INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Measuring spatial pattern in census units: Residential segregation in Franklin County, Ohio

Newsome, Tracy Hunter, Ph.D.

The Ohio State University, 1992
MEASURING SPATIAL PATTERN IN CENSUS UNITS:
RESIDENTIAL SEGREGATION IN FRANKLIN COUNTY, OHIO

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy in the Graduate
School of The Ohio State University

By

Tracy Hunter Newsome, B.S., M.A.

The Ohio State University
1992

Dissertation Committee:
Morton O'Kelly
Edward Taaffe
Randall Jackson
Mark McCord

Approved by
Morton O'Kelly
Adviser
Department of Geography
ACKNOWLEDGEMENTS

I wish to express thanks to Dr. Morton O'Kelly for his guidance and advice during this research. As well, I thank the other members of my committee, Dr. Edward J. Taaffe, Dr. Randall Jackson, and Dr. Mark McCord for their comments and suggestions. Furthermore, I express gratitude to Jonathan Comer for technical assistance, as well as to Joe Damico, Greg Orth, and other representatives of Academic Computing Services at The Ohio State University. I am also grateful for the support extended by the Committee on Urban Affairs of the Ohio State University Urban University Program during the project entitled "A Geographical Information System to Map 1980-1990 Block Group Changes in Franklin County". Finally, I thank Wayne Walcott for his encouragement and support throughout my academic endeavors.
VITA

August 19, 1961................................................................. Born - Eustis, Florida

1985................................................................. B.S., University of North Carolina at Charlotte, Charlotte, North Carolina

1988................................................................. M.A., University of North Carolina at Charlotte, Charlotte, North Carolina

1987-1989................................................................. Project Leader, University of North Carolina Institute for Transportation Research and Education, Charlotte, North Carolina

1989-Present................................................................. Graduate Research Associate, The Ohio State University, Columbus, Ohio

FIELDS OF STUDY

Major Field: Geography

  Studies in Location Analysis (Dr. Morton O'Kelly, Dr. Randall Jackson) and Transportation (Dr. Mark McCord and Dr. Morton O'Kelly)
TABLE OF CONTENTS

ACKNOWLEDGMENTS................................................................. ii
VITA............................................................................................. iii
LIST OF TABLES......................................................................... vi
LIST OF FIGURES....................................................................... vii

CHAPTER

I. INTRODUCTION........................................................................... 1

II. LITERATURE............................................................................... 9

2.1 Introduction............................................................................. 9
2.2 Aggregation Error................................................................. 10
2.3 Aggregation and Cluster Construction............................... 16
2.4 Clusters and Pattern Detection......................................... 22
2.5 Graph Theoretic Applications.......................................... 25
2.6 Homogeneity, Clustering, and Pattern Definition........... 26
2.7 Segregation and Related Indices........................................ 29
2.8 Conclusion............................................................................. 39

III. METHODS............................................................................. 41

3.1 Introduction............................................................................. 41
3.2 Data..................................................................................... 42
3.3 Methods............................................................................... 43
3.3.1 Moran's $I$...................................................................... 43
3.3.2 Minimum Spanning Tree............................................... 53
3.4 Geographic Data Problems.............................................. 61
3.5 Conclusion............................................................................. 63

IV. APPLICATION TO CONSTRUCTED DATA.............................. 65

4.1 Introduction............................................................................. 65
4.2 Data and Methods............................................................... 66
4.3 Results Using the Adjacency Criterion............................. 71
4.3.1 Moran's $I$...................................................................... 71
4.3.2 Minimum Spanning Tree............................................... 75
4.4 Results Using the Distance Contiguity Criterion............. 97
4.4.1 Moran's $I$...................................................................... 97
4.4.2 Minimum Spanning Tree ..................................................... 99
4.5 Conclusion ........................................................................ 101

V. APPLICATION TO CENSUS DATA: FRANKLIN
COUNTY, OHIO ................................................................. 104

5.1 Introduction ..................................................................... 104
5.2 Study Area and Data ...................................................... 105
5.3 Nonwhite Population Density Maps
and Segregation Indices ...................................................... 109
5.4 Results .......................................................................... 116
5.5 Comparison of Moran's $I$ and MST Methods ............. 124
5.6 Conclusion ................................................................... 127

VI. CONCLUSIONS .................................................................. 129

APPENDICES

A. Fortran Program for Moran's $I$ Calculation .................. 137
B. Fortran Program for MST Calculation ......................... 143

LIST OF REFERENCES .......................................................... 160
LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Results of Spatial Autocorrelation on Constructed Datasets: Adjacency Criterion</td>
<td>72</td>
</tr>
<tr>
<td>2. Results of Minimum Spanning Tree on Constructed Datasets: Adjacency Criterion</td>
<td>75</td>
</tr>
<tr>
<td>3. Average Edge Cost in MST for Constructed Datasets</td>
<td>78</td>
</tr>
<tr>
<td>4. Random Permutation Results for Constructed Datasets</td>
<td>81</td>
</tr>
<tr>
<td>5. Results of Spatial Autocorrelation on Constructed Datasets: Distance Criterion</td>
<td>97</td>
</tr>
<tr>
<td>6. Results of Minimum Spanning Tree on Constructed Datasets: Distance Criterion</td>
<td>100</td>
</tr>
<tr>
<td>7. Moran's $I$ and MST Results: Franklin County</td>
<td>117</td>
</tr>
<tr>
<td>8. Random Permutation Results for Franklin County</td>
<td>118</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURES</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Pattern Descriptions</td>
<td>28</td>
</tr>
<tr>
<td>2. Subareas of a Region</td>
<td>57</td>
</tr>
<tr>
<td>3. Connected Subareas of a Region</td>
<td>57</td>
</tr>
<tr>
<td>4. Resulting Graph of Region With Edge Costs</td>
<td>58</td>
</tr>
<tr>
<td>5. Graph With Evident Clustering</td>
<td>59</td>
</tr>
<tr>
<td>6. Graph With No Evident Clustering</td>
<td>60</td>
</tr>
<tr>
<td>7. Constructed &quot;Region&quot; and Aggregation Scheme</td>
<td>67</td>
</tr>
<tr>
<td>8. Positive (Clustered) Pattern</td>
<td>68</td>
</tr>
<tr>
<td>9. Negative (Checkerboard) Pattern</td>
<td>68</td>
</tr>
<tr>
<td>10. Results of Aggregating a Checkerboard Pattern</td>
<td>74</td>
</tr>
<tr>
<td>11. MST Permutation Frequencies: Block Level Positive Dataset, Without Actual MST Value</td>
<td>86</td>
</tr>
<tr>
<td>12. MST Permutation Frequencies: Block Level Positive Dataset, With Actual MST Value</td>
<td>87</td>
</tr>
<tr>
<td>13. MST Permutation Frequencies: Block Group Level Positive Dataset, Without Actual MST Value</td>
<td>88</td>
</tr>
<tr>
<td>14. MST Permutation Frequencies: Block Group Level Positive Dataset, With Actual MST Value</td>
<td>89</td>
</tr>
<tr>
<td>15. MST Permutation Frequencies: Tract Level Positive Dataset, Without Actual MST Value</td>
<td>90</td>
</tr>
<tr>
<td>16. MST Permutation Frequencies: Block Level Negative Dataset, Without Actual MST Value</td>
<td>92</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

This research compares the usefulness of two approaches to measuring spatial pattern for areal geographic units, specifically census units. The first method is a commonly used index of spatial autocorrelation, Moran's $I$ (attributed to Moran, 1948). The properties of Moran's $I$ are well known and this technique has been used to discover the degree to which data are interdependent over an area. Spatial autocorrelation is often viewed as a statistical problem for those interested in developing models (see, for instance, Anselin, 1988) and Moran's $I$ is used to check for pattern in such situations, as well as in other studies where areal units are tested for pattern. Because Moran's $I$ has known properties as well as common usage, it is used here to provide an appropriate comparison with the second method used.

The second method used in this research is a novel approach to measuring pattern within areal (census) units. It is a graph theoretical method based on the value of a minimum spanning tree created from a graphical depiction of the region under study. The approach is based on a materials science application by Shier (1982) and Filliben, Kafadar, and Shier (1983). The properties of this approach are less well known than those of Moran's $I$ and other spatial autocorrelation measures, but it has an intuitive appeal for those interested in discerning spatial pattern. The emphasis on this method is not necessarily as a replacement for Moran's $I$ or other measures, but rather to provide a complementary technique to be
used by spatial analysts. It is felt that this approach offers a useful alternative to measurements which are less geographically oriented (i.e. do not take explicit account of locational relationships between the areas under study) or are predicated on some hypothesized underlying distribution. Moran's $I$ is used as a benchmark method to help test the minimum spanning tree (MST) approach used in this research. Moran's $I$ provides an excellent method for comparison because of its similarities to the MST method, which are discussed in detail in Chapter III.

Moran's $I$ and the MST method are used here to examine the level of residential segregation of whites and nonwhites in Franklin County, Ohio. The outcome of this research and the usefulness of the methods presented are equally applicable to other geographic questions related to pattern. The question of residential segregation was chosen, in part, because it is so pattern oriented, yet many of the measures commonly used to measure it are essentially aspatial, such as the index of dissimilarity. In addition, the issue of segregation is an issue which continues to garnish attention in academic research and the public eye, and the development of additional techniques to investigate it is warranted. The emphasis of this research, though, is methodological. The purpose is to determine the usefulness of the approaches presented here for discerning spatial pattern in many possible applications.

While the general question of this research is to determine the effectiveness of the MST method, the stated purpose of this research is twofold: first, it seeks to determine the usefulness of the two approaches presented here in detecting spatial patterns, particularly the newer MST method. Second, in comparing the two methods, this research further seeks to determine the effect on these methods of using successively higher levels of aggregation. While the effects of aggregation on Moran's $I$ have been noted, the MST method is novel and aggregation effects are less known. It is hypothesized that the MST
method will compare well to Moran's \( I \) in terms of discerning pattern, but will be affected by aggregation. However, determination of the type and degree of aggregation effects will aid in the application of these techniques to study pattern. It is hoped that the outcome of this research is a further development of the understanding of the effects of using aggregated areal data as well as the addition of a new spatial analytical tool for those interested in all forms of spatial analysis.

This research is conducted in two stages, to analyze fully the applicability of the MST method and compare the results of that method to the Moran's \( I \) index. The first stage of this research consists of an application of the two methods in a simulated regional environment, created to emulate census-like data units. The results from this analysis lead to the second stage of the research, which consists of an application of the methods to determining residential segregation patterns in Franklin County, Ohio. While there are many possible applications of the methods presented here for determining spatial pattern, the issue of segregation provides an opportunity to improve upon commonly used existing methods. As is discussed in detail in Chapter II, many of the commonly used measures of segregation are aspatial or, for those methods which do consider spatial relations, require pre-defined classes or categories which are often the only units used in the analysis. The methods presented here overcome some of these drawbacks and should provide a useful approach to discerning segregation patterns in an urban area. Furthermore, the attention paid to aggregation effects should prove useful to those interested in the future application of such methods to pattern detection.

Inherent to the question of analyzing spatial pattern is the scale at which that pattern is measured. As is discussed in Chapter II, pattern is often measured for point distributions of objects and many measures are available for conducting point pattern analysis. On the
other hand, the use of point pattern analysis for many problems in geography is not necessarily practical. As Lee (1967) points out, the sheer size of many geographical problems or geographical data sets requires that the data be analyzed in a form that is more manageable. This "manageability" arises either from summarizing the data already collected, or from the fact that many of the data would not be available to the researcher unless it was previously summarized. This is usually the case for those using census data as the information presented in the census is summarized. However, the data available in such a summarized areal form represents a form of data aggregation which may obscure relationships between the data or, in the case of this research, spatial patterns in the data.

The issue of data aggregation in geographic studies is significant. That aggregation may take the form of purposely assigning demand to aggregate points, such as in some location models, or can arise from commonly used aggregate sources, such as the census. It is generally assumed that use of aggregated data covers underlying relationships (or, in this case, patterns) or in other ways disrupts the usefulness of certain measures. In addition, the level of aggregation can obscure spatial patterns, the subject of this research, or even cause misinterpretation of patterns. Chapter II provides a more detailed look at aggregation effects in geographic studies. Aggregation has been shown to affect the use of many measures, including the spatial autocorrelation measure used in this research, and can influence the results of relational models built upon such data. Even given this, the use of aggregated areal data, such as census data, is very common for both academic researchers and non-academic geographic practitioners. As Clark and Avery (1976) put it, even though aggregation effects are known for many measures, "...social scientists in general and geographers in particular continue to work with aggregate data as a surrogate for individual data (p. 429)." This research, then, is concerned not only with the measurement of spatial pattern, but specifically with its measurement using census units. As such, it is also
explicitly concerned with the effects of aggregation upon the ability of the measures employed here to discern spatial patterns using such census units. Therefore, this research tests for pattern at different levels in the census hierarchy to determine the effects of continued aggregation to higher levels.

Chapter II provides an introduction to the literature related to the use of Moran's $I$ and the proposed MST method for measuring spatial pattern. First among these is the general issue of data aggregation. This includes not only the use of census data, but also the purposeful aggregation of demand units in location models, and extends to the concept of determining the best method for aggregating data or even zones of data. The discussion on aggregating zones leads further to a discussion of creating clusters and finally to methods for discerning clusters or other patterns in data, whether they be on the landscape or in some other "space". The issue of pattern detection in general is raised, through a discussion of traditional methods for point pattern analysis. Pattern detection is related to pattern definition, which is discussed particularly in regard to the MST method and its proposed measure of homogeneity for regions. Further, some applications of graph theoretic tools are discussed, again in relation to the MST method used here. Finally, the issue of residential segregation is discussed, with an overview of the traditional methods for measuring such segregation, and a discussion of how the methods used here can provide useful additional tools for segregation analysis.

Chapter III presents, in detail, the methods used in this research, as well as an introduction to the data used for each phase of the research. The general topic of spatial autocorrelation is presented, particularly the spatial autocorrelation index Moran's $I$. Included is a discussion of some common problems related to the use of areal geographic data and how those may be handled by such measures. A brief introduction to graph theoretic definitions
leads to a discussion of the MST method, including illustration of how a region may be
depicted as a graph and how that graph can be used to detect patterns in the data. Finally,
some of the overlapping issues related to the use of areal data for both methods are
discussed.

Chapters IV and V present the results of the two stage analysis. The first stage, presented
in Chapter IV, consists of an experimental application of the two methods to constructed
data regions, arranged to depict specific patterns. The results of this first stage of the
analysis set the stage for the application of the two methods to determining residential
segregation patterns in Franklin County, Ohio. The results of the second phase of analysis
are presented in Chapter V. From these analyses, comparisons can be made between the
Moran's $I$ index and the MST method in terms of their usefulness for measuring pattern
over a region. Furthermore, these analyses include investigation of the effect of using
different aggregation levels for the census data with each method.

Chapter VI summarizes the findings of this research including an emphasis on the
similarities and differences between the two methods used, as well as the aggregation
effects found for each. From this, important implications for the use of either method are
presented. Finally, this research presents interesting avenues for future research questions
related to the use of these methods for discerning spatial pattern, particularly for the MST
method, which represents a novel approach to these types of applications.

This research is concerned with measuring spatial pattern using areal units of analysis and it
investigates one of the most essential aspects of geographic studies. Gould (1970)
suggests that without pattern and spatial interdependence, geographic research holds little
interest (in Goodchild, 1986). Spatial pattern and its analysis forms the basis of most
geographical studies, either directly or indirectly. Direct concern with spatial pattern extends to those interested in regional studies, where the pattern of certain measurable aspects may be used to define a region. Those interested in studying the urban landscape, whether at the intra-urban or interurban scale, are often concerned with the patterns found for the location of human activities. Those activities range from residential or population locations to the location of businesses and commercial facilities on the urban landscape. Still others are interested in spatial patterns in an indirect fashion. They concern themselves with the processes which create such patterns and the pictures that are formed of underlying relationships. These researchers view pattern as a way to help investigate such relationships. On the other hand, those interested in developing geographic models of such relationships may view underlying spatial patterns as a statistical problem, or something to be overcome in order to appropriately apply statistical models. Therefore, spatial pattern and the role it plays in geographic studies varies according to the expectations of a particular study. Many geographical models assume a uniform pattern of spatial data, particularly if they are using areal units as measures for the model (e.g., demand measures in location models). The examples given show that assumptions are often made concerning the spatial patterns of geographic data. In contrast, this research examines spatial pattern without any explicit assumptions. Of particular interest is measuring patterns within aggregated census data units.

The methods presented here should prove useful to many researchers and practitioners, if only to bring to the forefront the necessity to check for pattern. Spatial autocorrelation techniques are known to be useful in this regard, though they are often used to check for underlying patterns which are deemed nuisances in geographic modeling. The MST method should provide an additional tool for spatial analysts. The potential uses of this tool range from testing the validity of assumptions about underlying patterns in location
models using areal demand units to helping to define meaningful regions. This research focuses on the spatial pattern of a specific variable (nonwhite population density) measured over a large region. The possibility of using the MST method, perhaps in conjunction with other measures such as Moran's I, to examine patterns over smaller areas presents an exciting challenge. Application of these techniques to defining regions, whether they be political or school districts, or for determining the patterns within predefined regions, such as areal demand units, will provide further opportunities to apply these methods. Furthermore, this research provides added information about the use of aggregated census data, particularly in regard to examining spatial pattern.
CHAPTER II
LITERATURE

2.1. Introduction

This research seeks to compare the usefulness of a graph theoretic approach to the spatial autocorrelation statistic Moran's I for examining spatial pattern. In so doing, it takes explicit notice of the effects of aggregation upon detection of spatial pattern. While there is little in the literature which points directly to the use of graph theoretic approaches to examining spatial pattern, the basic issue of spatial pattern is intricately tied to several related questions in the literature. These include questions related to aggregation error, aggregation and clustering techniques, and pattern recognition and cluster definition. This chapter opens with a discussion of the effects of using aggregated data in research, which leads further to a discussion of how best to aggregate data including some examples of clustering techniques. The detection of clusters and other patterns is also discussed, with particular focus on the applications and methods used. This is followed by a brief discussion of applications of graph theoretic methods, most specifically the minimum spanning tree. Finally, since this research applies the methods to detecting patterns of residential segregation, the chapter concludes with a more detailed examination of spatial segregation issues and the traditional methods used to examine such segregation.
2.2. Aggregation Error

The use of areal data, such as census tracts, is common with many types of geographic studies as well as studies in the other social sciences. The values attributed to these areal units of study are based on aggregation of individual data values and should not be divorced from broader issues related to aggregation. An important question related to the use of aggregate data concerns the possible error introduced through their use. This potential for error exists for the location modeler interested in locating facilities relative to aggregated units of demand, to the sociologist or urban geographer interested in studying the social ecology of areas, and to the spatial analyst concerned with proper interpretation of pattern. In short, any research using census data, or any other aggregation scheme, is subject to the possible effects of that aggregation.

The topic of aggregation error is important to this research for two reasons. First, the methods described here use census data, the census being a common source of data. Census data are aggregated and it has been hypothesized that the level of aggregation can be important in discerning pattern using the methods described here. It is anticipated that this research will contribute to the general body of knowledge about aggregation effects. Second, while this research examines spatial pattern over a broad area and does not specifically consider patterns within smaller geographical units, an important future step would be in applying these methods with smaller units of study. This will allow researchers to more fully investigate assumptions made about the pattern of data within areal units such as census tracts.

Many studies have investigated data aggregation error in relation to various methodologies, and the topic is not new to researchers. Hannan (1991) provides an overview of the aggregation problem for social scientists. He focuses particularly on aggregation effects on
parameter estimates in linear models. Gehlke and Biehl (1934) early on pointed out the results of using aggregated data on correlation coefficients. They aggregated census tracts into groups of ever increasing size (thus decreasing the number of groups). Calculations of correlation coefficients for the various aggregation levels resulted in increasing $r$ values. They conclude that their "...results raise the question whether a correlation coefficient in census tract data has any value for causal analysis (p. 170)." Clark and Avery (1976) follow the same line of questioning when they examine the effect of using "macrolevel" data in comparison to expected results from "microlevel" data for correlation and regression analyses. Increasing levels of aggregation (or grouping of census tracts based on proximity) cause the correlation and slope coefficients to increase, though the pattern is not constant at all aggregation levels. They attribute the inconsistencies with the amount of spatial autocorrelation, or covariation of the data, present at different aggregation levels. Clark and Avery (1976) further point out the difficulty and danger of making inferences from data used at higher levels of aggregation. The term "ecological fallacy" refers to the mistake of inferring from analyses at higher levels of aggregation to lower levels (Robinson, 1950, in Clark and Avery, 1976).

Openshaw (1984) also examines the effects of aggregation on various statistics, by using aggregated data from the U.K. and disaggregate data from Italy. He examines aggregation effects in correlation, factor analysis, regression, and clustering or classification techniques. His findings suggest that there are varying degrees of severity of aggregation effects for different methods and that "...classification methods present the greatest opportunities for constructing ecological fallacies by inference (p. 28)." He concludes by calling for more studies of the effects of aggregation on various statistical methods.
Hillsman and Rhoda (1978) and Goodchild (1979) examine the problems associated with data aggregation for location modelers. In general, this aggregation refers to creating aggregate demand points for location analysis. In practice, this would include the use of census tract centroids, for example, as the locations of demand points. Hillsman and Rhoda (1987) identified three sources of error in the $p$-median problem, both direct and indirect, which can result from aggregating demand points. These formed the basis for later work by Casillas (1987), Current and Schilling (1987), Current and Schilling (1990), Daskin et al (1987) and others. The following description of these aggregation error types is based on the discussion by Current and Schilling (1987).

Source A, B, or C errors can arise from aggregation of demand points (Current and Schilling, 1987). Source A errors arise because distance is measured from an aggregate point rather than the actual disaggregate points. This results in misrepresentation of the true travel costs in the cost function. If a facility location is chosen based on distance from an aggregate point, the cost function for any disaggregate point assigned to that facility might be overestimated if the disaggregate demand point is closer to the facility than is the aggregate point. Conversely, the true cost may be underestimated if the disaggregate demand point is further from the facility than is the aggregate point (Current and Schilling, 1987).

Source B errors are considered a special case of the Source A error type (Current and Schilling, 1987). They occur when facilities are located at aggregation points. This causes a loss of locational information and obscures the distances and the subsequent weighted costs of the facility location. When a facility is located at an aggregate demand point, the travel cost is necessarily underestimated because the distance to the aggregate demand point
from the facility is zero, while the actual distances from the disaggregate points to the facility is greater than zero (Current and Schilling, 1987).

The final error type, Source C errors, occur when a demand point is allocated to the wrong facility because of the aggregation scheme (Current and Schilling, 1987). For example, a disaggregate demand point may be allocated to a nearby aggregate point, and that aggregate point may subsequently be assigned to a facility. However, the disaggregate point may actually be closer to a different facility than is the aggregate point. Therefore, Source C errors will likely result in an overestimation of weighted travel cost, assuming that the disaggregate demand will actually frequent the closer facility. Current and Schilling (1987) refer to this as an indirect consequence of aggregation.

These error types have been discussed in relation to the $p$-median problem, which was investigated by Casillas (1987) and Current and Schilling (1987). Current and Schilling (1990) and Daskin et al (1987) have also examined these errors in relation to covering models. Current and Schilling (1990) suggest that covering problems have more potentially serious problems related to data aggregation, because the distance is essentially converted to binary distances (Church and Revelle, 1976, in Current and Schilling, 1990), where the distance is zero if it is within the covering distance and infinity otherwise.

The three types of aggregation error can be related somewhat differently to the covering problem. In the covering problem, there are two possible Source A errors, Case 1 and Case 2 (Current and Schilling, 1990). Case 1 occurs when a disaggregate demand point is considered to be covered (assigned to a facility) because the aggregate demand point to which it is assigned is covered. The disaggregate point might not actually be covered. In Case 2, a disaggregate demand point is not counted as being covered, when in fact it is
covered. Source B errors are similar to the $p$-median problem, in that an aggregate demand point is used as a facility. Therefore, it is possible that a disaggregate demand point may be considered covered (since it is assigned to the aggregated point), but may actually be beyond the covering distance of the facility. Current and Schilling (1990) find no analogous Source C error for covering problems, since demand points are either in the covering area or they are not.

The point is not to belabor the types of aggregation error found in location models, but rather to show how aggregation can cause a loss of locational information and possibly introduce error into the models. This results in errors either in terms of chosen locational patterns or in the value of the objective function, though there is lack of agreement on which of these types of errors is most affected by aggregation (see, for example, Goodchild (1979) versus Casillas (1987)). Hillsman and Rhoda (1978) point out that errors will likely be larger for irregularly shaped areas, such as census tracts. Love, Morris, and Wesolowsky (1988) provide a method for using areal demand units and rectangular distance to locate facilities, but this assumes a uniform distribution over that area, which they state is not "overly restrictive (p. 44)." Both Current and Schilling (1990) and Daskin et al (1987) provide aggregation approaches for minimizing the aggregation effects in covering models. Current and Schilling (1987) provide methods for alleviating aggregation errors in the $p$-median problem. These are discussed in more detail below. A specialized form of aggregation relates to the issue of creating "optimal zones", which is also discussed in the following section.

An important aspect of aggregation, in terms of this research, is how it affects pattern interpretation in general. The issue of scale is not unimportant when trying to assess spatial pattern and is closely related to the previous discussion of data aggregation. The subject
arises particularly in reference to spatial autocorrelation measures (Goodchild, 1986; Upton and Fingleton, 1985; Cliff and Ord, 1981; Chou, 1991) where the scale of the study or of the spatial units themselves can obscure patterns and even extends to interpretation and classification of images in remote sensing (Woodcock and Strahler, 1987). Woodcock and Strahler use graphs of "local variation" of scenes to investigate how size and spatial arrangement of objects react under different resolutions, and how these graphs may aid in selecting appropriate scales for study.

