Toward the Development of an Exploratory Spatial Data Analysis Application for Reporting Sensitivity to the Modifiable Areal Unit Problem: A Conceptualization

by

Eleanor May Setton
B.A., University of British Columbia, 1994

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Geography

We accept this thesis as conforming
to the required standard

[redacted]

Dr. C.P. Keller, Supervisor (Department of Geography)

[redacted]

Dr. O. Niemann, Departmental Member (Department of Geography)

[redacted]

Dr. H. Müller, Outside Member (Department of Computer Science)

[redacted]

Dr. K. Stewart, External Examiner (Department of Economics)

University of Victoria

Supervisor: Dr. C.P.Keller

ABSTRACT

The modifiable areal unit problem (MAUP) has been the concern of geographers for many decades. As of this date, there is no generally applicable solution to MAUP. It has been suggested that in the absence of such a solution, the effect of MAUP on any particular analysis be reported explicitly. This research conceptualizes an exploratory spatial data analysis (ESDA) application for reporting MAUP sensitivity. The characteristics of MAUP and ESDA are identified through literature review, and these characteristics are used to formulate a conceptual framework for developing the ESDA application. The conceptual framework is extended by proposing measures for sensitivity and through the demonstration of the concepts using two existing software applications, Arc/Info and S+.

This research finds that there are three distinct levels for reporting MAUP sensitivity, each with distinct concepts of sensitivity and distinct measures. Reports can be generated for private or public use. It appears that the public report could prove to be lengthy and place an undue burden on analysts. It is suggested that interested parties should have access to the same ESDA application used to generate the report. Some limitations were imposed on this research by the software applications used for demonstration purposes. The applications do not share a common data model, data base, or user interface; therefore, dynamic linking between views in either application is not possible. Given that this configuration of software appears to be the only commercially available means of using both Arc/Info and a statistical program in tandem, it is concluded that the widespread use of ESDA and hence any application for reporting MAUP sensitivity based on the use of such software is not expected in the immediate future.

Examiners:

[redacted]

Dr. C.P. Keller, Supervisor (Department of Geography)

[redacted]

Dr. O. Niemann, Departmental Member (Department of Geography)

[redacted]

Dr. H. Müller, Outside Member (Department of Computer Science)

[redacted]

Dr. K. Stewart, External Examiner (Department of Economics)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

The author wishes to express gratitude to the many people who contributed to the ultimate success of this thesis.

I'd like to thank the members of my committee for their support and encouragement. Special thanks are due to my supervisor, Dr. Peter Keller, for his timely and insightful comments and questions which improved this thesis markedly. To Graham Garlick, I owe a true debt of gratitude for his patience, challenging questions and dedication while creating the Arc Macro Language programs which made much of this research possible. Thanks are also due to Rick Sykes and Rosaline Canessa for what seemed to be unending technical support and software advice.

Without the generous support from those at StatSci and ESRI, who generously provided S+ and Arc/Info respectively, this research would not have been possible. I especially owe thanks to Ms. Jacqueline LeFebvre at StatSci, and Mr. Doug Herman and Ms. Laura Martin at ESRI.

Without the friendship and support of my fellow graduate students and geography department faculty, this experience would have much less rich and rewarding. Finally, to Abraham, and to my family, for all their support in so many ways - Thank you.

CHAPTER 1

INTRODUCTION AND THESIS ORGANIZATION

1.1. INTRODUCTION

The use of statistical analysis techniques within the field of Geography has been criticized for many decades. The nature of geographic data creates real conceptual problems for the application of many existing statistical methods. Inferential statistics require that the observations analyzed are independent of each other, when in fact this is rarely the case with spatial data. A second fundamental problem exists when spatial data are collected for areas or regions, as opposed to unique points over space. Often, the areas or regions are arbitrarily defined and have no relationship to the spatial distribution of the phenomenon represented by the data. Such is the case when data are collected for areas or regions which reflect political or administrative boundaries, or for a regular grid of cells or pixels through remote sensing technology. The result of any analysis performed on data associated with a particular set of areas is unique to that set of areas. Changing the configuration of the boundaries creates a 'new' dataset, and if the same analysis is performed, a new and unique solution is often the result. This characteristic is known as the **m**odifiable **a**real **u**nit **p**roblem (MAUP). The existence of MAUP calls into question the results of any research based on areal data. If it is possible to produce a different result simply by changing an already arbitrary boundary system, uncertainty exists about the reliability and validity of the result.

MAUP has been recognized for some time. A number of studies exist which empirically demonstrate the effects of changing boundary systems. As well, a number of efforts have been aimed at explaining, predicting, correcting, or minimizing the effects of MAUP. At this time, no general solution to MAUP has been found. MAUP can be successfully avoided by restricting analysis to unaggregated data using individuals or points, or, as suggested by Tobler (1989), by using only those techniques which produce

stable results for any boundary system. Unfortunately, these strategies are restrictive since many data are defined by areas (such as densities) or are available only for areas (census data), and there appear to be no commonly used techniques which are not susceptible to MAUP. One option which has been suggested is that the sensitivity of results to boundary changes be explicitly reported for research conducted using areal data (Fotheringham 1989). Rather than ignoring MAUP, trying to avoid areal data, or waiting until robust techniques are developed, this approach allows for 'business as usual' while acknowledging the existence of MAUP.

Interest in MAUP research has increased over the last decade, due to the advent and popularization of GIS use and functionality. GIS allow users to perform a range of manipulations and analyses on areal data with relative ease, and so areal analyses have experienced a surge in popularity. Investigating MAUP has been identified as part of the current research agenda in GIS by a number of authors: Anselin and Getis (1992), Goodchild et al. (1992), Fotheringham (1993), Getis (1993), Unwin (1993), Bailey (1994), Batty and Xie (1994), Keller (1994) and Rogerson and Fotheringham (1994). These authors also identify the use of exploratory spatial data analysis (ESDA) in GIS as a current research topic. In particular, Keller (1994), Fotheringham (1993), Getis (1993), and Anselin and Getis (1992) suggest that exploratory spatial data analysis (ESDA) may provide a way for performing MAUP sensitivity analyses.

The primary objective of this research is to apply the ESDA methodology to MAUP sensitivity reporting in a conceptual manner, in order to identify characteristics of both MAUP and of ESDA which would define and guide the development of methods and processes for reporting MAUP sensitivity. An attempt is made to demonstrate the concepts developed in this research using existing software applications.

1.2 THESIS ORGANIZATION

Although presented as a series of chapters, the thesis can be organized into three logical sections, each with specific secondary objectives. The objective of the first section, which includes Chapters 2, 3 ,4 and 5, is to develop a conceptual framework that integrates the characteristics of MAUP and the characteristics of ESDA as a methodology. The conceptual framework is used to develop normative goals for a method or process of reporting MAUP sensitivity using ESDA techniques. Chapter 2 provides a definition of MAUP, reviews empirical research endeavors which document how MAUP effects are manifested, and summarizes the ways in which MAUP effects can be created. Chapter 3 identifies the literature which suggests that reporting the sensitivity of results to MAUP using ESDA as a methodology forms part of the current research agenda in the discipline of Geography, and reviews efforts which have been made to address MAUP using different strategies. Chapter 4 describes ESDA in terms of its characteristics in comparison to alternative solution approaches, and identifies a set of tools defined by ESDA as a methodology. Chapter 5 draws on the previous chapters in order to develop a conceptual framework which clarifies the concept of sensitivity reporting by identifying a set of characteristics which include three report levels and two report uses. These characteristics, in conjunction with the requirements for ESDA, are used to develop normative goals for a sensitivity report method or process. Finally, a number of existing ESDA applications are reviewed, and the technical environment which will be used to demonstrate the concepts developed in this research is identified.

The second section of this thesis is structured according to the conceptual framework identified in Chapter 5. The objective of this section, which includes Chapters 6, 7, 8, and 9, is to extend the conceptual framework by developing the concepts associated with different report levels and uses. Chapters 6, 7, and 8 each focus on a specific level of sensitivity reporting as identified in the conceptual framework, using a similar format. In each chapter, a concept of sensitivity appropriate for the level of

reporting is developed and existing measures which could be used for identifying sensitivity are reviewed. When existing measures are not adequate, new measures are proposed. The measures of sensitivity are demonstrated using maps, histograms, and tables. In all cases, the proposed measures are specific to cardinal level areal data in vector format; however, when appropriate, issues concerning the application of the concepts to raster format data and categorical data are discussed. The interpretation of the measures is not explicitly undertaken in these chapters, but is undertaken instead in the next chapter. The objective of Chapter 9 is to demonstrate the generation of two kinds of sensitivity reports by using the software chosen for this research. The proposed measures and a selection of ESDA techniques are used to demonstrate the exploration process and to identify the characteristics of each report type.

The third and final section consists of Chapter 10, which presents a summary of the research and discusses a number of areas for future research and/or further development. Some conclusions regarding the conceptual framework developed and the use of ESDA as a methodology are presented. As well, limitations of the chosen software are discussed and recommendations for future development are made.

CHAPTER 2

THE MODIFIABLE AREAL UNIT PROBLEM

2.1. INTRODUCTION

The objective of this chapter is to introduce the modifiable areal unit problem (MAUP). The following sections provide a definition of MAUP in terms of the current literature, review a series of empirical studies concerning the effects of MAUP on a range of common geographical analysis techniques, and describe some of the ways in which MAUP effects are created.

2.2. THE MODIFIABLE AREAL UNIT PROBLEM

Many geographical analyses are based on areal data. Researchers may study the spatial distribution of phenomena within a region, between one region and another, or perhaps the variation in the density of a particular phenomena over space. The data used for these kinds of analyses are linked to the geographic areas from which they are derived; for example, frequency counts such as population or derived values such as average income, are reported for defined geographic areas. Any changes in the definition of the areas (also called areal units) may change the calculated values. Areal data are in fact created by defining the areal unit, which may be a subjective process guided by a researcher or by political or administrative requirements. Areal units are modifiable, in that there are many, indeed infinite, ways of dividing space into exclusive areas. These characteristics give rise to the modifiable areal unit problem.

MAUP consists of two related problems which affect mathematical measures and analyses of areal data, generally called the "*scale problem*" and the "*aggregation problem*". Throughout the literature however, a number of authors have used a variety of terms to describe these two problems. To avoid any confusion, this thesis will use the following terms: the "*aggregation problem*" and the "*configuration problem*".

2.2.1. The Aggregation Problem

Spatial aggregation occurs when contiguous areal units within a study area are combined into fewer and usually larger areal units which cover the same study area. For example, 100 base areal units may be combined to create 10 new areal units within a particular region. Aggregation of this sort is most obvious in socio-economic data; for example, census data are available for enumeration areas, census tracts, census sub-divisions, federal electoral districts and so on, each of which represent a different aggregation level. The data for each boundary set are derived from the original census sample and reflect implicitly the boundary system used to calculate the derived data.

Figure 2.1 shows the process of spatial aggregation. An important feature to note is that base areal units are not always maintained as whole entities during the aggregation procedure. New aggregation boundaries may not coincide exactly with the base areal unit boundaries, and so a base areal unit may be split into two or more sections which are then aggregated to different groups (cases c and d).

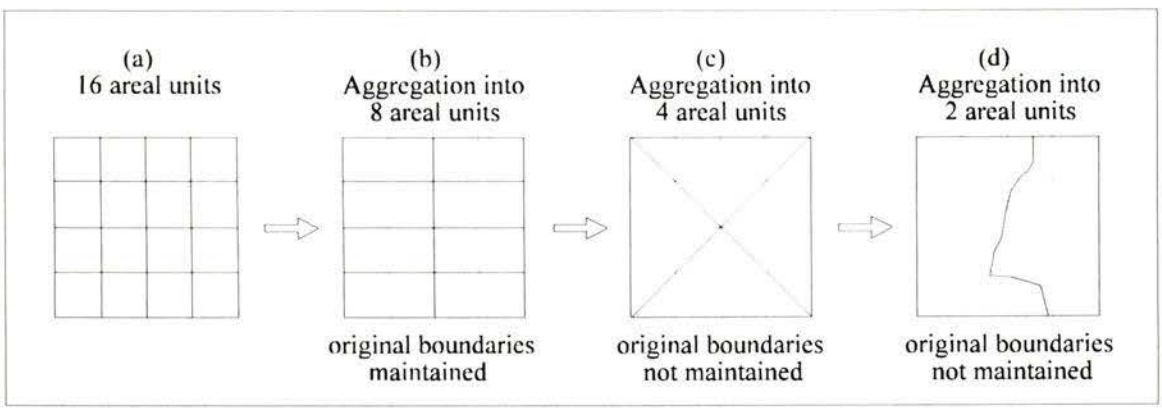


Figure 2.1. The Aggregation Process

It is important to note that MAUP is specific to spatial aggregation. The term 'aggregation' has been used in the past to describe several different processes. Generally, there are three types of aggregation: attribute, temporal, and spatial. Attribute aggregation

occurs when a single summary measure is used to represent a sub-group of individuals or points from a specified population, for example: average income of females in a sample population. In this case, some characteristic of the data defines the sub-group. Attribute aggregation may also occur when a single summary measure is used to represent individuals or points within a specified area, for example: average age of the population within a census tract. In this case, the area defines the population sub-group, and it is in this way that areal data are created: given a particular area, a single summary measure is used to describe all points or individual observations within that area.

Temporal aggregation occurs when individual observations are summarized for defined time periods, for example: average monthly rainfall, total hours worked per week. The time period of choice defines the population sub-group. These kinds of aggregation may certainly have effects on the data; however, MAUP is specific to spatial aggregation, and so this research focuses specifically on spatial aggregation. For this research, the term 'aggregation' will refer to spatial aggregation.

2.2.2. The Configuration Problem

Configuration refers to the way in which a particular study region is divided into a specific number of areal units; for example, it is possible to configure four areal units within a study area in a number of ways. The number of areal units remains the same; however, the study area is divided differently. Simply changing the boundaries of the existing areal units while maintaining the same number of areal units, creates a new configuration of the study area. Figure 2.2 shows four ways to configure four areal units. As noted for the aggregation process, there are two types of configuration procedures. The original areal units may be regrouped as whole entities (cases a, b, and c), or entirely new boundaries may be drawn which dissect the original areal units (case d).

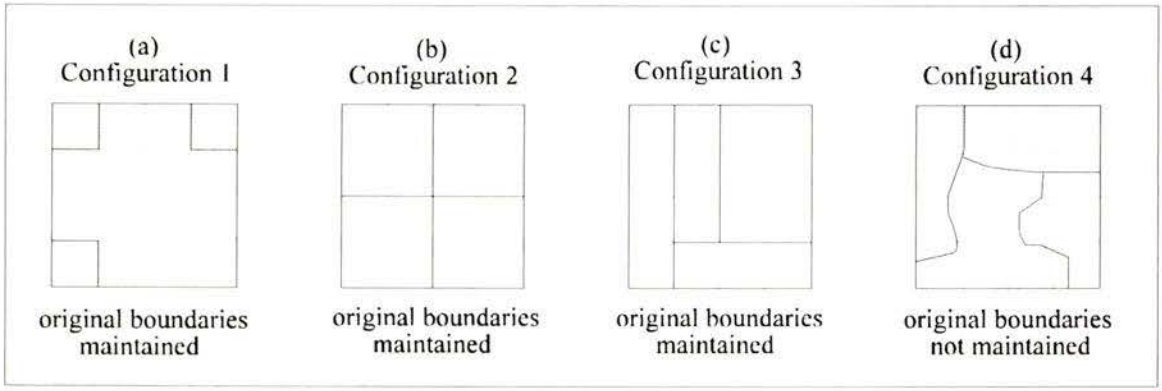


Figure 2.2 : Four Configurations of Four Areal Units

Although it is theoretically possible to create an infinite number of configurations for a given set of areal units, configuration effects may be more pronounced for socio-economic data, represented by irregularly shaped polygons such as enumeration areas or census tracts, than for biophysical data. Political or administrative boundaries are essentially 'meaningless' in that they are not derived from the spatial properties of the phenomenon being measured. They are, in fact, arbitrary. In contrast, biophysical data are often described with 'meaningful' boundaries, since boundaries try to follow the actual spatial distribution of the phenomenon as closely as possible. Although the exact placement of the boundary may be in question, there will generally be a limited number of logically possible configurations for the study area. It is suggested here that configuration effects play a lesser role in biophysical data in comparison to the effects of aggregation changes, while both configuration and aggregation have equal impact on socio-economic data.

2.3. EMPIRICAL INVESTIGATIONS OF THE EFFECTS OF MAUP

Although MAUP effects may not have been systematically studied prior to the 1930's, political redistricting for the benefit of partisan interests as early as 1812 shows that there has been an awareness of MAUP effects for almost 200 years (Morrill 1981). Generally called 'gerrymandering', the practice of selective political redistricting involves manipulating electoral boundaries in order to give one party an advantage over the others by ensuring a majority of supporting voters in as many electoral districts as possible. Since the 1930's, the effects of MAUP on statistical analyses have been studied for a variety of analytical techniques. The following survey focuses particularly on literature published within the last twenty years.

2.3.1. Correlation and Regression Analyses

Clark and Avery (1976) look at the effects of aggregation on a simple linear bivariate correlation model. They find that both the correlation coefficient and the slope estimate tend to increase as the data are aggregated into successively fewer units, but that these trends are not constant for the different levels of aggregation. They suggest that the observed deviations are a function of the changing relationship between the variables as aggregation increases, as well as changes in the expression of spatial autocorrelation. In conclusion, they warn that potential bias of correlation analyses must be recognized explicitly.

Openshaw (1984a) compares the results of correlation and regression analyses for individual data and aggregated data. He also finds that correlations become stronger with aggregation, and that this effect is more pronounced for those correlations that were near zero for the individual data. In regression analysis, he finds that r^2 increases with successive aggregation. He attributes this to the fact that more random variation is present at lower levels of aggregation. Openshaw also performs a regression on the correlation coefficients in an attempt to model the change due to aggregation, but the performance of

this model is not good, indicating there are a number of factors contributing to the effect of aggregation which do not necessarily behave in a linear fashion.

Fotheringham and Wong (1991) investigate the effects of both aggregation and configuration on a linear regression model and on a logit regression model. They report that these models give varying results at various aggregations and configurations, and that the effects are essentially unpredictable rather than systematic. They do not find any particular link between the level of spatial autocorrelation of a variable and its sensitivity to changes in aggregation or configuration, although this conclusion is later criticized by Arbia (1989). Arbia (1989) suggests their results show that "...wider ranges [of the correlation coefficient] are associated with negative spatial autocorrelation and the smaller ranges to positive spatial autocorrelation" (Arbia 1989, 19).

2.3.2. Factorial Analysis

Perle (1977) looks at the results of a factorial analysis based on two levels of aggregation of the same census data for two time periods, 1960 and 1970. The results show there are significant differences between the two sets of results, for both factor interpretations and inter-factor relationships. Significant inter-factor relationships at the higher level aggregation tend to disappear at the lower level. Similarly, distinct dimensions at the lower level tend to merge together at the higher level. At the higher level of aggregation several status factors are present in the 1960 data but these merge into one factor in the 1970 data, which indicates some process of social change occurring. At the lower level aggregation, these factors remain separate in both the 1960 and 1970 analyses, contradicting the conclusion supported by the higher level study. Perle concludes that the choice of aggregation level can lead to very different views on the structural configuration of the study area.

Openshaw (1984a) performs a similar experiment with census data, comparing results from individual data and aggregated data. He reports that the number of factors

identified decreases as the level of aggregation increase, and attributes this to the increasing association between factors as the zone size increases. Also noted is a change in the strength of the relationships between variables. Openshaw identifies the principle effect of aggregation as the creation of new factors from variables that are not strongly associated at an individual level.

2.3.3. Cluster Analysis

Openshaw (1984a) applies the approach of comparing the results of analyses on individual data with those of aggregated data to look at the impact of aggregation on cluster analysis. His results show that there is very little similarity between the two sets of results. He states that the classification changes of cluster analysis may, in fact, be the easiest way to invoke the potential ecological fallacies created by the effects of MAUP.

2.3.4. Spatial Interaction Models

Openshaw (1977) investigates the effects of both aggregation and configuration on four spatial interaction models, two of which are singly constrained and two of which are doubly constrained. The results show a large range of parameter estimates, and large standard deviations of the goodness-of-fit measure. He concludes that both aggregation and configuration have significant impacts on spatial interaction models.

Putnam and Chung (1989) look at aggregation and configuration impacts on a multivariate multiparametric spatial interaction model, focusing on the sensitivity of model parameters and the goodness-of-fit measure. They find that different aggregations result in different distributions of these statistics, and that some parameters are dependent on the configuration, while others are not. The results did not support their hypothesis that more reliable parameter estimates would produce better goodness-of-fit measures, as no trend or pattern was found.

2.3.5. Input-Output Models

Blair and Miller (1983) investigate the effects of aggregation on a multi-regional input-output model. They use the percentage of error present at various aggregation levels to make comparisons between levels. They conclude that the percent error present at each aggregation level is relatively small, and so spatially aggregated input-output models may provide satisfactory answers to research questions. However, they also report that error increases with aggregation, and that at the lowest aggregation level the magnitude of errors for within region '1' and outside region '1' are similar, but become less similar as aggregation increases. They cannot explain this trend, and state that the trend contradicts the results of a previous study of an inter-regional model. Fotheringham and Wong (1991) suggest that as this study is limited to a maximum of four regions, there are likely too few regions to assess the impacts of aggregation fully.

Miller and Shao (1990) perform a similar analysis, using a much larger dataset than in Blair and Miller (1983). They conclude that the percent error generated at different levels of spatial aggregation is relatively small, and that while there is sensitivity to the effects of aggregation, these models may be adequate.

2.3.6. Location-Allocation Models

Bach (1981) compares the location criteria produced given increasing aggregation of a dataset. The results show that when the level of aggregation changes, the criteria values are distorted and the optimal location patterns change, sometimes quite radically. Bach concludes that aggregation strongly influences both characteristics.

Goodchild (1979) performs a similar study and finds similar results. He concludes that since the results of location-allocation analyses are specific to a particular pattern of aggregation, the benefit of using this type of analysis as a way of making objective recommendations on solutions to planning problems is lost. He further states that the

solutions may be manipulated by using different aggregations, and so the usefulness of location-allocation models is doubtful.

2.3.7. Remotely Sensed Data

Woodcock and Strahler (1987) investigate the effects of changing the resolution of a remotely sensed image. Changing the resolution is analogous to changing aggregation levels in this case. Their goal is to identify a particular resolution level which optimizes the identification of specific objects, through the use of a local variance measure. Their findings indicate that the resolution level of the image affects the process of classification.

Marceau, Howarth, and Gratton (1994) investigate the impact of aggregation on the information content and classification accuracy of remotely sensed data, specifically airborne MEIS II data of a forested environment. Through four aggregations, they find that the peak in variance statistics differs between classes, indicating that some classes are better identified at low resolution (i.e. aggregations) while others are better identified at higher resolutions. They conclude that remotely sensed data are dependent on the sampling grid (i.e. areal units) used for collection, and should not be dissociated from the collection scale.

2.4. CREATING MAUP EFFECTS

The reviewed investigations of MAUP focus primarily on the results of statistical analyses which are applied to a number of potential aggregations or configurations. As such, these studies highlight how MAUP effects are *manifested*, rather than how MAUP effects are *created*. Sections 2.2.1 and 2.2.2 served to define aggregation and configuration, but a more in depth discussion of the nature of areal data and the creation of MAUP effects is warranted.

2.4.1 The Nature of Areal Data

Areal data are those which describe phenomena within specified areas on the earth's surface, rather than individual objects, points, or lines. An important exception to note here is point data which are derived from areal units. For example, the centroid of an area may be calculated and the area value attached to the centroid for analyses which require point data. Areal data are summary measures created through attribute aggregation. Briefly, attribute aggregation consists of assigning a single summary measure to a set of points or observations. When areal units are used, homogeneity of the area described is assumed. Areal data may be measured at either categorical or cardinal levels, depending on the phenomenon and the purpose for which the data are collected. At each measurement level, measures can be expressed in different ways. Table 2.1 summarizes measurement levels, measurement types (expressions), and gives associated examples of common summary measures for areal data.

Table 2.1. Common Areal Data Summary Measures

Measurement Level	Measurement Type	Examples
Categorical	Class or Category	vegetation types, soil types, climate regimes, income (high, medium, low)
Cardinal	Frequency	total population i.e. bears, people, trees, cars
	Percent	percentage of diseased trees, percentage of bears over 5 years old, percentage of car owners
	Average	average height of trees, average age of bears, average income of people
	Density	number of trees per hectare, number of bears per square kilometre, number of people per square metre

The examples given in Table 2.1 are drawn from both biophysical and socio-economic phenomenon. Data for biophysical phenomenon are often collected through field surveys and observations. Although the data can be presented for the collection points, they are often presented as categorical thematic maps using areas which are defined according to the spatial distribution of the phenomenon. Remote sensing technology also provides for the collection of biophysical data, in the form of reflectance values for a grid of pixels or cells. The pixels themselves represent areas, and so the raw reflectance values are areal data; however, it is common to create categorical thematic maps using the data collected through remote sensing. The practice of presenting biophysical areal data as categorical thematic maps limits the effects of the configuration problem. As discussed previously (see Section 2.2.2), the configuration of areas usually is defined by the phenomenon itself, and so there are relatively few logical configurations possible to represent the phenomenon. Aggregation, however, can have a substantial impact on biophysical data which are represented by areas.

Large datasets for socio-economic phenomenon are most often collected through regular census efforts. The data are collected for individuals; however, for reasons of privacy, the data are usually available only as summary measures for areas defined by political or administrative boundaries. These boundaries are often irregular in shape and size, as well as arbitrary. Most commonly, census data are expressed using cardinal measures such as frequencies, percentages, and averages to summarize the individual data. Both aggregation and configuration have a substantial impact on socio-economic data. It is for this reason that this research will focus specifically on socio-economic cardinal level data, collected for irregularly shaped, arbitrary areas.

2.4.2. The Creation of MAUP Effects

When areal data are aggregated or re-configured, new summary values are calculated for the resultant areal units. It is this re-expression of the data which causes MAUP. When aggregation or configurations are changed, the information provided by the dataset is changed by the summary process . Depending on the method used, aggregation usually creates a new dataset with less information about the inherent variability of the original areal units and about the distribution of the phenomenon in question, while configuration changes lead to the creation of a different dataset which may or may not contain as much information about variability and distribution. The results of any analysis which depends upon these data values will necessarily reflect the changes made to the data by aggregation or reconfiguration.

There are a variety of methods for calculating or assigning new cardinal summary values to the new areal units. The same methods are used for both the aggregation and the reconfiguration of areal units; however, the method used depends upon the type of measurement (i.e. frequency, average, etc.), and on the amount of information contained within the dataset. The method used may be adjusted when areal unit boundaries are split rather than maintained. The following section identifies common methods for calculating new summary measures for cardinal level data.

2.4.2.1. Methods for Calculating New Summary Values for Areal Units

When areal units are combined and the areal data are frequencies, the new summary values are simply the sums of the frequencies for each constituent areal unit. When the areal data are percentages, averages, or densities, different methods exist for calculating new summary measures, each of which make different implicit assumptions about the data and may produce different results even when the same base data are used. For example, one method for calculating new summary measures such as percentages or averages, is to take the average value of the constituent areal units. A second method

might be to areally weight the average in order to incorporate some of the spatial characteristics of the data. A third method, recalculation, may be used; however, the components of the measures must be known. For example, given a percentage of car owners, when the number of cars and the number of people are known for each area, these component frequencies can be first summed according to the new grouping, and then the percentage can be recalculated. Each of the above methods may produce different summary values, even though the same original dataset is used.

Table 2.2 summarizes these common methods for calculating new summary values, taking into account the type of measurement, the information needed, and the associated implicit assumptions. The assumptions noted and the adjustments needed when original area boundaries are split are addressed in the following discussion

Table 2.2. Methods for Calculating New Summary Values

Measurement Type	Method	Info. Needed	Assumptions
Frequency	sum	n/a	n/a
Percentage	simple average	n/a	equal populations* in all areas
	areally weighted average	areas	equal population density in all areas
	recalculation	components of measure	n/a
Average	simple average	n/a	equal populations in all areas
	areally weighted average	areas	equal population density in all areas
	recalculation	components of measure	n/a
Density	simple average	n/a	equal area for all areas
	recalculation	components of measure	n/a

* population refers to the observation population

Using a simple average assumes that the populations on which the averages are based are equivalent for each of the grouped areas; for example, an extremely low average will have equal importance in the calculation, even though it may be associated with an extremely low population. In this way, the new summary measure may be biased by any extreme values. When all populations are equal, each value does in fact have equal importance and the simple average results in an unbiased measure. Using a simple average for density measures is analogous: area acts in the same way that population acts for percentages or averages. The simple average accords the value of each area equal importance, and in this way it is implicitly assumed that the areas are equivalent.

Including the sizes of the areal units being combined as weights during averaging may be used when calculating a new value. The idea behind this method is as follows: if one very large areal unit with a very low average is being combined with a very small areal unit with a very high average, a simple average gives each value equal importance and is therefore biased. In fact, the majority of the new areal unit consists of the low average value, and this should be reflected in the new value. The assumption implicit in this method is that of equal population densities among the constituent areal units, in other words, the phenomenon population rises proportionally to area (see Goodchild, Anselin, and Deichmann, 1993). Areally weighting an average value makes sense if the very large areal unit with a low average contains a very large population, and the very small areal unit with a high average contains a very small population. Unfortunately, the restricted relationship between area and population may not always hold true.

The methods listed in Table 2.2 can be used for calculating new values for partial areal units; however, when areal units are split, new values must first be assigned for the portion of each split areal unit before the combined value can be calculated. When using frequencies as summary measures, a portion of the frequency of the 'parent' areal unit must be assigned according to area. For example, if a partial areal unit comprises 10% of the area of the 'parent' areal unit, 10% of the frequency of the 'parent' areal unit is assigned to

the partial areal unit. The implicit assumption in this method is that the frequency is evenly distributed within the 'parent' areal unit. When averages, percentages, or densities are used, the split portion retains the value of it's 'parent' areal unit, and the methods listed in Table 2.2 can then be applied.

There are a number of more sophisticated methods for calculating new values when the base areal units are not grouped as whole entities. A review of a range of areal interpolation methods can be found in Goodchild, Anselin, and Deichmann (1993). They include radially symmetric kernel functions (gravity modeling), maximally smooth estimation (pycnophylactic interpolation), and piecewise approximation (EM algorithms). These methods will not be specifically addressed in this research.

2.5. CHAPTER SUMMARY

The modifiable areal unit problem consists of two related problems: the aggregation problem and the configuration problem. Aggregation consists of combining a number of areal units into larger and fewer areal units which cover the same study area. Configuration changes consist of moving the boundaries of the existing areal units so that, although the number of areal units remains the same within a given study area, the space is divided differently.

Given a set of base areal units, an aggregation may combine those areal units as whole entities, or create new boundaries that cross directly through the existing base areal unit boundaries. Creating a new configuration of a particular aggregation may group different base areal units as whole entities, or again, create new boundaries which cross through the original base areal unit boundaries. Changing the configuration of the base areal units themselves requires, by definition, that the existing boundaries be changed.

Both aggregation and configuration impact socio-economic data collected for irregular and arbitrary boundaries. Aggregation may have a more substantial impact than

configuration on biophysical data, since they are commonly described using boundaries derived from the spatial distribution of the phenomenon in question.

A number of studies have identified the presence of MAUP in a wide range of analytical methods commonly used by geographers, including correlation, regression, factorial analysis, cluster analysis, spatial interaction models, input-output models, and location-allocation models. MAUP has also been shown to apply to remotely sensed data.

MAUP effects apply only to areal data. Areal data are those which summarize a set of observations within a specified area on the earth's surface, in contrast to data which describe individual objects, points, or lines. An important exception are point data which have been generated from areas. For example, the data value of an area can be assigned to the area's centroid point, and the centroid points may then be used for analysis. Areal data may be either categorical or cardinal levels of measurement. Categorical areal data are expressed as classes or categories, while cardinal areal data are commonly expressed as frequencies, percentages, averages, or densities.

MAUP effects are created when areal units are aggregated or re-configured, and new summary values are calculated and assigned to the resultant areal units. Information inherent in the original data may be lost or may simply be changed. The result of aggregation or configuration changes is the creation of a new and unique dataset. Common methods for calculating new summary values include summing for frequencies, and simple averaging, areally weighted averaging, or recalculation for averages, percentages, and densities. Different methods make different assumptions regarding the data, and influence the resulting new summary values.

CHAPTER 3

APPROACHES TO DEALING WITH THE MODIFIABLE AREAL UNIT PROBLEM

3.1. INTRODUCTION

The importance of the modifiable areal unit problem (MAUP) has not been underestimated, and a number of strategies for dealing with the effects of MAUP have been suggested. These strategies generally fall into three classes: strategies for avoiding MAUP, statistical approaches to solving or minimizing MAUP, and reporting sensitivity to MAUP. The objective of this chapter is to review suggested approaches to dealing with MAUP and its effects on analyses using areal data. The chapter is organized into three sections. The first two are concerned with reviewing the aforementioned approaches and identify sensitivity reporting as a relatively undeveloped concept. In the fourth, exploratory spatial data analysis (ESDA) is identified in the literature as a possible methodology for studying MAUP. This chapter defines the area of research for this thesis and concludes with a statement of research objectives.

3.2. SUGGESTIONS FOR AVOIDING MAUP

The most obvious way to avoid MAUP is to avoid using areal data altogether. This amounts to using only point or individual data. While this strategy provides a way to avoid MAUP, it is not often possible to implement. In many cases, data are available only in an already aggregated form. In other cases, the object of study is areal in nature, for example, measurements of density. Also, it is possible that individual or point datasets may be so extensive that some type of aggregation is necessary to facilitate analysis and understanding.

If areal units must be used, it has been suggested that these units should be formed so that they are homogeneous representations of the phenomenon under study. Again,

when possible, this could allow for analyses unaffected by MAUP. MAUP effects are created when areal units with different values are combined, and a new representative value must be assigned to the newly created areal unit. If the areal units combined have the same values, the new value would necessarily be representative. While this approach could be used for univariate analyses, it does not appear to be useful for multivariate analyses. It may be possible to delineate areal units homogeneously for one variable; however, it will be extremely difficult to delineate areal units which are completely homogeneous for two or more variables. As well, the definition of homogeneity may differ from study to study or variable to variable. Finally, although the use of homogeneous areal units may avoid aggregation effects, it does not necessarily avoid configuration effects (Sawicki 1973). For example, maximum homogeneity will be associated with only one configuration, therefore any configuration changes may result in heterogeneous areal units.

Another suggestion for avoiding MAUP is to use only 'meaningful' boundaries. This concept was introduced in Chapter 2. It has been suggested that MAUP effects may be more severe when the boundaries delineating the areal units are arbitrary, as is the case when administrative or political boundaries are used for socio-economic data. These boundaries are essentially meaningless in terms of the phenomenon represented by the data, and it is possible that any number of aggregation level or configurations could be used to structure the study area. 'Meaningful' boundaries are those which are defined by the phenomenon itself. Boundaries which identify a particular land use, or a particular vegetation type are 'meaningful' since the boundary must be defined according to the distribution of the phenomenon. As such, these 'meaningful' boundaries are 'correct'; no other set of boundaries is valid; and MAUP effects are theoretically nonexistent. In reality, it is extremely difficult, if not impossible, to define meaningful boundaries in an exact fashion. Also, aggregation changes will, of course, cause even 'meaningful' boundaries to change.

