fotheringham and Wong (1991) present the results of an analysis of the effects of aggregation on linear regression and logit models constructed from an 871 block group census dataset for the Buffalo Metropolitan Area. The models have four independent and one dependent variables, and all variables are proportions in which the numerator and denominator are aggregated separately and divided after aggregation. This may have affected the results because each number is the combination of two others, both of which are likely affected differently by the MAUP. A systematic variation of the parameters for both models with scale is found, with some becoming more negative and others more positive as the scale (i.e. the number of zones) decreases. To one degree or another, all show an increase in variation of values (and the standard errors of the parameters) with the decrease in scale. In an attempt to link the changes to spatial autocorrelation, the variation of the Moran Coefficient of the variables with aggregation is examined. Four of the five have curves that are approximately normal in shape, with the highest values in the intermediate levels of aggregation. This differs significantly from my results as shown in Figure 4.2 and in Reynolds and Amrhein (1998a), and may be due to the nature of the proportion variable that contains an implicit interaction between the spatial properties of two variables that are summed during aggregation. The coefficient of determination R^2 is found to increase significantly with the decrease in scale, which again differs from my results (Reynolds and Amrhein, 1996). Overall, Fotheringham and Wong are pessimistic about ever being able to deal with the MAUP in multivariate analysis. Again, my preliminary results indicate that this pessimism is probably unfounded.

2.3. Theoretical Work

The theoretical side of the research is represented in this review by three papers. Steel and Holt (1996) present a list of “rules” for random aggregation as a summary of their results, based

on the assumption that the groups are formed at random and that there is no association between the variate values and group membership. They are listed as follows.

(1) The expectations of weighted group-levels statistics are not affected by aggregation. Thus any observed change, as we change boundaries or scale, is caused by random variation.

(2) The variance of weighted group-levels statistics is determined mainly by the number of groups in the analysis. If the number of groups is small, this variation will be high and the likely range will be so large that in many cases useful inferences will not be possible.

(3) Valid confidence intervals and hypothesis tests can be obtained by means of weighted group-level statistics. Even if the unit-level distribution is nonnormal, the analysis of weighted group-level statistics can proceed with procedures associated with the normal distribution, provided that the sample size within groups is not very small.

(4) Unweighted statistics have the same expectation as their weighted counterparts, but larger variances. Unless the variation in group population sizes is small, standard confidence intervals will have less than the required coverage.

Holt et al. (1996) propose statistical models whose purpose is to explain the aggregation effect in populations composed of geographic groups. They conclude that the aggregation effects depend upon the sample sizes upon which the area means are based, the number of areas used in the analysis, and the strength of intra-area homogeneity on both variances and covariances for the variables of interest. Auxiliary variables are introduced that explain much of the intra-area homogeneity, which leads to a decomposition of the aggregation bias into two components, one attributed to a set of grouping variables and the other to a residual source of aggregation bias conditional on the grouping variables. With some information about the individual level covariance matrix of the grouping variables, it is believed that an adjustment can be made to eliminate the first component of the aggregation bias.

Steel, Holt, and Tranmer (1996) use the same model as Holt et al. (1996), but present a strategy for identifying adjustment variables for which an estimate of the unit-level covariance matrix is available and that account for group effects. First, one must identify a set of variables that covers the same subject area as the variables of interest, but for which both area level and unit level data are available from the past, such as previous census data. Variables (such as housing variables in their example) that are known to be strongly associated with areal differences can be added to this set, so long as estimates of both of the area and unit level covariance matrices are

available. A Canonical Grouping Variable analysis can then be carried out to identify the variables that load most strongly onto the most important CGVs. Finally, a set of adjustment variables from the CGV analysis that is available within the current dataset and for which the unit level co-variance matrix is available needs to be identified. These variables can then be used to adjust the aggregate analysis for the variables of interest.

This brief survey of the extensive literature, as well as the more comprehensive surveys by Dudley (1991) and Openshaw and Taylor (1981), indicate that little use has been made of numerical simulations in the study of the MAUP, primarily due to the computationally intensive nature of the simulations. The dataset generator and aggregation models described in Chapter 3 are a first step towards rectifying this deficiency.

