The Development of A Spatial Analysis Toolkit For Use in A Metropolitan Crime Incident Geographic Information System

by

Ned Levine, PhD
Ned Levine & Associates
Annandale, VA

Final Report

Grant Number 97-IJ-CX-0040
National Institute of Justice
Office of Justice Programs
Washington, DC

This report summarizes grant number 97-IJ-CX-0040, the development of a spatial analysis toolkit for use in a metropolitan crime incident geographic information system. This grant had the goal of developing a Windows-based spatial statistics program that could interface with crime mapping geographic information systems (GIS) and which could be linked with the crime mapping efforts of Baltimore County Police Department. The program, *CrimeStat*, was developed over a 21 month period, along with worked examples and an extensive user's manual/textbook.

Products of the Study

Attached to this final report are the products of the study. These include:

1. A copy of the final version of *CrimeStat* (version 1.0) on a CD-Rom disk;

2. A copy of the final version of the user's manual in printed form;

3. A digital copy of the final manual in Adobe 'pdf' format, also on the CD-Rom. There are nine separate 'pdf' files, one for each of the seven chapters, one for the references, and one for the appendix;

4. A copy of a quick instruction guide to the program (*CrimeStat* Quickguide) in printed form;

5. A digital copy of the *CrimeStat* Quickguide which is also on the CD-Rom; and

6. An example simulated data set from the Baltimore region which is also on the CD-Rom.

Description of the Project

The *CrimeStat* program was developed over a 21 month period, from November 1997 through August 1999. Mr. Long Doan of *Doan Associates*, Falls Church, VA, has been the key individual in the programming required to develop *CrimeStat*.

Input and Output

CrimeStat is a full-featured *Windows NT*® program using a graphical interface with database and expanded statistical functions. It can read files in *dBase*® (III or IV), which is a common file format in desktop GIS programs, as well as *ArcView* Shape (SHP) files directly (Borland.Com, 1998; ESRI, 1998a). In addition to printing tables, *CrimeStat* can write graphical objects to the *ArcView*®, *MapInfo*®, and *Atlas*GIS*™ GIS programs and can write interpolation files to these and to the *Surfer*® *for Windows* and *ArcView Spatial Analyst*© programs (Golden Software, 1994; ESRI, 1998a; 1998b; 1998c; 1997; MapInfo, 1998). The calculating algorithms, particularly for distances, are multi-threading, which means they can take advantage of multiple processors.

Statistical Routines

CrimeStat includes statistical routines for the mean center, standard distance deviation, standard deviational ellipse, center of minimum distance, Moran's I and Geary's C spatial autocorrelation indices, the angular mean and variance, Ripley's K statistic, a nearest neighbor clustering routine, a K-mean clustering routine, a local Moran test, and one and two variable kernel density interpolation routines with fixed and adaptive density estimation. Also, *CrimeStat* has Dynamic Data Exchange (DDE) capabilities so that it can be accessed from within another program. The user's manual describes the functionality of the program and the various routines.

Development Process

The program is written in Visual C++. It is made up of 176 separate file modules that are linked together during the compilation. Each of the files handles a specific function within the program (e.g., reading in a *dBase*® 'dbf' file; calculating a standard deviational ellipse; outputting a reference grid into *ArcView*® 'shp' format).

CrimeStat was developed with a series of specific compilations, what we call *builds*. With each new *build*, new functions were added to the program or existing functions were modified. Over the 21 month period, we produced over 110 separate builds. Aside from constructing each of the routines, extensive testing was done on each. Numerous modifications were made, both based on our experience and on the experience of Baltimore County Police Department, the GIS team at the criminal division of the U. S. Department of the Justice, and numerous 'beta' testers who helped evaluate the program.

The process that was followed was typically that I would design a statistical function. Then, Mr. Doan would do the programming and conduct preliminary tests of it. Then, I would conduct extensive tests of the routine and give feedback on how the routine needed to be modified. Mr. Doan would then fix any problems. After several iterations, the routine was then sent out to one of the program testers (see below). Upon receiving comments on individual routines, modifications were then implemented. This process was repeated until all routines were working properly and appeared to provide the required information.

Linking *CrimeStat* to Crime Mapping Efforts

During the development, *CrimeStat* was integrated into two crime mapping efforts. First, it has been used extensively by Baltimore County Police Department. Mr. Phil Canter of the Baltimore County Police Department has worked closely with us in developing the program and has developed applications using the program. Second, it has been integrated into the Regional Crime Analysis GIS (RCAGIS) that is being developed by the criminal division of the U.S. Department of Justice. We have worked closely with Mr. John DeVoe of USDOJ, head of the RCAGIS efforts, and Mr. Ron Wilson of Indus Corporation, who was the individual responsible for linking RCAGIS with *CrimeStat*.