A fourth way to avoid MAUP is to use only those analytical methods which are insensitive to MAUP effects. Tobler (1989) suggests that MAUP is a consequence of using inappropriate techniques, such as correlation or regression analysis. He suggests instead that the spatial cross-coherence function should replace correlation as a method of comparison. With respect to the studies performed using these statistical measures, he states "the fallacy in all of these studies is the assumption that the correlation coefficient is an appropriate measure of association amongst spatial units. Clearly, it is not ...but all of these authors put the blame on the spatial units" (Tobler 1989, 116). Tobler's argument is reasonable, but in practice, there are few if any known analytical techniques which are insensitive to MAUP, and those analytical techniques which are sensitive to MAUP are still widely used.

3.3. STATISTICAL APPROACHES TO SOLVING OR MINIMIZING MAUP

Statistical approaches to understanding MAUP have taken various forms. In general, there are two groups: those methods which attempt to devise a predictive or corrective method as a means of solving MAUP, and those which inspect the range of aggregation or configuration possibilities for some optimum situation as a means of minimizing MAUP by improving model performance.

3.3.1. Corrective Methods

An early attempt to correct for MAUP effects is made by Robinson (1956). He suggests that for correlation and regression analyses, the area of each areal unit in a specific aggregation can be used to modify the calculation of correlation and regression parameters. This can be described as a variable-centred approach, since some aspect of the variable is used to adjust the model. Using a simple example, Robinson shows that the results of these analyses are consistent for three aggregation levels when areal weights are used. Thomas and Anderson (1965) criticize Robinson's solution as being limited only to

very special cases where the variables X and Y have exactly the same distribution for all aggregation levels, and when the changes between aggregation levels in the numerators and denominators of the formulae used are exactly proportional. They go on to treat the different aggregation level areal units as random samples, and use this assumption to provide a rationale for using inferential statistics (tests of statistical significance) to study the regression parameters produced using three sets of data. This approach can be called result-centred, in that the focus is on the analytical results rather than on the variables used. In all cases, they report that the correlation coefficients, as well as the slope and intercept parameters of the regression line, are not significantly different over a range of aggregation levels. Based on these results, they conclude that when significance testing allows, the variation in results have no geographical significance. While not specifically corrective, their method would allow for the disregard of MAUP effects.

The assumption that areal units constitute a random sample of an underlying distribution is a contentious one. Larson (1986) summarizes the problems associated with making this assumption. Perhaps the most important problem is the need to assume statistical independence. He suggests that the definition of areal units is a form of sampling without replacement which cannot be corrected for since "in normal sampling without replacement, the selection of an individual precludes only the future selection of that same individual while in areal unit sampling, the selection of an areal unit normally precludes the future selection of all other areal units that contain any element of the original unit" (Larson 1986, 370).

Using a similar rationale to Robinson (1950), Bearwood and Kirby (1975) use a variable-centred approach to derive a 'coefficient of separation', which when used to adjust a simple gravity model, produces consistent results across a range of aggregation levels. Openshaw (1977) suggests that, although Bearwood and Kirby(1975) show that the model can be made independent of a coarse zoning system, "this is an accounting device of administrative rather than of analytical or applied value" (Openshaw 1977, 170). In a

series of papers, Batty and Sikdar (1982 a, b, c, d) make an important contribution to the corrective strategy. They link the variable-centred approach with the result-centred approach. A series of aggregation levels are created, and a summary information statistic, based on the probability density function, is calculated for each aggregation level. This information statistic is then used in combination with the model parameters which result from the corresponding aggregation level to create a predictive model using both regression and entropy techniques. Using a simple one dimensional interaction model (population density), Batty and Sikdar obtain a good approximation of the model parameters using this method; however, when a two dimensional model of trip distribution and trip location is used, the results are poor. They suggest that the poor results are due to the poor performance of the two dimensional model rather than the information statistic/parameter model.

An omission in all four of these studies is the investigation of the effects of configuration. In all cases, only the effects of aggregation are investigated. This omission is particularly important in terms of the method proposed by Bearwood and Kirby (1975). In this case, the coefficient of separation is calculated using the lowest aggregation level data and applied to higher aggregation levels. It is highly likely that, given a different configuration of the lowest aggregation level data, the coefficient would be different. In this instance then, although the results using the highest level aggregation can be made to be consistent with the results using the lower aggregation level, the coefficient of separation is not independent of the configuration at the lowest aggregation level.

Arbia (1989) explicitly includes configuration effects in an impressive study which uses the theory of stochastic process as a basis for studying the effects of a statistical manipulations on the mean, variance, skewness and kurtosis of the probability distributions of a variable. Also addressed are the variation observed in the spatial autocorrelation of a single variable, and the cross correlation of two variables. Arbia is able to provide theories which explain the empirical results of many studies, and to provide ways of reducing the

effects. Arbia's work is theoretically complex and it does not appear that it has yet been applied in general practice.

While all five studies have made important contributions, there are some limitations. As mentioned previously, the first four studies ignore configuration effects, and while Arbia includes this, the methods developed are based on using entire base areal units when creating new groupings. The idea that base areal units may be split when creating new aggregation levels and configurations was introduced in Chapter 2. The question of whether or not recombining portions of areal units may have some effect on Arbia's work is worth pursuing in the future. Secondly, the results of these studies are restricted to univariate and bivariate cases. The extension to multivariate cases has not yet been made, and it is unknown whether any of these methods or theories will be applicable.

3.3.2. Seeking Optimum Situations

If the assumption is made that MAUP effects cannot be solved in all cases, the next option is to minimize those effects. This approach has been taken by a number of researchers. Their work can be classified into two groups: identifying the level at which a process is operating, and identifying the level which optimizes the chosen model. The following descriptions clarify the differences between the two approaches.

3.3.2.1. Identifying the Process Level

The general hypothesis underlying this approach is: given a range of aggregation levels, that level which contains the highest variance or information indicates the level at which the process under study is operating, and is, therefore, the correct level to use for that process and model. In this way, the question of which aggregation level is appropriate is resolved, and the range of results in other levels can be ignored.

Moellering and Tobler (1972) use a variable-centred approach and employ analysis of variance techniques to examine scale effects on three different variables. Each

aggregation level is assigned a variance, and these variances are then compared to see "where the action is" (Moellering and Tobler 1972, 36). In this case, high variances indicate 'action'. For each of the three variables, a different aggregation level is indicated by the variance statistic. This suggests that the process governing the distribution of the variables is different in each case.

Batty (1976) also uses a variable-centred approach and introduces an information statistic as an alternative to the analysis of variance technique used by Moellering and Tobler (1972). Batty (1976) compares the performance of the information statistic with that of the Moellering and Tobler (1972) variation statistic, as well as with other existing information or variation measures. He finds the proposed information statistic performs comparably. The information statistic is then applied in Batty and Sammons (1978). Unfortunately, their results indicate that the highest level of information occurs at the lowest aggregation level, i.e. when the smallest areal units are used, and so this method may not contribute to the identification of a process level.

Moellering and Tobler (1972), Batty (1976), and Batty and Sammons (1978) do not address the potential effects of configuration. It is highly likely that a range of information or variation statistics could be produced for the same aggregation level, given a number of configurations, and the range may overlap with that of another aggregation level. Under such circumstances, it would not be possible to identify one specific level for analysis.

3.3.2.2. Seeking the Optimum Model Performance

The general hypothesis underlying this approach is: given a range of aggregation levels, that level which produces the best model performance is in fact the level at which the process under study is operating, and is, therefore, the correct level to use for that process and model. This approach then defines optimal model performance in some way, and finds that aggregation level which gives the result closest to the optimum.

Using a variable-centred approach, Masser and Brown (1975) apply an hierarchical aggregation procedure for optimizing spatial interaction models. The objective is to create the maximum interaction between base areal units which are grouped together at a different level, and to minimize the interaction between those groups of base areal units. In this way, the areal units chosen are optimal, and should produce optimal model results. This procedure is essentially a clustering technique. Masser and Brown (1975) find that the hierarchical aggregation procedure provides "...no straightforward or universal solution..."(Masser and Brown 1975, 523).

In a true departure, Openshaw (1977) suggests that the desired model parameters should be selected, and a range of aggregations and configurations should be tested against the desired performance. That aggregation and associated configuration which produces results which meet the requirements would then be the appropriate set of areal units for studying the process in question. Openshaw admits that this result-centred approach may have some problems: "some will argue that the whole concept of optimal zone design...is incompatible with...normal science...,[and] some will view it as gerrymandering" (Openshaw 1977, 182). Openshaw also describes another problem of the optimal zoning approach - the problem of optimal zone stability over time. It is possible, and quite likely, that the data within zones which create optimal conditions can change over time, and may no longer provide an optimal solution.

Of all the studies reviewed in this section, Openshaw (1977) is the only one to include the effects of configuration changes. He uses a result-centred approach, focusing on four different spatial interaction models. At each of two aggregation levels, he takes a 'random sample' of configurations, and summarizes the model results for the configurations. He concludes that, in this case, the effects of aggregation and configuration do indeed have an important effect on spatial interaction models. This observation is in contrast with Bearwood's and Kirby's (1975) statement that gravity models can be made independent of aggregation effects.

3.4. REPORTING SENSITIVITY

Fotheringham (1989) proposes that the effects of MAUP on model calibration should be documented by reporting the results for different levels of aggregation and for different configurations. This approach can be identified as result-centred. A sensitivity report could serve as a way of assessing reliability and validity. Results showing less sensitivity would be more reliable than those exhibiting high sensitivity, and results showing high sensitivity might indicate conceptual problems with the choice of variables or areal units. The sensitivity analysis could also generate further research questions regarding why some results are more or less sensitive.

Fotheringham does note an important weaknesses of this approach: while configuration effects can be investigated at any level of analysis, the data must be available at a more disaggregate level than the level desired for research in order to investigate the effects of aggregating up to the level of analysis. He suggests that data could be aggregated at higher levels, and the results assumed to hold true for lower levels, but this appears to have little validity as it has been shown that the effects are unpredictable in multivariate analysis, rather than systematic (Fotheringham and Wong 1991).

Although reporting sensitivity to MAUP may be restricted to configuration effects and higher level aggregation effects, this approach has some promise, in that the effects of MAUP are explicitly stated, and serve as a caution to avoid making inferences about other levels of aggregations or configurations. A further benefit of this approach would be the rapid increase in data about MAUP effects on a wide variety of analysis methods and datasets, as researchers using areal units would be reporting the effects for their particular problem.

At this time, no specific method has been suggested for reporting the effects of MAUP, however, a number of authors have identified the potential for using an exploratory spatial data analysis (ESDA) methodology within geographic information systems (GIS) for MAUP research. For example, see Anselin and Getis (1992), Goodchild

et al. (1992), Fotheringham (1993), Getis (1993), Unwin (1993), Bailey (1994), Batty and Xie (1994), Keller (1994) and Rogerson and Fotheringham (1994).

3.5. CHAPTER SUMMARY

MAUP can be avoided through the use of individual or point data, or by using analytical techniques which are not sensitive to MAUP; however, in many cases, it is impossible or undesirable to use individual or point data, and there are few known techniques which are not affected by MAUP. Much of the research about MAUP has therefore focused on providing a solution to MAUP or on minimizing the significance of MAUP on a range of analytical techniques.

Statistical approaches to solving MAUP have been used by Robinson (1956), Thomas and Anderson (1965), Bearwood and Kirby (1975), Batty and Sikdar (1982 a,b,c,d), and Arbia (1989). In all cases, these approaches are limited to univariate or bivariate cases, and none but Arbia (1989) include configuration effects within their respective approaches.

Approaches to minimizing the significance of MAUP have been suggested by Moellering and Tobler (1972), Batty (1976), and Batty and Sammons (1978). Each of these papers focus on the use of a statistic measure which identifies the aggregation level at which a process is operating. Again, the effects of configuration are not included in these research efforts. Masser and Brown (1975) and Openshaw (1977) suggest optimizing model performance through the selection of the 'correct' aggregation or configuration, although as Openshaw (1977) notes, this approach may be criticized as being data manipulation rather than objective analysis.

The focus of each of the reviewed articles has been placed on either the variables used for analysis or on the results of a particular analytical technique. These approaches have been identified in this research as variable-centred or result-centred. The variables used for analysis provide the data used in the variable-centred approach, while the results

produced using an analytical technique provide the data for the result-centred approach. One feature that all reviewed articles share is the use of a common data transformation process. In all cases, a 'base' dataset, i.e. one that is disaggregate, is successively aggregated and sometimes re-configured in order to perform the desired analysis of MAUP effects.

Fotheringham (1989) and (1991) suggests that reporting the sensitivity of analysis results to MAUP holds some potential. Rather than attempting to solve or minimize the problem for a specific analytical technique, MAUP effects are explicitly stated for any technique and can provide useful information about the robustness of an analysis. Currently, no standard method or process exists for reporting the effects of MAUP.

Reporting sensitivity to MAUP has been identified as an area of current research, and ESDA has been identified as having potential applicability to MAUP sensitivity reporting. The primary objective of this research then is to apply ESDA as a methodology for MAUP sensitivity reporting in a conceptual manner, in order to identify characteristics of both MAUP and of ESDA which would define and guide the development of methods and processes for reporting MAUP sensitivity.

CHAPTER 4

EXPLORATORY SPATIAL DATA ANALYSIS: A METHODOLOGY

4.1. INTRODUCTION

The objective of this chapter is to define and differentiate exploratory spatial data analysis (ESDA) from a number of alternative approaches to analysis. The term 'exploratory spatial data analysis' has become popular recently in geographical and statistical literature. In the interest of clarity, this chapter attempts to define the term and its component parts. To that end, the first section of this chapter will be devoted to describing and defining a number of approaches to data analysis: first, the overall concept of data analysis (DA) is covered, followed by confirmatory data analysis (CDA), exploratory data analysis (EDA), spatial data analysis (SDA), confirmatory spatial data analysis (CSDA), and finally exploratory spatial data analysis (ESDA). Although this may seem somewhat basic, there appear to be a variety of usages and definitions for the above terms throughout the literature. Clarification of these terms is thus useful to avoid confusion.

4.2. DATA ANALYSIS

The term 'data analysis' covers the full range of methods and techniques used to manipulate 'raw' data. These methods may be either descriptive or explanatory, qualitative or quantitative, and be applied to nominal, ordinal, interval or ratio levels of measurement. A simple count or classification of objects or observations is a form of data analysis, as is the use of a statistical technique for the purpose of finding or testing the relationships between a number of variables. In this sense, then, data analysis has been practiced for many centuries, and over time, has expanded to include ever more sophisticated techniques.

In general, it is possible to classify data analysis techniques into two categories - those that are based in mathematics, and those that are based on observation or visualization. All statistical techniques, whether they are descriptive or explanatory, are quantitative, and therefore spring from a mathematical base. These techniques use the numerical expressions of observations to produce new numerical expressions which may simplify (descriptive statistics) or explain (via the use of models) the original data. In contrast, visual techniques are based on graphic representations of data and are qualitative in nature, although they may use either qualitative or quantitative data. These techniques include graphs (i.e. histograms or scatterplots), diagrams, and maps. For example, a graph of temperature and energy input will show qualitatively that temperature increases as energy input increases. Mathematical expressions or operations will be required to quantitatively assess the relationship. Similarly, a map showing annual rainfall may lead to the qualitative observation that annual rainfall is higher in some regions than in others, but again, mathematical expressions or operations will be required to give a quantitative analysis.

DiBiase et al. (1992) suggest that visualization is used in two different ways: for visual thinking, and for visual communication. Visual thinking occurs when graphic methods of representing the data are used to find patterns or trends which may then be quantitatively analyzed using mathematical or statistical techniques. In this sense, visual thinking is exploratory, descriptive, and inductive. Visual communication occurs when the results of a mathematical or statistical analysis are presented graphically for ease of understanding. In practice, visualization techniques may be used in either fashion, or may be developed specifically for one particular use. Interactive graphics provide an example of the latter, in that through interaction with visual representations and statistical representations, a researcher may uncover patterns or trends of interest. Interactive graphics may have less application as communicative devices, simply due to constraints on accessibility to the interactive environment by a large audience.

The emphasis on the use of statistical techniques, visual communication, and visual thinking, depends on the approach adopted by a researcher. The differences between approaches will be drawn out in the following sections.

4.2.1. CDA - Confirmatory Data Analysis

Confirmatory data analysis (CDA) is not defined by any particular set of techniques, instead it is performed at a specific point during research. CDA takes place after an hypothesis has been formulated, and uses data collected or selected according to a specific research design. CDA is performed in order to either support or refute the stated hypothesis. In this sense then, CDA is more closely associated with mathematical forms of analysis that are explanatory (statistics and models), with visual communication, and with deductive reasoning.

4.2.2. EDA - Exploratory Data Analysis

Exploratory data analysis (EDA) is similar to confirmatory data analysis in that it is not defined by a specific set of techniques. EDA encompasses any analysis performed without a particular hypothesis in mind. The purpose of performing exploratory data analysis is to look for interesting characteristics or patterns within a set of data that may lead to the formation of an hypothesis which may then be tested using confirmatory data analysis. "EDA is concerned with observational data more than with data obtained by means of formal design of experiments" (Good 1983, 283), and in this way it is "data-driven" analysis (Anselin and Getis 1992). EDA is linked with the mathematical forms of analysis which are descriptive, with visual thinking, and with inductive reasoning. Tukey (1977) draws a simple analogy: EDA is like detective work, aimed toward finding and revealing clues, while CDA becomes judicial or quasi-judicial by assessing the strengths of the evidence. Unless the detective finds the clues, judge or jury has nothing to consider.

4.2.3 SDA - Spatial Data Analysis

Spatial data analysis (SDA) can be thought of as a subset of DA in that it, too, encompasses a variety of methods; however, SDA applies in particular to data which have a geographical distribution. Like DA, its techniques range from simple descriptive measures of patterns of events to complex statistical tests of relationships between variables. The development of SDA began with the recognition that traditional CDA and EDA techniques were not valid when the data had spatial characteristics. Openshaw states that "conventional statistical theory and normal science paradigm led to acceptance of the importance of introducing scientific methods and quantitative techniques in geography under the hopes that various unrealistic statistical and geographical assumptions could be relaxed later"(Openshaw 1984b, 6). This has not been the case, and likely never will be, since geographic data possess characteristics which violate the assumptions of most statistical methods. One of the most basic assumptions of many statistical methods is that the data are independent of each other. Geographic data are generally autocorrelated, and therefore dependent, since observations made in close proximity to each other generally will be similar. The presence of autocorrelation has led to the development of a variety of spatial data analysis techniques which explicitly include the spatial relationships between observations, usually in the form of proximity measures and distance measures. Like DA, SDA can be approached from either the confirmatory or exploratory viewpoints.

4.2.4. CSDA - Confirmatory Spatial Data Analysis

CSDA techniques are similar to CDA techniques in that they are linked with the mathematical forms of analysis that are explanatory, to visual communication and to deductive reasoning. The essential difference is that CSDA uses statistical techniques that explicitly include the spatial characteristics of the data under study. For example, CSDA might consist of testing a model which includes a distance decay function on a set of

geographical data. The results may be presented in graphic forms, such as scatterplots or maps, for the purpose of communication.

4.2.5. ESDA - Exploratory Spatial Data Analysis

ESDA is analogous to EDA in that it is linked with the mathematical forms of analysis which are descriptive, to visual thinking, and to inductive reasoning. Developed in answer to the criticism that EDA techniques are aspatial, ESDA incorporates the spatial characteristics of geographical data specifically by including the use of maps with other visual thinking devices. Through the use of a map, simple descriptive but aspatial statistics and graphic techniques can be evaluated within their geographical context, thus incorporating the spatial characteristics of the data.

4.2.6. Definitions

So far, this chapter has examined the defining characteristics of six categories of analysis: data analysis (DA), confirmatory data analysis (CDA), exploratory data analysis (EDA), spatial data analysis (SDA), confirmatory spatial data analysis (CSDA), and exploratory spatial data analysis (ESDA). From these descriptions, then, it is possible to derive specific definitions for each. While the following definitions may seem redundant, each differs in a substantial way.

> Data analysis: the practice of manipulating data using a variety of methods/techniques for the purpose of increasing the understanding of particular phenomena, through describing, explaining, or predicting occurrences of those phenomena.

Confirmatory data analysis: the practice of manipulating data which result from experimentation, or are selected from observations under the guidance of an existing hypothesis, using a variety of methods/techniques for the purpose of supporting or refuting that hypothesis.

Exploratory data analysis: the practice of manipulating data resulting from observation, using a variety of methods/techniques for the purpose of identifying patterns within the data which may suggest further investigations or hypotheses.

Spatial data analysis: the practice of manipulating spatial data and associated attribute information using a variety of methods/techniques designed to include the spatial characteristics of phenomena, for the purpose of increasing understanding of particular phenomena, through describing, explaining, or predicting occurrences of those phenomena.

Confirmatory spatial data analysis: the practice of manipulating spatial data and associated attribute information which result from experimentation, or are selected from observations under the guidance of an existing hypothesis, using a variety of methods/techniques designed to include the spatial characteristics of phenomena, for the purpose of supporting or refuting that hypothesis.

Exploratory spatial data analysis: the practice of manipulating and producing a map view of spatial data and associated attribute information resulting from observation using a variety of methods/techniques designed to explore the spatial characteristics of phenomena, for the purpose of identifying patterns within the data which may suggest further investigations or hypotheses.

It is important to note that any researcher may use more than one type of analysis when conducting research. Figure 4.1 shows the relationships between these approaches.

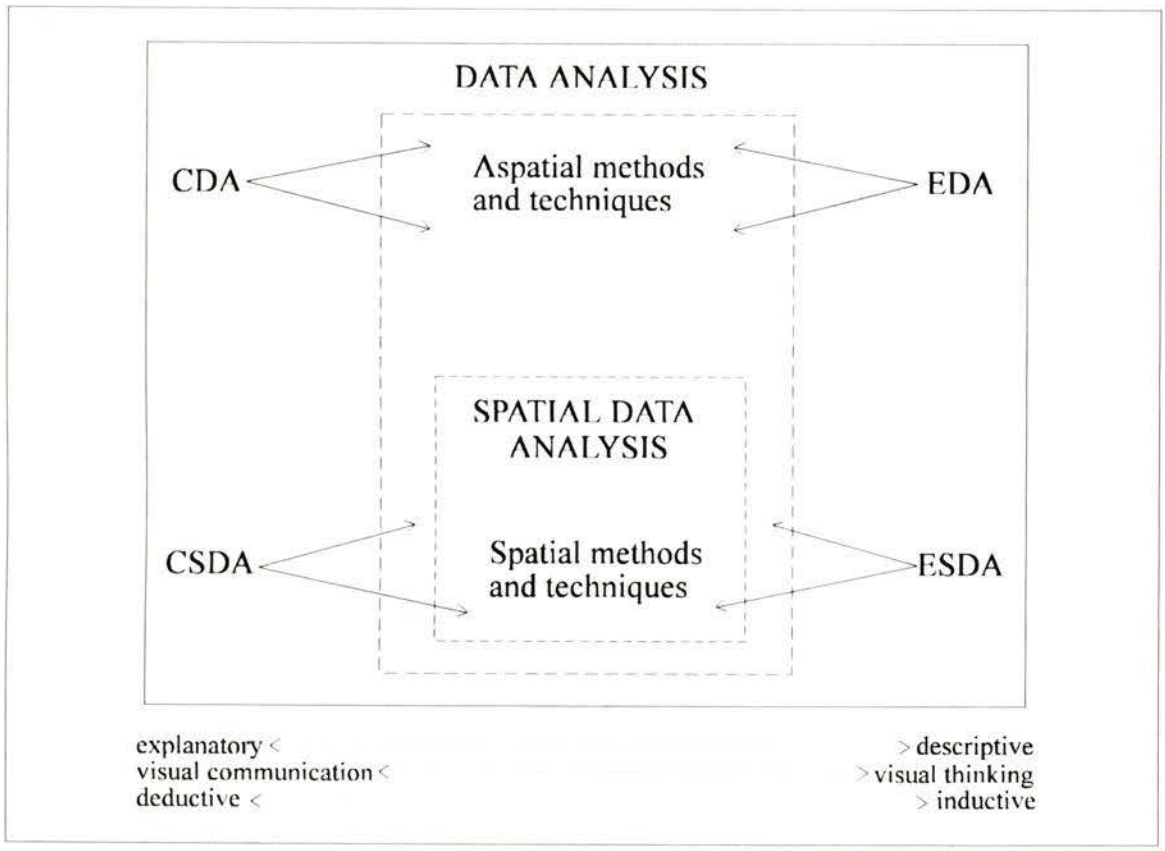


Figure 4.1. The Relationships Between Components of Data Analysis

Several important features should be noted. Firstly, data analysis is an inclusive term from which all subsequent approaches are derived. It is an 'umbrella', covering a wide range of techniques and methods of analysis, both spatial and aspatial. Secondly, while CSDA and ESDA may draw from spatial and aspatial methods and techniques, CDA and EDA are confined to using aspatial methods and techniques. Thirdly, there are close theoretical links between CDA and CSDA, and EDA and ESDA. The aim of research is

the same for both CDA and CSDA. In this case, both are used for confirming or refuting an a priori hypothesis. Similarly, the aim both EDA and ESDA, is to gain a fuller understanding of, and to uncover patterns within the data without an a priori hypothesis.

4.3. EXPLORATORY SPATIAL DATA ANALYSIS

The previous section has introduced specific definitions for a variety of approaches within data analysis. This section focuses specifically on expanding on the description of ESDA. ESDA was been defined on page 37 as:

> "the practice of manipulating and producing a map view of spatial data and associated attribute information resulting from observation using a variety of methods/techniques designed to explore the spatial characteristics of phenomena, for the purpose of identifying patterns within the data which may suggest further investigations or hypotheses".

As such, ESDA has been described as being linked to mathematical forms of analysis which are descriptive, to visual thinking, and to inductive reasoning. This is a distinct contrast to CSDA, but as stated before, while CSDA and ESDA possess distinct and differing characteristics, both approaches can be used within the scope of any one research effort. Fotheringham (1993) describes this relationship by making a distinction between pre-confirmatory ESDA and post-confirmatory ESDA, both of which work in a close iterative loop with CSDA toward developing geographic theory. In practice, this would consist of using pre-confirmatory ESDA techniques to formulate a model or hypothesis, followed by the use of CSDA techniques to test the model or hypothesis. The results of CSDA may generate 'new' data to be used for post-confirmatory ESDA. The results of post-confirmatory ESDA might then suggest new models or hypotheses, which can be tested using CSDA techniques, and so on, until some solid theory can be stated. In this way, the practice of ESDA is firmly embedded within the normal science paradigm.

An alternative to the normal science paradigm which has been suggested is the 'visualization paradigm', under which "it ...is acceptable to visually inspect an information database from as many angles as possible to seek a better understanding of that data, and to [arrive] at insights and research questions that may lead to observations and conclusions hitherto undiscovered" (Keller 1994, 298-299). MacEachern and Monmonier (1992) suggest that this "philosophical change...affords perception a more equal footing with mathematical and formal logic in scientific analysis...[and is] both driving and being facilitated by an explosion of observational...data," and that "human vision, instead of being considered a potential source of bias, has come to be recognized as a powerful tool for extracting patterns from chaos" (MacEachern and Monmonier 1992, 197).

An important idea implicit in these descriptions is that ESDA can be 'stand alone'. It is not necessary to follow ESDA with CSDA within a research effort; it is valid to conduct research for which the desired result is the production of new research questions and not the confirmation or refutation of any existing hypotheses. This is the essential difference from the normal science paradigm, which supports a process of hypothesis formation followed by confirmation or refutation via experiment and testing.

Whether one chooses to practice ESDA under either the normal science paradigm or the visualization paradigm, a key aspect of the ESDA methodology is the techniques, or methods used to perform ESDA. While there is no definitive set of techniques which apply specifically to ESDA, it is possible to divide those techniques which are commonly suggested or used into three categories: statistics, graphic views, and interaction techniques.

4.3.1. Statistics

Depending upon the nature of the data to be explored and the reason for exploration, most existing statistics, either spatial or aspatial, can be used for ESDA in conjunction with graphic and visualization techniques. Table 4.1 summarizes those statistics which have been identified as being useful for ESDA.

Table 4.1. Statistics Suggested or Used for ESDA

Author	Statistical Techniques
Goodchild (1992)	- smoothing/filtering algorithms - autocorrelation tests - correlation functions - outlier and spatial outlier detection
Haining (1994)	- aspatial descriptive measures: mean, median, mode variance, standard deviation, quartiles outlier tests - spatial measures tools to check heterogeneity tools to examine trends outlier tests spatial autocorrelation tests
Fotheringham (1993)	- spatially disaggregated measures: $G_i(d)$ of Getis and Ord (1992) - range from simple object count to regression parameter estimates - spatial outlier detection
Haslett et al (1991)	- anomaly detection

A wide range of statistical techniques have been identified, but there is an obvious emphasis in two areas: outlier/anomaly detection, and spatial autocorrelation measures. The objective of ESDA is to find patterns or trends within a dataset, and so techniques which identify spatial outliers, anomalies, and patterns such as clustering have obvious value.

4.3.2. Graphic Views

In the context of this research, graphic views refer to those which provide a view of the data or the statistical analysis results, other than as a table or spreadsheet. Table 4.2 summarizes those views which have been identified as useful within ESDA for areal data. Note that basic map views such as vector or raster representations are not listed as they are common to all cases. This summary is not exhaustive, but serves to identify common and well known techniques. A comprehensive survey can be found in Cleveland and McGill (1988).

Table 4.2. Graphic Views Suggested or Used for ESDA

Author	Graphic Techniques
Brodlie (1994)	- plots: scatterplots histograms - maps: contours surfaces
Haining (1994)	- plots: scatter histograms box plots stem and leaf diagrams spatial correlogram
Haslett et al. (1991)	- plots: scatterplots histograms line graphs variogram clouds vector with point and line overlays
Keller (1994)	- plots: use of Chernoff faces
Unwin (1993)	- plots: scatterplots, scatterplot matrices histograms boxplots bar charts

A key aspect of ESDA is the inclusion of one or more maps as visual presentations of the data. This allows aspatial statistical and graphic techniques to be evaluated within their geographic context, and thus incorporate the spatial characteristics of the data. Figure 4.2 shows how the inclusion of a map gives a different perspective on a data point which has been identified as an outlier through a scatterplot of the data. What appears to be an obvious outlier on the scatterplot (observation B) is in fact consistent with the surrounding areas, whereas a point which appears to be consistent with the data in the scatterplot (observation A) is an obvious anomaly on the map. Haslett et al. (1991) state that when searching for local or global anomalies, the map view is the single most important view, since it allows for the identification of anomalous regions which may not have particularly extreme data values. They cite the example of geological research, in which the subject of interest is areas which are unlike surrounding areas and so are anomalous, rather than those areas which are described only as having high or low values for some characteristic.

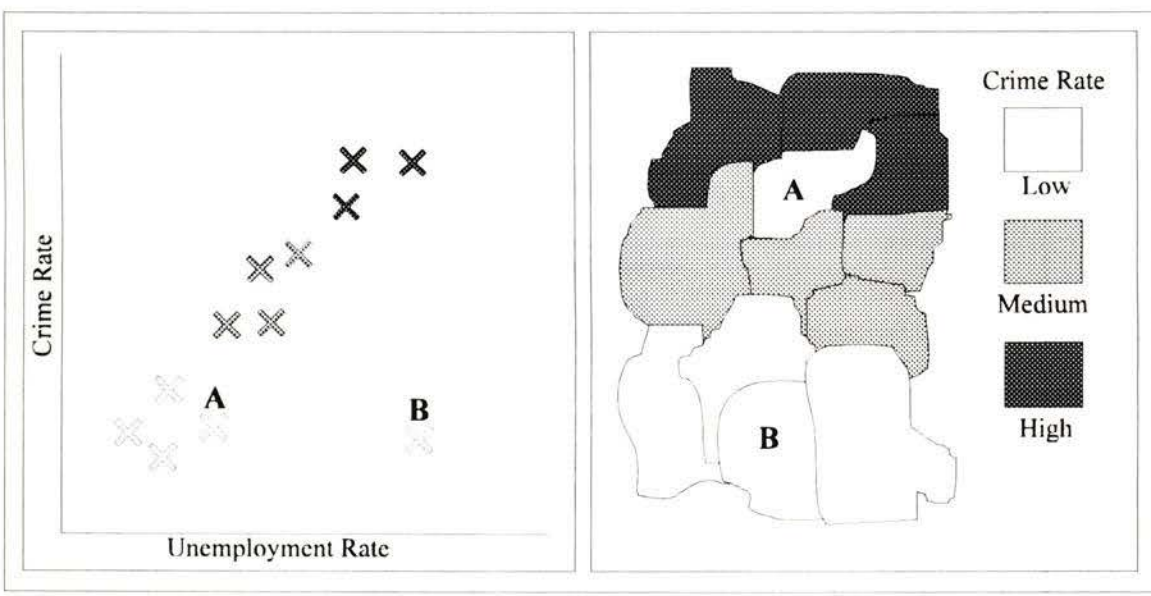


Figure 4.2. Two Views of the Same Data
(Adapted from Fotheringham 1993)

4.3.3 Interaction Techniques

As the name suggests, interaction techniques are those that allow the user to interact with the views in some fashion. This interaction between the results of statistical analyses and graphic views of the data is key to performing ESDA. As such, interactive techniques are comprised of methods for directly manipulating the statistical and/or graphic views in 'real time', so that the results of the manipulation are immediately apparent all views. Table 4.3 summarizes interaction techniques which may be used for ESDA.

Table 4.3. Interaction Techniques for ESDA

Author	**Interaction Techniques**
Becker, Cleveland, Wilks (1988)	- identification - deletion - linking - brushing - scaling (of graphs) - rotation
Dorling (1992)	- zoom in/zoom out - pan
Fotheringham (1993)	- brushing/windowing
Goodchild (1992)	- moving windows - deleting cases
Haslett et al (1991)	- moving windows - selecting - highlighting - linking
Unwin (1993)	- interrogation/selection - amending plots - rotation of plots - aggregation

For ESDA, the most innovative technique suggested to date is that of 'brushing'. Given a visual display which consists of several linked scatterplots representing various properties of the data, the analyst is able to select portions of the data on one scatterplot, and instantly see the same data high-lighted on all other scatterplots (Monmonier 1989, Becker et al. 1988). The important ideas here are: linked views, selection, and high-lighting. Extensions of this concept include attaching basic statistical operations to the 'brush', so that selected statistics are calculated from the data that fall within the 'brush' area, and are recalculated as the brush moves over the study area (Haslett et al. 1991).

4.4. CHAPTER SUMMARY

ESDA is a relatively new concept; it is only within the last decade that literature regarding ESDA has begun to appear. While ESDA falls within the general realm of data analysis, it is a unique approach. Under the rubric of data analysis (DA) fall a series of related approaches. The first is confirmatory data analysis (CDA) which is hypothesis confirming and associated with explanatory statistical techniques such as inferential tests, with visual communication through graphics, and with deductive reasoning. Exploratory data analysis (EDA) is the opposite of CDA, in that it is hypothesis generating, and associated with descriptive statistical techniques, visual thinking through graphics, and inductive reasoning. Spatial data analysis (SDA) is a subset of DA, and is defined by explicitly including the spatial characteristics of the phenomenon under study within the analyses performed. Confirmatory spatial data analysis (CSDA) is essentially CDA, with the spatial characteristics of the data providing important inputs to the statistical techniques used, while exploratory spatial data analysis (ESDA) is associated with EDA and is differentiated specifically due to the inclusion of the spatial characteristics of the data. The inclusion of maps as views of the data is perhaps the key defining characteristic of ESDA. Visualization, through dynamic and interactive linking of multiple 'views' of the data, allows for exploration of the associations between the views.