De Cola (1991) uses fractal analysis to measure the effects of scale on spatial autocorrelation techniques for point data. A fractal is "...a phenomenon that does not change in form or appearance across scales...(p. 546)". Using quadrat counts of point data at different scales and join-count statistics, De Cola shows that fractals can be used to determine the effects of different scales on measuring patterns among point objects. Allen and Turner (1992) examine the use of tract scale census data for depicting ethnic patterns and Chou (1991) illustrates how scale across a regular region can influence Moran's $I$, finding that the index decreases as resolution decreases. This is an important aspect of this research, as the methods presented are tested at different levels in the census hierarchy. Alexander et al (1991) dealt with the general issue of scale explicitly in trying to devise a method to identify small clusters of cancers. They suggest that larger scale (scale referring to the overall size of the region in question) study areas make identification of clusters easier. As such, they examine clusters at both a within groups scale and a between groups scale.

Finally, Odland (1988) states

...sizes and shapes limit the scale of the patterns that can be detected. Patterns that are manifest at small scales will not be evident when the information is aggregated
into larger areas. The average resolution of a set of subregions is defined as the square root of the average of their areas (Tobler, 1984) and pattern elements less than twice this size cannot be detected by examining data for the subregions (p. 27)."

This represents clearly one problem created by aggregation. As aggregation increases, underlying patterns are likely obscured and the entire issue of pattern detection and definition becomes even more complex.

2.3. Aggregation and Cluster Construction

Since aggregation is a necessary component of much geographic and social research, some have suggested that aggregation error may be reduced by using appropriate aggregation schemes. This was the approach used by Daskin et al (1987) to test aggregation schemes and by Current and Schilling (1987, 1990) to introduce aggregation "rules" to alleviate error in location modeling problems. Current and Schilling (1987) suggest an improvement in measuring aggregate distances to facilities to eliminate Source A and Source B errors in the p-median problem, though this will not eliminate Source C errors. Their improvement requires measuring the sums of the true distances between all of the units aggregated to one point and the facility. Their findings indicate that this measure reduces error and leads away from a bias toward higher numbers of facilities.

Current and Schilling (1990) also introduce aggregation "rules" for dealing with error in covering problems. Their rules state that: (1) only disaggregate demand points should be considered as potential aggregate demand points, (2) a disaggregate point should not be aggregated to any point that is more distant than the covering distance, and (3) a disaggregate demand point should be assigned to an aggregate demand point only "...if the set of potential facility sites which cover each of them are identical (p. 122)." The first two
rules will eliminate some errors, while the third rule will eliminate all aggregation error types in the covering problems.

Daskin et al (1987) test three aggregation schemes at the national level to alleviate errors in the maximal covering problem. These schemes are based on the method of selection of the aggregate points. The first aggregation scheme is based on the relative demands of the disaggregate points. Those points with the highest demand are selected as aggregate points. The second scheme is sequential and is based on the distance between disaggregate points. After initially choosing the point with highest demand as an aggregate point, the scheme then chooses the farthest disaggregate point as the next aggregate point and so on. The third scheme considers both the demand and the distances and is a combination of the first two methods. Here, the "next" aggregate point is not the farthest, but rather the point with the largest product of demand and distance from the current set of aggregate points. Their findings suggest that the first and third approaches are most useful in reducing aggregation "coverage or optimality errors (p. 32)."

Goodchild (1979), while recognizing that the modeler may have no choice in how the data are aggregated, also suggests that some forms of aggregation are safer than others. He reports Openshaw's (1977) argument that, since zones are often created for certain purposes, they will have built in bias and suggests examining how zoning and aggregation level may cause bias. Rodriguez (1991) also discusses Openshaw's (1977) "optimal-zones" and suggests they be constructed using the smallest aggregation units possible to avoid error. This leads to a special form of aggregation, that of creating appropriate zones.

While the effects of aggregation for location modelers and their responses are discussed above, others have argued that common forms of areal data are not properly aggregated. It
is important to understand that this concern is not just in terms of pure aggregation, or what geographers might consider scale, but also in terms of the actual zoning system used (Openshaw and Taylor, 1981). In other words, the zones used to aggregate the data may not be suited for a particular research purpose. Openshaw (1984) states that

> Important geographical issues, such as the purposeful definition of meaningful areal units for reporting census data and the magnitude of aggregational biases and ecological fallacies that characterise particular sets of areal definitions, have not been thoroughly investigated or indeed subjected to much study (p. 17).

In fact, Openshaw suggests that "...zone design should be viewed as a spatial complement to model calibration and the statistical process of parameter estimation (1978, p. 782)."

The understanding that zoning systems are not unique and, therefore, can present different results under differing zoning systems, is referred to as the modifiable areal unit problem (Openshaw and Taylor, 1981). Anselin (1988) describes the problem as arising from

> the fact that statistical measures for cross-sectional data are sensitive to the way in which the spatial units are organized. Specifically, the level of aggregation and the spatial arrangement in zones (i.e., combinations of contiguous units) affects the magnitude of various measures of association, such as spatial autocorrelation coefficients and parameters in a regression model (p. 26).

This concern about proper zonal aggregation includes standard census areas, as well as construction of many other types of zones (Openshaw and Taylor, 1981). Openshaw and Taylor (1981) point out that the modifiable areal unit problem has a long history in electoral districting. Along these lines, Morrill (1981), in his discussion of political districting, points out that, while compact districts are desirable, the patterns within them should be considered. A geometrically compact district may have a very uneven population distribution within it. This type of zonal aggregation is somewhat different in that it is based on population alone, although efforts are sometimes made to create specific "types" of districts (based on racial composition for instance). Interestingly, he suggests, for this
type of districting, starting with larger units and refining them, as opposed to Rodriguez' (1991) approach of starting with small units and building up to create optimal zones. Again, a distinction must be made between mere aggregation of zones and the actual choice of zoning systems, though the two are related and present similar problems.

Openshaw (1977, 1978) uses examples of a linear regression model (1978) and spatial interaction models to show how a zoning scheme can affect results. He found that, in spatial interaction models, the choice of zoning system had more effect than was originally expected (1977). In his linear regression example, he attempts to devise "optimal zones" using different zone design criteria (1978) and concludes by stating that, while the zoning system can affect results, and while it may be possible to create the proper zoning system, there is "...no simple or general-purpose solution to the problem. There is a great need for theoretical studies to identify meaningful zone-design criteria for specific models, but the problem is proving to be highly intractable and more amenable to empirical study (p. 793)."

The difficulties in creating optimal zones are apparent in Batty and Sammons' (1978) attempt to use an information theory approach to create zones which best reveal the underlying "pattern" of the data. They attempt to aggregate the zones in such a way that the loss of information from aggregating is more than offset by the creation of zones approximating equal areas, thereby enhancing the information available. The results of their study, which was applied to seven urban areas, using varying spatial configurations, indicate that attempts to increase information do not work. Their information measure did not improve the aggregation scheme. Rather, "...the original 274-zone system is the most informative; and in this sense, the most efficient, and hence optimal (p. 776)."
More recently, Rodriguez (1991) approached the issue of zoning aggregation in his attempt to use GIS and database models to relate aggregated units to one another. He suggests that, in aggregating zones, it is important to start with a low level and build from there to create larger, but more homogeneous, areal units. As he puts it, "...technology has eliminated the need to rely upon pre-existing political boundaries for reporting geographic information (p. 46)." This goes beyond Openshaw's (1978) statement that "...there can no longer be any reasonable excuse for the researcher who continues to use spatially aggregated data in a naive fashion by failing to investigate the empirical consequences of choosing a particular zoning system (p. 793)."

However, many researchers and practitioners use aggregated areal data, particularly census data, with zones which are created without their specific application in mind. As Hannan (1991) writes, even simple problems such as missing data at appropriate aggregation levels (for instance, block level data withheld for privacy) may force a researcher to use a different aggregation level. Openshaw and Taylor (1981) strongly oppose the continued use of standard or arbitrary zones and attribute their continued use partially to the insistence of geographers to adhere to "normal science", and their fear "...that to acknowledge its [the modifiable areal unit problem] existence would cast doubts on the acceptability of nearly all applications of quantitative techniques to zonal data (p. 67)." Yet they also recognize the usefulness to many researchers of using standard areal data (they also mention the confidentiality issue) when they state that "...it is hardly surprising that the vast majority of applications of quantitative techniques have used zonal data. If the problem of studying data for modifiable units cannot be avoided, then it is essential that the consequential limitations of such studies are clearly understood (p. 60)." The argument is that those using aggregated census data must avoid "naive" mistakes in interpretation. This is an important issue and is raised in this research as well.
The issue of appropriate aggregation methods for minimizing error or ecological fallacy is clearly related to the creation of groups or clusters. Cluster creation is relevant partially because clusters on the landscape represent spatial patterns of interest to those who expect clustering as well as those who see clustering as a deviation from uniformity assumptions. Just as Daskin et al (1990) and others are interested in how best to aggregate demand data or other data, cluster analysis is concerned with how best to create groups of objects (be they points representing tree species, traffic zones, households, or flowers). Masser and Brown (1975) present two possible methods for creating zones for spatial interaction modeling. They apply hierarchical aggregation schemes based on Ward's (1963) heuristic to create optimal zones for labor market areas and "migration subsystems" in Merseyside and Greater London, respectively. Their aggregation scheme is based on similarities between pairs of data units, which is the basic approach used in this research. Hirst (1977) presents a similar approach which takes into account different magnitudes of row and column totals in the interaction matrix during the hierarchical grouping procedure. Gower and Ross (1969) and Johnson (1967) offer clustering schemes applied to such diverse topics as biological taxonomy and psychology. As with the above studies, these schemes are based on similarity between data pairs, with Johnson (1967) using hierarchic clustering and Gower and Ross using information from the minimum spanning tree of the distance between cluster members. Jardine and Sibson (1968a) discuss the uses of both hierarchic and nonhierarchic clustering schemes based on dissimilarity between the objects to be clustered.

Jardine and Sibson (1968b) further advocate a nonhierarchic clustering scheme for purposes of biological taxonomy. The nonhierarchic clustering schemes allow overlapping clusters, unlike hierarchical approaches. The problem is to find a balance between the more
informative overlapping (nonhierarchic) group of clusters and the less complex, nested (hierarchic) group of clusters. More recently, Openshaw, Sforzi, and Wymer (1985) use a nonhierarchic taxonomic clustering routine to create classifications of Italian households. They are interested in using individual data to create more informative clusters of households than might be found with other aggregation schemes. Their nonhierarchic clustering method created groups of 14 and 30 clusters, based on dichotomous variables. They suggest using these types of clusters to provide additional information about household variability within regions, which is unlike standard uses of census data. This relates to the concept of creating more informative "optimal zones" discussed above.

2.4. Clusters and Pattern Detection

The opposite issue of how to create clusters is that of how to recognize clusters in data units. This is closely related to the research topic here, as one goal is to discern spatial patterns over areas. There are many approaches to discerning clusters or pattern detection and the issue is one of interest to a broad range of researchers. The detection of clusters may apply to points or to values spread over an area, though Upton and Fingleton (1985) point out that such values "ostensibly" represent another type of point pattern. Therefore, approaches used to detect pattern overlap between those concerned with point patterns and those concerned with areal patterns. This research is primarily concerned with the patterns exhibited through areal data.

Rogers (1974) discusses the use of quadrats for point pattern analysis (see also Upton and Fingleton, 1985 and Cliff and Ord, 1981). Quadrat analysis involves dividing the study region into grids of equal size and counting the numbers of points in each grid. Comparison of grid counts provides information about the point pattern exhibited, since different patterns (based on different processes) are expected to exhibit different
distributions of grid counts. One difficulty with quadrat analysis is that placement of and size of the grids used influences the findings (Rogers, 1974). This relates to the earlier concern with optimal zones and the modifiable areal units problem.

Another approach to analyzing point patterns is referred to as nearest-neighbor analysis, wherein the distribution of distances between neighboring points is used to indicate pattern (see Rogers, 1974; Cliff and Ord, 1981; Upton and Fingleton, 1985). Many applications of cluster and pattern detection methods are in the realm of the physical sciences. Examples given in Upton and Fingleton (1987) and Cliff and Ord (1981) are often concerned with detecting clusters of biological specimens such as trees and birds. Pielou (1961) shows how nearest-neighbor statistics can be used to discern pattern and, subsequently, interrelationships between two species of plants. That approach considers not only the distance to the nearest neighbor, but the "type" of the nearest neighbor, by comparing the expected and observed probabilities of nearest neighbor pairings of two species. Similar approaches, based on distance, have been used to examine interdependence between types of firms (Lee, 1979; Okabe and Miki, 1984). Other common applications and variations on these methods are in medicine, where attempts are made to detect significant clustering of childhood leukemia (Cuzick and Edwards, 1990) and small clusters of malignancies (Alexander et al, 1991). Alexander et al (1991) present a nearest neighbor technique based on areal units rather than points (the nearest neighbor area method) to investigate clustering both within and between areas. In general, these types of techniques are used not just to find clusters or patterns, but to understand the underlying processes which create those patterns.

Another form of distance based analysis of point patterns, second-order analysis, was developed by Ripley (1977) and is referred to as second order because it uses the variance
rather than the mean to examine point pairs within specified distances (in Getis and Franklin, 1987). Getis and Franklin (1987) expand on second-order analysis with their second-order neighborhood analysis, which examines neighbors of individual points, using specified distances to define neighboring points. This work evolved into the \( G \) statistics, which can be used with areal data values when they are assigned to points, and can examine values for all points within a given distance of other points or only those within a given distance of a particular point (Getis and Ord, 1992). Again, these statistics are based on distances between neighboring points.

Upton and Fingleton (1985), Cliff and Ord (1981) provide good reviews of the techniques used to discern spatial point patterns, and also provide reviews of methods for areal pattern analysis. A set of techniques for detecting spatial clustering and other patterns with areal data are termed spatial autocorrelation techniques and many examples are offered in the realm of biological or medicine sciences applications. A discussion of these techniques is presented in more detail in the following chapter, as it is one of the approaches used in this research to discern spatial pattern.

Filliben, Kafadar, and Shier (1983) use cluster techniques to determine the spread of materials across a disk created from mixing two materials. One of their approaches uses a minimum spanning tree in the same fashion as Shier (1982) and represents the basic technique presented here. Zahn (1971) uses a similar innovative approach to examining specific point patterns, focusing on applications in biological taxonomy. Particularly interesting to this research are those cluster detection applications where graph theoretic techniques are used (Shier, 1982; Zahn, 1971; Gower and Ross, 1969; Royaltey et al, 1975). Gower and Ross (1969) actually show how clusters can be formed as is discussed above, but their technique is directly related to cluster detection as well.
2.5. Graph Theoretic Applications

The work by Zahn (1971) and others discussed above in relation to cluster detection provides one example of how graph theoretic concepts can be used in diverse applications. Some basic concepts of graphs and graph theory are provided in the following chapter, but it is interesting to briefly examine previous applications of these techniques. Phillips and Garcia-Diaz (1981) provide a comprehensive overview of various graph theoretic techniques as well as examples of applications of those techniques. While it is beyond the scope of this research to enumerate them, those using minimum spanning trees are of special interest.

Some graph theoretic applications are already discussed in the previous section. These include Zahn's (1971) application of minimum spanning tree algorithms for detecting clusters in a biological taxonomic application and Filliben, Kafadar, and Shier's (1983) and Shier's (1982) use of minimum spanning trees to determine how well two materials mix, on which part of this research is based. Shier's basic approach is discussed in more detail in Chapter III. Royaltey, Astrachan, and Sokal (1975) also test for patterns in two-dimensional space using a graph approach. They depict a region as a graph or network, calculate expected statistical values for edge lengths in the graph, and compare these with the actual values found for a given series of patterns. They are able, with some success, to discern certain types of patterns using this method. Their method for depicting regions as graphs is essentially the same as that provided in Shier (1982), Filliben, Kafadar and Shier (1983), and presented in this research and is discussed in more detail in the following chapter.
In other applications, Held and Karp (1970; 1971), use a minimum spanning tree approach to attempt to solve the traveling salesman problem (actually, they use a "1-tree", which is a variation of the minimum spanning tree). Saltman, Bolotsky, and Ruthberg (1973) use the minimum spanning tree as the first step in a procedure for optimizing the network of telephone connections based on miles over the network, rather than calls made over the network. The minimum spanning tree is used in initially setting up the graph or network of cities to be served. From there, a group of routines, including shortest path algorithms, are used to further optimize the entire network. Phillips and Garcia-Diaz (1981) also comment on Dei Rossi, Heiser, and Kings' (1970) application of a minimum spanning tree for optimally connecting television stations in a network.

These represent but a few of the possible applications of graph theoretic techniques to common problems, particularly those involving optimization. A more detailed discussion of minimum spanning trees and graph theoretic descriptions are presented in the following chapter.

2.6. Homogeneity, Clustering, and Pattern Definition

The discussion so far has centered on aggregation issues and techniques for creating and discerning clusters, which represent one form of pattern. However, it is important to look at how pattern is defined in the literature, particularly in relation to the methods used here. Different assumptions about expected spatial patterns can lead to different measurement approaches and pattern definitions. For example, existence of spatial clusters in the data deviates from some models' assumed uniformity of data spread. This is the assumption made by Shier (1982). In his research, he measured the "homogeneity" of disk materials. He described a checkerboard pattern (high values next to low values which are, in turn, next to high values) as homogeneous. Therefore, Shier identifies homogeneity with
uniformity, or lacking clusters of high or low values. In commonly used measures of spatial autocorrelation, this is considered negative spatial autocorrelation. On the other hand, a random distribution indicates a lack of spatial autocorrelation, while Goodchild (1986) describes positive spatial autocorrelation as "...when black and white cells cluster together into homogeneous regions (p.4)." These different terms for the same pattern introduce an element of confusion into pattern description and detection.

Filliben, Kafadar, and Shier (1983) provide added explanation of this definition of homogeneity. In terms of a material, where mixing has occurred, they write

Loosely speaking, a surface in which the percentages [of mixed materials] are highly variable within small areas may be considered "homogeneous," if such highly-varying subareas occur consistently throughout the material. In this regard, one may more easily characterize the lack of homogeneity as distinct patches of consistently small/large values. In trying to achieve homogeneity by mixing, a few number of large sparsely or densely populated patches would suggest insufficient mixing, whereas a large number of small patches would indicate that homogeneity had been more or less achieved (p. 169).

The terms homogeneity and heterogeneity in relation to spatial data are used throughout the literature in differing ways and add to the complexity of examining different spatial patterns. If we accept Shier's (1982) description of homogeneity, we equate homogeneity with uniformity. This is in keeping with Alexander et al (1991) who seem to equate heterogeneity with non-uniformity, and Cuzick and Edwards (1990) who call a non-uniform density inhomogeneous. Royaltey et al (1975) refer to their "crazy quilt" pattern as uniform and only one type of deviation from expected randomness. Figure 1 shows a graphical representation of the different terminology used to describe the same pattern.

Accepting the assumption that homogeneity is the same as uniformity is somewhat counter-intuitive. Intuition would say that, as shown in Figure 1, areas with similar data values are
- "checkerboard"
- "crazy-quilt"
- homogeneous
- uniform
- negative spatial autocorrelation

- clustered
- inhomogeneous
- positive spatial autocorrelation

intuitive homogeneity

- high values
- low values

Figure 1. Pattern Descriptions.
homogeneous. This is a common interpretation of homogeneity, but is the opposite of the form of homogeneity previously described. This represents just one possible pattern, yet clearly demonstrates the complexity inherent in defining and describing spatial pattern.

Another problem with developing a more intuitive feel for the above definitions of homogeneity relates to the scale of the area under consideration and the level of aggregation. This research is concerned not just with spatial pattern, but with the effects of using different census aggregations. Shier (1982) and Royaltey et al (1975) used differences in neighboring areas' values to discern pattern. Both assumed that small differences indicate similar neighbors and large differences indicate dissimilar neighbors. Similar neighbors indicate deviation from uniformity or, more specifically, clustering. The problem is that, particularly for human systems, it is unlikely for a small area to exhibit the type of homogeneity described. Homogeneity in a small area is likely to arise from similar values of the variable under consideration. Large deviations in a variable are less likely within a census block group than in a group of census tracts or a city. This is what leads Rodriguez (1991) to suggest using the smallest aggregated units to devise optimal-zones, as they are more intuitively homogeneous. If we go back to the definition by Filliben, Kafadar, and Shier (1983), we can equate their "homogeneity" with a well-mixed (at a small scale) area. For an urban application such as this research uses, this would imply that homogeneity occurs if dissimilar people live in close proximity to one another. This is not the generally accepted implication of homogeneity.

2.7. Segregation and Related Indices

While this research considers the issue of spatial pattern in general and without preconceived expectations, there are many possible applications of techniques for discerning pattern. One important application is in examining residential segregation. This
research uses the spatial autocorrelation and minimum spanning tree approaches described in the following chapter to examine segregation patterns in Franklin County, Ohio. Sociologists, geographers, and other social scientists are concerned with levels of segregation, whether in terms of race, income, or ethnicity in urban areas and the possible effects of that segregation. Segregation has even been studied in relation to plant species (Pielou, 1961), where, just as in human systems, the purpose is to discern underlying processes indicated by segregation patterns. Along these lines, White (1984) makes a distinction between "sociological segregation", or "...the absence of interaction among social groups (p. 1009)" and "geographic segregation", or "...an unevenness in the distribution of social groups across physical space (p. 1009)." It is this geographical segregation which creates patterns in an urban area, and with which this research is concerned.

White (1984) suggests that, while the existence of neither social or geographic segregation necessarily coincides with existence of the other, the two types are most likely correlated. He further states that spatial or geographic segregation may be unrelated to social segregation, which is of underlying concern, but

...spatial proximity is probably important. How far other people live from you and whether they are black or white, rich or poor, is likely to make a difference in the character of your urban social life. Persons you may not speak with daily may frequent the same stores and parks, and their nearness may affect the value of your home (p. 1009).

Therein, he states the crux of the matter. It is often assumed that geographic proximity or lack thereof affects how society functions, as demonstrated by studies examining the effects of segregation on the ability of segments of society to compete for jobs (Kain, 1968; Jencks and Mayer, 1990). Hence, planners and policy makers might respond with
deliberate geographic attempts to create social integration, such as the desegregation of schools (White, 1984).

The manner in which segregation occurs can vary. Clark (1991) quotes Schelling's (1971) contention that "...people get separated along many lines and in many ways. There is segregation by sex, age, income, language, religion, color, taste...and the accidents of historical location (1971, p. 143)." Schelling goes further to state that such segregation may arise from "discriminatory" behavior, which he describes as an "awareness" of others' differences that affects residential location decisions (in Clark, 1991). His definition of awareness is based on individual behavior and does not include organized discriminatory behavior (Clark, 1991) such as loan practices or real estate transactions. Stearns and Logan (1986), discussing Burgess (1925) and Hawley (1950, 1956), describe the "traditional ecological model (p. 29)", which states that residential segregation results from competition between blacks and whites for desirable housing locations, and that this results in "Segregation [as] part of the normal life cycle of a metropolitan area, a manifestation of individual affinities expressing themselves in a free market place (p. 29)." More recently, the concept of a dual housing market has been proposed, where blacks and whites are not operating in a free market place, but rather blacks operate under a different set of rules than those used for whites (Stearns and Logan, 1986).

The issue of segregation patterns, particularly of racial or ethnic groups, cannot be divorced from those broader issues that influence and are influenced by the levels of segregation found in a society. Massey and Denton (1987), in their discussion of residential segregation pattern trends of ethnic groups, discuss five sets of historical changes which have affected segregation in American cities in various ways. They include "Lessening prejudice against blacks and other minorities, the ongoing impact of civil rights legislation,
and the rise of the black middle class... (p. 803)” as having positive impacts on reducing segregation. Conversely, "...rapid immigration and metropolitan decentralization... (p. 803)” likely had the opposite effect of actually increasing segregation.

These segregation issues can be tied to other social issues, such as black unemployment. Jencks and Mayer (1990) examine whether segregation affects employment levels of blacks. They are essentially responding to the "spatial mismatch" that Kain (1968) suggests resulted from blacks' segregation in the inner-city and the growth of jobs at the periphery of cities and which others still view as a reason for joblessness among inner-city blacks (Jencks and Mayer, 1990). Their findings are mixed, in that segregation likely reduces employment opportunities for blacks in white areas, but also increases employment opportunities for blacks in black areas, with no clear-cut net or aggregate result. Other factors must be considered, such as employers' attitudes toward hiring blacks, the public's attitudes toward dealing with members of other racial groups, employees' attitudes about working in areas not dominated by their own racial group, and how possible differences in firms' productivity affects employers' preferences for hiring black versus white laborers (Jencks and Mayer, 1990). Segregation, even after three decades of concerted policy efforts and increased awareness, still exists, still influences the American urban pattern, and is likely more complex than ever.

The literature on segregation follows two basic lines: segregation based on income (or examination of "poverty areas") and segregation based on race or ethnicity. The former studies (see for example Massey and Eggers, 1990; Greene, 1991; Jargowsky and Bane, 1990) often require a priori definitions of what constitutes poverty or poverty areas and also often incorporate racial or ethnic data. Clark (1991) states that "...income is correlated both with race and with residence, it is not surprising to find that even if residential choices
were unconstrained by racial residential preferences, ethnic groups would not be distributed randomly in the city (p. 4)."

There is great overlap between those studies that consider segregation based on income and those that consider racial or ethnic segregation. It is not surprising, given this, that many of the same types of techniques are used to analyze these various forms of segregation. These techniques are essentially descriptive, while those using them are seeking to analyze the processes underlying the patterns. While other measures have been proposed and used to examine segregation as a form of spatial pattern, the most common is the index of dissimilarity (White, 1984). Massey and Denton (1987) refer to the index of dissimilarity as a way to measure the "evenness" of a distribution, where "Evenness is maximized and segregation minimized when all tracts have the same relative number of minority members as the whole urban area (p. 805)." Stearns and Logan (1986) also refer to this as a measure of unevenness. This concept of using the index of dissimilarity for measuring evenness or the distribution of two groups is easily applied to both ethnic/racial segregation (Massey and Denton, 1987) or to income/poverty segregation (Massey and Eggers, 1990), though Massey and Eggers (1990) suggest that the Gini coefficient is used more often when examining income differentials.

