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

0-612-35299-4

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

List of Tables

List of Figures

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 aggregated 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 covariance 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

[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.

Huijbregts, 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.

$\mathbf{C}$ = 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.

$\mathbf{C_S} = \frac{\mathbf{1}^T\mathbf{1}}{\mathbf{1}^T\mathbf{C1}}\mathbf{C}$, the scaled connectivity matrix, used in computing the Moran Coefficient

Σ_1 = the covariance matrix of the intermediate variables $\mathbf{V}$

Σ_2 = the desired covariance matrix of the final variables $\mathbf{X}$

$\mathbf{V}$ = matrix of intermediate variables $\mathbf{v}_i$

$\mathbf{A}$ = scaling matrix

$\mathbf{X}$ = 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 $\mathbf{V}$, each with n observations, is postmultiplied by a p×p matrix $\mathbf{A}$ to form $\mathbf{X} = \mathbf{VA}$. It is easy to show that the covariance matrix of $\mathbf{X}$ is $\Sigma_2 = \mathbf{A}^T\Sigma_1\mathbf{A}$. To solve for $\mathbf{A}$, define $\Sigma_1 = \mathbf{B}^T\mathbf{B}$ and $\Sigma_2 = \mathbf{D}^T\mathbf{D}$, i.e. find the Cholesky decompositions of the covariance matrices. It quickly follows that $\mathbf{A}$

$= \mathbf{B}^{-1}\mathbf{D}$. Changing a variable's mean requires nothing more than adding (μ_2-μ_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.

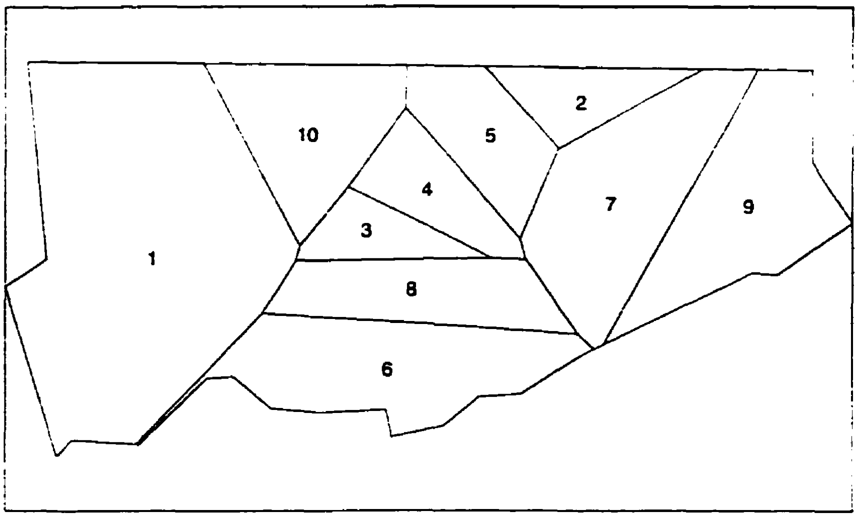
Unfortunately, the Moran Coefficient is not as readily bent to our will. Written in matrix notation, its formula is $\mathrm{MC}(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{MC_S Mx}}{\mathbf{x}^T\mathbf{Mx}}$. There is no simple general way to represent the MC of a variable that is a linear combination of two or more other variables as a function of the MCs of these variables. Suppose, however, that we compute the eigensystem of $\mathbf{MC_SM} = \mathbf{E\Lambda E}^T$, 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: $\mathrm{MC}(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{E\Lambda E}^T\mathbf{x}}{\mathbf{x}^T\mathbf{Mx}}$ (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 $\mathbf{e}_i^T\mathbf{E\Lambda E}^T\mathbf{e}_i$ reduces to λ_i and $\mathbf{e}_i^T\mathbf{Me}_i$ reduces to one. Hence, the Moran Coefficient of an eigenvector of $\mathbf{MC_SM}$ 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 $\mathrm{MC}(\mathbf{y}) = \frac{a^2\lambda_i + b^2\lambda_j + c^2\lambda_k + \cdots}{a^2 + b^2 + c^2 + \cdots}$. Thus, the key to creating variables with specified Moran Coefficients lies in selecting appropriate linear combinations of the eigenvectors of $\mathbf{MC_SM}$.

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

e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	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 $\left\{\begin{matrix} \mathbf{x}_1 = a_{11}\mathbf{v}_1 \\ \mathbf{x}_2 = a_{21}\mathbf{v}_1 + a_{22}\mathbf{v}_2 \\ \mathbf{x}_3 = a_{31}\mathbf{v}_1 + a_{32}\mathbf{v}_2 + a_{33}\mathbf{v}_3 \\ \mathbf{x}_4 = a_{41}\mathbf{v}_1 + a_{42}\mathbf{v}_2 + a_{43}\mathbf{v}_3 + a_{44}\mathbf{v}_4 \end{matrix}\right\}$.

Since the $\mathbf{v}_j$ are eigenvectors, the MCs of the $\mathbf{x}_j$ are, using the relation previously defined,

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

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:

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

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 $a^2 = \left(\dfrac{\lambda_2 - MC}{MC - \lambda_1}\right) b^2$ (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	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
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

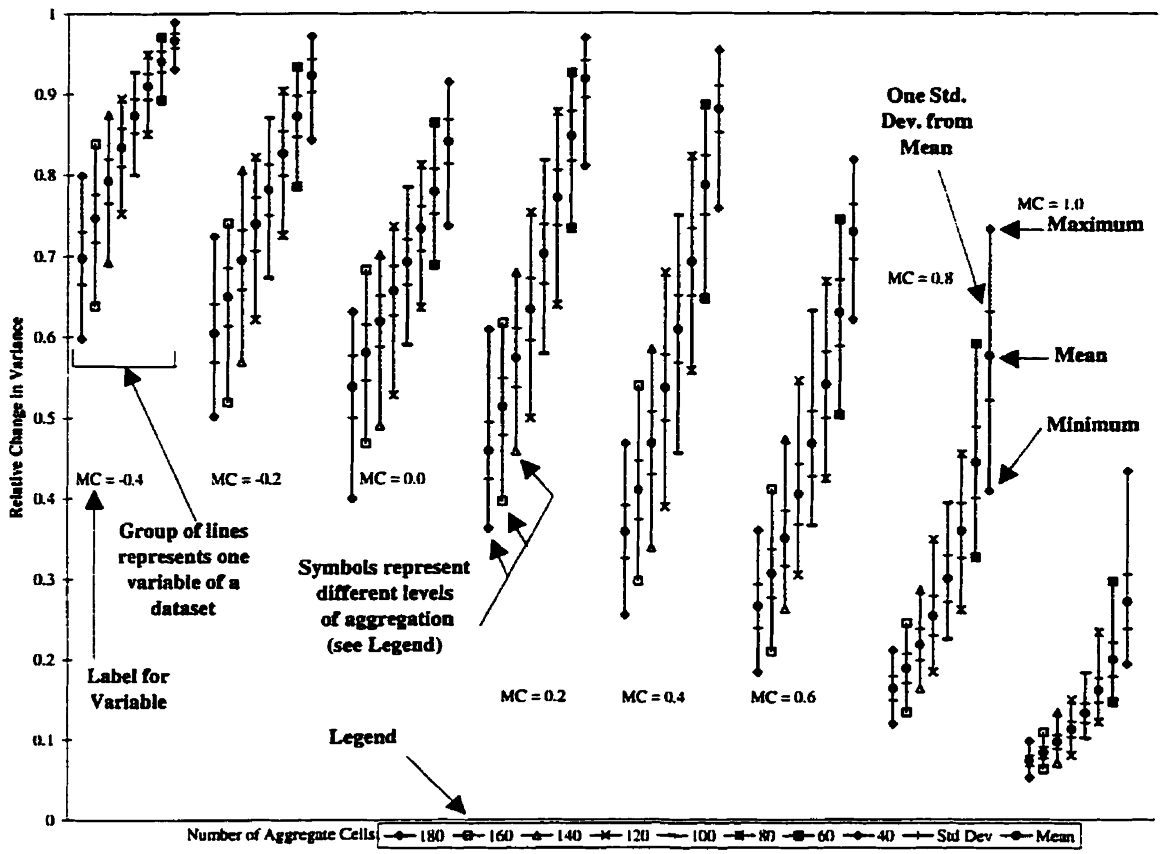
3.2. The Aggregation Model

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.

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 spatial 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 dataset 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:

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

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

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

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,

---

[2] Described in detail in Chapter 3.

rather than 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 $\sigma_{agg}^2 = \frac{1}{N}\sum_{i=1}^{M} n_i\left(x_i - \bar{x}\right)^2$ 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 $2\hat{\gamma}(h) = \frac{1}{N(h)} \sum_{N(h)} \left(Z(s_i) - Z(s_j)\right)^2$ (Cressie, 1993, p. 69), where h is the Euclidean distance between the points $s_i=(x_i, y_i)$ and $s_j=(x_j, y_j)$ and Z(s) is the variable value at point s. Because the data locations are regions, their centroids are used for the values of 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 $\left(Z(s_i) - Z(s_j)\right)^2$ where the Euclidean distance between s_i and 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, $\alpha=0.5$ 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, 1997: Some effects of spatial aggregation on multivariate regression parameters. To appear in *Festschrift for Jean Paelinck*, Griffith, Amrhein, and Huriot, eds.

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

4.10. Figures for Chapter 4

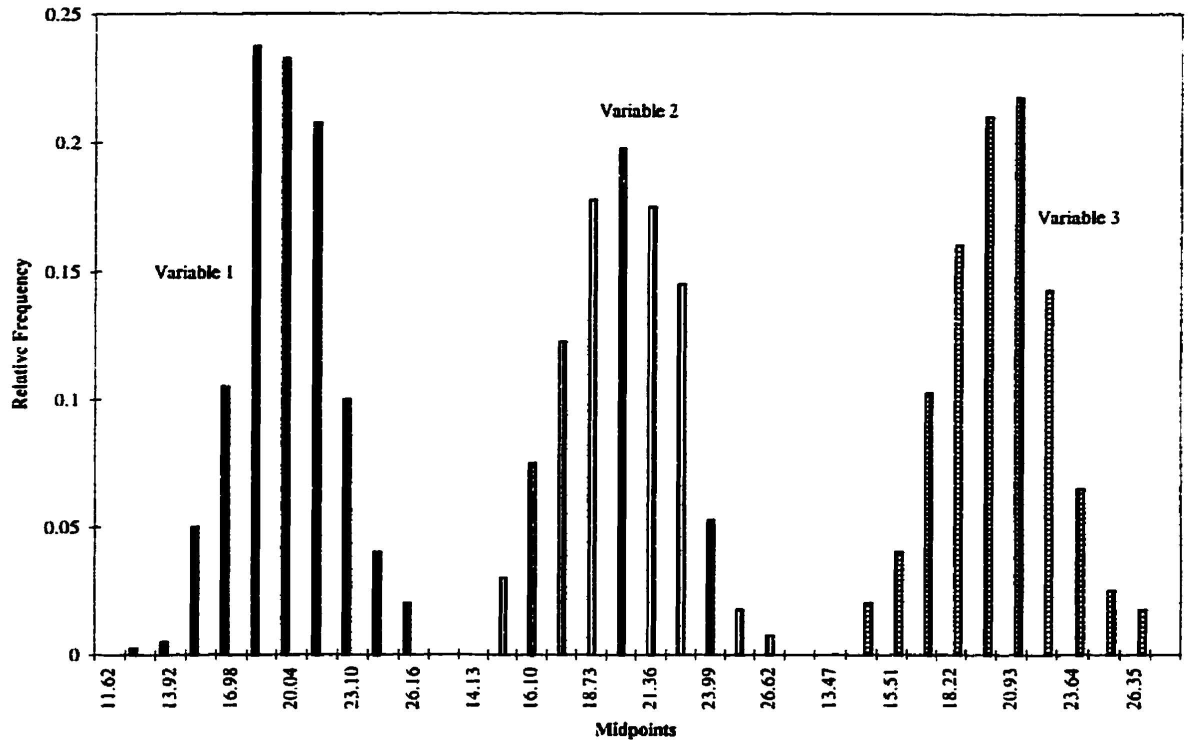


Figure 4.1: Frequency distributions of three variables generated by the new synthetic dataset generator. The variables have Moran Coefficients of -0.4, -0.2, and 0.0 respectively. The distributions are clearly mound-shaped, but are not normal.

a)

b)

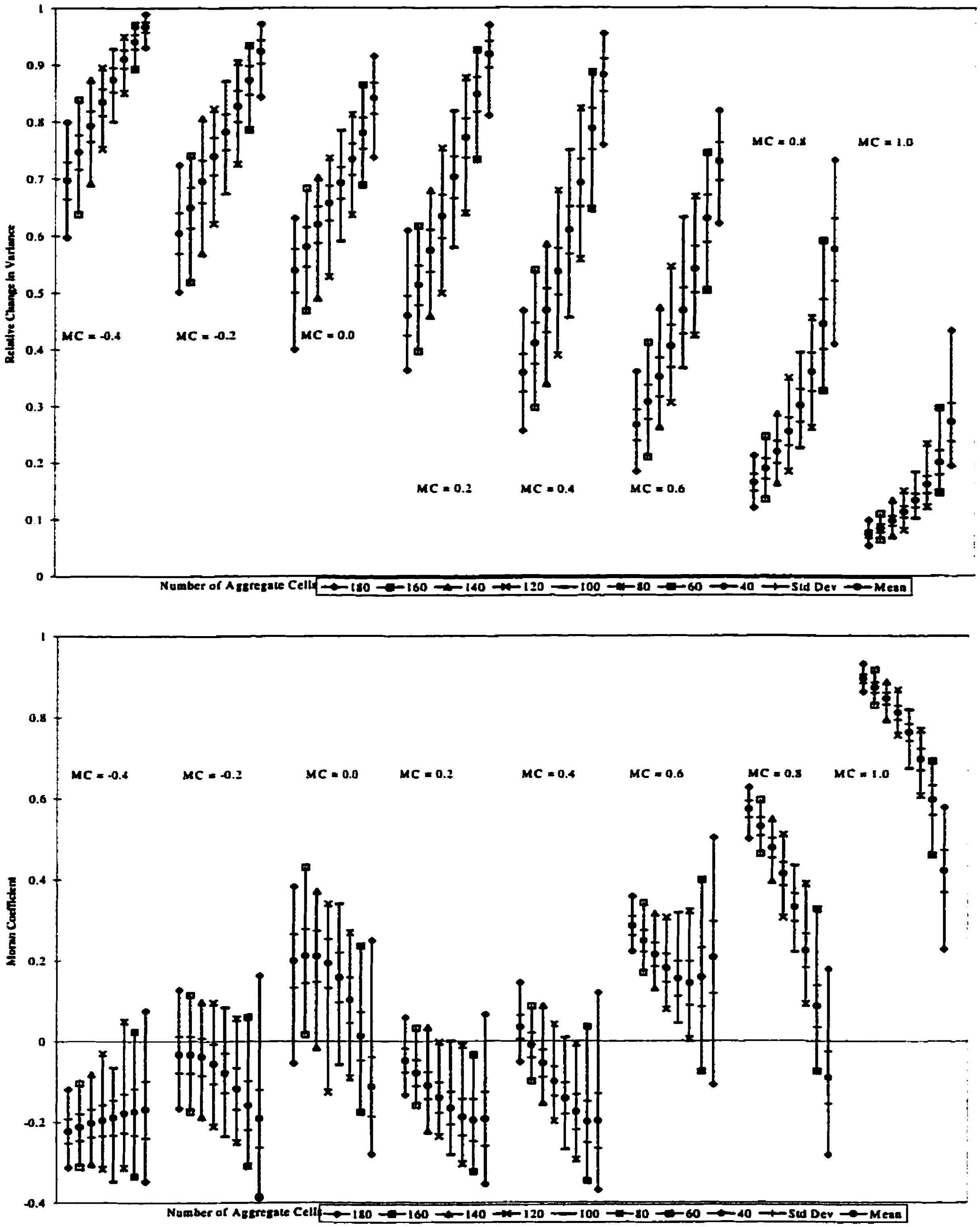


Figure 4.2: Variation of relative change in variance RCV (top) and MC with initial MC and aggregation. Note how the RCV has a well-defined variation with MC, but the aggregated MC does not.

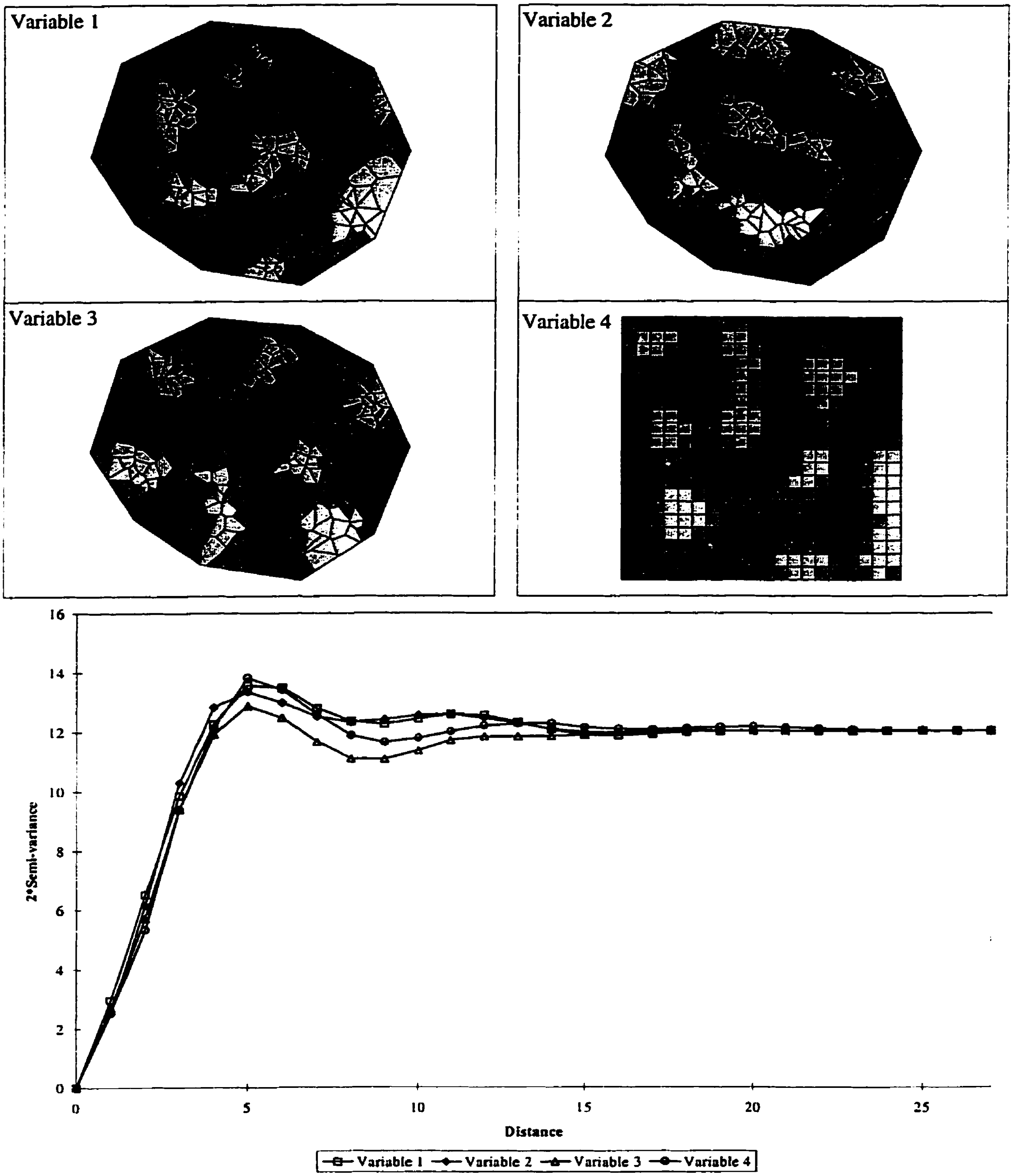


Figure 4.3a: Examples of variables with Moran Coefficients of 0.8 (top) and the variograms of the variables (bottom). These variables all have a large number of small clusters of high and low values, indicating short length scales and hence aggregation effects will be noticeable even for relatively small aggregated zones.

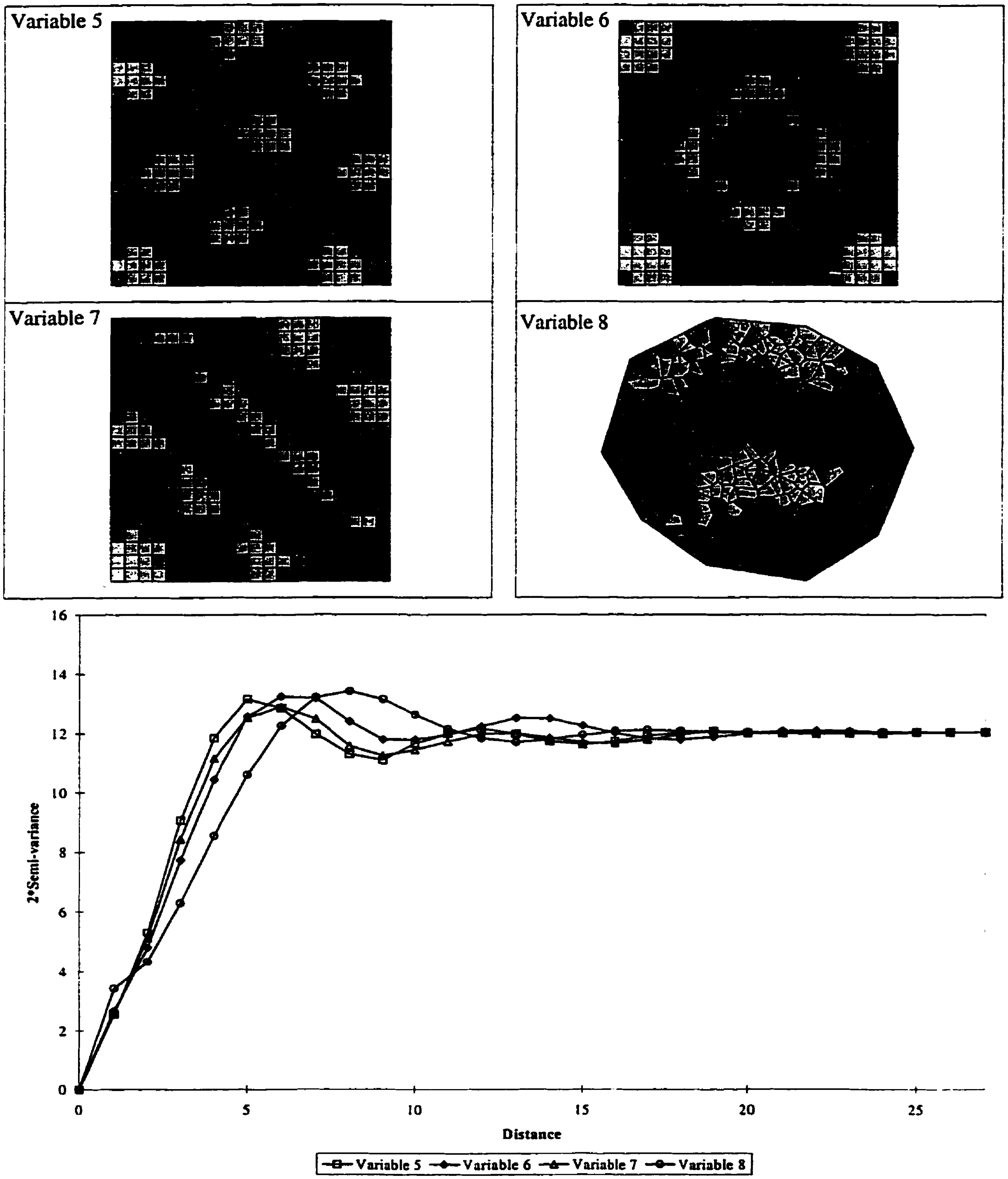


Figure 4.3b: Four more variables with MCs of 0.8 with length scales longer than those of Figure 3a. Note how the length scale is related to the number and positioning of clusters of similar values.