ESDA can operate under one of two research paradigms, either the normal science paradigm, or the visualization paradigm. Under the normal science paradigm, ESDA is used in conjunction with CSDA, with ESDA generating hypotheses and CSDA verifying those hypotheses. The final product is an explanatory and/or predictive result. Under the visualization paradigm, ESDA is a 'stand-alone' process which provides for a fuller understanding of the data and generates potential hypotheses. The end product in this case is increased understanding and new questions rather than a single 'correct' result.

As a methodology, ESDA consists of a set of tools and defines how those tools are used to provide methods for conducting research. Three tool sets of ESDA have been identified: statistical measures, graphic views, and interactive techniques. Statistical measures range from simple descriptive measures to more sophisticated outlier and anomaly detection methods. Graphic views consists of maps and a variety of plots such as histograms and scatterplots. Interactive techniques include selection, highlighting, identification, and rotation of plots. When used in conjunction, these tools provide for a powerful way of performing ESDA.

CHAPTER 5

A CONCEPTUAL FRAMEWORK FOR SENSITIVITY REPORTING

5.1. INTRODUCTION

In Chapters 2 and 3, it was established that MAUP affects a wide range of statistical analyses, from simple descriptive measures to more sophisticated models, and that in many cases the effects are unpredictable. A number of approaches have been used to investigate MAUP, for the purpose of either solving the problem or minimizing the effects. The alternative of reporting MAUP effects explicitly has been suggested. An ESDA methodology has been identified as having potential use in MAUP research, and a number of authors have suggested the use of ESDA for performing MAUP sensitivity analyses. This chapter focuses on developing a conceptual framework which identifies the characteristics of sensitivity reporting using an ESDA methodology, and the technical environment required for using ESDA methods.

5.2. LEVELS OF REPORTING MAUP SENSITIVITY

Fotheringham (1989) suggests that the effects of MAUP on model calibration should be documented by reporting the results for different levels of aggregation and for different configurations. It is suggested here that this may be a somewhat limited view of MAUP sensitivity reporting. In Chapter 3, three approaches to MAUP investigation were identified: variable-centred, result-centred, or a combination of variable and result-centred (see pg. 31). As well, each author uses a common data transformation process of creating multiple aggregations and/or configurations of a set of base areal units to provide the datasets for the chosen analytical technique. Figure 5.1 shows how the common data transformation process and the general MAUP investigation approaches coincide to identify three levels at which MAUP sensitivity can be reported.

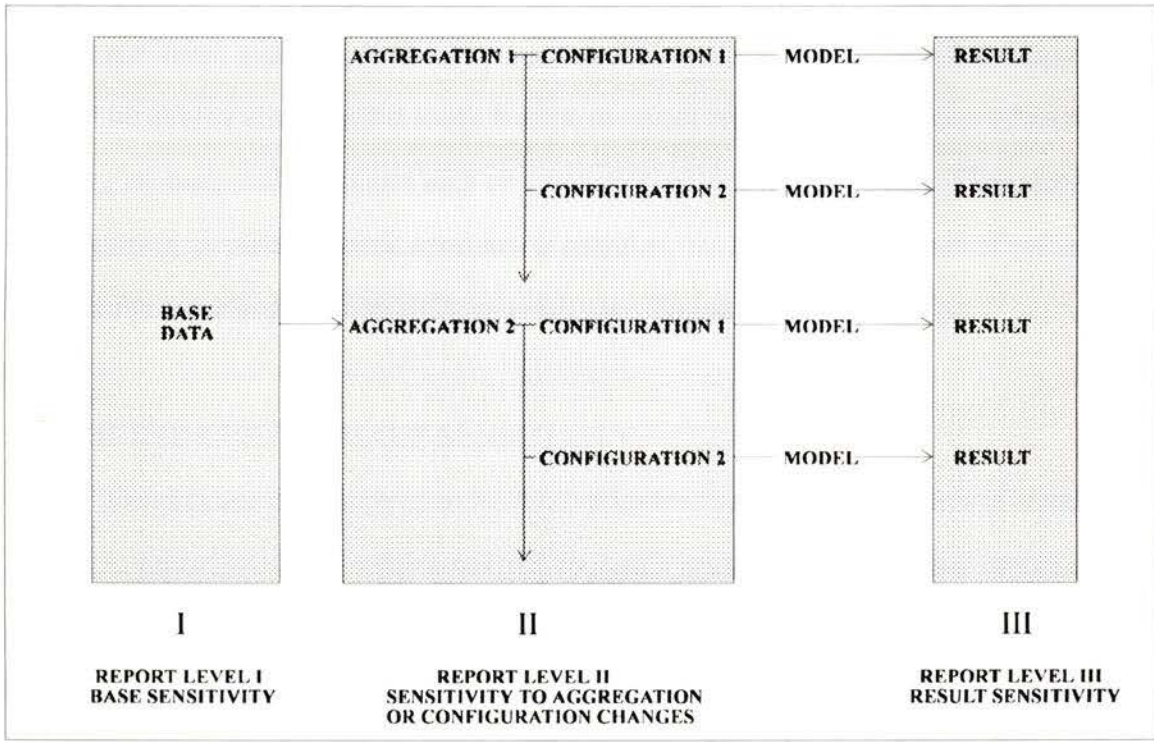


Figure 5.1. Three Levels for Reporting MAUP Sensitivity

Figure 5.1 suggests the hierarchical creation of multiple aggregations, and for each level of aggregation, multiple configurations. Analytical results can be derived for each unique combination of aggregation and configuration. There may exist an infinite number of possible combinations of aggregations and configurations. This reflects the data transformation process used by all of the authors reviewed in this thesis.

At report level I, the variables of the base dataset provide the basis for exploration and reporting. At this level, no aggregation or configuration changes have been made, nor have any models been applied. It is proposed here that the sensitivity of the original base variables should be addressed. Chapter 6 addresses potential measures which can be applied to the base variables. At report level II, aggregation and/or configuration changes have occurred, and it is proposed that the actual changes in the values of the variables due

to aggregation or configuration changes can be investigated. Again, a variable-centred approach is used, and Chapter 7 addresses potential measures which can be used to report MAUP sensitivity at this report level. Finally, at report level III, the results produced by applying an analytical technique such as a statistical model, are the subject of exploration and reporting. A result-centred approach is useful at this level, and methods of presenting information regarding the sensitivity of results are discussed in Chapter 8.

5.3. THE CONCEPT OF MAUP SENSITIVITY

The concept of sensitivity to MAUP varies for each of the report levels identified above. At report level I, a variable-centred approach is used to calculate and report the sensitivity of the base variables. Sensitivity at this level can be thought of as the *potential* for information change or loss exhibited by the data at its most disaggregate level. High sensitivity equates to high potential for information change or loss.

At the second report level, a variable-centred approach is used to calculate and report the sensitivity of the variables to changes in either aggregation and/or configuration. In this context, sensitivity measures will be based on the actual information change or loss exhibited by the base data under a set of aggregations or configurations. Those base areal unit values which exhibit a high degree of 'instability' can be designated as highly sensitive.

At the third report level, a result-centred approach is used to report the sensitivity of results obtained for the different combinations of aggregations and configurations. Sensitivity will be based on the range of results produced under a set of aggregations and/or configurations. A wide range of results for a particular set of aggregation and configuration combinations will identify that set as highly sensitive.

5.4 THE USES FOR A MAUP SENSITIVITY REPORT

MAUP sensitivity reporting, as suggested above, can be used in two ways. Firstly, a sensitivity report can be produced for the individual user's private use as an exploratory device. The user can explore and identify how sensitive variables are to MAUP, how different aggregations or configurations behave with respect to MAUP, and how sensitive analytical model results are to MAUP. Used in an iterative fashion, this type of sensitivity reporting allows for 'what-if' scenarios, and may help shape the final choice aggregation level and areal unit boundary definition.

Secondly, a sensitivity report may be used in a more public way, that is, as an accompaniment to a specific research endeavor. In this case, a MAUP sensitivity report allows for independent evaluation of the research in terms of the stability of results, and acts as a metadata statement. This use equates to the current implicit definition of sensitive reporting found in Fotheringham (1989).

5.5. NORMATIVE GOALS FOR A METHOD

The characteristics associated with MAUP sensitivity reporting can be combined with the techniques associated with the ESDA methodology to provide normative goals which guide the development of a method for reporting MAUP sensitivity. The following discussion identifies normative goals which arise from the characteristics of MAUP sensitivity reporting and which are consistent with the use of the ESDA methodology.

Previous investigations of MAUP effects tend to have been specific to a particular analytical technique. It is proposed here that a method for reporting MAUP sensitivity should be generic enough to be used for any analytical technique applied in any particular study. This implies that the method also should be data generic, to the extent that all types of areal data are susceptible to MAUP effects.

Three levels for reporting MAUP sensitivity have been identified: base variable sensitivity, variable sensitivity to aggregation/configuration, and result sensitivity. MAUP

reporting, therefore, should facilitate all three levels of sensitivity reporting, and incorporate both variable-centred and result-centred approaches.

MAUP sensitivity reports are required for different uses, and a method should be capable of facilitating the specific requirements of a variety of users. The ESDA methodology will define the range of exploratory and analytical elements for consideration, including map views and graph views such as scatterplots and histograms. Dynamic exploration of these elements should allow for private reporting, and for the static communication of MAUP sensitivity for public reports.

A method for MAUP sensitivity reporting requires that sensitivity to MAUP effects are quantified in some way. Different report levels and data types change the concept of sensitivity and so a number of measures rather than one universal measure may be required. The ESDA methodology defines the nature of these calculated measures: they should be primarily descriptive and include a means for outlier/anomaly detection. The exploration of these measures should include the use of a map view, therefore some of the measures should be spatially disaggregate, describing sensitivity for each areal unit.

New summary values derived for combined areal units can be calculated in a variety of ways, many of which make implicit assumptions about the nature of the data. Measures of sensitivity should reflect the method of calculating new summary values when appropriate.

Implicit in the concept of a MAUP sensitivity report is the evaluation of a range of aggregations and/or configurations. In conjunction with a method for creating MAUP sensitivity reports, a user should be able to use the ESDA methodology to create different aggregations and/or configuration with the associated new summary values calculated as required.

5.6. A TECHNICAL ENVIRONMENT FOR USING ESDA METHODS

In practice, ESDA must be conducted in a computerized environment, with linked access to attribute data, location data, statistical operations, and maps. There have been a number of research efforts toward performing ESDA in this type of environment, using a number of solution approaches. These approaches can be loosely categorized as either general ESDA applications or problem specific applications. This section focuses on describing these past efforts, and on describing and justifying the technical environment selected for this research.

5.6.1. General ESDA Applications

Xia and Fotheringham (1993) develop an ESDA module for GIS as a means of demonstrating the possibility of integrating GIS and ESDA. As such, their research is general, and the choice of ESDA techniques is arbitrary. Included in their module are a map view and four graph types - general scatterplots, histograms, spatial scatterplots, and weighted spatial scatterplots. Measures of spatial autocorrelation are also included - Moran's **I**, Geary's **c**, and both $G_i(d)$ and $G_i^*(d)$. Although their demonstration data are census based, centroid points are used for statistical calculations. Interaction techniques are limited to selection and highlighting. The module is based on the GIS package Arc/Info for SUN workstations. Both Arc Macro Language (AML) and C++ object-oriented language are used to create linked multi-window graphic and statistical views of the data, as well as a menu driven user interface.

The development of a different general ESDA application can be traced through three papers: Haslett, Wills, and Unwin (1990), Haslett et al. (1991), and Unwin (1993). The following summary draws on all three papers to describe SPIDER/REGARD, the ESDA application developed by these authors. The objective of these works is to demonstrate the possibilities of using statistical graphics as further views to a map. The tools developed emphasize the spatial query aspect of ESDA and provide for selecting and

high-lighting across linked views. In addition, a moving average, or moving region tool has been developed. The user interactively defines a circle of the desired distance, and a statistic is calculated using the data points within the circle. This statistic changes to reflect only the points within the circle as it is moved across the reference map, and a graph view or 'trace' of the generated values is available. Graphics include scatterplots, histograms, and variogram clouds. In the first two papers, only point data can be used, and the map view is created by plotting the X,Y coordinates of each sample point. Lines and images can be used, although they are solely for context and lack topology. By 1993, Unwin (1993) states that the capability to handle points, lines, and regions has been added to the application. Also included in this latter version is a functional language for calculating statistics, a wider range of graph types including boxplots and barcharts, and a prototype aggregation tool. The application was initially developed on a MAC II, using Pascal under MacApp as a programming language. It is not clear if the most recent version is running in the same environment.

In another general ESDA application, MacDougall (1992) uses JMP, a statistical package which supports interactive linking, to demonstrate the practice of ESDA. MacDougall uses the scatterplot function to produce a map view of point data by using the X,Y coordinates as axes. There is no capability of overlaying reference features. Raster data can be explored in the same manner as irregularly spaced point data, through conversion of the raster data into point data by calculating explicit locational coordinates from the row and column position of each cell. MacDougall suggests that this data transformation can be accomplished via a short custom program, or by the program's data transformation functions. MacDougall (1992) also describes a prototype program, called Polygon Explorer for the Macintosh computer which incorporates a map display, bar charts, histograms, scatterplots, and cluster analysis capabilities. Included are frequency counts and percent area reports, the ability to re-express the data as logarithms or square roots, and outlier detection using Tukey fences. Cluster analysis can be performed using

either the hierarchical or block methods. The program is limited to the use of one categorical and two continuous variables, and the map view consists only of the original polygons used. Interaction tools include selection and highlighting, as well as rotation of a three dimensional plot.

5.6.2. Problem Specific Applications

Batty and Xie (1994) use ARC/INFO running on a SUN microsystem platform as a software base and add a variety of functions for modelling urban density through the use of Arc Macro Language and C programming. The primary focus of their work is to develop urban modelling capabilities within GIS, with ESDA providing information for designing zoning systems under a set of requirements. Specifically, maps, surfaces and scatterplots are linked and used for the detection of outliers by visual inspection. It also is possible to create multiple aggregations or configurations of whole areal units using a variety of methods, including on-screen selection, location seeding, attribute seeding, or interpolation in which attribute values from one set of areal units are translated for a different set of areal units. In this way the results of a chosen model can be explored for sensitivity to changes in aggregation or configuration.

Scott (1994) describes an effort at integrating ARC/INFO and STATA through MS-Windows, in order to perform exploratory analysis of data quality. The specific objective is to identify outliers or anomalies which may represent data entry errors. The prototype includes map and graph views of the data produced by simple overlays, regression analysis, and trend surface analysis; however, none of these views are interactively linked. Each program runs in a separate window, but all can be displayed at the same time. The program does allow for full GIS and STATA functionality with no additional programming.

A series of papers by Openshaw deserve mention here. In Openshaw et al. (1987) a Geographical Analysis Machine (GAM) is developed specifically for finding significant patterns in point data. Openshaw, Cross and Charlton (1990) describe a Geographical Correlates Exploration Machine (GCEM) which is used for exploring the correlations between a point dataset (the dependent variable) and a series of areal datasets (the independent variables). Openshaw (1993) describes the Space-Time-Attribute Analysis Machine (STAM) which searches for patterns in geographic, temporal, and attribute space. The STAM is refined further by creating a Space-Time-Attribute Creature (STAC), which is used to search for patterns intelligently. Although stated explicitly only for the GCEM, the development environment is assumed to consist of Arc/Info and an external programming language, running on a Cray X-MP/48. These papers present interesting and original work; however, Openshaw's emphasis clearly is different than that of the other authors reviewed here. Openshaw's objective is to automate the exploration process and produce a summary of statistically significant results. These results may then encourage the formation of hypotheses about the results. While the GAM, GCEM, and STAM/STAC certainly are exploratory, the automation of the process precludes any interactive exploration or visualization of the process by the user.

5.6.3. The Research Development and Demonstration Environment

A number of possible technical environments can be used to implement ESDA applications. Firstly, an existing statistical application which supports dynamic linking can be used as a base, although this poses limits on the potential for generating map views of the data. This approach is demonstrated by MacDougall (1992). Secondly, an existing GIS can be used as a base, with statistical capabilities and interactive links specifically programmed. This option may be limiting in the range of statistics which can be easily calculated, and may require programming outside of the GIS environment. Xia and Fotheringham (1993) use Arc/Info as a base and create additional programming through

AML and C++. Batty and Xie (1994) use Arc/Info as a base and create additional programming through AML and C. Thirdly, a 'stand-alone' system can be developed which incorporates both statistical and GIS functionality; however, this approach essentially 're-invents the wheel' as there are existing software packages which contain some if not all of the required functionality. The latter approach is used, however, by Haslett et al. (1991) to develop SPIDER/REGARD, and by MacDougall (1992) for the development of the Polygon Explorer prototype application. Fourthly, an existing GIS and an existing statistical package may be used concurrently, which allows for full functionality of both applications. Scott (1994) demonstrates this approach; however, the applications run side-by-side rather than in an integrated fashion. Interactive, dynamic linking was not possible.

The potential for using both a GIS and a statistical application has become more attractive with the recently developed link between S+ and Arc/Info. This link allows for full functionality of both applications from within Arc/Info. It is possible to run Arc/Info with S+ in the 'background. In this way the user can send commands to S+ directly from the Arc/Info prompt, without switching windows or exiting the Arc/Info session. Data may be transferred between the applications as a matter of course, again by using commands from the Arc/Info prompt. This link represents a major extension of Arc/Info in terms of statistical capabilities. At this time, S+ appears to be the only commercially available software which provides a ready-made link to Arc/Info. Available hardware and software at the University of Victoria, where this research is being conducted, currently supports Arc/Info and S+/S+Spatialstats, running on a SUN Sparc workstation under Solaris 2.4. Since these software applications are available and fulfill many of the requirements for using an ESDA methodology, the methods developed in this research will be tested and demonstrated using Arc/Info and S+/S+Spatialstats in an integrated fashion.

5.7. CHAPTER SUMMARY

A number of characteristics are embodied within the concept of MAUP sensitivity reporting. Most importantly, there are three levels at which sensitivity can be reported: base variable sensitivity (report level I), variable sensitivity to aggregation/configuration (report level II), and result sensitivity (report level III). These three levels form the basis for developing the methods of sensitivity reporting since each level has unique characteristics. The concept of sensitivity changes at each reporting level, and may require different sensitivity measures. The way in which a sensitivity report is used also can vary. A sensitivity report can be private, in that the information provided is used in an iterative fashion to explore MAUP effects, or a report can be public when it is included with a specific study and can provide a mechanism for evaluating validity.

The characteristics of a MAUP sensitivity report, and the methodology of ESDA can be combined to form a set of normative goals for a method of reporting MAUP sensitivity. The method should be generic to a range of analytical techniques and areal data types; incorporate variable-centred and result-centred approaches; produce private and public reports; use descriptive, spatially disaggregated measures which reflect different concepts of sensitivity and different methods for calculating new values for combined areal units; and allow the user to create new aggregations and/or configurations of the base data.

There are a number of approaches to implementing ESDA applications: statistics application based, GIS based, stand-alone, or concurrent use of a statistic package and GIS. The newly available S+/S+Spatialstats package supports a wide range of both aspatial and spatial statistics and a provides a link to Arc/Info. Since the current technical environment at the University of Victoria supports both S+/S+Spatialstats and Arc/Info, methods for reporting MAUP sensitivity will be demonstrated using these software packages.

CHAPTER 6

BASE VARIABLE SENSITIVITY

6.1. INTRODUCTION

This chapter is the first of three which extend the conceptual framework by developing each of the three report levels identified in Chapter 5 more fully. This chapter is specifically concerned with the first level of reporting, identified previously as base variable sensitivity (see Figure 6.1). The following sections address the concept of sensitivity associated with this level, review existing descriptive statistics which may contribute to measuring sensitivity, propose measures specific to MAUP sensitivity, and demonstrate the use of the proposed measures.

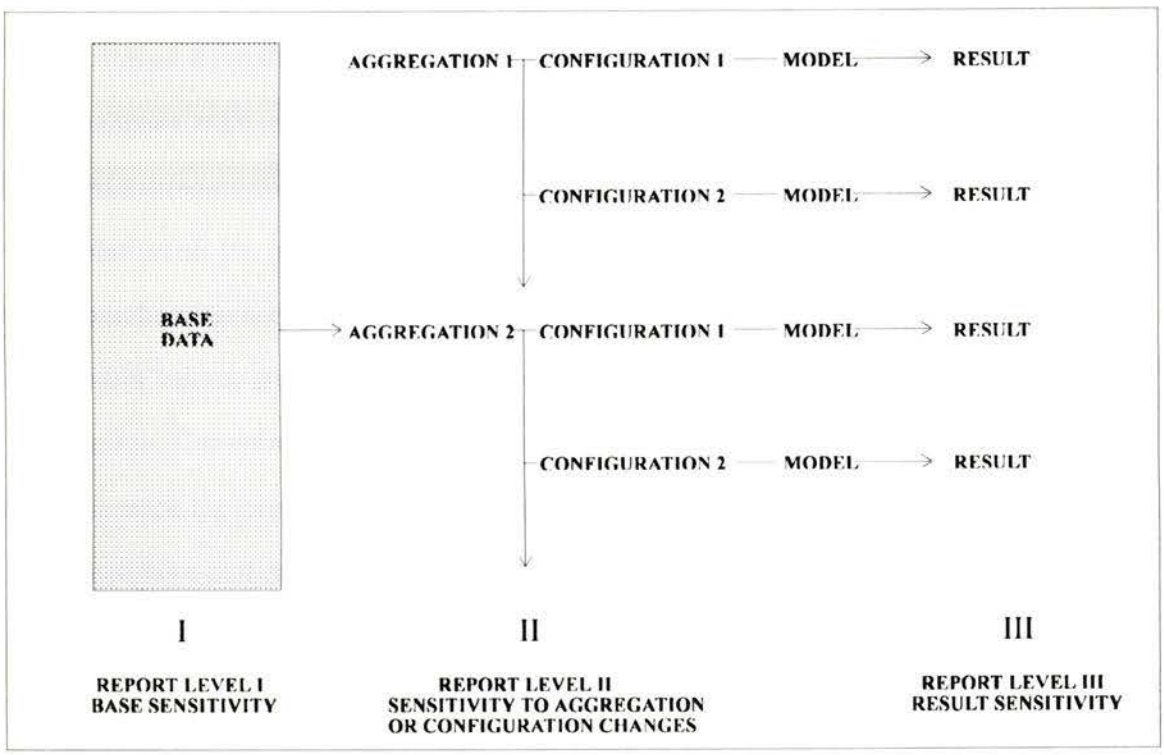


Figure 6.1. Report Level I - Base Variable Sensitivity

It is important to note that the following discussions are relevant only to cardinal level areal summary measures for irregular polygons. Issues concerning the application of the demonstrated measures and methods to categorical data and to raster format data are addressed at the end of this chapter.

6.2. A CONCEPT OF BASE VARIABLE SENSITIVITY

The creation of MAUP effects is attributed to the change in the summary values of a variable resulting from successive aggregation and/or configuration changes of the base areal units. At report level I, no aggregation or configuration changes have occurred, nor has any analytical technique been applied. For this reason, a variable-centred approach is appropriate for this level of reporting, focusing on the *potential* for changes to the summary values given aggregation and/or configuration changes.

Recall from Chapter 3 that one way to avoid MAUP effects is to group only homogeneous areal units. For example, given a group of three areal units, each with an assigned average value of 10, the use of a simple average, an areally weighted average, or recalculation of the average to derive a new summary measure results in a new average of 10. When homogeneous areal units are grouped, the new value truly represents the constituent values, and thus loss of information is limited to a decrease in distribution information. Conversely, when heterogeneous areal units are grouped, MAUP effects may increase since the new value is not truly representative of the constituent values. It is suggested here that as the variation between grouped areal unit values increases, the new value becomes less representative overall and the effects of MAUP increase.

Recall also that MAUP is specific to *spatial* aggregation and configuration. The *potential* sensitivity of any particular areal unit, therefore, depends on how its own value compares to the values of its contiguous neighbours, that is, those areal units with which grouping is likely. The use of 'neighbours' is common in certain types of spatial analysis.

When data are areal in nature, two methods for defining neighbours are generally used, one based on proximity, the other based on connectedness.

Defining neighbours by proximity uses distance as the primary measure. Centroid points for each areal unit are calculated, and only those areal units with centroids within a specified distance of the 'target' areal unit centroid are considered as neighbours.

Defining neighbours based on connectedness consists of identifying those areal units which share either a common boundary, a common vertex, or both with respect to the 'target' areal unit. These three options are known as the rook's case, the bishop's case, and the queen's case respectively (see Cliff and Ord 1973, 16-17). When using connectedness as a basis for defining neighbours, multiple levels of neighbours can be identified and are generally differentiated as 'orders'. For example, a first order neighbour is one that directly shares a boundary with the target areal unit. A second order neighbour is one that directly shares a boundary with the first order neighbour of the target areal unit, a third order neighbour directly shares a boundary with the second order neighbour, and so on (see Arbia 1989, 5). In this research, the queen's case of connectedness is used to define neighbours - any areal unit which shares either a boundary segment or a vertex is considered a 1st order neighbour with respect to the target areal unit.

An areal unit with a low value surrounded by contiguous areal units with generally high values has potentially high sensitivity to MAUP. The single low value areal unit can be described as a spatial outlier. Figure 6.2 is a simple depiction of both potential high and potential low sensitivity. In example (A), the central areal unit is the spatial outlier: its value is unlike the values of its contiguous neighbours. When grouped with its contiguous neighbours, a simple average calculation results in a new value of 6.7. The new value is fairly representative of all of constituent areal units with the notable exception of the central areal unit, the spatial outlier. In example (B), the central areal unit is not a spatial outlier, and the group value calculated is fairly representative of all constituent areal units.

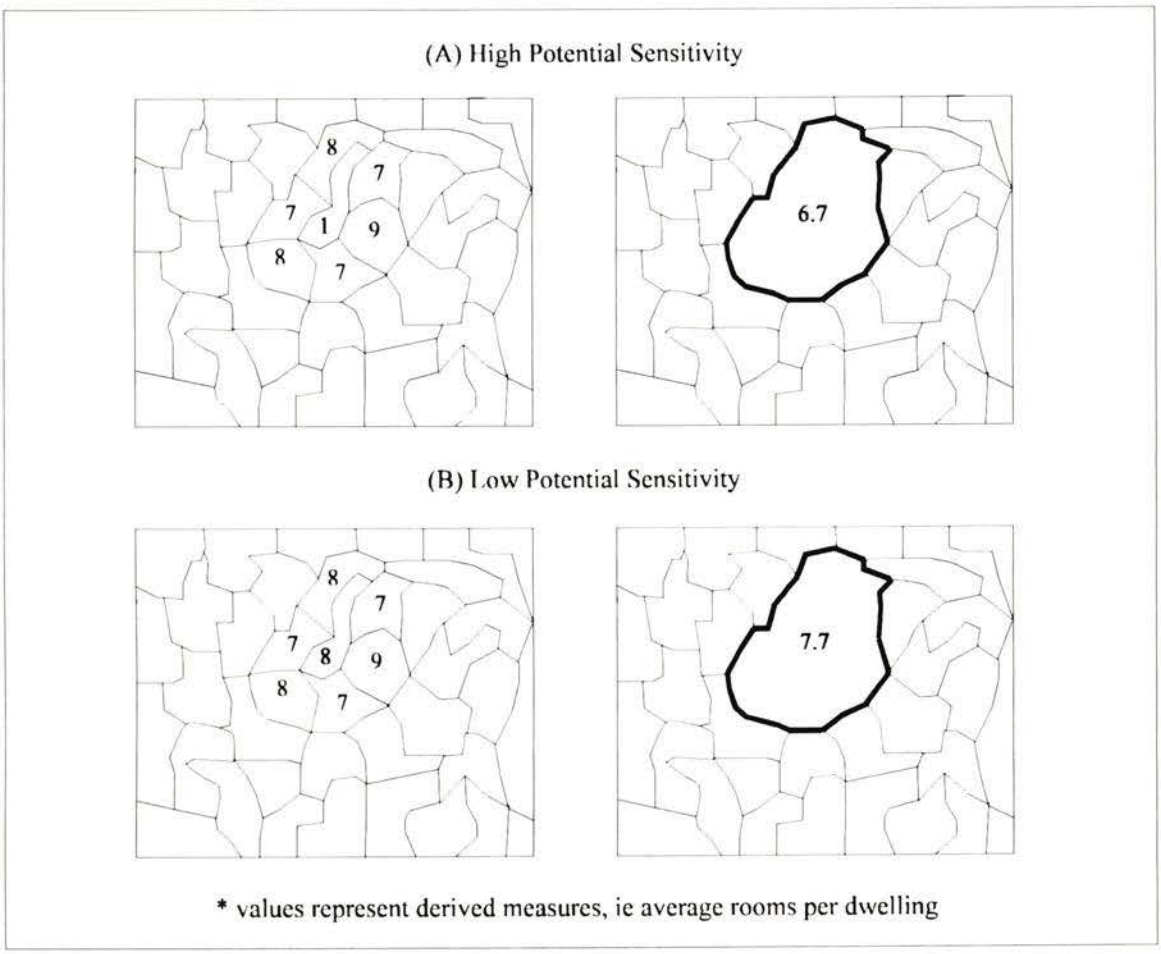


Figure 6.2. High and Low Potential Sensitivity Situations

There are two important exceptions to be noted. Firstly, recall from Chapter 2, cardinal level areal data are most commonly expressed as frequency counts, percentages, averages, or densities. Areal data which consist of frequency counts are not subject to the same problem of representativeness. When frequency counts for areal units are combined, the new value is a sum of the constituent values. For example, two areal units, with populations of 100 and 200 respectively, can be combined to form a new areal unit with a population of 300. In this case, the question of whether or not the new value, 300, is representative of the original values is meaningless. The problem is one of loss of distributional information, rather than one of representativeness. The loss of distributional

information may impact analytical results, and so a result-centred approach may be more appropriate for reporting this sensitivity. Secondly, point data derived from areal units (i.e. polygon centroids), are affected by aggregation and/or configuration changes, and any analyses using this type of point data will be sensitive to MAUP effects. Again, a result-centred approach is felt to be more appropriate for reporting this sensitivity. Incorporating either frequency data or point data sensitivity into a report is addressed in Chapter 8. The following discussion pertains then only to cardinal level summary measures such as averages, percentages, or densities.

Recall that there exist different methods for calculating new summary measures given aggregation or configuration changes. The method used to calculate new values affects the definition of potential sensitivity to MAUP. For example, when calculating new values for average income (all income/total population), three methods can be used: a simple average, an areally weighted average, or recalculation based on the components. When simple averaging is used, extreme data values will have the most influence on the new value. When areally weighted averaging is used, values associated with large areas will have the most influence. When new values are recalculated based on the component values, density measures associated with large areas will have the most influence while for average or percent measures, values associated with large observation populations will have the most influence.

6.3. EXISTING MEASURES

The concept of *potential* sensitivity to MAUP can be linked to that of spatial autocorrelation. When spatial autocorrelation is highly positive for a particular variable, it can be stated that the values of neighbouring areal units or points are very similar i.e. correlated; when spatial autocorrelation is highly negative, the likelihood of the values of neighbouring areal units or points being similar is low. The possibility of using a measure of spatial autocorrelation as an indicator of potential sensitivity is discussed in this section.

Two commonly used measures of spatial autocorrelation are Moran's **I** and Geary's **c**. Both measures operate in a similar way. Both Moran's **I** and Geary's **c** compare the variability of the values in a specified neighbourhood of areal units to the variability of the values for the entire study region. Moran's **I** uses the cross product of deviations from the mean for each comparison pair, while Geary's **c** uses the squared difference between each comparison pair. A significant component in each equation is a weighting factor which identifies the areal units to be considered as neighbours and their relative importance. The weighting factor can be based on simple adjacency (1 if areal units are adjacent, 0 if not) or can be a spatial function. Cliff and Ord (1973) give the example of using distance from centroid points and length of common boundary between comparison pairs to create a weighting factor. Complete descriptions of these measures and the associated theory can be found in Griffith (1987) or Cliff and Ord (1973). In this form, both Moran's **I** and Geary's **c** are global measures since the final result is a single descriptive measure for the entire study area. This limits the utility of these specific measures for reporting MAUP sensitivity, since the interest at this level of reporting is in the sensitivity of each specific areal unit.

Getis and Ord (1992) note that along with Moran's **I** and Geary's **c**, semi-variance measures and methods for estimating spatial autocorrelation coefficients of regression equations are also applied globally. As an alternative, Getis and Ord (1992) develop a spatially disaggregate measure, $G_i(d)$, based on the use of distance statistics. In essence, $G_i(d)$ is the ratio of the sum of all distance weighted variable X values within a neighbourhood to the sum of all variable X values within the neighbourhood. $G_i(d)$ does not include the value of the 'target' areal unit in the calculation, however an alternative measure, $G_i^*(d)$ does include the 'target' value. A value of $G_i(d)$ or $G_i^*(d)$ is assigned to each areal unit and therefore can be mapped. Getis and Ord (1992), Anselin (1993), and Getis (1994) demonstrate the $G_i(d)$ measure as a map view. The values of $G_i(d)$ can be either negative or positive: a highly negative $G_i(d)$ indicates the clustering of small variable

X values, whereas a highly positive $G_i(d)$ indicates the clustering of large variable X values (Anselin 1993, 2). The use of $G_i(d)$ and $G_i^*(d)$ is limited to ratio level data with no negative values (Getis and Ord, 1992). Also, the measure is dependent on measuring distances between centroid points of areal units, and so may be subject to variation given different distance measures.

Dykes (1994), although mainly concerned with using autocorrelation measures as a way to aid in the classification of choropleth maps, notes that:

> " restrictions of the spatial autocorrelation index involve uncertainty surrounding the definition of a standard weight function meaning that Moran's **I** is no different from many other weighted spatial statistics in that it may be manipulated to 'prove' that a dataset behaves in a particular manner" (Dykes, 1994, 107).

Along the same line of reasoning Dykes (1994) states that:

> "...areal value data...exhibit an intrinsic paucity of distance-related information, and attempts to compensate for this such as centroid calculation, $1/d^2$ modelling, and common boundary length coefficients are inadequate..[therefore] the only remaining spatial parameter is adjacency ..." (Dykes 1994, 108).

Dykes uses a multinominal adjacency-based function, adapted from an image processing technique to demonstrate the mapping of co-occurrence values between choropleth classes.

While Moran's **I**, Geary's **c**, $G_i(d)$, and Dykes function can be used as disaggregate measures, the planned method of calculating new summary measures needs to be incorporated. The measures described above do not reflect the possibility of different sensitivities, and so the next section proposes a potential measure which can be adjusted to incorporate data types and calculation methods.

6.4. PROPOSED MEASURES

The proposed measure, SB_t, for the sensitivity of a cardinal level summary measure such as average, percentage or density, is:

$$SB_t = \frac{\sum_{i=1}^{n} W_{ti} \left(| V_t - V_i | \right) x_{ti}}{\sum_{i=1}^{n} x_{ti}} \qquad \forall t, (t = 1...n) \quad (6.1)$$

where:

V_t = variable value for areal unit t
V_i = variable value for areal unit i
x_{ti} = 1 if areal unit i is adjacent to areal unit t, 0 otherwise
W_{ti} = weighting factor
n = number of areal units in study area

Note that: $\sum_{i=1}^{n} x_{ti}$ will equal the number of areal units neighbouring areal unit t.