2.4. References

Amrhein, C. G., 1993: Searching for the elusive aggregation effect: Evidence from statistical simulations. *Environment and Planning A*, **27**, 105-119.

Amrhein, C. G., and H. Reynolds, 1996: Using spatial statistics to assess aggregation effects. *Geographical Systems*, **2**, 83-101.

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..

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

Dudley, G., 1991: Scale, aggregation, and the modifiable area unit problem. *The Operational Geographer*, **9(3)**, 28-33.

Fotheringham, A. S., and D. W. S. Wong, 1991: The modifiable area unit problem in multivariate analysis. *Environment 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.

Holt, D., D. G. Steel, M. Tranmer, and N. Wrigley, 1996: Aggregation and ecological effects in geographically based data. *Geographical Analysis*, **28**, 244-261.

King, G., 1997: *A Solution to the Ecological Inference Problem*. (Princeton: Princeton University Press).

Okabe, A., B. Boots, and K. Sugihara, 1992: Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. (London: Wiley)

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, (London: Pion), 127-144.

Openshaw, S., and P. Taylor, 1981: The modifiable area unit problem, in *Quantitative Geography: A British View*, N. Wrigley, ed., (London: Routledge and Regan Paul).

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.

Steel, D. G., and D. Holt, 1996: Rules for random aggregation. *Env. and Planning A*, **28**, 957-978.

Steel, D. G., D. Holt, and M. Tranmer, 1996: Making unit-level inferences from aggregated data. Submitted to *Survey Methodology*.

3. Technical Details

This chapter describes the spatial dataset generator, the aggregation model, and the output diagrams in detail. It replaces technical descriptions that were present to varying degrees in the three papers that form the next three chapters.

3.1. The Spatial Dataset Generator

3.1.1. Introduction

The need for a systematic study of the effects of the MAUP on summary statistics is clear. The literature, some of which is discussed in the previous chapter, contains many case studies of the effects of aggregation on various statistics using a single dataset for each study. Each set comes with its own connectivity matrix and the variables have parameter values that are totally out of the control of the researcher. A researcher reviewing the literature is likely to wonder if the results found from dataset X will be replicatable for dataset Y, even though the initial correlations (for example) of the variables are completely different. Furthermore, many papers, such as Clark and Avery (1976), discuss the possible effects of spatial autocorrelation on their results in passing, but since they have no control over it, little more than speculation can be stated. To date, there has been no attempt to systematically vary the dataset parameters in order to test their effects on the aggregated statistics, and it is this deficiency that my research is redressing.

The method of generating synthetic spatial datasets discussed below is chosen because it allows the user to create a set of variables with specific levels of spatial autocorrelation (as measured by the Moran Coefficient) *and* Pearson correlations exactly and directly, as opposed to other methods that take a set of existing values and rearranges them. Control over the spatial autocorrelation of the variables is a requirement for my research, as it plays an important role in the effect of spatial aggregation on statistics[1], while control over Pearson correlations was required for the bivariate and multivariate experiments. Other methods of generating spatial data, such as the turning band method (see for example Bras and Rodriguez-Iturbe, 1985), work with only one variable at a time and make the data fit to a particular type of variogram (Journel and Huijbregts,

[1] A highly spatially autocorrelated variable will tend to suffer less from aggregation than one that is randomly or negatively autocorrelated because the observations that are aggregated tend to be similar to one another, hence less information (i.e. variability) is lost. Section 6.4 discusses this in more detail.

1978, p. 12), but this is not satisfactory because it is advantageous for this research to deal with a single number rather than a graph when attempting to describe spatial organization and link it to the behaviour of statistics under aggregation, and it is not intuitive how to link a variogram to a specific level of spatial autocorrelation. Using one of these methods also works on only one variable at a time, making the specification of correlations between them difficult.