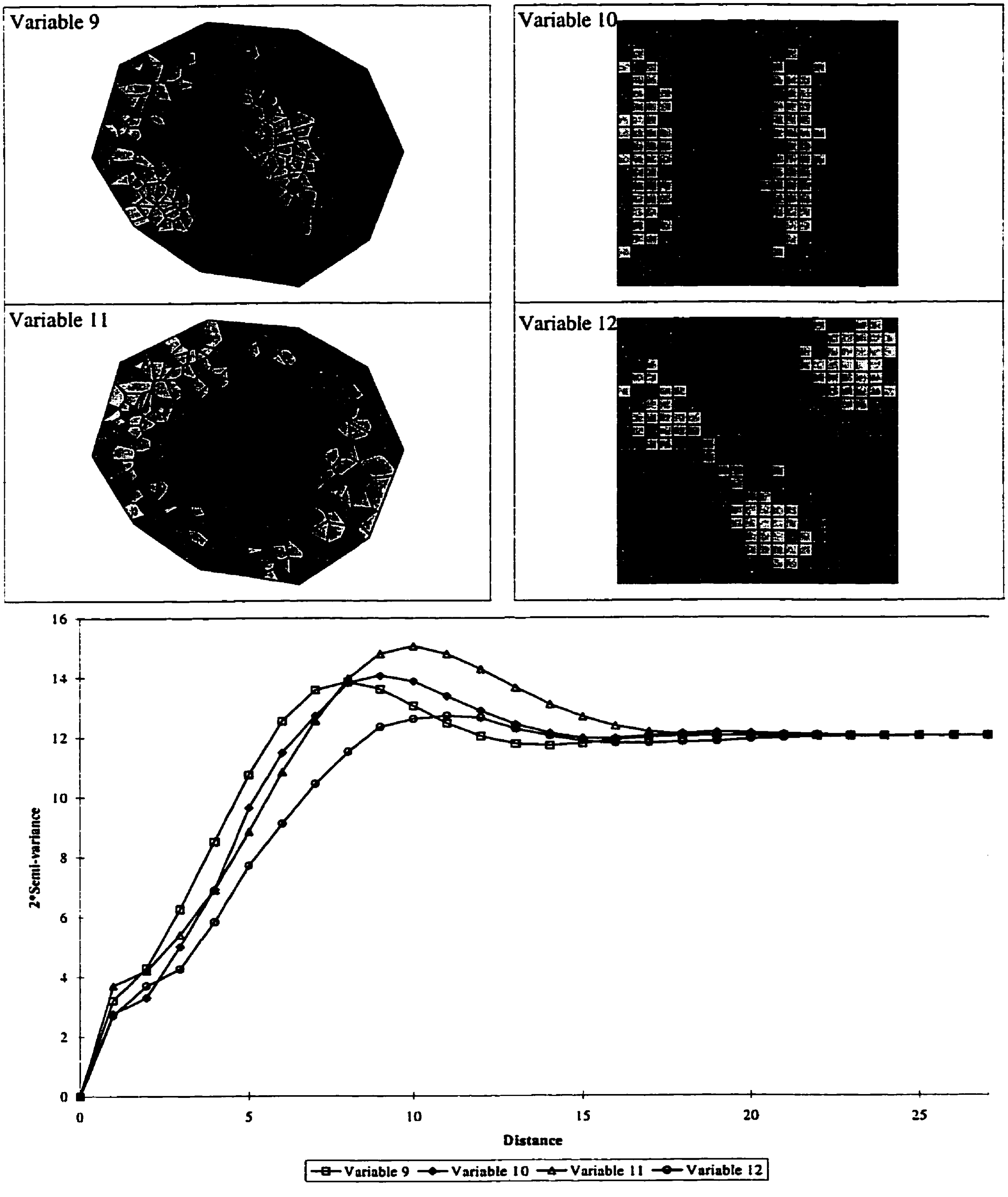


Figure 4.3c: Four more variables with MCs of 0.8, all with longer length scales. Note the lack of oscillation of the variograms after the peaks, compared to those of the previous figures.

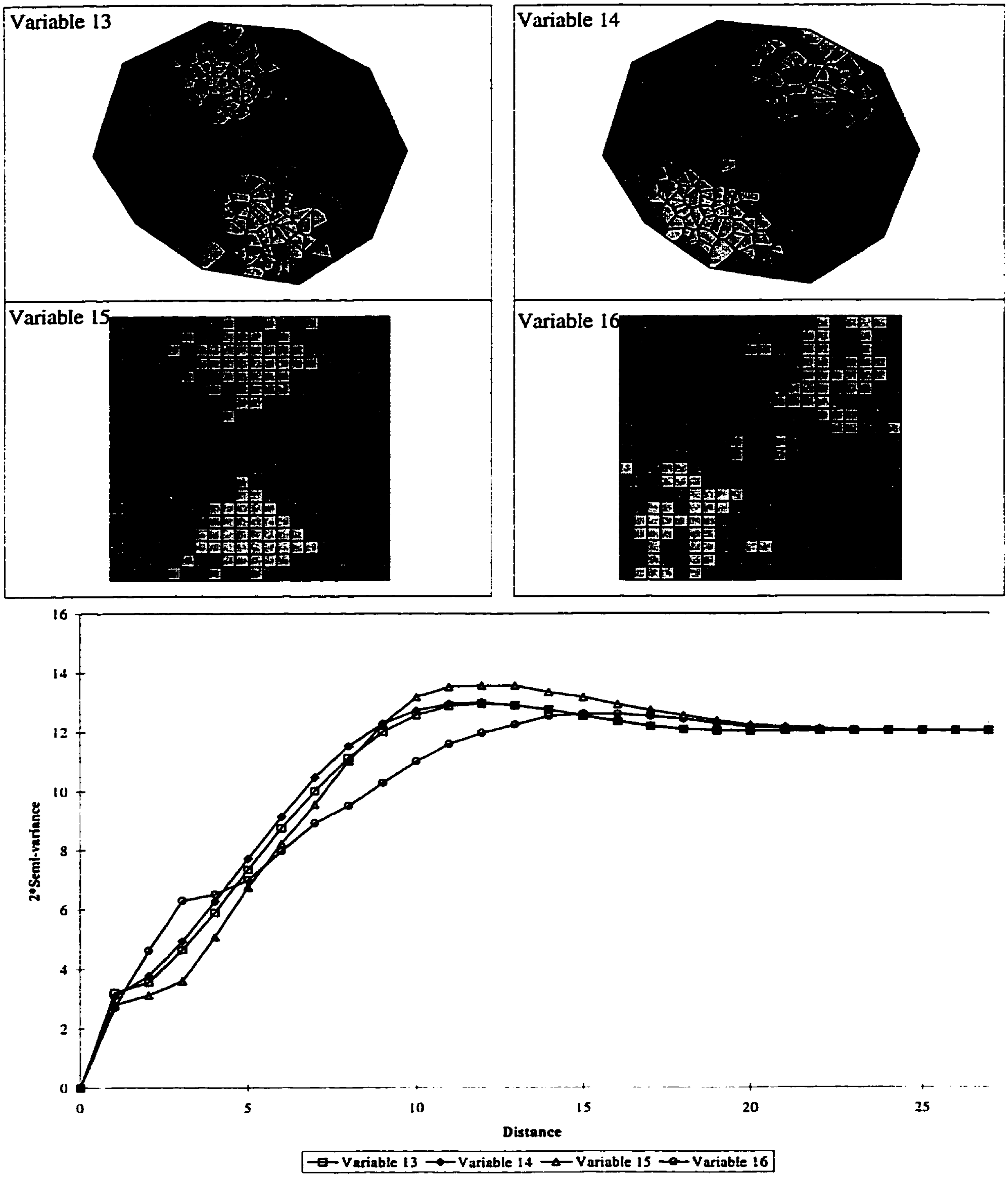


Figure 4.3d: The final four variables with MCs of 0.8, all with long length scales. On average, aggregation effects manifest themselves more slowly for these variables than for those with shorter length scales.

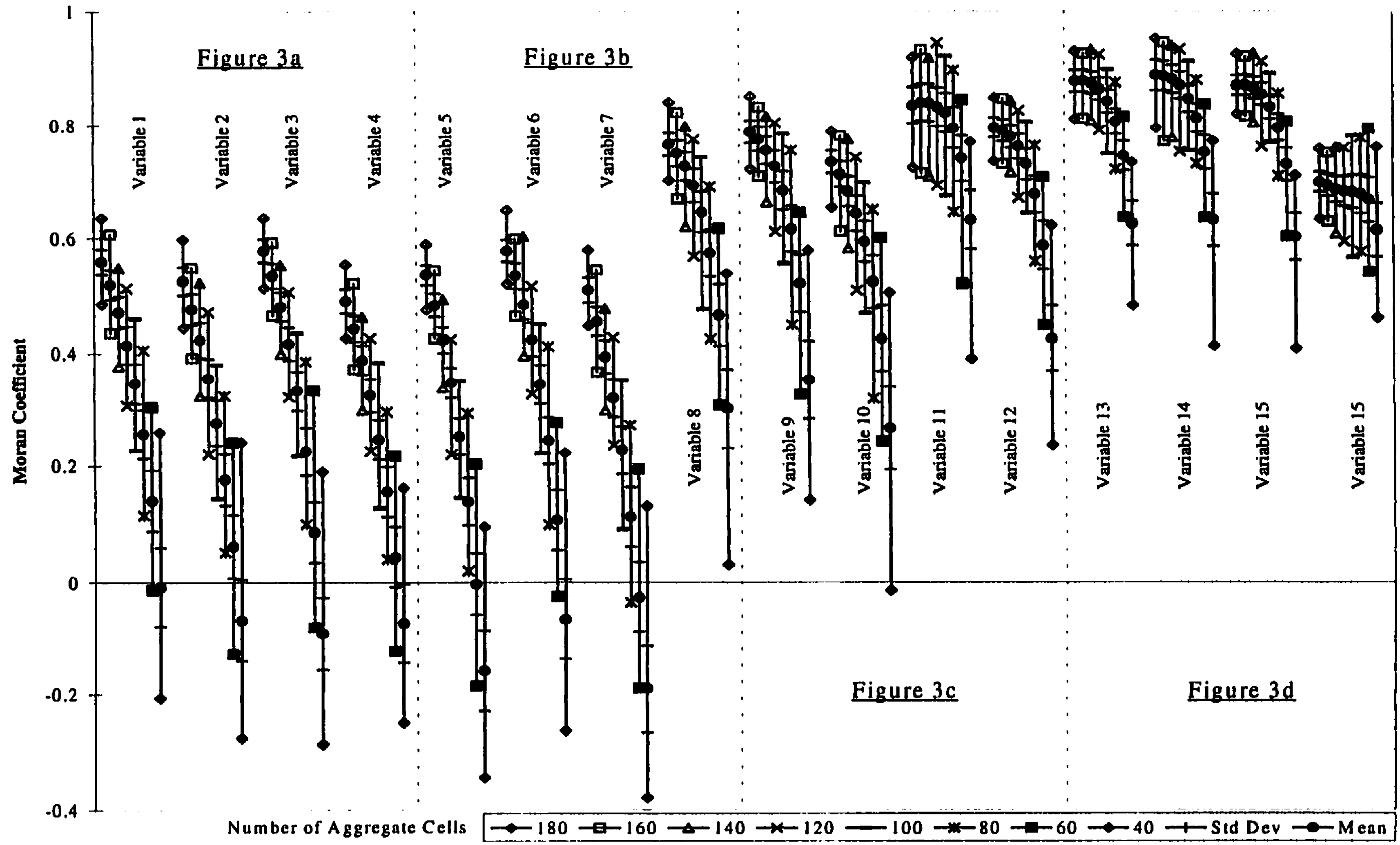


Figure 4.4a: Variation of the MCs of the variables in Figures 3a to 3d. It can be seen that the longer the length scale, the larger the region must be before aggregation effects become severe and the slower the rate at which the aggregated MC decreases. Each group of lines is labeled with the variable number; each set of four groups is labeled with the figure in which they appear.

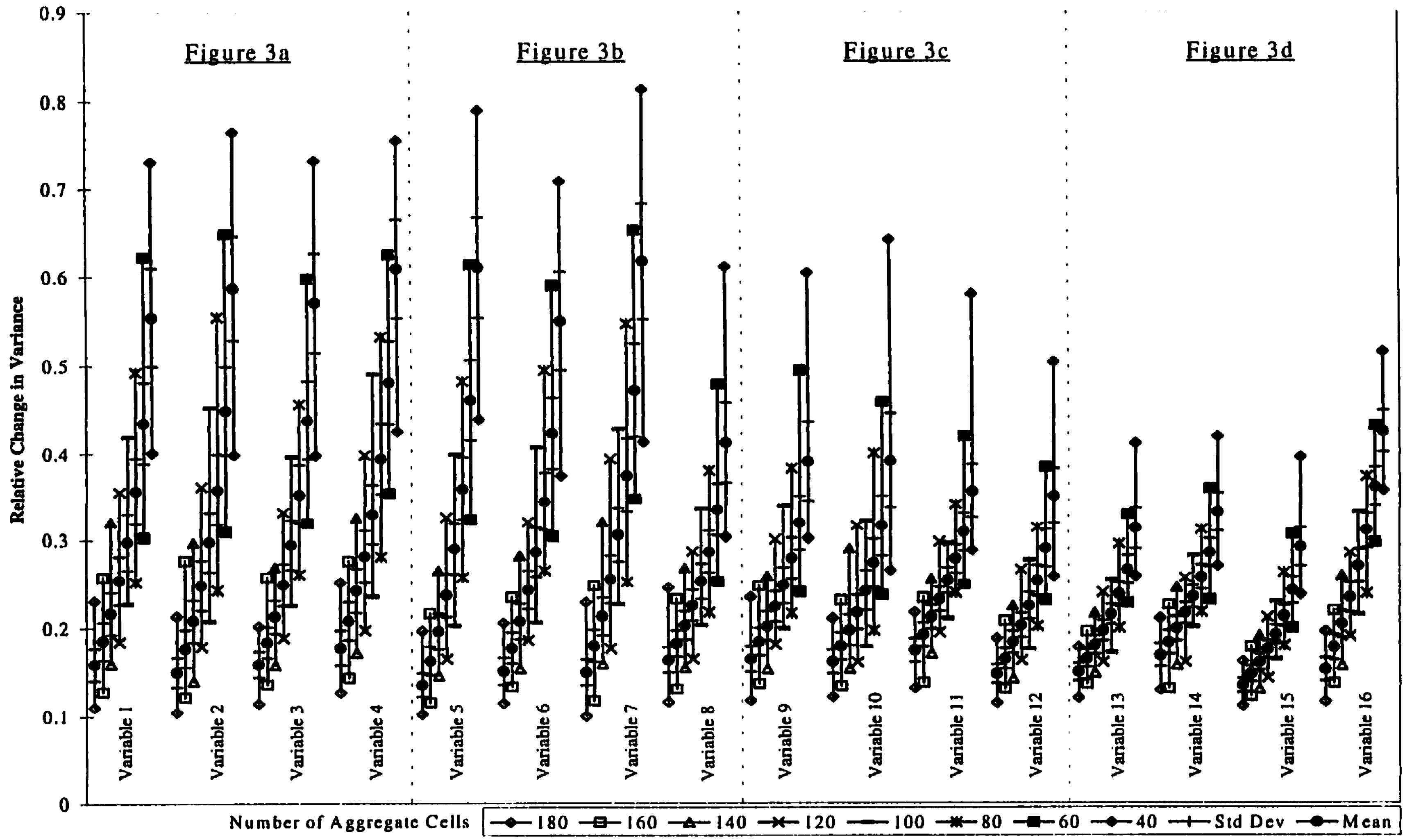
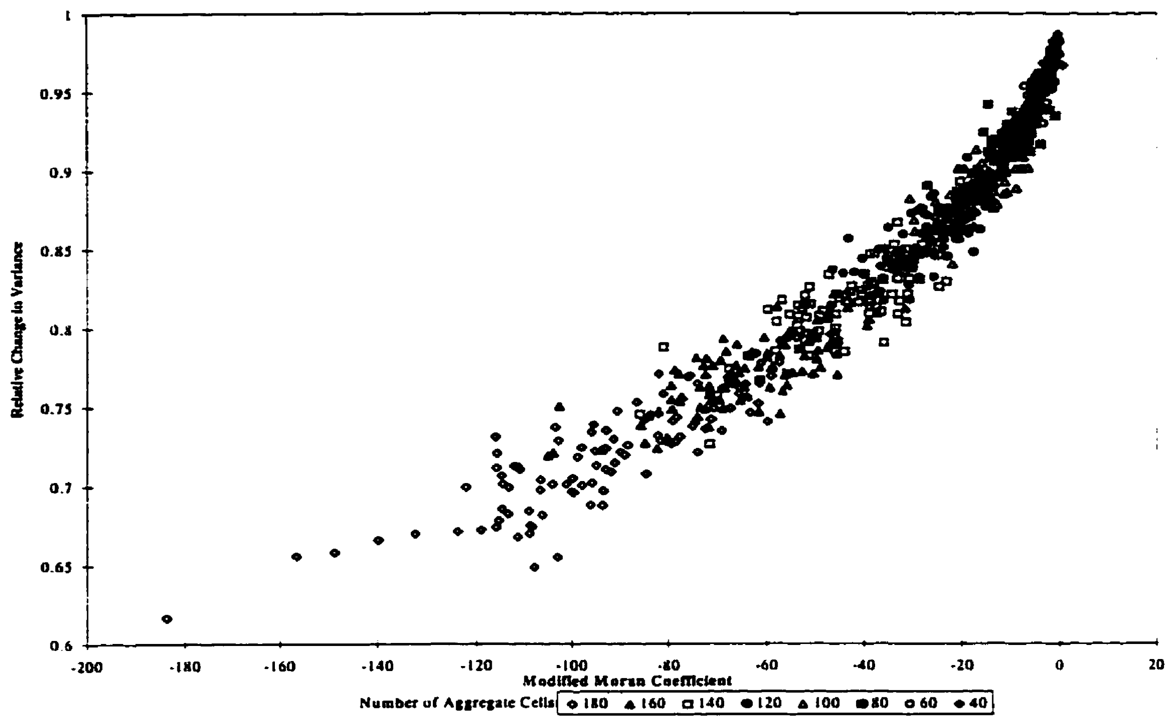


Figure 4.4b: Variation of the variances of the variables in figures 3a to 3d. Results here correspond with those in Figure 4a: the longer the length scale, the less the variable is affected by aggregation.

a)



b)

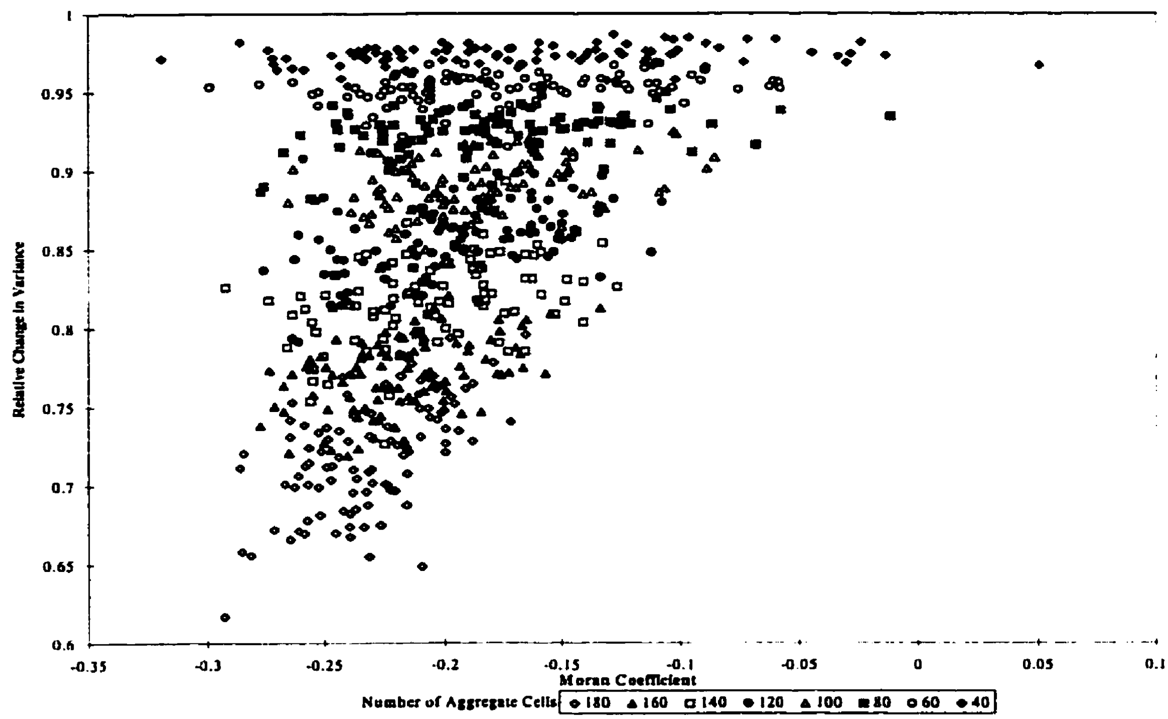
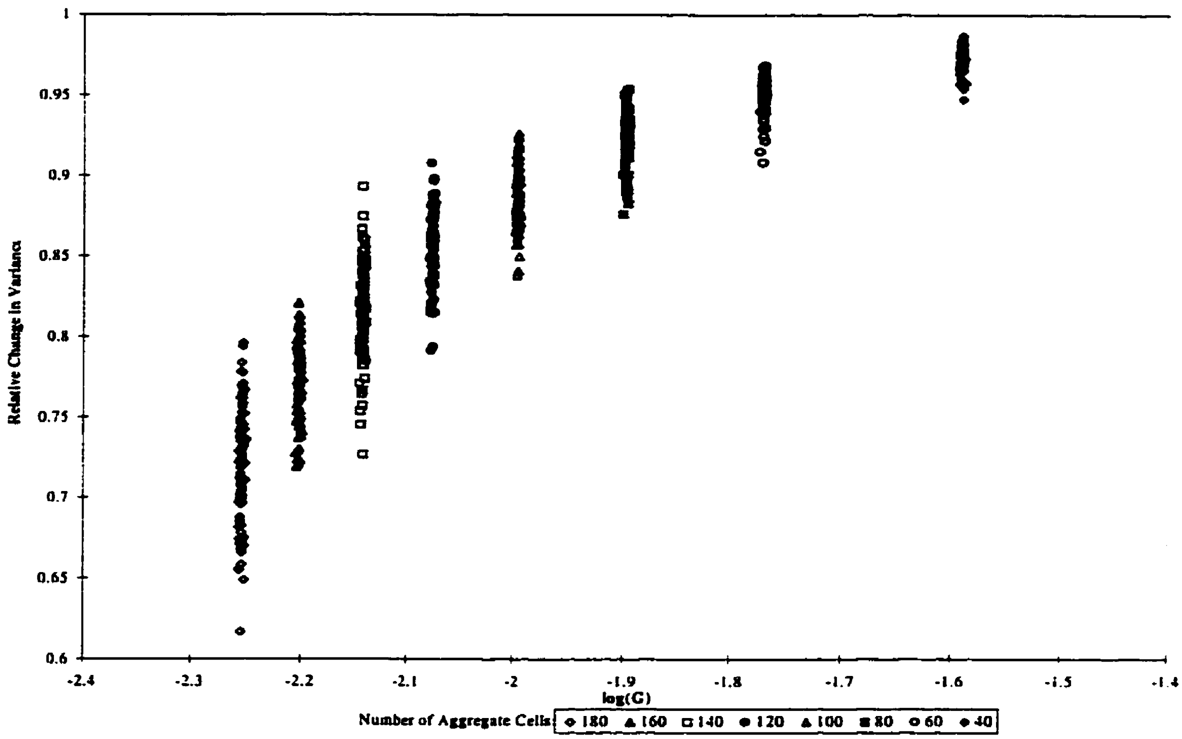