In essence, the measure is the average difference between the target areal unit variable value and the corresponding values of any contiguous neighbours. In this research, neighbours are identified based on their connectedness to the target areal unit. Any areal unit sharing a boundary or a vertex is considered a 1st order neighbour. Any areal unit sharing a boundary or a vertex with a 1st order neighbour is considered a 2nd order neighbour and so on. Other options should be included in an ESDA application for exploring and reporting sensitivity to MAUP, such as using a distance measure based on areal unit centroids, or the interactive choice of neighbours. The weighting factor is used to adjust the influence of neighbouring values according to the data type and the planned method of calculating new values for grouped areal units.

Table 6.1 lists suggested weighting functions according to data type and planned calculation method.

Table 6.1. Suggested Weights for SB_t

Planned Calculation Method:	Data Type density	Data Type % or average
Simple Average	$W_{ti} = 1$	$W_{ti} = 1$
Areally Weighted Average	N/A	$W_{ti} = \frac{A_i}{A_t}$
Recalculation	$W_{ti} = \frac{A_i}{A_t}$	$W_{ti} = \frac{P_i}{P_t}$

A_i = area of neighbour areal unit P_i = observation population of neighbour areal unit
A_t = area of target areal unit P_t = observation population of target areal unit

Recall that when using a simple average, all values are accorded an equal influence on the resultant summary measure, and so the weighting factor is set at 1. When areally weighted averaging is used, values associated with large areas will have the most influence on the resultant summary measure, and so the new summary value may be more representative of the influential constituent values. In this case, the weight factor proposed is a ratio of a neighbour areal unit area and the target areal unit area, A_i / A_t. This adjusts SB_t by increasing the sensitivity value when the target areal unit is small in comparison to one or more neighbouring areal units, and by decreasing the sensitivity value when the target areal unit is large in comparison to one or more neighbouring areal units. When new values are recalculated based on the component values, density measures associated with large areas will have the most influence on the resultant summary measure, while for

averages or percentages, values associated with large observation populations will have the most influence on the resultant summary measure. In all cases, the new summary measure may be more representative of the influential constituent values, and less representative of the non-influential values.

For density measures, the weighting factor suggested, A_i / A_t, adjusts SB_t by increasing the sensitivity value when the target areal unit is small in comparison to one or more neighbouring areal units, and by decreasing the sensitivity value when the target areal unit is large in comparison to one or more neighbouring areal units. For averages and percentages, the weighting factor proposed is P_i / P_t . This adjusts SB_t by increasing the sensitivity value when the observation population of the target areal unit is small in comparison to one or more of the neighbouring areal units, and by decreasing the sensitivity value when the observation population of the target areal unit is large in comparison to one or more of the neighbouring areal units.

A specific problem is encountered when using observation populations to weight SB_t. It may occur that some areal units in a study area do not contain the phenomenon of interest, and so there are no observations associated with those areal units. Given an observation population of 0 for a target areal unit, the population weighted calculation of SB_t cannot be performed since 0 becomes the denominator of the weighting factor. In these cases, it is suggested that any target areal unit with an observation population of 0 be assigned an SB_t value of -1. Since the natural range of SB_t values is from 0 to some upper limit defined by the data, a value of -1 would serve to indicate an unusual situation. In fact, target areal units with 0 observation populations will be highly sensitive, since the areal unit will take on the value of any other areal unit with which it is combined. The calculation of SB_t is not affected adversely when an areal unit with a 0 value occurs as a neighbour.

The problem of 0 observation populations does not occur when any of the other weighting functions are used for the calculation of SB_t; however, the more general issue

of missing data must be addressed. Recall that SB_f measures potential sensitivity for densities, averages, or percentages. In all cases, the data range from 0 upwards. It is not possible to have a negative density, average, or percentage measure. When data are missing, there a several options for noting the occurrence. Commonly, a value is assigned which falls outside the normal range of the data, such as -9999. In this way, 'nodata' can be differentiated numerically. The use of such an extreme value, however, would exert undue influence on the calculation of SB_f. It is suggested that areal units with missing data be assigned a value of 0 when calculating SB_f. In this way, an areal unit with missing data is treated in the same fashion as an areal unit with no observation population. In essence, this entails the creation of a new dataset to use for calculation purposes, while keeping the raw data, which differentiates between 0 values and 'nodata' values, available for future reference.

An important feature of SB_f is that it does not reflect the sensitivity of an areal unit to actual aggregations or configurations, but rather reflects the sensitivity of the areal unit to its general surroundings as defined by the choice of neighbours. SB_f describes local sensitivity when low order neighbourhoods are used, or regional sensitivity when higher order neighbourhoods are used. The choice of neighbourhood levels may be guided by the planned aggregation or configuration change. For example, when configuration changes are planned, rather than aggregation changes, it may suffice to use 1st order neighbours. The spatial characteristics of each areal unit will also affect the neighbourhood size. For example, a large areal unit may have a large number of neighbours and cover a large portion of the study area. Similarly, an areal unit which is contained entirely within another areal unit will have only one st order neighbour. An areal unit which is separated from all others by some physical barrier such as water will have no neighbours at all, and so SB_f cannot be calculated when neighbours are defined by order. Instead, some other method of choosing neighbours must be available, such as manual selection or distance measures.

SB_t is univariate since it describes the sensitivity of a specified variable. For multivariate analyses, it would be useful to compare SB_t values for a number of variables. Of particular interest are areal units which have spatially coincident high values of SB_t for a set of variables, or spatially coincident low values of SB_t for the variable set. An average of SB_t values for each of the chosen variables provides an indication of the spatial coincidence of SB_t values; therefore, the proposed measure for multivariate sensitivity is:

$$SC_t = \frac{\sum_{j=1}^{m} SB_{tj}\, x_{tj}}{\sum_{j=1}^{m} x_{tj}} \qquad \forall t, (t = 1 \ldots n) \qquad (6.2)$$

where:
SB_{tj} = SB_t for variable j
x_{tj} = 0 if SB_{tj} = -1, 1 otherwise
n = number of areal units in the study area
m = number of variables

Note that : $\sum_{j=1}^{m} x_{tj}$ will equal the number of variables summed

SB_t naturally ranges from 0 to some upper limit defined by the range of the data, and may include negative values when observation population based weights are used. For this reason, some standardization of SB_t should occur prior to calculating SC_t. There are a number of ways which can be used to create standard scales; however, in this research, normal distribution of the raw data is assumed and Z scores are used as a way of standardizing the raw data prior to any calculation of the proposed measures.

The use of Z scores to standardize the data does limit the calculation of SC_I to using only those SB_I values calculated with the same weighting function. SB_I values derived using different weighting functions can be used, although it would be necessary to first calculate SB_I from the unstandardized data and then standardize the resultant values. In either case, only those values which result from the use of the same neighbour definition should be averaged to produce SC_I.

6.5. USING THE PROPOSED MEASURES

For demonstration purposes, the software packages Arc/Info and S+ are used. The calculation of SB_I was performed in Arc/Info using a short Arc Macro Language (AML) program written under the author's supervision by G. Garlick. SC_I was calculated in SPSS prior to the installation of S+ during the later stage of this research. For this thesis, SB_I measures are based on 1st order neighbours and are unweighted; all new summary values are calculated using a simple average.

The data used for demonstration are four cardinal level variables from the 1991 Census of approximately 484 enumeration areas which comprise the Victoria Census Metropolitan Area:

A) average value of owned dwellings,

B) average household income,

C) average persons per room, and

D) percent owner one family households.

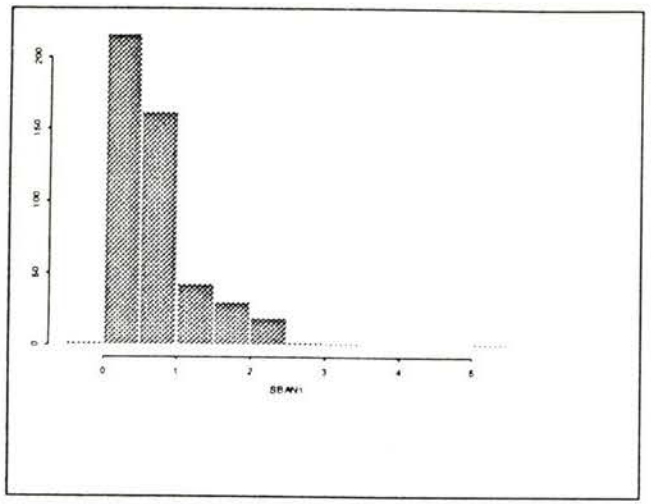
Approximately 70 enumeration areas have no population and so have 0 values for all four variables. In addition to the 0 value data, variables B, C, and D are missing values for 21, 1, and 40 additional enumeration areas respectively. All occurrences of 'nodata' values have been assigned a value of 0 prior to conversion of the raw data to Z scores. There are 9 enumeration areas which are islands; these have been removed from the data. Similarly, there are 5 enumeration areas which have no observation population and are

contained wholly within larger enumeration areas with no observation populations. These 5 enumeration areas have also been removed from the data, leaving a total of 470 enumeration areas in the dataset.

There are two general questions about the sensitivity of base variables which can be addressed using the proposed measures. The first question is: where are the potentially sensitive areal units located? The second question asks: where are areas of coincident potential sensitivity for the chosen set of variables? The following two sections address each question in turn.

6.5.1. Locating Potentially Sensitive Areas

Figure 6.3 shows three elements which represent different views of the unweighted SB_f measure for variable A. The first is a map of the enumeration area boundaries which define the study area. The two supplementary views are a histogram and a table. The SB_f measure for average income is shown as "SBAN1". This naming convention identifies the measure (SB_f), the variable (A), and the neighbour definition (N1, all first order neighbours). In this example, values for SB_f which are greater than 1 are have been selected on the histogram, the associated areas on the map are shaded, and the table lists the data associated with the selected areas. This approximates the ESDA technique of brushing. Values greater than 1 have been selected because the histogram shows that the majority of values fall between 0 and 1, and so values greater than 1 might be classified as having relatively high sensitivity. Note that an ESDA application supporting data brushing should allow any single value or range of values to be selected from the histogram, or any region to be selected from the map. The shaded areas on the map indicate those areas or regions which perhaps should be left unchanged when aggregating or re-configuring the base areal units. When leaving potentially sensitive areas or regions unchanged is not an option, these areas should be noted for further exploration after aggregating or re-configuring changes have been made, or addressed explicitly.



A	Az	SBAN1
188000	6.92	5.09
86151	2.24	2.27
99631	2.86	1.98
67359	1.38	1.97
81019	2.00	1.96
126000	4.07	1.95
74543	1.71	1.87
33071	-.20	1.72

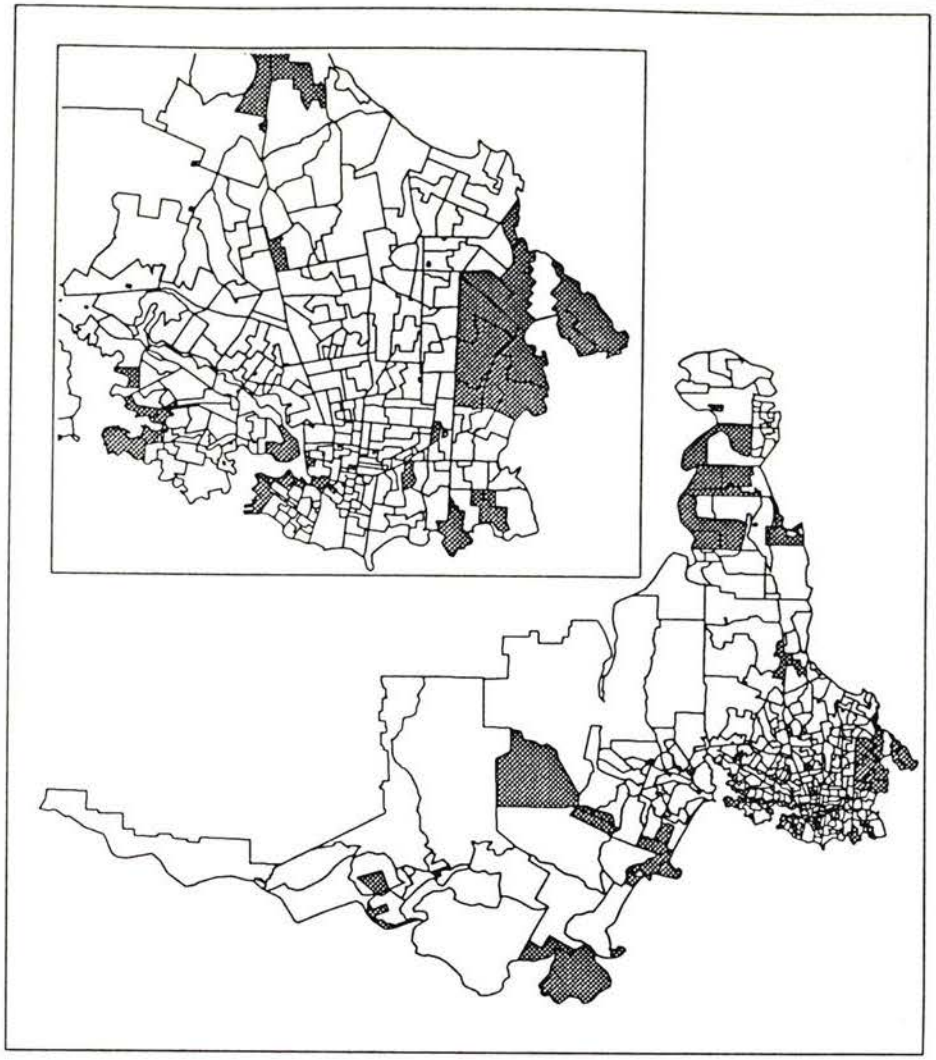


Figure 6.3. Sbt for Variable A - SBAN1 greater than 1

6.5.2. Locating Areas of Coincident Potential Sensitivity Between Variables

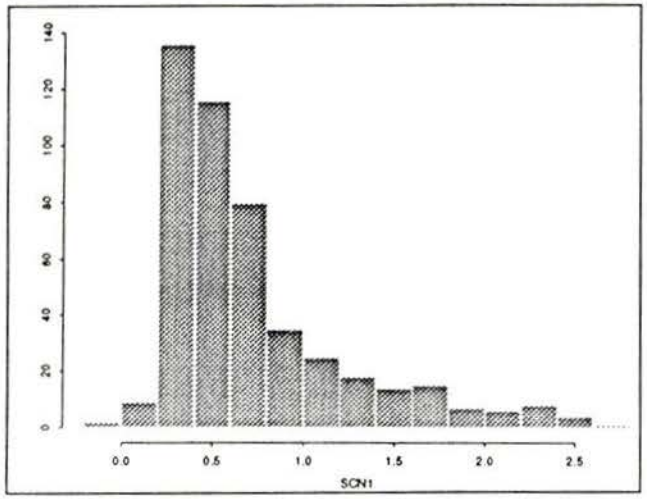
When the absolute values of SB_f for a set of variables are averaged, the combined measure, SC_f is the result. Figure 6.4 shows three views of SCN1: a map, histogram, and table. Note again the naming convention used for SC_f : SC identifies the measure as SC_f , and N1 identifies the neighbourhood level. SC_f is the average of SB_f values for variables A, B, C, and D. SC_f values greater than 1 have been selected on the histogram, shaded on the map, and associated data are shown in the table. Again, this approximates brushing.

The shaded areas on the map represent areas of coincident sensitivity between the four variables. The interpretation of this pattern is similar to that for SB_f. Areas which show high SC_f values contain at least one variable which has high potential sensitivity. These areas could be left unchanged by subsequent aggregation or reconfiguration; however, when this is not an option, these areas should be explored more fully in order to ascertain the cause of the sensitivity indicated.

6.6. APPLICATION TO RASTER FORMAT AND CATEGORICAL DATA

This chapter so far has been limited to the development and demonstration of sensitivity measures for cardinal level summary measures at report level I, the base variable sensitivity level. The following discussion presents several issues concerning the application of base sensitivity measures to raster format data and/or categorical data.

Cardinal level measures in raster format data such as reflectance measures from remote sensing can be incorporated easily. The measure of SB_f remains the same; however, a weight function based on area is no longer necessary since all areal units have the same area. The SB_f measurement acts as a filtering algorithm when applied to raster data. There is, however, an important distinction between SB_f and those filtering algorithms commonly used for image processing. The objective of using a filter on an image is to assign a new value to a cell based on the values of its neighbours in order to



SBAN1	SBBN1	SBCN1	SBDN1	SCN1
2.70	3.34	2.12	2.72	2.72
2.30	3.52	1.96	2.49	2.57
2.45	2.62	2.61	2.60	2.57
2.46	2.68	2.61	1.74	2.37
2.42	2.18	2.44	2.30	2.34
2.18	1.53	3.00	1.89	2.15
2.18	1.90	2.61	1.39	2.02
5.09	2.78	0.00	0.11	1.99

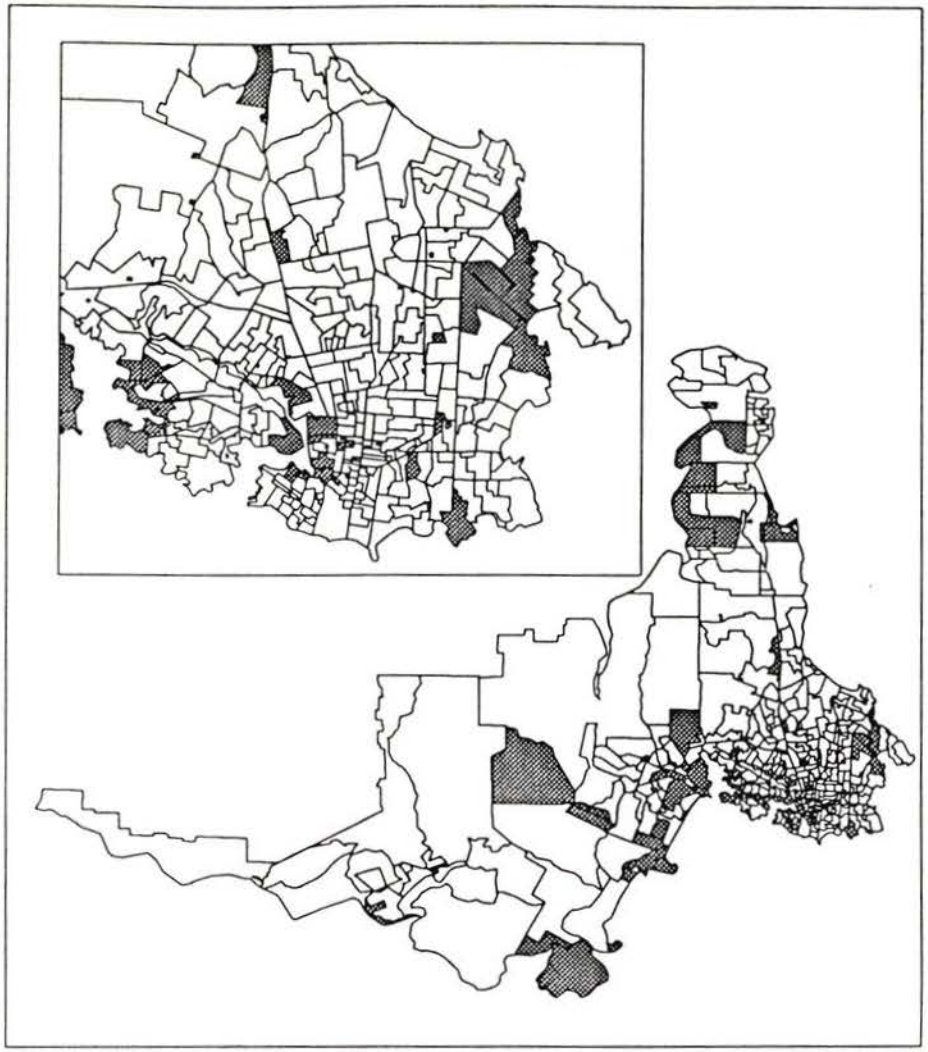


Figure 6.4. SCt for Four Variables - SCN1 greater than 1

smooth the image or incorporate the additional information available in nearby cells. These new values are then used to provide a basis for creating a thematic classification of the image. The objective of using SB_I is to measure the *potential for change* when a particular cell is merged with its neighbours to create a larger cell, and so indicates potential information loss.

A more complex situation occurs when categorical data are used in either vector or raster format. The SB_I measure demonstrated above is based on the numerical difference between values of a specified variable. It is not possible to measure a numerical difference between categories, and so the measures become binary in that values are either 'the same' or 'different'. One option is to assign any comparison of two values a 1 if the values are different, or a 0 if the values are the same. When a single variable is evaluated, SB_I could consist of a simple total of the binary measures for all contiguous neighbours, divided by the total number of neighbours. For example, 3 out of 4 different contiguous neighbours would produce an SB_I of 0.75. The coincidence of sensitivity could be a simple average of these measures.

Methods for assigning new summary measures, in this case categories, to grouped areal units differ, and will influence the corresponding concept of sensitivity. For example, using a presence/absence method, those areal units which contain the desired category will have no sensitivity, while all those areal units which do not have the desired category will be highly sensitive since their classification value will change. Alternatively, when dominant area is used to assign a new category, target areal units which are surrounded by areal units of different categories will be sensitive and those surrounded by areal units with the same categories will not be sensitive. In essence then, a different sensitivity measure may be required for each method used, as was the case for cardinal data.

Although there may be some potential in these suggestions, there is at least one pitfall as well. If this type of measure is applied to cardinal data, 'different' must be numerically defined; for example, any value that is greater than 2 units high or lower than

the target unit is considered different. This, in effect, creates a classification of the data. Questions concerning appropriate methods for classifying data have occupied not only geographers, but researchers in almost every discipline.

6.7. CHAPTER SUMMARY

The first level of MAUP sensitivity reporting identified in the conceptual framework developed in this research is that of base variable sensitivity. Since no aggregations or configurations have occurred at this level, nor has any analytical technique been applied, a variable-centred approach is used to develop the concept of sensitivity. The focus is on *potential* sensitivity and is related to the degree of similarity or dissimilarity between any particular areal unit value and the values of contiguous areal units. A sensitive area is one which is very unlike its neighbours in terms of value, and is therefore a spatial outlier. There are two exceptions to this definition: data in the form of frequency counts and point data which are derived from areal units.

The concept of base variable sensitivity developed in this chapter suggests that a spatially disaggregate measure of sensitivity should be used which allows for the identification of spatial outliers. The measure should incorporate the different concepts of sensitivity which arise from the use of different calculation methods which may be applied later.

Measures of spatial autocorrelation are used to describe areas in terms of their neighbourhoods, and so may be of potential use. Moran's **I** and Geary's **c** are global in nature in that a single measure for a study area is the result. Both may be adjusted to give more spatially disaggregate results; however, there is no potential for including the influence of attribute calculation methods. $G_i(d)$, developed by Cliff and Ord (1992) is a disaggregated spatial autocorrelation measure; however, as in the previous example, $G_i(d)$ cannot be easily adjusted to include the influence of attribute calculation methods.

Dykes (1994) suggests that the only useful spatial information inherent in areal units is adjacency, also called 'connectedness' in this thesis.

In order to meet the requirements set out by the characteristics of MAUP and the ESDA methodology, two measures are proposed. The first, referred to as SB_f, is the average difference between a target areal unit and its contiguous neighbours. The measure can be adjusted using a weighting function based on the planned method for calculating new values in subsequent aggregations or re-configurations of the data. No weighting factor is used when a simple average is planned. A weighting factor based on the areas of the target areal unit and the areas of the contiguous areal units is suggested when areally weighted averages are planned. When recalculation is planned, a weighting factor based on areas is used for density measures, and a weighting factor based on observation populations is used for percentages and averages. High SB_f values indicate areas which have high potential sensitivity to aggregation or configuration changes.

When an observation population based weight is used specifically, as would be the case when the data are averages or percentages and the new summary measures will be recalculated using the measure components, it is suggested that those areas which have no observation population can be explicitly noted through assigning - 1 as a specific SB_f value. A negative value would not normally occur in SB_f, and so -1 serves to indicate an unusual situation. This adjustment is necessary only in the case where the data are averages or percentages, observation population based weights are used, and recalculation is the planned method for calculating new summary measures.

Missing data also complicates the calculation of SB_f. In general, missing data can be noted with an extreme value such as - 9999; however, it is suggested that missing data be given a value of 0 and be treated in the same fashion as true 0 value data. This necessitates the creation of a new dataset to be used specifically for calculation purposes while the original dataset can be retained for future reference. Similarly, any areal units which are 'islands' create a problem for the calculation of SB_f when neighbours are based

on orders (i.e. connectivity). Other methods for choosing neighbours should be available, such as manual selection or distance based measures which could deal with such 'islands'.

A second measure, SC_f, is proposed to measure coincident sensitivity between variables. The measure is the average of the SB_f values associated with each variable in a given set for a given areal unit. The scales of the data should be standardized in some way prior to calculating SC_f. In this research, normal distribution of the raw data has been assumed, and the raw data have been converted to Z scores prior to any calculations. SB_f is expressed in standard units, and all SB_f measures can be directly averaged to produce SC_f without further scale standardization. Both 0 population and 'nodata' values were assigned a value of 0 prior to conversion to Z scores. In this way, both types of data are given the lowest possible Z scores and so act as highly sensitive areas, as is appropriate.

This research focuses on cardinal level data in vector format, and the application of the proposed measures to raster format data or to categorical data raises several issues. While the application of SB_f and SC_f to raster format cardinal data is relatively unproblematic, the issues which arise when the measures are applied to categorical data are much more complex. Both SB_f and SC_f are based on the numerical difference between values for areal units. A numerical difference is not possible given nominal or ordinal data and so the measurement of difference could become binary in nature, i.e. different or not different. As is the case for cardinal data, the method for calculating new values influences the definition of sensitivity, and so a number of measures may be necessary. Extending the 'binary' measure to cardinal data is particularly problematic in that 'difference' must be defined for cardinal data. Defining 'difference' is in fact a classification of the data, and there are an almost infinite number of ways in which this could be performed.

CHAPTER 7

SENSITIVITY TO AGGREGATION OR CONFIGURATION CHANGES

7.1. INTRODUCTION

This chapter is the second of three which extend the conceptual framework by developing each of the three sensitivity report levels identified in Chapter 5 more fully. This chapter is concerned specifically with the second level of reporting, identified previously as variable sensitivity to aggregation/configuration (see Figure 7.1). The following sections address the concept of sensitivity associated with this level, review existing descriptive statistics which may contribute to measuring sensitivity, propose measures specific to MAUP sensitivity, and demonstrate the use of the proposed measures.

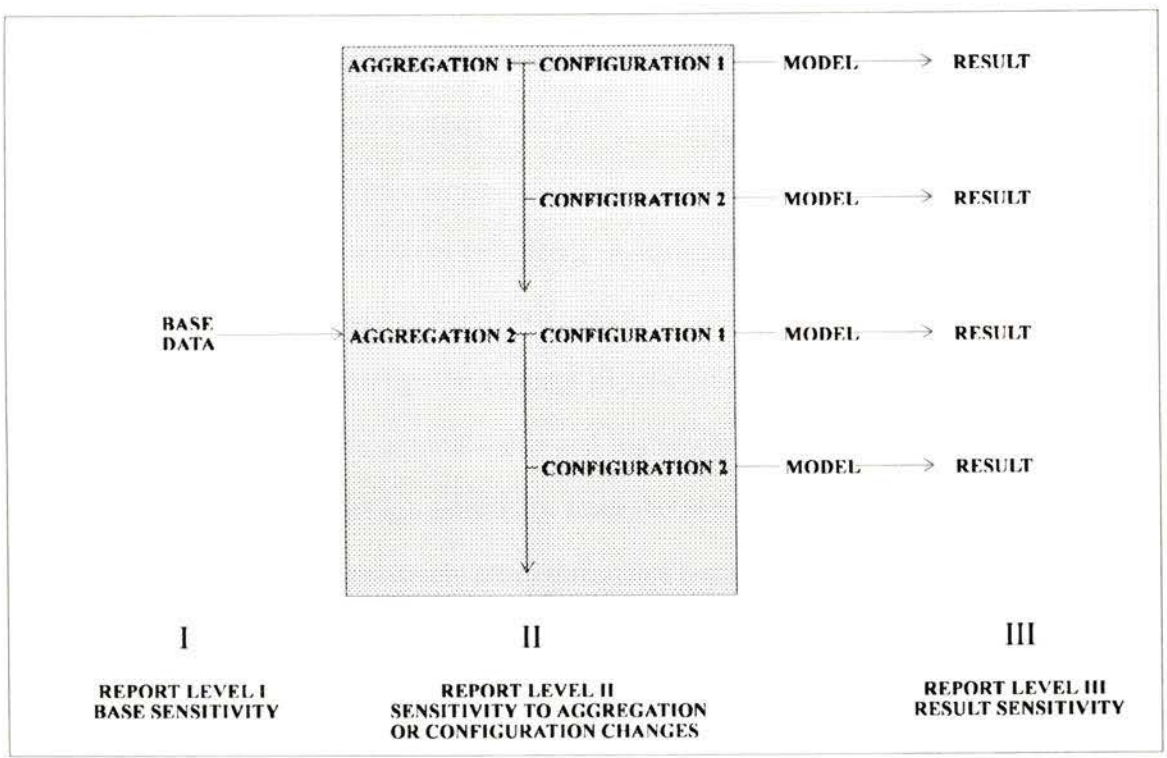


Figure 7.1. Report Level II - Variable Sensitivity to Aggregation or Configuration

In all cases, the focus is on cardinal level measures in vector format. The application of the demonstrated measures and methods to categorical data and to raster format data is addressed at the end of this chapter.

7.2. A CONCEPT OF SENSITIVITY TO AGGREGATION OR CONFIGURATION CHANGES

MAUP effects are created when new values for variables are calculated according to aggregation or configuration changes. In Chapter 5, since no aggregation or configuration changes had been made, the concept of sensitivity was derived from the characteristics of the base variables themselves, and was referred to as *potential* sensitivity. At this level of reporting, aggregation and/or configuration changes have been made, and so it is possible to look at *actual* sensitivity. Previously, the method of calculating new summary measures was identified as an important factor in defining the sensitivity of base variables; however, this is not the case for sensitivity to aggregation or configuration. Given the new summary values for each aggregation or configuration, it is possible to look directly at the changes which occur in the variable values, regardless of how the new summary values were calculated. A variable-centred approach is appropriate, since at this level no analysis has been undertaken.

At this level of reporting, the concept of sensitivity is straightforward: a relatively large difference between the original value and the newly calculated value for any particular variable indicates high sensitivity since the new value is not representative of the original value. Similarly, a relatively small difference between these values indicates low sensitivity since representation is better. Figure 7.2 shows this relationship diagrammatically. There are two important features to note. Firstly, the base areal units and associated values provide the reference for comparing new values. Figure 7.2 shows only one reference unit, but for each grouping, all base areal units can be assigned a specific sensitivity in comparison to the value of the new areal unit in which they are

included. Secondly, sensitivity is dependent on the specific aggregation or configuration and may vary accordingly. For Aggregation X, the reference areal unit with a value of 1 has a relatively low sensitivity since there is a relatively small difference between the original and the new value of 1.33. For Aggregation Y, the reference areal unit has a relatively high sensitivity since there is a relatively large difference between the original value of 1 and the new value of 5.

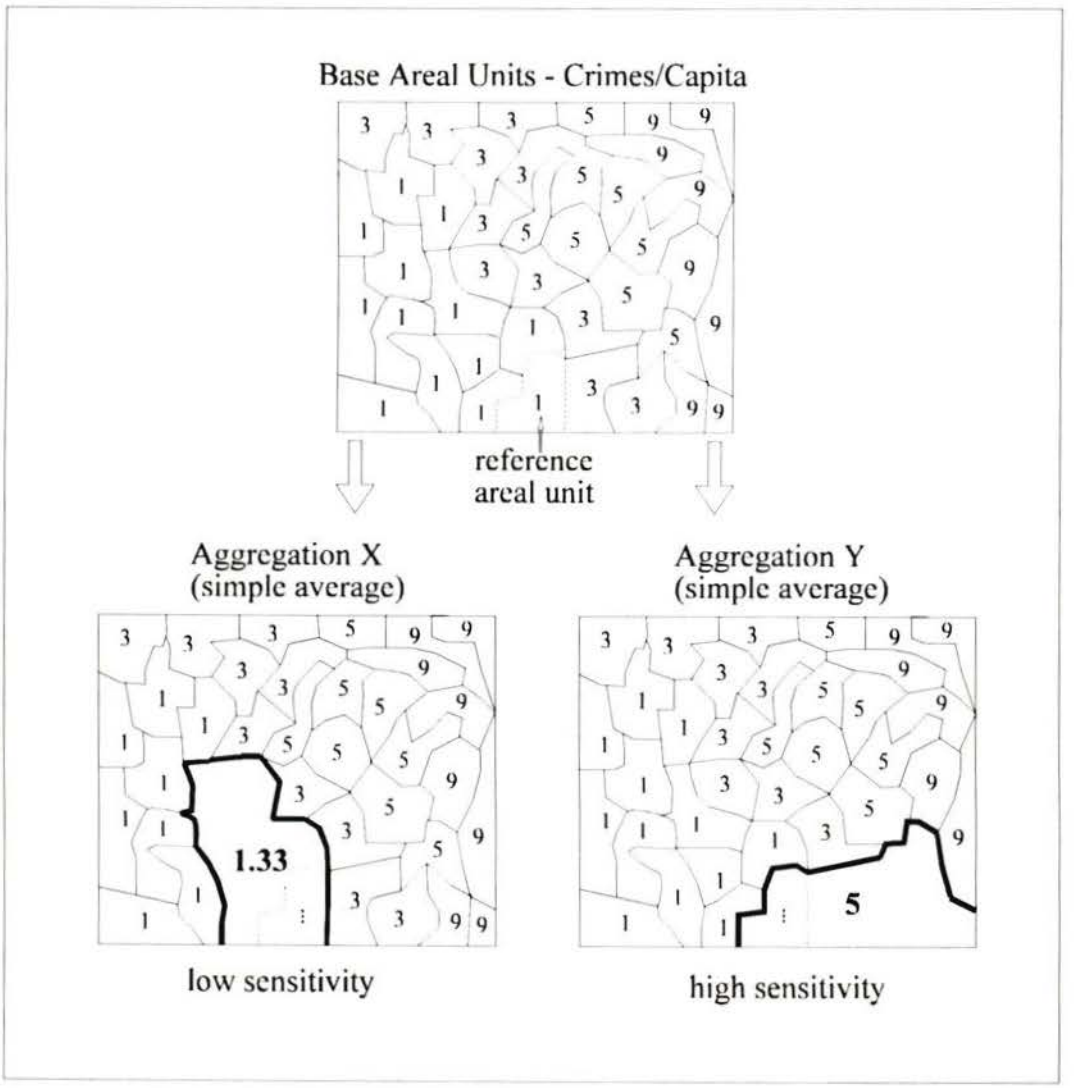


Figure 7.2. Sensitivity to Aggregation or Configuration Changes

In general, when a new grouping consists of base areal units with a wide range of values (i.e. the grouping is heterogeneous), the new value will be less representative of each constituent value and sensitivities will be higher. When a new grouping consists of base areal units with similar values (i.e. the grouping is homogeneous), sensitivities will be lower. Each specific aggregation or configuration will create a unique spatial pattern of sensitivities.