Testing of the Program

During the development of the program, the above individuals provided extensive testing of the routines. This was a continual process and their efforts were essential for a successful completion of the program. In addition, upon completion of the first beta version of the program (March 1999), a number of other individuals acted as 'beta' testers. Again, as comments about the program were received, modifications to the final version was implemented. The author would like to acknowledge the efforts of Professor Karl Kim, Professor Luc Anselin, Professor Richard Block, Dr. Carolyn Block, Dr. Lee DeCola, Dr. Eric Jefferis, Professor Bob Langworthy, Professor Jim LeBeau, and Dr. Joseph Szakas Jefferis in providing these informal tests. Finally, two anonymous reviewers evaluated the draft program and manual for NIJ. Again, the author would like to thank those reviewers for their efforts. Also, the continual feedback by my program manager, Ms. Cynthia Mamalian, and Dr. Nancy LaVigne, head of the crime mapping research center at NIJ throughout the development process helped to improve the program.

Formal Presentations About the Program

Finally, during the development process, the author made numerous presentations at formal meetings about the program. Each of these presentations provided an opportunity to get feedback on the program, a process that was critical in articulating the design details. A brief listing of the presentations is as follows:

1. Formal presentation and poster session at the 4th International Conference on GeoComputation. U. S. Army Corps of Engineers. Fredericksburg, VA. July 1999.

2. Presentation to the Virginia State Crime Commission. Fairfax, VA. June 1999.

3. Presentation at the Towson University GIS Conference. Baltimore. June 1999.

4. Presentation at the U. S. Department of Transportation, Washington, DC. May 1999.

5. Presentation at the U. S. Geological Survey, Reston, VA. March 1999.

6. Presentation at the Transportation Research Board Annual Meeting. Washington. January 1999.

7. Three presentations at the National Institute of Justice Crime Mapping Research Conference, Arlington, VA. December 1998.

8. Two-day course and one-day workshop at the Justice Research Statistics Association annual meeting, San Diego. September 1998.

9. Presentation at Research and Evaluation Conference. National Institute of Justice, Washington, DC. July 1998.

10. Presentation at the Towson University Annual Geographic Information Systems Conference, Towson, MD, June 1998.

11. Presentation at the NIJ Cluster Conference on the Development of Spatial Analysis Tools. Washington, DC. February 1998.

12. Presentation at the Applied Geography Annual Meeting. Albuquerque, NM, November 1997.

In addition, I will make a formal presentation at the National Center for Health Statistics in Hyattsville in late September (simultaneously broadcast to the Centers for Disease Control in Atlanta). In short, there have been numerous presentations where *CrimeStat* was presented to an audience and in which feedback about the program has been received.

Distribution and Follow-Up

Upon completion of this project, NIJ will distribute the program, most probably through a web site. The author will follow up on this distribution and will seek to make additional modifications to the program as feedback from general law enforcement and

criminal justice users comes in. It is hoped that there will be a second grant whereupon improvements to the program can be made.

CrimeStat

(Version 1.0)

A Spatial Statistics Program for the Analysis of Crime Incident Locations

Ned Levine & Associates
Annandale, VA

The National Institute of Justice
Washington, DC

August, 1999

Table of Contents

Table of Contents i

Acknowledgments ix

License Agreement and Disclaimer xi

Part I: Program Overview

Chapter 1: Introduction to *CrimeStat* 1
- Uses of Spatial Statistics in Crime Analysis 1
  - Input and Output 2
  - Routines 2
- What the Program Does and Does Not Do 2
- Program Requirements 3
  - Required Hardware 3
  - Required Software 4
    - Windows 95 5
    - Windows 98 5
- Installing the Program 5
  - Adding an Item to the Start Menu 5
  - Adding an Icon to the Desktop 6
  - Installing the Sample Data Set 6
- Step-by-step Instructions 6
- On-line Help 7
- Chapter 1 Endnotes 8

Chapter 2: Quickguide to *CrimeStat* 9
- Primary File 9
  - Select Files 9
  - Variables 9
  - Column 11
  - Directional 11
  - Type of Coordinate System and Data Units 11
- Secondary File 11
  - Select Files 11
  - Variables 13
  - Column 13
  - Type of Coordinate System and Data Units 13
- Reference File 13
  - From File 13
  - Generated 13
- Measurement Parameters 16
  - Area 16