The Moran Coefficient (MC) is a convenient tool for measuring spatial autocorrelation in discretized surfaces, and for the purposes of this research it is also convenient for generating variables with specific levels of autocorrelation. It is, however, a first-order spatial statistic, since it only deals with immediate neighbours to a cell, and this, among other things, means that it is not unique. That is, many different spatial arrangements of a set of numbers can produce similar or equal values of the MC. The data generation algorithm discussed below unfortunately lacks the ability to select a desired type of spatial arrangement (or even a specific one). This poses a minor problem, as the research shows that the arrangement of the values, especially for higher levels of spatial autocorrelation, affects the behaviour of the MC and the various bivariate statistics and interferes with the ability to draw highly general conclusions about their behaviour under aggregation. As the conclusions drawn are no less valid for this lack of control, a more systematic attempt to study the effects of spatial arrangement on the behaviour of moderately to strongly autocorrelated variables under aggregation can be postponed as a topic for future research. Since the generator is capable of producing a variety of spatial arrangements, it may be possible to modify it in the future to control just which arrangement it produces. This weakness does, unfortunately, make the dataset generator unhelpful in efforts to simulate real-world datasets, since it is very often the arrangement of the values that is as much of interest as the values themselves.

Each synthetic variable created is a linear combination of eigenfunctions of the connectivity matrix, making control of the resulting frequency distribution not possible with the current algorithm. The distributions are mound-shaped and unimodal, but not necessarily normal (see Figure 4.1 for examples). Certain combinations of MC and Pearson correlation are also found to be incompatible, such as two variables with widely differing MCs but a high level of correlation. This is reasonable because if the two variables were highly correlated then one would expect their spatial arrangements to be similar, something which is not possible with widely differing MCs. The requirement that the covariance matrix be positive definite, which it must be by definition,

makes it difficult to create a large number of combinations of MCs and negative correlations. Finally, although it is theoretically possible to create spatial datasets of any size, the effort required to compute and decompose $\mathbf{MC_SM}$ (defined below) increases extremely rapidly with size. These drawbacks and restrictions aside, the spatial dataset generator has proven to be a useful tool for this preliminary empirical research into the effects of aggregation on statistics.

3.1.2. Some Symbols Used in the Derivation

The derivation of the method used to generated geo-referenced data uses the following symbols:

n = number of zones in a geo-referenced dataset

p = number of variables in a geo-referenced dataset

$\mathbf{M} = \mathbf{I}-\mathbf{11}^T/n$ is a projection matrix commonly found in statistics and is used for the matrix equivalent of sum of squares of deviations from the mean.

C = the binary spatial connectivity matrix of the region, where $c_{ij}=1$ if region i is next to region j, otherwise $c_{ij}=0$. Most of the experiments are performed using an irregular ten-sided convex polygon illustrated in Figures 4.3 and 6.1 that is divided into 400 random Voronoi polygons. Some experiments in Chapter 4 are performed on a square region of dimension 20.

$\mathbf{C}_S = \frac{\mathbf{1}^T\mathbf{1}}{\mathbf{1}^T\mathbf{C}\mathbf{1}}\mathbf{C}$, the scaled connectivity matrix, used in computing the Moran Coefficient

$\mathbf{\Sigma_1}$ = the covariance matrix of the intermediate variables **V**

$\mathbf{\Sigma_2}$ = the desired covariance matrix of the final variables **X**

V = matrix of intermediate variables $\mathbf{v_i}$

A = scaling matrix

X = matrix of variables with desired properties $\mathbf{x_i}$; $\mathbf{X}=\mathbf{VA}$.

3.1.3. The Dataset Generator

Their aspatial nature makes setting means, variances, covariances, and correlations of variables to prespecified values a relatively simple task, as follows. Suppose a set of p variables **V**, each with n observations, is postmultiplied by a p×p matrix **A** to form $\mathbf{X} = \mathbf{VA}$. It is easy to show that the covariance matrix of **X** is $\mathbf{\Sigma_2} = \mathbf{A}^T\mathbf{\Sigma_1}\mathbf{A}$. To solve for **A**, define $\mathbf{\Sigma_1} = \mathbf{B}^T\mathbf{B}$ and $\mathbf{\Sigma_2} = \mathbf{D}^T\mathbf{D}$, i.e. find the Cholesky decompositions of the covariance matrices. It quickly follows that **A**

$= \mathbf{B}^{-1}\mathbf{D}$. Changing a variable's mean requires nothing more than adding $(\mu_2-\mu_1)$ to each observation, where μ_1 is the current mean and μ_2 is the required mean. To change a single variable's variance, each observation must be multiplied by σ_2/σ_1, where σ_1 is the current standard deviation and σ_2 the desired one.