The index of dissimilarity is given by:

\[ D_{xy} = 0.5 \sum \left| \frac{x_i}{X} - \frac{y_i}{Y} \right| \]  

\[ (1) \]

\( X \) and \( Y \) are the area-wide totals for each of two groups, and \( x_i \) and \( y_i \) are the members of each group in a given subarea, such as a tract (Massey and Denton, 1987). The \( D \) value is high (close to 100) when there are high levels of segregation as "...when some communities
house a disproportionately large percent of a region's blacks while others house a disproportionately large percent of a region's whites (Stearns and Logan, 1986, p. 36)."

This is also the proportion of either group which would have to be redistributed for each tract or subarea to mirror the area-wide proportions, which are used as the basis for comparison (White, 1984).

Other indices can be used to examine segregation or other types of patterns, besides the index of dissimilarity. However, the index of dissimilarity is still the most widely used. Other indices are similar, such as the index of segregation, which differs from the index of dissimilarity only in that the proportion of members of one group in an area are compared to the proportion of total members of all groups in the area, rather than to the proportion of a second group in the area. The index of segregation is interpreted in much the same way as the index of dissimilarity (Wheeler and Muller, 1981). Also used in segregation studies is the P* index, which can be used to measure the concentration of groups (Massey and Eggers, 1990), whether they be poverty groups or ethnic groups. The P* index is referred to as an "exposure" index because it measures the probability of contact occurring between members of different groups, by measuring the degree to which members of different groups will reside in the same tracts (Massey and Eggers, 1990; Massey and Denton, 1987). The P* index actually equates to four indices when two groups are being compared, because each group can be compared with itself or asymmetrically with the other group. When a group is compared with itself, this is referred to as an "isolation" index and when it is compared to the other group, this is referred to as an "interaction" index (Massey and Denton, 1987). Though Stearns and Logan (1986) find the P* index to be less useful than the related correlation ratio (which incorporates overall population composition), Massey and Denton (1987) found that they are similar enough that the P* index may be preferred for analytical ease and comparison with other studies.
Stearns and Logan (1986) suggest that other studies have shown that the index used may affect results, but Massey and Denton (1987) found that various "measures of evenness" were highly correlated with one another and, therefore, their empirical differences are slight. They use the index of dissimilarity to allow comparison with earlier studies, most of which used that index. Even though some have suggested that the index of dissimilarity is not the best index possible, there is a certain amount of inertia in its use. For instance, Stearns and Logan (1986) use the correlation ratio rather than the index of dissimilarity because the index of dissimilarity does not reflect the addition of new black population if the distribution does not change. They want to capture the effects of that added population. The correlation ratio does so by incorporating the overall composition of the total population. White (1984) finds fault with the index of dissimilarity because it does not explicitly consider the spatial arrangement of the population. He suggests a new index based on spatial proximity, which is discussed below. However, even recent studies continue to use the index of dissimilarity (e.g. Massey and Denton, 1987; Allen and Turner, 1992), suggesting inertia in its use.

The above measures of segregation involve comparing ratios to ratios and are, as White (1984) points out, basically aspatial. Some other measures of segregation, particularly those involving income segregation or poverty areas, rely on a priori definitions of what constitutes a poverty area. For instance, Jargowsky and Bane (1990) define and map "ghetto tracts" based on the rate of poverty within the tract, then compare population in ghetto tracts with the overall population for MSAs. Greene (1991), in a centrographic approach to measuring poverty concentration, defines "extreme poverty areas" in the same manner as Jargowsky and Bane and uses only the qualifying tracts in his measure, thereby incorporating a spatial aspect, but neglecting the overall urban arrangement of
"disqualified" tracts. Jargowsky and Bane (1990), while recognizing that arbitrary definitions of poverty areas might not be desirable, argue that these methods allow mapping and thus a spatial element that is missing in the standard segregation indices, where the geographic pattern is not examined and can vary for the same index value. Therefore, while some commonly used measures are aspatial, others rely on arbitrary definitions for comparison.

White (1984) responds to the arguments against the index of dissimilarity by proposing a segregation measure based on distance which, he says, explicitly incorporates the spatial aspect of segregation. While he allows that there are some advantages to using the index of dissimilarity, such as ease of computation and interpretation, he also argues that various other problems with the index (and other similar indices) make it less suitable. An important disadvantage of these indices is their lack of a spatial component, which has been mentioned above as a criticism of the index.

White's (1984) segregation statistic, $P$, incorporates the spatial proximity of the overall population, the spatial proximity of the members of one group, the spatial proximity of members of the second group, and the spatial proximity of members of the first group to the second group. Furthermore, his index allows incorporation of different distance functions when calculating the proximity values, though there is no clear indication of which is most appropriate. However, this method of including the spatial aspect is limited to distance between areas and does not appear easily transferable to other methods of representing spatial arrangement, such as contiguity. White (1984) shows that the proximity measure $P$ can discern clustering based on race for an urban area and is preferable because of its spatial dimension, yet he recognizes that the index of dissimilarity
has advantages. These advantages are likely the reason for the continued use of $D$ in the segregation literature.

While White (1984) advocates the use of the proximity measure described above, Greene (1991) responds to problems with standard segregation indices with a centrographic approach. He argues that centrographic techniques, which are statistics describing the "center" of a given population, allow more interpretative advantages than White's approach because large amounts of data are summarized and cartographic representation allows easier interpretation. According to Lee (1967), centrographic measures "...are directly analogous to concepts in univariate statistics, and many are defined in exactly the same way or are two-dimensional generalizations of univariate concepts (p. 18)" and "...are applications of these concepts to the locations of discrete objects on a plane (p. 18)." He uses six centrographic techniques to study spatial segregation of various ethnic groups in different U.S. cities. The following measures are included as examples of useful centrographic techniques:

1. **Mean Center**, a measure of the location of a population.
2. **Standard Distance**, a linear or one-dimensional measure of the spread of the population.
3. **Standard Radius**, a two-dimensional measure of the spread of the population...
4. **Principle Axes**, indication of the major direction of the dispersion.
5. **Standard Ellipse or ellipse of density**, a graphic portrait of the other measures.
6. **Coefficient of Circularity**, a measure of out-of-roundness (p. 35).

Lee (1967) states that the potential for these measures is great, but that they are most likely to be used along with other techniques, since varied circumstances will make them more or
less useful for a given study. He further points out that the methods used are purely descriptive and that careful consideration should be given before assuming underlying behavioral causes of a specific pattern. This is true as well of the research reported here.

Greene (1991) presents a recent application of centrographic techniques to segregation issues. He uses the standard radius of the distribution of "extreme poverty areas" to examine poverty concentration in 30 U.S. cities. He espouses the use of centrographic techniques for their cartographic interpretive ability and the fact that they consider the spatial arrangement of poverty areas. This is important because the actual spatial pattern of poverty areas can "...play a significant role in the degree to which the poor are isolated from the social and economic mainstream (p. 243)." His findings suggest that concentration of the poor in poverty areas is increasing and that these areas are becoming larger. One difficulty with his approach is that it does not explicitly consider the poverty areas' spatial relationship to non-poverty areas. In addition, the poverty areas are based on a pre-defined notion of what constitutes such an area. Though this method does incorporate the spatial aspect missing from other indices, it might ignore the overall spatial relationships and/or include an arbitrary assessment of poverty. In general, while centrographic techniques consider the spatial aspect of segregation more than other indices, they do not offer the flexibility found in the methods used in this research. Their measures are distance-based, which incorporates a spatial component, but they do not lend themselves to other representations of locational relationships among the areas under study. Furthermore, they are best used with graphic or mapped representation of the pattern and Lee (1967) points out that interpretation can be difficult if they are not used properly in conjunction with one another or their results are oversimplified.
Even given the number of indices available and widely used to examine segregation, it is possible to see the advantages of the spatial autocorrelation and minimum spanning tree (MST) approaches used in this research, particularly if they are seen as useful additions to the tools rather than as replacements. Both explicitly consider the spatial arrangement of the subareas under study. Furthermore, they allow a great deal of flexibility in how that spatial arrangement is incorporated and are not necessarily based on distance measures, such as $P$ and centrographic techniques. Each can include any number of ways of defining neighbors. Finally, they do not depend on an a priori definition of the subareas.

2.8. Conclusion

The issue of spatial pattern within and among areal units is broad and one of interest to many researchers. This research specifically examines the use of Moran's $I$ and the graph theoretic approach of minimum spanning tree (MST) construction to measure pattern over census units. While the properties of Moran's $I$ are well known, the MST method represents a novel approach to examining spatial pattern from a geographical point of view. To compare and contrast the use of these two methods in this research, several related areas of the literature have been discussed. Among these is the broad subject of data aggregation, important to this research because a focal point is how aggregation affects performance of the two methods tested. This discussion leads to an examination of appropriate aggregation schemes to atone for aggregation problems in geographical models. Included in this is the modifiable areal unit problem and the desire to create "optimal zones", as well as other techniques and applications for clustering.

Cluster detection is closely related to overall pattern detection. Here, various methods of determining the existence of clusters are described, particularly in relation to point patterns. A more thorough discussion of areal pattern detection, focusing on spatial autocorrelation
techniques, is presented in the following chapter. Related to the issue of pattern detection is that of pattern definition, which is discussed in particular relation to Shier's (1982) definition of homogeneity and his approach for measuring it, on which this research is partially built. Also, since this research uses a graph theoretic technique as one approach to examining spatial pattern, some other applications of these types of techniques are briefly discussed.

Finally, this research focuses on the applicability of the Moran's $I$ and MST methods for determining the degree of residential segregation of white and nonwhite population in Franklin County, Ohio. Some of the more traditional approaches to measuring segregation are presented in a general discussion of segregation issues. It is suggested that the two methods here, one commonly used in other applications and the other less tested, provide useful alternative measurements. Both techniques reach beyond limitations found in currently accepted segregation measures, including the failure to incorporate locational relationships and the necessity to predefine categories, or to consider only certain areas in a given measure of segregation. Furthermore, it is argued, both provide flexibility not found in currently used segregation methods.
CHAPTER III
METHODS

3.1. Introduction
This research uses two methods to examine spatial pattern: the spatial autocorrelation statistic Moran's I and an exploratory graph theoretic approach using a minimum spanning tree. There are other methods that can be used to examine measures of association and/or pattern across geographic areas. However, the two methods used here differ from many of those in that they both explicitly consider the data attributes and the locational relationships of all of the units in a given region or geographic study area. Therefore, they give an overall indication of pattern based on the attributes of the entire area, with explicit attention given to locational arrangements of the subareas within a study region. An important objective of this research is to determine the usefulness of the minimum spanning tree (MST) approach for measuring pattern. Therefore, the Moran's I statistic is used as a benchmark for comparison. It was chosen over other methods because of its similarities to the MST method. Both fully consider locational attributes and spatial arrangements.

This chapter consists of a brief discussion of the data used for the research. A more thorough discussion of the data and their formation for each research phase are given in their respective analysis chapters. This is followed by an introduction to spatial autocorrelation and a discussion of the Moran's I spatial autocorrelation statistic, including the steps taken to set up the program used in this research. Added attention is given to the
formation of the spatial weighting function to be used both for Moran's I and for the
minimum spanning tree (MST) method. The proposed MST method is also described,
with a simple example of how a region can be depicted as a network. Finally, some basic
and shared issues of geographic data problems related to pattern detection using these
methods are discussed.

3.2. Data
The research presented here consists of two basic steps, one exploring spatial pattern
through a constructed example and the other through application to segregation patterns in
Franklin County, Ohio. Two forms of data are used in the research, corresponding to two
phases. The first phase consists of a simulated census environment with constructed data
patterns and the second phase applies the methods to actual census data. The data
requirements and data construction for each phase are discussed in detail in their respective
chapters. This section merely serves as a brief introduction of the data on which to base a
discussion of the methods themselves.

The first phase of the research uses "constructed" data as a first step toward determining the
effectiveness of each of the two spatial methods. The data were created so as to depict two
distinct patterns at the lowest aggregation level and were also specifically designed to
emulate the census style of hierarchical aggregation, with blocks aggregating to block
groups and block groups aggregating to tracts. This allows both methods to be tested at
several "census" aggregation levels. For both the constructed and actual census datasets,
one of three possible data files is read into the program. These files contain the data at
specified aggregation levels, since the programs do not actually aggregate the data. The
data are previously aggregated using the identification numbers (which are designed to
identify blocks in the same block group and block groups in the same tract) for the
subareas. As is illustrated in the following chapter, the constructed data are aggregated in a very regular fashion due to the scheme for assigning identification numbers and data values. The constructed data are also formatted in the same way as the census data used in the second research phase. The only substantial differences between the constructed data and the actual census data are that the constructed data apply to a very regular grid-like regional arrangement and the values of the various subareas are for a simulated variable. The details of the constructed region and the data patterns are given in Chapter IV.

The second research phase, which is an actual application of the methods, uses 1990 first release census data for Franklin County, Ohio. The data were read from the STF1 tape, with the variable "nonwhite population density" serving as test variable for the two methods. In addition, locational attribute data were obtained from the 1990 Topologically Integrated Geographic Encoding and Referencing (TIGER) System files provided by the U.S. Bureau of the Census (U.S. Bureau of the Census, 1991). Full details of the data collection procedures and results are presented in Chapter V. Use of both the constructed data and the census data provide insight into the usefulness of each of the two methods described in this chapter for discerning spatial pattern.

3.3. Methods

3.3.1. Moran's $I$

The spatial autocorrelation statistic Moran's $I$ is used in this research as a comparison for the usefulness of the MST method. In addition, this research considers more fully the usefulness of Moran's $I$ itself for discerning spatial pattern, particularly in terms of the application used here: that of determining levels of residential segregation in an urban area. Using Moran's $I$ as a benchmark against which to test the MST method helps indicate not
only whether the two methods correctly discern pattern, but also shows the effect of aggregation on the ability of each to do so, and points out any differences between the methods. Moran's I was chosen because of similarities between it and the MST method. Since both methods use complete locational and attribute information for a region, both consider the spatial arrangement of values over the entire region explicitly. The intent here is not to disprove the effectiveness of Moran's I, but to use it to show how the MST approach can contribute to discerning spatial patterns in data and the effects of aggregation upon interpreting such patterns.

Moran's I is a member of a family of statistics which are referred to as spatial autocorrelation measures. Spatial autocorrelation measures are descriptive statistics that measure the similarity of objects (or areas) to nearby objects (or areas) (Goodchild, 1986). They measure interdependence between the variable attributes of nearby areas. Because they measure this interdependence, they can also be used to measure spatial pattern. Upton and Fingleton (1985) specifically define spatial autocorrelation as "...a property that mapped data possesses whenever it exhibits an organized pattern...(p. 151)." For instance, if it is found, using a spatial autocorrelation statistic, that there is a high degree of positive interdependence (or similarity among nearby areas) we can assume that the data values are clustered on the map. This type of interdependence is referred to as positive spatial autocorrelation. On the other hand, a negative type of interdependence, where neighboring values are extremely dissimilar, would signify more of a checkerboard pattern. This type of interdependence is referred to as negative spatial autocorrelation.

Spatial data interdependence is common and, indeed, forms the backbone of many geographic and regional science studies (Anselin, 1988). Quoted in both Cliff and Ord (1981) and Goodchild (1986) is Gould (1970), who stated that
Why we should expect independence in spatial observations which are of the slightest intellectual interest or importance in geographic research I cannot imagine. All our efforts to understand spatial pattern, structure, and process have indicated that it is precisely the lack of independence - the interdependence - of spatial phenomena that allows us to substitute pattern, and therefore predictability and order, for chaos and apparent lack of interdependence - of things in time and space (pp. 443-444).

Also quoted is Tobler's (1970) "...first law of geography: everything is related to everything else, but near things are more related than distant things (p. 8 in Cliff and Ord, 1981; p. 3 in Goodchild, 1986; p. 8 in Anselin, 1988)." While this may be the case, those seeking to understand or model processes and relationships often view spatial autocorrelation as a nuisance in the data and something that must be accounted for (Anselin, 1988; Odland, 1988). In Odland's (1988) discussion of spatial autocorrelation, he points out that

Spatial autocorrelation has, therefore, a dual nature. Autocorrelation, or dependence among places, is a basic characteristic of most geographic processes and most spatial distributions. Autocorrelated data, on the other hand, make it difficult to investigate these same processes and distributions by using standard statistical methods. This duality has caused autocorrelation to be widely regarded as a statistical difficulty rather than a reflection of spatial processes. It is, in fact, both of these things, and the appropriate treatment of autocorrelation will depend on objectives of a particular research project (pp. 15-16).

It is important to remember that this research is interested only in discerning spatial pattern. The intent is not to view spatial autocorrelation as it affects other forms of statistical modeling, nor is the intent to make inferences about the processes leading to the spatial patterns indicated. Instead, the research focuses exclusively upon the use of the methods presented for discerning spatial patterns. Upton and Fingleton (1985), Cliff and Ord (1981), Griffith (1987), and Odland (1988) provide comprehensive reviews of spatial analytical techniques including spatial autocorrelation, but the formulation used here is based on the presentation in Goodchild (1986).
Spatial autocorrelation measures are particularly suited to examining both attribute and distance relationships between areal data measured on an interval scale, though some can be used at the ordinal scale (Goodchild, 1986) and simple join count statistics can be used for nominal level data (Upton and Fingleton, 1985). Two commonly used spatial autocorrelation statistics for interval level data are Geary's c and Moran's I. Both are cross products of two matrices, one measuring the geographical "distance" (or proximity) between any two areas (referred to as the W matrix), and the other measuring the attribute "distance" (or proximity) between any two areas (referred to as the C matrix). The two matrices define the "map" of a variable over a region. Hubert, Golledge, and Costanzo (1981) show that both Moran's I and Geary's c belong to a family of generalized cross product statistics, which measure spatial distance as well as distance of another variable between areas. A brief review of similar statistics, such as the statistic r (a basic cross product statistic which simply consists of multiplying the attribute and locational matrices), is given in Upton and Fingleton (1985).

The major difference between the Geary's c and Moran's I statistics is the manner in which the attribute distances or differences for the C matrix are measured. For Geary's c statistic, the attribute difference between areas (c_{ij}) is calculated as the squared distance between the areas' attribute values (z):

\[
c_{ij} = (z_i - z_j)^2
\]  \hspace{1cm} (2)
For Moran's $I$, $c_{ij}$ is the product of the $i$ attribute value's and the $j$ attribute value's respective differences from the mean ($\bar{z}$) of the attribute under consideration:

$$c_{ij} = (z_i - \bar{z})(z_j - \bar{z}) \quad (3)$$

Thus, $c_{ij}$ for the Moran's $I$ statistic measures the covariance of the variable of interest at one place versus the other (Goodchild, 1986).

Moran's $I$ and Geary's $c$ offer similar insights into the degree of spatial autocorrelation for an area, but only Moran's $I$ is used here. Cliff and Ord (1981) suggest that Moran's $I$ is the more powerful of the two measures (in Upton and Fingleton, 1985). The "differences-squared form" of Geary's $c$ likely affects the distribution of that statistic more than the covariance form used in Moran's $I$ (Cliff and Ord, 1981). In addition, interpretation of the Moran's $I$ statistic is easier since the values of the index correspond more logically to the general definitional values of positive and negative spatial autocorrelation (Goodchild, 1986). For instance, a checkerboard type pattern, which would be termed negative spatial autocorrelation, is indicated by a Moran's $I$ value of less than zero. The same pattern would be indicated by a Geary's $c$ value of greater than one. While this is not overly problematic, the combination of added strength and more intuitive interpretation leads to the selection of Moran's $I$ over Geary's $c$ as the benchmark statistic for the MST method.

The method for measuring the $c_{ij}$ (attribute distances) value for both Geary's $c$ and Moran's $I$ has already been discussed. The $wij$ values, which measure the spatial proximity of any two areas or $ij$ pair, may be defined in many ways. The choice of a contiguity criterion and spatial weighting function, if used, is generally considered the most important aspect of
calculating spatial autocorrelation statistics (Odland, 1988). The purpose of the spatial weighting function is to

represent nearness and distance as consistently as possible for a set of irregular regions or irregularly spaced points. Other hypotheses about the relations among places can also be represented by spatial weighting functions, however, and the flexibility in defining the weights makes spatial autocorrelation statistics a useful means of investigating alternative hypotheses about the relations among places (Odland, 1988, pp. 29-30).

This flexibility extends to the MST approach described in this research as well. Also, the choice of the spatial weighting function is relatively more important when the purpose of the research is to derive inferences about underlying relationships between areas or to test hypotheses about such relationships. The simplest approach to a contiguity function is a binary (0 or 1) entry into the matrix, depending on whether two areas are neighbors, though other, weighted approaches can be used.

It is first necessary to define what constitutes a "neighbor" before one even considers the type of value to include in the W matrix. In other words, it is necessary to define, for each area, i, each other area, j, which can be considered its neighbor, thereby defining the ij pairs in the locational matrix. Upton and Fingleton (1985) provide a good discussion of common types of contiguity definitions based on contiguous quadrats. In other words, these definitions apply most formally to regularly spaced and shaped areas on a grid, such as the constructed data used in the first phase of this research. Tobler (1979) refers to such a grid configuration with each area having the same shape, size, and (for the non-border areas) number of neighbors as exhibiting spatial neighborhood stationarity (in Odland, 1988). Under these conditions, the rook's definition of contiguity assumes that neighbors are those quadrats or areas that share a boundary. The bishop's definition of contiguity assumes that neighbors are those areas that touch corners, making this definition useful for
those interested in "diagonal trends" (Upton and Fingleton, 1985). Finally, the queen's
definition combines the rook's and bishop's definition, but in so doing loses much of its
ability to discern pattern, as the two forms of contiguity cancel each other out. These form
the basic definitions of contiguity for regularly shaped areas and, in a simple binary
locational matrix, those ij pairs that meet the selected contiguity criterion would be indicated
by a 1 in the W matrix, while those not meeting the criterion would receive a 0.

While the above binary approach to formulating the locational matrix is straightforward, the
realities of using spatial areal data can create difficulties in using such an approach for some
types of studies. In most geographic studies, the areas under consideration are irregularly
shaped or sized, and do not meet the conditions of spatial neighborhood stationarity
discussed above (Odland, 1988), thereby clouding the issue of what constitutes a "rook's
move" for instance. In other studies, the simple binary indication of contiguity is not
precise enough for those interested in inference or model estimation for particular
applications. For instance, some studies may be concerned with the effects of spatial
interaction among areas, rather than merely the spatial pattern of a given variable. Given
this, alternative measures of contiguity (or alternative forms of spatial weighting) have been
offered.

Contiguity between areas can be measured or weighted according to the distance between
the areas (Upton and Fingleton, 1985) and may also be weighted by the amount of shared
boundary between two areas (Cliff and Ord, 1981; Goodchild, 1986). The latter approach
is appealing, but there is little reason why boundary length in and of itself should indicate a
more precise measure of proximity. A long boundary may merely indicate that two
neighboring areas are large, while their centers are actually quite distant. Cliff and Ord
(1981) account for this by suggesting a combination of the distance between neighboring
areas' centers and the proportion of boundary they share as the spatial weighting function. Another method is to use some actual measure of the distance (or the inverse of distance) between the two areas' centroids or some other point. This measure of distance may be altered for some studies to include "perceived" distance or travel time (Upton and Fingleton, 1985).

Most of these alternative contiguity measures consist of replacing the simple binary indication of contiguity with an actual value or weight indicating the degree of proximity, which is chosen based on the particular study and its requirements. The point is to not only define "neighbors", but to define degrees of neighbors, wherein closer neighbors (however that is measured) are more significant. Anselin (1988) points out that many defend the use of a "neutral" (binary contiguity) weighting system, but provides an argument against using simple binary contiguity exactly because so many possible approaches to defining neighbors can be used. However, he is specifically interested in examining the underlying processes of spatial autocorrelation and states that

In line with a model driven approach to spatial econometrics, the weight matrix should bear a direct relation to a theoretical conceptualization of the structure of dependence, rather than reflecting an ad hoc description of spatial pattern (p. 21).

For the purposes of this study, that pattern is precisely what is being investigated and the notion of binary contiguity is sufficient for determining such pattern. Furthermore, as Goodchild (1986) points out, entries into the distance matrix "...can be calculated in any suitable way (p. 17)."

There is no strong indication that a specialized spatial weighting function is required to discern spatial pattern for the applications used here. Therefore, for the purposes of this research, a binary contiguity or locational matrix is used, wherein areas are either
considered neighbors or they are not. Two different approaches for defining contiguity are used, each incorporating different aspects of geographical proximity. This combines some of the aspects of a binary matrix (with an absolute definition of contiguity) with the flexibility and deference to spatial relationships inherent in other forms of weighting. The first contiguity definition is a distance based criterion, wherein the distances between areas' centroids are calculated and the four closest areas to each area are considered its neighbors. This is analogous to the rook's definition of contiguity for regularly shaped areas, which is the situation in the first phase of this research. For instance, if the region is represented as a regular grid, then for any area, i, not on the edge of the region, its neighbors are those areas, j, immediately to the west, north, east, and south of i. These distance based "neighbors" or ij pairs are indicated in the W matrix with a binary value of 1 or 0.

The second method of determining entries into the locational matrix, the adjacency contiguity criterion, uses the TIGER/Line files to indicate which areas are adjacent to, or share a common boundary with, each area in the region. For a regular region, such as that described above, these neighbors would be the same for any area, i, as those found with the distance contiguity criterion described above, with the exception of those areas on the edge of the region. Future applications of the methods could use the lengths of the common borders to devise neighbors' "importance" or to introduce weights based on boundary length, if that is deemed necessary. This is an important point. Both the Moran's I method and the MST method discussed below are flexible enough to incorporate spatial weights in the W matrix if they are required in future applications. In this case, a binary indication of contiguity is sufficient.