The concept of sensitivity to aggregation and configuration changes developed here does not apply to all types of data, a characteristic shared with the concept for base variable sensitivity developed in Chapter 5. When the data associated with the base areal units represent frequency counts, the idea that the new value calculated for a given aggregation or configuration is not representative is meaningless. Recall that when two or more areal units are combined, the frequency counts are simply summed and the new value represents a total rather than an average value, and so is not intended to summarize those constituent values with a new, representative value. When points which have been derived from areal units are used, comparison across layers is difficult since the point location itself may move according to the centroid position of the new group. For these reasons, both frequency data and point data are not addressed here, but will be discussed in Chapter 8, which focuses on sensitivity of analysis results.

7.3. EXISTING MEASURES

The sensitivity of any particular base areal unit variable to MAUP can be conceptually linked to the variance exhibited within new groups. Measures of variance have been used quite widely for geographic analyses. More specifically, a number of authors have proposed measures based on variance to study a geographic distribution at a number of aggregation levels. This seems to be an appropriate place to look for a suitable measure for sensitivity to aggregation and configuration changes.

Moellering and Tobler (1972) present a method for using analysis of variance to identify the aggregation level which exhibits the highest variance, and therefore indicates the level at which a process is operating. This work is linked to those strategies which attempt to minimize MAUP by identifying the 'correct' aggregation level at which to perform a particular analysis. A number of restrictions apply to the method proposed by Moellering and Tobler (1972): input data must be totally enumerated, any spatial unit must be totally contained in one and only one unit on the spatial scale above it, and each spatial unit at one level must contain at least one unit of the next lower scale. It has been identified in Chapter 2 that, in fact, aggregation or configuration changes may occur which do not use the base areal units as whole entities, but rather reassign portions to new groups. The method used by Moellering and Tobler (1972) results in a 'global' measure for each aggregation level. Batty (1976) presents a method which makes use of entropy statistics. The method results in a global measure which can be 'decomposed' into the proportion contributed to the variance for each aggregation level used, just as that of Moellering and Tobler (1972). A major drawback in each of these methods is the omission of configuration effects. It may be possible to configure one aggregation level in such a way as to produce a higher variance measure than that of a different level. This would create real ambiguity in the analysis.

The concept developed for sensitivity at this level is embedded in the base areal units and their associated original values. As such, a spatially disaggregate, variable-centred measure is indicated, as opposed to a global measure. Rather than relying on relatively sophisticated statistics, a more tractable solution may lie within the field of cartographic modelling. Cartographic modelling "is a general methodology for the analysis and synthesis of geographical data..[employing] what amounts to...algebra..."(Tomlin 1991, 361).

In practice, cartographic modelling is one of the fundamental functions of many GIS and can be conducted through the use of the common 'overlay' function for vector

In (c), the maps (a) and (b) have been 'overlayed', producing the map shown which contains new polygons, each with a particular combination of data - one value from map (a), and one value from map (b). A table is shown in (d) which lists the data for each new polygon in clockwise fashion. It is a simple matter to perform algebraic operations on the two columns of data. In (d), the third column shows the numerical difference between values from (a) and values from (b). The value given as '(a) - (b)' should not be interpreted as a new density measure. Instead, the values identify where, and by how much, the density measures of map (a) differ from the density measures of map (b). The use of subtraction is often used in GIS to compare maps which make use of the same measures but may have been produced using different classifications or at different times. Tomlin (1991) presents a range of cartographic modelling techniques which include 'local' operations such as:

> "LocalCombination [which] computes a value that uniquely identifies the particular combination of existing values that are associated with each location on two or more specified layers...LocalDifference [which] subtracts from each location's value on one specified layer its value(s) on one or more additional layers...[and] LocalSum [which] adds each locations' value on one specified layer to its value(s) on one or more additional layers..."
>
> (Tomlin 1991, 367-368).

The type of operations used in cartographic modeling are well suited to producing the spatially disaggregate measures required by ESDA. Similarly, the technical environment for this research, particularly the software application Arc/Info, supports the overlay functions necessary. The following proposed measures for sensitivity to aggregation and configuration changes therefore are based on the techniques associated with cartographic modelling.

7.4. PROPOSED MEASURES

Given an overlay of a set of base areal units and an aggregation or new configuration, the proposed measure of sensitivity is:

$$D(l)_{kbj} = (l)_{kj} - (l)_{kb} \qquad \forall k, (k = 1,...,n) \qquad (7.1)$$

where:
n = number of polygons in overlay map
(l) = a specific variable
$(l)_{kj}$ = new summary value associated with aggregation or configuration
$(l)_{kb}$ = base value.

Recall that the overlay operation creates a 'new' set of boundaries which are defined by both the base and the aggregation boundaries of the two component maps, and therefore $D(l)_{kbj}$ values are associated with these newly created '*k*' polygons. This measure is simply the difference between the two sets of values within each 'k' polygon, similar to the operation illustrated in Figure 7.2. Note that the direction of change in value between the first map and the second map is retained.

When the sensitivity of a particular variable to more than one aggregation or configuration is measured, a multiple overlay is performed. It is proposed that the average of the absolute $D(l)_{kbj}$ values for each aggregation of configuration can be used to represent sensitivity to the series:

$$DM_k = \frac{\sum_{j=1}^{p} |D(l)_{kbj}|}{p} \qquad \forall k, (k = 1,...,n) \qquad (7.2)$$

where:
n = number of polygons in overlay map
p = number of aggregations and/or configurations evaluated

A large value indicates that the variable is relatively unstable, or sensitive to one or more of the aggregations or configurations, while a low value indicates relative stability. The direction of change is no longer available due to the use of absolute values.

In some cases, it will be desirable to compare the sensitivity of a set of variables to a single aggregation or configuration change in order to identify any areas of coincident sensitivity among the variables. The measures proposed above can be extended to cover this situation. Given a specific aggregation or configuration change, a single overlay can be performed, and the average of absolute $D(l)_{kbj}$ values can be calculated:

$$MD(l)_{kbj} = \frac{\sum_{l=1}^{s} |D(l)_{kbj}|}{s} \qquad \forall k, (k = 1,\ldots,n) \qquad (7.3)$$

where:
n = number of polygons in overlay map
s = number of variables evaluated

$MD(l)_{kbj}$ values for a set of variables can indicate whether there is relatively little change, or relatively high change, in the set of variables.

When the interest lies in evaluating a set of variables for a series of aggregation or configuration changes, an average can again be used. The proposed measure is:

$$MDM_k = \frac{\sum_{j=1}^{p} MD(l)_{kjb}}{p} \qquad \forall k, (k = 1,\ldots,n) \qquad (7.4)$$

where:
n = number of polygons in overlay map
p = number of aggregations and or configurations evaluated

High values of MDM_k indicate that one or more of the variables has undergone relatively large changes due to the aggregations or configurations imposed, while low values indicate that relatively little change was caused in the variables by making any of the aggregation or configuration changes. Table 7.1 summarizes the proposed measures.

Table 7.1. Summary of Proposed Measures for Aggregation/Configuration Sensitivity

Scenario	Proposed Measure*
Evaluation of the change in one variable with respect to one specific aggregation or configuration	$D(l)_{khi}$: new value - base value, where (l) refers to a specific variable
Evaluation of the change in one variable with respect to a series of aggregations or configurations	DM_k: average of $D(l)_{khi}$ values for a specific variable where each $D(l)_{khi}$ is associated with the same aggregation or configuration
Evaluation of the changes in a set of variables with respect to one aggregation or configuration	$MD(l)_{khi}$: average of all $D(l)_{khi}$ values where each $D(l)_{khi}$ is associated with a unique variable given a specific aggregation or configuration
Evaluation of the changes in a set of variables with respect to a series of aggregations or configurations	MDM_k: average of $MD(l)_{khi}$ values where each $MD(l)_{khi}$ is associated with a set variables for a unique aggregation or configuration

As was the case for the measures of base variable sensitivity described in Chapter 6, it is suggested that the values of $D(l)_{khj}$ be standardized in some fashion prior to the calculation of all following measures. For this research, missing data have been assigned a

value of 0, and the raw data have been converted to Z scores. $D(l)_{kbj}$ values therefore are standardized and all following measures can be calculated directly. An ESDA application should provide for other means of standardizing the data, and for other means of identifying 0 observation population areas and 'nodata' values.

One drawback of these measures is that, with the exception of $D(l)_{kbj}$, the direction of change is not apparent. It is not possible to decipher whether the aggregation value is lower or higher than the base value. This is not seen as a severe problem since this information can be retrieved during the exploration process. A second drawback is created when multiple aggregations, configurations, or variables are measured. A single sensitivity value results, however, it does not indicate which of the aggregations, configurations, or variables is most affected. Again, this is not seen as a major problem, since the information can be retrieved through the exploration process.

Other measures exist which may produce additional information, for example: the range of $D(l)_{kbj}$ values for any given overlay polygon, or the variance associated with a grouping of areal units. While these are not demonstrated in this research, an ESDA application should make available any number of potentially useful descriptive measures.

7.5. USING THE PROPOSED MEASURES

In order to demonstrate the measures and subsequent explorations, a series of aggregations were created using the base dataset already mentioned in Chapter 6, consisting of 470 enumeration areas for the Victoria CMA (census metropolitan area) and four associated socio-economic variables:

A) average value of owned dwellings,

B) average household income,

C) average persons per room, and

D) percent owner one family households.

The enumeration areas were aggregated into 4 levels: 200 and 100 arbitrary areal units, 65 census tracts, and 4 federal electoral districts. As well, three alternate configurations were created at the census tract level: two using the base units as whole entities, and one using a 5 km square grid. With the exception of the grid configuration, the aggregations and configurations were created in Arc/Info by selecting groups of base areal units and then performing a 'merge', until the desired number of new areal units was reached. The grid was created using a short Arc Macro Language (AML) program to generate a square grid; this grid was then intersected with the outline of the study area to create a polygonal coverage based on the grid pattern. The grid boundary system is the only one which splits the base areal units rather than grouping them as whole entities.

Each operation (merge and intersect) resulted in a new map of areal unit boundaries with no attributes associated. An AML program which calculates the appropriate values for the new boundaries using the methods described in Chapter 2 (see page 17) was written by G. Garlick, under the author's supervision. This AML was used to create the appropriate data files for each new map. The calculation of the proposed measures was performed in SPSS prior to the installation of S+ during the later stages of this research.

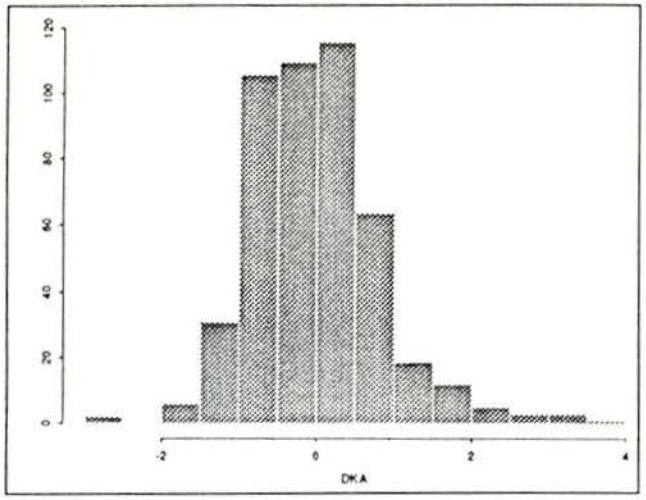
The following demonstrations address four potential questions about sensitivity to aggregation changes. The first question is: where are the areas most affected by a single aggregation or configuration change for a given variable? The second question is: are any areas relatively stable/unstable over a range of aggregation or configuration changes given a particular variable? The third question asks: are there any areas of coincident sensitivity between a set of chosen variables for that aggregation or configuration change? Finally, the fourth question asks: are any areas relatively stable/unstable over a range of aggregation or configuration changes given a set of variables? In all demonstrations, the new values shown were calculated as a simple average of the constituent areal unit values.

7.5.1. Locating Sensitive Areas

Figure 7.4 shows three views which have been generated based on the aggregation from 470 enumeration areas to 65 census tracts. The first is a map of the study area using the census tract boundaries for reference. The second and third are a histogram and a table respectively. The measure $D(l)_{kbj}$, shown for variable A, is noted as DKA. In this example, values for DKA which are greater than 1 have been shaded on the map and histogram. The table lists data associated with the selected areas. The shaded areas identify those areas on the map which have increased in value by more than 1 standard unit, and so may be identified as 'unstable'. The definitions of stable and unstable are subjective; however, an ESDA application should allow for exploration of different stability thresholds simply by reselecting using different threshold values.

7.5.2. Locating Stable and Unstable Areas

Evaluating a single variable over a series of configurations is the basis of this example. Figure 7.5 depicts a map view, a histogram and a table. The measure of interest in this case is DM_k, which gives the average of absolute $D(l)_{kbj}$ values over a series of configurations. In this example, variable A and three configurations at the census tract level are used. One configuration represents the actual census tracts, one is an arbitrary reconfiguration of the enumeration areas as whole units, and the third is a 5 kilometre square grid pattern. Note that DM_k is presented as DMKA. DMKA values greater than 1 have been selected on the graph and shaded on the map view. The boundaries shown on the map are defined by the intersections of the three configurations used in this example. Shaded areas indicate instability for the series of configurations. Unshaded areas indicate relative stability given any configuration in the series.



Ab	Abz	AJ	AJz	DKA
74543	1.71	114347	5.42	3.72
80500	1.97	114347	5.42	3.45
0	-1.69	57402	1.42	3.12
0	-1.69	46731	0.67	2.37
33070	-0.18	67235	2.11	2.30
0	-1.69	40948	0.27	1.97
55436	0.83	70557	2.34	1.51
33075	-0.18	54406	1.21	1.40

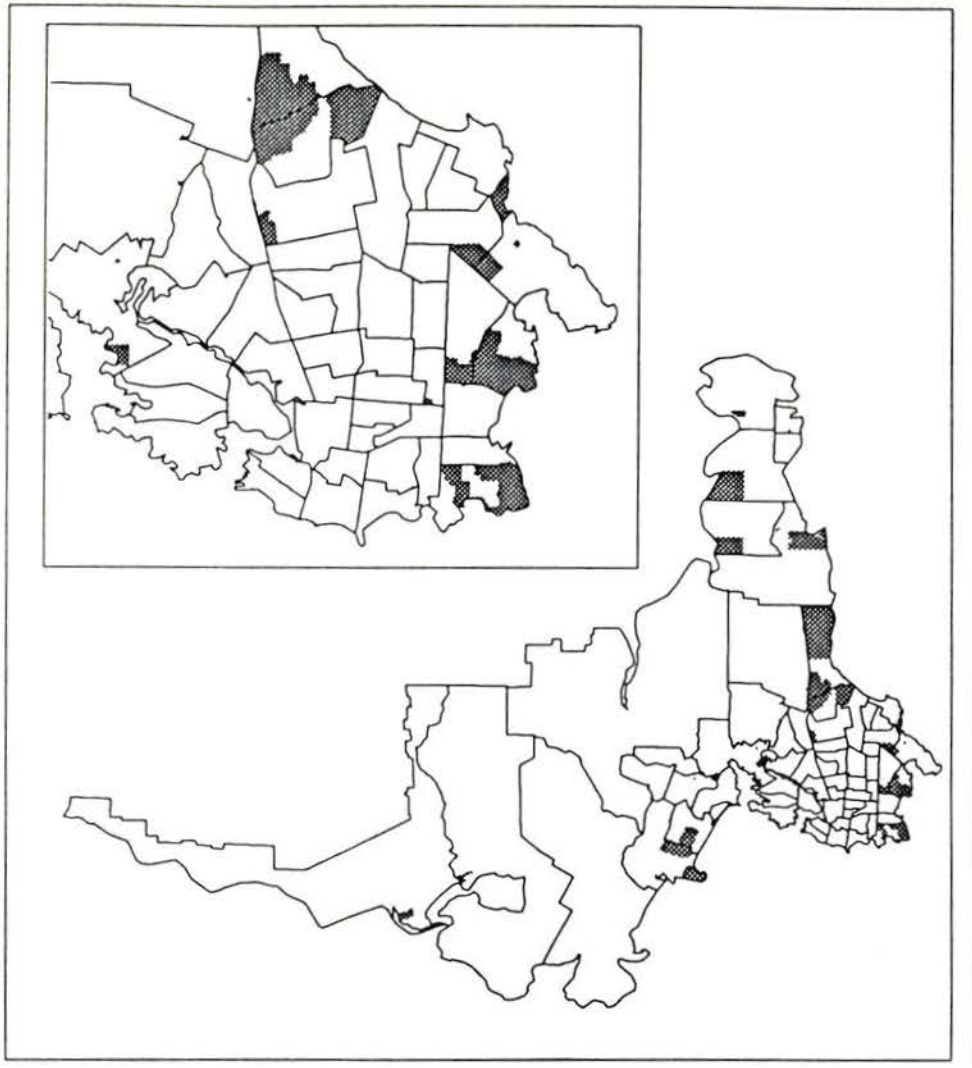
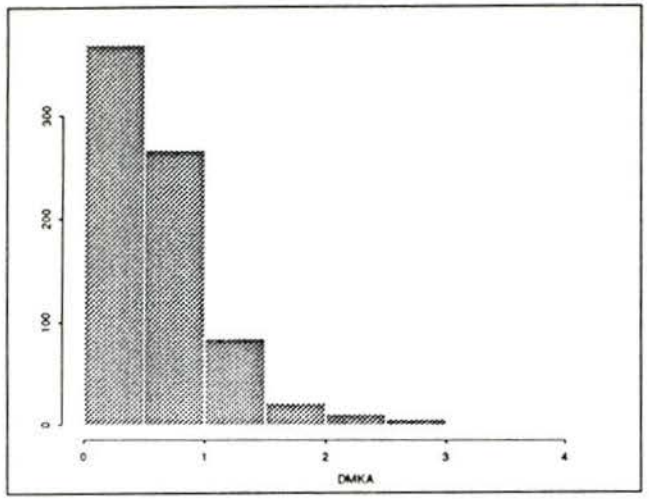


Figure 7.4. D(l)kbj for Variable A - DKA greater than 1



DKA1	DKA2	DKA3	DMKA
-1.49	3.44	-7.15	4.03
3.78	0.89	6.51	3.73
2.73	3.05	3.04	2.94
2.71	2.51	3.34	2.85
-2.58	-2.39	-3.04	2.67
2.75	2.15	3.07	2.66
2.47	2.69	2.80	2.65
2.12	2.79	2.80	2.57

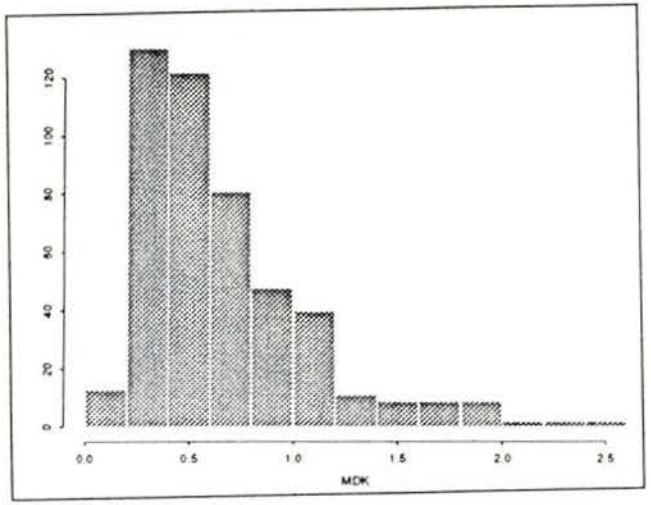
Figure 7.5. DM(l)kbj for Variable A - DMKA greater than 1

7.5.3. Locating Areas of Coincident Sensitivity Between Variables

The three views shown in Figure 7.6 reflect the coincident sensitivity of all four variables to the single aggregation from enumeration areas to census tracts. Census tract boundaries are used as a visual reference. The data represent the measure $MD(l)_{kbj}$, shown as MDK, which indicates the average of absolute $D(l)_{kbj}$ values originating from the four variables used for demonstration. In this example, MDK values greater than 1 have been selected on the histogram and shaded on the map. Again, this is an arbitrary selection, and other thresholds could be used for further exploration. The map view shows clearly those areas which have coincident sensitivity among the four variables given the aggregation to census tract level.

7.5.4. Locating Areas of Coincident Stability and Instability

Figure 7.7 shows three views based on the measure MDM_k, (shown as MDMK), which indicates the average of the $MD(l)_{kbj}$ measures for each of the four variables for the series of configurations. The boundaries shown on the map are defined by the intersections of the three configurations used. As in previous examples, shaded areas on the map view correspond to values of MDMK over a certain value, in this case, greater than 1. The shaded areas indicate the locations where each of the four variables have coincident high $MD(l)_{kbj}$ values for the configuration series. When one or more of the $MD(l)_{kbj}$ values is high, a high MDM_k value results. Conversely, unshaded areas indicate the locations where each of the four variables have coincident low $MD(l)_{kbj}$ values for the series of configurations. The boundaries used for visual reference on the map are defined by the intersections of the three configurations used. These views allow for the identification of areas for further exploration.



DKA	DKB	DKC	DKD	MDK
2.63	3.52	1.61	2.42	2.55
2.65	3.37	1.39	2.77	2.55
2.37	2.56	2.27	1.83	2.26
2.61	3.22	0.69	2.51	2.26
1.92	2.78	1.61	2.50	2.20
1.83	0.93	3.02	2.28	2.02
2.01	1.51	2.46	2.01	2.00
2.05	3.18	1.83	0.79	1.96

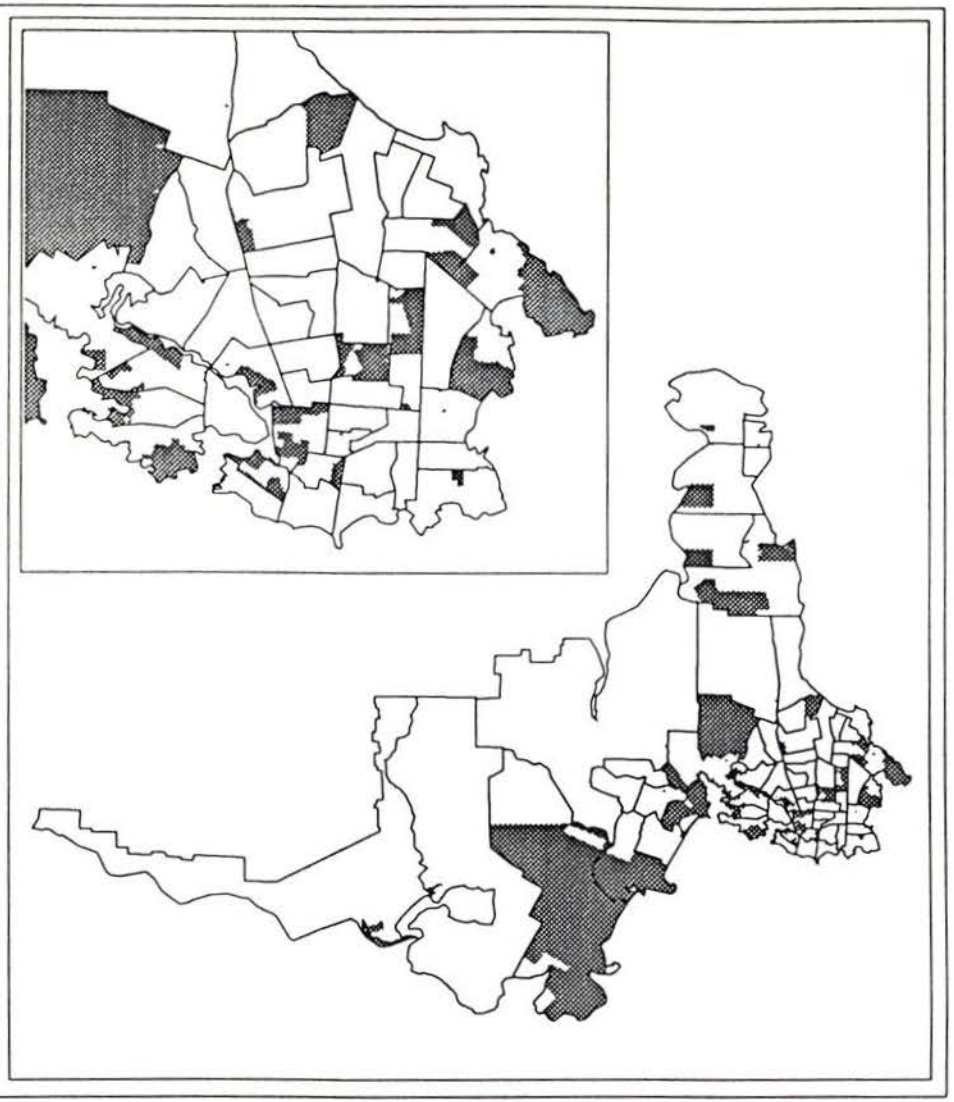
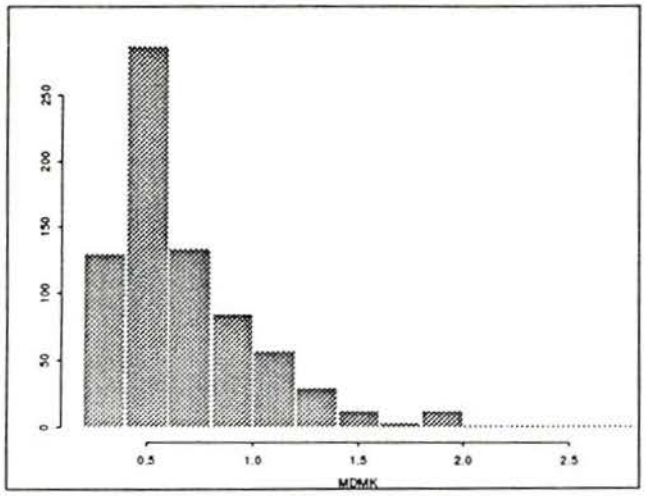


Figure 7.6. MDk for Four Variables - MDK greater than 1



MDKA	MDKB	MDKC	MDKD	MDMK
2.94	3.39	1.97	2.56	2.71
2.66	3.47	1.67	2.80	2.65
2.23	2.87	2.06	2.59	2.44
2.85	2.77	1.67	2.38	2.42
2.65	2.48	2.25	2.02	2.35
2.57	2.18	2.39	1.83	2.24
1.94	1.95	2.61	1.82	2.08
2.03	1.33	2.94	2.00	2.08

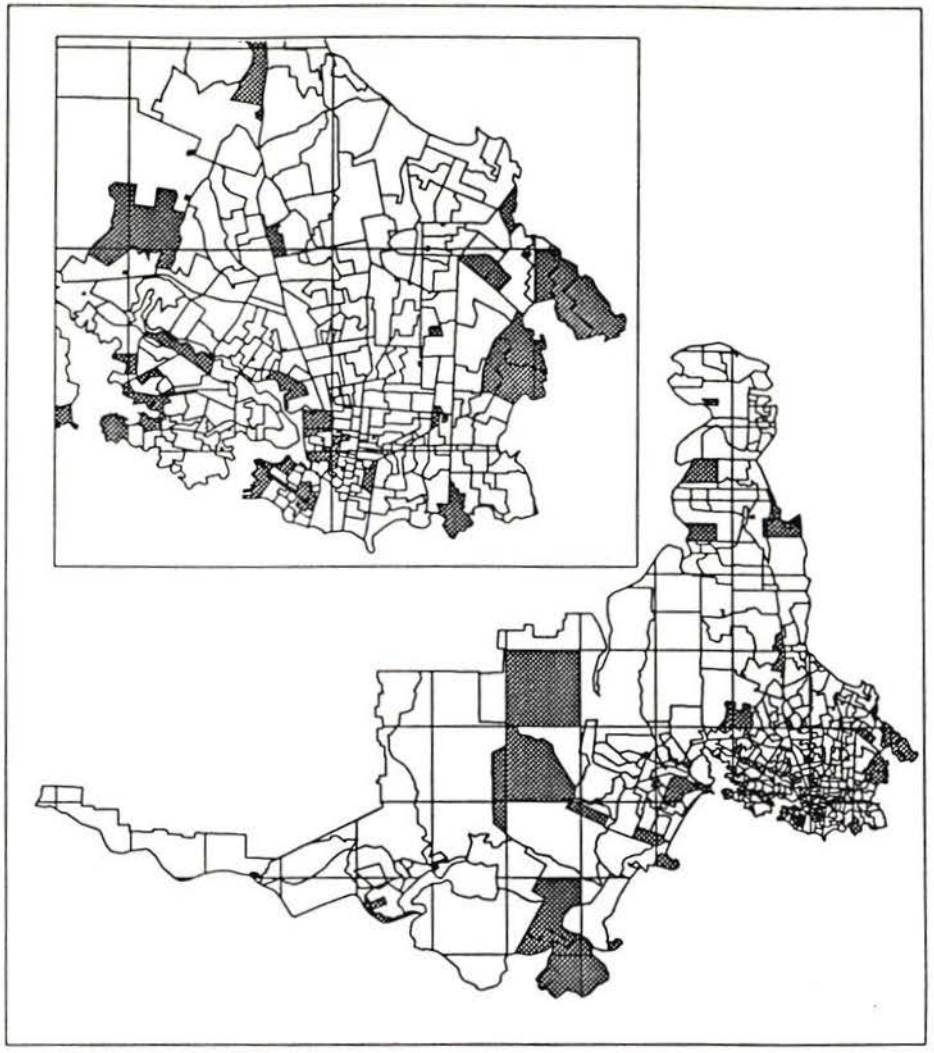


Figure 7.7. MDMk for Four Variables - MDMK greater than 1

7.6. APPLICATION TO RASTER FORMAT AND CATEGORICAL DATA

The application of the measures discussed in this chapter to raster format data is unproblematic. The raster format is particularly well suited to the kind of operations suggested here. Current GIS applications incorporate all the necessary functions for performing mathematical operation between map layers. While some GIS allow for the use of maps with different pixel sizes, others restrict operations to the use of maps with the same pixel size. In this case, it would first be necessary to calculate new values for the larger (aggregated) pixels, and then create a map with the same pixel size as the base map, in order to use the proposed measures. In essence, a group of four base pixels would remain as four separate pixels, however a value based on those of the four pixels would be calculated and would then replace the original values. This results in a four contiguous pixels with identical values.

The application of the proposed measures to categorical data is more problematic, as was the case for the base sensitivity measures, due to the lack of numerical differences. Again, the use of a binary measure may provide some indication of sensitivity to aggregation or configuration changes. A measure could consist of assigning any comparison of two values a 1 if the values are different, or a 0 if the values are the same. When a single variable is evaluated for a single aggregation, a map of 1's and 0's result, with the 1's identifying any areas which changed in value. For a series of aggregations or configurations, a simple sum of the 1's could give an indication of how stable the variable is, for example: if the value changes 3 out of 4 times, the variable is relatively unstable, while if the value never changes, the variable is stable. The real problem arises when any attempt is made to combine values for different variables. There may be no standard for 'different' between variables. For example, the difference between coniferous trees and deciduous trees has no relation to the difference between 200 mm and 50 mm of rainfall in a given area. Any further mathematical analysis, such as summing between a number of

variables would be of questionable validity given the violation of the mathematical assumptions inherent in each level of measurement.

7.7. CHAPTER SUMMARY

The second level of MAUP sensitivity reporting identified in the conceptual framework is that of variable sensitivity to aggregation or configuration changes. Although aggregation and/or configuration changes have been made at this level, no analytical technique has been applied, and so a variable-centred approach is used to develop the concept of sensitivity. At this level, the focus is on *actual* sensitivity and is based on the difference between the original values of the areal units and the new values given the imposed boundary system. Since the actual values can be compared, the method of calculation for the new values does not need to be incorporated into any measure at this level. The concept of sensitivity is straightforward: a sensitive area is one which undergoes a relatively large change under the new boundary system. This indicates the new value is not particularly representative of the original value. There are two exceptions to this definition: data in the form of frequency counts, and point data which are derived from areal units.

Measures based on analysis of variance statistics and entropy statistics have been used to study the differences between aggregation levels, since they reflect the amount of variance at each level. Both methods produce a single value for each of the aggregation levels studied and so are global in nature. A further disadvantage is that these methods have been developed specifically to identify the most informative aggregation level, i.e. the 'best' level for analysis and so do not take the effects of configuration changes into account. Techniques associated with cartographic modeling produce a variety of spatially disaggregate measures which can be applied to this level of reporting. Of particular interest is the 'local difference' measure described by Tomlin(1991). This measure provides the basis for the proposed measures which follow.

Four preliminary measures are proposed. The first, referred to as $D(l)_{kbj}$, is the difference between the new value and the original value. This measure serves to identify areas for which the associated variable undergoes relatively small or relatively high change given a change to a specified aggregation or configuration. The second measure, DM_k refers to the average of the absolute $D(l)_{kbj}$ values for a specific variable for a series of aggregations or configurations. The measure can be used to indicate areas for which a variable is relatively sensitive or insensitive to any of the boundary systems imposed. The third measure is $MD(l)_{kbj}$, and represents the average of the absolute $D(l)_{kbj}$ values for a set of variables given a change to a specified aggregation of configuration. Again, this measure identifies areas for which each of a set of variables remain relatively unchanged or become relatively different given a boundary system change. Finally a fourth measure, MDM_k, is proposed. MDM_k is the average of $MD(l)_{kbj}$ values and can be used to identify areas for which all associated variables are either sensitive or insensitive to any in a series of aggregations and/or configurations.

Although not demonstrated, other descriptive measures which may provide additional useful information could include the range of sensitivity values rather than the average, or could be based on the variance of the constituent raw data values in any one grouping. An ESDA application for exploring and reporting MAUP sensitivity should make available a wide range of descriptive statistics.

The application of the above measures to cardinal level raster data is relatively simple. In fact, the raster format is particularly well suited to performing the type of mathematical operations suggested here. The application to categorical data is more problematic. As was the case for the measures suggested in the previous chapter, these measures are based on the numerical difference between the original values and new values associated with a new boundary system. A binary measure could be used, although issues regarding the meaning of 'different' and the validity of summing measures produced by comparing categorical values become a concern.

CHAPTER 8

RESULT SENSITIVITY

8.1. INTRODUCTION

This chapter is the last of three which extend the conceptual framework of this thesis by developing each of the three sensitivity report levels identified in Chapter 5 more fully. This chapter is specifically concerned with the third level of reporting, identified previously as result sensitivity (see Figure 8.1). The following sections address the concept of sensitivity associated with this level, review existing methods which may contribute to measuring sensitivity, and demonstrate the proposed methods for presenting the analysis results. As in the previous two chapters, the data used for demonstration in this chapter consist of the same four variables and a series of aggregations and configurations created from the base dataset of 470 enumeration areas in the Victoria CMA.