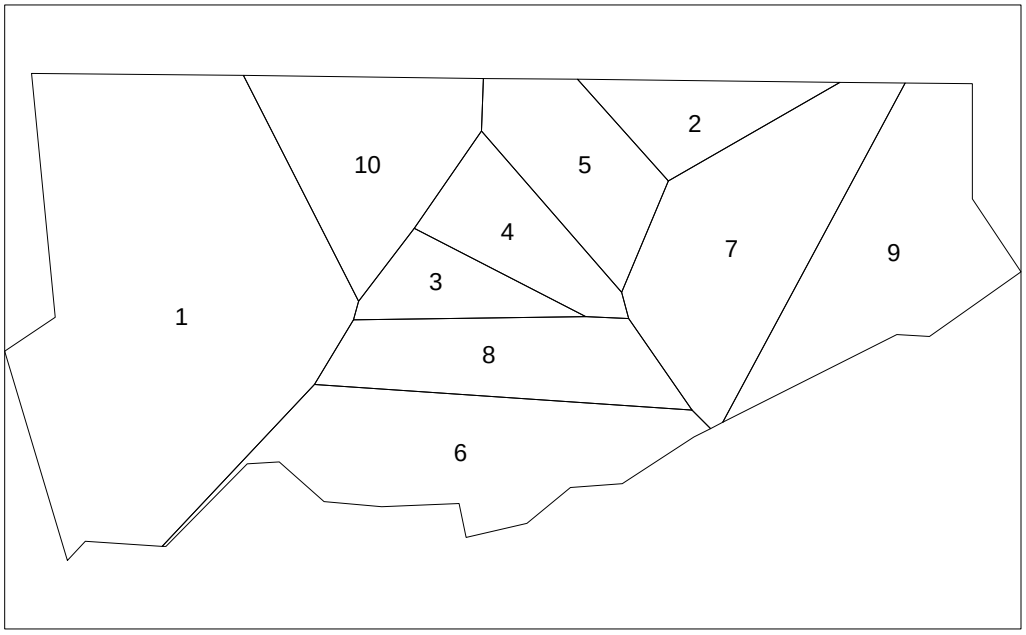
Unfortunately, the Moran Coefficient is not as readily bent to our will. Written in matrix notation, its formula is $MC(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{M}\mathbf{C}_S\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{M}\mathbf{x}}$. There is no simple general way to represent the MC of a variable that is a linear combination of two or more other variables as a function of the MCs of these variables. Suppose, however, that we compute the eigensystem of $\mathbf{M}\mathbf{C}_S\mathbf{M} = \mathbf{E}\Lambda\mathbf{E}^T$, where $\mathbf{E}$ is the matrix of eigenvectors and Λ is a matrix with the diagonal elements equal to the eigenvalues and the rest zero. Hence we can rewrite the formula for the Moran Coefficient: $MC(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{E}\Lambda\mathbf{E}^T\mathbf{x}}{\mathbf{x}^T\mathbf{M}\mathbf{x}}$ (Tiefelsdorf and Boots, 1995; Griffith, 1996). Let $\mathbf{x}$ be one of the eigenvectors $\mathbf{e_i}$. By definition, the eigenvectors are all orthonormal, so that $\mathbf{e_i}^T\mathbf{E}\Lambda\mathbf{E}^T\mathbf{e_i}$ reduces to λ_i and $\mathbf{e_i}^T\mathbf{M}\mathbf{e_i}$ reduces to one. Hence, the Moran Coefficient of an eigenvector of $\mathbf{M}\mathbf{C}_S\mathbf{M}$ is just its corresponding eigenvalue. Using similar arguments, it can be shown that the MC of a linear combination of eigenvectors $\mathbf{y} = a\mathbf{e_i} + b\mathbf{e_j} + c\mathbf{e_k} + \ldots$ is $MC(\mathbf{y}) = \frac{a^2\lambda_i + b^2\lambda_j + c^2\lambda_k + \cdots}{a^2 + b^2 + c^2 + \cdots}$. Thus, the key to creating variables with specified Moran Coefficients lies in selecting appropriate linear combinations of the eigenvectors of $\mathbf{M}\mathbf{C}_S\mathbf{M}$.