The use of the distance based contiguity criterion and the adjacency based contiguity criterion are discussed further in the analysis chapters which follow. Briefly, the distance
criterion is used because it is easily applicable for any user, since the coordinates of each area's centroid are provided on census tape. Also, the first phase of research consists of constructed datasets that meet the assumption of spatial neighborhood stationarity described above and, therefore, lend themselves to the use of such a contiguity definition. The adjacency criterion is used because it provides a realistic view of what constitutes a neighbor, given the irregular boundaries and areas of typical census data.

For both the Moran's $I$ and the minimum spanning tree (MST) method, contiguity between areas is calculated, whether using the distance contiguity criterion or the adjacency contiguity criterion, and the results are recorded within the program in matrix form. Therefore, this step in the calculation of contiguity is much the same for both methods used in the research. The spatial autocorrelation FORTRAN program used here (presented in Appendix A) is based on the program presented in Goodchild (1986).

The final Moran's $I$ statistic is given by:

$$I = \frac{\sum_i \sum_j w_{ij} c_{ij}}{s^2 \sum_i \sum_j w_{ij}} \quad (4)$$

where

$$s^2 = \frac{\sum_i (z_i - \bar{z})^2}{n} \quad (5)$$

Once Moran's $I$ is calculated, it is interpreted in the following way. A positive $I$ value indicates positive spatial autocorrelation (or a clustering of values) while a negative $I$ value
indicates negative spatial autocorrelation (some form of checkerboard pattern, with very dissimilar values for neighbors). A value close to 0 indicates randomness, or independence, between the areas' values (Goodchild, 1986). Interpretation of the resulting Moran's I values provides insight into the spatial pattern of a given variable, given that this measure of interdependence is based on the spatial configuration of a variable's measured values.

3.3.2. Minimum Spanning Tree

The second method used here to measure spatial pattern is a graph theoretic method that uses a minimum spanning tree. This section provides background on graph theory and a discussion of the minimum spanning tree method. It is not intended to provide a full review of graph theory, but rather to first define some terms related to the use of graphs. Minieka (1978) provides a short history of graph theory. He defines a graph as consisting of "...two parts, points and arrows joining these points (p. 2)", where the "points" are referred to as vertices and the "arrows" are referred to as arcs. This discussion uses the terms nodes instead of vertices and edges instead of arcs, as the terms are often used interchangeably. The edges connecting nodes may be directed or undirected, where an undirected edge refers to a connection which flows both ways, rather than in one direction. Phillips and Garcia-Diaz (1981) define a tree as "...a set of connected undirected edges (arcs) that contains no cycles (p. 91)." Connected means that all of the nodes are accessible or are joined by the set of edges and a cycle is a sequence of edges on a graph that begins and ends at the same node (Minieka, 1978). Cycles are undesirable for optimization applications because they can be redundant, as it is possible to traverse the same part of the graph more than once.
Taking the definition of a tree one step further, a spanning tree is "...any tree formed from the arcs [edges] of the graph that includes every vertex [node] in the graph (Minieka, 1978; p. 7)." Therefore, a spanning tree consists of the nodes of the graph \( n \) and exactly \( n-1 \) edges connecting those nodes. If every edge in the spanning tree has a cost associated with it, then a minimal (or minimum) spanning tree is the spanning tree with the lowest total cost when the edges' costs are summed (Phillips and Garcia-Diaz, 1981). Knowing the minimum spanning tree of a graph (or network) can provide information about a system if the problem can be expressed as a graph or topological problem. This is the approach taken in this research. Other examples of the use of minimum spanning trees for optimization are given in Chapter II.

To create a minimum spanning tree, it is necessary to know the cost or weight associated with each edge in the graph. The cost of an edge, in general, refers to the effort it takes to move along the edge between any two nodes. Phillips and Garcia-Diaz (1981) give the common example of using distance as the cost of an edge. For example, in a transportation or optimization application, the edge cost between any two nodes on the graph might refer to the time or distance between two places on the network. In this research the edge cost, as with Shier's (1982) example, is comprised of the absolute value of the difference between the two end nodes' values of the variable of interest. Therefore, for every node \( i \) connected to every node \( j \), the edge cost is given by

\[
|x_i - x_j| \quad (6)
\]

where \( x \) is the value of the variable under consideration. This edge cost definition is used because differences in values between neighboring nodes indicate how similar or dissimilar
the nodes are and therefore provide information about the spatial pattern of the region. This is shown graphically and discussed more thoroughly below.

Once a problem is depicted as a graph or topological problem, with nodes, edges, and costs, it is possible to construct the minimum spanning tree. Both Minieka (1978) and Phillips and Garcia-Diaz (1981) provide algorithms for construction of the minimum spanning tree of a graph. The minimum spanning tree subroutine used here is taken from Camerini, Galbiati, and Maffioli (1988) and is based on Kruskal's (1956) algorithm. Without going into specifics of the program (which is presented in Appendix B), the algorithm includes the following simplified steps. The input arguments consist of a list of the edges in the network (indicated by a left node and a right node) and the cost of each edge. The cost of the edge connecting each node pair is equal to the absolute value of the difference between the nodes' values. These input arguments are calculated in much the same way as the matrices for the spatial autocorrelation program are calculated. A locational or distance matrix is constructed, just as for the spatial autocorrelation procedure, and the entries in this matrix (based on either the distance contiguity criterion or the adjacency contiguity criterion) are used to create the list of input arguments to the minimum spanning tree algorithm. The minimum spanning tree routine selects the lowest cost (or shortest) edge and, if it does not create a cycle, it is included in the minimum spanning tree. This process iterates until the n-1 (n being the number of nodes in the network) lowest cost edges are included in the tree. Since the edges are considered in ascending order of cost, the resulting tree is minimized in terms of cost.

The graph theoretic approach used here is based on the application by Shier (1982) and Filliben, Kafadar, and Shier (1983), who used minimum spanning trees to measure the "homogeneity" of disks composed of two different materials (this presentation is based
upon Shier's discussion of the problem). In that case, homogeneity meant that the two materials were evenly (uniformly) spread throughout the disk, with high and low ratios of the two materials interspersed throughout. Regardless of the terminology used to describe the pattern (which is discussed more thoroughly in Chapter II), the problem was set up as a network problem wherein the ratios of the two materials were computed for areas within a superimposed grid. Since Shier (1982) wanted to find whether neighboring subareas had similar values, the differences between the areas' ratios constituted the costs along the network connecting the subareas. By constructing a minimum spanning tree (which can be defined more simply here as the shortest path connecting all nodes in the network) and summing the edge costs of the tree, he developed an index of "homogeneity" based on the total cost (TC) of the minimum spanning tree. High TC values indicate homogeneity (uniformity), since the difference between neighboring nodes is large (remember the checkerboard pattern). Low TC values indicate some other pattern or the existence of clusters, since neighboring nodes have similar values. This is the approach used here to examine spatial pattern using census units as the areal units of analysis.

To use the minimum spanning tree to examine spatial pattern, the region under study must be depicted as a graph. Figures 2-4 show how a region can be analyzed as a graph or network. Figure 2 displays a "region" for which we may be interested in creating a graphic representation. The region is divided into subareas. In the case of a typical geographic study, the region may be a county or MSA and the subareas may be census tracts, block groups, or blocks. Each subarea has associated with it a value. In addition, each subarea is considered a node on an overall network. The value associated with each subarea is assigned to the node for that subarea. The node may be represented in any way, but it is convenient to think of the centroid of an area as its node. For the application presented in this research, the centroid is used because it is readily available from census data.
Figure 2. Subareas of a Region.

Figure 3. Connected Subareas of a Region.
Once the region's subareas have been depicted as nodes, it is possible to connect neighboring nodes, as shown in Figure 3. The subareas that are neighbors are connected by an undirected edge, which is defined as an edge with no specified direction (Minieka, 1978). There are many possible ways to determine which subareas (or nodes) are neighbors. For this simplified example, subareas that share a common boundary are considered neighbors and are connected by an edge. Figure 3 shows the edges of the graph created for the sample region. Note that each subarea is directly connected only to its adjacent neighbors.

![Graph of Region With Edge Costs](image)

**Figure 4. Resulting Graph of Region With Edge Costs.**

The region may now be thought of as a graph, with nodes representing the centroids and edges connecting the neighbors. To complete the graph for calculation of the minimum spanning tree, the edge costs must be calculated. The costs are simply the absolute value of the difference between connected nodes' values. As will be shown, this provides
information about pattern, since similar neighbors will have low edge costs, based on this
definition of cost. Figure 4 shows the sample region's resulting graph with the subareas
represented by nodes on the network (with their assigned values), the neighboring subareas
connected by edges, and the associated edge costs.

![Graph With Evident Clustering](image)

Figure 5. Graph With Evident Clustering.

Now that the region has been depicted as a graph, it is possible to see how the MST
method can indicate pattern. The edge costs can tell much about the arrangement of values
throughout the region. Some of the edge costs are relatively low and others are relatively
high. A low edge cost indicates similar neighbors. Conversely, a high edge cost indicates
dissimilar neighbors. For example, Figure 5 shows the same spatial arrangement of nodes
on a graph (the same "region" shown in Figure 2). However, the values assigned to nodes
are clustered, with low values on the left portion of the graph and higher values on the right
portion of the graph. The edge costs reflect this clustering. There are five edges that can
be characterized as "low cost" edges, compared with four "high cost" edges. The high cost
edges are those that connect the two clusters. When the minimum spanning tree is created,
only the lowest cost "connecting" edge between the two main clusters will be included. The remaining edges in the minimum spanning tree will have low edge costs, thereby creating a low total cost for the minimum spanning tree.

![Graph With No Evident Clustering.](image)

Figure 6 shows the opposite pattern. The values in Figure 6 are arranged so that high values are connected to low values. The resulting graph includes only two "low cost" edges. While these low cost edges will be included in the minimum spanning tree, the other edges included will, of necessity, be higher cost edges. Once the minimum spanning tree is computed, the total value of the edges in the tree will be higher than the total edge costs from Figure 5. This shows how total edge costs are related to the pattern of values in a region.

The minimum spanning tree is the lowest cost path connecting all nodes in the network. Therefore, summing the edge costs in the resulting tree indicates the overall similarity or dissimilarity of neighboring nodes in the network: a low total minimum spanning tree cost
indicates many similar neighbors, while a high total cost indicates many dissimilar neighbors. Many similar neighbors indicate a clustering of values, as opposed to a checkerboard or even a random pattern, where higher edge costs would occur.

Interestingly, Shier (1982) states that a high TC (total cost) value indicates a checkerboard or what he called "homogeneous" pattern. However, a more intuitive type of homogeneity would be attained by a uniform pattern of values. If the data values were absolutely uniform (each subarea had the same value), the minimum spanning tree cost would be zero. Low minimum spanning tree costs supposedly indicate clustering of values. Therefore, the lowest possible value could be interpreted as one big cluster, rather than uniformity. Indeed, Shier would not consider this "homogeneous", given his definition of homogeneity. In reality, this situation is not likely to occur and the interpretation of relative minimum spanning tree total costs can proceed as described. The issue is not with Shier's interpretation of high and low values, but rather with his definition of homogeneity. This definition of homogeneity is only appropriate if very small subareas are considered over a very large region, and if high and low values are interspersed over that region. This is not likely to occur within human systems.

3.4. Geographic Data Problems

This discussion has raised several issues related to the use of geographic data for areas. These issues are common to both the spatial autocorrelation measure Moran's I and the MST method and it is helpful to reiterate both the problems and the measures used to account for them in this research. Briefly, and in summary, the problems include irregular spacing of areas, irregular sizes, resolution or scale effects, boundary effects, and sample sizes of areal units (Odland, 1988).
The first two problems of irregular spacing and size (as well as shape) are pointed out clearly in Cliff and Ord (1981), mainly in terms of using counties as the units of study. Odland (1988) also discusses these issues chiefly in relation to comparison of county level or other large areas. These are important issues and do not disappear for smaller areas, but for census data measured at a lower level of aggregation, the differences are not likely to be as significant. Therefore, for a descriptive measurement of pattern, these issues can be related to the issue of scale or aggregation, which is specifically tested in this research in terms of its effect on discerning spatial pattern. Furthermore, the use of two different contiguity criteria, as well as the very regular constructed data in the first research phase, allow the methods here to be tested first without these issues arising in a significant fashion.

Boundary effects occur when areas on the edges of the region under study may have a different number of neighbors than areas elsewhere in the region. This occurs for both irregularly shaped areas and for regular areas. However, Odland (1988) points out that boundary effects are not prohibitive when spatial autocorrelation statistics are used "...merely as measurements of pattern for a region (p. 28)." Therefore, when using these methods to discern pattern, it is unlikely that the edge effects will significantly alter results. However, those using spatial autocorrelation statistics to estimate parameters in other models should be aware of the edge effects (p. 29).

The issue of sample size is also less relevant when considering spatial pattern, since there are no attempts to draw inferences or test hypotheses from the measurements. The only concern in this research is to measure the spatial pattern of a variable over a region and the uses of Moran's $I$ and the MST method are purely descriptive. This points out a major difference between this treatment of pattern versus the use of spatial autocorrelation or other
techniques to make inferences about relationships between places or to measure the effects of autocorrelation on model estimation. The difference in using a spatial weighting matrix for specific problems is one example of how important these issues are if one is drawing inferences from the results, yet a binary contiguity matrix is suited for this research. However, the remaining locational problems for measuring pattern can be traced to the effects of aggregation or scale on measuring that pattern. This is the last geographical "problem" listed above and, it is argued here, many of the other problems are offshoots of the scale issue as well, for those studies interested in spatial pattern alone.

Therefore, the geographical problem of most concern to this research is that of scale or resolution. The aggregation or scale issue is discussed more thoroughly in the preceding chapter. Chou (1991) has shown the effects of scale on Moran's $I$ (higher resolution caused higher $I$ values) and it is hypothesized that the MST method is affected as well. Therefore, the principle concern when using these methods becomes that of the effect of the level of aggregation used, which is a main question of this research.

3.5. Conclusion

This chapter presented an introduction to the methods used here to examine spatial pattern. A brief discussion of the data used in the two research phases, one using constructed data and the other using census data, is provided. A general discussion of spatial autocorrelation and related statistics provide background for the Moran's $I$ discussion. This also provides background into some of the common considerations for both Moran's $I$ and the MST approach, such as definition of and representation of neighboring areas for the respective methods. In this research, contiguity is measured in two different ways: as a function of the distance between neighboring areas and as a function of the existence of a common boundary between neighboring areas. These contiguity definitions are used in the
same fashion for both the Moran's $I$ statistic and the MST method, though the MST requires that the information eventually be input as a list to the minimum spanning tree algorithm, while Moran's $I$ retains the original matrix representation.

The minimum spanning tree method is described in relation to a brief introduction to graph theory and graph theoretic definitions. It is shown how a region can be depicted as a graph for analysis with a minimum spanning tree and how those graphs can indicate pattern. Similarities between the MST method and Moran's $I$ are discussed as well, particularly in regard to inclusion of both locational differences and attribute differences between neighbors and appropriate representation of those neighbors. Finally, some geographical considerations are discussed and related to the use of Moran's $I$ and the MST method as descriptive approaches to discerning spatial pattern. The most important of these for determining pattern is scale or level of aggregation, which constitutes an important aspect of this research. Through description of the two methods as well as comparisons between them, this chapter has shown how both the MST and Moran's $I$ statistics can provide useful tools for analyzing spatial patterns, as well as how they are each applied in this research.
CHAPTER IV
APPLICATION TO CONSTRUCTED DATA

4.1. Introduction
To test the two methods, the spatial autocorrelation (Moran's I) and MST programs were run on constructed datasets exhibiting specific patterns. This step provides several advantages prior to using the programs on actual census data for Franklin County. First, using constructed datasets allows an experimental application of the methods without attendant problems associated with actual census units, such as uneven areas and irregular boundaries. It has been described how these aspects of areal geographic data might influence research results for various types of studies. While these influences for this research are more likely to manifest themselves in terms of aggregation effects, the constructed data allow a controlled examination of the methods and resulting aggregation effects. Second, since the data are specifically designed to mirror certain spatial patterns, use of the constructed datasets provide expected and more easily interpreted results in the first phase of the research. Finally, the constructed dataset results can highlight potential problems with interpreting the methods' results and allow for more informed use of the methods with the actual census data.

This chapter describes the analysis of the constructed data. After a discussion of the types of constructed datasets used and variations within the methods, the results of using the
spatial autocorrelation (Moran's I) and minimum spanning tree (MST) methods on the constructed datasets are presented. This is followed by a brief discussion of the usefulness of the findings in regard to the application of these methods to the actual census data in the second research phase.

4.2. Data and Methods

The constructed datasets consist of values arranged on a regular grid-like area designed to emulate the census aggregation hierarchy of a region. Figure 7 shows the constructed region used. The smallest units (blocks) aggregate to block groups and the block groups aggregate to tracts, just as with the census. Note that the smaller units making up any single larger unit are arranged in a square, due to the grid-like nature of the "region". The grid region consists of 256 blocks (i.e. 64 block groups and 16 tracts). The data values for the three aggregation levels are contained in separate files, along with the locational attributes of the region and its subareas. For example, each subarea (at each aggregation level) has a centroid based on its grid location and a separate adjacency file shows which subareas share borders. This type of constructed area is obviously unlike the irregular regions typically found with census data, but the purpose is to provide a similar hierarchy to test the methods under controlled circumstances.

Two basic constructed spatial patterns are used to test the two methods. Values of a mock variable are assigned in such a way as to create the specific patterns. The values of the variable range from 1 to 100, with low values falling between 1 and 10 and high values falling between 90 and 100. The values are assigned in a regular fashion for ease, since regularity will not hurt the outcome. The values are assigned sequentially beginning in the upper left corner of the region and high and low values are either deliberately clustered or interspersed. The values do not extend beyond the specified ranges (1-10 for low values
Figure 7. Constructed "Region" and Aggregation Scheme.
Figure 8. Positive (Clustered) Pattern.

Figure 9. Negative (Checkerboard) Pattern.
and 90-100 for high values). The intent is to create a regular pattern, since the interest is in discerning the specific patterns, rather than examining spatial relationships or processes between the areas in the grid. The low and high value ranges are dissimilar enough to produce the expected results from the two methods. In other words, these value ranges are dissimilar enough that a cluster of similar values will result in a positive Moran's I value and a low MST value. Conversely, dissimilar neighboring values will result in a negative Moran's I value and a high MST value.

The first pattern tested, shown in Figure 8, is a clustered pattern, consisting of high values (dark coloring) in the lower half of the region and low values (light coloring) in the top portion of the region. The values are assigned sequentially and the 1 to 10 value range is repeated throughout the white portion of the pattern. Likewise for the high values in the lower portion, where the 90 to 100 value range is repeated. The two clusters are sharply delineated. This pattern represents a high degree of positive spatial autocorrelation. For this reason, this pattern is referred to here as the "positive" pattern, which is synonymous with "clustered" for the purposes of this research. The other pattern, as shown in Figure 9, represents a "checkerboard" pattern, with high and low values very regularly interspersed throughout the region. Figure 10 (later in this chapter) shows an example of the value assignment scheme used to create the checkerboard pattern. This is the opposite of the positive pattern and, since it represents almost perfect negative spatial autocorrelation, is referred to here as the "negative" pattern.

While there are other patterns or data arrangements that might be tested, these two patterns provide the best means of distinguishing the potential usefulness of the minimum spanning tree method, particularly in relation to the spatial autocorrelation measure. Other, less obvious patterns would introduce added uncertainty in interpretation. Since the constructed
data represent the first step in this research for evaluating the MST method, it is important that the patterns be as distinct as possible. Furthermore, these patterns provide an easy gauge for comparison between the two methods because they correspond to specific expected spatial autocorrelation values.

With the constructed data, two different contiguity criteria are used for determining which subareas would be considered "neighbors". That is, for each area i, each contiguous area, j, is determined and a 1 is placed in the appropriate ij location in the locational (or distance) matrix W. The contiguity criteria are the distance contiguity criterion and the adjacency contiguity criterion described in the preceding chapter. In this simulated environment, the task of determining neighbors would not be difficult to accomplish simply by observation, because of the regularity of the areas in terms of size and spatial configuration. However, the main purpose of using the constructed data is to test the methods' usefulness for actual census data and it is necessary to discover the best means for that application. Therefore, the constructed data analysis uses a distance based contiguity criterion and an adjacency based contiguity criterion to determine the subareas' neighbors. These criteria are described in Chapter III and both sets of results are presented here, though the distance based results are less detailed.

While both approaches to determining neighbors are used on the constructed data, this is to highlight any significant differences in the results which can be attributed to the contiguity criterion used. It is felt that the adjacency criterion will be most useful for the actual census data, since the census units can be irregular in terms of size, shape, and number of neighbors. The adjacency files provide a more realistic view of the locational relationships within irregular regions, particularly if the aggregation level is high enough that such irregularities are magnified. However, the constructed data provide an easy avenue to test
possible differences in the results. Both sets of results are presented, with more detail given to the first set of results, which are based on the adjacency criterion.

4.3. Results Using the Adjacency Criterion

4.3.1. Moran's $I$

Table 1 shows the results of the spatial autocorrelation (Moran's $I$) program using the adjacency criterion on the constructed datasets. Remember, the Moran's $I$ measure is used as a benchmark to adequately compare the MST method results as well as to show the effects of aggregation on each. As shown in Table 1, the results at the block level are as expected for both the positive (clustered) and the negative (checkerboard) patterns. The positive pattern has a Moran's $I$ value of +.93, while the negative pattern results in a Moran's $I$ of -.99. These are very close to the values of +1.0 and -1.0 which would be expected in a perfectly spatially autocorrelated pattern. The failure to reach those "perfect" levels is a result of the edge effects found in any regional pattern. Those subareas on the edge of the region are likely to have fewer neighbors, thereby creating fewer "joins" and influencing the value of the index. Even given this, the Moran's $I$ value clearly indicates the expected results for each pattern. That is, the clustered pattern has a positive value approaching 1, and the checkerboard pattern has a negative value approaching -1. The minimal edge effects do not create difficulties in properly discerning the patterns.

While the results of the spatial autocorrelation method are as expected at the block level for both the positive and the negative datasets, there are aggregation effects when the block group and block level values are calculated. For the positive dataset, Table 1 shows that the block group and tract level Moran's $I$ values decrease from the high positive value found at the block level (.93). At the block group level Moran's $I$ is .87 and at the tract level it is .67. This tendency for the Moran's $I$ value to decrease with increased
aggregation further substantiates Chou's (1991) findings in his study of resolution effects for Moran's $I$. While these values are not as close to the +1.0 mark as that found at the block level, they still clearly indicate a positive or clustered pattern. Therefore, while there are aggregation effects on Moran's $I$ for the positive dataset, they are not strong enough to cause interpretive problems. The same cannot be said of the aggregation effects on the negative dataset.

Table 1. Results of Spatial Autocorrelation on Constructed Datasets: Adjacency Criterion.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Aggregation Level</th>
<th>Moran's $I$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Block</td>
<td>.930070519</td>
</tr>
<tr>
<td></td>
<td>Block Group</td>
<td>.856477141</td>
</tr>
<tr>
<td></td>
<td>Tract</td>
<td>.667429805</td>
</tr>
<tr>
<td>Negative</td>
<td>Block</td>
<td>-.993741632</td>
</tr>
<tr>
<td></td>
<td>Block Group</td>
<td>-.036523763</td>
</tr>
<tr>
<td></td>
<td>Tract</td>
<td>-.086419702</td>
</tr>
</tbody>
</table>

The Moran's $I$ values for the negative dataset are -.04 at the block group level and -.09 at the tract level. When interpreting Moran's $I$, a value close to zero indicates randomness in the study area or a lack of identifiable pattern. This result is not desirable if the purpose is to discern underlying patterns, but is also not totally unexpected. Chou (1991) shows how, when different resolution levels are used, a negative pattern can be misinterpreted as a positive pattern, given enough resolution levels. A quick glance at the checkerboard pattern shows what happens as aggregation occurs. The constructed datasets consist of a very regular grid system with a regular aggregation scheme and systematic value assignment. This means that, even at the first aggregation level (block group), the block groups have
similar values due to aggregation of two low values with two high values of the variable of interest. Figure 10 shows a hypothetical checkerboard pattern similar to the block level checkerboard pattern used here. The upper diagram shows the regularly interspersed low and high values. The lower portion of the diagram shows the results of aggregating the original values to the next higher level in the "census" hierarchy, where four smaller areas become one larger area. The original black and white pattern is replaced by an overall gray pattern, where neighboring areas have similar values.

Differences between the values at the block group and tract levels are less the result of pattern differences and more a result of differences in number of subareas and the overall magnitude of the variable values. This result provides the first solid indication in this research that aggregation has a significant effect, even when using a standard method such as Moran's $I$. This is particularly true if the pattern is not clustered in the manner of the positive dataset.

The Moran's $I$ statistic is able to discern both patterns clearly at the block level, with no room for misinterpretation of the resulting $I$ value. The positive (clustered) pattern is also easily discernable, though with a lower index, at the block group and tract level. Even given the lower $I$ values at these levels, though, the aggregation effect is not enough to hinder interpretation. Moran's $I$ is not able to adequately discern the negative (checkerboard) pattern at the block group or tract levels however. Given the extremely regular nature of the data, aggregation totally obscures the underlying pattern with the resulting $I$ value indicating randomness. This result provides an important point of comparison for examining the minimum spanning tree results.
Figure 10. Results of Aggregating a Checkerboard Pattern.
4.3.2. Minimum Spanning Tree

Having shown that the constructed patterns do, at low aggregation levels, produce the expected results when using Moran's $I$, and having developed expectations of aggregation effects based on the $I$ values, it is possible to use those results for comparing the minimum spanning tree (MST) method results. The value of the MST is expected to be low for a positive (clustered) pattern and high for a negative (checkerboard) pattern, since similar neighbors will have low cost edges between them and dissimilar neighbors will have high cost edges between them. The discussion first consists of a comparison between MST values of the two patterns and later extends to a closer examination of and an approach for interpreting actual MST values for the separate patterns.