Figure 4.5: Relative change in variance (RCV) as a function of the aggregated MC without the sum of squares of deviations term (top) and of regular aggregated MC (bottom), for variable with initial MC of -0.4. Note how adding the term significantly worsens the relationship.

a)



b)

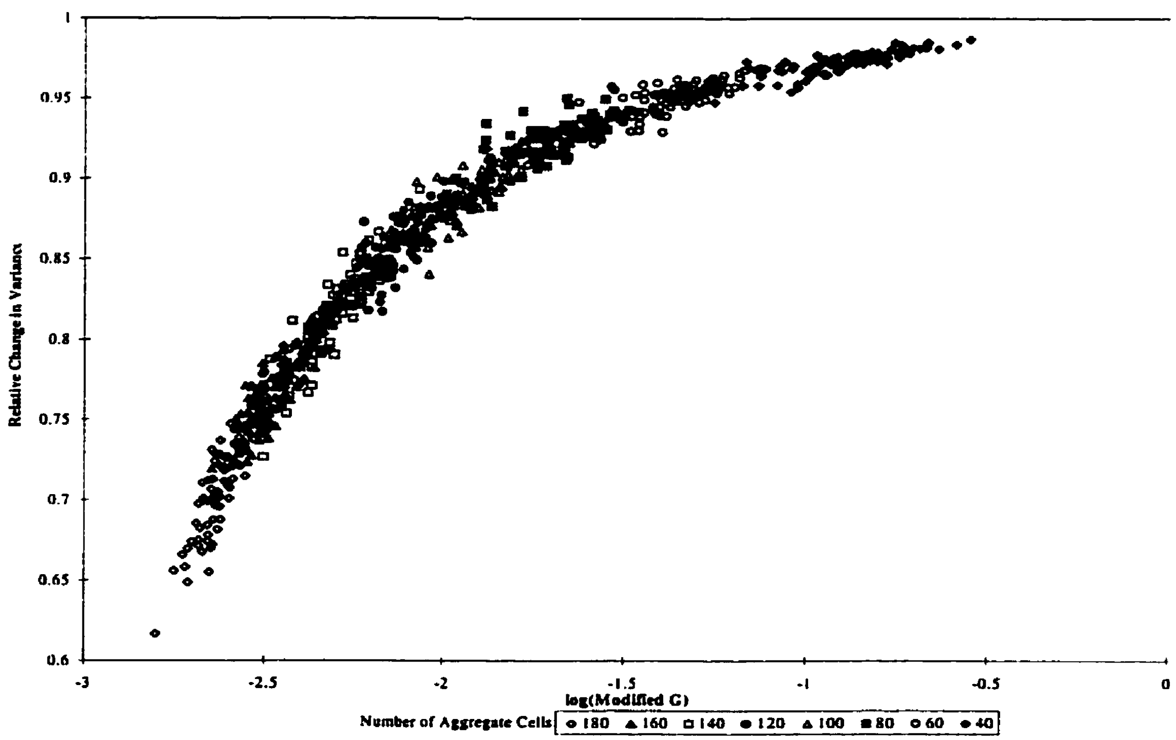
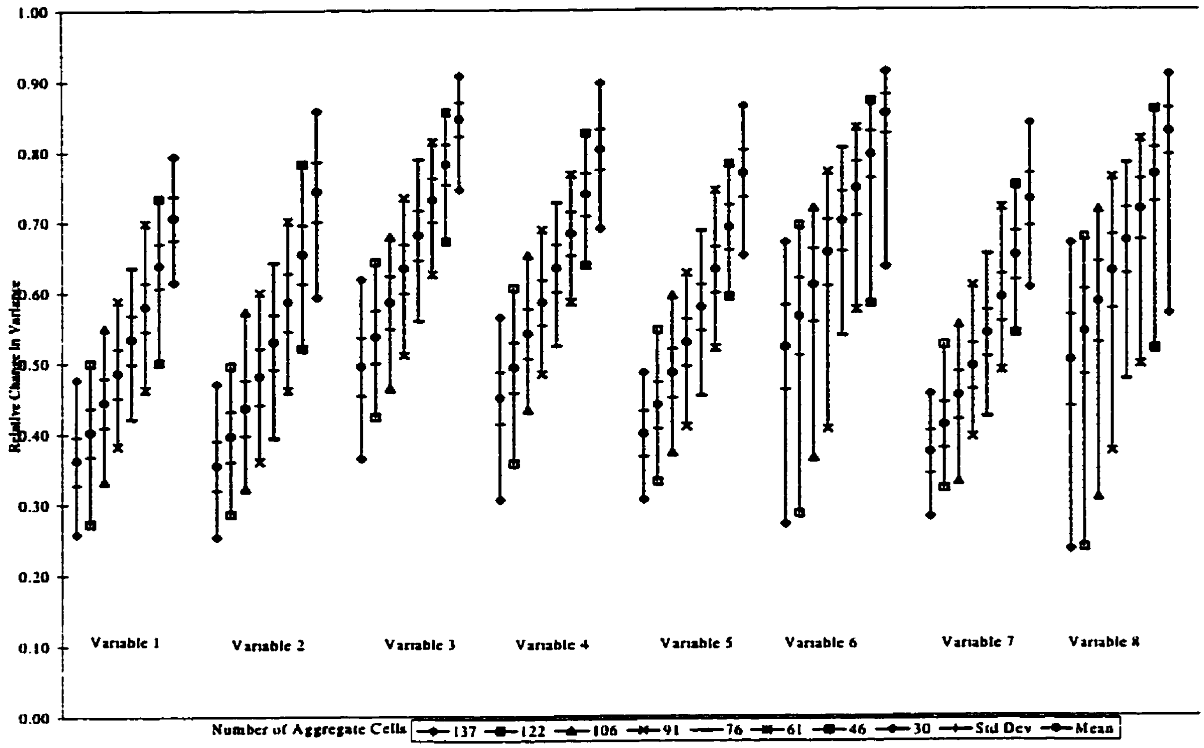


Figure 4.6: Relative change in variance (RCV) as a function of $\log_{10}(G)$ (top) and $\log_{10}$(modified G) (bottom). Notice how, unlike Figure 5, adding the variance (sum of squares of deviations divided by M, the number of cells) improves the relationship.

a)



b)

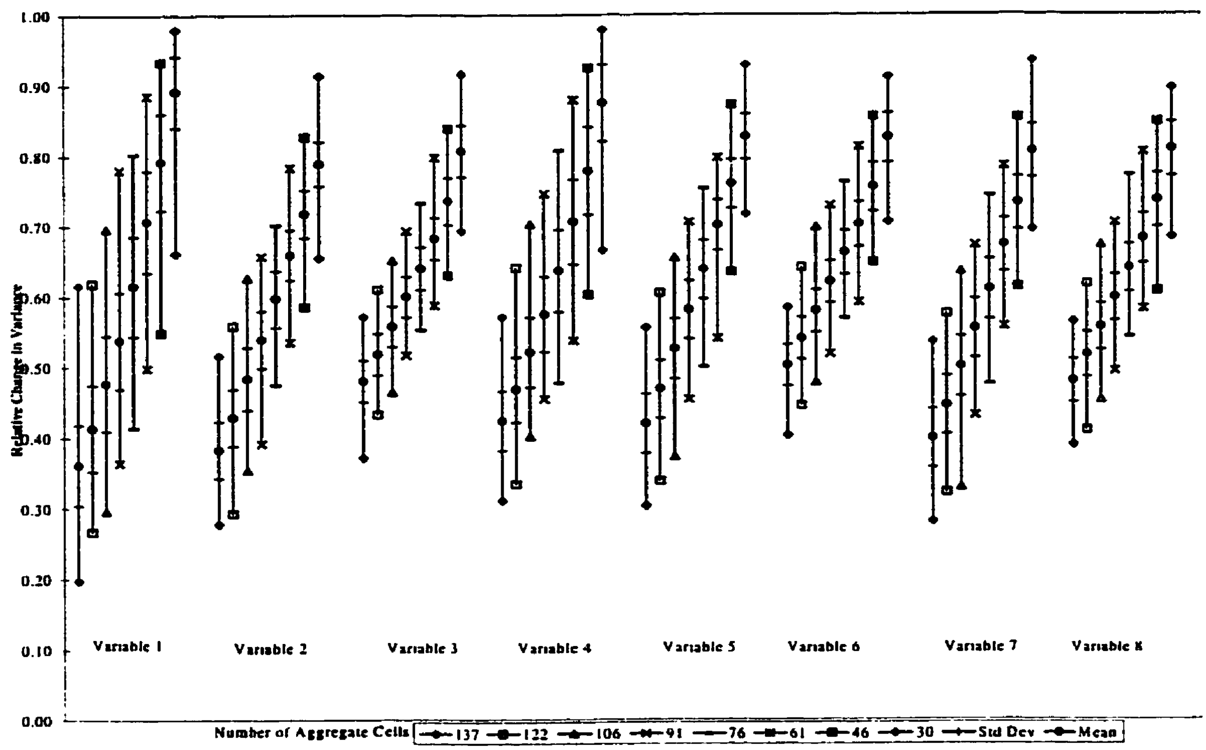
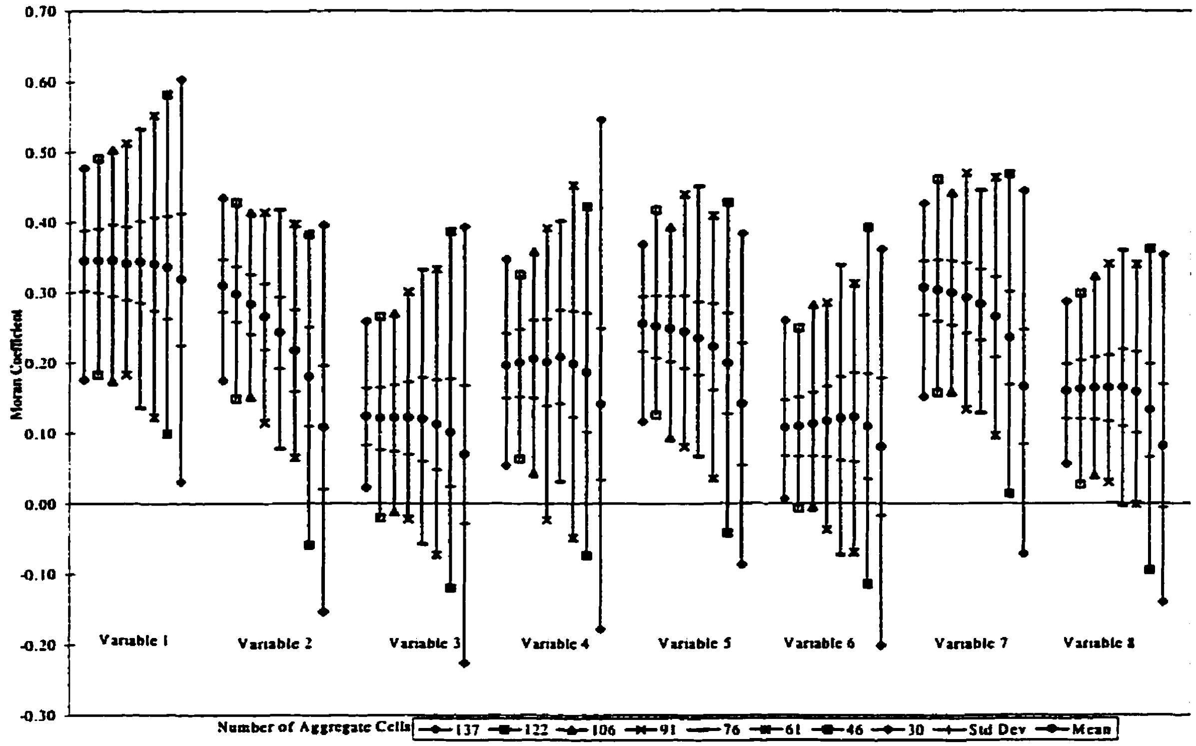


Figure 4.7: Behaviour of the Relative Change in Variance with aggregation for the actual Lancashire dataset (top) and a synthetic Lancashire dataset (bottom). Differences exist, but the general patterns of behaviour are quite similar.

a)



b)

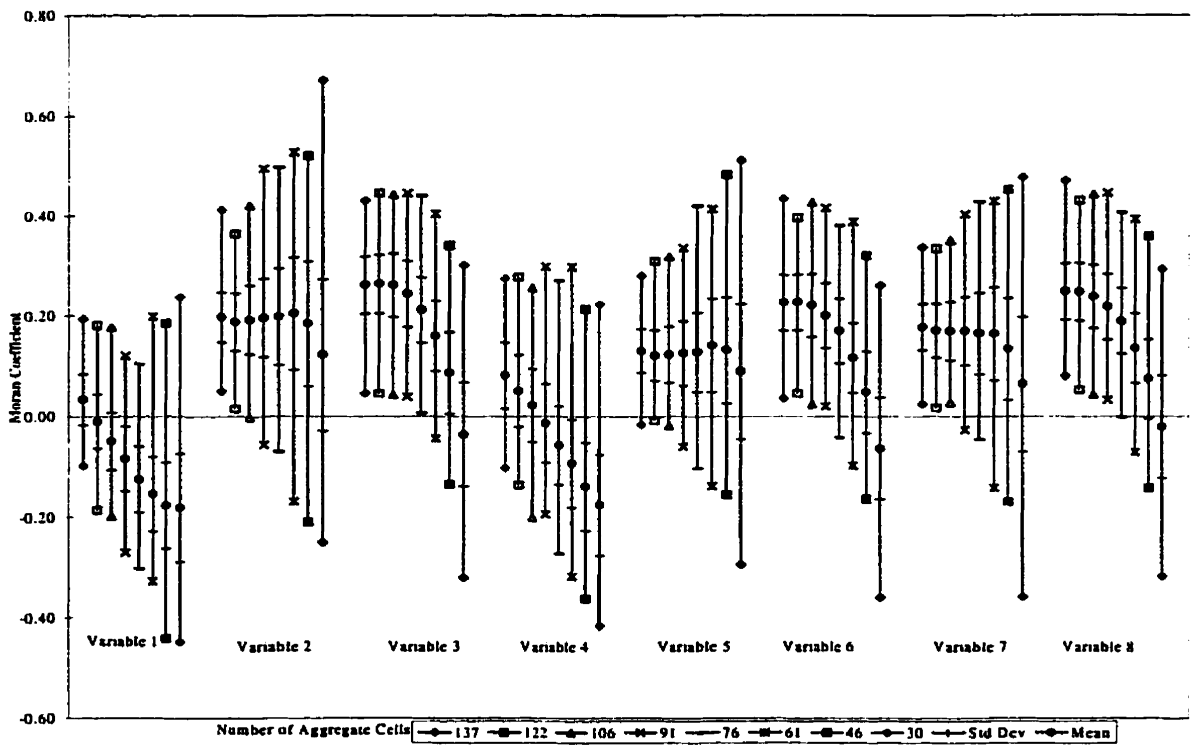
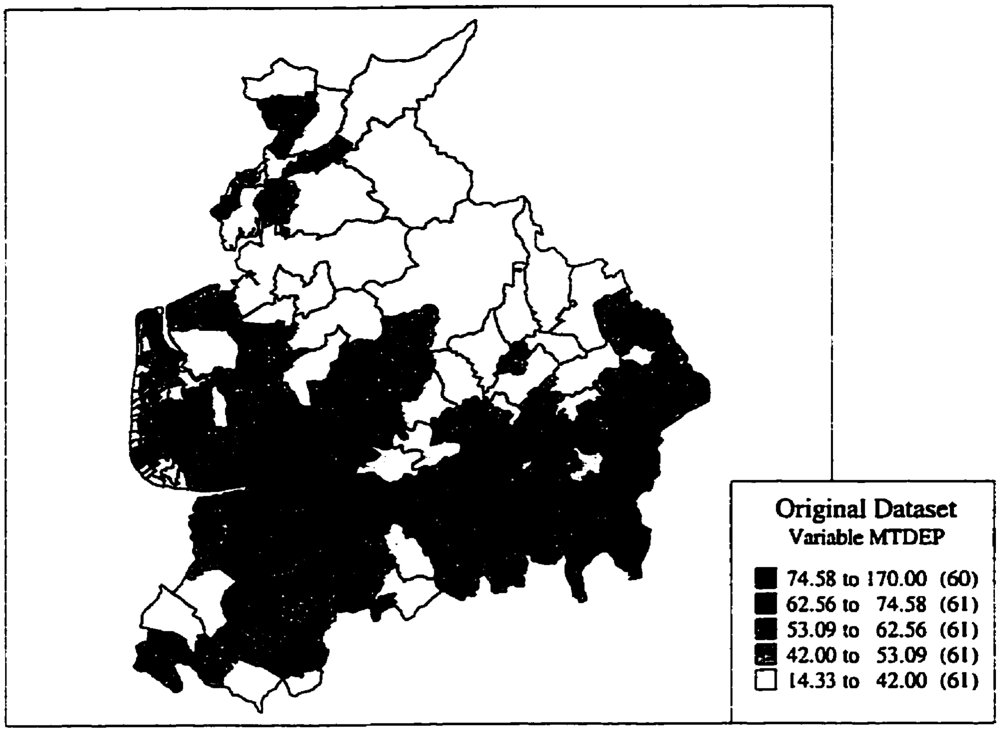


Figure 4.8: Behaviour of the aggregated Moran Coefficients for the actual Lancashire dataset (top) and a synthetic Lancashire dataset (bottom). The differences in behaviour are most likely due to the different spatial configurations of the values, as shown in Figure 4.9.



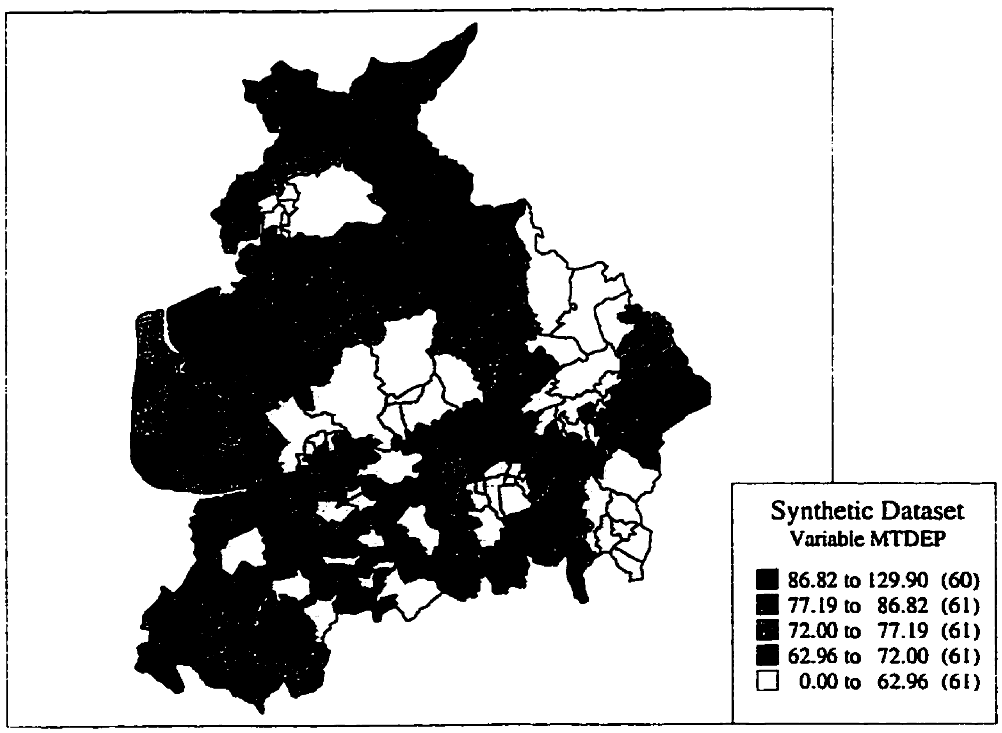


Figure 4.9: Comparison of the original and synthetic variable MTDEP in the Lancashire dataset.