3.1.4. Worked Example

The detailed description of the method below includes a worked example for the set of regions illustrated in the diagram on the next page. The desired values of statistics are:

Variable	Mean	Variance	Moran Coef	Correlations				
1	20	6	0.4	1.0	-0.6	0.4	-0.4	-0.8
2	20	6	0.2	-0.6	1.0	0.0	0.8	0.6
3	20	6	-0.2	0.4	0.0	1.0	-0.2	0.2
4	20	6	0.0	-0.4	0.8	-0.2	1.0	0.3
5	20	6	0.13	-0.8	0.6	0.2	0.3	1.0

The diagram of the region (a random Voronoi tessellation of Metro Toronto) is below.



1. Compute the eigensystem of $\mathbf{MC_SM}$.

Eigenvalues

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
-0.5263	-0.5263	-0.4649	-0.3942	-0.1166	0.0000	0.0540	0.0770	0.3796	0.5177

Eigenvectors

$\mathbf{e_1}$	$\mathbf{e_2}$	$\mathbf{e_3}$	$\mathbf{e_4}$	$\mathbf{e_5}$	$\mathbf{e_6}$	$\mathbf{e_7}$	$\mathbf{e_8}$	$\mathbf{e_9}$	$\mathbf{e_{10}}$
-0.3271	0.4228	-0.1569	-0.4261	0.0060	-0.3162	0.0253	0.4405	-0.1371	0.4411
-0.1319	-0.2147	-0.2663	0.1397	-0.5951	-0.3162	-0.2390	0.3868	0.0501	-0.4274
0.1319	0.2147	0.4576	0.3879	-0.4777	-0.3162	0.1249	-0.1477	0.2024	0.4125
-0.5909	-0.0066	-0.2787	0.0991	0.1440	-0.3162	-0.1274	-0.5740	0.3109	0.0133
0.0000	0.0000	0.6121	-0.3530	0.2670	-0.3162	-0.2223	0.1272	0.3747	-0.3513
-0.1319	-0.2147	0.1923	0.5045	0.4133	-0.3162	-0.2474	0.1864	-0.5183	0.0983
0.3957	0.6441	-0.2047	0.0708	0.0851	-0.3162	-0.0564	-0.2378	-0.2468	-0.3921
0.3271	-0.4228	-0.0411	-0.4733	-0.1860	-0.3162	-0.2494	-0.3720	-0.3082	0.2418
-0.1319	-0.2147	0.0802	-0.0959	0.0073	-0.3162	0.8442	-0.0478	-0.2096	-0.2488
0.4590	-0.2081	-0.3946	0.1463	0.3360	-0.3162	0.1474	0.2384	0.4819	0.2126

2. One can create the covariance matrix $\mathbf{\Sigma_1}$ by placing the variance of $\mathbf{e_2}$ on the diagonal of a p×p matrix, where p is the number of variables. This can be done because the eigenvectors are all uncorrelated, as well as orthonormal. We must do this step because we need to compute the scaling matrix **A** so that the needed values of the MCs can be calculated in Step 4.

Diagonal of Σ_1	0.1000	0.1000	0.1000	0.1000	0.1000

3. Next one can create the scaling matrix $\mathbf{A} = \mathbf{B}^{-1}\mathbf{D}$, where **B** and **D** are the Cholesky decompositions of Σ_1 and Σ_2 respectively.