Table 2. Results of Minimum Spanning Tree on Constructed Datasets:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Aggregation Level</th>
<th>MST Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Block</td>
<td>466</td>
</tr>
<tr>
<td></td>
<td>Block Group</td>
<td>884</td>
</tr>
<tr>
<td></td>
<td>Tract</td>
<td>1560</td>
</tr>
<tr>
<td>Negative</td>
<td>Block</td>
<td>22478</td>
</tr>
<tr>
<td></td>
<td>Block Group</td>
<td>312</td>
</tr>
<tr>
<td></td>
<td>Tract</td>
<td>144</td>
</tr>
</tbody>
</table>

Table 2 shows the results of applying the MST method to the constructed datasets using the adjacency criterion. It is useful to first examine the block level results. The block level MST value is 466 for the positive dataset. When compared with the 22,478 MST value for
the negative dataset, it is clear that the MST approach is able to differentiate between the two patterns in the expected manner. That is, the positive dataset has a "small" MST value and the negative dataset has a "large" MST value.

While the two patterns result in the expected MST values when compared at the block level (a low value for the positive pattern versus a high value for the negative pattern), aggregation to the block group and tract levels obscures the expected relationship between the two patterns' MST values. The positive pattern should have a lower value than the negative pattern. As shown on Table 2, the block group level MST value of 884 for the positive dataset is higher than the value of 312 for the negative dataset. This is exactly the reverse of the expected relationship between the two patterns' MST values. This effect compounds at the tract level with MST values of 1560 and 144 for the positive and negative datasets respectively. This reversal of the expected relationship between the two patterns' MST values is less a result of aggregation effects on the positive dataset than the aggregation effects on the negative dataset. In other words, while there might be a slight aggregation effect on the positive dataset's MST value, it would not be significant enough to cause the positive pattern's MST value to be higher than the negative pattern's value. However, as with the Moran's $I$ values for the negative dataset, the aggregation effects on the negative pattern is significant to cause an MST value lower than that found for the positive pattern. While the positive pattern's MST value did not increase much, the negative pattern's MST value decreased dramatically.

The reason for such a dramatic decrease in the negative pattern's MST value is similar to the aggregation effect on Moran's $I$. As the blocks are aggregated to block groups and then to tracts, each of the region's units or subareas has similar values. Therefore, the relative magnitude of the edge costs (the absolute value of the difference in the neighbors' values)
in the graph declines, even though the subarea's values are much higher at higher aggregation levels. Therefore, while the positive dataset is only slightly affected by the higher overall values (hence, higher possible edge costs) caused by aggregation, the negative dataset is overwhelmingly affected by the aggregation to very similar values for each subarea. This obscures the differences between the patterns and makes interpretation between the two patterns very difficult, for both Moran's I and the MST method.

Another way to explain differences between the MST values of the positive and negative datasets, when taking aggregation into account, is to look at other ways the MST value is affected. For instance, in this case, the MST value is a result of the differences between neighbors' values. This is what allows pattern interpretation. On the other hand, the MST value is also affected by the overall magnitude of the value ranges for the variable of interest, since these values create the edge costs in the minimum spanning tree. For instance, very large values of the variable of interest might create a larger MST value than a graph with smaller overall variable values, even though there may be little difference in their relative patterns. This can be accounted for when comparing two regions' patterns, if necessary. However, the MST value can also be affected by the number of edges included in the final minimum spanning tree. For instance, a smaller spanning tree, such as that for the tract level data, has fewer edges for which to add costs. Therefore, the final MST value is a combination of the actual pattern as indicated by edge costs, the magnitude of values in the region, and the number of subareas, though the pattern is the most important contributor to the value, as shown with the block level data. This has different ramifications, depending on whether a positive (clustered) or negative (checkerboard) pattern is present. It is important to keep in mind all of the possible reasons for any given MST value, besides the ones specifically associated with the pattern.
One way to account for the complexity of the final MST value, given the above considerations, is to compare the negative and positive dataset results in terms of the average cost of the edges included in the minimum spanning tree. The average edge cost for the positive pattern's minimum spanning tree should be lower than the average edge cost for the negative pattern's minimum spanning tree. As shown on Table 3, the average edge cost in the positive dataset's block level MST is 1.8, which is much lower than the 88.1 value found for the negative dataset at the block level. This is as expected when comparing the two patterns. However, the block group level average edge cost is 14 for the positive dataset and 5 for the negative, which is the opposite from expected. This result continues at the tract level, with values of 104 and 9.6 for the positive and negative datasets respectively. Again, this is a result of aggregation effects on the negative dataset rather than on the positive dataset. The average edge cost comparison depicts the same effects found by comparing the MST values, but demonstrates how future applications may interpret the MST value for regions with varying numbers of subareas.

**Table 3. Average Edge Cost in MST for Constructed Datasets.**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Aggregation Level</th>
<th>Average Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Block</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>Block Group</td>
<td>14.0</td>
</tr>
<tr>
<td></td>
<td>Tract</td>
<td>104.0</td>
</tr>
<tr>
<td>Negative</td>
<td>Block</td>
<td>88.1</td>
</tr>
<tr>
<td></td>
<td>Block Group</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>Tract</td>
<td>9.6</td>
</tr>
</tbody>
</table>
The previous discussion focused on comparison between MST values and average edge costs for the positive and negative datasets at the three different aggregation levels. The results indicate that the MST method provides the anticipated results at lower levels of aggregation and, further, how subsequent aggregation affects the MST values. However, the comparison between the two constructed patterns' MST values does not indicate much about the values themselves. In other words, what exactly does the block level negative dataset's MST value of 22,478 mean? Since the MST method does not create a descriptive index which falls within expected bounds (like Moran's $I$ does), it is necessary to examine the actual MST values further and to determine what a given MST value for a specific set of data means, when there is no other constructed pattern with which to compare it.

Random permutations of MST values provide a comparison for the two patterns' MST values and a method for detecting the pattern of any region or dataset. These permutations provide expected MST values for the region given a more random pattern than that actually found. In other words, the randomly generated MST values provide a gauge against which to measure the actual MST value for a pattern and, therefore, substantiates whether the MST has actually discerned any pattern at all. Examples of using permutations of randomly generated or simulated values to develop expected values or distributions are given in Shier (1982), Filliben, Kafadar, and Shier (1983), and Casetti and Krakover (1990). Getis and Ord (1992) use permutations of existing values to develop expectations for their $G_f(d)$ statistic.

The random permutations of the constructed data are computed in the following way. The MST program is run for 1000 iterations. On the first run through the MST program, the locational attributes of the region are determined. That is, the adjacency file is used to create the matrix of connected subareas or neighbors and the matrix is fixed for the entire
 iterative procedure. These locational relationships are maintained throughout the random permutations. However, for each subsequent iteration, the actual values of the variable of interest previously assigned to the subareas are randomly shuffled using the IMSL FORTRAN subroutines RNUN, SVRBP, and PERMU (Appendix B shows the entire MST program). RNUN generates random numbers from a uniform distribution, SVRBP sorts these random numbers to create a permutation vector, and PERMU rearranges the vector of the original values, based on the permutation vector from SVRBP (IMSL, 1989). The locational relationships of the subareas are not affected, so the only change with each permutation are the locations of the values of the variable. The result of the these permutations is a set of MST values given the actual locational attributes and the randomly shuffled and re-shuffled data values. The random permutations are computed for each constructed dataset at each aggregation level (block, block group, and tract) and can be examined in both a tabular and graphical form.

Table 4 shows the actual MST results found earlier for the two patterns, but with additional information about the set of randomly generated MST values at each aggregation level. This makes it possible to compare the actual MST value for each dataset with random values for the same data rather than simply to compare the MST values of the positive and negative datasets. This provides a clear indication of whether a pattern exists in the data. Table 4 indicates the actual MST value as well as the mean MST value, the minimum and maximum MST values, and the standard deviation for each pattern at each aggregation level. Note that these values are shown both with and without the actual MST value included. Except where indicated, the values are discussed including the actual MST values for the constructed patterns.
Table 4. Random Permutation Results for Constructed Datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Agg Level</th>
<th>MST Value</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>Stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Block</td>
<td>466</td>
<td>4455.74</td>
<td>466</td>
<td>6228</td>
<td>568.94</td>
</tr>
<tr>
<td></td>
<td>*w/o</td>
<td>4459.73</td>
<td>2774</td>
<td>6228</td>
<td>555.04</td>
<td></td>
</tr>
<tr>
<td>Block Group</td>
<td>884</td>
<td>5055.30</td>
<td>884</td>
<td>10136</td>
<td>1206.53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>w/o</td>
<td>5059.48</td>
<td>2100</td>
<td>10136</td>
<td>1199.89</td>
<td></td>
</tr>
<tr>
<td>Tract</td>
<td>1560</td>
<td>6503.57</td>
<td>1512</td>
<td>17228</td>
<td>2636.41</td>
<td></td>
</tr>
<tr>
<td></td>
<td>w/o</td>
<td>6508.51</td>
<td>1512</td>
<td>17228</td>
<td>2633.09</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>Block</td>
<td>22478</td>
<td>4477.49</td>
<td>2940</td>
<td>22478</td>
<td>794.81</td>
</tr>
<tr>
<td></td>
<td>w/o</td>
<td>4459.49</td>
<td>2940</td>
<td>6208</td>
<td>554.69</td>
<td></td>
</tr>
<tr>
<td>Block Group</td>
<td>312</td>
<td>338.63</td>
<td>240</td>
<td>456</td>
<td>34.61</td>
<td></td>
</tr>
<tr>
<td></td>
<td>w/o</td>
<td>338.65</td>
<td>240</td>
<td>456</td>
<td>34.62</td>
<td></td>
</tr>
<tr>
<td>Tract</td>
<td>144</td>
<td>132.93</td>
<td>56</td>
<td>248</td>
<td>28.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>w/o</td>
<td>132.92</td>
<td>56</td>
<td>248</td>
<td>28.36</td>
<td></td>
</tr>
</tbody>
</table>

* w/o under "MST Value" indicates that the following statistics do not include the actual MST value.
For the positive dataset, we expect low MST values, which indicate that neighbors with similar values are clustered together. The block level MST value for the positive dataset is 466. This value can be compared with the mean MST value of the 1000 random permutations of the data. As shown in Table 4, the mean MST value for the random permutations is 4456. This is much higher than the actual value computed, which was based on the positive pattern. Remember that the random permutations hold the same locational attributes as the positive pattern, but the data values are arranged randomly over those locations. Furthermore, the standard deviation of the MST values derived by random permutations is 569 (as opposed to 555 when the actual value is not included). This shows that the actual value of 466 is well below the range of what might be expected given a random pattern of the variable of interest. In addition, as shown on Table 4, the minimum MST value found for the random permutations is 2774. This shows clearly that the MST method has properly identified the clustered (positive) pattern because the MST value for that pattern is dramatically lower than those found for the random permutations.

Similar results are found when the block group and tract level MST values for the positive dataset are analyzed in this fashion. The block group MST value of 884 for the positive pattern is much lower than the random permutations' MST mean value of 5055. The standard deviation of 1207 for the permutation MST results shows how far from the mean the actual MST value falls. As well, the tract level MST value of 1560 for the positive pattern is much lower than the randomly generated mean MST value of 6504, but the random permutations have a standard deviation of 2636. This places the actual MST value within two standard deviations of the mean of the random permutations' values. This indicates that the value could possibly have resulted from some random arrangement of the data. However, this does not hamper interpretation of the actual MST value. When using the MST value as a descriptive indicator, the mere fact that the actual MST value for the
pattern is so much lower than the mean value of the random permutations indicates a strong degree of clustering.

Therefore, as the aggregation level increases, the actual MST value falls closer to the random permutations' mean MST value (in terms of standard deviation), though still well below the expected value given a random pattern. This can also be seen through examination of the minimum and maximum MST values found with the random permutations. While the means and standard deviations are shown both with and without inclusion of the actual MST value, the value of these statistics does not differ significantly with inclusion of the actual value (except for the block level standard deviation, which still does not affect interpretation of the actual MST value). However, note that the tract level random permutations include at least one value less than the actual MST value, as shown by the 1512 value listed in Table 4. Therefore, even though the mean and standard deviation indicate that the actual tract level MST value is much lower than the random expectation, that relationship is less strong than for the block and block group level values. This is in keeping with the above discussion of the tract level standard deviation and indicates the existence of aggregation effects for the positive pattern. However, these effects do not interfere with the ability of the MST method to discern pattern in the expected way.

Table 4 also shows the results of the random permutations for the negative dataset. We expect a high MST value for the negative (checkerboard) pattern. It has already been shown that the negative pattern's block level MST value of 22,478 is much higher than the block level MST value for the positive dataset. That value is also considerably higher than the mean MST value computed for the random permutations. The mean MST value for the random permutations of the negative dataset (including the actual MST value) is 4468, with
a standard deviation of 795. This indicates that the actual MST value for the negative pattern is significantly higher than would be expected for a random arrangement of the data over the same region. The high MST value clearly indicates that neighbors are dissimilar. Without including the actual MST value for the negative dataset, the second highest value for the random permutations is 6208. It is highly unlikely that the actual MST value of 22,478 could have occurred by chance. Therefore, at the block level, the MST method indicates a nonrandom, non-clustering pattern for the negative dataset, in keeping with expectations. Given the extreme MST value for the negative dataset and the standard deviation of the random permutations, the MST method is able, at the block level, to discern the checkerboard pattern.

The results are not as clear as the aggregation level increases. As shown on Table 4, the block group level MST value for the negative pattern is 312. The mean MST value of the random permutations is 339, with a standard deviation of 35 and a range between 240 and 456. The actual MST value is not clearly distinguishable from what might be expected with a random arrangement of data over the constructed region. This is also true at the tract level. The actual tract level MST value of 144 is very close to the random permutations' mean MST value of 133. Furthermore, the standard deviation of 28 and range between 56 and 248 show that the actual MST value does not suggest either of the two patterns tested or, indeed, any pattern other than what would be expected in a completely random distribution of values of the variable of interest. These simple statistics bear out the earlier assumption that the MST method would not distinguish the negative pattern when aggregation occurred and that the inability to compare the positive and negative patterns at higher aggregation levels results from aggregation problems with the negative pattern rather than with the positive pattern. While the results indicate that aggregation affects the positive MST values too, the effects are not enough to interfere with or cause
While the above statistics allow easy interpretation of the calculated MST values for the constructed patterns, it is also possible to graphically examine the effectiveness of the MST method in this regard. Figures 11-19 show frequency histograms of the randomly generated sets of MST values. Figure 11 shows the histogram for the block level positive dataset, not including the actual MST value for the positive pattern. As shown in Figure 11, the frequency histogram for the random permutations of the block level positive dataset indicates a normal-shaped distribution about the mean MST value of 4456. The actual MST value for the positive pattern (466) is well off the left side of the graph and shows up better in Figure 12, which illustrates the distribution when the actual MST value of 466 is included. The graphs illustrate the degree to which the actual MST value is removed from that which might be expected given a random geographical arrangement of the same data. The block group level histograms again show that the actual MST value (884) is far off the left end of the graph in Figure 13 (which does not include the actual MST value) or at the very left in Figure 14 (which does include the actual value). Both figures indicate a less normal distribution than did the block level histogram, but both show the distributions clearly centered about the mean of about 5055, which is considerably higher than the actual MST value found for the positive pattern. Both graphs illustrate clearly that the actual MST value is much lower than those obtained from the random permutations.

Figure 15 shows the frequency histogram for the tract level positive dataset without the actual MST value of 1560 (the histogram including the actual MST value is not shown, because it is the same graph). The distribution of the MST values for the random permutations at the tract level is less normal than at the block or even block group levels.
Figure 11. MST Permutation Frequencies: Block Level Positive Dataset, With Actual MST Value.
Figure 12. MST Permutation Frequencies: Block Level Positive Dataset, Without Actual MST Value.
Figure 13. MST Permutation Frequencies: Block Group Level Positive Dataset, Without Actual MST Value.
Figure 14. MST Permutation Frequencies: Block Group Level Positive Dataset, With Actual MST Value.
Figure 15. MST Permutation Frequencies: Tract Level Positive Dataset, Without Actual MST Value.
The bulk of the values are clustered about the mean MST value of 6509 (the mean without inclusion of the actual value), but the distribution is skewed toward the higher end of the graph and the mean does not fall into the largest frequency category. However, the graphic shows that the actual MST value of 1560 falls in the lower end of what would be expected for a random pattern, strongly suggesting a clustered pattern. The actual MST value for the positive pattern at the tract level falls within two standard deviations of the mean, so there is a chance that the value could have resulted from a random arrangement of the data values over the constructed region, though this is not likely. In this case, the frequency histograms illustrate the existence of aggregation effects for the positive dataset at the tract level, though these effects are not enough to hamper pattern interpretation. Whether this is all true aggregation error or the result of fewer subareas is not clear.

Figures 16-19 show the frequency histograms for the negative (checkerboard) dataset. As shown in Figure 16, the block level permutation MST values for the negative dataset are normally distributed about the mean of 4479. The actual MST value found for the block level negative dataset (22,478) is well off of the right-hand side of the graph, indicating (as the earlier statistics did) that the actual value is out of the range of that expected for a random arrangement of the data. Clearly, the pattern is not random. Furthermore, the high value indicates a total lack of clustering, as was dictated through use of the checkerboard pattern. At the block level, the MST method performs very well in discerning the checkerboard pattern in the expected fashion. Figure 17, which shows the frequency histogram including the actual MST value, shows clearly how far removed from the randomly generated distribution the actual pattern is.

The difficulties of the MST method in discerning the negative (checkerboard) pattern at higher aggregation levels, similar to those of the spatial autocorrelation statistic Moran's I,
Figure 16. MST Permutation Frequencies: Block Level Negative Dataset, Without Actual MST Value.
Figure 17. MST Permutation Frequencies: Block Level Negative Dataset, With Actual MST Value.
Figure 18. MST Permutation Frequencies: Block Group Level Negative Dataset, Without Actual MST Value.
Figure 19. MST Permutation Frequencies: Tract Level Negative Dataset, Without Actual MST Value.
has been demonstrated through comparison and simple statistical analysis. Figure 18 shows these results graphically, with the frequency histogram of the block group level MST values of the random permutations of the negative dataset. The actual MST value of 312 for the tract level negative pattern is in the middle of a normal distribution about the permutations' mean MST value of 340 and falls within one standard deviation of the mean. The histogram which does include the actual MST value has not been included here because it is the same histogram. To indicate the expected (negative) pattern, the actual value would have to fall in the right hand portion of the graph, which it does not. Aggregation to the block group level has masked evidence of the checkerboard pattern which was so clear at the block level. This is also evident at the tract level, as shown in Figure 19. The actual tract level MST value for the negative pattern (144) is in the same frequency category as the mean value for the random permutations (133). The distribution is not as normal, but the graph indicates that, at the tract level also, the actual MST value does not differ from one that might be expected for any given random arrangement of the data values. If aggregation had not been considered, the MST results would be misinterpreted and the underlying checkerboard pattern totally obscured.

The results so far indicate that the MST value is useful for discerning spatial pattern when using low aggregation levels, as the resulting values are much higher or lower than the values obtained through random permutations of the data for both datasets. However, aggregation effects occur with both the positive and the negative patterns. The aggregation effects do not appear until the tract level aggregation for the positive dataset and are not significant enough to interfere with pattern interpretation. The aggregation effects show up clearly at the block group level for the negative dataset. The same is true for the Moran's I results. These results are based on the use of the adjacency criterion for determining neighbors. As was mentioned earlier, a distance criterion is also used. The results are
similar and are only briefly discussed here, particularly as it is felt that the adjacency criterion provides a more realistic approach to determining neighbors for census data.

4.4. Results Using the Distance Contiguity Criterion

4.4.1. Moran's $I$

Table 5 shows the results of calculating the Moran's $I$ statistic for the constructed datasets using the distance contiguity criterion. As with the adjacency criterion results, the Moran's $I$ values are as expected. The Moran's $I$ value for the positive dataset at the block level is +.93. At both the block group and tract levels the values also indicate positive spatial autocorrelation or clustering, with +.86 and +.63 values respectively for block groups and tracts. As before, the value decreases with aggregation. Since the Moran's $I$ statistic is, to a certain extent, a function of the number of high and low "joins", this is not surprising and follows the trends indicated in the earlier results. However, even given the decrease in the Moran's $I$ value with aggregation, the value still strongly indicates a clustered pattern.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Aggregation Level</th>
<th>Moran's $I$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Block</td>
<td>.930754125</td>
</tr>
<tr>
<td></td>
<td>Block Group</td>
<td>.860247433</td>
</tr>
<tr>
<td></td>
<td>Tract</td>
<td>.625918329</td>
</tr>
<tr>
<td>Negative</td>
<td>Block</td>
<td>-.869459629</td>
</tr>
<tr>
<td></td>
<td>Block Group</td>
<td>-.092871666</td>
</tr>
<tr>
<td></td>
<td>Tract</td>
<td>-.074074030</td>
</tr>
</tbody>
</table>
The values for the negative dataset are similar to those found with the adjacency contiguity criterion results. At the block level, the Moran's $I$ value is -.87, which approaches the perfectly negative spatial autocorrelation value of -1.0. At both the block group and tract levels, the $I$ values fall off rapidly from the expected -1.0 value. Rather, the values of -.09 and -.07 for block group and tract respectively, are so close to zero as to indicate randomness in the pattern (a value near zero for Moran's $I$ indicates randomness). The fact that the values are both on the negative side of zero has little interpretive meaning, and is too close to zero to indicate a checkerboard or negative pattern. These are the same basic results found using the adjacency criterion for determining neighbors.

This quick comparison between the distance criterion and the adjacency criterion shows that the same aggregation effects hold true for both the distance results and the adjacency results. The one major exception is that the adjacency criterion returns a value closer to -1.0 for the block level negative dataset (-.99 for the adjacency criterion versus -.87 for the distance criterion). This is the result of edge effects, which occur because those subareas on the edge of the region are likely to have fewer neighbors, which results in fewer potential "joins" and affects the resulting value.

Both methods of determining neighbors introduce the possibility of edge effects, but the checkerboard pattern is more affected when the distance criterion is used than when the adjacency criterion is used. This is because of the form of the distance criterion. The four closest subareas are considered neighbors. For those blocks at the edge of the region, this means that the fourth neighbor is further away than if adjacent blocks are used. In that case, particularly for the checkerboard pattern in the constructed region, the fourth neighbor is likely to be a similar neighbor rather than a dissimilar neighbor, thereby affecting the Moran's $I$ value. While this indicates that the distance criterion is less suitable than the
adjacency criterion, it must be emphasized that these effects do not interfere with interpretation of pattern using these methods. At the block level, both the Moran's $I$ and the MST methods are able to clearly differentiate between the two patterns. However, these results provide added proof that the adjacency criterion provides a more realistic approach to defining neighbors and supports the use of that criterion through the remainder of the research. The difference between the distance and adjacency criteria for the block level positive data are minimal. Any edge effects are much the same for either criterion and both still provide a Moran's $I$ indicating a positive pattern.

Given the results shown in Table 5, several conclusions are warranted. As with the earlier reported results, the constructed patterns perform as expected at the block level using Moran's $I$. The positive pattern also works as expected at the block group and tract level as well, though the degree of positiveness decreases with aggregation. However, the negative pattern does not provide the expected results when aggregated to the block group and tract levels. This is because, as was hypothesized, aggregation obscures the underlying pattern. As the blocks are aggregated to block groups and then to tracts, the values of the aggregated units are too similar to indicate any pattern. This effect is exacerbated by the use of regular data in the constructed dataset. Aggregation creates units with roughly the same data values for the entire study area and effectively masks the underlying pattern shown by the block level Moran's $I$ statistic.

4.4.2. Minimum Spanning Tree

The same aggregation effects are shown by the results of the minimum spanning tree (MST) method on the constructed data. Having established that the adjacency criterion is a better method for determining neighbors, the distance contiguity criterion results are only
briefly discussed and are limited to a comparison between the values found for the positive and negative datasets at the three aggregation levels.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Aggregation Level</th>
<th>MST Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Block</td>
<td>442</td>
</tr>
<tr>
<td></td>
<td>Block Group</td>
<td>700</td>
</tr>
<tr>
<td></td>
<td>Tract</td>
<td>1492</td>
</tr>
<tr>
<td>Negative</td>
<td>Block</td>
<td>17,110</td>
</tr>
<tr>
<td></td>
<td>Block Group</td>
<td>276</td>
</tr>
<tr>
<td></td>
<td>Tract</td>
<td>96</td>
</tr>
</tbody>
</table>

Table 6 shows the results of the MST runs on the two constructed patterns using the distance criterion. A positive (clustered) pattern should result in a lower MST value than that for a negative (checkerboard) pattern. As shown on Table 6, this expected relationship between the positive and negative patterns' MST values holds well at the block level. The block level MST value for the positive dataset is 442, while the block level MST value for the negative dataset is 17,110. The negative dataset has a much higher MST, indicating high costs between neighbors and suggesting a pattern which is obviously not clustered, unlike the positive pattern.

As with the results reported earlier, aggregation beyond the block level makes interpretation more difficult. At the block group level, the positive dataset has an MST value of 700
compared to the negative dataset's value of 276. This is the opposite of what is expected, which is that the clustered pattern has a lower MST value than does the checkerboard pattern. As before, this is compounded at the tract level with values of 1492 and 96 for the positive and negative datasets, respectively.