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:

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

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 $r_{xy} = \frac{\text{Cov}(X, Y)}{s_x s_y}$, 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 $b_{xy} = \text{Cov}(X, Y) / s_x^2$ 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 aggregathe 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 co-variance 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 (1997) 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

Cells	Original MC = -0.4			Original MC = 0.8		
	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

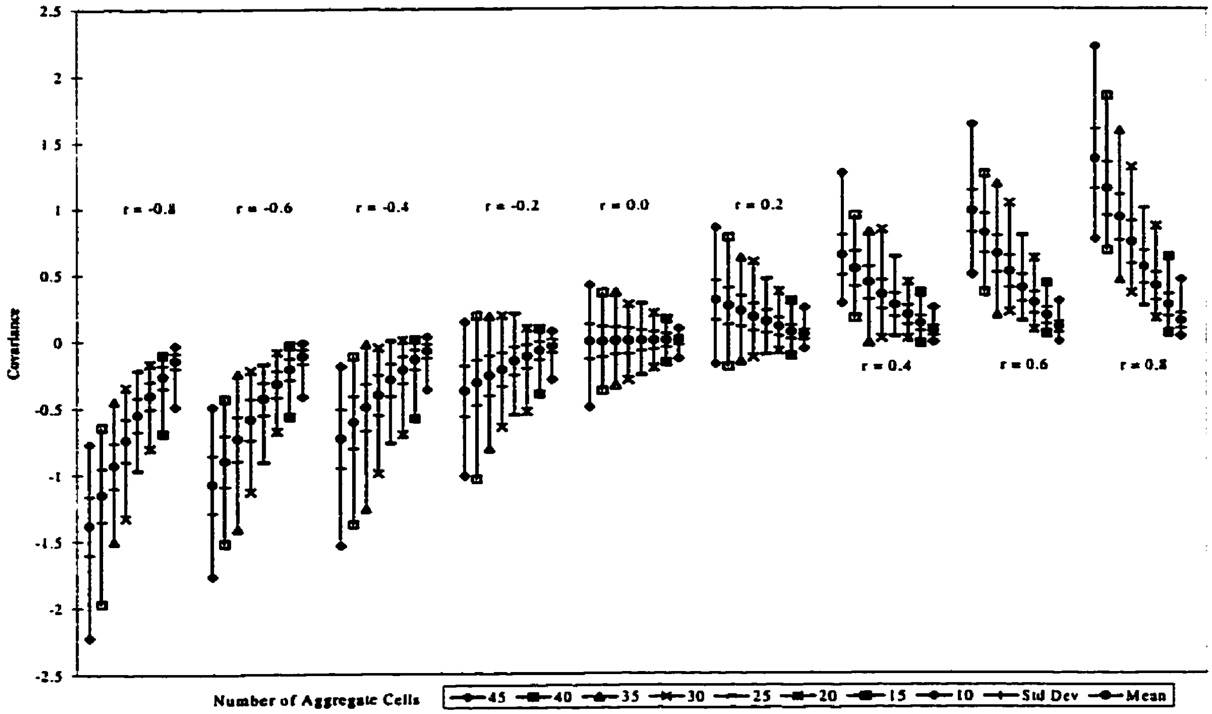
Cells	Original MC = -0.4			Original MC = 0.8		
400	r = -0.6	r = 0.4	r = 0.8	r = -0.6	r = 0.4	r = 0.8
180	-0.6202	0.3899	0.8008	-0.6041	0.4008	0.8011
160	-0.6238	0.3911	0.8020	-0.6035	0.4002	0.8000
140	-0.6240	0.3874	0.8040	-0.6030	0.3994	0.7984
120	-0.6289	0.3881	0.8041	-0.6013	0.3983	0.7956
100	-0.6220	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

5.10. Figures for Chapter 5

a)



b)

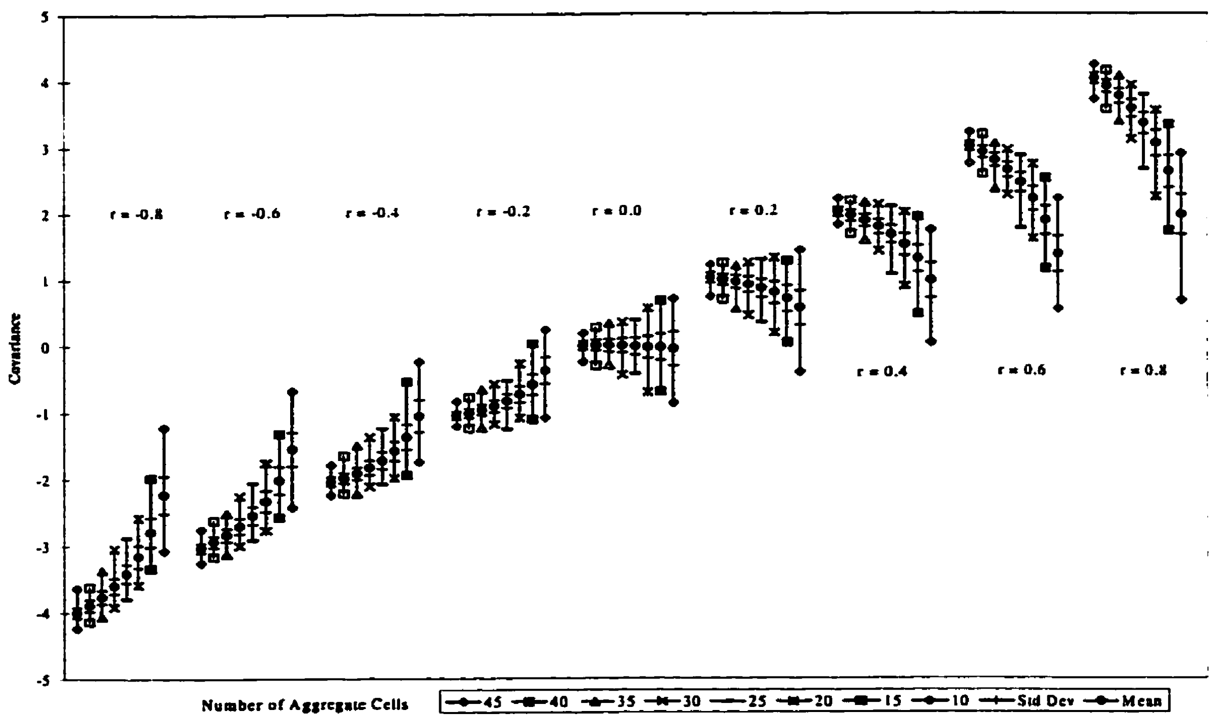
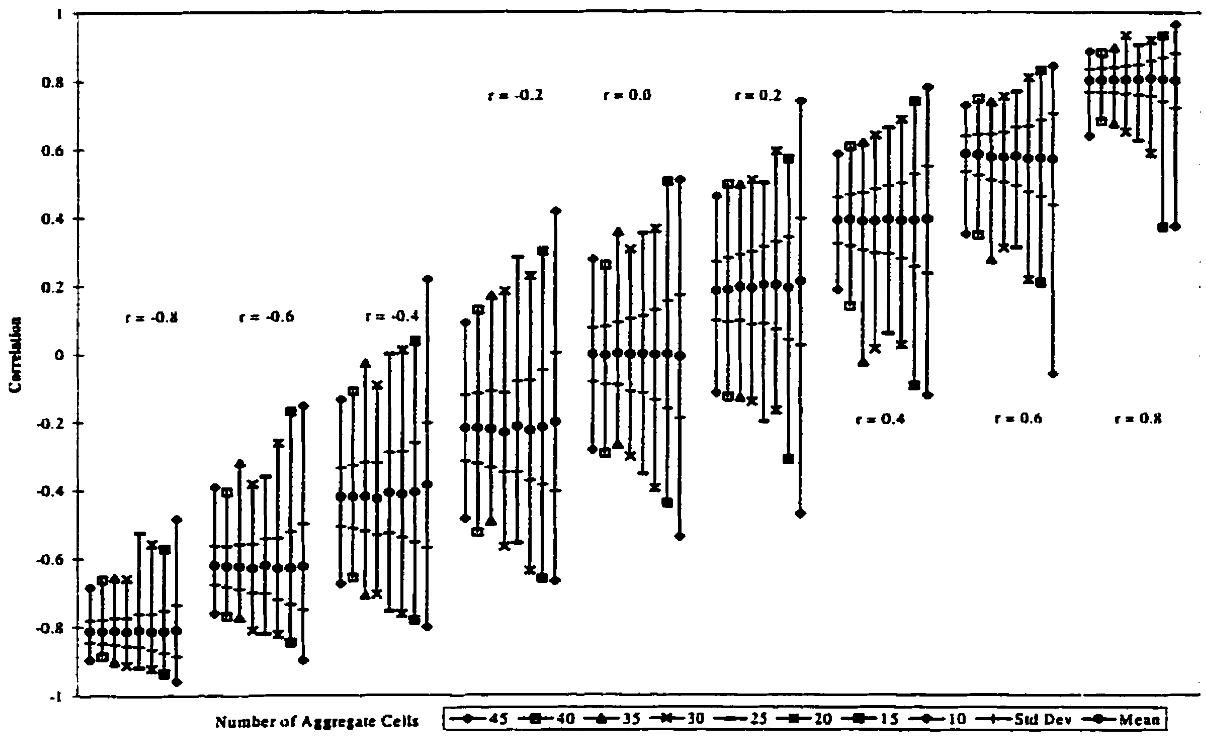


Figure 5.1: Variation of aggregated covariance with initial correlation where dependent and independent variables have MCs of (a) -0.4 and (b) 0.8. Note how the concavity of the line joining the heavy dots changes between the diagrams.

a)



b)

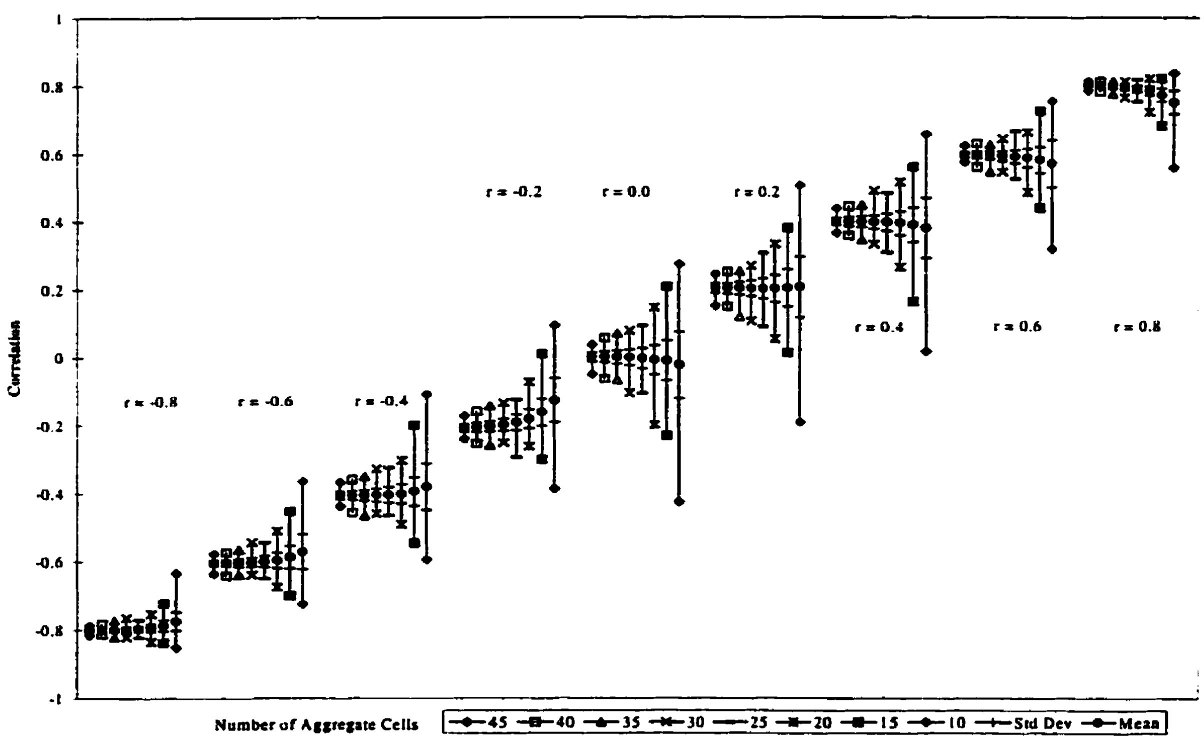
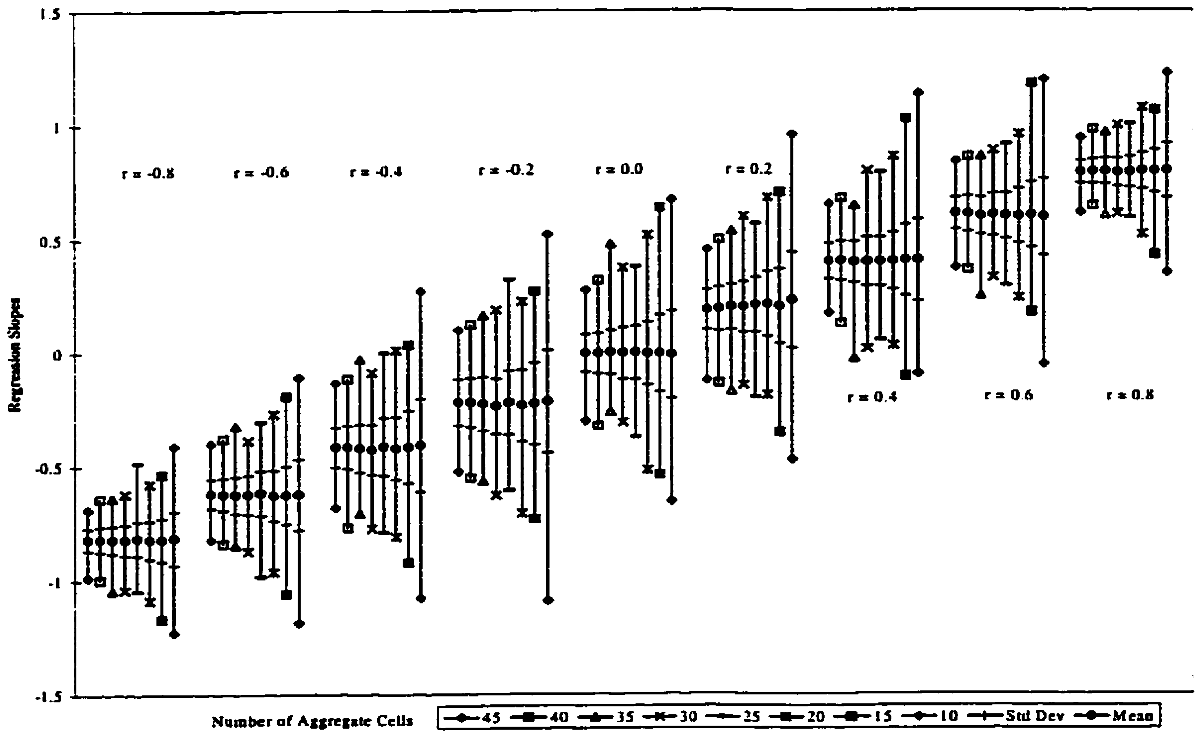


Figure 5.2: Variation of aggregated correlation with initial correlation where dependent and independent variables have MCs of (a) -0.4 and (b) 0.8. Note the symmetry of the ranges, and how the ranges decrease with increasing MC of the variables.

a)



b)

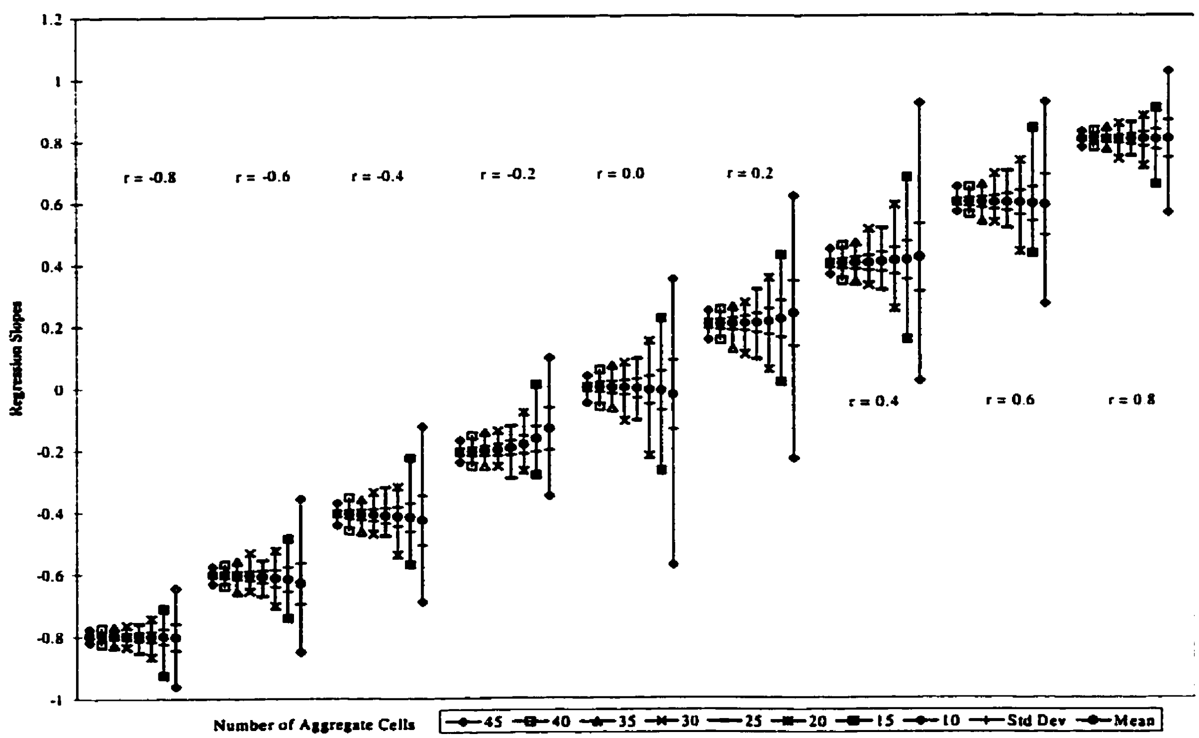
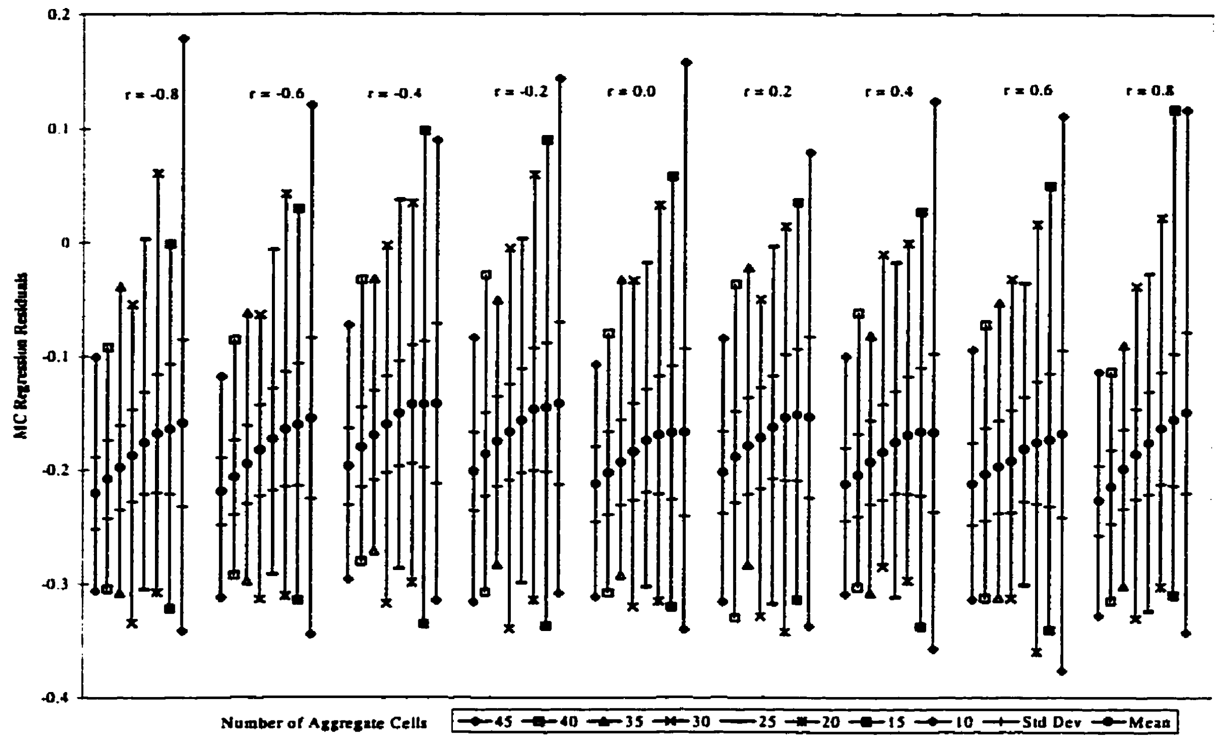


Figure 5.3: Variation of aggregated upper triangle (row is independent, column dependent) of the matrix of regression slope parameters with initial correlation, where dependent and independent variables have MCs of (a) -0.4 and (b) 0.8. Note the general lack of dependence on initial correlation. The lower triangle behaves similarly.

a)



b)

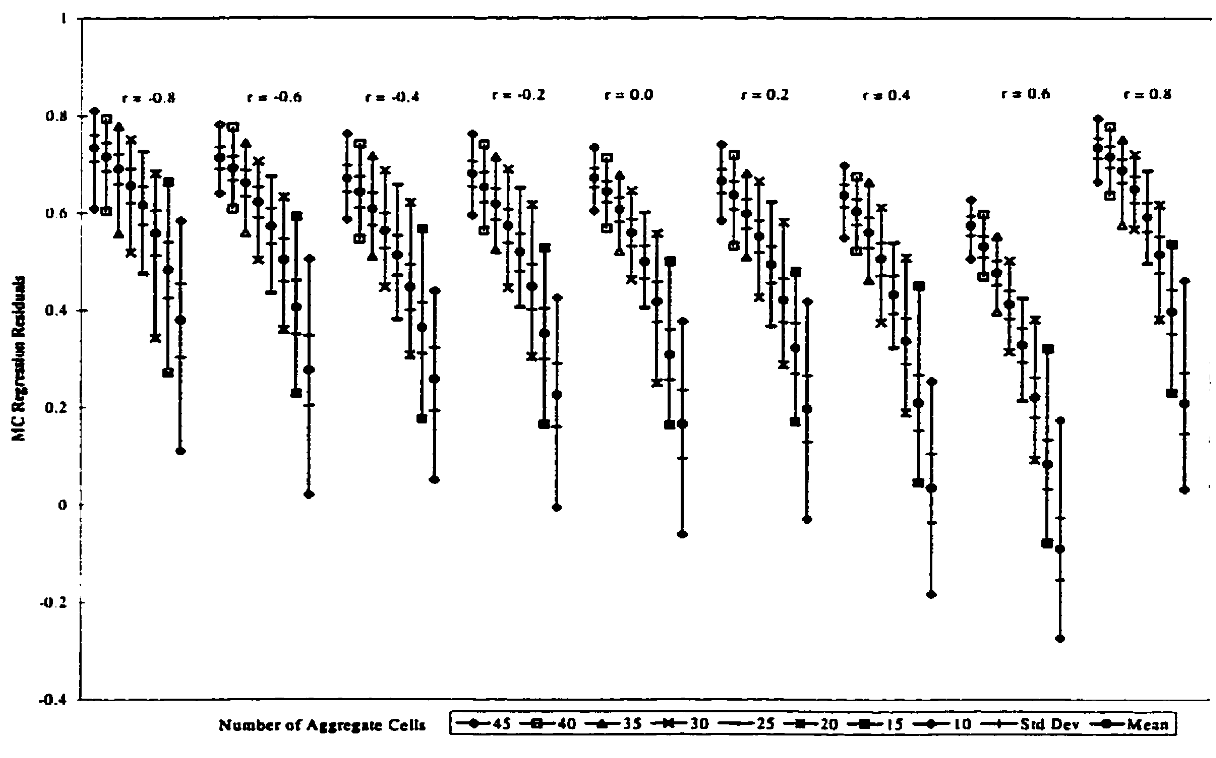
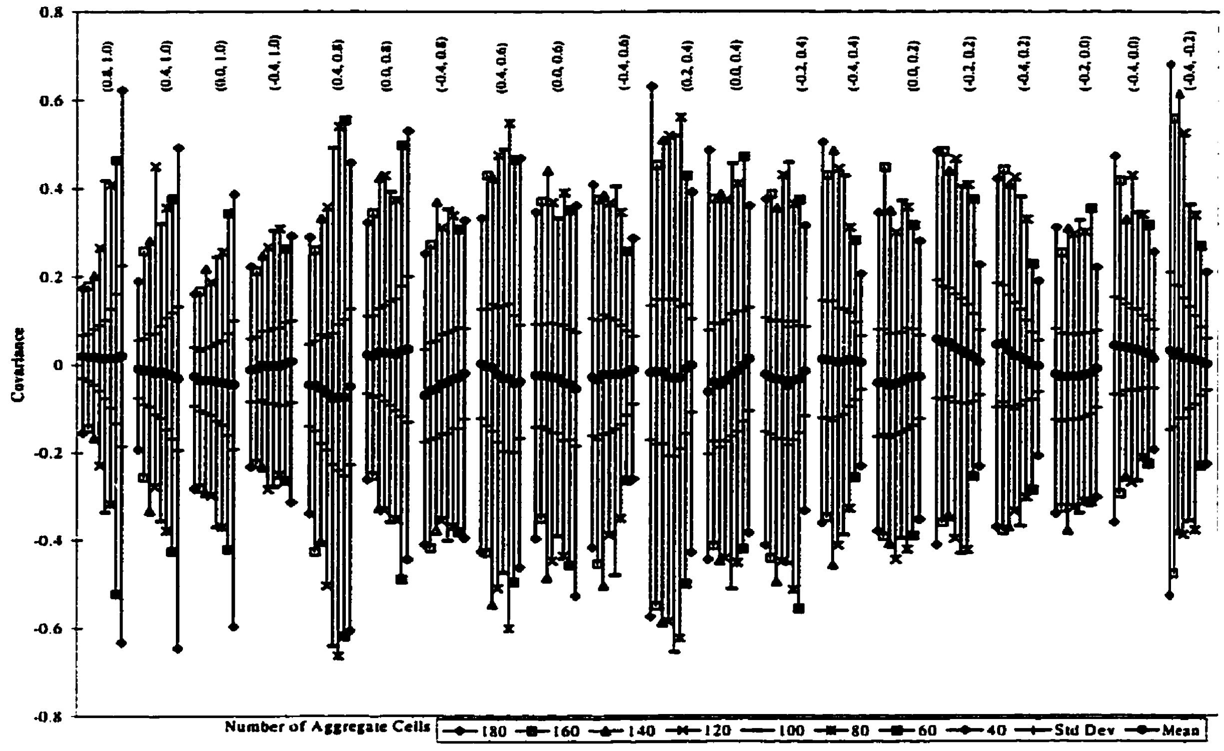