7.746	-4.6476	3.0984	-3.0984	-6.1968
0	6.1968	2.3238	5.4222	1.1619
0	0	6.7082	-2.2361	4.2485
0	0	0	4	0.5000
0	0	0	0	1.3964

4. Compute the MCs that each variable $\mathbf{v_i}$ must have in order for the equivalent $\mathbf{x_i}$ to have the desired MC. This must be done because multiplying **VA** will change the MCs for all but the first variable. The procedure is as follows. Recalling that **X** and **A** are composed of p vectors of length n, write $\mathbf{X} = \mathbf{VA} \Rightarrow (\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}) = (\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3}, \mathbf{v_4})\mathbf{A}$. Using the upper-triangular form of A to simplify, we get $\left\{ \begin{aligned} \mathbf{x_1} &= a_{11}\mathbf{v_1} \\ \mathbf{x_2} &= a_{21}\mathbf{v_1} + a_{22}\mathbf{v_2} \\ \mathbf{x_3} &= a_{31}\mathbf{v_1} + a_{32}\mathbf{v_2} + a_{33}\mathbf{v_3} \\ \mathbf{x_4} &= a_{41}\mathbf{v_1} + a_{42}\mathbf{v_2} + a_{43}\mathbf{v_3} + a_{44}\mathbf{v_4} \end{aligned} \right\}$.

Since the $\mathbf{v_j}$ are eigenvectors, the MCs of the $\mathbf{x_j}$ are, using the relation previously defined,

$$M_1 = \lambda_1$$
$$M_2 = \left(a_{12}^2\lambda_1 + a_{22}^2\lambda_2\right) / \left(a_{12}^2 + a_{22}^2\right)$$
$$M_3 = \left(a_{13}^2\lambda_1 + a_{23}^2\lambda_2 + a_{33}^2\lambda_3\right) / \left(a_{13}^2 + a_{23}^2 + a_{33}^2\right)$$
$$M_4 = \left(a_{14}^2\lambda_1 + a_{24}^2\lambda_2 + a_{34}^2\lambda_3 + a_{44}^2\lambda_4\right) / \left(a_{14}^2 + a_{24}^2 + a_{34}^2 + a_{44}^2\right)$$

where M_j is the Moran Coefficient for variable j, and λ_j is the MC which $\mathbf{v_j}$ must have so that $\mathbf{x_j}$ will have the MC that is desired. Solving for λ_j gives:

$$\lambda_1 = M_1$$
$$\lambda_2 = \left[M_2\left(a_{12}^2 + a_{22}^2\right) - a_{12}^2\lambda_1\right] / a_{22}^2$$
$$\lambda_3 = \left[M_3\left(a_{13}^2 + a_{23}^2 + a_{33}^2\right) - \left(a_{13}^2\lambda_1 + a_{23}^2\lambda_2\right)\right] / a_{33}^2$$
$$\lambda_4 = \left[M_4\left(a_{14}^2 + a_{24}^2 + a_{34}^2 + a_{44}^2\right) - \left(a_{14}^2\lambda_1 + a_{24}^2\lambda_2 + a_{34}^2\lambda_3\right)\right] / a_{44}^2$$
$$\lambda_j = \left[M_j \sum_{i=1}^{j} a_{ij}^2 - \sum_{i=1}^{j-1} a_{ij}^2\lambda_i\right] / a_{jj}^2$$

As can be seen, the required MC for variable j depends on the values of the MCs of the previous variables. If a value exceeds the bounds $\lambda_1 \le MC \le \lambda_n$, it means that the desired MC is not attainable with the current configuration of correlations and MCs.

Variable	1	2	3	4	5
Required MC	0.4000	0.0875	-0.3625	-0.2875	-0.5263

5. Randomly select the eigenvalues λ_{1i} and λ_{2i} that bracket each of the required MCs. Select the value of b from a uniform random distribution and compute the required value of a using the formula $a^2 = \left(\frac{\lambda_2 - MC}{MC - \lambda_1}\right) b^2$ (hence the need for the MC to be bracketed by the eigenvalues).

	Lower eigenvalue		Upper Eigenvalue			
Required MC	Index	Value	Index	Value	a	b
0.4000	7	0.0540	10	0.5177	0.2968	0.5088
0.0870	3	-0.4649	9	0.3796	0.7037	0.9676
-0.3620	2	-0.5263	5	-0.1166	0.7974	0.6509
-0.2870	4	-0.3942	8	0.0770	1.5589	0.8435
-0.5260	1	-0.5263	1	-0.5263	-1.0000	0.8027

6. Create the variables $\mathbf{v_i}$ using $\mathbf{v_i} = a\mathbf{e_{li}} + b\mathbf{e_{ui}}$, where $\mathbf{e_{li}}$ is the eigenvector of the lower eigenvalue and $\mathbf{e_{ui}}$ is that of the upper eigenvalue. Scale the $\mathbf{v_i}$ so that their variances match the variance of $\mathbf{e_2}$.