It is not necessary to go further with interpreting the results found using the distance criterion to determine neighbors. Random permutations were not run on the data using this method of neighbor determination. It has been shown that the basic positive and negative pattern comparison results are much the same using either contiguity criterion. Furthermore, the Moran's I results suggest that the adjacency criterion provides less susceptibility to edge effects (though these do not hinder interpretation with either criterion) than does the distance criterion, especially for the negative pattern. Finally, as has been discussed, the adjacency criterion is more likely to find the true "neighbors" of each subarea, given the less regular nature of census data when compared to the constructed datasets. In this case, the constructed data provide an excellent means of examining each contiguity approach before application to the census data. Given the results, logic dictates that the remainder of the research use only the adjacency criterion for determining neighbors.

4.5. Conclusion

Overall, both the Moran's I and MST methods have been shown to be successful in discerning the two test patterns at lower levels of aggregation. However, it has also been shown that aggregation affects the ability of either method to discern pattern. Aggregation has some affect for the positive (clustered) dataset, but not enough to cause misinterpretation of either the Moran's I or the MST value. On the other hand, both methods are significantly affected by aggregation with the negative (checkerboard) pattern.
Moran's $I$ fell close to zero at the block group and tract levels, indicating a random pattern. This occurs using both the distance contiguity criterion and the adjacency contiguity criterion. The MST value is not different from what would be expected with a randomly generated pattern at the block group and tract levels. This result is not unexpected when one logically considers the effect of aggregating any pattern resembling a checkerboard. Though it is unlikely that any real-life pattern would so closely approximate the checkerboard pattern, there are patterns that more closely resemble a checkerboard than a strict set of clusters as in the positive dataset. Therefore, understanding the effects of aggregation on the two extreme pattern types can prevent misinterpretation when aggregated data are used.

Testing the two methods on constructed datasets provides several insights into their use with actual census data. First, we know that aggregation affects the values found for both the positive and negative datasets, but affects interpretation of the negative dataset more. Therefore, we should be careful to consider what type of pattern exists, so as to avoid possible misinterpretation. For most human systems, it is likely that a positive pattern exists, particularly at lower levels of aggregation. Therefore, if that is the case, the MST method will discern such a pattern. Most importantly, if there is no expectation of the type of pattern present, then the lowest possible level of aggregation should be used. This holds true for the Moran's $I$ value as well.

Secondly, the MST approach works as well for discerning spatial pattern as does Moran's $I$. Moran's $I$ is easier to interpret if it stands alone, since it is an index with expected bounds. The MST requires random permutations of the data over the region to provide a comparison and allow pattern interpretation. Obtaining these permutations is easy for reasonable size areas and it is probable that the MST method can be made to work with
larger regions than Moran's $I$. This is discussed more thoroughly in the final chapter.

Once the random permutation results are obtained, interpretation of the MST is as expected: a clustered pattern has a relatively low MST value and a checkerboard pattern has a relatively high MST value, when aggregation levels are low. Given this, the MST method provides a useful addition to spatial analytical tools when properly applied. Proper application, as with Moran's $I$, includes the use of as disaggregate a level of data as possible.

Finally, comparison between the two contiguity criteria (distance versus adjacency) indicate few obvious differences for the constructed data region. There are some advantages to the adjacency criterion in terms of edge effects for the negative pattern, though these effects did not hinder pattern interpretation. This result, in addition to the obvious differences between the constructed data and the census data, and the fact that the adjacency criterion simply makes more sense, indicates that it is appropriate to use the adjacency criterion on the actual census application which follows.
CHAPTER V
APPLICATION TO CENSUS DATA: FRANKLIN COUNTY, OHIO

5.1. Introduction
Results from the previous chapter using constructed data arranged in specific patterns indicate that the type of pattern expected, the aggregation level, and, to a lesser extent, the method of determining subarea neighbors, may affect interpretation and/or usefulness of both the MST value and Moran's $I$ statistic. In particular, higher levels of aggregation have been shown to mask the checkerboard type pattern that is apparent at the block level. This information allows more informed application of the methods to census data in the second phase of research. While the constructed data patterns provide a controlled means of testing the two methods' pattern detection abilities, it is still necessary to examine the practicality and reliability of the methods with census data. This is particularly true for the MST method, as it represents a more novel approach to measuring spatial pattern than does Moran's $I$. This chapter presents results of applying the Moran's $I$ and MST methods to determining segregation patterns in Columbus, Ohio. The first section of this chapter discusses the study area and data used for the two methods. A graphical examination of the degree of segregation in the County, along with a brief discussion of dissimilarity indices, is presented to provide expectations of the prevailing pattern of residential segregation. This does not represent an hypothesis to be tested, but rather allows more informed interpretation of the results for this application, particularly in the case of the MST method.
Results of the Moran's $I$ and MST analyses are presented, along with some conclusions about the practical uses of both methods based on results from this application.

To test the usefulness of the MST method and further test the usefulness of Moran's $I$ for discerning spatial pattern, the methods are applied to an examination of residential segregation patterns in Columbus, Ohio. Residential segregation provides only one possible application of these methods, but one that is highly appropriate given the nature of segregation and segregation measures. The goal is to show how both methods may be used to investigate the degree of clustering in the segregation pattern of an urban area. Other more commonly used measures of segregation have already been presented in the literature review. Some of these methods, such as the index of dissimilarity and index of segregation, are aspatial in that they do not consider the actual locational relationships of the areas involved. Rather, they deal with relative proportional distributions of two groups for a given region or group of areas. Still others require some a priori categorization of the areas involved or choose only specific types of tracts or areas to consider. This is true even of approaches that do consider spatial arrangement, as they often consider only the spatial arrangement of specific types of areas or tracts. Both Moran's $I$ and the MST method presented here specifically take into account the locational relationships and attribute distributions of all of the region's subgroups, while offering flexibility in how those relationships are represented. These should provide a better indication of any overall pattern for the region.

5.2. Study Area and Data

The study area for this application of the MST and Moran's $I$ methods is Franklin County, Ohio, which includes the capital city of Columbus. Franklin County has a 1990 population of 961,437, of which approximately 18.5% are nonwhite. This portion of the research
uses the Moran's $I$ and MST programs to determine the extent to which that nonwhite population is clustered or is scattered throughout the county. In other words, this is an attempt to discover a numerical indication of overall segregation for Franklin County using first Moran's $I$ and then the MST method. This application considers only white versus nonwhite segregation. Many studies examine segregation patterns for different racial or ethnic groups, but the intent of this research is strictly to determine the usefulness of these methods in examining spatial patterns such as those found in segregation studies, not to develop theories underlying those patterns. There is no reason why these methods could not be applied to particular ethnic groups in the same fashion as presented here.

Figure 20 shows Franklin County, along with the 252 census tract divisions for the 1990 Census. Maps of the nonwhite population density patterns are presented later in this chapter, for comparison with analysis results. The data used to examine nonwhite population density patterns are from the 1990 first release TIGER/Line files and Summary Tape File 1 (STF1) census data tapes. Data at both the block group and tract levels were read from the STF1 census tapes. The block group and tract aggregation levels provide sufficient data for comparison of the methods at different aggregation levels, because of an expectation that the existing pattern of nonwhite population density is clustered rather than scattered throughout the county. The constructed data results indicate that a clustered pattern is less likely than a checkerboard or scattered pattern to be masked by aggregation. Rather, this type of pattern, while still subject to aggregation effects, will appear less strongly clustered at higher aggregation levels. Using both aggregation levels allows further comparison between the two methods of the aggregation effects demonstrated with the constructed data.
Figure 20. Franklin County, Ohio: 1990 Census Tracts.
To create the study variable "nonwhite population density", the following variables are read from the 1990 STF1 tapes at both the tract and block group levels: total population (for each census unit level studied) and total white population. Subtracting the white population from the total population creates the variable "nonwhite population". It is possible to examine segregation of specific groups, but that is not necessary for this research. These data are then combined with data from the TIGER files, which provide the area of each census unit. Dividing the nonwhite population of each unit by its area produces the final variable "nonwhite population density". In addition to providing the area information, the TIGER/Line files provide other necessary topological information about the County's census units for use in both the MST and Moran's' I calculations and are used to create the adjacency files listing the neighbors of each subarea in the county.

Using the variable "nonwhite population density" helps accommodate the fact that census units have irregular boundaries and sizes, as well as obviously different value ranges within them. While these problems are more significant at higher aggregation levels and for model oriented studies, it is desirable to account for them where possible. Use of the density variable helps account for different sized regions in measuring the variable of interest. Therefore, any given spatial unit (e.g. tract) may have a low nonwhite population density for three reasons: it may have a low nonwhite population overall, regardless of the size of the area; or it may have a medium nonwhite population spread over a large area; finally, it may have a small area, but a very low or nonexistent nonwhite population. For the purposes of initially discerning segregation patterns, it is not important which of these conditions exists, but rather that the method be able to handle them. Use of the density variable aids in accounting for these problems.
Only the adjacency criterion for determining neighbors is used here, since the findings from the previous chapter offer no obvious advantage to using the distance criterion (in fact, the Moran's $I$ values for the constructed data indicate that there may be a disadvantage). The adjacency files identify actual boundary sharing neighbors, making it possible to handle irregular boundaries and uneven numbers of neighbors without having to resort to visual inspection, which would be an enormous task in most practical applications of these methods. The combination of using the density variable along with the adjacency files for contiguity goes a long way toward alleviating concerns about the irregularity of census areas as units of study, particularly in this case, where the main concern is identifying spatial pattern.

The final data input to the two methods consist of the nonwhite population density for each census tract and block group, as well as the adjacency files for both the 252 census tracts and 942 block groups in the 1990 Census for Franklin County. Again, only the adjacency contiguity criterion for determining neighbors is used on the Franklin County census data. These data are used to calculate both the MST values and the Moran's $I$ values to determine the amount of clustering or segregation evident in Franklin County, as well as the effects for each method of using different aggregation levels (i.e. block groups versus tracts).

53. Nonwhite Population Density Maps and Segregation Indices

Prior to discussing the Moran's $I$ and MST results, it is helpful to graphically examine the degree of clustering of nonwhite population in Franklin County. Since this research tests the methods, particularly the MST method, in terms of their ability to discern spatial pattern, it is useful to have a basis for comparison. While the ability of Moran's $I$ to indicate certain patterns is known, both it and the density maps provide needed comparison to test the MST method. Figure 21 shows the nonwhite population density for Franklin
Figure 21. Franklin County: Nonwhite Population Density by Tract, 1990.
Figure 22. Areas of Higher Nonwhite Population Density: Tracts.
County at the census tract level for 1990. The average nonwhite population density per square mile for Franklin County is roughly 327. While it is hypothesized that nonwhite population is clustered, or that a relatively high degree of segregation exists, the map provides graphic indications of that segregation and shows a clearly clustered pattern of high nonwhite population density tracts. Though the pattern is by no means as strong as that of the "clustered" pattern used in the constructed data phase of this research, there is strong evidence of clustering.

As shown in Figure 21, the lowest nonwhite population densities are for census tracts located in the suburban fringes of Columbus and Franklin County. The highest densities are evident in the central city, with a distinct band east of the city center and along the major north/south corridors. In fact, all of the three highest density categories are located in the eastern portion of the city, with the exception of a very high density tract along West Broad Street. Overall, the pattern can be described as clustered, with strong clusters of high nonwhite population density tracts in the central city and low nonwhite population density tracts at the edges of the County.

Figure 22 shows a close-up of the areas with the higher nonwhite population density clusters. The larger scale makes it possible to see even more clearly the cluster of higher density tracts to the east and northeast of the city center. Again, with the exception of the high density tract along West Broad, all of the highest nonwhite population density tracts are in the eastern central portion of the city and are surrounded by other relatively high density tracts. Nonwhite population density declines as distance from the city center increases, particularly to the west and northwest. This type of clustering is typical of what might be found in terms of residential segregation for many U.S. cities.
It is important to keep in mind that the central city tracts may have higher nonwhite population densities because they are smaller tracts and have higher population densities in general. This is usually the case with tracts located in more urbanized areas. On the other hand, the patterns may arise because of societal trends of whites residing in suburban tracts while nonwhites occupy the central city. In fact, both of these issues probably contribute to the clustered pattern shown. This research is concerned only with discerning the pattern present and demonstrating the usefulness of the methods presented here for those who are interested in why such a pattern exists.

The clustered pattern described above is also seen at the block group level. Figure 23 shows the nonwhite population density by block group for the higher nonwhite population density areas. The block group map for the entire county is not shown because it is difficult to discern pattern due to the much larger number of block groups compared to tracts. The same basic overall clustered pattern of nonwhite population densities is shown in Figure 23, with nonwhite population density decreasing as the block groups are further removed from the central city. One reason for possible aggregation effects is apparent from this map. The fact that there are more block groups than tracts raises the possibility that there are more similar areas next to one another. This creates more similar "joins" and makes either method more likely to detect clustering. For example, the single very high density tract on the west side is composed of several high density block groups, which form a "cluster" of their own. Conversely, if the pattern for an area is more like a checkerboard, that pattern will likely be obscured at the tract level. Some of the higher nonwhite population density tracts, when observed at the block group level, are comprised of some high and some lower density block groups. We cannot infer too much from these maps due to their different scales and density categories, but we can see some examples where aggregation will affect results. In general, the block group and tract level maps
Figure 23. Areas of Higher Nonwhite Population Density: Block Groups.
shown here indicate that Franklin County has an overall clustered pattern of nonwhite population density.

The maps provide a graphical representation of nonwhite population clustering in Franklin County, but it is also possible to compare the methods used here with expectations derived using the dissimilarity index described in Chapter II, which measures the degree of concentration of a group (in this case, nonwhites). Overall, U.S. cities have experienced a decline in residential segregation over the past few decades based on this index (Yeates and Garner, 1980), though Massey and Denton (1987) found that black segregation declined less in large, northern urban areas than in smaller urban areas in the south and west. Columbus (and Franklin County) mirrors the trend toward less segregation. Massey and Denton (1987) found a dissimilarity index value, using tract data, of .818 for blacks in Columbus in 1970. This dropped to .714 in 1980. The index can be interpreted as the proportion of the group under study that would need to relocate to create a desegregated pattern (Yeates and Garner, 1980).

To compare with these earlier indices, the index of dissimilarity can be calculated (using the TRANSCAD GIS) from Franklin County census data for 1980 and 1990. The resulting index for 1980, at the tract level is .667 and the value for 1990 is .579. Note that these values are not directly comparable to the values listed above, as Massey and Denton (1987) measured the index for blacks (as well as other groups) over the metropolitan area, while the index here is for all nonwhite population measured over the county. However, the trend is the same described by Yeates and Garner (1980) and, to some extent, Massey and Denton (1987), indicating decreasing segregation. The overall indication, based on the index of dissimilarity, is that residential segregation exists in Franklin County. However,
these index values are based only on relative proportions and do not explicitly consider the spatial arrangement of that segregation, as do Moran's $I$ and the MST method used here.

5.4. Results

Table 7 shows the Moran's $I$ spatial autocorrelation and MST results for Franklin County, Ohio, at the block group and tract levels. The Moran's $I$ value at the block group level is +.63, which indicates clustering. This is expected, given the variable of interest, nonwhite population density, and the fact that visual inspection of the nonwhite population density maps for the county also indicates clustering. The positive Moran's $I$ value further substantiates that a clustered pattern exists.

The $I$ value of +.56 at the tract level also indicates a clustered pattern, though less strongly than at the block group level. The Franklin County Moran's $I$ values, while clearly indicating clustered patterns, are not as highly positive as those found for the positive pattern of constructed data in the preceding analysis. This is not surprising, since the constructed data are specifically designed to produce as distinct (or as high) a result as possible. The census data here have the inherent irregularities of such data, as well as the important fact that the clustering is not total. Some of the lower density block groups, for instance, have no housing units or population at all, which affects the degree of similarity or dissimilarity between them and their neighbors and, consequently, the overall Moran's $I$ value obtained. For the study area of Franklin County, the Moran's $I$ values reinforce well the expectations of clustering, and provide a good level of comparison for the results of the MST method.
Table 7. Moran's I and MST Results: Franklin County.

<table>
<thead>
<tr>
<th>Aggregation Level</th>
<th>Moran's I</th>
<th>MST Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block Group</td>
<td>.626174390</td>
<td>421,238</td>
</tr>
<tr>
<td>Tract</td>
<td>.575251639</td>
<td>91,799</td>
</tr>
</tbody>
</table>

The MST results can be compared with the Moran's I results and the density maps, which both indicate a clustered pattern of residential segregation. As shown on Table 7, the MST value at the block group level is 421,237. This value means little until it is compared with the results of random permutations of the data, as was shown in the preceding chapter for the constructed datasets. The random permutations were performed in the same manner as for the constructed data. The locational relationships between the areas (i.e. block groups) were maintained while the data values were randomly shuffled for each iteration of the MST algorithm. The constructed data and the tract level census data were subjected to 1000 random permutations but, due to the sheer size of the problem, only 418 permutations were possible at the block group level. However, this provides ample data for evaluation of the actual MST value versus the randomly generated MST values.

As shown on Table 8, the mean MST value of the random permutations on the block group level data for Franklin County is 878,211 (including the actual MST value). This is compared to the actual MST value of 421,237. Without the actual MST value included, the mean value for the randomly generated results is 879,304, with a range of random MST values from 790,296 to 943,261. The standard deviation of 32,236 places the actual MST value well below the range which could be expected from a random arrangement of the data. Since low MST values indicate clustering, the results of the randomly generated MST...
values clearly indicate a clustered pattern and further substantiate the results of the Moran's I method.

Table 8. Random Permutation Results for Franklin County.

<table>
<thead>
<tr>
<th>Agg Level</th>
<th>MST Value</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>Stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block Group</td>
<td>421,237</td>
<td>878,210.96</td>
<td>421,237</td>
<td>943,261</td>
<td>32,235.66</td>
</tr>
<tr>
<td></td>
<td>w/o</td>
<td>879,304.20</td>
<td>790,296</td>
<td>943,261</td>
<td>23,230.43</td>
</tr>
<tr>
<td>Tract</td>
<td>91,799</td>
<td>166,533.02</td>
<td>91,799</td>
<td>191,190</td>
<td>8954.90</td>
</tr>
<tr>
<td></td>
<td>w/o</td>
<td>166,607.76</td>
<td>136,073</td>
<td>191,190</td>
<td>8641.42</td>
</tr>
</tbody>
</table>

* w/o under "MST Value" indicates that the following statistics do not include the actual MST value.

Figure 24 shows graphically the relationship between the actual MST value, based on the actual calculated pattern of nonwhite population density, and the random permutations of MST values for the block group level data. The frequency histogram in Figure 24 shows the distribution of the random permutations of the MST results not including the actual MST value. The distribution is fairly normal (with one outlier category) and is clustered about the mean value of about 879,304. The actual MST value of 421,238 would be well off of the low end of the graph were it included. In fact, Figure 25 shows the resulting frequency histogram when the actual value is included. The value is shown in the far left, though, as expected, the distribution is not normal when the actual value is included. The very low value of the actual MST value in relation to the random permutations of the
Figure 24. MST Permutation Frequencies: Franklin County Block Groups, Without Actual MST Value
Figure 25. MST Permutation Frequencies: Franklin County Block Groups, With Actual MST Value
Figure 26. MST Permutation Frequencies: Franklin County Tracts, Without Actual MST Value
Figure 27. MST Permutation Frequencies: Franklin County Tracts, With Actual MST Value
nonwhite population density variable strongly illustrates the clustered pattern indicated by the MST results.

The computed MST value at the tract level is 91,799 (Table 8). This, when compared to the random permutations' mean value of 166,533, again indicates a clustered pattern. With a standard deviation of 8955 (including the actual value), it is highly unlikely that the low MST value found for the actual configuration of the data could have resulted from any random arrangement of the same data. Figure 26 shows the frequency histogram for the 1000 tract level random permutations, not including the actual MST value of 91,799. In general, the distribution of the randomly generated MST values is normal and centered about the mean MST value of 166,608. As figure 27 (which includes the calculated MST value) shows, the actual MST value is well out of range of the randomly generated MST values' distribution. Even with inclusion of the actual MST value, the distribution of the main cluster of values is normal about the mean.

One difference between the random permutation results for the constructed data and the Franklin County census data are that the frequency distributions for the constructed data were more normally distributed at the lower aggregation levels. For the Franklin County data, the frequency distribution was closer to normal at the tract level than at the block group (lower aggregation) level, though the block group level distribution was not significantly abnormal. The reason for this might be traced to the relative numbers of each for the constructed versus the actual census data. This is not problematic when one considers that, at both the block group and tract levels, the actual calculated MST value falls well outside a relatively normal distribution of randomly generated MST values. Therefore, the MST results show that clustering of nonwhite population density is strongly evident at the block group level, as well as at the tract level in Franklin County.
5.5. Comparison of Moran's $I$ and MST Methods

A stated purpose of this research is to compare the usefulness of Moran's $I$ and the MST method for discerning spatial pattern in general, and residential segregation patterns in particular. The results presented here using Franklin County census data and in the preceding chapter using constructed data patterns show that both Moran's $I$ and the MST method provide valid indications of spatial pattern under certain conditions. These conditions, as hypothesized, are related to the level of aggregation present in the data. Census data are available at different aggregation levels, and this research indicates that the choice of aggregation level can cause misinterpretation or the inability to interpret spatial patterns for both of the methods used here. These effects are similar for both the more commonly used Moran's $I$ (which has been documented to exhibit aggregation effects) and the MST method proposed in this research.

Moran's $I$ has been used here as an easily comparable method to the MST method and also to demonstrate its usefulness in segregation studies. By using both Moran's $I$ and the MST approach, it is possible to gauge the effectiveness of the MST method with a standard measure of spatial autocorrelation. In the applications presented in this research, Moran's $I$ and the MST method provide similar results. For instance, both are able to easily detect the patterns presented here at low levels of aggregation. This supports their use for future studies of spatial pattern (though Moran's $I$ is already proven in this regard) and illustrates their usefulness particularly in segregation studies, which have been conducted using common indices which fail to consider the spatial arrangement of the units under study.

The methods presented are also similar in terms of their usefulness when using more highly aggregated data units. Both methods have slight aggregation effects if the underlying
pattern is clustered, with Moran's $I$ decreasing in absolute value and the MST method producing MST values only slightly closer to the average values found for random permutations of the data. In the case of clustered patterns, however, these aggregation effects do not interfere with interpretation, as both methods clearly indicate clustering. In fact, graphing the MST method still provides, even under these aggregation effects, a very clear illustration of how clustered the pattern is in relation to random permutations of the data. On the other hand, for patterns which are not clustered, such as the constructed checkerboard pattern, both methods are subject to aggregation errors which obscure the underlying pattern. Aggregation of areal census data must be considered when applying these methods, further substantiating the body of literature which points to data aggregation as a serious problem for spatial analysts.

The census results reported here for the block group and tract level data are sufficient to show that the MST method works as expected in discerning clustered patterns and to point out possible aggregation effects. Given the above description of those aggregation effects, it is apparent that the lowest level of aggregation possible should be used, particularly if the pattern is not likely to be clustered. The block level results are not presented here because both methods, as they are used in this research, require a matrix representation of the locational relationships among the areas. Recall that for both Moran's $I$ and the MST method, a matrix, $W$, is created indicating the $ij$ pairs between all areas in the region. For the tract level data, this presents no problem. Even at the block group level (942 block groups), both methods work well. However, at the block level, this matrix becomes so large (there are approximately 11,500 census blocks in Franklin County) that it is computationally infeasible to use the matrix within any reasonable amount of computer time or memory. For example, the approximately 11,500 blocks for Franklin County requires
an 11,500 * 11,500 matrix, which is the equivalent of over 500 million bytes of memory on the IBM 3090 mainframe computer.

Herein lies an important difference between the Moran's I index and MST method. While the formulation of Moran's I requires the use of the n * n matrix (n being the number of areas), it should be possible to reformulate the program used for the MST method in future applications to read the same input information in a vector format. While the n * n matrix is used in this research for compatibility with the Moran's I method, future applications could reduce the matrix in the MST program to the comparable 3 vectors of length n, representing the list of neighbors (ij pairs from the W matrix) and their respective edge costs. This will greatly reduce the memory and computation time required to calculate the MST value. Therefore, the MST method is potentially capable of discerning pattern in very large areal datasets. This should prove useful to those interested in determining patterns over large regions or, as is strongly suggested by the results of this research, using the lowest level of census aggregation possible to investigate pattern.

While the ability of each method to discern pattern at low levels of aggregation is similar, one difference between the methods is in their ease of interpretation. While Moran's I has readily definable bounds and values to indicate spatial patterns, the MST method requires either a comparison between MST values for two regions or a comparison of the MST value actually found for a region with the results of random permutations of the data. This is not prohibitive for reasonable size datasets. Indeed, it is probable that even larger datasets can be processed using the MST method than with Moran's I. Furthermore, the resulting statistics and graphical representations of the randomly generated MST values are quite useful and capable of detecting pattern when aggregation levels are low. In fact, this
A useful avenue for further research using the MST method is to examine the sensitivity of the MST value to the size of the dataset used (i.e. the size of the graph and the number of edges) and the range of data values for the study area, particularly if comparative studies are desired. These considerations can affect the magnitude of the MST value for a given region. Assuming the further refinement of the MST method to allow application to small datasets as well as extremely large datasets, there is no reason why the MST value cannot be used to aid in actually devising regions such as the "optimal zones" discussed in Chapter II. In other words, the value of the minimum spanning tree for alternative zonal aggregation schemes can be compared to find the most homogeneous zones for a particular study or application.