Table of Contents (continued)

Length of Street Network 16
Type of Distance Measurement 16
Direct 16
Indirect 16
Spatial Distribution 17
Mean Center and Standard Distance(Mcsd) 17
Standard Deviational Ellipse (Sde) 17
Median Center/Center of Minimum Distance (Mcmd) 19
Directional Mean and Variance (DMean) 19
Spatial Autocorrelation Indices 20
Moran's I (MoranI) 20
Adjust for small distances 20
Geary's C (GearyC) 20
Adjust for small distances 21
Distance Analysis 21
Nearest Neighbor Analysis (Nna) 21
Number of nearest neighbors 23
Linear Nearest Neighbor Analysis 23
Number of Linear Nearest Neighbors 23
Ripley's K (RipleyK) 24
Distance Matrices 24
Within File Point-to-Point (Matrix) 24
From All Primary File Points to All Secondary File Points (IMatrix) 25
'Hot Spot' Analysis 25
Nearest Neighbor Hierarchical Spatial Clustering (Nnh) 25
Significance level 25
Minimum points per cluster 27
Output sizes for ellipses 27
K-means Clustering (KMeans) 27
Local Moran Statistics (L-Moran) 27
Adjust for small distances 28
Interpolation 28
Single Kernel Density Estimate 28
File to be interpreted 28
Method of interpolation 28
Choice of bandwidth 28
Adaptive bandwidth 30
Fixed bandwidth 30
Output units 30
Use intensity variable 30
Use weighting variable 30
Calculate densities or probabilities 30
Output 30
Duel Kernel Density Estimate 31

Table of Contents (continued)

File to be interpreted 31
Method of interpolation 31
Choice of bandwidth 31
Adaptive bandwidth 31
Fixed bandwidth 31
Variable bandwidth 32
Output units 32
Use intensity variable 32
Use weighting variable 32
Calculate densities or probabilities 32
Output 32
Dynamic Data Exchange (DDE) Support 32

Part II: *CrimeStat* Instructions and Statistics

Chapter 3: Reading Data into *CrimeStat* **35**
Required Data 35
Coordinates 35
Intensities and weights 37
'Clean' Data 38
Primary File 38
Input File Formats 40
ArcView 40
MapInfo 40
Atlas*GIS 40
ASCII 40
Identifying Variables 41
Weight Variable 41
Intensity Variable 41
Coordinate System 43
Spherical coordinates 43
Projected coordinates 43
Directional coordinates 43
Secondary File 46
Reference File 46
Existing Grid File 46
Generating a Reference File 50
Measurement Parameters 52
Area and Length of Street Network 52
Direct and Indirect Distance 52
Distance Calculations 56
Direct, projected coordinate system 56
Direct, spherical coordinate system 56
Indirect, projected coordinate system 57

Table of Contents (continued)

Indirect, spherical coordinate system 57
Endnotes for Chapter 3 58

Chapter 4: Spatial Distribution 67
Centrographic Statistics 67
Mean Center 67
Weighted Mean Center 70
Center of Minimum Distance 73
Standard Deviation of the X and Y Coordinates 79
Standard Distance Deviation 79
Standard Deviational Ellipse 82
Table Outputs 84
Selecting Output Objects 87
Calculating the Statistics 87
Output Files 87
Statistical Tests 90
Differences in the Mean Centers of Two Samples 90
Significance levels 90
Tests 91
Example 1: Burglaries and robberies in Baltimore County 93
Differences in the Standard Distance Deviations of Two Samples 97
Differences in the Standard Deviational Ellipses of Two Samples 97
Differences in the mean centers 98
Differences in the angle of rotation 98
Differences in the standard deviations along the transformed axes 98
Differences in the areas of the two ellipses 98
Significance levels 99
Decision-making Without Formal Tests 99
Example 2: June and July auto thefts in precinct 11 99
Example 3: Serial burglaries in Baltimore City and Baltimore County 101
Example 4: Auto thefts over time in Baltimore County 108
Directional Mean and Variance 108
CrimeStat Input and Output for Directional Mean and Variance 111
Example 5: Directional mean and variance 114
Statistical Test of Differences in Mean Direction Between Two Groups 114
Example 6: Angular comparisons between two groups 118
Spatial Autocorrelation 120
Indices of Spatial Autocorrelation 121
Moran's I 121
Adjustment for small distances 122
Testing the significance of Moran's I 123