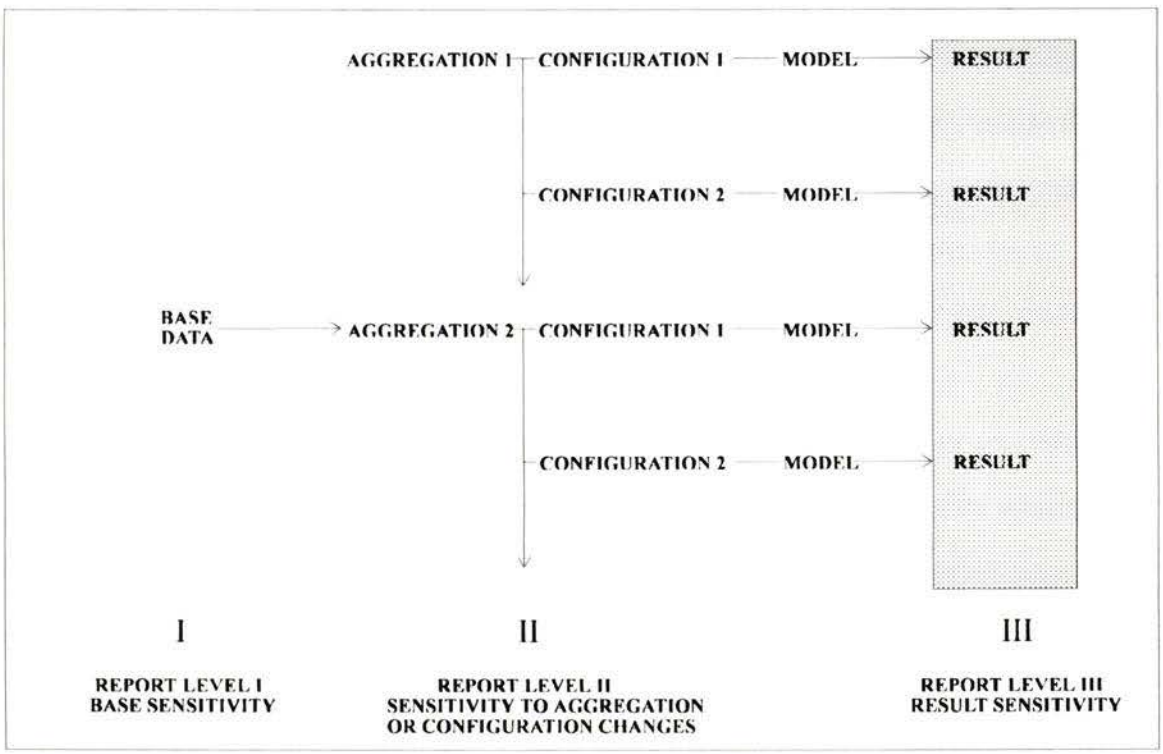


Figure 8.1. Report Level III - Result Sensitivity

8.2. A CONCEPT OF RESULT SENSITIVITY

When a fixed set of areal units and associated variables are used as input for analysis, a unique solution will result. It seems intuitively obvious then that performing the same analysis using a different set of areal units and associated variables will most likely result in a second unique solution. These solutions may or may not be similar. Although each set of areal units and associated variables may be derived from the same original dataset, each can be seen as a different 'expression' of the original, with each expression reflecting the aggregation or configuration characteristics of the areal units and the method used to calculate new summary variable values.

A definition of result sensitivity to MAUP is implicit in a wide range of literature, much of which is concerned with empirically demonstrating the sensitivity of a particular analytical method (for example: Clark and Avery 1976, Openshaw 1984b, Fotheringham and Wong 1991). It is simply this: The result obtained using any particular analytical technique is sensitive to MAUP if a different result can be obtained simply by changing the aggregation or configuration of the areal units. It follows then that high sensitivity would be indicated by the potential for a wide range of results, and that low sensitivity would be indicated by the potential for only a narrow range of results. If the results produced are identical regardless of the aggregation or configuration used, the result has no sensitivity to MAUP.

In the previous two chapters, the developed measures for sensitivity are restricted to cardinal level data and do not include either frequency counts or point data which have been derived from areal units. These restrictions are not necessary at this level, since the object of interest is the analytical result. While many analytical techniques are restrictive in terms of data requirements, there is a broad range of techniques available which use a broad range of data types, measurement levels and formats.

8.3. EXISTING MEASURES

The sensitivity of any particular analytical technique to MAUP can be demonstrated by repeating a specific analysis using a variety of aggregations and/or configurations of the original dataset. This process produces a set of results which can be presented in a number of ways. Openshaw (1984b) uses simple descriptive measures such as the mean and the standard deviation to summarize the values of the correlation coefficient between two variables for 10,000 random configurations and aggregations of the original data. An early attempt by Thomson and Anderson (1965) to use inferential tests to make further analysis of the results was based on defining each aggregation or configuration as a random sample of some universal population. Both Summerfield (1983) and Larson (1986) make convincing arguments concerning the invalidity of this approach when using areal units.

The approach of using common descriptive summary measures is felt to have great utility for reporting result sensitivity to MAUP, however, not all analytical methods produce a single value as a result. For example, a number of locations may be identified using location/allocation methods, as well as associated summary statistics of locational efficiency. In this case, a numerical summary or description of the range of results is possible; however, a visual summary could be more useful, and could be accomplished by viewing maps of the results concurrently or in combination through overlays. This research takes the approach that the most important information at this level consists of the results obtained and the corresponding aggregation or configuration system which generated the result. This information can be easily provided in a tabular format, simply by listing the result and the name of the associated aggregation or configuration. When the number of results is high enough to make a table too large to easily comprehend, a histogram of the results could provide for a simpler summary presentation.

8.4. DOCUMENTING RESULT SENSITIVITY

For demonstration purposes, the following arbitrary multiple regression model is employed using the same four variables used in the previous two chapters:

$$B = f\{A, C, D\}$$

where:

B = average value of owned dwellings - dependent

A = average income - independent

C = average persons per room - independent

D = percent of owner one family households - independent

A number of aggregations and configurations have been created and are used to provide data for the model. No claims are made here as to the conceptual soundness of the model as it is purely illustrative. The model results are not intended to contribute in any way to the understanding of the relationships and interactions of the variables, nor are they used to make any generalizations or predictions about the chosen variables.

Table 8.1 presents the r^2 obtained for both single and multiple regression analyses using data from a series of aggregations and configurations. The table can be used as an index during exploration. For example, selecting an entry in the table would generate a map view as specified, and perhaps a second table with more extensive result information. If the purpose of the analysis is to compare results at this level, the information provided allows for a quick evaluation of the model performance given the potential boundary systems. This information allows the analyst to identify areas for further exploration.

Table 8.1. Single and Multiple Regression r^2 for a Series of Boundary Systems

NAME	multiple r^2	A r^2	C r^2	D r^2
1. 480 EA Areas	.49	.66	.24	.48
2. 200 Areas	.72	.67	.19	.59
3. 100 Areas - Conf. 1	.75	.66	.21	.68
4. 100 Areas - Conf. 2	.79	.73	.21	.58
5. 100 Areas - Conf. 3	.77	.75	.33	.55
6. 65 CT Areas - Conf. 1	.80	.78	.03	.52
7. 65 Areas - Conf. 2	.86	.78	.15	.66
8. 65 Areas - Conf. 3	.84	.79	.07	.59
9. 65 Areas - Conf. 4 grid	.84	.78	.01	.50
10. 19 CSD Areas - Conf.1	.93	.84	.41	.81
11. 19 CST Areas - Conf. 2	.90	.87	.73	.84
12. 4 FED Areas	.99	.97	.85	.67

EA = Enumeration Areas CT = Census Tracts CSD = Census Subdivisions
CST = Census Subdivision Types FED = Federal Electoral Districts

When the number of results is large enough to make the use of a table more difficult, a simple histogram may offer a better method for presenting the range of a particular model parameter or result. Figure 8.2 shows a histogram of the multiple regression r^2 values obtained for each boundary system used above. This example is not particularly powerful; however, when a large number of results are produced, the histogram view may be the best way of initially presenting those results. The histogram view can be used in conjunction with the table to identify specific results and the aggregation or configuration that produced the result. Although the histogram presents the results in a categorized fashion, the ESDA technique of brushing would allow for selection and identification of any particular result value or range of values.

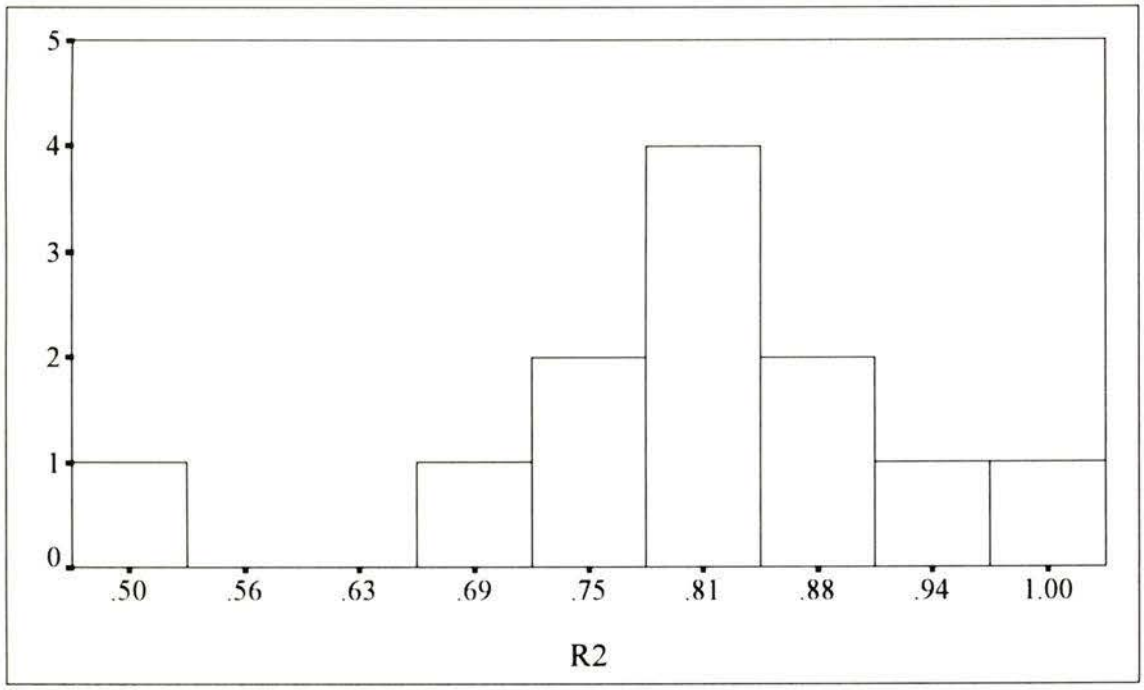


Figure 8.2. Histogram of r^2 for multiple regression results

A third presentation option is used when the results refer to spatial locations. Although not demonstrated here, the result for each aggregation or configuration can be mapped using different symbols or perhaps colours. The specific aggregation or configuration label could be encoded as an attribute of each spatial location in the result and thereby be available through a simple query to the analyst. In this way, further exploration can be identified and pursued.

8.5. CHAPTER SUMMARY

This chapter is concerned with the sensitivity of analysis results to MAUP. Any analytical technique and any type of areal data can be tested for sensitivity at this level of reporting. The concept of sensitivity is straightforward: analyses for which a wide range of results can be obtained, simply by changing the aggregation or configuration of the areal

units, are sensitive to MAUP. Conversely, analyses which produce the same result regardless of the boundary system used are not sensitive to MAUP.

The most common approach to measuring sensitivity is to report the range of results produced, rather than any quantitative measure of the results themselves. The reporting approach is adopted for this research. The report can function as a starting point for exploration. Linking the table entries to the associated views of the sensitivity measures could allow for the application of the selection and identification techniques of ESDA. When a large number of data cases exist, the use of a table for reporting becomes less effective. In this situation, a histogram view may be more easily comprehended, and the ESDA technique of brushing could be used for exploration. Rather than be used as a single view, the histogram view should be linked to the table for easy indexing of the results. Not demonstrated here is the map view which should be used when the results of any analytical technique are identified with specific spatial locations, as would be the case in a location/allocation analysis. Result information could be encoded as attributes of the solution locations, and be accessed through simple querying functions common in GIS.

CHAPTER 9

CREATING MAUP SENSITIVITY REPORTS

9.1. INTRODUCTION

Two kinds of sensitivity reports were identified in Chapter 5 - private and public. The objective of this chapter is to demonstrate, through the use of Arc/Info and S+, the generation of each report type using the proposed measures and ESDA techniques. Section 9.2 presents a review of the proposed measures and applicable ESDA techniques. Section 9.3 gives a particular scenario and describes an exploratory process used for generating the private report given the stated scenario. The results produced and the characteristics of the private report are summarized. Section 9.4 describes the public report which would accompany any published results derived from the scenario given in Section 9.3 and summarizes the characteristics of the private report. Since this chapter focuses on demonstrating the exploration and reporting processes, any strengths and/or weaknesses of the technical environment which are identified through the demonstration process will be discussed in Chapter 10.

9.2 A REVIEW OF THE PROPOSED MEASURES AND ESDA TECHNIQUES

In Chapter 5, three levels for reporting MAUP sensitivity were identified. Report level I addressed base variable sensitivity, report level II addressed variable sensitivity to aggregation and/or configuration changes, and report level III addressed the sensitivity of analytical results to MAUP. In Chapters 6, 7, and 8, potential measures and methods for presenting the sensitivity measurement results were introduced. Although each report level was discussed separately, it is envisioned that any given private or public report would make use of the proposed measures from all report levels as applicable. Table 9.1 summarizes the proposed measures; Table 9.2 summarizes those ESDA techniques which appear to be most applicable.

Table 9.1. Summary of Proposed Sensitivity Measures

Report Level I	**Sensitivity of Base Variables**
SB_l	weighted average difference between a target areal unit and all selected neighbours given a specific variable
SC_l	average of SB_l values for a set of variables associated with the target areal unit
Report Level II	**Sensitivity of Base Variables to Aggregation or Configuration Changes**
$D(l)_{kih}$	the difference between the values of variable(l) for a base dataset and an aggregation or reconfiguration
DM_k	the average of $D(l)_{kih}$ values of variable(l) for a set of aggregations or configurations
$MD(l)_{kih}$	the average of $D(l)_{kih}$ values for a set of (l) variables given one aggregation or configuration
MDM_k	the average of $MD(l)_{kih}$ values for a set of (l) variables and a number of aggregations or configurations
Report Level III	**Result Sensitivity**
N/A	no specific measures proposed

Table 9.2. Summary of ESDA Techniques

Views	maps, histograms, scatterplots, tables
Statistics	descriptive and/or spatially disaggregate, as proposed
Interaction	brushing, selection, highlighting, identification, zoom in/out,

9.3. GENERATION OF A PRIVATE REPORT

The private report is generated solely for the analyst; the process is iterative and meant to allow for flexibility in exploring the data and the effects of MAUP. The end product is increased awareness rather than a formal static report. The process followed by any analyst will differ according to the data, the interest of the analyst, the type of analysis planned and the context of the analysis. The following scenario and description of one possible exploration process serve to illustrate the potential of using the proposed sensitivity measures and ESDA techniques for private reporting of MAUP effects.

Assume that an analyst wishes to perform the following multiple regression analysis at a census tract level:

$$B = f\{A, C, D\}$$

where:

B = average value of owned dwellings - dependent

A = average income - independent

C = average persons per room - independent

D = percent of owner one family households - independent

The raw data are available for enumeration areas, so in order to perform the analysis, the data must first be aggregated to the census tract level. New values are to be calculated using a simple average. The analyst is aware of MAUP and the potential effects on the analysis, and so wishes to explore the behaviour of the variables with respect to MAUP. The analyst begins with an exploration of base variable sensitivity, followed by an exploration of the effects of aggregation to the census tract level. Finally, the results of the multiple regression for a number of census tract configurations are explored. Each of these stages is described in the following sections.

9.3.1 An Exploration of Base Variable Sensitivity

For this first step in the exploration process, the brush function provided by S+ is used. The brush function, when invoked, creates a matrix of scatterplots using all variables in the dataset, histograms for each variable, and a table which lists a numerical label of each data point. When the brush is placed over one or more data points in any scatterplot, or on a histogram, the same points are highlighted in all other scatterplots and the position in each histogram is adjusted to show the positions of the data point(s). Figure 9.1 shows the result of using the brush function on the dataset which consists of the individual SB_f values (SBAN1, SBBN1, SBCN1 and SBDN1), and the combined SC_f values (SCN1). Note that measures are named Var1, Var2, Var3 and Var4 for SBAN1, SBBN1, SBCN1 and SBDN1 respectively, and that SCN1 is named Var 5. These names are generated by S+. A legend has been added to Figure 9.1 for clarity, and the first variable names listed will be used in this discussion. High values of SCN1 have been selected with the brush (located in the lower left scatterplot), and are highlighted in all scatterplots. On the histograms, those bars which fall below the horizontal axis correspond to the highlighted points, and the table highlights those entries which correspond to the selected data points.

The distribution of highlighted points in each scatterplot in the lowest row gives an indication of the relative contribution of each SB_f measure to SCN1. Of interest is the apparent fact that high SCN1 values are not generally made up of consistently high SB_f values for all of the variables. Instead, the highlighted points for each variable seem to be fairly evenly distributed along the horizontal axis of each scatterplot, suggesting that a range of both low and high SB_f values for any given variable contribute to high values of SCN1. For all variables, there exists at least one SB_f measure of greater value than the highlighted points. A preliminary interpretation of this pattern is that high values of SB_f for any given variable are not necessarily spatially coincident with high values of SB_f for any other given variable.

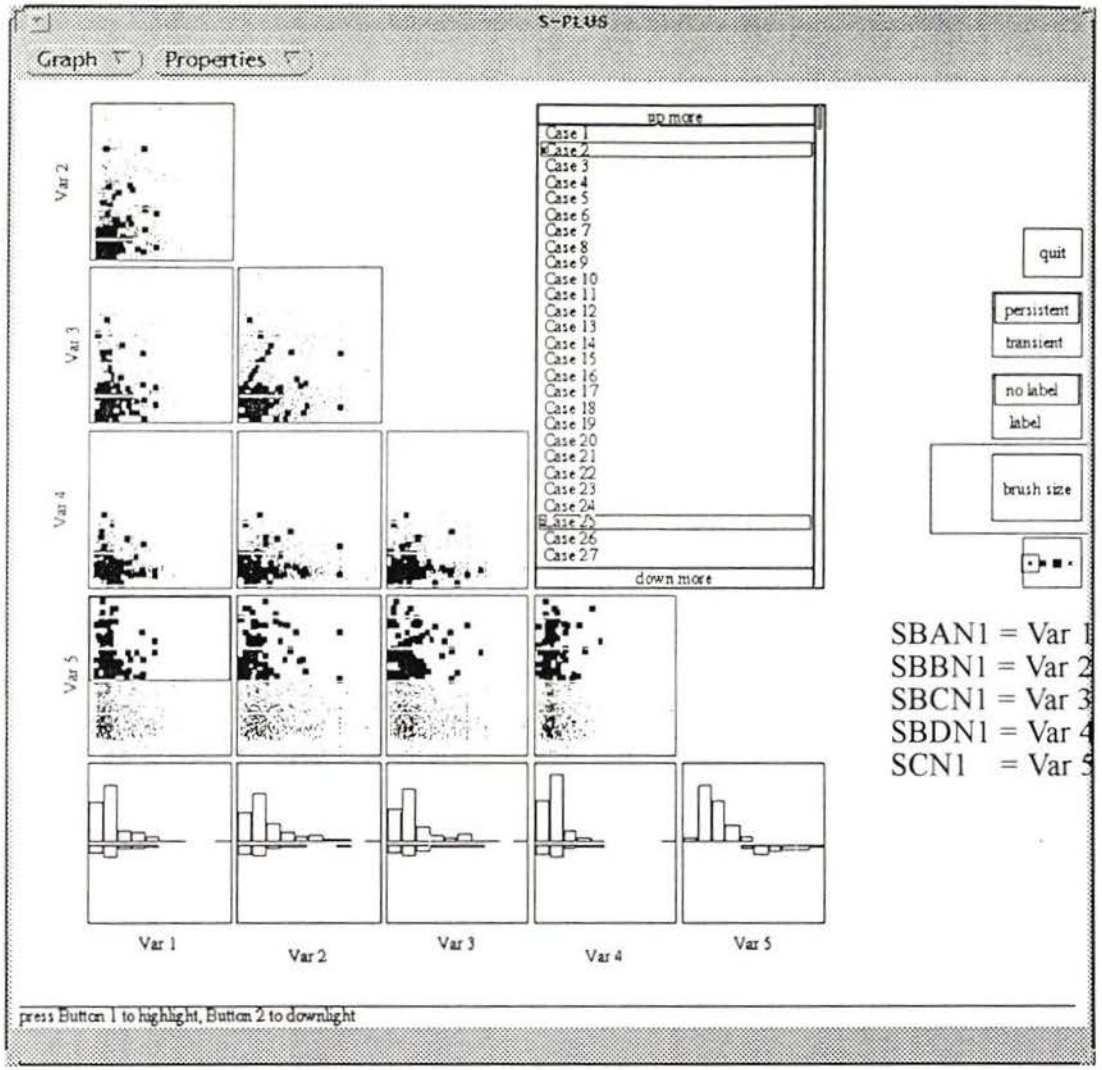


Figure 9.1. Scatterplot Matrix for Base Variables

The histogram views along the bottom of the scatterplot matrix support this interpretation. The position of the highlighted points in the distribution are depicted by the histogram columns which fall below the horizontal axis. For SBAN1, SBBN1, SBCN1 and SBDN1, those portions of the histogram which fall below the axis are relatively evenly distributed; however, in all four cases, the extreme high values remain above the axes and so are identified as not contributing to any of the high SCN1 measures.

The hypothesis that the potential sensitivity measures for each of the variables have unique spatial distributions can be investigated through the use of map views. Figure 9.2 shows map views for SBAN1, SBBN1, SBCN1, and SBDN1. All values greater than 1 have been shaded. This is an arbitrary selection, since the lack of scales on the axes of the scatterplot view make it impossible to select exactly the same numerical range of high values for the map views. It is readily apparent that the distribution of SBAN1 and SBBN1 values greater than 1 appear to be quite similar, while the distribution of SBCN1 and SBDN1 values greater than 1 appear to be dissimilar in comparison to the first two variables.

The map views show that, although the distributions of potential sensitivity appear dissimilar in some cases, there are a number of shaded areas common to all four maps. Table 9.3 lists the original variable values, generated by randomly selecting areas with high coincident potential sensitivity for all variables. The table shows that a number of areal units selected have original values of 0. Either the data are missing, or there are no observation populations associated with the areal units. Since the data were converted to Z scores prior to the calculation of any sensitivity measures, the Z scores for 0 values are the lowest found in the data, and so behave as extremes with high potential sensitivity, as is appropriate. The identification of areal units which do not contain the phenomenon under study, or have missing data allows the analyst to make some decision about the treatment of these areas.

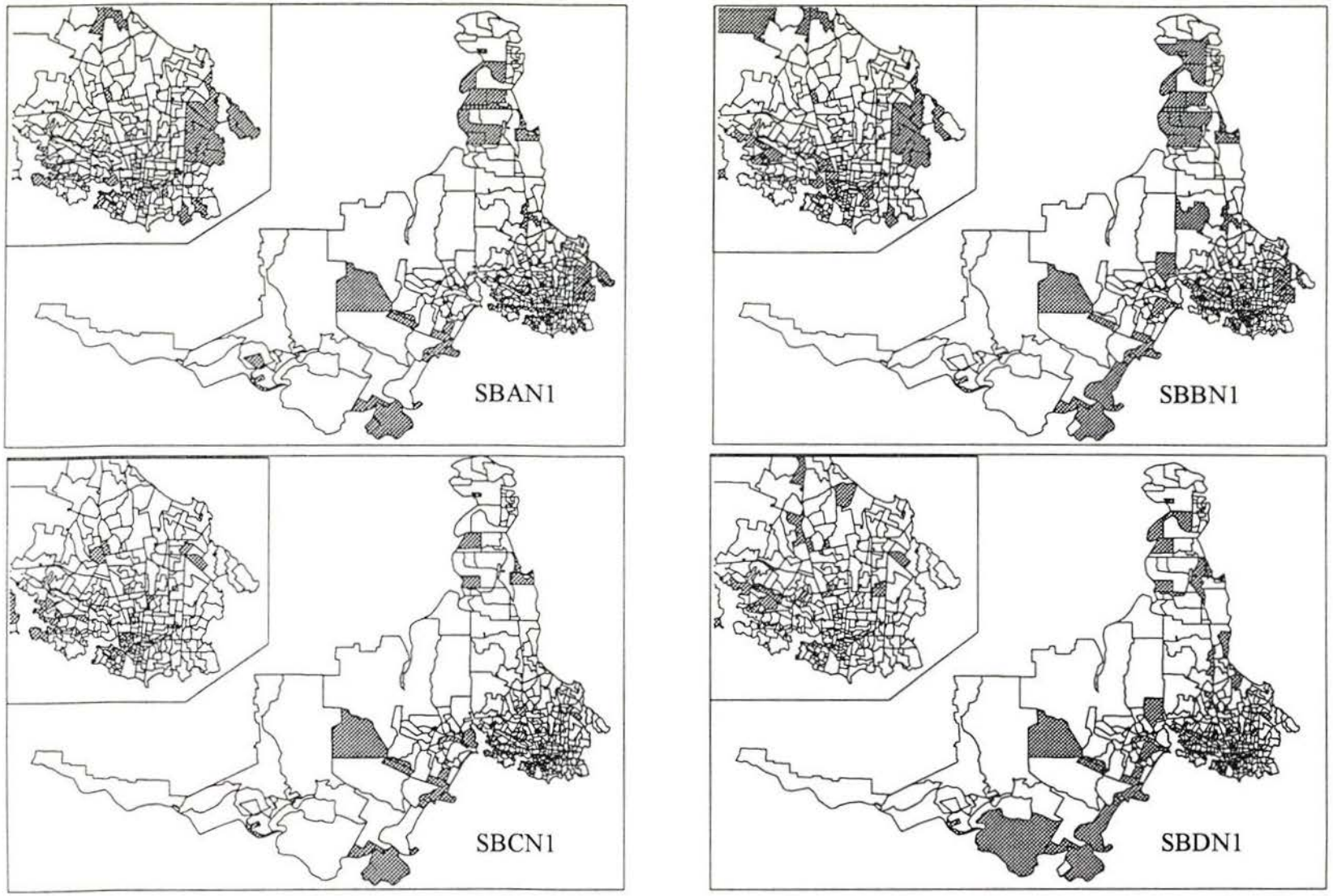


Figure 9.2. Map Views for SBAN1, SBBN1, SBCN1, and SBDN1 - Greater than 1 Shaded

Table 9.3. Original Values for High Potential Sensitivity Areas

A	B	C	D
0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00
86151.00	298701.00	0.30	80.26
0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00
33067.00	444320.00	0.30	80.00
65755.00	284086.00	0.30	75.76
46841.00	199553.00	0.40	64.71
0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00
33071.00	114662.00	0.40	25.53
33071.00	400271.00	0.30	70.59
33080.00	51667.00	0.50	30.00
0.00	0.00	0.00	0.00

With areas of 0 value identified, returning to the map views allows for the identification of 'true' spatial outliers. In this sense, 'true' means that the areal units do indeed have data which are in fact extreme in comparison to the surrounding areal units. Figure 9.3 shows four maps which correspond to the potential sensitivity for each variable for areal units which have non-zero original values. A visual comparison re-confirms the idea that SBAN1 and SBBN1 are similarly spatially distributed, while SBCN1 and SBDN1 differ in spatial distribution. This illustrates an interesting problem. If the analyst wishes to create an optimal (i.e. homogeneous) configuration in order to minimize MAUP effects, it will be necessary to choose between optimizing for variables A and B, or optimizing for either C or D. In any case, the optimal solution for A and B will be decidedly non-optimal for C or D, and vice verse.

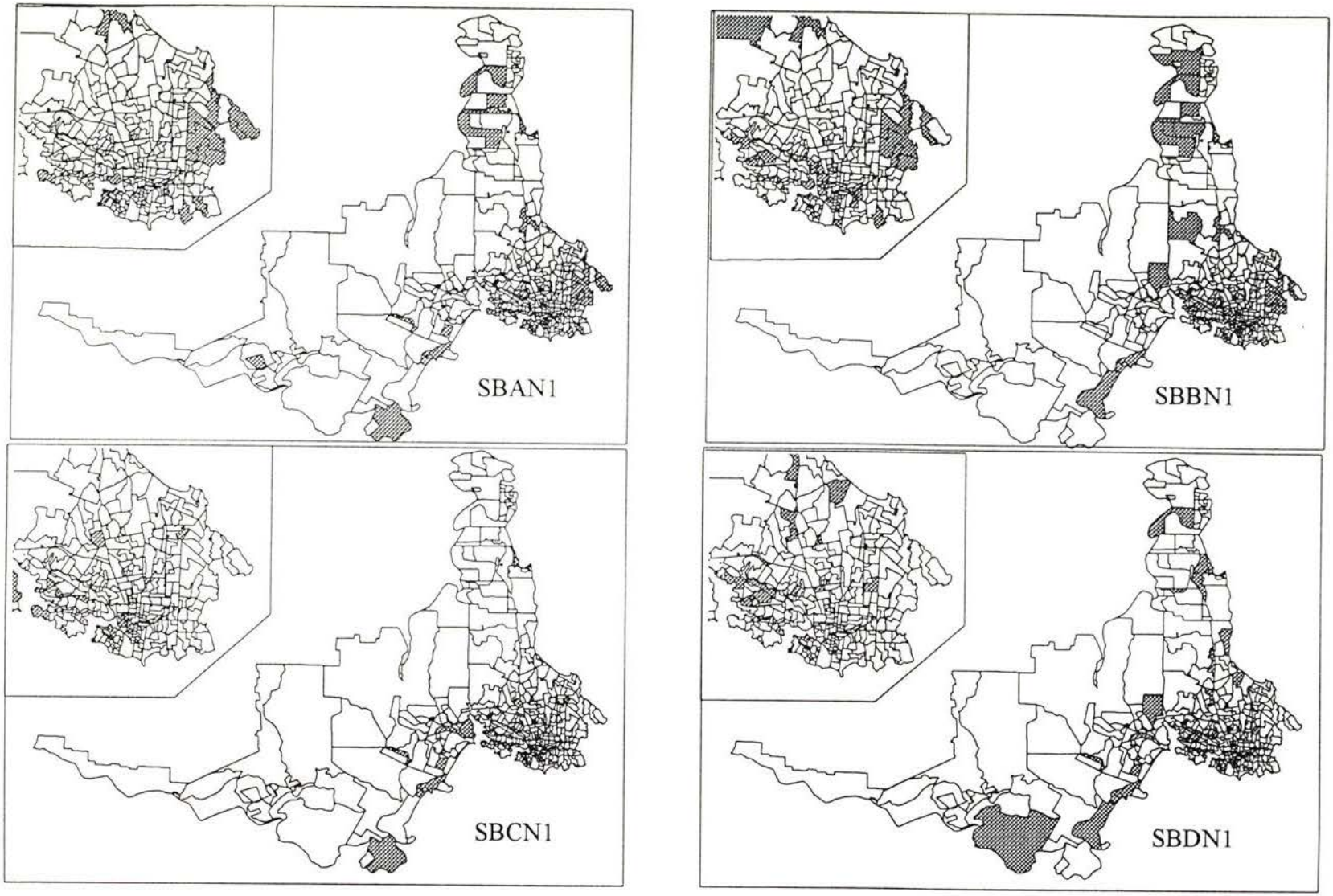


Figure 9.3. Map Views for SBAN1, SBBN1, SBCN1, and SBDN1 - No 0 Values, Greater than 1 Shaded

9.3.2 An Exploration of Variable Sensitivity to Aggregation

Assume at this point that the analyst has chosen to include all areas with no observation populations or with missing data, and has performed the aggregation of 470 enumeration areas to 65 census tracts. A simple average of the enumeration area values has been used to calculate the new values for the census tracts. It is now possible to explore the actual sensitivity of the variables to the aggregation performed.

Figure 9.4 shows the scatterplot matrix generated by the S+ brush function. The measures of interest are $D(l)_{kbj}$, noted as DKA, DKB, DKC and DKD for each variable, and MD_k, noted as MDK. Also, note the addition of a legend of variable names. High values for MDK have been selected using the histogram view and are highlighted in all other views. As was the case for the base sensitivity exploration, the distribution of highlighted points in the scatterplots and the distribution of histogram bars which fall below the horizontal axes show that a range of D(l)kbj for each variable are included in the composite measure MDk.

The top scatterplot in the first row shows that DKA has a fairly linear relationship with DKB. The first scatterplot in the second row suggests that the relationship of DKC to DKB is the least linear. This leads to the conclusion that the actual sensitivity of A and B are spatially distributed in a similar pattern, while the actual sensitivity for C and D are not coincident. Of interest is the pattern shown in the second scatterplot in the first row, the first scatterplot in the second row, and the first scatterplot in the third row. These scatterplots show the relationship between variable C and the other three variables. In each case, the relationships are less linear than those shown for the combinations of the other three variables. It can be surmised then, that out of this set of variables, the actual sensitivity of variable C shows a lack of spatial coincidence with the actual sensitivity of variables A, B, and D.

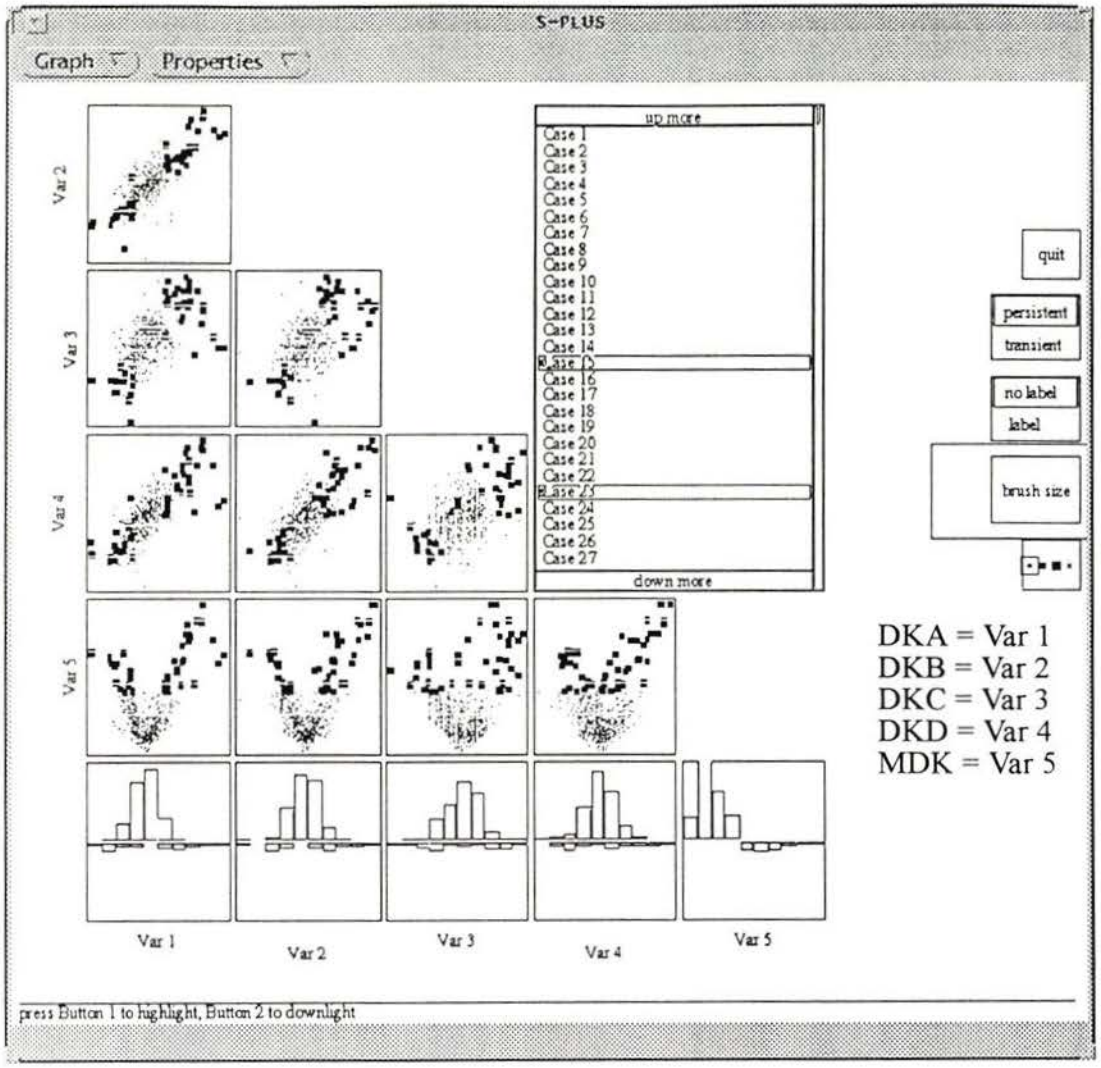


Figure 9.4. Scatterplot Matrix for Actual Sensitivity

Given the information gained through the use of the brush function in S+, maps of the actual sensitivity of each variable can be used to further explore the spatial distribution of actual sensitivity. Figure 9.5 shows a map view of D(l)kbj for each variable. D(l)kbj is noted as DKA, DKB, DKC and DKD for variables A, B, C and D respectively. In each view, values for D(l)kbj which are either less than -1 or greater than +1 have been shaded. These areas indicate where the original variable values have either decreased or increased by 1 or more standard units. Although not shown here, the choice of these values is based on histogram views of D(l)kbj values for each variable.