Zone	$\mathbf{v_1}$	$\mathbf{v_2}$	$\mathbf{v_3}$	$\mathbf{v_4}$	$\mathbf{v_5}$
1	0.3938	-0.2032	0.3313	-0.1651	-0.3271
2	-0.4896	-0.1161	-0.5427	0.3070	-0.1319
3	0.4192	0.4328	-0.1357	0.2708	0.1319
4	-0.0527	0.0875	0.0859	-0.1860	-0.5909
5	-0.4154	0.6631	0.1689	-0.2499	0.0000
6	-0.0397	-0.3061	0.0950	0.5324	-0.1319
7	-0.3672	-0.3200	0.5528	-0.0509	0.3957
8	0.0832	-0.2734	-0.4451	-0.5933	0.3271
9	0.2104	-0.1223	-0.1617	-0.1071	-0.1319
10	0.2579	0.1577	0.0513	0.2422	0.4590

7. Compute $\mathbf{X} = \mathbf{VA}$ and shift the values of the $\mathbf{x_j}$ so that their means equal the desired means. This is done by adding the difference between the desired mean and the current mean to each observation of $\mathbf{x_j}$.

Zone	x_1	x_2	x_3	x_4	x_5
1	23.0506	16.9106	22.9705	16.2769	18.1917
2	16.2078	21.5560	14.5732	23.3288	20.5627
3	23.2474	20.7334	21.3941	22.4346	17.6478
4	19.5917	20.7871	20.6165	19.7014	19.8752
5	16.7820	26.0396	21.3864	23.5052	23.9373
6	19.6924	18.2880	19.8033	20.3804	20.3762
7	17.1560	19.7235	21.8268	17.9628	24.7789
8	20.6446	17.9192	16.6368	16.8820	17.4359
9	21.6296	18.2641	19.2829	18.6180	17.6295
10	21.9978	19.7785	21.5095	20.9099	19.5649