Figure 5.4: Variation of the MC of regression residuals with the original correlation, where dependent and independent variables have the original MC of a) -0.4 and b) 0.8. Note the general lack of dependence on correlation.

a)



b)

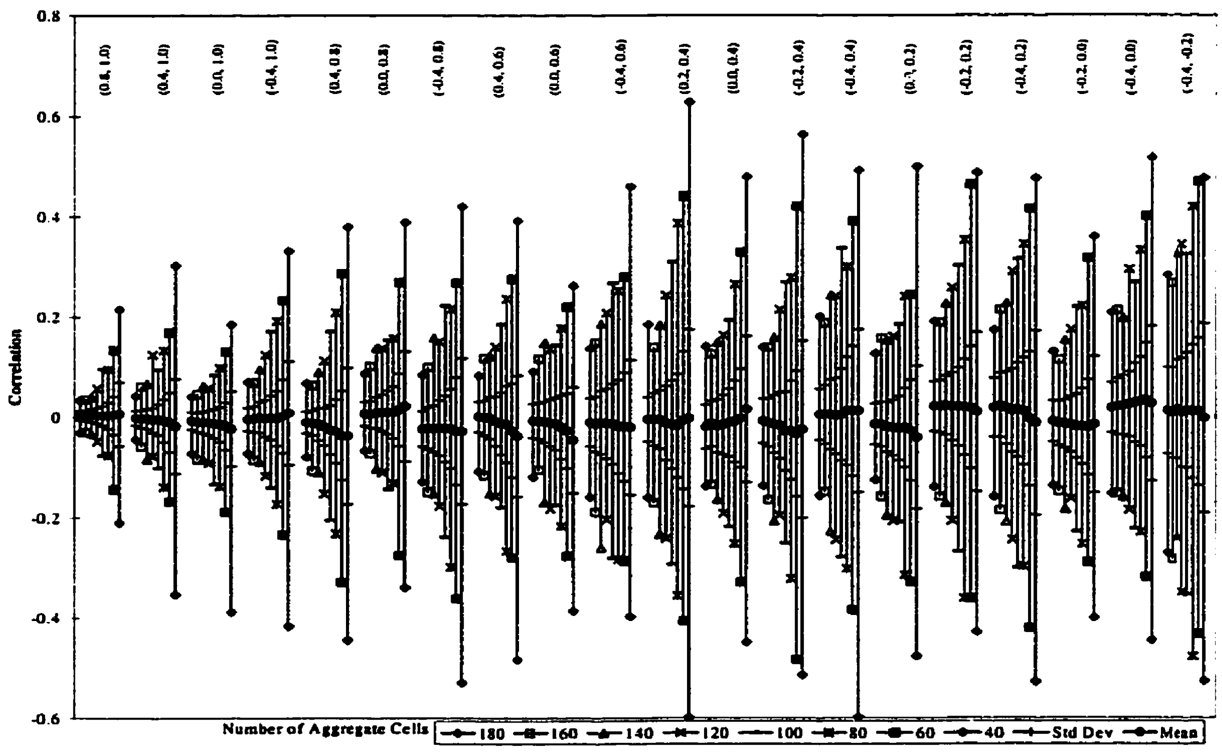
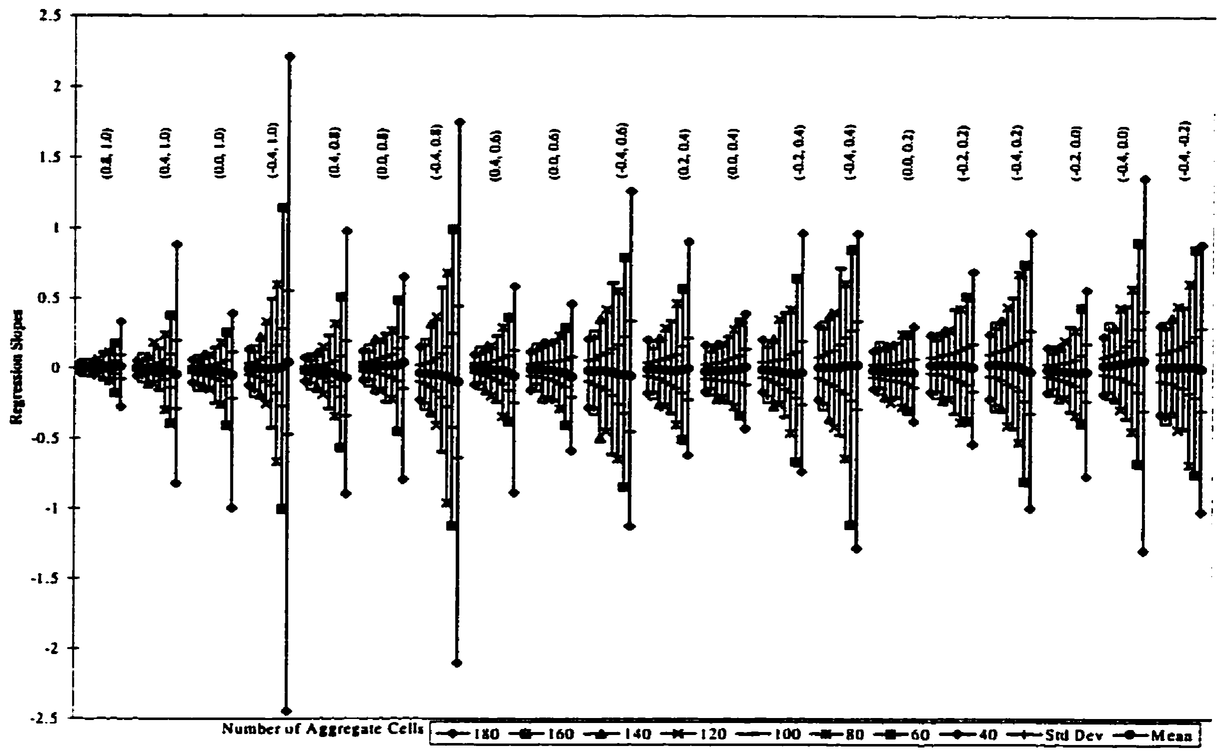


Figure 5.5: Variation of covariances (top) and correlations with the (MC independent, MC dependent) variables for an initial correlation of 0.0. Note how the pattern of change in a) is similar to that between Figures 5.1a and 5.1b.

c)



d)

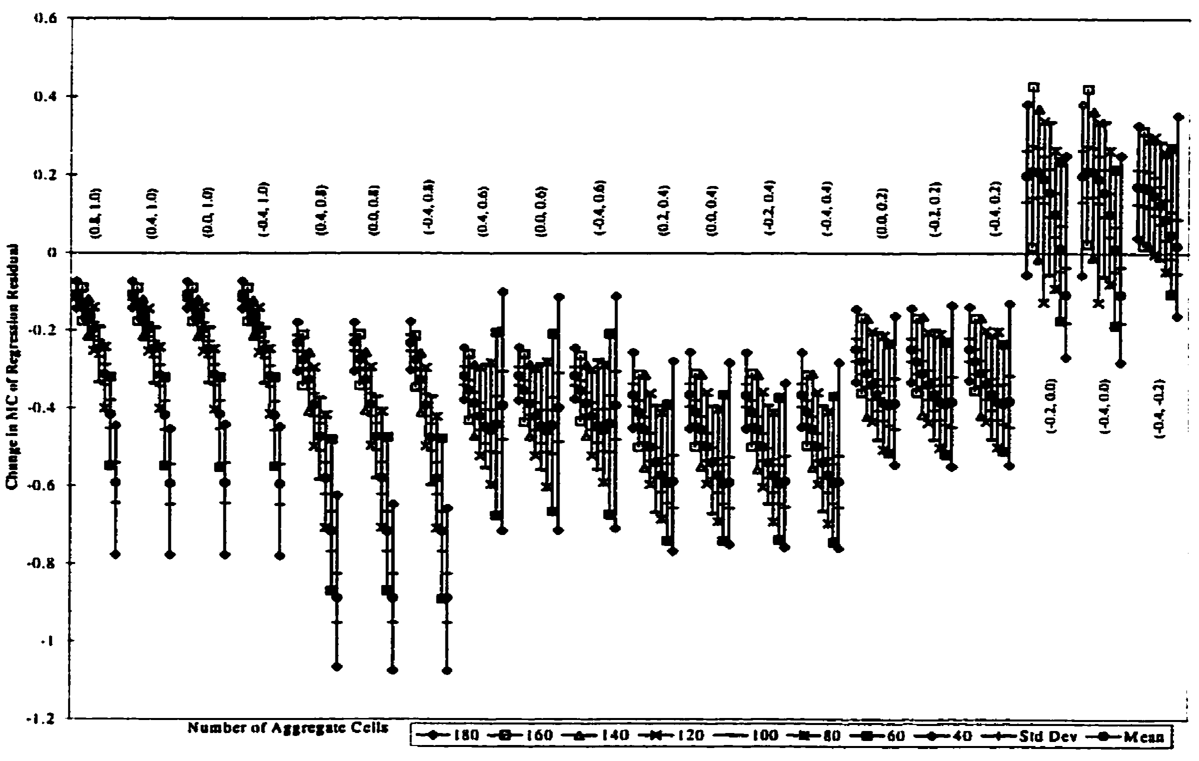


Figure 5.5, con't: Variation of upper triangle of regression slope parameters (top) and change in MC_{RR} with the (MC independent, MC dependent) variables for an initial correlation of 0.0. Note the lack of dependence of MC_{RR} on the MC of the independent variable.

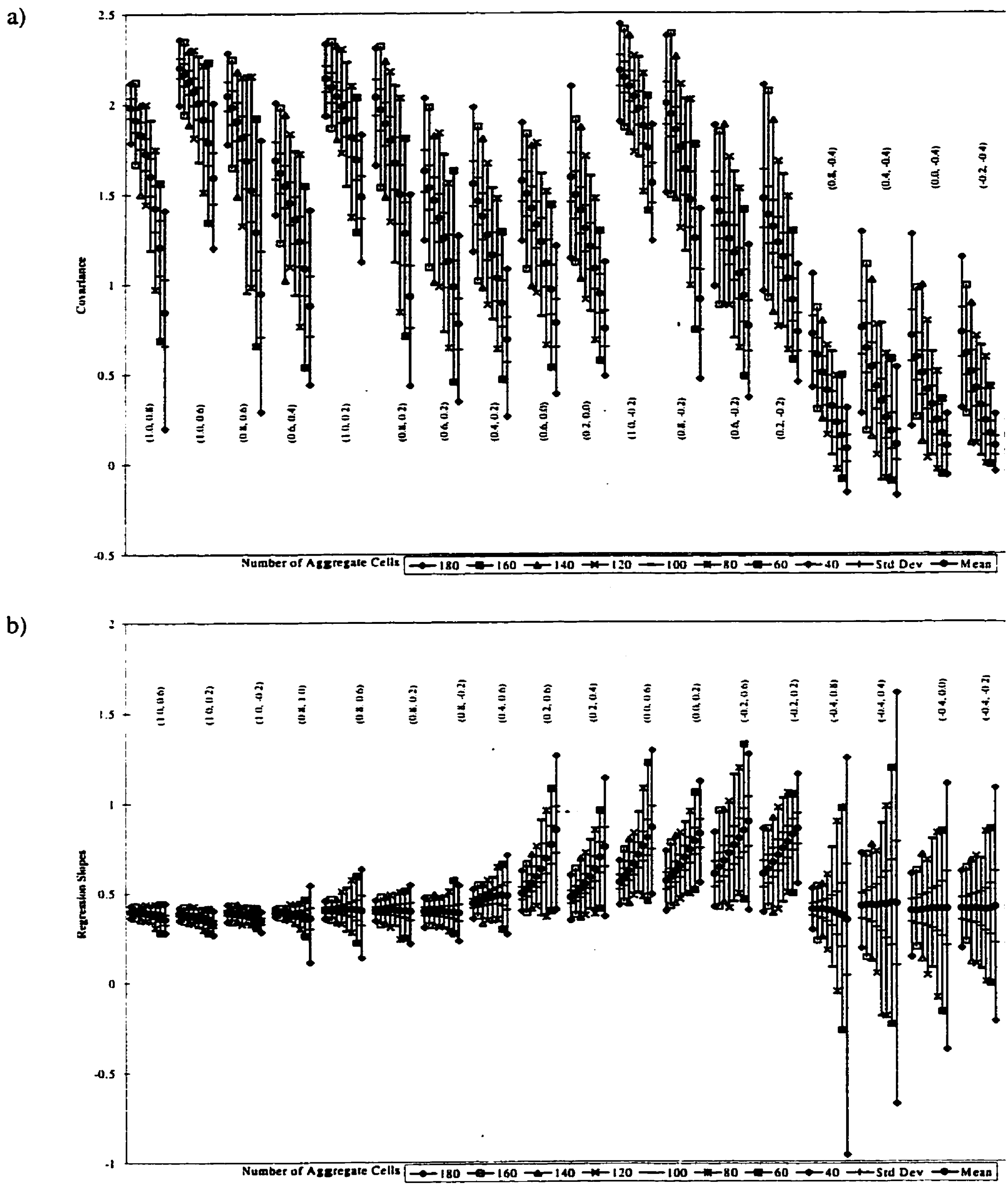


Figure 5.6: Variation of covariance (top) and upper triangle of the matrix of regression slope coefficients with the (MC independent, MC dependent) variables for an initial correlation of 0.4.

c)

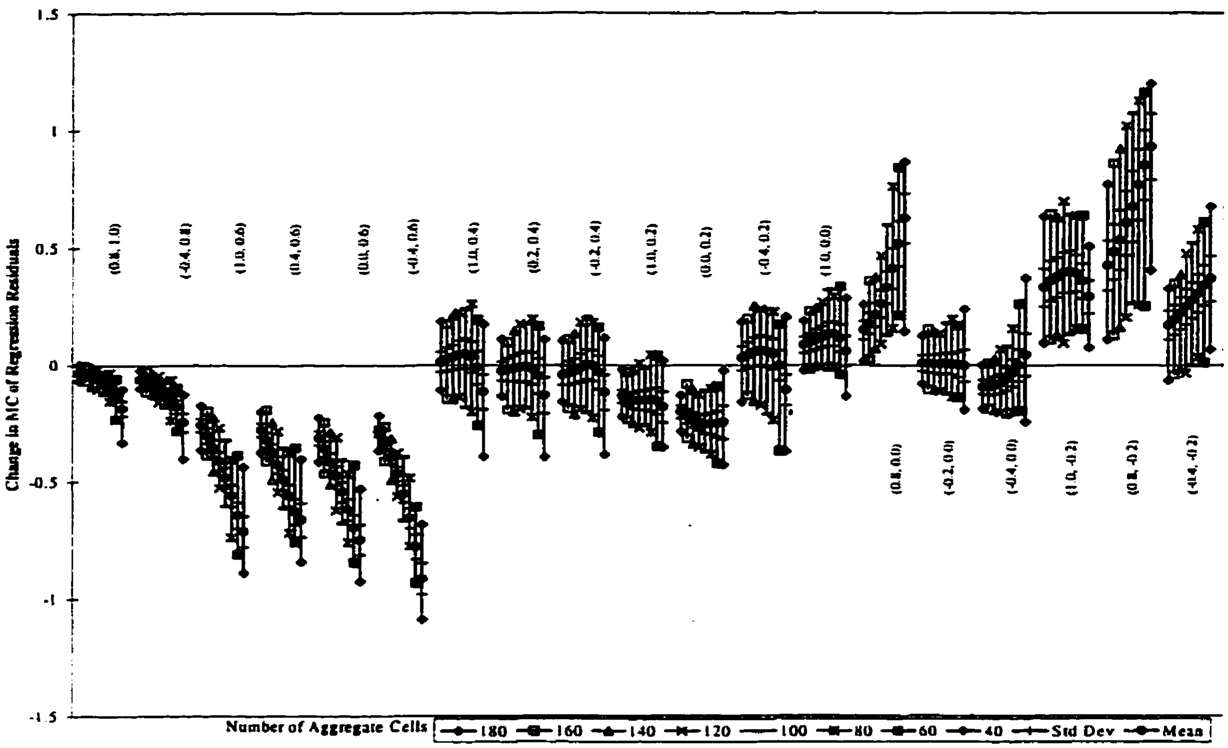


Figure 5.6 (con't): The change in the MC_{RR} with the (MC independent, MC dependent) variables for an initial correlation of 0.4. Again note the general lack of dependence on the independent variable, and how it tends to decrease for the high MCs and increase for the low MCs, indicating a general trend towards random autocorrelations.

a)

b)

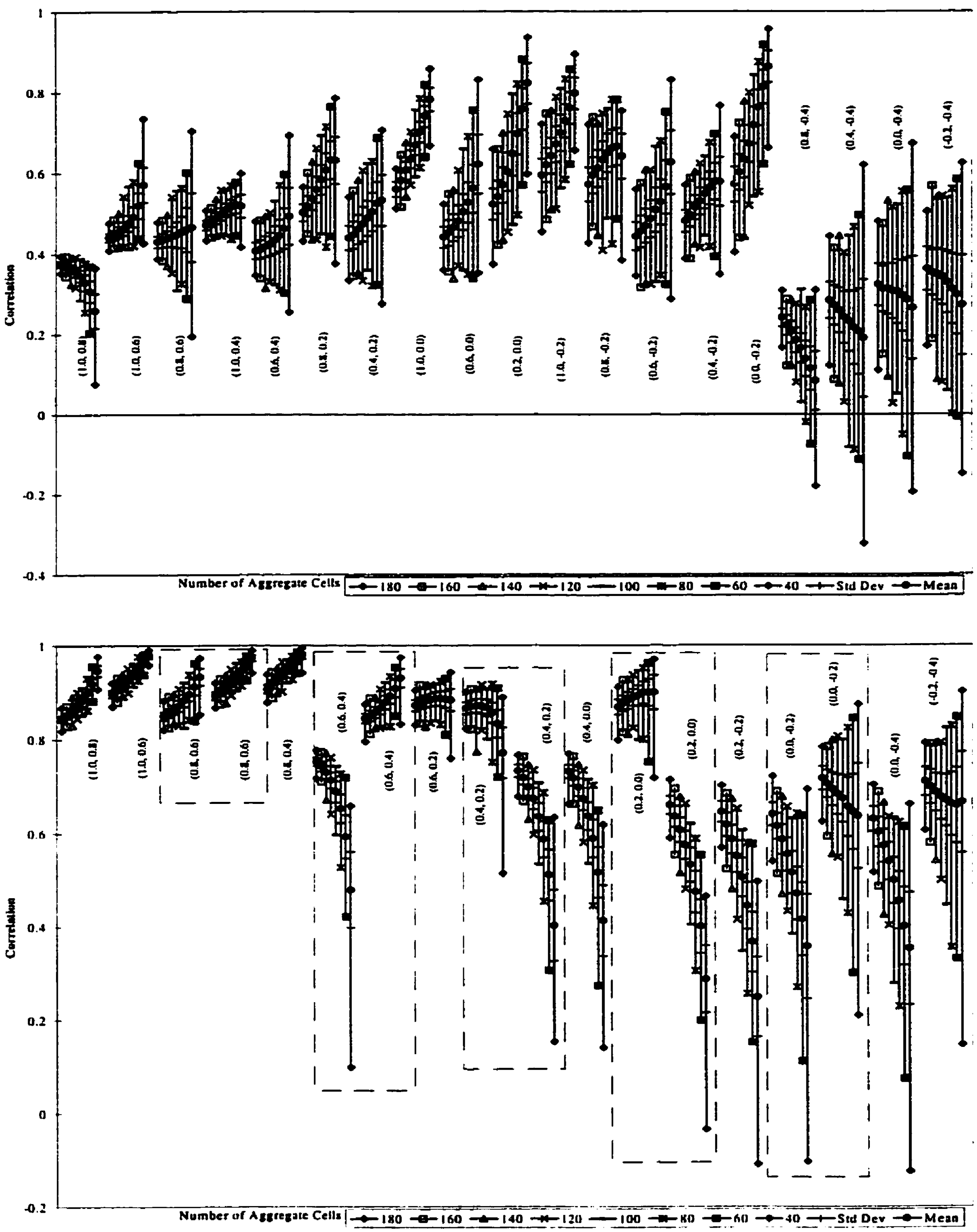


Figure 5.7: Variation of correlation with the (MC independent, MC dependent) variables for initial correlations of 0.4 (top) and 0.8. Note the often wide variation in behaviour of correlations in the dashed boxes where the dependent and independent variables have the same MCs, likely caused by differences in spatial arrangements of the variables.

a)

b)

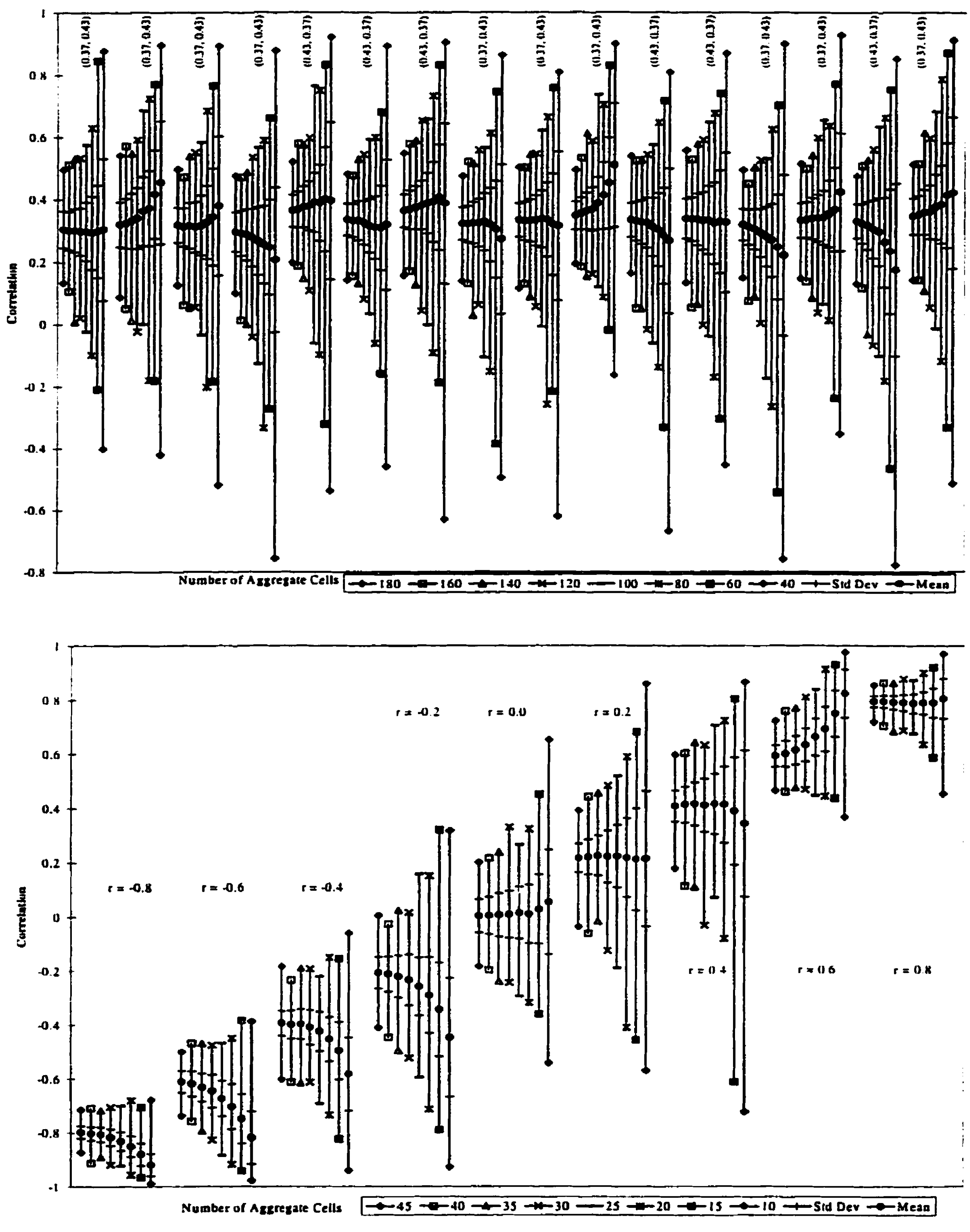


Figure 5.8: Variation of correlation for several combinations of variables whose MCs and correlations mimic those used in Openshaw and Taylor (1979) (top), and for a set of variables with MCs of 0.4 and different correlations (bottom). These results generally agree with those of Openshaw and Taylor.