While there are common areas in all maps, it is again apparent that the distribution of actual sensitivity for C has less in common with the spatial distributions of actual sensitivity for A and B. It also appears that D is less sensitive than the other variables as there are fewer shaded areas. While the sensitivity within the shaded areas may be high, it can generally be stated that there are fewer regions of actual sensitivity for D, given the aggregation to census tracts.

9.3.3 An Exploration of Result Sensitivity to Configuration

Since the analyst is aware of MAUP and the associated effects, the analyst is interested in the robustness of the multiple regression results given alternative configurations of the census tracts. Three different configurations are created and the associated data are used for each respective multiple regression analysis. Figure 9.6 shows each of the four configurations, and Table 9.4 presents selected parameters from the results of the multiple regression analyses.

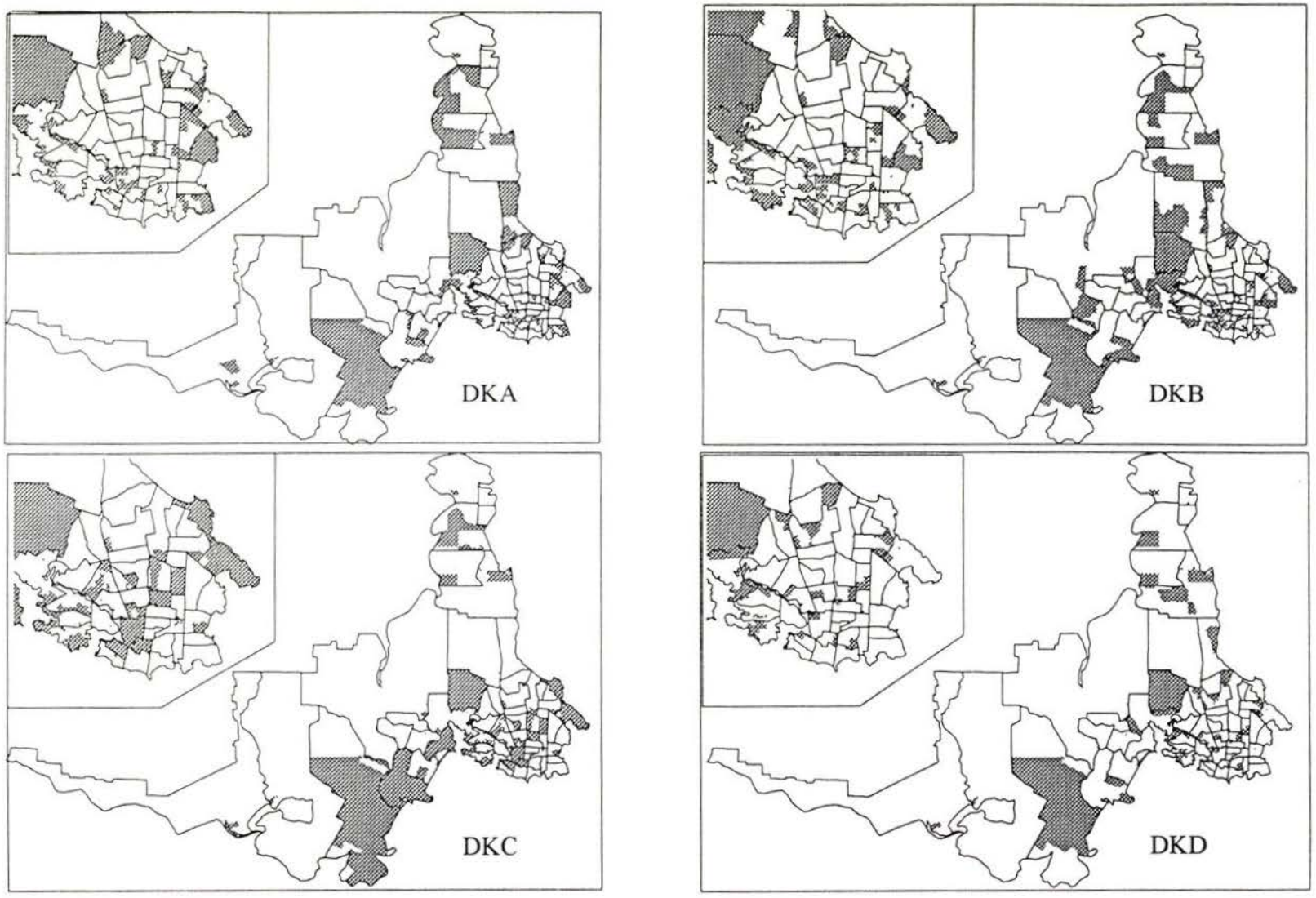


Figure 9.5. Map Views for DKA, DKB, DKC, and DKD - Values < -1 or > 1 Shaded

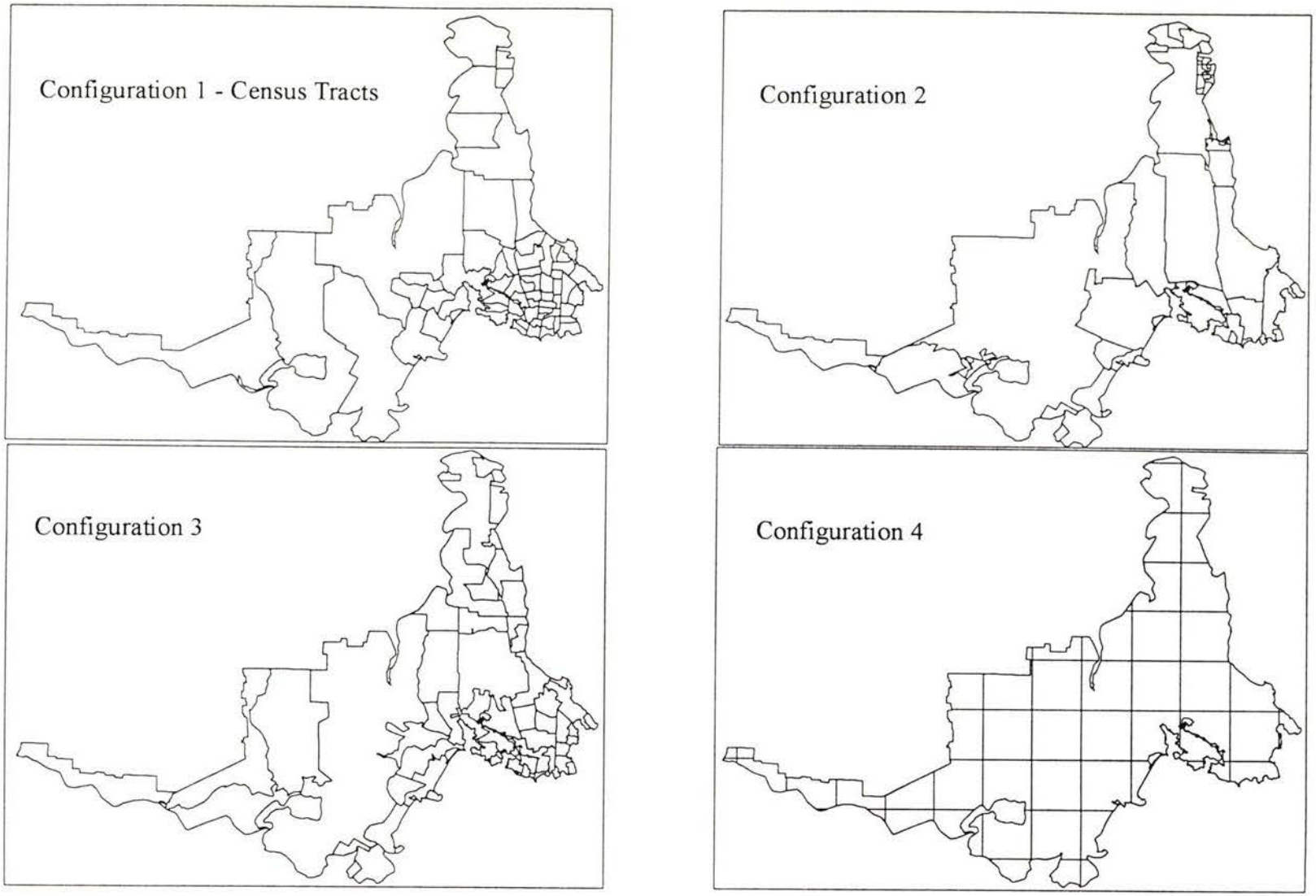


Figure 9.6. Four Configurations of 65 Areal Units

Table 9.4. Selected Parameters of Four Multiple Regression Analyses

Configuration Name	r^2	Correlation Coefficient			Significant T		
		A	C	D	A	C	D
Configuration 1 - CT*	.80	.885	.184	.723	.0000	.0531	.4816
Configuration 2	.86	.883	.388	.810	.0000	.0001	.0000
Configuration 3	.84	.889	.271	.767	.0000	.3264	.0000
Configuration 4	.84	.885	.088	.232	.0000	.0002	.0004

*CT = Actual Census Tracts

The best model performance is associated with Configuration 2, an arbitrary reconfiguration of the actual census tracts. The r^2 is highest and all three independent variables are significant contributors. Configuration 4, 3, and 1 perform increasingly poorly respectively, based on the parameters shown here. Of interest is the poor performance of model when the actual census tracts and the similar reconfiguration are used. Also of interest are the correlation coefficients for each variable. Note that the correlation coefficient for C is highest for Configuration 2 and lowest for Configuration 4. This is also true of the correlation coefficient for D. Previously it was suggested that A and B behave in a similar fashion in terms of potential sensitivity, while C and D are less similar. The correlation coefficients also suggest that C and D may be contributing in some way to the differences observed in model performance. In fact, C was identified as having the most dissimilar spatial distribution of both potential and actual sensitivity, while it was suggested that D may be less sensitive in general. The next step in the exploration might be to investigate the behaviour of C more fully.

Figure 9.7 shows each of the configurations with DKC values less than -1 or greater than 1 shaded. Again, it is apparent that Configurations 2 and 4 have caused changes in C in different locations when compared with the changes made given Configurations 1 and 3. Of interest is the amount of area shaded for Configuration 4.

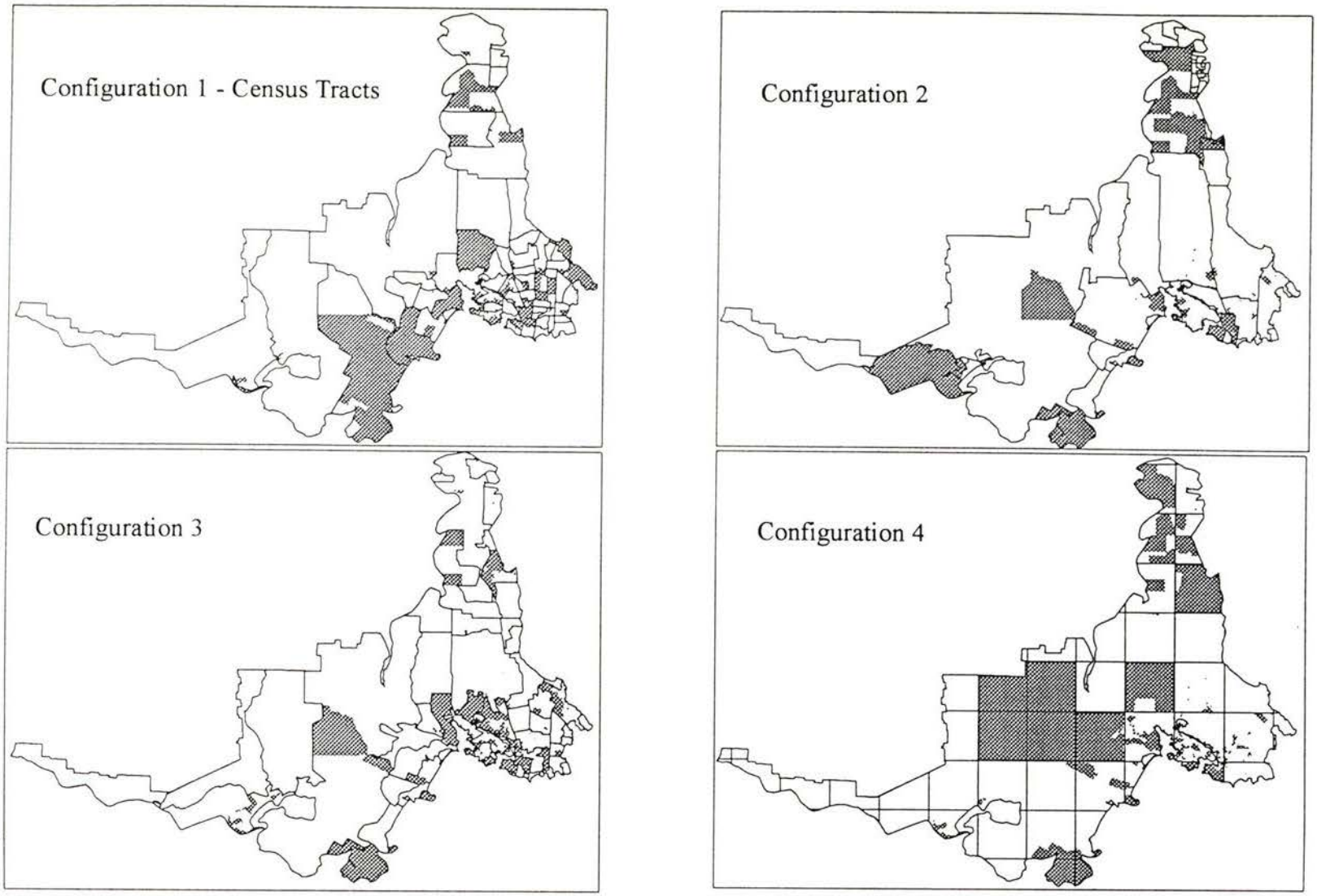


Figure 9.7. DKC for Four Configurations - Values < -1 or > 1 Shaded

A much larger area is shown as being relatively 'unstable' when compared to the other three views. Since the analyst has included all 0 value areas in the aggregation, a quick look at how the areas of 0 value for C have been affected by the configurations may be identify whether the inclusion of those areas may be responsible in some part. Figure 9.8 shows each configuration with all 0 value areas for C that have changed by either +1 or -1 standard unit shaded. In fact, the inclusion of 0 values appears to cause changes of more than +1 or -1 standard units in relatively few areas. Instead, those areas which have positive data values for C appear to be more important in terms of actual sensitivity.

The use of some simple descriptive measures may be useful at this point to further explore the patterns shown. Table 9.5 lists the coefficient of variation for the data associated with each configuration, as well as the coefficient of variation for the original base data. The values for Configurations 2 and 4 are shown in bold type.

Table 9.5. Coefficient of Variation for Each Variable for Five Datasets

Configuration Name	A	B	C	D
Enumeration Areas	.58	.64	.46	.82
Configuration- CT*	.42	.43	.27	.54
Configuration 2	**.65H**	**.80H**	**.53H**	**.74**
Configuration 3	.62	.60	.36	.59
Configuration 4	**.40L**	**.49**	**.25L**	**.42L**

*CT = Actual Census Tracts **H** = highest value in column **L** = lowest value in column

With the exception of D, Configuration 2 has maximized the variation of the variables. Conversely, with the exception of B, Configuration 4 has minimized the variation of the variables. This presents an interesting paradox: the best model results are associated with both the maximization and the minimization of variation in the dataset.

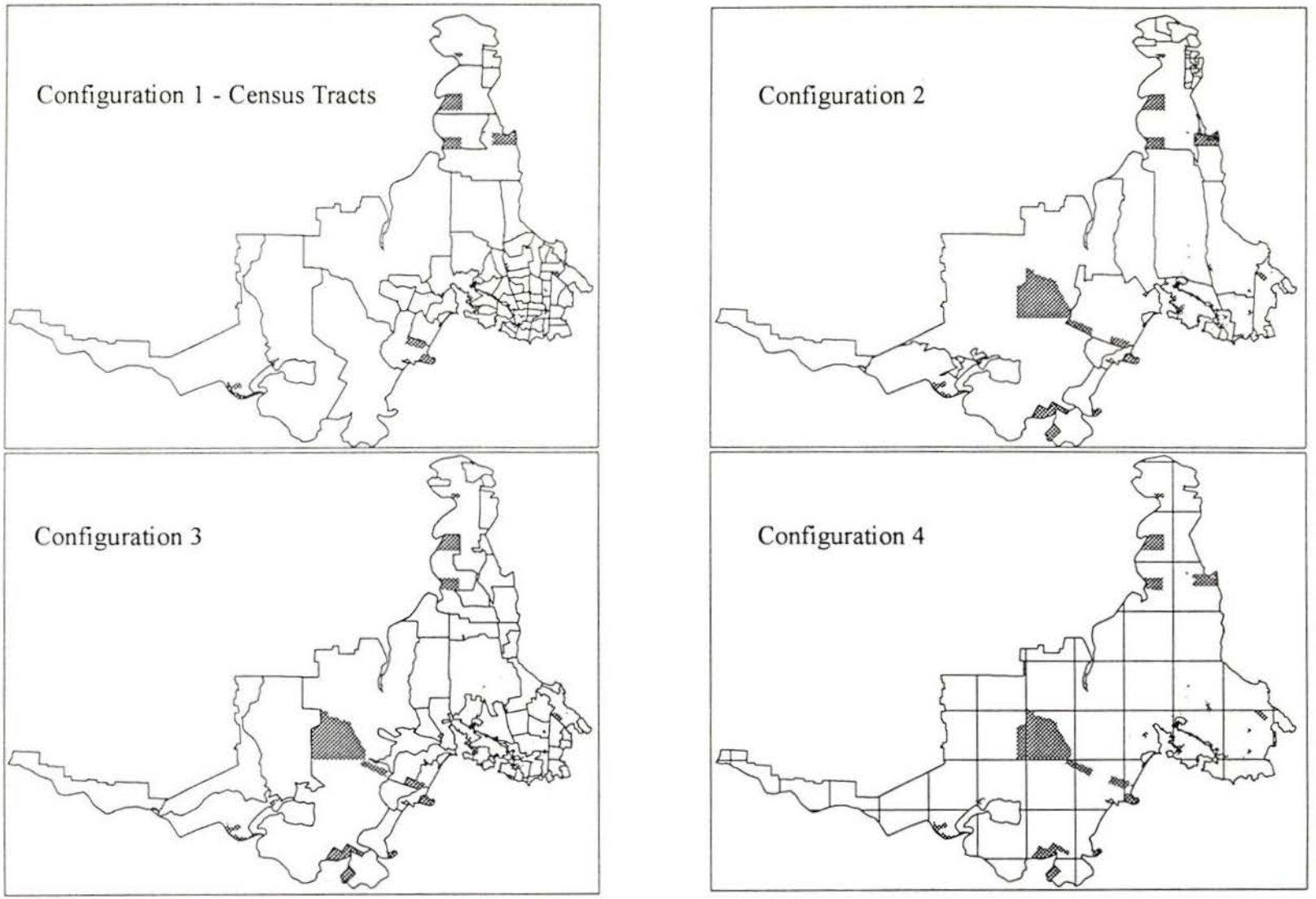


Figure 9.8. DKC for Four Configurations - Areas with 0 Original Value and DKC Values < -1 or > 1 Shaded

Further exploration may uncover additional information; however, the demonstration will conclude at this point and the following discussion will focus on the exploration process used for the demonstration. Figure 9.9 depicts the demonstrated process and summarizes the questions and information generated at each point.

It is difficult to convey in writing the actual interactiveness of this process since the description uses a linear path and shows views separately in the interest of clarity. It is not suggested here that this is the only sequence for exploration, nor is it suggested that these are the only views which are useful. Different users would have different reasons for using such an application and as mentioned previously, the exploration process would be guided by the type of analysis, the potential use of the analysis results, and on the interest of the user themselves.

A number of ESDA techniques have been applied in this demonstration. Statistical techniques include the use of the proposed measures which identify both data and spatial outliers/anomalies and summarize univariate measures when multivariate comparisons are required. Also introduced was the use of existing descriptive measures, i.e. correlation coefficients and coefficients of variation. Data views include tables, maps, scatterplots, and histograms; interactive techniques include brushing, highlighting, selection, and identification. Although each view has been presented separately, these views actually co-exist on the display screen at the same time. The exploration process can be relatively quick when it is not necessary to generate each view separately, and regenerate the same view later in the process. The impact of seeing all views simultaneously cannot be underestimated. Each view provides pieces of information which, when seen together, allow for potentially easier interpretation of the individual views.

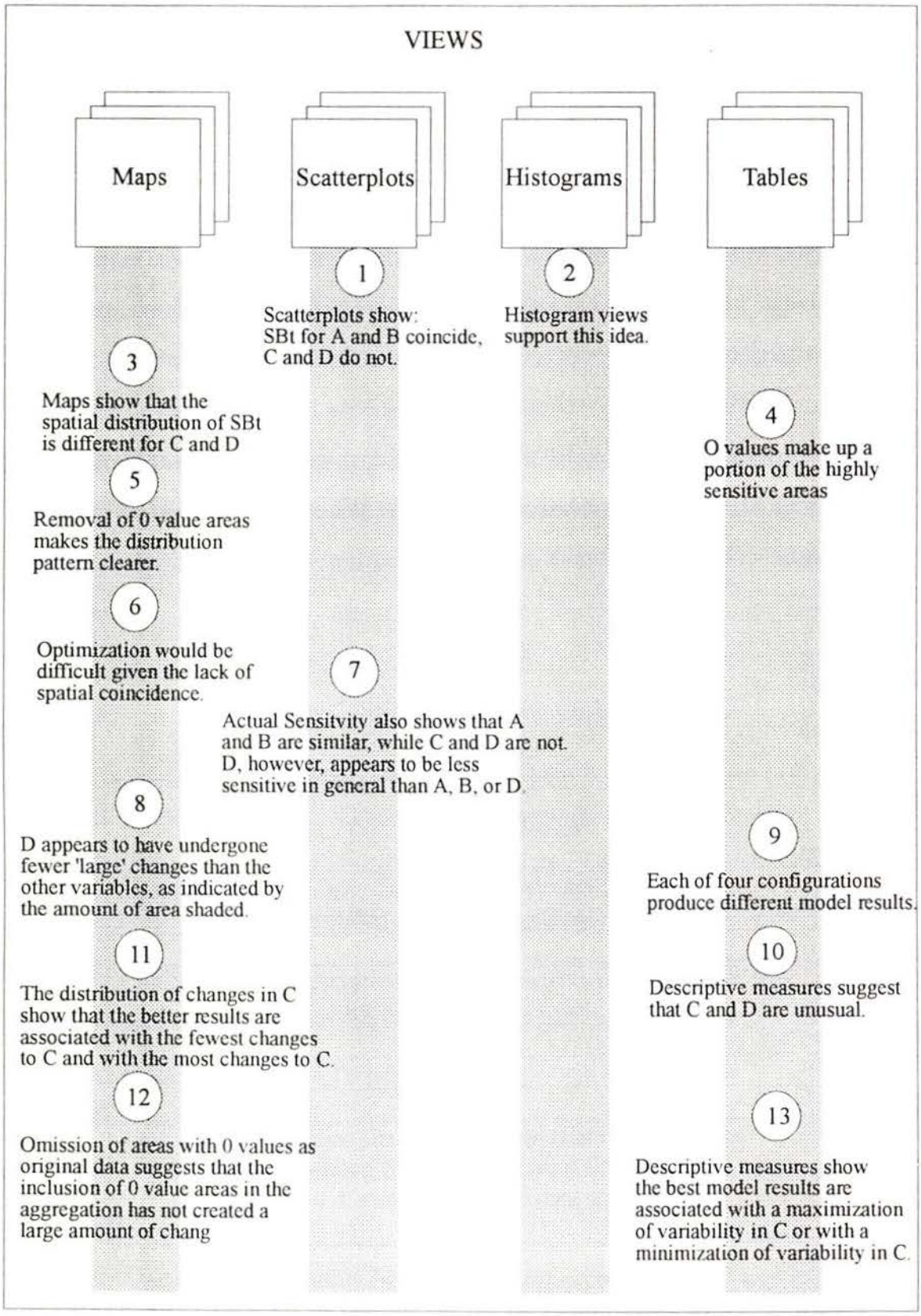


Figure 9.9. The Exploration Process

9.4. GENERATION OF A PUBLIC REPORT

The public report is generated to accompany any analysis using areal data. The report documents any aggregation or configuration changes made by the analyst, and provides an indication of the result sensitivity using measures and views of the data as applicable. A reader can then use the information provided to make some subjective assessment of the effects of MAUP on the presented analysis. When an ESDA application is available, the analysts need only provide the datasets and the associated boundary maps; however, when an ESDA application is not available, the report should be available in a static form.

Given the same scenario described in Section 9.3, assume that the analyst wishes to publish the results of the multiple regression analysis using the actual census tract boundaries. Table 9.6 lists the report elements which should be included in the static public report.

Table 9.6. Public Report Elements Required

REPORT ELEMENTS
Actual Sensitivity Maps: Choropleth representations for each variable transferred and all associated histograms
Coincident Actual Sensitivity Map: Choropleth representation and the associated histogram
Model or Analytical Technique Results
Text Report: including the method for calculating new values, the method of handling missing data, and the method of standardizing the data scales if used.

Note that the sensitivity maps are in choropleth form. This is to accommodate the static report format. Histograms of the values for the same classes used to create the choropleth map must be included in order to allow for an independent evaluation. The potential sensitivity is not listed as a required element, since the actual sensitivity is more useful given the analysts choice, prior to generating the report of a particular aggregation or configuration. It is suggested here that, for the static public report, the actual sensitivity provides more useful information than does the potential sensitivity.

Perhaps the most important feature of Table 9.6 is that the listed elements are required when only one aggregation or configuration change is considered. The same report element will be required for every aggregation or configuration evaluated in terms of result sensitivity. In the previous demonstration, four configurations were evaluated. If all the report elements listed were included in a static public report, 20 maps and 20 histograms would be required, given the use of four variables and four configurations.

It quickly becomes apparent that a comprehensive static public sensitivity report may be difficult to accomplish and may actually be longer than the published research. Several options exist for managing the static public report. The first is to allow the analyst to choose those report elements for the report. This option essentially defeats the purpose of producing the report. The information included in Table 9.6 allows readers to make their own assessment of MAUP sensitivity without interference from the analyst. The inclusion of histograms with each of the maps makes the classification system used explicit, and the reporting of each boundary system used to test result sensitivity allows for independent evaluation of the choice of alternate boundary systems.

A second option exists in the form of a disclaimer. Published analyses which use areal data can be explicitly identified as such and a brief text report of sensitivity analysis results can be substituted for the full report. However, details for accessing the full report would be made available to the readers. The full report might reside with the publishers of

the research or with the researcher themselves, and could be made available in digital format on the Internet.

The potentially serious burden of producing a static public report may be the root cause of the general lack of such sensitivity reporting in the past and current literature. Certainly the optimal solution is to allow any interested reader access to an ESDA application for exploring MAUP sensitivity, thereby allowing them to generate their own private reports. The analyst need only provide the original dataset and the aggregation or configuration boundaries chosen for the analysis. Of course, reticence to provide data which may have been collected at great cost could create logistical problems for the kind of data sharing needed for this solution to work.

9.5. CHAPTER SUMMARY

The conceptual framework developed in Chapter 5 identified two types of reports: private and public. The private report is generated by and for the analyst using an iterative and flexible approach. The end result is to increase awareness of the data characteristics and behaviour in terms of MAUP sensitivity. The public report is produced by the analyst to accompany published analyses using areal data. The function of the public report is to allow for a reader to make an independent, albeit subjective, evaluation of the effects of MAUP on the published results.

The generation of a private report is demonstrated using Arc/Info and S+, the software applications chosen for performing this research. An hypothetical analysis, consisting of multiple regression using four variables, forms the basis of the exploration. The potential sensitivity of the base variables is investigated first, followed by the actual sensitivity given the aggregation to census tract. Finally, the result sensitivity is evaluated using three alternative configurations of the census tracts.

A number of ESDA techniques are used during the demonstration. Statistical techniques consist of the proposed measures which are descriptive in nature and allow for

the identification of data and spatial outliers, as well as other descriptive measures such as correlation coefficients and coefficients of variation. Graphic views include maps, histograms, scatterplots, and tables. Interactive techniques include brushing, highlighting, selection, and identification.

The use of ESDA techniques provides some interesting information regarding the data used for the demonstration. Areas of potential and actual sensitivity are easily identified, and the lack of spatial coincidence between the sensitivities of the variables used is apparent. The exploration of result sensitivity shows that two of the arbitrary configurations improve the model performance. The use of descriptive measures other than the proposed measures shows that the configurations which produce better model results either maximize or minimize the variation of the two variables previously identified as having different spatial distributions of sensitivity. The inclusion of zero value areal units is shown to have a relatively small impact on the aggregated data for this case study.

The required elements for a static public report of the previous multiple regression analysis are listed. Even for this simple example, it is very apparent that the minimum requirements produce a lengthy, perhaps overwhelming, public report. It is suggested that this may be the underlying reason for the lack of sensitivity reporting in the past and current literature. A reasonable way of managing the report may be to 'label' any analysis using areal data explicitly and provide a brief text report which summarizes the information regarding MAUP sensitivity. The test report would include information regarding access to the full report for interested parties, perhaps through the Internet. The optimal solution, however, would be to provide access to an ESDA application for exploring MAUP sensitivity to any interested reader of the published analysis. In this way, readers could create their own private report; however, the requirement of providing the necessary data could create logistical problems.

CHAPTER 10

CONCLUSION

10.1. INTRODUCTION

This final chapter is intended to present both a general summary and some general conclusions regarding this research. The primary objective of the research is stated in Section 2 and a brief summary of the research is presented followed by a subjective evaluation of the conceptual framework. This serves to identify areas which require further development and/or research. These areas are discussed in Section 3. Section 4 provides a subjective evaluation of ESDA as a methodology and identifies both positive and negative aspects of using this methodology.

10.2. THE CONCEPTUAL FRAMEWORK

The primary objective of this research was to apply ESDA as a methodology to MAUP sensitivity reporting in a conceptual manner, in order make a first step toward developing a functional ESDA application for MAUP sensitivity reporting. The research therefore focused on developing a conceptual framework which integrated both the characteristics of MAUP and the requirements of the ESDA methodology. The conceptual framework provided a guide for further development and extension of the ideas generated by that process. Where applicable, the concepts generated were applied to an actual dataset using a hypothetical problem for demonstration and evaluation purposes. The following sections describe the various stages in the development of the framework and its subsequent extension.

10.2.1. The Characteristics of MAUP

Chapter 2 was devoted to defining and exploring MAUP in terms of its effects on analyses and of the ways in which MAUP effects are created. A number of empirical

studies were reviewed which established the extent of MAUP effects, and a summary of the ways in which MAUP effects are created was presented. The key aspects of MAUP identified are: the types of analyses which are susceptible to MAUP; the types of data which are susceptible to MAUP; and the existence of a number of methods for combining values associated with areal units, each of which make implicit assumptions about the data and create MAUP effects in different ways.

Chapter 3 focused on approaches suggested for dealing with the effects of MAUP. The approaches included methods for avoiding MAUP, methods for solving MAUP, and methods for minimizing MAUP. The reviewed methods used three general approaches: a variable-centred approach in which the methods used the variables themselves for analysis; a result-centred approach in which the results of a particular analytical techniques were used for analysis; or a combination of both approaches. In all cases, MAUP effects were studied by creating a series of aggregations and associated configurations from a base dataset. The key aspects of MAUP that were identified are: MAUP effects are more common when arbitrary areal unit boundaries are used; MAUP can be avoided only in restricted circumstances; methods for solving MAUP are non-existent; and methods for minimizing MAUP are far from having general applicability.

10.2.2. The Characteristics of ESDA

Chapter 4 presented an in-depth discussion of ESDA, first in relation to other approaches to data analysis, and secondly in terms of its use as a methodology. The defining characteristics of ESDA are: the use of descriptive methods of analysis; graphics, which must include maps as aids to visual thinking; interactive and dynamic techniques; and the goal of increasing awareness or knowledge about the data rather than producing a single quantitative result.

As a methodology, ESDA defines the kinds of techniques or methods which are appropriate for research use. Statistical methods are used for identifying patterns or trends

and so consist of data outlier and spatial outlier detection methods, descriptive measures such as mean and standard deviation, and spatially disaggregate measures which provide for meaningful map views. Graphic methods include maps, scatterplots, histograms, boxplots, and a range of other techniques for graphically displaying the data. Interaction methods range from selection, identification, and highlighting, to brushing, rotation of plots and moving statistical windows. The real power of ESDA comes from using these techniques in conjunction with each other so that multiple views (graphics) are displayed in a linked fashion which allows for interaction with one view to be reflected immediately in all other views.

10.2.3. The General Conceptual Framework

A general conceptual framework was developed from the characteristics identified above. The potential for MAUP to be more common when arbitrary areal units, such as political or administrative boundaries, are used was the basis for restricting the conceptual framework to addressing vector format data specifically. Similarly, the type of data often associated with political or administrative boundaries are generally concerned with socio-economic phenomena, and so only cardinal level data were specifically addressed.

The idea that MAUP can be studied by creating a hierarchy of datasets, each of which correspond to a particular aggregation and configuration of the base data, and the idea that either a variable-centred or a result-centred approach can be used, were combined to identify three levels for reporting MAUP sensitivity. Given three levels for reporting and different approaches, it was suggested that the concept of 'sensitivity to MAUP differs between report levels. Finally, the type of report suggested and implied in the current literature is for 'public' use, since it serves to document MAUP effects in order to provide the reader with some way of assessing the result quality. Using an ESDA methodology, the product of analysis is not necessarily a static or formal result, but rather an increased awareness of the behaviour of the data. This suggested that a second type of

report existed - a private report, which is generated by and for the analyst in an iterative and flexible manner.

A set of normative goals for developing a method or methods for reporting MAUP sensitivity were developed using the conceptual framework as a guide. These goals state that a method for reporting MAUP sensitivity should: allow for the use of a wide range of analytical techniques and data types; allow for both variable-centred and result-centred approaches; allow for the generation of both private and public reports; include appropriate spatially disaggregated measures of sensitivity; allow for the creation of aggregations and configurations of a base dataset; and provide new data associated with the new aggregations or configurations.

The chapter concluded with a review of research efforts aimed at implementing ESDA methods in a computerized environment. These efforts were categorized as being either general ESDA applications or problem specific applications. Finally, the technical environment selected for this research was identified.