Table of Contents (continued)

Example 7: Testing auto thefts with Moran's I 123
Geary's C Statistic 128
Adjustment for small distances 128
Testing the significance of Geary's C 130
Example 8: Testing auto thefts with Geary's C 130
Endnotes for Chapter 4 132

Chapter 5: Distance Analysis 137
Nearest Neighbor Index 137
Testing the Significance of the Nearest Neighbor Index 139
Calculating the statistics 140
Example 1: The nearest neighbor index for street robberies 140
Nearest Neighbor Analysis is Not a Test for Complete Spatial Randomness 141
Edge Effects 141
Example 2: The nearest neighbor index for residential burglaries 142
K-Order Nearest Neighbors 143
Linear Nearest Neighbor Index 145
Testing the Significance of the Linear Nearest Neighbor Index 145
Calculating the statistics 146
Example 3: Auto thefts along two highways 147
K-Order Linear Nearest Neighbors 150
Ripley's K Statistic 152
Potential Bias in the Statistic 153
Comparison to a Spatially Random Distribution 155
Specifying simulations 156
Comparison to Baseline Populations 156
Distance Matrices 157
Endnotes for Chapter 5 161

Chapter 6: 'Hot Spot' Analysis 163
Statistical Approaches to the Measurement of 'Hot Spots' 163
Types of Cluster Analysis ('Hot spot') Methods 164
Optimization Criteria 167
Cluster Routines in *CrimeStat* 168
Nearest Neighbor Hierarchical Clustering 168
Nearest Neighbor Criteria 169
First-order clustering 170
Second and higher-order clusters 171
Guidelines for Selecting Parameters 171
Nnh Output Files 172
Example 1: Nearest neighbor hierarchical clustering of burglaries 173

Table of Contents (continued)

Advantages of Hierarchical Clustering 173
Limitations of Hierarchical Clustering 177
K-Means Partitioning Clustering 178
CrimeStat K-means Routine 179
K-means Output Files 179
Example 2: K-means clustering of street robberies 180
Advantages and Disadvantages of the K-means Procedure 180
Local Moran Statistics 184
Formal Definition of Local Moran Statistic 185
The I_i statistic 185
Distance weights 186
Small distance adjustment 186
Similarity or dissimilarity 186
Example 3: Local Moran statistics for auto theft 187
Some Thoughts on the Concept of 'Hot Spots' 190
Advantages 190
Disadvantages 190
Endnotes for Chapter 6 193

Chapter 7: Kernel Density Interpolation 199
Kernel Density Estimation 199
CrimeStat Kernel Density Methods 206
Single Density Estimates 209
File to be Interpolated 209
Method of Interpolation 209
Choice of Bandwidth 209
Fixed interval 211
Adaptive interval 211
Output Units 211
Intensity or Weighting Variables 211
Calculations 212
Output Files 212
Example 1: Kernel density estimate of street robberies 212
Dual Density Estimates 217
File to be Interpolated 217
Method of Interpolation 217
Choice of Bandwidth 217
Fixed interval 217
Variable interval 217
Adaptive interval 218
Output Units 218
Intensity or Weighting Variables 218
Calculations 218

Table of Contents (continued)

Output Files 219
Example 2: Kernel density estimates of auto thefts relative to population 220
Conclusion 224
Endnotes for Chapter 7 226

References **227**

Appendix A: Dynamic Data Exchange Support **235**

Acknowledgments

CrimeStat was developed under the direction of Dr. Ned Levine of *Ned Levine & Associates*, Annandale, VA, with a grant from the *National Institute of Justice* (NIJ). The developer wishes to give special thanks to Mr. Long Doan of *Doan Consulting*, Falls Church, VA, the chief programmer for the project, and to Mr. Phil Canter of the *Baltimore County Police Department*, Towson, MD, who provided support and data for analysis. Both these individuals were completely essential to the development of the program, Mr. Doan through his talented programming and Mr. Canter through his efforts at building a regional crime analysis system. Acknowledgments are also given to Ms. Cynthia Mamalian, project director at NIJ, Dr. Nancy LaVigne, Director of the Crime Mapping Research Center at NIJ, and Mr. John DeVoe of the Criminal Division, U. S. Department of Justice, who has integrated *CrimeStat* into their Regional Crime Analysis Geographic Information System. Finally, Ms. Sandra Wortham of *Wortham Design*, Wilmington, DE, designed the graphical icons used in the program.