6. The Effects of Aggregation on Multivariate Regression Parameters[1]

6.1. Summary

Several empirical studies of the Modifiable Area Unit Problem (MAUP) have been performed on census data, one of which has been about its effects on multivariate regression analysis. Recognizing that as much control as possible needs to be exerted in order to effectively study the MAUP, a spatial dataset generator was created that allows the user to construct sets of variables with various spatial and aspatial properties. The effect of aggregation on multivariate regression parameters, with special attention to the influence of spatial autocorrelation, is studied using a number of synthetic datasets created by the data generator. It is found that the effects depend on the combinations of autocorrelations of the unaggregated dependent and independent variables. It is also found that aggregation introduces collinearities between independent variables where none existed before. The patterns displayed provide hope that the effects of the MAUP on multivariate regression may not be as unpredictable as was once feared.

6.2. Introduction

The Modifiable Area Unit Problem (MAUP), a term introduced in Openshaw and Taylor's (1979) classic chapter, has long been recognized as a potentially troublesome feature of spatially aggregated data, such as census data. Aggregation of high-resolution (i.e. a large number of small spatial units) data to lower resolution (i.e. a smaller number of larger spatial units) 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 more than likely change because of a loss of information. This phenomenon is called the *scale effect*. The N 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 ways to partition a region into N cells. Each partitioning will result in different values for the aggregated statistics; this variation in values is known as the *zoning effect*. The two effects are not independent, because the lower-

[1] This chapter is based on Reynolds and Amrhein, 1998b, and was actually written before the other papers.

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 choice of aggregation.

Several studies (for example, Amrhein and Reynolds, 1996, 1997; Fotheringham and Wong, 1991; Amrhein and Flowerdew, 1993; Openshaw and Taylor, 1979) have been published that study the effects of the MAUP on a number of census datasets. Of these, only Fotheringham and Wong (1991) have examined the effects of the MAUP on multiple regression parameters, pessimistically concluding that its effects on multivariate analysis are essentially unpredictable. Amrhein (1993) presents the results of a statistical simulation of the MAUP by aggregating randomly-generated point data into square grids of various sizes, thus avoiding many of the problems associated with the use of census data. This chapter expands upon the ideas from both, using statistical simulations to study the effects of the MAUP on multivariate analysis. The fact that Steel and Holt's (1996) analytically derived rules for random aggregation agree with Amrhein's (1993) empirical rules corroborates that simulations are an effective tool for examining the effects of the MAUP.

6.3. The synthetic spatial dataset generator

The use of census data imposes a serious constraint upon those who seek to understand the mechanics of the MAUP simply because there is no control over the nature of a region's overall shape; the shapes, sizes and connectivities of its subregions; or the ranges, means, variances and covariances, frequency distributions, and spatial autocorrelations of the variables. The effects of aggregation on a given census variable can be determined readily enough, but few clues to underlying processes can be gleaned because the data cannot be systematically varied to test for the effects of changes. Other weaknesses of census data, such as random rounding and values missing due to the absence or suppression of data, only serve to make the drawing of any conclusions even more difficult. In order to study the MAUP, it is therefore advantageous to be able to construct synthetic spatial datasets over which a researcher can control and systematically vary all of the above features. This chapter employs the dataset generator described in detail in Chapter 3. Figure 6.1 illustrates the region used for the experiments, which is divided into 400 subregions, along with three sample aggregations.

6.4. The experiments

Spatial autocorrelation is known to play a key role in the MAUP, as is illustrated in the following experiment. Consider a spatial dataset that contains negative spatial autocorrelation; that is, numbers that are dissimilar are located in adjoining regions. In the aggregation process, contiguous regions are joined and the individual variable values are (in this case) replaced by their average, hence creating a new dataset with a reduced variance. With some algebra, it is easy to show that the difference between the original variance and the aggregate variance (weighted by the number of units in each cell) is the sum (again weighted by the number of units) of the variance of the regions within each cell. For the negatively autocorrelated dataset, it is expected that the values in each cell will have a high variance, and hence the change in variance will be relatively large. As the spatial autocorrelation becomes more positive, the expected internal variance within each cell should decrease, since similar values will tend to become more likely to be adjacent, and hence the change in variance should become less. The influence of spatial autocorrelation on the behaviour of bivariate and multivariate statistics is more difficult to assess, however, as Chapter 5 demonstrates for the bivariate case, since each variable's MC and spatial pattern will cause it to respond to aggregation differently.

The experiments in this chapter explore the effects of aggregation on the various parameters of the linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. Three independent parameters are considered to be sufficient to capture enough of the complexities involved in multivariate linear regression without creating excessive computational and analytical overhead. Fotheringham and Wong (1991) use a four-variable regression model, in which the variables are all proportions; their results are compared to ours here.

Three different experiments are performed. In the first, y, x_1, x_2, and x_3 are all assigned the same level of spatial autocorrelation (as measured by the MC). Eight datasets are created in which all four variables have MCs of -0.4, -0.2, 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0 respectively, and have zero correlation between them. In the second experiment, x_1, x_2, and x_3 are assigned the same MC, while y is given a different one and again all variables are uncorrelated. Datasets are created with MCs for dependent and independent variables chosen from -0.4, 0.0, 0.4, and 0.8, for a total of twelve combinations. The third experiment counts the number of statistically significant changes in correlations between variables for the datasets of the first experiment in order to estimate the potential for introduced collinearities. Obviously, having variables with no collinearity is an idealized case, since most variables will have some degree of correlation between them, but it is a good place to start.

The aggregation algorithm is described in detail in Chapter 3. For these experiments, as in Chapters 4 and 5, the regions are aggregated to M = 180, 160, 140, 120, 100, 80, 60, and 40 cells, representing from 45% to 10% of the original 400 regions, in order to assess the scale effect of the MAUP. All of these aggregations are performed independently in a run of the model, and each run is independent of the previous runs. To account for the variability of results introduced by the zoning effect, 1000 runs of the model are performed. After each aggregation, the data are fitted to the multiple linear regression model and the resulting parameters, plus the Moran Coefficient of the regression residuals (MC_{RR}), are saved.

Once all aggregations are completed, the maximum, minimum, mean and standard deviation of each parameter for each scale of aggregation are computed and saved for analysis. The analysis plots (see Figure 6.2b as an example, and Chapter 3 for a more detailed description) are arranged in groups of eight lines, one line for each scale of aggregation, with the labels for each line being listed in the plot's legend. Each group represents a set of initial conditions for an experiment, and is labeled on the plot with (MC_x, MC_y), where MC_x is the MC of the independent variables and MC_y that of the dependent variable. Each line represents the range of values of the parameter that are obtained for the scale over all the runs, and is also marked by the mean value (a heavy dot) and at the mean ± 1 standard deviation (a small horizontal line) to give a rough idea of the distribution of values.

6.5. Results

The results from the first experiment, in which the Moran Coefficients for the dependent and independent variables are the same, show that all of the multivariate regression parameters vary systematically with a change of scale and also with the level of spatial autocorrelation latent in the data. Figures 6.2 to 6.4 illustrate the variations in R^2, the MC of the residuals, and the values for β_0, β_1, and their standard errors; figures for β_2, and β_3 are similar to those of β_1, and are not shown. All of the figures show the same pattern, with the ranges for all scales decreasing with increasing spatial autocorrelation. This conforms to expectations, since we expect the scale effect to be less severe with greater positive autocorrelation due to more similar values tending to be aggregated. The figures also show that the variation of all parameter values increases with the magnitude of the scale effect over all levels of spatial autocorrelation. This again agrees with expectations, since more information is lost as the data values are aggregated into fewer cells, and with a larger number of regions going into each cell it is expected that there would be a greater degree of variation in results caused by the choice of partition, even for highly spatially autocorrelated data.

Since all the variables are generated randomly and are mutually uncorrelated, the values of R^2 for the unaggregated datasets are all close to zero. Figure 6.2a illustrates that aggregation can produce a model that can have, in extreme cases, from 20% to even 70% of the variation explained by the model, depending on the scale of aggregation and the spatial autocorrelation of the data. The distance of the maximum extreme values from the mean plus one standard deviation mark indicate they are all outliers in the frequency distributions, and as such they will tend to increase the mean value. But even with that in mind it is still apparent that aggregation tends to give models with better fits than the original data, with better fits being associated with greater aggregation. This agrees with expectations, since a reduction in the variability of the data values will tend to produce a better-fitting model (if covariance is also not reduced), but the loss of information caused by reducing the sample size offsets any apparent gain.

Figure 6.2b illustrates the change of the MC_{RR} with aggregation. One of the basic assumptions of a linear regression model is that the residuals are independent, and it is clear that this assumption is being violated since spatially autocorrelated residuals are not independent[2]. Since the initial correlations between the variables are all zero, all of the regression slope parameters are also initially zero so that the initial MC_{RR} will simply be the MC of the deviation of y about its mean, which equals the MC of y. The diagram illustrates the tendency for the regression residuals to become more randomly autocorrelated, with that for the initially negative residuals tending to increase, while that for the initially positive ones tending to decrease. The change in residuals for the MC of 1.0 does not follow the pattern of the rest of them, but still does tend to decrease slightly. As with the findings of Chapter 5, it appears that aggregation

---

[2] Since each observation can be partly predicted from its neighbours, the information content of observations is reduced. See Section 5.3, Griffith (1988, pp. 82-83), and Cliff and Ord (1981, p. 199) for details.

tends to improve the statistical quality of linear regression, even though it changes all of the parameter values.

Figures 6.3 and 6.4 show that the regression coefficients and their standard errors behave similarly under aggregation. The mean values of the β_0 and β_1 estimates b_0 and b_1 remain close to their unaggregated values over all levels of spatial autocorrelation and all scales. In contrast, the average value of the standard error for all coefficients shows a definite increase with the scale effect. This is not unexpected, as Fotheringham and Wong (1991) point out, since the standard error depends partly on the number of aggregated units. Interestingly, even though the range of variation of the standard error due to the zoning effect decreases with increasing spatial autocorrelation, the mean value for a given scale remains essentially constant. The β_2 and β_3 coefficient estimates b_2 and b_3 and their standard errors behave similarly and are not shown.

The results of the second experiment, in which the independent variables x_1, x_2 and x_3 contain the same level of spatial autocorrelation, while y has a different one, are presented in Figures 6.5 to 6.7. Each plot consists of 12 groups of lines, with each group representing a combination of MCs for the dependent and independent variables. The groups are organized in four sets of three, with each set's dependent variable having the same Moran Coefficient.

As before, the range of variation of the various parameters increases as the scale decreases. Figure 6.5a shows that the range of R^2 decreases as the MC of both the independent and dependent variables increases, though it appears to decrease faster with the increase in the independent variables' MC than with the dependent variable's. This is consistent with the results shown in Figure 6.2a and indicates that, as before, less information is lost when the variables are highly autocorrelated, resulting in smaller variations of the aggregated statistic values.

By examining Figure 6.5b and comparing it to Figure 6.2b, it is apparent that the behaviour of the MC of the residuals depends more on the spatial autocorrelation of the dependent variable than that of the independent variables, since the distributions do not change significantly with the MC of the independent variables. As explained above, this is due to the initial values of the slope parameters being zero, resulting in the initial MC_{RR} being the MC of the dependent variable. As before, the behaviour will depend on the spatial pattern of the variables, not just on their MCs.

As with the first experiment, the regression coefficients and their standard errors each behave in roughly the same way for each combination of spatial autocorrelations. There are three clearly visible patterns, aside from the usual increase in variability with decreasing aggregation scale. First, the mean values of the distributions for the regression coefficients tend to remain fairly stable as the number of aggregate cells decreases, while the means of the standard errors tend to increase. Second, for a given MC of the independent variable, the variability of the ranges increases with increasing MC of the dependent variable, though this effect becomes much less dramatic as the MC of the independent variables increases. The size of some of the ranges is interesting, especially with the intercept parameter b_0 which can be almost 80 above or below the mean of 20 for the 40-cell case in the third from last group in Figure 6.6a. Third, for a given MC of the dependent variable, the range decreases with increasing MC of the independent variables. The patterns are reflected in the those for the standard errors, as shown in Figures 6 and 7 for b_0 and b_1 (those for b_2 and b_3 are similar and not shown). Since the multivariate linear regression model parameter estimates are of the form $\mathbf{b}=(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})$, it is expected that variations in the spatial autocorrelation of the independent variables **X** will influence the outcome more than those of the dependent variable **Y**. These figures should serve as a clear warning to those who would blindly use multivariate regression methods on aggregated georeferenced data and then expect the results to apply to a higher resolution!

Comparison of these results with those of Fotheringham and Wong (1991) is difficult because the dependent and each of their four independent variables had a different MC, ranging from almost 0.9 for their P^{black} to about 0.25 for P^{eld}. Even from the very simple second experiment, it is clear that having the dependent and independent variables with different MCs increases the complexity of the response of the regression parameters to aggregation. Differences in the spatial patterns of the variables, as shown above, can also hamper comparisons, as results may be very different for variables with the same MCs.

Fotheringham and Wong's (1991) (hereafter referred to as FW for brevity) analysis of the change in Moran Coefficients of the variables can be compared with experimental results, however, using the diagrams of Chapter 4. Even though the change in the MC depends on the spatial arrangement of the variable, Figures 4.2b, 4.4a, and 4.8 show that the distributions widen as the number of aggregate cells decreases (also shown in FW's Figure 6), and that the mean value either decreases or increases monotonically, unlike most of the examples in their Figure 6 which increase and then decrease. These differences could be the result of FW's performing only 20 random aggregations for each spatial scale (20 being not nearly enough to approximate the true distribution of aggregate values), having more than twice the number of base units as we used, and using proportional variables (i.e. numerator and denominator are aggregated separately and the results divided) rather than variables that are simply summed or averaged, or perhaps to unknown violations of the regression model assumptions. Further research needs to be done to study the effects of the MAUP on proportion-type variables.

Also of interest in a study of multivariate linear regression are conditions that violate the assumptions of the model. The easiest one to study is collinearity, the presence of correlation between the independent variables[3]. For this experiment, the datasets used in the first experiment, which all have zero correlation between the variables, are aggregated in the model as before and the number of correlations that are statistically significantly different from zero are counted for each level of aggregation. Table 6.1 summarizes the results for the sets that have MCs of -0.4, 0.2, and 0.8 for the aggregation levels of 180, 100, and 40 cells, while Figure 6.8 illustrates the variation of correlation with MC for the datasets whose variables have the MCs of -0.4 and 0.8. Note that the values in the row labelled *Any* will be less than the sum of the values in the columns if more than one of the correlations is significant at the same time, which occurs frequently for the -0.4 MC case at all levels of aggregation, but less so for the other datasets.

Figure 6.8 and Table 6.1 demonstrate that the ranges of the introduced correlations decrease as the MCs of the variables increase, while as usual the ranges increase with decreasing numbers of cells. The reduction in the range is caused by the decreasing amount of variability lost as the variables become more positively spatially autocorrelated, so as the range decreases fewer values in the distribution cross over into the critical range. As illustrated in Chapter 5, predicting how a pre-existing non-zero correlation between two of the variables will be affected by aggregation is not simple, as the change will depend on the interaction between the spatial

[3] Note that the paper which forms this chapter was initially written before my more detailed analysis of bivariate statistics in Chapter 5. Since the counting of significant changes in r was not a topic discussed in Chapter 5, I decided to leave this in as is.

distributions of the variables. The fact that there can be significant changes in the collinearities reinforces the need for caution when using multivariate regression techniques on aggregated data.

6.6. Conclusions

In order to systematically examine the role of spatial autocorrelation in the data on the response of multivariate regression parameters to aggregation, a multiple linear regression model of the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ was employed, as three independent variables are sufficient to capture much of the complexity of multivariate regression while minimizing the computational and analytical overhead. The first two of the three experiments performed were designed to test the effect of various spatial autocorrelation levels in the independent and dependent variables on the variation of the regression parameters with aggregation. The third experiment tests to see how much collinearity is introduced between independent variables with increasing aggregation, when there was none in the unaggregated data.

When all variables have the same spatial autocorrelation, as measured by the Moran Coefficient, the variation of the parameters tends to decrease as the Moran Coefficient increases, as expected, indicating that more positively autocorrelated data are less affected by the MAUP. For all values of MC tested, the mean values of the coefficient estimates b_0, b_1, b_2 and b_3 are found to be essentially constant over all levels of resolution, even as the range of the distributions increases. Change in the variability is reflected in the standard errors for the coefficients, whose mean values and ranges tend to increase with decreasing spatial resolution. The mean value of R^2 shows a very large variability for negatively autocorrelated data that tends to decrease with increasing values of the Moran Coefficient. The change of the MC of the residuals depends on the MC of the dependent variable more than that of the independent variable, since the initial values of the β coefficients are zero and hence the initial MC_{RR} is that of the dependent variable.

When all of the independent variables have a particular Moran Coefficient, and the dependent variable has a different one, it appears that the MC of the independent variables tends to play a larger role in the variation of the regression coefficients, R^2, and the MC_{RR}, than does the MC of the dependent variable. For a given MC of the dependent variable, the variability in the coefficients and their standard errors tend to decrease with increasing MC of the independent variables. However, for a given MC of the independent variables, the variability tends to *in-*

crease with increasing MC of the dependent variable. The range of R^2 decreases as the MC of either the dependent or independent variables increase. It appears that the change in MC_{RR} depends on the MC of the dependent variable for initially uncorrelated variables.

Results from the third experiment reveal that collinearities between independent variables can be introduced by aggregation. The mean values of the ranges of correlations remain at or very near 0.0 for all resolutions and MCs of the variables. As one would expect, the ranges of the aggregate correlations are much greater for the variables with low or moderate MC than for those that are more highly autocorrelated, resulting in more statistically significant changes of correlations, many of which will occur simultaneously. Of course few datasets have no correlations between the variables, but it will be difficult to predict the change in a non-zero correlation until a way to incorporate the spatial patterns of the variables into the analysis is found.

The results of the experiments in this chapter only scratch the surface of the behaviour of multivariate regression parameters when data are aggregated from one level of spatial resolution to another. It is clear that the spatial autocorrelation of each of the variables involved influences the behaviour, and that if each variable has a different autocorrelation it will be difficult to predict ahead of time what the behaviour of the regression parameters will be. Exploration of the effect of the MAUP on multivariate regression using variously autocorrelated variables and various degrees of collinearity is a focus for future research.

The variables used in these experiments are all variables that were averaged during the aggregation process. The behaviour of variables that are proportions, in which numerator and denominator are aggregated individually, and variables that are summed in aggregation, also needs to be examined. Comparison of FW's results to ours indicates that multivariate models constructed with variable other than averaged variables may behave differently under aggregation from the model described in this chapter. Models that involve combinations of different variable types may behave even more differently. All of these require further research.

The ultimate goal of the research is, of course, to see if it is possible to empirically estimate error in a spatial dataset that has been introduced by aggregation, and the presence of recognizable patterns indicates that the prospects are perhaps not as gloomy as FW first believed.

Table 6.1: Total number of statistically significant correlations between the variables created by the aggregation process. The number of instances when any of the combinations produced a significant correlation is recorded in the row labelled Any.

	MC = -0.4			MC = 0.2			MC = 0.8		
Cells	180	100	40	180	100	40	180	100	40
y, x_1	20	52	64	1	2	12	0	0	3
y, x_2	13	60	65	0	1	6	0	0	2
y, x_3	27	42	79	0	2	10	0	0	2
x_1, x_2	20	57	60	0	2	1	0	0	0
x_1, x_3	33	45	61	0	0	10	0	0	4
x_2, x_3	14	54	71	0	0	6	0	0	1
Any	120	279	337	1	7	33	0	0	12

6.7. 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.

Amrhein, C. G., and R. Flowerdew, 1993: Searching for the elusive aggregation effect: Evidence from British census data. Unpublished manuscript available from the authors.

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

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

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

Reynolds, H., and C. G. Amrhein, 1998b: 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.

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.