5.6. Conclusion

The results given here indicate that, based on the MST values at both the block group and tract levels, a strong degree of nonwhite population clustering exists in Franklin County. This is supported by the Moran's $I$ statistic and the density maps provided. Furthermore, the MST method is able to detect the clustered pattern quite well. Both the MST method and the Moran's $I$ index are capable of discerning a clustered pattern using census data, which have irregular properties which must be taken into consideration, particularly at higher levels of aggregation. The combination of the density variable, accounting for different sized census units, and the adjacency criterion for determining neighbors, helps both methods discern the exhibited pattern.
The aggregation effects found with the constructed data patterns are not as obvious for the Franklin County data. While the Moran's $I$ value does decrease with aggregation to the tract level, the MST values computed for the data fall well below those found for the random permutations of the data. There is no indication that the tract level value is any closer to the "random" distribution than is the block group level MST value. The tract level value for the constructed data fall closer to the distribution of random MST values, indicating some aggregation effect, but this is not evident with the Franklin County census data used here. Indications of this analysis are that, if the pattern is reasonably expected to be clustered, both methods are able to discern that clustering, even at higher aggregation levels. Since Moran's $I$ is an index with a fixed range, the aggregation effects are more easily spotted than with the MST value.

Finally, some comparisons between the Moran's $I$ and MST methods are presented. Both methods have similar aggregation effects and are similarly able to detect pattern at low levels of aggregation, and clustered patterns at higher levels of aggregation. The Moran's $I$ value is readily interpretable, but the use of random permutations with the MST method provides ample information for discerning pattern and is not difficult or prohibitive. The lowest level of aggregation is recommended for either method. Use of the lowest census aggregation level (i.e. block level) might require a change in the MST method away from a matrix notation of "neighbors" and towards a vector notation, thereby creating the ability to work with very large datasets using the MST method. A further researchable question relates to the use of these methods for investigating patterns over smaller areas. If the sensitivity to problem size for the MST method can be determined, this method should provide a useful analytical tool for devising homogeneous zones.
CHAPTER VI
CONCLUSIONS

This research has provided a comparison of the usefulness of two measures, the spatial autocorrelation statistic Moran's $I$ and the proposed minimum spanning tree (MST) method, for determining spatial pattern. The specific spatial pattern examined here is the degree of residential segregation, or clustering of white and nonwhite population, in Franklin County, Ohio. While the ability of Moran's $I$ to discern pattern is not in doubt, the use of that measure provides a very important benchmark to test the usefulness of the MST method. In addition, an express purpose of this research was to determine the effects for each method of using different levels in the census hierarchy (i.e. different aggregation levels). Aggregation or scale effects have been investigated for Moran's $I$ previously and it was hypothesized that aggregation effects would arise with the MST method as well. The goal was to discover not only whether the MST method can detect pattern as proposed, but how each method is affected by scale or aggregation of data, and specifically whether that aggregation interferes with the method's ability to discern spatial pattern.

The research was carried out in two phases, one consisting of the application of the two methods to measuring specifically designed data patterns. This allowed a controlled application of the methods and provided additional information to be compared with the findings of the second phase of research. The second phase consisted of the application of
the two methods to determine residential segregation patterns in Franklin County, Ohio. This application represents only one of many for which these methods could be used, as the issue of spatial pattern is of interest to many social scientists. However, the use of these methods to examine segregation was deemed appropriate in part because of the fact that many commonly used measures of segregation, particularly the index of dissimilarity, fail to incorporate the spatial aspect of pattern. Both Moran's $I$ and the MST method explicitly consider the spatial arrangement and locational relationships between the areas under study. This is one reason the Moran's $I$ statistic was chosen as the most appropriate benchmark against which to compare and contrast the MST method.

After a discussion of the problem to be considered, that of measuring spatial pattern with the methods described, Chapter II provided a review of the issues surrounding the measurement of pattern, including the general issue of data aggregation. Data aggregation was discussed both in relation to modeling using aggregated data and to discerning patterns. The effects of using aggregated data with various measures and statistics was discussed along with the issue of properly using aggregated data. Since aggregated data are a fact of life for many social scientists and certainly geographers, some issues related to proper aggregation techniques for modelers were presented as well as a discussion of the desire to create appropriate zoning schemes using aggregated data and of creating the most informative and, therefore, "optimal" zones (Openshaw, 1977). Also discussed were methods for clustering data, including some that are similar to the MST method described here.

The opposite issue of creating clusters is that of discerning them or of discerning pattern in general. A brief discussion about point pattern analysis and methods to discern such patterns was followed by examples of how certain patterns have been defined. This is a
complex issue, since various names have been attached to the same patterns by different researchers. Specifically, the issue of what constitutes homogeneity arises in reference to Shier's (1982) description of homogeneity as resembling a checkerboard pattern, with dissimilar neighbors. This does not match the more intuitive view that homogeneity indicates similar neighbors.

After a brief discussion of past applications of graph theoretic techniques to other problems, the general issue of segregation was discussed, with particular attention paid to how that segregation is measured. Though there have been attempts to use spatial approaches for measuring such segregation (most based on distance related measures), there is a tendency to use what are essentially aspatial measures, specifically the index of dissimilarity or related indices. It is suggested that this continued use of the index of dissimilarity is based partially on inertia and its computational ease. While the general spread or concentration of population over areas is considered in these indices, the actual locational relationships of the areas is not. Both methods presented here provide appropriate alternatives or complements to such measures.

Chapter III presents the methods and a brief discussion of the data used in the research. The chapter includes a discussion of spatial autocorrelation in general and Moran's I in particular. Included in that discussion are issues related to appropriate definitions of contiguity when measuring pattern. Various alternatives were proposed and it was pointed out that, in terms of providing a descriptive measure of pattern, the binary contiguity matrix provides a good formulation. This research examined the use of two alternative measures of contiguity, one based on distance and the other on existence of a common boundary. Both alternatives result in a binary matrix representation.
Chapter III also provided a brief discussion of graph theoretic definitions required to explain the second approach used here, that using the minimum spanning tree (MST). Having defined the necessary terms, examples were given to illustrate how the MST method can detect pattern. In particular the MST's ability to indicate a clustered versus checkerboard type pattern was described and illustrated, where a low total cost for the minimum spanning tree indicates similar neighbors (hence, a clustered pattern) and a high total MST cost indicates dissimilar neighbors (or a checkerboard pattern). Finally, Chapter III provided a discussion of some common concerns for both methods and the use of geographic areal data in general. It is argued that, for the purely descriptive purposes of this research, the problem of aggregation or scale is the most relevant and some of the other issues can be related to this. Other problems, such as edge effects and sample size, become more significant if the express purpose of the research is in helping to formulate relational models using areal data. Those problems deemed important to this experimental, descriptive application were accounted for through the use of regular areas in the constructed data, and adjacency files and density measures for the census application.

Analyses of the two research phases were presented in Chapters IV and V. Chapter IV presents results from the application of Moran's $I$ and the MST method to constructed data patterns. This step in the research provided a controlled approach to examining both the usefulness of each method in detecting the patterns and the effects for each of using different levels of aggregation. Both approaches to determining neighbors were used in this phase, one a distance based contiguity criterion and the other an adjacency based criterion. Both methods were able to discern the patterns tested at lower levels of aggregation. This was not in doubt for Moran's $I$, but use of Moran's $I$ and the
constructed data patterns provided the best gauge against which to measure the effectiveness of the MST method.

While both methods are able to detect the patterns tested at low levels of aggregation, the use of higher aggregation levels makes interpretation more difficult. Results of this research indicate that, for a clustered pattern, aggregation has less effect than for a checkerboard pattern. Aggregation reduces the intensity of the results, but does not interfere with interpretation. Both methods, even at the highest level of aggregation can discern a clustered pattern. However, aggregation can totally obscure the checkerboard type pattern and the results for both methods indicate a random or "nonpattern". This finding emphasizes the desirability of using the lowest level of census aggregation possible for a particular study. Finally, the adjacency contiguity criterion was found to be more realistic and was used exclusively in the census application.

Results from the second research phase, the census application, were presented in Chapter V. This application of the Moran's I and MST methods to Franklin County census data builds on and reinforces the findings from the constructed data analysis. Since the underlying pattern of the test data (nonwhite population density) was clustered, this analysis phase substantiates the previous findings relative to a clustered pattern type. Both methods clearly detect the clustered pattern at both the block group and tract levels of aggregation. Moran's I decreased somewhat with aggregation, but both it and particularly the MST method indicate clustering of nonwhite population density. This phase of the research shows that either method provides a good indication of residential segregation patterns, particularly at low levels of aggregation. In addition, either method is an improvement on techniques that do not consider the spatial arrangement of the areas under study and should provide a useful, complementary measure of residential segregation.
While the results of this research show similarities between Moran's $I$ and the MST method, there are also some basic differences. In terms of interpretation, Moran's $I$ presents a fixed range of possible values, with associated levels of interdependence and, hence, associated patterns. In contrast, the MST method provides a number rather than an index. This number must be compared with other MST values to provide any interpretative ability. The MST values for two patterns over the same region can be compared, as with the constructed data. For any single region, random permutations of the data provide a distribution of possible MST values with which to compare the actual value. This is not at all prohibitive and provides an array of information from which to glean evidence of pattern. Furthermore, this distribution of random permutation values provides a way to graphically illustrate the relative magnitude of the actual MST value found for a region.

The results given in Chapter V indicate another important difference between the two methods. In this research, an $n \times n$ matrix (n being the number of areas in the study region) was used to depict the locational relationships between areas in the region. While calculation of the Moran's $I$ statistic requires the use of that $n \times n$ matrix, future applications of the MST method can be formulated to allow the use of larger data sets. The MST method has the potential to be used with either larger areas or with lower levels of aggregation (which equate to larger areas). This latter potential, in particular, is an important ability, as it has been shown here that aggregation can obscure underlying patterns. It is preferable to use the lowest possible level of aggregated data and the MST method provides the potential to do so even when examining pattern over large areas.

The MST method has been shown to be an effective indicator of pattern when data are examined at lower aggregation levels. The implication is that this method provides the
spatial analyst with another tool, particularly for descriptive purposes. There are several interesting avenues for future research using this method. Foremost among these is the potential application of this method to larger data files, opening the way for its use with block level data in segregation or other pattern oriented studies. Conversely, the MST method's ability to work on small data sets is unclear, since the number of edges in a region will affect the MST value. Therefore, future research directions include refinement of the technique to be used on small areas as well as larger data sets. If the sensitivity of the MST value to region or graph size can be determined, this should provide an additional approach to devising homogeneous regions such as those discussed in Chapter II, by comparing MST values found for alternative aggregation schemes.

Furthermore, determining the sensitivity of the MST method to some of the other problems with areal units of analysis provides a further research question. The issue of scale or aggregation level has been investigated here and found to be an important consideration. Though this was not an express purpose of the research, application of the MST method to the regular areas of the constructed data in comparison to the census data do not indicate substantial differences in the ability of the method to discern pattern related to other areal issues (such as irregular areas) in a purely descriptive application such as this. However, further study can substantiate the sensitivity of the MST value to these problems for those interested in more than descriptive pattern detection. Finally, these results provide further indication of the aggregation effects on geographical models in general and spatial analytical tools in particular. These effects were noted for two distinct patterns and future applications can further examine these effects on different types of patterns.

The methods presented here are not proposed necessarily as replacements for existing methods, but they provide very appropriate alternatives or complementary approaches in
studies such as the residential segregation application presented here. They do not attempt to solve, in this application, all of the problems inherent to using areal data such as census units. However, the use of such data by social scientists, particularly geographers, will likely continue. For those interested in determining the underlying pattern exhibited by census data, the MST method has been shown to be effective if aggregation levels are considered. This holds as well for the Moran's $I$ index. Furthermore, consideration of the open and researchable questions discussed above should improve a useful tool for those interested in detecting spatial pattern in general.
1. // JOB
2. // REGION=3072K
3. // EXEC VSF2CLG
4. //FORT.SYSIN DD *
5. COMMON W(942,942), SWIDOT(942), SWDOTI(942)
6. CHARACTER*4 TSTRING(252), TLEFT, TRIGHT
7. CHARACTER*5 BGSTRING(942), BGLEFT, BRIGHT
8. CHARACTER*8 BSTRING(942), BLEFT, BRIGHT
9. REAL LAT(942), LONG(942), POP(942), NONWHT(942)
10. REAL DENSITT(942)
11. INTEGER L, N, I, J
12. N=0
13. L=5
14. C NOTE: The assignment to L above depends on the lowest level
15. C OF GEOGRAPHICAL UNITS USED. USE A VALUE OF 4 FOR
16. C TRACTS, 5 FOR BLOCK GROUPS, OR 8 FOR BLOCKS.
17. 100 FORMAT(2X,A4,1X,F9.6,1X,F10.6,1X,F9.0,1X,F9.0,1X,F9.1)
18. 200 FORMAT(2X,A5,1X,F9.6,1X,F10.6,1X,F9.0,1X,F9.0,1X,F9.1)
19. 300 FORMAT(2X,A8,1X,F9.6,1X,F10.6,1X,F9.0,1X,F9.0,1X,F9.1)
20. C The following IF structure sends the appropriate information
21. C TO SUBROUTINE COMPUTE1 BASED ON THE ASSIGNMENT OF L.
22. IF (L.EQ.4) THEN
23. 1 N=N+1
24. READ(1,100,END=999) TSTRING(N), LAT(N), LONG(N), POP(N),
25. + NONWHT(N), DENSITY(N)
26. GO TO 1
27. 999 N=N-1
28. CALL COMPUTE1(TSTRING, LAT, LONG, POP, NONWHT, DENSITY, N, L, TLEFT,
29. + TRIGHT, W, SWIDOT, SWDOTI)
30. ELSE IF (L.EQ.5) THEN
31. 2 N=N+1
32. READ(2,200,END=998) BGSTRING(N), LAT(N), LONG(N), POP(N),
33. + NONWHT(N), DENSITY(N)
34. GO TO 2
35. 998 N=N-1
36. CALL COMPUTE1(BGSTRING, LAT, LONG, POP, NONWHT, DENSITY, N, L, BLEFT,
37. + BRIGHT, W, SWIDOT, SWDOTI)
38. ELSE IF (L.EQ.6) THEN
39. 3 N=N+1
40. READ(3,300,END=997) BSTRING(N), LAT(N), LONG(N), POP(N),
41. + NONWHT(N), DENSITY(N)
42. GO TO 3
43. 997 N=N-1
44. CALL COMPUTE1(BSTRING, LAT, LONG, POP, NONWHT, DENSITY, N, L, BLEFT,
45. + BRIGHT, W, SWIDOT, SWDOTI)
ELSE

PRINT *, 'IMPROPER VALUE FOR L'
END IF
END

SUBROUTINE COMPUTE1(STRING, LAT, LONG, POP, NONWHT, DENSITY, N, L, LEFT, + RIGHT)

COMMON W(942,942), SWIDOT(942), SWDOTI(942)
INTEGER I, J, K, L, NR, NC
CHARACTER*(* ) LEFT, RIGHT, STRING(N)
REAL LAT(N), LONG(N), POP(N), NONWHT(N), DENSITY(N)
REAL IDENOM, S0, S1, S2, DIST, SUM2, ZBAR, KR, SIGSQ, TSQ
REAL M2, M4, B2, MORAN, Z1, VARN1, VARR1, SNI, SRI, SUM1
SUM1=0.0
SUM2=0.0

DO 71 I=1, N
SUM1=SUM1+DENSITY(I)
SUM2=SUM2+(DENSITY(I)**2)
CONTINUE

KR=N
ZBAR=SUM1/KR
SIGSQ=(SUM2/KR)-(ZBAR**2)
SIGSQ=(SUM2-(ZBAR*SUM1))/(KR-1)

DO 74 I=1, N
DO 74 J=1, N
W(I,J)=0.0
SWIDOT(I)=0.0
SWDOTI(I)=0.0
CONTINUE

NR=0
NC=0
S0=0.0
S1=0.0
S2=0.0
IDENOM=0.0

IF (L.EQ.4) THEN
201 READ (11,101,END=996) LEFT, RIGHT, DIST
IF (LEFT.EQ. ' O' .OR. RIGHT.EQ. ' O') THEN
GO TO 201
ELSE IF (DIST.EQ.0) THEN
GO TO 201
ELSE
DO 76 I=1, N
IF (STRING(I).EQ.LEFT) THEN
NR=I
GO TO 77
END IF
CONTINUE
DO 78 I=1, N
IF (STRING(I).EQ.RIGHT) THEN

END
NC = 1
GO TO 79
END IF
123. 78 CONTINUE
124. END IF
125. W(NR, NC) = 1.0
126. GO TO 201
127. ELSE IF (L .EQ. 5) THEN
128. READ (12, 102, END=996) LEFT, RIGHT, DIST
129. IF (LEFT .EQ. ''' 0' OR RIGHT .EQ. ''' 0' ) THEN
130. GO TO 301
131. ELSE IF (DIST .EQ. 0) THEN
132. GO TO 301
133. ELSE
134. 80 DO 80 I = 1, N
135. IF (STRING(I) .EQ. LEFT) THEN
136. NR = I
137. GO TO 81
138. END IF
139. CONTINUE
140. 81 DO 82 I = 1, N
141. IF (STRING(I) .EQ. RIGHT) THEN
142. NC = I
143. GO TO 83
144. END IF
145. CONTINUE
146. 82 END IF
147. 83 W(NR, NC) = 1.0
148. GO TO 301
149. ELSE IF (L .EQ. 8) THEN
150. READ (13, 103, END=996) LEFT, RIGHT, DIST
151. IF (LEFT .EQ. ''' 0' OR RIGHT .EQ. ''' 0' ) THEN
152. GO TO 401
153. ELSE IF (DIST .EQ. 0) THEN
154. GO TO 401
155. ELSE
156. 84 DO 84 I = 1, N
157. IF (STRING(I) .EQ. LEFT) THEN
158. NR = I
159. GO TO 85
160. END IF
161. CONTINUE
162. 85 DO 86 I = 1, N
163. IF (STRING(I) .EQ. RIGHT) THEN
164. NC = I
165. GO TO 87
166. END IF
167. CONTINUE
168. 86 END IF
169. 87 W(NR, NC) = 1.0
170. GO TO 401
171. ELSE
172. PRINT * , 'L IS NOT AN ACCEPTABLE VALUE'
173. END IF
174. 101 FORMAT (2(6X, A4), 6X, F4.0)
175. 102 FORMAT (2(5X, A5), 6X, F4.0)
176. 103 FORMAT (2(2X, A8), 6X, F4.0)
177. 996 CONTINUE
DO 88 I=1, N
DO 88 J=1, N
SWIDOT(I)=SWIDOT(I)+W(I,J)
SWDOT(J)=SWDOT(J)+W(I,J)
80=S0+W(I,J)
81=S1+(W(I,J)+W(J,I))**2
82=IDXMT((DENSITY(I)-ZBAR)*(DENSITY(J)-ZBAR)*W(I,J))
88 CONTINUE
MORAN=IDXMT/(S8Q*S0)
89 CONTINUE
S1=S1/2
89 CONTINUE
EI=1/(KR-1)
90 M4=S8Q
M4=0.0
91 CONTINUE
90 M4=M4+((DENSITY(I)-ZBAR)**2)**2
EI=1/(KR**2)
M4=M4/(EI**2)
SNI=(M4-KR**2)/(S1-KR**2+3*(S0**2))
SRI=(SNI-KR**2)/(S1-KR**2+3*(S0**2))
91 CONTINUE
90 SNI=(M4-KR**2)/(S1-KR**2+3*(S0**2))
91 CONTINUE
90 SRI=(SNI-KR**2)/(S1-KR**2+3*(S0**2))
91 CONTINUE
END IF
90 IF (L.EQ.4) THEN
PRINT *, 'Statistics for County measured at Tract Level'
90 ELSE IF (L.EQ.5) THEN
PRINT *, 'Statistics for County measured at Block Group Level'
90 ELSE IF (L.EQ.8) THEN
PRINT *, 'Statistics for County measured at Block Level'
91 END IF
90 PRINT *, 'MORANS I = ', MORAN
PRINT *
PRINT *, 'VARNI (Resampling) = ', VARNI
PRINT *, 'VARNI (Randomization) = ', VARNI
PRINT *, 'SNI = (MORAN-EI)/SQR(VARNI)
PRINT *, 'SNI = (MORAN-EI)/SQR(VARNI)
PRINT *, 'SRI (Resampling) = ', SRI
PRINT *, 'SRI (Randomization) = ', SRI
240.             RETURN
241.             END
242.             // GO.FT01F001 DO DSN=TNEWSO.DIS.TRACT.DATA2,
243.             // DISP=SHR
244.             // GO.FT02F001 DO DSN=TNEWSO.DIS.BKGRP.DATA2,
245.             // DISP=SHR
246.             // GO.FT03F001 DO DSN=TNEWSO.DIS.BLOCK.DATA,
247.             // DISP=SHR
248.             // GO.FT11F001 DO DSN=TNEWSO.DIS. ADJNEW. TRACT,
249.             // DISP=SHR
250.             // GO.FT12F001 DO DSN=TNEWSO.DIS. ADJNEW. BKGRP,
251.             // DISP=SHR
252.             // GO.FT13F001 DO DSN=TNEWSO.DIS. ADJNEW. BLOCK,
253.             // DISP=SHR
254.             // EXEC WNOTIFY
// JOB,
// REGION=6144K.TIME=(10,0)
// JOBPARM LINES=10000,DISKIO=10000,SERVICE=*
// EXEC VSF2CLG,PARM,FORT=DC(TRACY),OPT(3),
// PARM,LIKE=ANMODE=31,AMODE=ANY,TIME=10
//FORT.SYsin DD *
C PROCESS DC(TRACY)
C COMBINED FROM DIS.PROG AND KRU.S.FORT
C
CHARACTER*4 TSTRING(252),TLEFT,TRIGHT
CHARACTER*5BGSTRING(942),BGLEFT,BRIGHT
CHARACTER*BSTRING(942),BLEFT,BRIGHT
REAL LAT(942),LONG(942),POP(942),NONWHT(942)
REAL DENSITY(942)
INTEGER L,N,I,J
COMMON /TRACT/ W(942,942),W(942,942)
N=0
L=8

C NOTE: The assignment to L above depends on the lowest level
C OF GEOGRAPHICAL UNITS USED. USE A VALUE OF 4 FOR
C TRACTS, 5 FOR BLOCK GROUPS, OR 8 FOR BLOCKS. IF THE
C tract vs. block level is desired, use value 3.
C
100 FORMAT(2X,A4,1X,F9.6,1X,F10.6,1X,F9.0,1X,F9.0,1X,F9.1)
200 FORMAT(2X,A5,1X,F9.6,1X,F10.6,1X,F9.0,1X,F9.0,1X,F9.1)
300 FORMAT(2X,A8,1X,F9.6,1X,F10.6,1X,F9.0,1X,F9.0,1X,F9.1)
27.
C The following IF structure sends the appropriate information
C TO SUBROUTINE COMPUTE1 BASED ON THE ASSIGNMENT OF L.
C
IF (L.EQ.4) THEN
1 N=N+1
READ(1,100,END=999) TSTRING(N),LAT(N),LONG(N),POP(N),
+ NONWHT(N),DENSITY(N)
GO TO 1
999 N=N-1
ELSE IF (L.EQ.5) THEN
2 N=N+1
READ(2,200,END=998) BGSTRING(N),LAT(N),LONG(N),POP(N),
+ NONWHT(N),DENSITTY(N)
GO TO 2
998 N=N-1
ELSE IF (L.EQ.8) THEN
3 N=N+1
READ(3,300,END=997) BSTRING(N),LAT(N),LONG(N),POP(N),
+ NONWHT(N),DENSITY(N)
GO TO 3
997 N=N-1
ELSE
PRINT *, 'IMPROPER VALUE FOR L'
END IF
CALL COMPUTE1(BSTRING,LAT,LONG,POP,NONWHT,DENSITY,N,L,BLEFT,
SUBROUTINE COMPUTE1(STRING, LAT, LONG, POP, NONWHT, DENSITY, N, L, 
+ LEFT, RIGHT)
  INTEGER LX(4250),RY(4250),IT
  DIMENSION WW(4250)
  INTEGER I, J, K, NR, NC, IT
  REAL LAT(N), LONG(N), POP(N), NONWHT(N), DENSITY(N)
  REAL DIST, RAND(942), RAND2(942), DENS2(942)
  COMMON /TRACT/ W(942,942), W2(942,942)
  DO 56 I=1,N
    DENS2(I)=DENSITY(I)
  56 CONTINUE
  DO 55 I=1,N
    PERM(I)=I
  55 CONTINUE
  DO 65 IT=1,1001
    IF (IT.EQ.1) THEN
      GOTO 7
    ELSE
      CALL RNUN(N,RAND)
      CALL 3VRBP(N,RAND,RAND2,PERM)
      CALL PERMU(N,DENSITY,PERM,1,DENS2)
      END IF
    IF (IT.EQ.1) THEN
      DO 71 I=1,N
        DO 71 J=1,N
          W(I,J)=0.0
          W2(I,J)=0.0
        71 CONTINUE
        IF (L.EQ.4) THEN
          CALL 201
        ELSE IF (DIST.EQ.0) THEN
          GO TO 201
        ELSE IF (L.EQ.5) THEN
          CALL 301
        ELSE
          GO TO 88
        END IF
      70 IF (IT.EQ.1) THEN
        DO 71 I=1,N
          DO 71 J=1,N
            W(I,J)=0.0
            W2(I,J)=0.0
          71 CONTINUE
          IF (L.EQ.4) THEN
            CALL 201
          ELSE IF (DIST.EQ.0) THEN
            GO TO 201
          ELSE IF (L.EQ.5) THEN
            CALL 301
          ELSE
            GO TO 88
          END IF
        70 CONTINUE
      79 W(NR,NC)=1.0
      GO TO 201
    ELSE IF (L.EQ.5) THEN
      CALL 301
      GO TO 201
  END IF
120.  GO TO 301
121.  ELSE IF (DIST.EQ.0) THEN
122.    GO TO 301
123.  ELSE
124.    DO 80 I=1, N
125.    IF (STRING(I).EQ.LEFT) THEN
126.      NR=I
127.      GO TO 81
128.    END IF
129.    CONTINUE
130. 80  DO 82 I=1, N
131.    IF (STRING(I).EQ.RIGHT) THEN
132.      NC=I
133.      GO TO 83
134.    END IF
135. 82  CONTINUE
136.  END IF
137. 83  W(NR,NC)=1.0
138.  GO TO 301
139.  ELSE IF (L.EQ.0) THEN
140.    READ (13,103,END=996) LEFT, RIGHT, DIST
141.    IF (LEFT.EQ.'O'.OR.RIGHT.EQ.'O') THEN
142.      GO TO 401
143.    ELSE IF (DIST.EQ.0) THEN
144.      GO TO 401
145.    ELSE
146.    DO 84 I=1, N
147.    IF (STRING(I).EQ.LEFT) THEN
148.      NR=I
149.      GO TO 85
150.    END IF
151. 84  CONTINUE
152. 85  DO 86 I=1, N
153.    IF (STRING(I).EQ.RIGHT) THEN
154.      NC=I
155.      GO TO 87
156.    END IF
157. 86  CONTINUE
158.  END IF
159. 87  W(NR,NC)=1.0
160.  GO TO 401
161.  ELSE
162.    PRINT *, 'L IS NOT AN ACCEPTABLE VALUE'
163.  END IF
164. 101  FORMAT (2(6X,A4),6X,F4.0)
165. 102  FORMAT (2(5X,A5),6X,F4.0)
166. 103  FORMAT (2(2X,A8),6X,F4.0)
167. 996  CONTINUE
168.  DO 88 I=1, N
169.  DO 88 J=1, N
170.  IF (W(I,J).EQ.1.0.AND.W(J,I).EQ.1.0) THEN
171.    W(J,I)=0.0
172.  END IF
173.  CONTINUE
174.  END IF
175.  CONTINUE
176.  END IF
177.  DO 37 I=1, M
178.  DO 37 J=1, N
W2(I,J)=0.0