10.2.4. Report Levels

Chapter 6 was the first in a set of three chapters which extended the general conceptual framework by focusing on each of the report levels identified respectively. The first level of reporting, base variable sensitivity was the subject of this chapter. The general conceptual framework identified the areas for further development: the concept of sensitivity associated with report level I, the approach used (variable-centred or result-centred), and the identification of spatially disaggregate measures for MAUP sensitivity.

At this level, the concept of sensitivity was based on the potential for change and was linked to the characteristics of the data for a specific areal unit in comparison to the data for the contiguous neighbours of that areal unit. As well, the planned method for calculating new values for aggregations or configurations was identified as having an influence on sensitivity. Spatial autocorrelation measures were reviewed; however, none

could be easily adjusted to allow for the incorporation of the influence of the planned calculation method. Two preliminary measures were proposed and demonstrated using the software selected for this research.

Chapter 7 focused on the second level of reporting - variable sensitivity to aggregation and configuration. Areas identified for further development were: the concept of sensitivity, the approach used (variable-centred or result-centred), and the identification of spatially disaggregate measures for MAUP sensitivity. At this report level, the concept of sensitivity was based on the actual change caused by aggregation or reconfiguration. Sensitivity was liked to the homogeneity of the new grouping and so methods based on variation measures were reviewed. These methods included analysis of variance statistics and entropy statistics; however, they did not allow for a spatially disaggregated measure. Four measures, based on cartographic modeling techniques, were proposed and demonstrated by creating a series of aggregations and configurations with the software selected for this research.

Chapter 8 was concerned with the third level of reporting, identified as result sensitivity. At this level the concept of sensitivity was identified as the same concept implicit in the current literature regarding MAUP. That is, the range of possible results indicates the degree of sensitivity. Methods for summarizing the range of results were reviewed and adopted for this research. These methods were demonstrated for a hypothetical analysis using the software selected for this research.

10.2.5. Report Uses

The general conceptual framework identified two kinds of reports: private and public. Chapter 9 was concerned with demonstrating a possible process for generating these reports using the previous work as a guideline. With the use of an hypothetical multiple regression analysis, graphic views and interactive techniques as required by ESDA were used to demonstrate the creation of a private report. The demonstration

provided some insight into the process of generating the report, the uses for different views and techniques, and the kind of information gained through the exploration of the data and the proposed measures. In this case, it was identified that no specific linear process is suitable, rather each individual may explore the information in a unique way. The exploration did provide some interesting information regarding the demonstrated analysis which could be useful to the analyst and contribute to a better understanding of the behaviour of the data in terms of MAUP.

Given the information produced through the demonstration of the private reporting process, the information necessary for a comprehensive public report was identified. Required data included all information which would allow a reader to make an independent subjective evaluation of the result quality in terms of MAUP. The information identified as necessary for the relatively simple analysis demonstrated was voluminous and it was suggested that, rather than including all the information required with the published report, a summary of the report could be provided along with the necessary directions for accessing the full report. It was also suggested that all data used for the analysis be made available for those with access to an ESDA application which would allow for independent exploration and circumvent the need for a comprehensive static public report.

10.2.6. A Subjective Evaluation of the Conceptual Framework

The general conceptual framework provides a way of integrating the characteristics of MAUP and the characteristics of ESDA as a methodology. This has proven to be a fairly powerful organizing concept. The definition of MAUP sensitivity implicit in the literature regarding MAUP has been identified as being incomplete and somewhat restrictive. Rather than pertaining strictly to the range of results produced by using a number of aggregations or configurations, the concept of sensitivity varies depending on the data type, the planned or used method for calculating new values, the approach used (i.e. variable-centred or result-centred), and the level in the data

provided some insight into the process of generating the report, the uses for different views and techniques, and the kind of information gained through the exploration of the data and the proposed measures. In this case, it was identified that no specific linear process is suitable, rather each individual may explore the information in a unique way. The exploration did provide some interesting information regarding the demonstrated analysis which could be useful to the analyst and contribute to a better understanding of the behaviour of the data in terms of MAUP.

Given the information produced through the demonstration of the private reporting process, the information necessary for a comprehensive public report was identified. Required data included all information which would allow a reader to make an independent subjective evaluation of the result quality in terms of MAUP. The information identified as necessary for the relatively simple analysis demonstrated was voluminous and it was suggested that, rather than including all the information required with the published report, a summary of the report could be provided along with the necessary directions for accessing the full report. It was also suggested that all data used for the analysis be made available for those with access to an ESDA application which would allow for independent exploration and circumvent the need for a comprehensive static public report.

10.2.6. A Subjective Evaluation of the Conceptual Framework

The general conceptual framework provides a way of integrating the characteristics of MAUP and the characteristics of ESDA as a methodology. This has proven to be a fairly powerful organizing concept. The definition of MAUP sensitivity implicit in the literature regarding MAUP has been identified as being incomplete and somewhat restrictive. Rather than pertaining strictly to the range of results produced by using a number of aggregations or configurations, the concept of sensitivity varies depending on the data type, the planned or used method for calculating new values, the approach used (i.e. variable-centred or result-centred), and the level in the data

transformation process (i.e. base level, aggregation/configuration level, or result level). The identification of these characteristics through the use of the conceptual framework suggests that reporting sensitivity to MAUP is not a simple endeavor, and the development of a single procedure or method is unlikely and perhaps even undesirable.

Although the use of the conceptual framework creates a complex set of requirements for reporting MAUP sensitivity, the complexity is based on using the information inherent in the dataset. For example, the definition of sensitivity associated with the results of an analysis is restricted to using only the information regarding the distribution of the results produced. Extending the definition of sensitivity to include a variable-centred approach incorporates information about the spatial distribution and numerical characteristics of the original data, and information about the actual changes caused by aggregating or re-configuring the original data.

The extensions of the conceptual framework made in Chapters 6 through 9 are specific to summary measures for cardinal level areal data in vector format. As such, only the general framework can be applied to a wide range of data types and formats. Where appropriate, some conclusions have been made regarding the application of the extended framework to raster format and categorical data. These conclusions, specifically in Chapters 6 and 7, suggest that cardinal level raster data can be incorporated within the extended framework presented in this research; however, categorical data prove to be much more problematic. In order to identify potential methods or processes for reporting sensitivity for these kinds of data, it would be necessary to begin with the general conceptual framework and develop the extensions of each report level specifically for nominal and ordinal data.

10.3. AREAS FOR FURTHER DEVELOPMENT AND RESEARCH

In the previous section it was identified that the application of the extended framework to nominal and ordinal would require further research, and that the general conceptual framework could provide an appropriate starting point for that research. A number of other issues and areas for future development arise from this research and are addressed below.

10.3.1. How Many Aggregations or Configurations Should Be Explored?

For both private and public reporting, it is implicit in the conceptual framework that the analyst will make use of a number of aggregations and/or configurations in order to gain a better understanding of the data and to provide evidence that published results are stable. An important question arises at this point: how many aggregations or configurations should be used in order to investigate the data adequately?

As discussed previously, the answer to this question depends upon the type of analysis applied, the context of that analysis, the data, and the interest of the analyst. In the case of private reporting, exploration is performed specifically to increase the analysts understanding of the behaviour of the variables and analytical technique given changes in aggregation or configuration. In this case, the number of aggregations or configurations evaluated will be entirely up to each individual analyst.

In the case of public reporting, the emphasis is on the results of the analysis and so the number of aggregations and/or configurations evaluated will again depend upon the types of analysis, and on the potential use for the results. For example, if a location/allocation analysis is performed in order to identify ten potential sites for a new school, it would not make sense to test the result sensitivity using data aggregated into ten or fewer areal units. In this way, the number of aggregation levels which are suitable for testing may be logically restricted. It is suggested here that in many instances, there is a logical aggregation level defined by the study itself. For example, an analysis of

neighbourhood crime rates would not be carried out using data from municipal boundaries: a lower aggregation level is required.

Although it is conceivably possible to create an infinite number of configurations for any given aggregation level, many of those configurations would not represent logical boundaries. While it is not within the scope of this research to state that certain aggregations or configurations are better than others, within the context of a particular analysis some aggregations or configurations would seem to be inappropriate. For example, Figure 10.1 shows two configurations of ten areal units. Imagine that Configuration 1 shown below represents the actual census tracts of a study area, and that the purpose of the analysis is to identify a potential site for a new hospital. A solution based on the second configuration would most likely be unacceptable to many people as it may be perceived as being an unfair or unequal representation of the study area. There is no easy answer then, which prescribes a particular number aggregations or configurations to use when investigating the sensitivity of results.

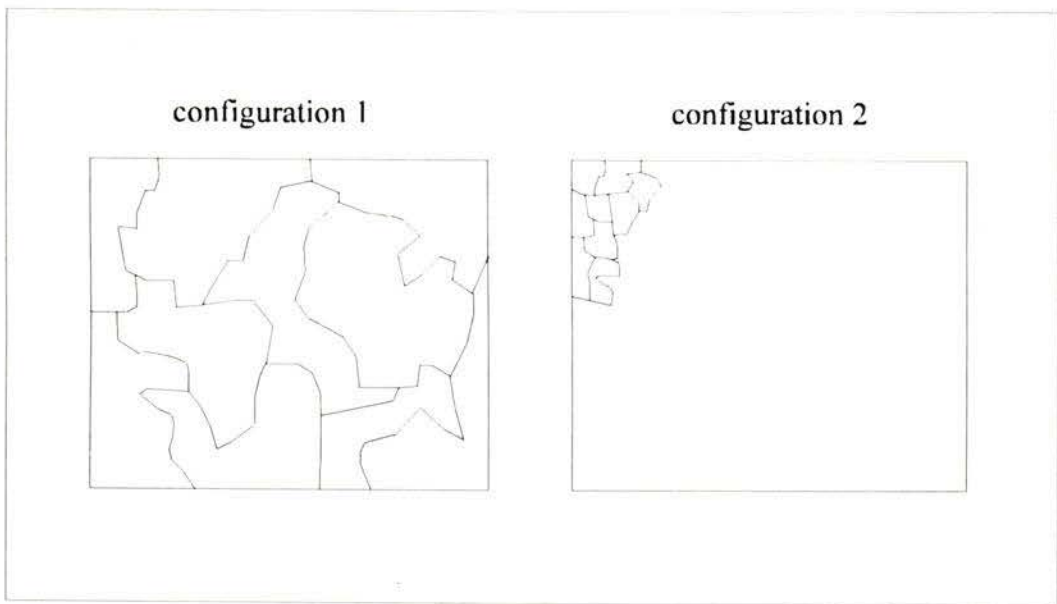


Figure 10.1. - Two Configurations of 10 Areal Units

10.3.2. Functionality

The process of simulating ESDA throughout this research has led to the identification of a number of issues which would need to be addressed and accommodated in the design of an ESDA application for exploring and reporting MAUP sensitivity, using Arc/Info and S+. There are three general groups of functions: aggregation and configuration, views, and analysis and modelling. The following sections discuss each in turn, and a discussion of the limitations of the chosen software follows.

10.3.2.1. Aggregation and Configuration Creation

One of the normative goals for the development of an ESDA method for reporting MAUP sensitivity was that any method developed should allow the user to create any number of aggregations and or configurations for use in the subsequent generation of either private or public reports. A number of methods for performing this operation should be available for use. For this research, new groupings were created in Arc/Info by using the 'select, 'merge', and 'intersect' functions. Essentially, groups of base areal units were selected either by using the cursor, or by using logical expressions based on the enumeration area number of each areal unit to form the subsequent aggregations and configurations. For example, a list of the enumeration area numbers which belonged in a particular census tract were used to form the census tract boundary system. Once a group was identified, the selected base areal units were merged. This eventually created a new map with no associated attributes. This new map was then imposed on the base map by 'intersection'. An AML program was then used to calculate the new values for the new map.

Other options for creating aggregations or configurations exist within Arc/Info. For example, areal units can be selected using the 'passthru' method in which the cursor is used to draw a 'lasso around a series of areal units, resulting in all areal units which fall completely or partially within the circle defined are selected. Selecting areal units within a

particular radius can be accomplished using proximity criteria. Batty and Xie (1994) developed a 'seeding' option to create aggregations of areal units. In practice, the user selects a certain number of areal units interactively, and the associated algorithm is based on clustering techniques. Given some parameters, the areal units are assigned to seeds in a variety of ways. This could be a very useful and efficient method of creating a number of arbitrary groupings. In general, methods for grouping areal units are tied to the functions available with Arc/Info, due to the fact that MAUP is specific to spatial aggregation and so topological information is required. A connectivity matrix may allow for some selection outside of the GIS environment, however, since Arc/Info has well developed functions for selection, it is suggested that methods be developed within Arc/Info. Although not the focus of this research, an in-depth discussion of all possible strategies for generating aggregations or configurations is sufficient material for a study in itself.

10.3.2.2 Views

Arc/Info supports the creation of map views; however creating graphs is not a particular strength. S+, on the other hand, provides a wide range of graph types. When used together, it is possible to produce a variety of views of the data. Although not demonstrated here, Arc/Info can be used to generate surface views and contour maps of the data, which can be 'viewed' from different perspectives. S+ can be used to generate simple graphs using lines or point symbols, as well as box, pie, and dot charts, box plots, density plots, quantile-quantile plots, and Chernoff faces or stars. Perspective plots such as surfaces or contours can be generated using the centroid position of each areal unit when the X Y coordinates of the centroids are available. Included in the brush function is the option of producing a 3-dimensional spin diagram which can be rotated interactively. Grey scale image maps are also supported. Between the two software packages, a wide range of options exists for creating different views of any dataset.

10.3.2.3. Analysis and Modelling

Both Arc/Info and S+ include a wide range of functions which allows for both spatial and aspatial analysis. This is the real utility in using these two software applications in tandem. Most importantly, S+ includes a number of descriptive aspatial and spatial statistics and a range of EDA techniques which include outlier and anomaly detection methods through both numeric and graphic means. These functions have not been demonstrated in this research, but include: variogram estimation, kriging, neighbour analysis, spatial autocorrelation measures, point pattern analysis techniques, and stem and leaf diagrams. The potential for the use of these additional methods should be further investigated as they can provide additional information regarding the views generated within this research.

10.3.3. Limitations of the Technical Environment

Although the software chosen for this research provides most of the functions required for ESDA, a number of limitations became apparent during the demonstration of report generation in Chapter 9. This section discusses those limitations and, where applicable, makes recommendations for future development of the software if ESDA techniques are to be truly available. These recommendations are shown in italic font.

One of the main limitations of the software has been the lack of a truly dynamic and interactive link between Arc/Info and S+. This configuration of software however, appears to be the only commercially available product which permits the use of both packages from one window, allows for the simple transfer of data between the two applications and provides many of the statistical and graphic techniques required for ESDA.

The link between the two programs is basically a data transfer link and a command link which allows for S+ commands to be entered from within an Arc/Info session. In this way, it is not necessary to leave Arc/Info in order to access the functionality of S+. This is

certainly an advance in terms of incorporating advanced statistical analysis and modelling capabilities in a GIS; however, both applications are generally lacking any kind of general user interface other than a command line. While Arc/Info does have an available module, ArcView, which provides a reasonable graphic user interface, the current version of S+GISLINK does not support the use of ArcView. *It is recommended that the link between S+ and Arc/Info be further developed to allow for linked Arc/Info and S+ views. This would necessitate the development of a common user interface, a common data model and the concurrent sharing of a single database.*

All of the views demonstrated in this research were created as separate, unlinked entities. The only exception is the brush function in S+, which generates a scatterplot matrix and associated histograms. The individual scatterplots and histograms are linked and activated by actions in any element of that display. Unfortunately, the scatterplot matrix display is not without problems.

Firstly, because the scatterplot matrix cannot be linked to an Arc/Info map view, any selection made on the scatterplots must be made separately for the map view using Arc functions such as 'select' and 'reselect'. This task is very difficult since axis scales for the scatterplots are neither produced nor available. The exact numerical range selected using the brush is not apparent; however, it is necessary to enter an exact numerical range when selecting in Arc/Info. Any direct correspondence between the highlighted scatterplot points and the selected areas on the Arc/Info map is therefore difficult to establish. The lack of scales on the scatterplot axes also deprives the user of comparing the magnitude of separate variable measures. It is possible that the high measure in one scatterplot is 10 times higher then the high measure in another; however, this is not immediately apparent. *It is recommended that, in the absence of a truly interactive link, the scatterplot matrix generated by S+ include axes scales.*

Secondly, although a table is included in the scatterplot view, the only information available is the case number of each data point. These case numbers are generated by S+.

Although other information exists within the data matrix, it is not displayed in the table view. *It is recommended that the user be given the option of choosing the information displayed in the brush function table view.*

A third problem in using the brush function concerns the graphic display. When only a few variables are used, the scatterplot matrix is reasonably sized; however, when more than three or four variables are used, the matrix quickly becomes quite large. Each individual scatterplot therefore becomes smaller and more difficult to read. Once the scatterplot matrix is generated, the graphic window cannot be resized to improve readability. As well, the brush function uses all variables present in the data matrix, and cannot be adjusted. Any addition or deletion of variables from the scatterplot view requires that a new data matrix be created and a new scatterplot matrix generated. *It is recommended that the user have the capability of interactively adding or deleting variables from the scatterplot view, and have the capability of resizing the display as needed.*

In Arc/Info, a serious limitation existed due to the length of time required for calculating the proposed measures. The proposed measures were calculated both in Arc/Info and in SPSS; S+ was not used in this case due to lack of availability during the earlier stages of this research. An AML program was created in Arc/Info to calculate SB_I since topological information regarding adjacency was required. Unfortunately, AML is not source code and must be interpreted before execution. Due to the extra processing required, the calculation of SB_I for 1st order neighbours generally took several hours to complete. When 2nd order or 3rd order neighbours were included, the time needed rose to 12 hours or more. This is a serious limitation to performing ESDA 'on the fly'. It is first necessary to calculate all the planned measures in advance, which restricts the flexibility of any subsequent exploration.

Measures for actual sensitivity to aggregations or configurations were calculated in SPSS in order to circumvent the additional processing required by the use of AML

programming. The first step in this process used an AML to intersect and calculate the new values for each boundary system. Again, this procedure required an inordinate amount of time - generally over 2 hours with the time increasing when the number of variables increased. Once the new boundary system had appropriate values, the data file was exported to SPSS where the simple differences and averages were calculated. The updated data field was then transferred back to Arc/Info and rejoined to the map.

The lack of reasonable calculation capabilities in Arc/Info when AML is used is a serious limitation. It would be preferable to have the ability to simply choose which variables and boundary systems to use in an interactive fashion, and have the appropriate measures calculated as needed. *It is recommended that the use of an external programming language be explored for use in place of Arc Macro Language, with the objective of allowing for 'immediate' calculation of any required measure.*

10.4 ESDA AS A METHODOLOGY

The use of ESDA as a methodology provides for great flexibility in the manner in which research is performed. ESDA can be used under either the 'normal science' paradigm, or the 'visualization' paradigm. Within the normal science paradigm, ESDA is used in conjunction with CSDA (confirmatory spatial data analysis). Ideas and hypotheses generated through the use of ESDA techniques are formally tested using CSDA techniques. In this sense, the end goal of the process is not just to increase understanding, but to also attempt to make generalizations through explaining the patterns identified through ESDA. Under the visualization paradigm, ESDA is 'stand-alone' - the generation of ideas is not followed by CSDA within any given research effort, and so formal testing of hypotheses is not performed and general theories are not produced.

Although ESDA allows for flexibility in the approach to research, the dependence on visualization as a key technique creates both strengths and weaknesses. A particular strength lies in the use of maps (including surfaces), which add an important dimension to

the analysis of spatial data and it is hard to formulate any argument against their use. Similarly, the interaction techniques allow for visual data analysis which truly extends the usefulness of static graphics. It is suggested here, however, that this dependence on data views and interactive techniques creates a restriction on the applicability of ESDA in general. For example, one technique for visualizing multivariate data is the use of Chernoff faces (Chernoff 1973). Specialized symbols are created for each data point within the dataset and graphed. In a Chernoff face, the 'eyes' may represent variable 1, the 'mouth' may represent variable 2, the 'head' may represent variable 3. Each of the 'features' changes in form to reflect the underlying data values. This is an interesting method, but is, in practice, restricted to being used on very small data sets. Interpretation of more than a dozen or so 'faces' becomes very difficult, and as the size of the dataset increases, either the size of the 'faces' must decrease or the size of the display must increase in order to represent all data points. This example serves to highlight the general problem of visualization: the readability and associated interpretability of the views generated.

Using Arc/Info and S+ in this research allowed for the relatively quick generation of a number of data views. Unfortunately, both the map views generated in Arc/Info and the scatterplot matrices generated in S+ were extremely difficult to read accurately, even when the display window covered the entire screen. This severely impacted this researcher's ability to concurrently evaluate more than one view of the data at a time. It is suggested here that this problem is not specific to the research performed here, but rather may occur in any given situation. The utility of ESDA therefore may be diminished when large datasets associated with complex maps are used.

10.5. CONCLUSION

The integration of the characteristics of MAUP and the characteristics of ESDA as a methodology has identified a number of areas and issues which need to be addressed if an ESDA application for reporting MAUP sensitivity is to be developed. Perhaps the

most fundamental idea generated here is that there are many different concepts of sensitivity, each of which require unique and specific measures to identify areas of sensitivity. Secondly, any measure developed for this purpose is most useful in a spatially disaggregate form. Thirdly, the use of ESDA techniques for generating private and public reports appears to provide information which could be used to guide the aggregation or configurations process, and allow for subjective evaluation of published results by readers.

While the private report can be generated in a flexible and unstructured way, a static public report which accompanies published research, such as would be required by readers without access to an ESDA application, appears to be inordinately large and therefore would not likely become general practice. One option is to provide the full report in a digital format, such as through the Internet, as a means of giving interested parties access. The optimal solution would be to provide access to an ESDA application for exploring and reporting MAUP sensitivity to any interested party; however, this would require the analyst to provide all the data used for the analysis. This may also create a barrier to the widespread acceptance of reporting MAUP sensitivity.

These issues may have serious ramifications for the development of any ESDA application based on the conceptual framework developed here. The range of functions needed to create an ESDA application which incorporates all potential sensitivity measures and ESDA techniques suggests that such an application should be developed as a distinct software module rather than being embedded in current software. Limiting the functionality in order to 'piggyback' the application on some existing software would necessarily omit some of the options identified by the conceptual framework and decrease the utility of the ESDA application.

This research was conducted using what appears to be the only commercially available software configuration which links Arc/Info with a statistical package in such a way as to allow for the full functionality of both applications in a robust and easy to use way. This is certainly an important development in terms of extending the analytical

capabilities of Arc/Info; however, the weak user interface and lack of an interactive link between the applications limits the use of ESDA in practice. It appears then that generally available software applications which allow for full GIS and statistical application functionality and for a truly interactive ESDA environment are yet to be developed.

LITERATURE CITED

Anselin L, Getis A. (1992). Spatial Statistical Analysis and Geographic Information Systems, in *Annals of Regional Science*. V26 N1. pp. 19-33.

Anselin L. (1993). The Moran Scatterplot as a Means to Visualize Instability in Spatial Autocorrelation, in ESDA and GIS Workshop Proceedings. NCGIA Publication Series, USCB : Santa Barbara, CA.

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

Bach L. (1981). The Problem of Aggregation and Distance for Analysis of Accessibility and Access Opportunity in Location-Allocation Models, in *Environment and Planning A*. V13. pp. 955-978.

Bailey TC. (1994). A review of statistical spatial analysis in geographical information systems, in GIS and Spatial Analysis. Eds. S Fotheringham and P Taylor. Taylor and Francis : London. pp. 13-45.

Batty M. (1976). Entropy in Spatial Aggregation, in *Geographical Analysis*. V8. pp.1-21.

Batty M, Sikdar PK. (1982a). Spatial Aggregation in Gravity Models: 1. An Information-theoretic Framework, in *Environment and Planning A*. V14. pp. 377-405.

Batty M, Sikdar PK. (1982b). Spatial Aggregation in Gravity Models: 2. One Dimensional Population Density Models, in *Environment and Planning A*. V14. pp. 525-553.

Batty M, Sikdar PK. (1982c). Spatial Aggregation in Gravity Models: 3. Two Dimensional Trip Distribution and Location Models, in *Environment and Planning A*. V 14. pp. 629-658.

Batty M, Sikdar PK. (1982d). Spatial Aggregation in Gravity Models: 4. Generalizations and Large Scale Applications, in *Environment and Planning A*. V14. pp. 795-822.

Batty M, Sammons R. (1978). On Searching for the Most Informative Spatial Pattern, in *Environment and Planning A*. V10. pp. 747-779.

Batty M, Xie YC. (1994). Modelling Inside GIS 1. Model Structures, Exploratory Spatial Data Analysis and Aggregation, in *International Journal of Geographic Information Systems*. V8 N3. pp. 291-307.

Bearwood JE, Kirby HR. (1975). Zone Definition and the Gravity Model; The Separability, Excludability and Compressibility Properties, in *Transportation Research*. V9. pp. 363-369.

Becker RA, Cleveland WS, and Wilks AR. (1988). Dynamic Graphics for Data Analysis, in Dynamic Graphics for Statistics. Eds. WS Cleveland and ME McGill. Wadsworth : Pacific Grove, CA. pp. 1-50.

Blair P, Miller RE. (1983). Spatial Aggregation in Multi-Regional Input-Output Models, in *Environment and Planning A*. V 15. pp. 187-206.

Brodlie K. (1994). A Typology for Scientific Visualization, in Visualization In Geographical Information Systems. Eds. H Hearnshaw and D Unwin. John Wiley and Sons : Chichester, UK. pp. 34-41.

Chernoff H. (1973). The Use of Faces to Represent Points in k-Dimensional Space Graphically, in *Journal of the American Statistical Association*. V68. pp. 361-368.

Clark WAV, Avery KL. (1976). The Effects of Data Aggregation in Statistical Analysis, in *Geographical Analysis*. V8. pp. 428-438.

Cleveland WS, McGill ME. (1988). Dynamic Graphics for Statistics. Wadsworth : Pacific Grove, CA.

Cliff AD, Ord JK. (1973). Spatial Autocorrelation. Pion Ltd. : London.

DiBiase D, MacEachren AM, Krygier JB, and Reeves C.. (1992). Animation and the Role of Map Design in Scientific Visualization, in *Cartography and Geographic Information Systems*. V19 N4. pp. 201-214.

Dorling D. (1992). Stretching Space and Splicing Time: From Cartographic Animation to Interactive Visualization, in *Cartography and Geographic Information Systems*. V19 N4. pp. 215-227.

Dykes J. (1994) Area-Value Data: New Visual Emphases and Representations, in Visualization In Geographical Information Systems. Eds. H Hearnshaw and D Unwin. John Wiley and Sons : Chichester, UK. pp. 103-114.

Fotheringham S. (1989). Scale-Independent Spatial Analysis, in Accuracy of Spatial Databases. Eds. M Goodchild and S Gopal. Taylor and Francis : New York. pp. 221-228.

Fotheringham S. (1993). GIS and Exploratory Spatial Data Analysis, in ESDA and GIS Workshop Proceedings. NCGIA Publication Series, USCB : Santa Barbara, CA.

Fotheringham S, Wong DWS. (1991). The Modifiable Areal Unit Problem in Multivariate Statistical Analysis, in *Environment and Planning A*. V23 N7. pp. 1025-1044.

Getis A. (1993). Untitled Position Paper, in ESDA and GIS Workshop Proceedings. NCGIA Publication Series, USCB : Santa Barbara, CA.

Getis A. (1994). Spatial Dependence and Heterogeneity and Proximal Databases, in Spatial Analysis and GIS. Eds. S Fotheringham and P. Taylor. Taylor and Francis : London. pp. 105-120.

Getis A, Ord JK. (1992). The Analysis of Spatial Association by Use of Distance Statistics, in *Geographical Analysis*. V24 N3. pp. 189-206.

Good IJ. (1983). The Philosophy of Exploratory Data Analysis, in *Philosophy of Science*. V50. pp. 283-295.

Goodchild M. (1979). The Aggregation Problem in Location Allocation, in *Geographical Analysis*. V11. pp. 240-255.

Goodchild M, Haining R, Wise S, and 12 others. (1992). Integrating GIS and Spatial Data Analysis - Problems and Possibilities, in *International Journal of Geographical Information Systems*. V6 N5. pp. 407-423.

Goodchild M, Anselin L, and Deichmann U. (1993). A Framework for the Areal Interpolation of Socioeconomic Data, in *Environment and Planning A*. V25. pp. 383-397.

Griffith DA. (1987). Spatial Autocorrelation A Primer. Resource Publications in Geography, Association of American Geographers : Washington.

Haining R. (1994). Designing Spatial Data Analysis Modules for Geographical Information Systems, in Spatial Analysis and GIS. Eds. S Fotheringham and P. Taylor. Taylor and Francis : London. pp. 45-63.

Haslett J, Bradley R, Craig P, Unwin A, and Wills, G. (1991). Dynamic Graphics for Exploring Spatial Data with Application to Locating Global and Local Anomalies, in *American Statistician*. V45. pp. 234-242.

Haslett J, Wills G, and Unwin AR. (1990). SPIDER - An Interactive Statistical Tool for the Analysis of Spatially Distributed Data in, *International Journal of Geographical Information Systems*. V4. pp. 285-296.

Keller CP. (1994). Exploratory Spatial Data Analysis (ESDA) - The Next Revolution in GIS, in Symposium Proceedings Volume 1. 8th Annual Symposium on Geographic Information Systems, Vancouver, BC February 21-24, 1994. pp.297-302.

Larson RC. (1986). The Invalidity of Modifiable Areal Unit Randomization, in *Professional Geographer*. V38 N4. pp. 369-374.

MacDougall EB. (1992). Exploratory Analysis, Dynamic Statistical Visualization, and Geographic Information Systems, in *Cartography and Geographic Information Systems*. V19. pp. 237-246.

MacEachren AM, Monmonier M. (1992). What's Special About Visualization, in *Cartography and Geographic Information Systems*. V 19. pp. 197-200.

Marceau DJ, Howarth PJ, and Gratton, DJ. (1994). Remote Sensing and the Measurement of Geographical Entities in a Forested Environment, in *Remote Sensing of the Environment*. V29 N2. pp. 93-104.

Masser I, Brown PJB. (1975). Hierarchical Aggregation Procedures for Interaction Data, in *Environment and Planning A*. V7. pp. 509-523.

Miller RE, Shao G. (1990). Spatial and Sectoral Aggregation in the Commodity-Industry Multiregional Input-Output Model, in *Environment and Planning A*. V22 N12. pp. 1637-1656.

Moellering H, Tobler W. (1972). Geographical Variances, in *Geographical Analysis*. V21. pp. 34-50.

Monmonier M. (1989). Geographical Brushing: Enhancing Exploratory Analysis of the Scatterplot Matrix, in *Geographical Analysis*. V21. pp. 88-84.

Morill R. (1981). Political Districting and Geographical Theory. Resource Publications in Geography Series. Association of American Geographers : Washington, DC.

Openshaw S. (1977). Optimal Zoning Systems for Spatial Interaction Models, in *Environment and Planning A*. V9. pp.169-184.

Openshaw S. (1978). An Empirical Study of Some Zone-Design Criteria, in *Environment and Planning A*. V10. pp. 781-794.

Openshaw S. (1984a). Ecological Fallacies and the Analysis of Areal Census Data, in *Environment and Planning A*. V16. pp.17-31.

Openshaw S. (1984b). The Modifiable Areal Unit Problem. Concepts and Techniques in Modern Geography (CATMOG) 38. Geoabstracts : Norwich, UK.

Openshaw S. (1993). Two Exploratory Space-Time-Attribute Pattern Analysers Relevant to GIS, in Spatial Analysis and GIS. Eds. S Fotheringham and P. Taylor. Taylor and Francis : London. pp. 83-104.

Openshaw S, Charlton M, Wymer C, and Craft A. (1987). A Mark I Geographical Analysis Machine for the Automated Analysis of Point Data Sets, in *International Journal of geographical Information Systems*. V1 N4. pp. 355-358.

Openshaw S, Cross A, and Charlton M. (1990). Building a Prototype Geographical Correlates Exploration Machine, in *International Journal of Geographical Information Systems*. V4. pp. 297-311.

Perle ED. (1977). Scale Changes and Impacts on Factorial Ecology Structures, in *Environment and Planning A*. V9. pp. 549-558.

Putnam SH, Chung S-H. (1989). Effects of Spatial Systems Design on Spatial Interaction Models 1: The Spatial Definition Problem, in *Environment and Planning A*. V21. pp. 27-46.

Robinson AH. (1956). The Necessity of Weighting Values in Correlation Analysis of Areal Data, in *Annals of the Association of American Geographer*. V46. pp.233-236.

Robinson WS. (1950). Ecological Correlations and the Behaviour of Individuals, in *American Sociological Review*. V15. pp. 351-357.

Rogerson PA, Fotheringham SA. (1994). GIS and Spatial Analysis - Introduction and Overview, in Spatial Analysis and GIS. Eds. SA Fotheringham and P Taylor. Taylor and Francis : London. pp. 3-10.

Sawicki DS. (1973). Studies of Aggregated Areal Data - Problems of Statistical Inference, in *Land Economics*. V49. pp. 109-114.

Scott LM. (1994). Identification of GIS Attribute Error Using Exploratory Data Analysis, in *Professional Geographer*. V46. N3. pp. 378-386.

Summerfield MA. (1983). Populations, Samples and Statistical Inference in Geography, in *Professional Geographer*. V35. pp. 143-149.

Thomas EN, Anderson DL. (1965). Additional Comments on Weighting Values in Correlation Analysis of Areal Data, in *Annals of the Association of American Geographers*. V55. pp. 492-505.

Tobler W. (1989). Frame Independent Spatial Analysis, in The Accuracy of Spatial Databases. Eds. M Goodchild and S. Gopal. Taylor and Francis : New York. pp. 115-122.

Tomlin CD. (1991). Cartographic Modelling, in Geographical Information Systems Principles and Applications Volume 1 : Principles. Eds. DJ Maguire, MF Goodchild, and DW Rhind. Longman Scientific and Technical with John Wiley and Sons Inc : NY. pp. 361-374.

Tukey JW. (1977). Exploratory Data Analysis. Addison-Wesley : Reading, MA.

Unwin A. (1993). Interactive Statistical Graphics for GIS - Current Status and Future Potential, in ESDA and GIS Workshop Proceedings. NCGIA Publication Series. USCB : Santa Barbara, CA.

Woodcock CE, Strahler AH. (1987). The Factor of Scale in Remote Sensing, in *Remote Sensing of the Environment*. V21. pp. 311-332.

Xia FF, Fotheringham AS. (1993). Exploratory Spatial Data Analysis With GIS --The Development of the ESDA Module under Arc/Info, in GIS/LIS '93 Annual Conference and Exposition Proceedings Volume II. Minneapolis, Minnesota. pp. 801- 810.

VITA

Surname : Setton Given Names : Eleanor May

Place of Birth: Victoria, British Columbia, Canada

Educational Institutions Attended:

University of Victoria	1994 to 1996
University of British Columbia	1991 to 1994

Degrees Awarded:

B.A.	University of British Columbia	1994

PARTIAL COPYRIGHT LICENSE

I hereby grant the right to lend my thesis to users of the University of Victoria Library, and to make single copies only for such users or in response to a request from the Library of any other university, or similar institution, on its behalf or for one of its users. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by me or a member of the University designated by me. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Title of Thesis:

Toward the Development of an Exploratory Spatial Data Analysis Application for Reporting Sensitivity to the Modifiable Areal Unit Problem: A Conceptualization

Author 

Eleanor May Setton
June 19, 1996