Many others helped in the development of this program: Professor Karl Kim of the Department of Urban & Regional Planning at the University of Hawaii in Honolulu, HI, provided support for an early prototype. The early programming was conducted by Mr. Chai Khoo and Ms. Donna Okazaki of the University of Hawaii. Some refinements were made by Mr. Kwee Phua and Mr. Steven Wojnarowski of Oz Info Pty Ltd. of Melbourne, Australia. Professor David Wong of the Department of Geography & Earth System Sciences at George Mason University, Fairfax, VA, provided statistical advice.

Finally, I want to thank my wife, Dr. C. Elizabeth Castro, for being so supportive and analytically critical throughout this process. Without her patience, I would never have undertaken such an ordeal.

License Agreement and Disclaimer

This project was supported by Grant No. 97-IJ-CX-0040 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. Points of view in this document are those of the author and do not necessarily represent the official position or policies of the US Department of Justice.

Any questions about the use of this program should be directed to either:

Dr. Ned Levine
Ned Levine & Associates
Annandale, VA
ned@nedlevine.com

Dr. Nancy La Vigne
Crime Mapping Research Center
National Institute of Justice
U. S. Department of Justice
810 7th St, NW
Washington, DC 20531
cmrc@ojp.usdoj.gov

Part I: Program Overview

Chapter 1
Introduction to *CrimeStat*

CrimeStat is a spatial statistics package which can analyze crime incident location data. Its purpose is to provide a variety of tools for the spatial analysis of crime incidents or other point locations. It is a stand-alone *Windows NT®* program that can interface with most desktop geographic information systems (GIS). It is designed to operate with large crime incident data sets collected by metropolitan police departments. However, it can be used for other types of applications involving point locations, such as the location of arrests, motor vehicle crashes, emergency medical service pickups, or facilities (e.g., police stations).

Uses of Spatial Statistics in Crime Analysis

Most GIS packages, such as *MapInfo®*, *ArcView®*, *ARC/INFO®*, *Atlas*GIS™*, and *Maptitude®*, have very sophisticated data base operations. They do not, however, have statistical methods other than means and standard deviations of variables. For most purposes, GIS can provide great utility for crime analysis, allowing the plotting of different incident locations and the ability to select subsets of the data (e.g., incidents by precinct, incidents by time of day). Most crime analysts visually inspect incident maps and, based on their experience, draw conclusions about shifts over time, 'hot spots' and other patterns suggested by the data.

There are times, however, when a more quantitative approach is needed. For example, an analyst wishing to examine patterns of streets robberies over time will need indices which document how the robberies may have shifted. For a neighborhood showing an apparent sudden increase in auto thefts, there needs to be a quantitative standard to define the 'typical' level of auto thefts. In assigning police cars to patrol particular major arteries, the center of minimum travel needs to be identified in order to maximize response time to calls for service. For research, as well, quantification is important. In examining correlates of burglaries, for example, a researcher needs to determine the exposure level, namely how many residences or commercial buildings exist in a community in order to establish a level of burglary risk. Or a precinct may want to target areas for which there is a high concentration of incidents occurring within a short time ('hot spots'). While some of these analyses can be conducted with GIS queries, quantification can allow a more precise identification and the ability to compare different types of incidents. In short, there are many uses for quantitative analysis for which a statistical program becomes important.

CrimeStat is a tool designed to provide statistical summaries and models of crime incident data. The tool kit provides crime analysts and researchers with a wide range of spatial statistical procedures that can be linked to a GIS. The procedures vary from the simple to some very sophisticated 'cutting edge' routines. The reasoning is that different audiences vary in their needs and requirements. The program should be of benefit to different organizations. For many crime analysts, simple descriptions of the spatial distribution will be sufficient with the aim being practical intervention over a short time

period. For these persons, many of the techniques provided in *CrimeStat* will be unnecessary.

For other analysts, statistical tools can supplement a much larger GIS effort, such as the Regional Crime Analysis System (RCAGIS) being developed by the U.S. Department of Justice in cooperation with a number of police departments in the Baltimore-Washington metropolitan area. For other researchers, even more demanding techniques may be needed to detect the underlying spatial structure as a means for formulating a temporal-spatial theory. A pattern in and of itself has little meaning unless it is linked to some framework. The ability to quantify relationships with a large amount of data can address problems that previously were avoided and can be a first step in developing an explanatory framework or interventionist strategy. *CrimeStat* attempts to address both types of needs by providing statistics in a 'toolbox' framework. We recognize that today's exotic statistical techniques may become tomorrow's practical diagnostics and want the program to be useful for many years.