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

6.8. Figures for Chapter 6

a)

b)

c)

d)

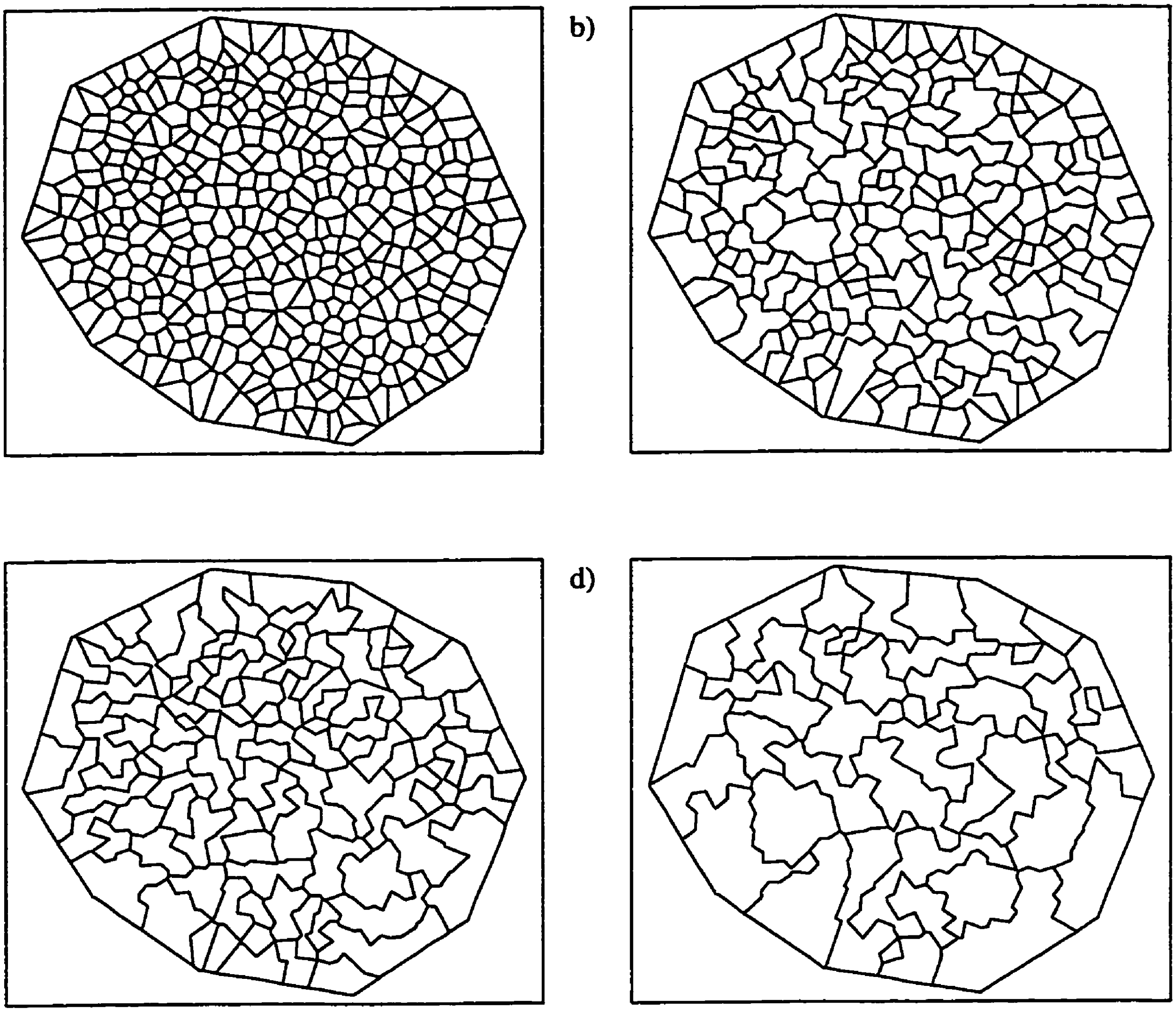


Figure 6.1: The synthetic region used in all of the experiments, with its 400 cells (a) and a sample aggregations to 180 cells (b), 100 cells (c) and 40 cells (d).

a)

b)

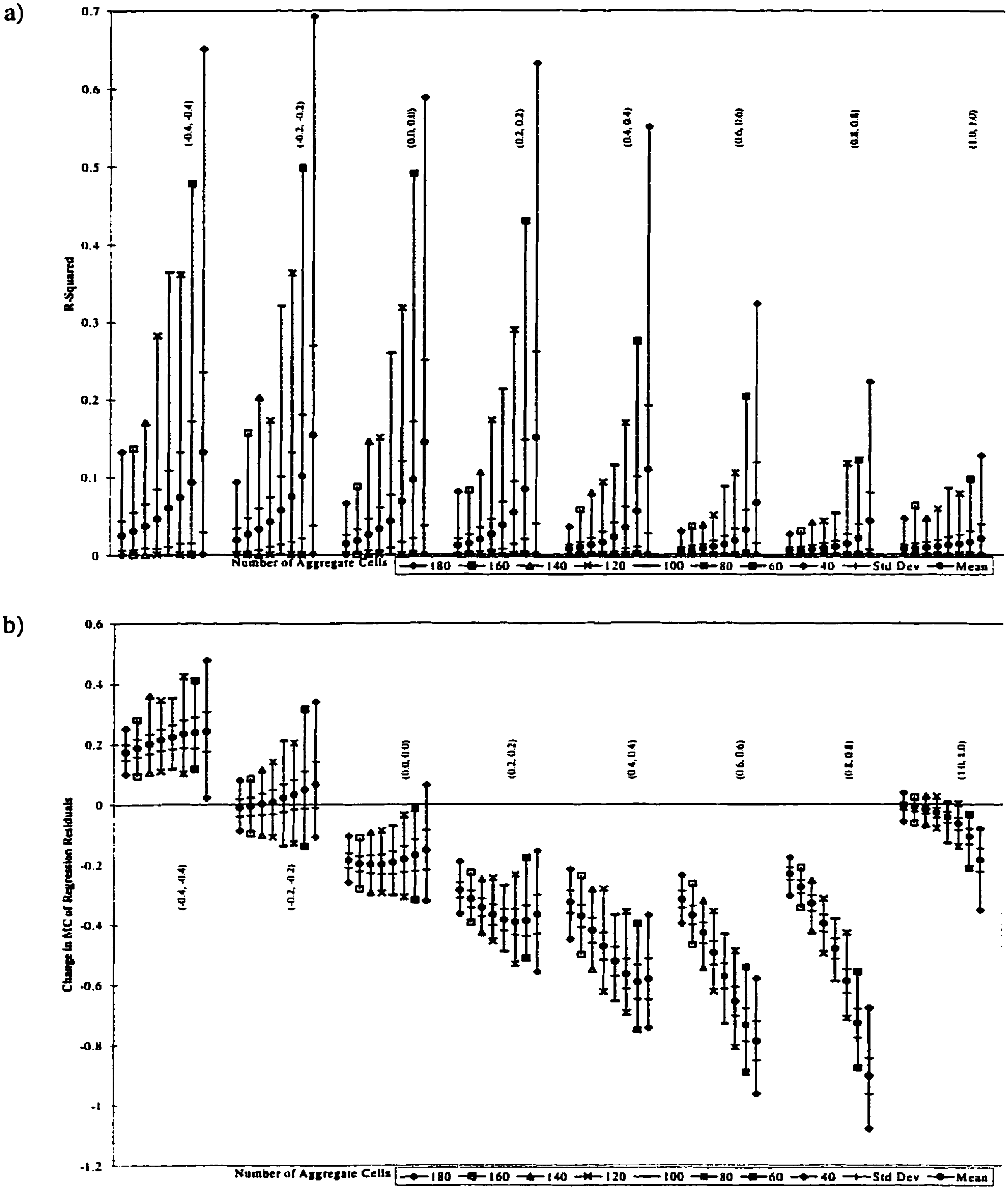
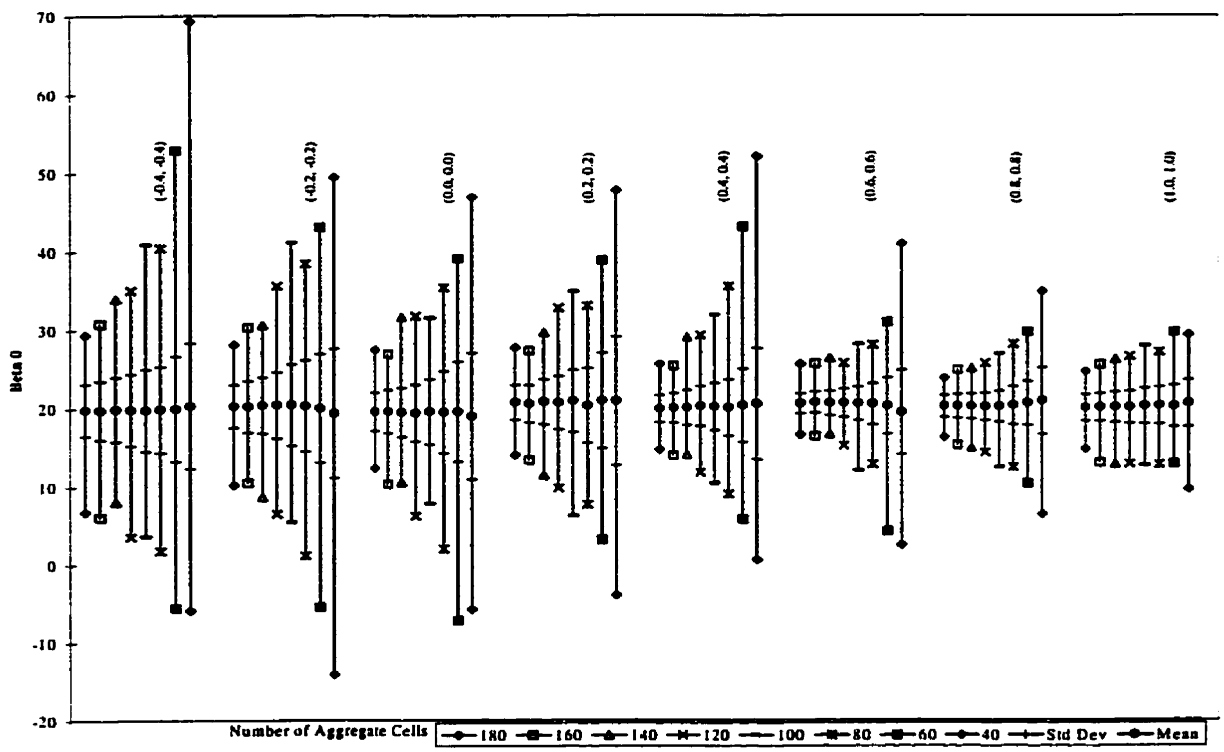


Figure 6.2: Variation of R^2 (top) and the change of Moran Coefficient of the multivariate regression residual with aggregation over 1000 runs of the aggregation model, with dependent and independent variables having the same Moran Coefficient.

a)



b)

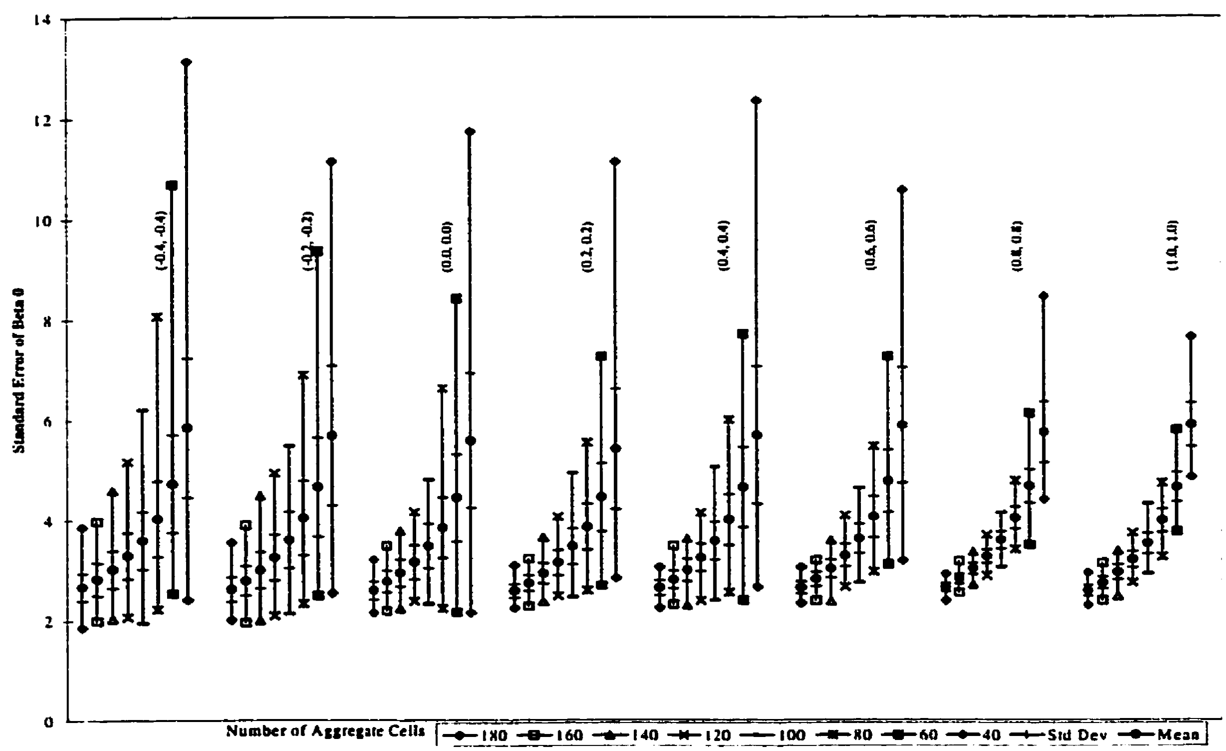
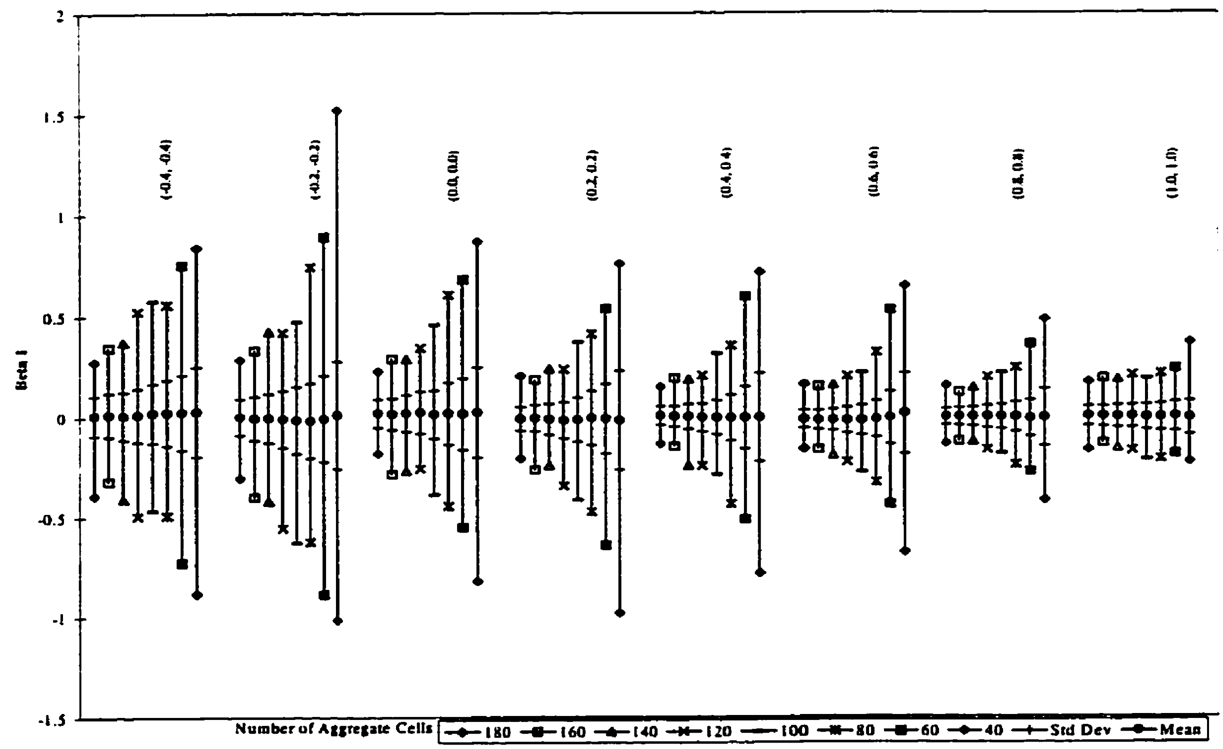


Figure 6.3: Variation of the multivariate regression parameter β_0 and its standard error over 1000 runs of the model, with dependent and independent variables having the same Moran Coefficients. Note how the variability decreases with increasing MC.

a)



b)

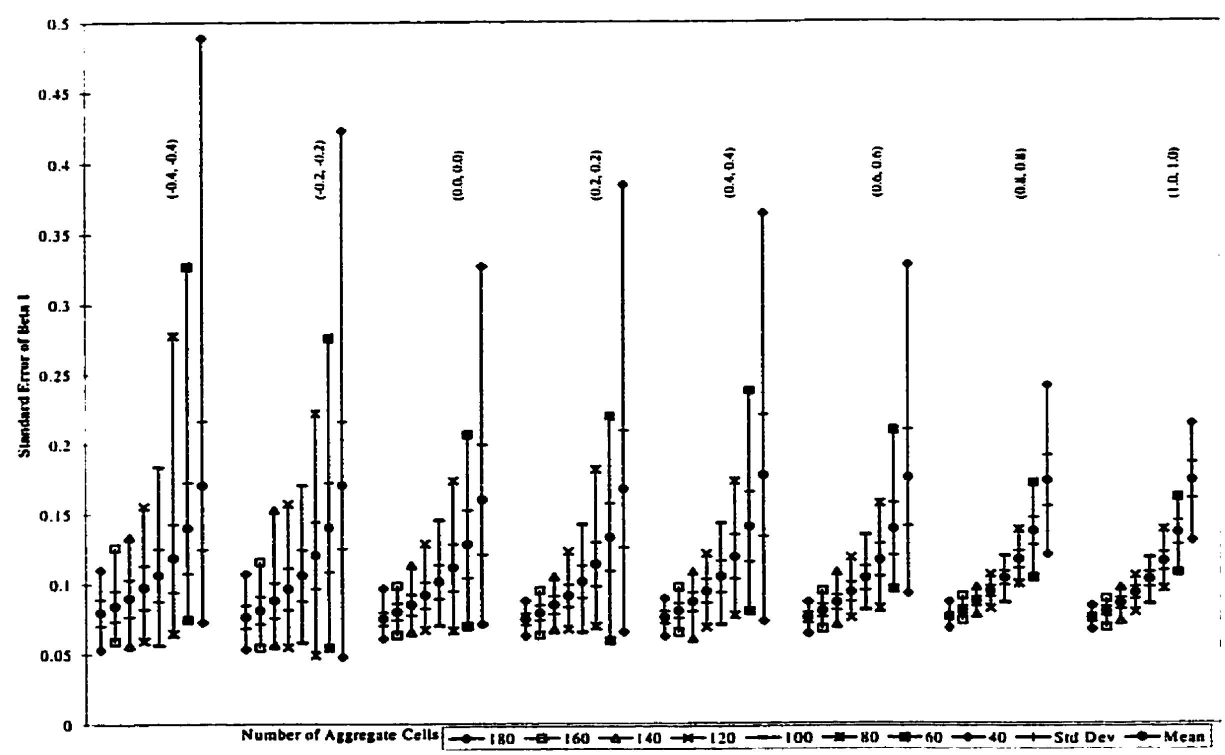


Figure 6.4: Variation of the multivariate regression parameter β_1 and its standard error over 1000 runs of the model, with dependent and independent variables having the same Moran Coefficients. Note how the variability decreases with increasing MC.

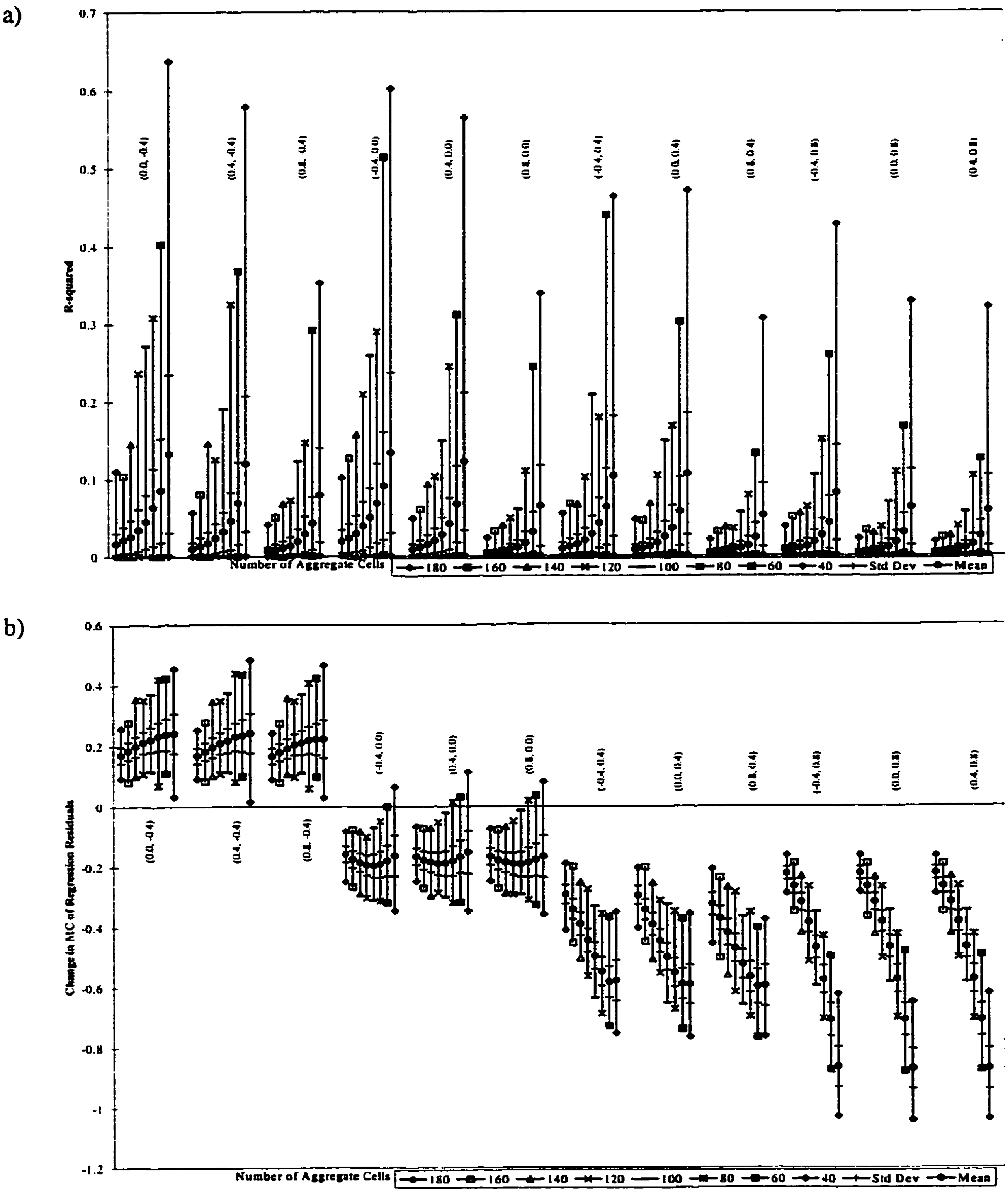
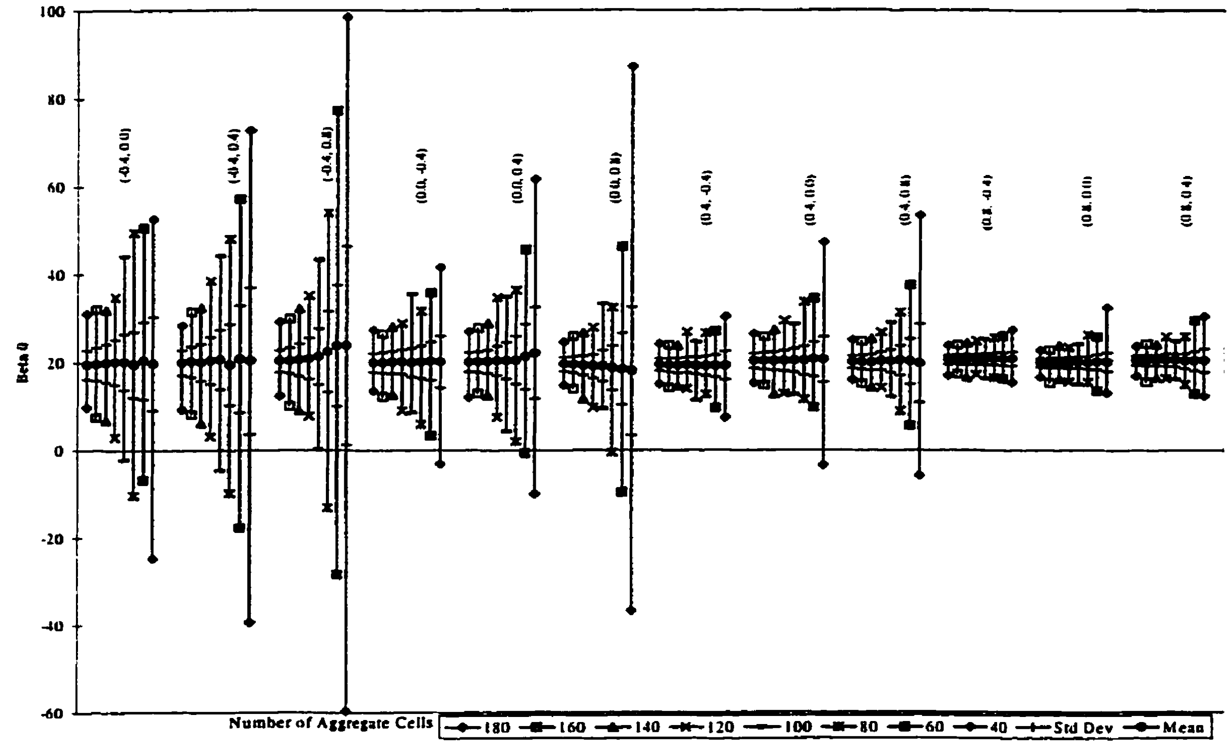