DO 89 I=1, N
DO 89 J=1, N
IF (W(I,J).EQ.1.0) THEN
   W2(I,J)=ABS(DENS2(I)-DENS2(J))
END IF
89 CONTINUE

SUMW=0
DO 90 I=1, N
DO 90 J=1, N
IF (W(I,J).EQ.1.0) THEN
   SUMW=SUMW+1
END IF
90 CONTINUE

WRITE (6,104) N, SUMW
104 FORMAT(2(I10))
NN=N
HH=SUMW
I9=0
DO 91 I=1, N
DO 91 J=1, N
IF (W(I,J).EQ.1.0) THEN
   WRITE (6,105) I, J, W2(I,J)
   I9=I9+1
   LTH(I9)=I
   RTH(I9)=J
   WW(I9)=W2(I,J)
END IF
91 CONTINUE
105 FORMAT (I10,I10,F9.0)
CALL BIGKRUS(NH,m,LT,RT,WW)

C SUBROUTINE COMPUTE2(STRING,LAT,LONG,POP,NONWHT,DENSITY)
C RETURN
C SUBROUTINE BIGKRUS(NH,M,LT,RT,WW)

C ******** SAMPLE CALLING PROGRAM FOR SUBROUTINE KRUS ********
C (SHORTEST SPANNING TREE PROBLEM)  
C THE PROGRAM IS BASED ON THE PAPER
C F.M. CAMERINI, C.M. GALBIATI, F. MAFFIOLI
C "ALGORITHMS FOR FINDING OPTIMUM TREES:
C DESCRIPTIONS, USE AND EVALUATION",
C ALL THE SUBROUTINES ARE WRITTEN IN AMERICAN STANDARD FORTRAN AND ARE ACCEPTED BY THE
C PFORT VERIFIER.
C QUESTIONS AND COMMENTS SHOULD BE DIRECTED TO
C PAOLO CAMERINI AND FRANCESCO MAFFIOLI
**DIPARTIMENTO DI ELETTRONICA**

**POLITECNICO DI MILANO, MILANO, ITALY.**

---

SEE SUBROUTINE KRUS AND SPSIMP FOR THE MEANING OF ARGUMENTS

- **LOCAL VARIABLES**
  - **I** CURRENT INDEX
  - **K** INDEX
  - **LIN** INPUT UNIT
  - **LOU** OUTPUT UNIT
  - **MRD** MAXIMUM NUMBER OF ARCS
  - **NRD** MAXIMUM NUMBER OF NODES

- **CALLED SUBROUTINES**
  - SPSIMP DELETES MULTIPLE ARCS AND LOOPS FROM THE GRAPH
  - KRUS

- **TYPE AND DIMENSION STATEMENTS**

  - INTEGER CHECK(942, 942, F0, F(942), FATHER(942), RT(4250), LT(4250),
    - SIZE(942), Q0, Q(4250))
  - REAL SUM

- **GRAPH SIMPLIFICATION**
  - READ (LIN, 100) N, M
  - IAZZ = 1, ICK = 1, IORD = 0
  - IF (M .LE. 0 .OR. M .GT. MRD .OR. N .LE. 0 .OR. N .GT. NRD) STOP
  - WRITE (LOU, 110) N, M

  - CALL SPSIMP(IAZZ, ICK, IORD, N, M, LT, RT, W, CHECK)
    - IF (M .LE. 0) STOP
  - CALL KRUS (N, M, LT, RT, W, F0, F, SIZE, FATHER, Q0, Q)
  - WRITE (LOU, 160) F0
  - SUM = 0
  - SUM = SUM + W(K)
  - WRITE (LOU, 119) SUM

  - 1020 I = 1, F0
  - K = F(I)

  - 100 FORMAT (2110.314)
  - 110 FORMAT (1H ,18HNUHBER OF NODES », 14, 10X, 17HNUHBER OF ARCS », 14)
  - 119 FORMAT(2X, ** * * * SUM OF WEIGHTS », ', F12.5)
  - GO TO 99999
  - 120 FORMAT (2110, F9.0)
  - 150 FORMAT (1H ,22HEND-POINT OUT OF RANGE)
  - 160 FORMAT (1H ,17HNUHBER OF ARCS », 14)
  - 99999 RETURN
SUBROUTINE KRUS (N,M,LT,RT,W,F0,F,SIZE,FATHER,Q0,Q)

SUBROUTINE KRUS COMPUTES A SPANNING TREE OF MINIMUM TOTAL WEIGHT BY KRUSKAL'S METHOD

WARNINGS

THE GRAPH MUST BE CONNECTED

THE END-POINTS OF EACH ARC NEED NOT BE IN ORDER, I.E. LT(I) MAY BE GREATER THAN RT(I) FOR ANY ARC I

LOOPS AND MULTIPLE EDGES ARE ALLOWED, BUT IN THIS CASE IT IS ADVISABLE TO UTILIZE SUBROUTINE SPSIMP

INPUT ARGUMENTS

N NUMBER OF NODES OF THE GRAPH (N.LT.1 IS FORBIDDEN)
M NUMBER OF ARCS OF THE GRAPH (M.LT.0 IS FORBIDDEN)
LT(I) FIRST END-POINT OF ARC I (DIM. M)
RT(I) SECOND END-POINT OF ARC I (DIM. M)
W(I) WEIGHT OF ARC I (DIM. M)

OUTPUT ARGUMENTS

F(I) I-TH ARC OF THE SOLUTION (DIM. N)
F0 CARDINALITY OF F
IF N.LE.1 KRUS RETURNS WITH F0=0

WORKING ARGUMENTS

SIZE(.) TREES DIMENSION VECTOR (DIM. N)
FATHER(.) TREE FRED. VECTOR (DIM. N)
Q(.) VECTOR OF HEAP STRUCTURED ARCS (DIM. M)
QO CARDINALITY OF Q

LOCAL VARIABLES

A CURRENT ARC
J LOOP INDEX
NF NUMBER OF TREES IN THE CURRENT FOREST
X NAME OF THE TREE CONTAINING LEFT NODE OF ARC A
Y NAME OF THE TREE CONTAINING RIGHT NODE OF ARC A

CALLED SUBROUTINES AND FUNCTIONS

INTEGER FUNCTION DFIND, FINDS TREE CONTAINING GIVEN NODE
SUBROUTINE DINIT, Initializes DATA STRUCTURE
SUBROUTINE DUNION, PERFORMS UNION OF TWO TREES
SUBROUTINE HINIT, Initializes HEAP STRUCTURE FOR ARCS
INTEGER FUNCTION HMIN, Extracts AN ARC OF MIN. WEIGHT FROM HEAP

TYPE AND DIMENSION STATEMENTS

INTEGER LT(1),RT(1),W(1),F0(1),FATHER(1)
INTEGER F0,A,X,Y,DFIND,HMIN,Q0,Q(1)
DIMENSION W(1)
SOLUTION SET OF ARCS IS INITIALIZED

\( F_0 = 0 \)

IF \( N \leq 1 \) RETURN

THE FOREST OF \( N \) DISJOINT TREES IS INITIALIZED AND CONTAINS ONE NODE OF THE GIVEN GRAPH FOR EACH ELEMENT

AN ORIENTED TREE STRUCTURE IS USED FOR THE SET OF NODES OF THE TREE. \( NF \) IS THE NUMBER OF TREES IN THE FOREST

CALL \texttt{DINIT}(N,SIZE,FATHER)

\( NF = N \)

HEAP STRUCTURE CONTAINING THE ARCS IS INITIALIZED

CALL \texttt{HINIT}(W,Q0,Q)

ITERATIVE PHASE BEGINS

\[ 10 \] CONTINUE

IF \( NF \leq 1 \) RETURN

\[ 20 \] CONTINUE

A NEW ARC IS EXTRACTED FROM HEAP

\( A = \texttt{HMIN}(W,Q0,Q) \)

THE SETS CONTAINING THE ARC END-POINTS ARE SEARCHED

\( X = \texttt{DFIND}(\text{LE}(A),\text{FATHER}) \)

\( Y = \texttt{DFIND}(\text{RT}(A),\text{FATHER}) \)

THE ARC IS REJECTED IF IT HAS END-POINTS IN THE SAME SET

IF \( X = Y \) GO TO 20

\( F \) IS UPDATED WITH ARC FOUND

\( F_0 = F_0 + 1 \)

\( F(F_0) = A \)

THE TWO SETS OF NODES CONTAINING ARC \( A \) END-POINTS ARE MERGED AND THE NUMBER OF TREES OF THE FOREST IS DECREASED BY 1

CALL \texttt{DUNION}(X,Y,X,SIZE,FATHER)

\( NF = NF - 1 \)

GO TO 10

END

SUBROUTINE \texttt{SPSIMP}(IAZZ,ICK,IORD,N,M,LT,RT,W,CHECK)

SUBROUTINE \texttt{SPSIMP}

(A) ORDERS THE END-POINTS OF EACH ARC (THE LEAST FIRST)

(B) INITIALIZE TO ZERO (IF REQUESTED) ALL OR PART OF THE ENTRIES OF MATRIX CHECK (SEE COMMENTS ON SUBROUTINES AZCHEC & APCHEC FOR)

(C) CALLS SUBROUTINE SIMPL TO DELETE MULTIPLE ARCS AND LOOPS

(D) RESETS TO ZERO ALL OR PART OF THE ENTRIES OF CHECK.
INPUT ARGUMENTS
IAZE FLAG FOR REQUESTING ZERO SETTING OF MATRIX CHECK 0:NO,1:YES
ICK TYPE OF ZERO SETTING 0:PARTIAL 1:TOTAL
IORD FLAG INDICATING IF END-POINTS ARE IN ORDER 0:NO 1:YES
N NUMBER OF NODES

INPUT/OUTPUT ARGUMENTS
M NUMBER OF ARCS
LT(I) FIRST END-POINT OF ARC I (DIM. M)
RT(I) SECOND END-POINT OF ARC I (DIM. M)
W(I) WEIGHT OF ARC I (DIM. M)

WORKING ARGUMENTS
CHECK(,,) WORKING MATRIX - SEE COMMENTS IN SUBROUTINE SIMPL FOR MEA
(DIM. N X N)

CALLED SUBROUTINES
APCHEC PARTIAL ZERO SETTING
AECHEC TOTAL EERO SETTING
ORDNOD ORDERING OF END-POINTS
SIMPL DELETING OF MULTIPLE ARCS AND LOOPS

TYPE AND DIMENSION STATEMENTS
INTEGER CHECK(N,N),RT(1)
DIMENSION LT(1),W(1)

ORDER END-POINTS IF NECESSARY
IF ( IORD .EQ. 0 ) CALL ORDNOD(M,LT,RT)

IF IAZZ=0 THE INITIAL EERO SETTING IS NOT PERFORMED
IF ( IAZZ .EQ. 0 ) GO TO 20
IF ( ICK .EQ. 0 ) GO TO 10
CALL AECHEC(N,CHECK)
GO TO 20

10 CONTINUE
CALL APCHEC(M,LT,RT,CHECK,N)
20 CONTINUE

DELETION OF MULTIPLE ARCS AND LOOPS
CALL SIMPL(M,LT,RT,W,CHECK,N)

ZERO RESETTING
IF ( ICK .EQ. 0 ) GO TO 30
CALL APCHEC(M,LT,RT,CHECK,N)
GO TO 99
30 CONTINUE
CALL APCHEC(M,LT,RT,CHECK,N)
99 RETURN

END


SUBROUTINE SIMPL(M,LT,RT,W,CHECK,N)

SIMPL DELETES MULTIPLE ARCS AND LOOPS

WARNINGS

THE END-POINTS OF EACH ARC MUST BE ORDERED (THE LEAST FIRST)

THE ENTRIES OF MATRIX CHECK MUST BE ZERO (AT LEAST THOSE
CORRESPONDING TO THE END-POINTS OF EACH ARC)

INPUT ARGUMENTS

M NUMBER OF NODES (FOR DIMENSIONING CHECK)

INPUT/OUTPUT ARGUMENTS

M NUMBER OF ARCS

LT(I) FIRST END-POINT OF ARC I

RT(I) SECOND END-POINT OF ARC I

W(I) WEIGHT OF ARC I

CHECK(.,.) WORKING MATRIX - SEE COMMENTS BELOW FOR MEANING

LOCAL VARIABLES

H CURRENT ARC

I FIRST END-POINT OF ARC H

J SECOND END-POINT OF ARC H

K VALUE OF CHECK(I,J)

TYPE AND DIMENSION STATEMENTS

INTEGER LT(M),RT(M),CHECK(N,N),H

DIMENSION W(I)

IF ( M .LT. 1 ) RETURN

FOR EACH I LESS THAN J,

CHECK(I,J)=0 INDICATES THAT NO ARC WITH END-POINTS I,J

HAS BEEN FOUND

CHECK(I,J)=K INDICATES THAT K IS THE LEAST WEIGHTED ARC

CURRENTLY FOUND WITH END-POINTS I,J.

ARCS ARE EXAMINED IN SEQUENCE

H = 1

10 CONTINUE

IF ( H .GT. M ) RETURN

I = LT(H)

J = RT(H)

IF ( I .NE. J ) GO TO 20

DELETE THE LOOP, INSERT THERE THE M-TH ARC

AND UPDATE THE NUMBER M OF ARCS

GO TO 10
20 CONTINUE
K = CHECK(I,J)
IF ( K .EQ. 0 ) GO TO 40

CC DELETE THE MULTIPLE ARC WITH GREATER WEIGHT
CC IF ( W(K) .LE. W(H) ) GO TO 30
CC SUBSTITUTE ARC K WITH ARC H

30 CONTINUE
LT(H) = LT(M)
RT(H) = RT(M)
W(H) = W(M)
M = M - 1
GO TO 10

40 CONTINUE
CHECK(I,J) = H
M = M + 1
GO TO 10

END

SUBROUTINE ORDNOD(M,LT,RT)
ORDNOD ORDERS THE END-POINTS OF EACH ARC - THE LEAST FIRST
I.E. SETS LT(I).LE.RT(I) FOR EACH ARC I.

INPUT ARGUMENTS
M NUMBER OF ARCS

INPUT/OUTPUT ARGUMENTS
LT(I) FIRST END-POINT OF ARC I
RT(I) SECOND END-POINT OF ARC I

LOCAL VARIABLES
H CURRENT ARC
I FIRST END-POINT
J SECOND END-POINT

TYPE AND DIMENSION STATEMENTS
INTEGER LT(1),RT(1),H

DO 1000 H=1,M
I = LT(H)
J = RT(H)
IF ( I .LE. J ) GO TO 1000
LT(H) = J
RT(H) = I
1000 CONTINUE
RETURN
END

SUBROUTINE AECHEC(M,CHECK)
ALCHEC SETS TO ZERO ALL VALUES OF CHECK(I,J) WITH I LESS THAN J
AND J LESS THAN OR EQUAL TO N.

INPUT ARGUMENTS

N    NUMBER OF NODES

OUTPUT ARGUMENTS

CHECK(,,) MATRIX TO BE SET TO ZERO

LOCAL VARIABLES

I    ROW INDEX
J    COLUMN INDEX
JJ    CURRENT COLUMN INDEX - 1

TYPE AND DIMENSION STATEMENTS

INTEGER CHECK(N,N)

IF ( N. LT. 2 ) RETURN
DO 1010 J = 2, N
   JJ = J - 1
   DO 1000 I = 1, JJ
      CHECK(I,J) = 0
   1000 CONTINUE
1010 CONTINUE
RETURN
END

SUBROUTINE APCHEC(M,LT,RT,CHECK,N)

APCHEC SETS TO ZERO ALL VALUES CHECK(I,J) SUCH THAT I AND J
ARE END-POINTS OF SOME ARC.

INPUT ARGUMENTS

M    NUMBER OF ARCS
LT(I)    FIRST END-POINT OF ARC I
RT(I)    SECOND END-POINT OF ARC I
N    NUMBER OF NODES

INPUT/OUTPUT ARGUMENTS

CHECK(,,) MATRIX TO BE SET TO ZERO

LOCAL VARIABLES

H    CURRENT ARC
I    ROW INDEX
J    COLUMN INDEX

TYPE AND DIMENSION STATEMENTS

INTEGER H,LT(1),RT(1),CHECK(N,N)

IF ( M. LT. 1 ) RETURN
DO 1000 H = 1, N
660. I = LT(H)
661. J = RT(H)
662. CHECK(I,J) = 0
663. 1000 CONTINUE
664. RETURN
665. END
666. C
667. INTEGER FUNCTION DFIND(I,FATHER)
668. C
669. DFIND RETURNS THE SET CONTAINING I
670. C
671. INPUT ARGUMENTS
672. C
673. I ELEMENT TO LOCATE
674. C
675. WORKING ARGUMENTS (SHARED BY DINIT, DFIND AND DUNION)
676. C
677. FATHER(.) VECTOR CONTAINING PREDECESSORS OF EACH ELEMENT
678. C
679. TYPE AND DIMENSION STATEMENTS
680. C
681. INTEGER FATHER(I),R,H
682. C
683. R = I
684. 10 IF (FATHER(R) .EQ. 0 ) GO TO 20
685. R = FATHER(R)
686. GO TO 10
687. 20 DFIND = R
688. C
689. THE PATH COMPRESSION PHASE IS NOW EXECUTED
690. C
691. H = I
692. 30 IF (FATHER(H) .EQ. 0 ) GO TO 40
693. K = FATHER(H)
694. FATHER(H) = R
695. H = K
696. GO TO 30
697. 40 RETURN
698. END
699. C
700. SUBROUTINE DINIT(N,SIZE,FATHER)
701. C
702. DINIT Initializes a set J with a tree oriented
703. structure containing the element J, J=1,...,N
704. C
705. INPUT ARGUMENTS
706. C
707. N NUMBER OF TREES
708. C
709. WORKING ARGUMENTS (SHARED BY DINIT, DFIND & DUNION)
710. C
711. FATHER(.) VECTOR CONTAINING THE PREDECESSOR OF EACH ELEMENT
712. C
713. SIZE(.) VECTOR CONTAINING THE DIMENSIONALITY OF EACH SET
714. C
715. LOCAL VARIABLES
716. C
717. I CURRENT TREE
718. C
719. TYPE & DIMENSION STATEMENTS
INTEGER SIZE(1), FATHER(1)

DO 1000 I = 1, N
  FATHER(I) = 0
  SIZE(I) = 1
1000 CONTINUE
RETURN
END

SUBROUTINE UNION(I, J, K, SIZE, FATHER)

DUNION MERGES SET I AND SET J, FORMS SET K, AND DESTROYS I AND J

INPUT ARGUMENTS
I
  SET TO BE MERGED WITH J
J
  SET TO BE MERGED WITH I
K
  RESULTING SET

WORKING ARGUMENTS (SHARED BY DINIT, DPRINT, & DUNION)
FATHER(.) VECTOR CONTAINING THE PREDECESSORS OF EACH ELEMENT
SIZE(.) VECTOR CONTAINING THE SIZE OF EACH SET

TYPE AND DIMENSION STATEMENTS

INTEGER SIZE(1), FATHER(1), P, Q, S
P = I
Q = J
ISIZE = SIZE(I)
JSIZE = SIZE(J)
IF (ISIZE .LE. JSIZE) GO TO 10
S = P
P = Q
Q = S
10 CONTINUE
K = Q
SIZE(K) = SIZE(I) + SIZE(J)
FATHER(P) = Q
RETURN
END

SUBROUTINE HEAPIF(W, I, J, Q)

HEAPIF ESTABLISHES THE HEAP PROPERTY BETWEEN LOCATIONS I AND J
PROVIDED IT HOLDS ALREADY BETWEEN LOCATIONS I+1 AND J

INPUT ARGUMENTS
I
  INITIAL LOCATION
J
  FINAL LOCATION
W(.) WEIGHT VECTOR

I/O ARGUMENTS
Q(.) HEAP VECTOR

LOCAL VARIABLES
IDIS INDEX OF ODD POSITIONS OF Q
IPAR INDEX OF EVEN POSITIONS OF Q
C K = 2*IN OR 2*IN+1
C S USED TO EXCHANGE Q(K) WITH Q(IN)
C
C TYPE AND DIMENSION STATEMENTS
C
C INTEGER Q(1), S
C DIMENSION W(1)
C
C IN = I
10 CONTINUE
K = 2* IN
IF ( K .LT. J ) RETURN
IF ( K .EQ. J ) GO TO 20
IPAR = Q(K)
IDIS = Q(K+1)
IF ( W(IDIS) .GE. W(IPAR) ) GO TO 20
K = K + 1
20 CONTINUE
C
C HEAPIF RECURSION
C
C IN = K
GO TO 10
END
C
C SUBROUTINE HINIT(W,M,Q0,Q)
C
C SUBROUTINE TO INITIALIZE AND ORGANIZE AS HEAP STRUCTURE
C AN INDEX VECTOR Q OF A WEIGHT VECTOR W
C
C INPUT ARGUMENTS
C M INDEX NUMBER
C W WEIGHT VECTOR
C
C OUTPUT ARGUMENTS
C Q(.) HEAP VECTOR
C Q0 CARDINALITY OF Q
C
C LOCAL VARIABLES
C I LOOP INDEX OR ARGUMENT OF HEAPIF
C K LOOP INDEX FOR HEAPIF
C L M/2 TRUNCATED TO INTEGER
C
C CALLED SUBROUTINE
C
C HEAPIF, TO ORGANIZE A VECTOR AS HEAP STRUCTURE
C
C TYPE AND DIMENSION STATEMENTS
C
C INTEGER Q0,Q(1)
C DIMENSION W(1)
C

840. C
841. C INITIALISATION OF VECTOR Q
842. C
843. Q0 = M
844. Q(1) = 1
845. IF (M.LE.1) RETURN
846. DO 1000 I=2,M
847. Q(I)=I
848. 1000 CONTINUE
849. C
850. C ORGANIZATION OF VECTOR Q AS HEAP STRUCTURE
851. C
852. L = M / 2
853. I = L
854. DO 1010 K = 1,L
855. CALL HEAPIF (W,I,Q0,Q)
856. I = I - 1
857. 1010 CONTINUE
858. 200 CONTINUE
859. RETURN
860. END

C

861. INTEGER FUNCTION HHIN(W,Q0,Q)
862. C
863. C FUNCTION WHICH RETURNS THE ELEMENT OF SMALLEST WEIGHT
864. C AND REORGANIZES AS HEAP STRUCTURE VECTOR Q
865. C
866. INPUT ARGUMENTS
867. C
868. W(.) WEIGHT VECTOR
869. C
870. I/O ARGUMENTS
871. C
872. Q(.) INDEX VECTOR
873. C
874. Q0 CARDINALITY OF Q
875. C
876. CALLED SUBROUTINE
877. C
878. HEAPIF, TO REORGANIZE VECTOR Q AS HEAP STRUCTURE
879. C
880. TYPE AND DIMENSION STATEMENTS
881. INTEGER Q0,Q(1)
882. DIMENSION W(1)
883. IF (Q0.GT.0) GO TO 10
884. HMIN = 0
885. RETURN
886. 10 CONTINUE
887. HMIN = Q(1)
888. Q(1) = Q(00)
889. Q0 = Q0-1
890. IF (Q0.GE.2) CALL HEAPIF(W,1,Q0,Q)
891. RETURN
892. END

/*
894. 
895. //GO.FT01F001 DO DSN=TNW5O.SIM.TRACT.NEG,
896. DISP=SHR
897. //GO.FT02F001 DO DSN=TNW5O.SIM.BKGRP.POS,
898. DISP=SHR
899. //GO.FT03F001 DO DSN=TNW5O.SIM.BLOCX.NEG,
900.  // DISP=SHR
901.  //GO.FT11F001 DD DSN=TNEWSO.SIM.ADJ.TRACT,
902.  // DISP=SHR
903.  //GO.FT12F001 DD DSN=TNEWSO.SIM.ADJ.BKGRP,
904.  // DISP=SHR
905.  //GO.FT13F001 DD DSN=TNEWSO.SIM.ADJ.BLOCK,
906.  // DISP=SHR
907.  //GO.FT04F001 DD DSN=TNEWSO.HST.SIM6,UNIT=USER80,
908.  // DCB=(RECFM=FB, LRECL=80, BLKSIZE=6160),
909.  // DISP=(NEW,CATLG,DELETE),SPACE=(.TRK,(10,2),RLSE)
910.  // EXEC WNOTIFY
911.  /*
912.  //