3.2. The Aggregation Model

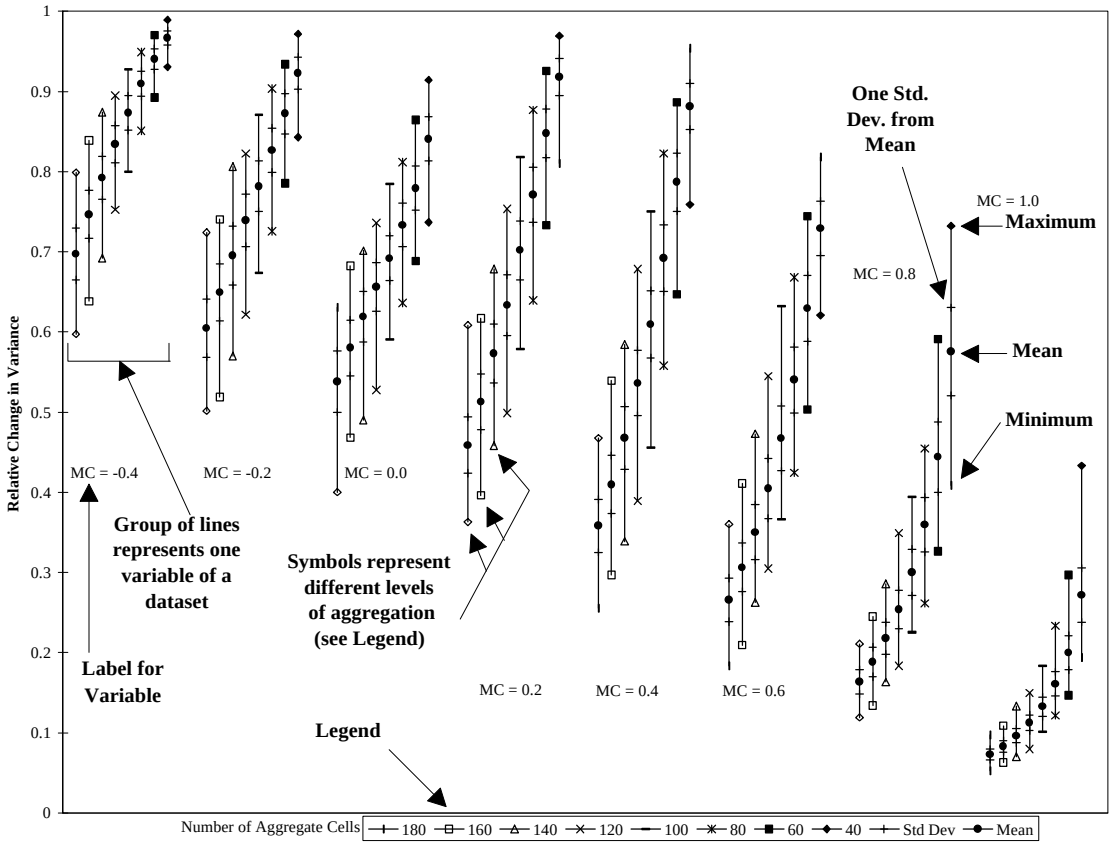
Because nearly all spatial aggregations are performed by aggregating a number of contiguous spatial units into one unit, the aggregation program does the same. An aggregation is initiated by the random selection of M seed regions from the N regions of the spatial dataset, which are copied into an array of "just aggregated" regions. In each pass of the routine, the neighbours of all of the recently aggregated regions are examined. Any neighbour that borders only one of the expanding cells automatically becomes a member of the new cell, while any neighbour that borders more than one cell is assigned to that cell currently having the fewest regions, in an attempt to keep the number of regions per cell as equal as possible. In either case, the region is added to the "just aggregated" region list for the next pass. Aggregation passes continue until no more free regions remain. The assignment process for region j consists of setting element j of an index array to the identifier of the seed region around which the cell is built. The new connectivity matrix is built by looking at the neighbours of the regions within each cell. The cell IDs of those neighbours that are outside the cell are added to the new neighbours list. The new cells are then renumbered, the cell averages are computed, and the various statistics are computed using these average values, and then are stored.

One "run" of the model consists of a set of eight independent aggregations, one to each of 40%, 35%, ..., 10% of the original number of cells. One "experiment" consists of 1000 runs performed on a given dataset. The 1000 values of each statistic for each level of aggregation are processed to produce the mean, standard deviation, maximum and minimum values that are used

to plot the summary diagrams (see below). Each distribution is also tested for normality using both the Kolmogorov-Smirnov and Shapiro-Wilk test statistics.

3.3. Interpretation of the Diagrams

Consider the sample diagram below, which is a replica of Figure 4.2a. All figures consist of sets of eight lines, where each set is based on the results for a particular variable, or in the case of the bivariate and multivariate experiments, a pair of variables. Each line in a set represents a distribution of statistic values for a given aggregation level as indicated in the legend at the bottom of the figure. Each line is marked with the extremes of the distribution (a symbol keyed to the level of aggregation), the mean (a heavy dot), and the mean plus and minus one standard deviation (small horizontal lines), included to give an idea of the shape of the distribution. The standard deviation is chosen instead of the interquartile range that is used in the more standard box plots because it requires less effort to compute, it encloses more values, and the diagrams are also often so dense that a box plot would make them even harder to read.



Each set of lines is labeled according to the nature of the experiment, either with the Moran Coefficient(s) of the variable(s), or initial correlation of the variables in some of the bivariate experiments. This format is chosen because it allows a lot of information to be displayed compactly yet legibly, an important feature given the very large volumes of numbers the model produces. It would not be feasible to use three-dimensional plots, as it would be difficult to plot all of this information legibly, especially for comparing results over different levels of aggregation.

3.4. References

Bras, R. L., and I. Rodriguez-Iturbe, 1985. *Random Functions and Hydrology*. (Reading, Mass: Addison-Wesley), pp. 310-314.

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

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

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