Figure 6.5: Variation of the multivariate R^2 (top) and the change of the Moran Coefficient of the regression residual (bottom), with the independent variables having the same MC and the dependent variable having a different MC.

a)



b)

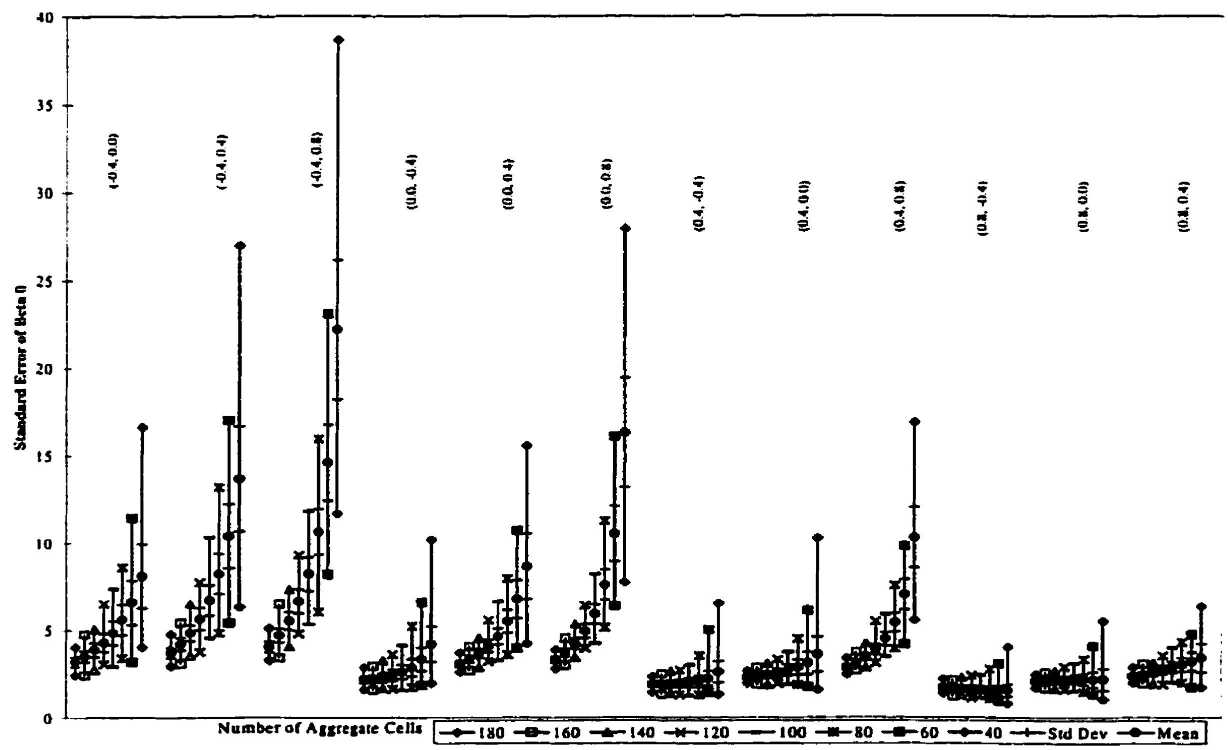
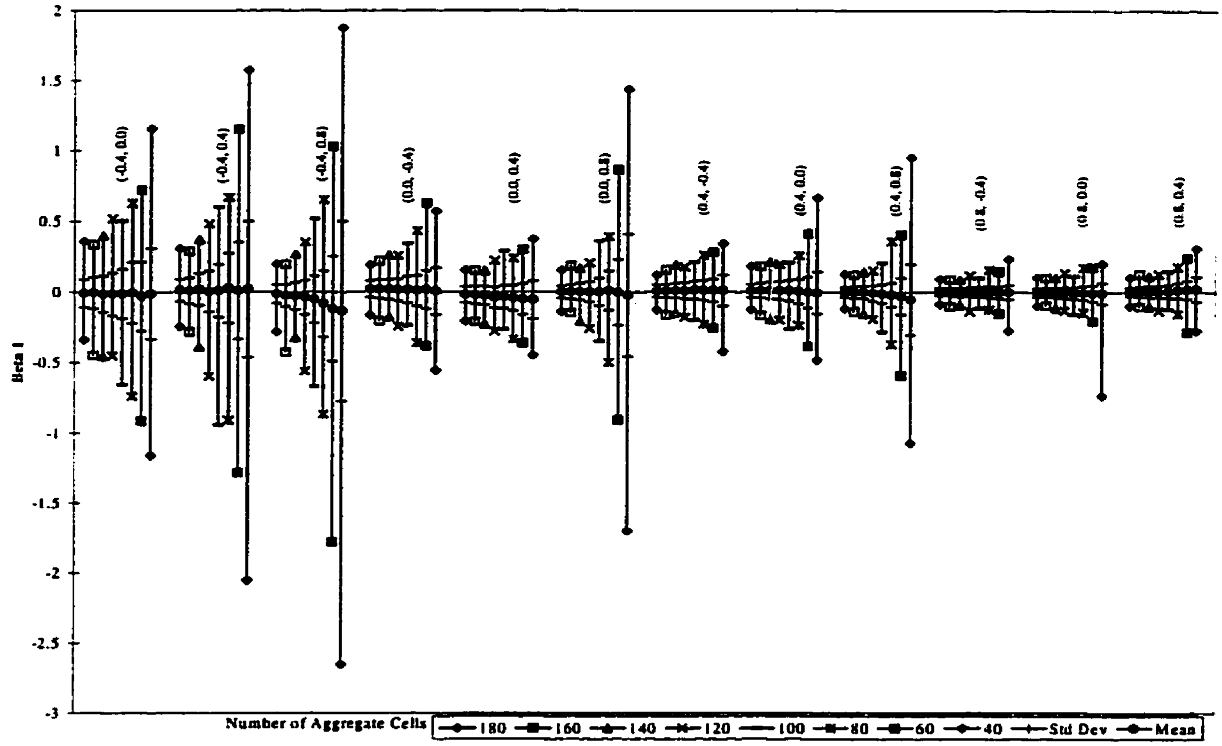


Figure 6.6: Variation of the multivariate regression parameter β_0 (top) and its standard error, with the independent variables having the same MC and the dependent variable having a different MC.

a)



b)

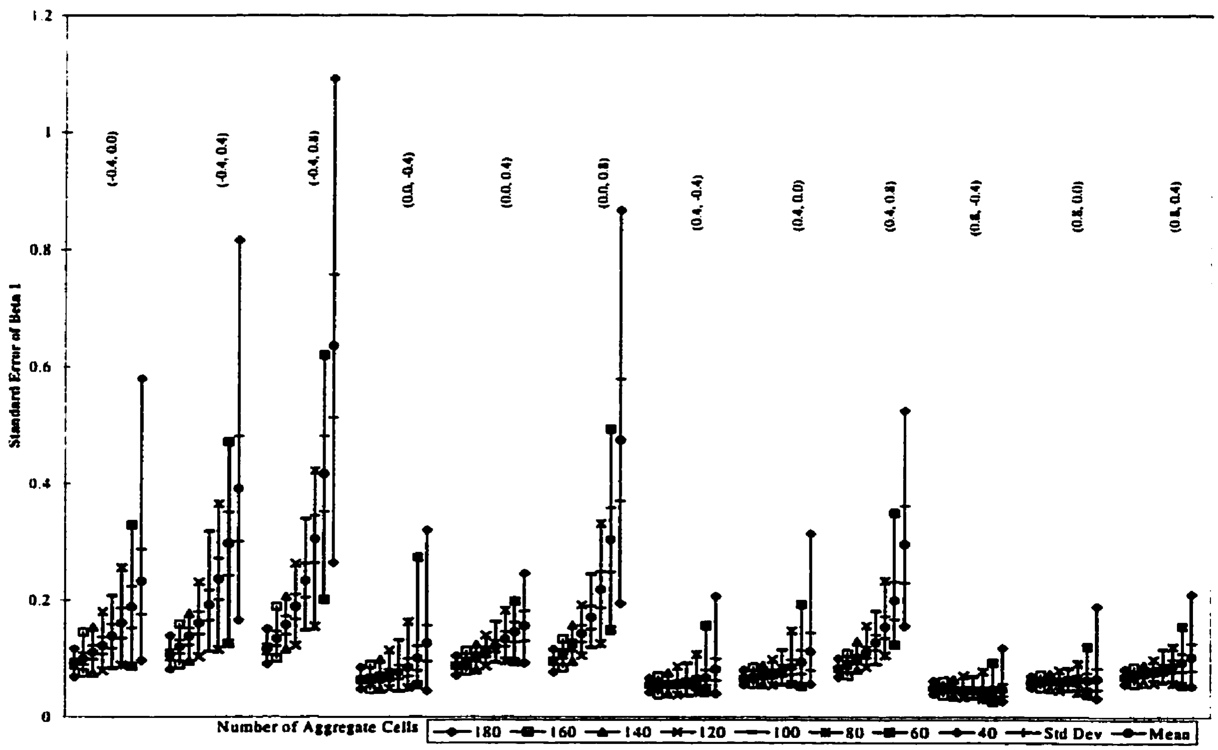


Figure 6.7: Variation of the multivariate regression parameter β_1 (top) and its standard error, with the independent variables having the same MC and the dependent variable having a different MC.

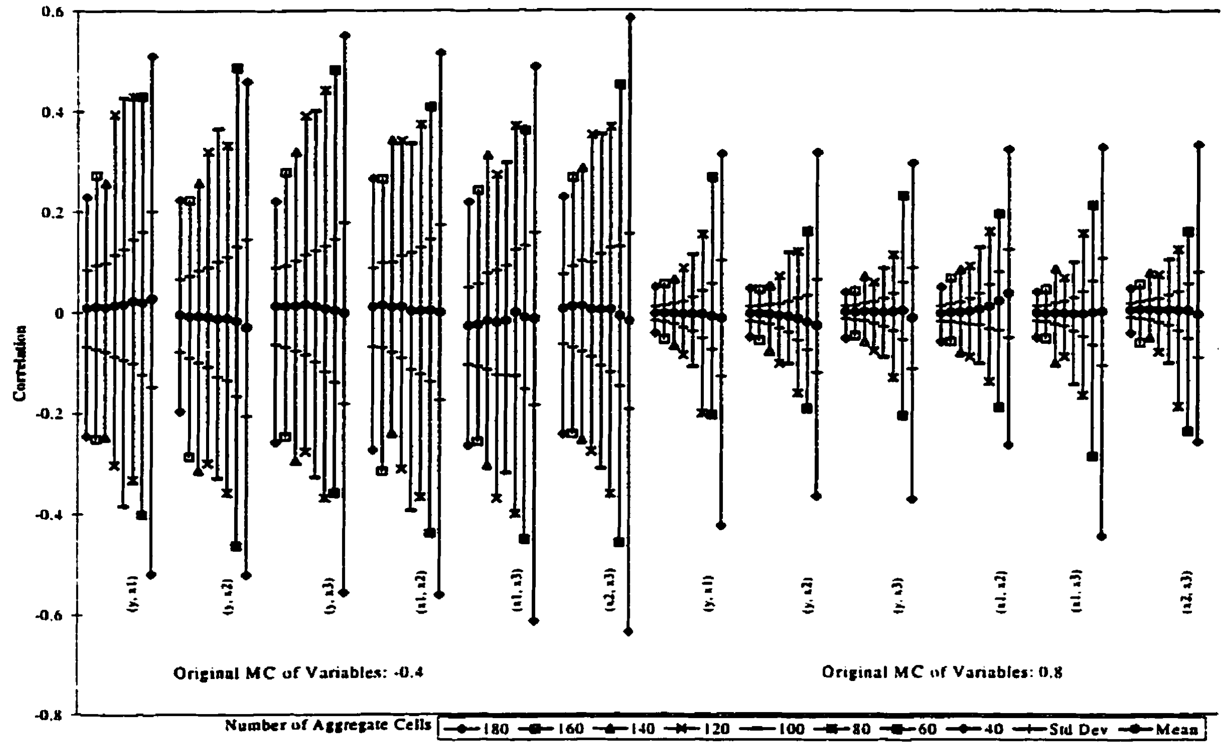


Figure 6.8: Variation of correlation with aggregation for the datasets of experiment 1 in which the original MCs of the variables are -0.4 (left) and 0.8 (right).

7. Summary of Conclusions

The results of this research clearly demonstrate why the Modifiable Area Unit Problem has been such a source of frustration for spatial analysts for so long. Even a relatively simple statistic like the weighted variance behaves in a complex manner, influenced by the spatial autocorrelation and arrangement of the unaggregated variable. More complex statistics, like the Moran Coefficient, correlation, covariance, and the bivariate regression slope parameters, are affected by the spatial arrangements of both variables, while the multivariate regression parameters are affected by those of all variables involved. Unfortunately, results reported in Chapter 4 amply indicate that the MC is not a sufficient measure of spatial organization for the purposes of prediction of results, since many different types of arrangement can have the same MC, and it is often the arrangement for the given MC that determines how a variable will behave under aggregation. Even so, it is still useful as a first approximation in most cases, and further research may be able to provide a summary statistic that can include pattern as well as spatial autocorrelation.

One of the common features to all the experiments is that the frequency distributions (which are a result of the zoning effect) of all of the aggregated statistics are either normally distributed or nearly so. The assumption of a normal distribution plays a pivotal role in most inferential statistical theory, so this empirical finding may help to further advance theoretical investigations of the MAUP. The finding is surprising, especially for something as complex as a MC of a regression residual, because due to Murphy's Law I would expect a distribution that would make the analysis of the MAUP with statistical theory even more difficult[1].

The relative change in variance shows a strong dependence on the spatial autocorrelation of the original variable, which of course is no surprise, but it also depends on the spatial arrangement of values. The aggregated Moran Coefficient depends not just on the initial spatial autocorrelation, but also on the spatial arrangement of the values, especially as the original MC increases and patterns become more distinct. Patterns with a large number of small clusters of similar values will show the greatest change in aggregate univariate statistics as the number of cells decreases because as the cell size increases, the likelihood of including regions with dissimilar values increases faster than it does when there are only a few large clusters. A more precise definition of the relationship must await a better way to describe the spatial arrangement of

[1] OK, this is a bit cynical. Maybe I have been a post-graduate for too long.

the data values, perhaps by using two or more spatial autocorrelation statistics in conjunction with each other.

The relative change in variance is strongly non-linearly correlated to the G statistic, which has been modified by dividing by the unweighted aggregate variance. This dependence does not appear to be because the unweighted aggregate variance is present on both sides of the regression equation, though what causes it and how it can be exploited are worth future research.

The covariance tends to behave in a similar way to the variance under aggregation, in spite of the possibility for it to increase or decrease. The range of the distributions of both statistics decreases with the decreasing number of aggregate cells for low values of spatial autocorrelation of variables, since increasing the cell size will not appreciably increase the (co)variation within each cell that can be lost by aggregation. As the MC increases, the within-cell variability will tend to increase with an increase in cell size as more dissimilar values are included, with the rate of this increase depending on the spatial arrangment (many small or fewer larger clusters).

When both variables have the same MC, the ranges of the covariance, correlation and regression slope parameter tend to increase as MC decreases, and to increase as the number of aggregate cells decreases. The MC of the regression residual (MC_{RR}) is not much affected by the initial correlation of the variables, but changes considerably with the increase in MC of the variables, showing a marked tendency to decrease as the number of aggregate cells decreases. This indicates that the statistical quality of regression results can actually be improved with aggregation, even though the values of the parameters are quite different from the original. This apparent improvement is offset by the loss of information caused by the reduction in sample size. When the variables have different MCs and the initial correlation is zero, the behaviour is still reasonably regular. The range of correlations tends to increase as the MC of the variables decreases, and the range of regression slope parameters is greatest when the MCs of the variables are the most different, and again tends to increase as either variable's MC decreases. The change in the MC_{RR} appears to depend primarily on the MC of the dependent variable. When the variables have different MCs and the initial correlation is non-zero, prediction of the statistics, and especially MC_{RR} and correlation, becomes difficult due to differences that are caused by the differences in spatial patterns of variables that have the same MC. Having a smaller number of initial zones in the aggregation increases the ranges of the aggregated statistics for variables with the same MC because dissimilar values are closer together, increasing the chances of having aggregate cells with larger internal variations.

When the dependent and three independent variables in the multiple regression experiments have the same MCs, the variation of the statistics tends to decrease as MC increases. The mean of the distributions of the regression parameters remains essentially constant as the number of aggregate cells decreases. As with the bivariate case, the change of the MC_{RR} seems to be independent of the MC of the independent variables, but again this is caused by the initial correlations between variables being zero and so the initial MC_{RR} is the MC of the dependent variable. When the dependent variables have one MC and the independent variable has another, the MC of the independent variables tends to have more of an effect on the regression statistics than does that of the dependent variable. For a given MC of the dependent variable, the variability in the coefficients and their standard errors tends to decrease with increasing MC of the independent variables. However, for a given MC of the independent variables, the ranges of the statistics increase with an increase in the MC of the dependent variable. As the results from the bivariate analysis indicate, collinearities between variables are introduced when the initial correlations are zero. However, only 2 to 8 percent of the aggregations produce correlations that are statistically significantly different from zero.

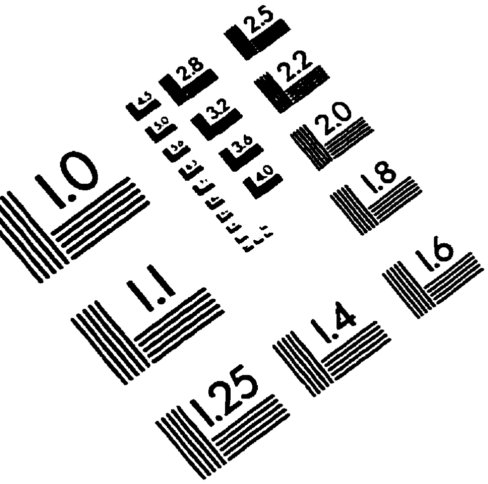
The results of this research make it abundantly clear that those who use spatially referenced data should not try to extend any conclusions they draw to levels of spatial resolution that are different from the resolution of the data. As yet there is no way to estimate the value of a statistic computed at a finer scale of resolution (larger number of smaller regions) from aggregated data, applying results derived from a coarser spatial resolution will most likely lead to the drawing of erroneous conclusions.

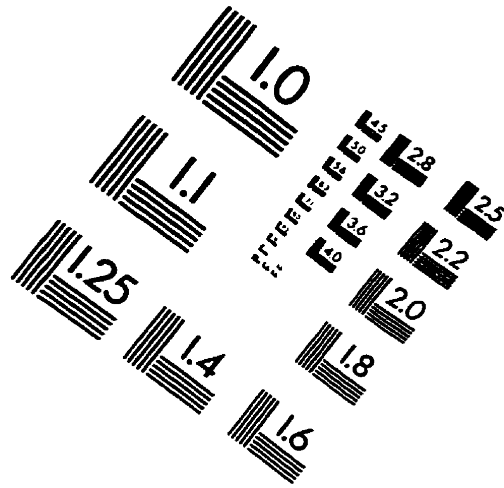
8. Topics for Future Research

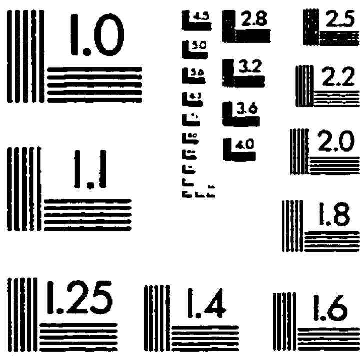
This research represents the first step in the systematic empirical exploration of the Modifiable Area Unit Problem, and much remains to be explored. All of the research work in this thesis is for variables that are averaged during aggregation, and it is suspected that variables that are summed or that are proportions (i.e. numerator and denominator aggregated separately) will not behave in the same way. Only a few of the possibilites have been explored for the multivariate regression statistics, and more complex multivariate procedures such as factor analysis have not been tested at all. Before such analysis can properly proceed, however, a better way is required to numerically quantify spatial arrangements than the Moran Coefficient. A variogram certainly contains a complete description of the spatial structure, but then a way to describe the variogram would have to be concocted and we are no better off. The MC itself is not sufficient to describe the spatial arrangement, but perhaps using it in conjunction with other spatial autocorrelation statistics that describe the pattern differently will work.

It is hoped that my research will lead to further advances in the theoretical as well as empirical exploration of the MAUP, and that the knowledge that it is not totally intractable and chaotic might be enough to renew interest and research in this challenging statistical phenomenon.

IMAGE EVALUATION TEST TARGET (QA-3)

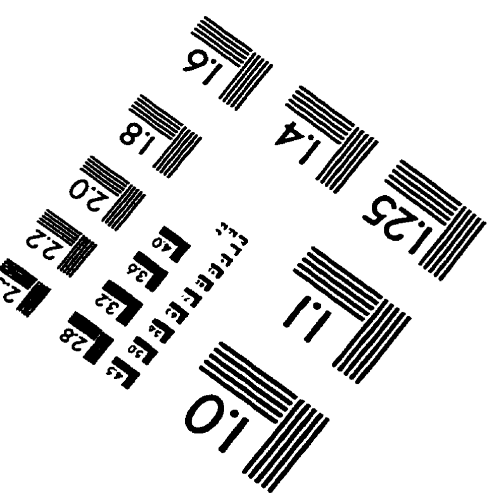


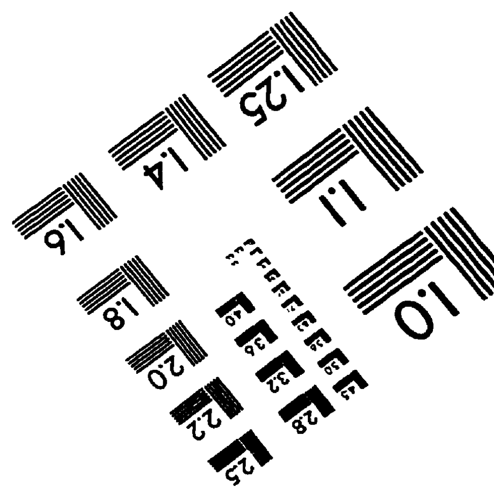




150mm

6″





APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989