Input and Output

CrimeStat is a full-featured *Windows NT®* program using a graphical interface with database and expanded statistical functions. It can read files in *dBase®* (III or IV), which is a common file format in desktop GIS programs, as well as *ArcView* Shape (SHP) files directly (Borland.Com, 1998; ESRI, 1998a). In addition to printing tables, *CrimeStat* can write graphical objects to the *ArcView®*, *MapInfo®*, and *Atlas*GIS™* GIS programs and can write interpolation files to these and to the *Surfer® for Windows* and *ArcView Spatial Analyst©* programs (Golden Software, 1994; ESRI, 1998a; 1998b; 1998c; 1997; MapInfo, 1998). The calculating algorithms, particularly for distances, are multi-threading, which means they can take advantage of multiple processors.

Routines

CrimeStat includes routines for the mean center, standard distance deviation, standard deviational ellipse, center of minimum distance, Moran's I and Geary's C spatial autocorrelation indices, the angular mean and variance, Ripley's K statistic, a hierarchical nearest neighbor clustering routine, a K-mean clustering routine, a local Moran test, and one and two variable kernel density interpolation routines with fixed and adaptive density estimation. Also, *CrimeStat* has Dynamic Data Exchange (DDE) capabilities so that it can be accessed from within another program.

What the Program Does and Does Not Do

CrimeStat provides descriptions of the spatial arrangements of crime incidents. There are a variety of tools that can be used to describe these arrangements from the analysis of central tendency and one- and two-dimensional dispersion to the analysis of the distances between incident locations to identification of collections of incidents which cluster together ('hot spots') to three-dimensional models of crime density. These tools are useful in helping crime analysts detect patterns of crime and provide different perspectives

on the arrangements. In this sense, it is a tool for analyzing one or, at most, two variables affecting crime incidence - the incidents themselves and a secondary variable that can be used for comparison.

On the other hand, *CrimeStat* is not a standard statistical package aimed at modeling correlates or determinants of crime incidents. It does not have a regression module nor other multivariate techniques quantifying the predictors of crime locations. Also, it only works with point data and not with characteristics of zones or line segments. Users who want to model determinants of crime can use specialized regression packages, such as *SpaceStat*© (Anselin, 1992) or *S-Plus*® (MathSoft, Inc., 1998).

CrimeStat is a program that specializes in the analysis of point locations. Over the years, many statistical tools have been developed for analyzing point locations. Many of these have either not been implemented as computer programs or were collected together as part of a specialized statistical system. They have been typically unavailable to crime analysts and the major statistical packages (e.g., *SAS*®, *SPSS*™, *Systat*®) do not include these routines. Consequently, we have collected those that are most appropriate for crime analysis and detection and organized them into a single package with a common graphical interface. There are statistics that are not included that have been used for crime analysis (e.g., 'hot spot' analysis with the *Spatial and Temporal Analysis of Crime*© *(STAC)* program or two dimensional spectral analysis). In a later version, *CrimeStat* will be broadened to include such routines. Nevertheless, those that are currently in *CrimeStat* represent a wide variety of tools that can be used for crime analysis.

Program Requirements

Required Hardware

CrimeStat runs on a *Windows NT* system; it is not hardware dependent so that any processor that can run *Windows NT* will suffice. While it can run on a relatively slow computer (e.g., 75 MHZ clock speed) with limited RAM (e.g., 8 MB), it will run much better on a 200 MHZ Pentium II computer (or faster) with more than 16 MB of RAM. The faster the processor used, the quicker the program will run. The more RAM the computer has, the quicker the program will run. The program is very intensive with respect to calculations. Some of the statistics produce large matrices (e.g., the distance from every point to every other point). Depending on the size of the data files that will be processed, there may be hundreds of millions of calculations on any one run. It is critical, therefore, that the computer be fast and have sufficient amounts of RAM. The program was designed on an *NT* system with 64 MB of RAM, but was tested on an *NT* system with 256 MB of RAM.

In addition, *CrimeStat* is designed to be multi-threading which means that it will take advantage of multiple processors in a *Windows NT* environment. *Windows NT 4.0* (to be known as Windows 2000 in the next version) supports multiple processors (Microsoft, 1998a). *Windows NT Workstation 4.0* supports two processors. *Windows NT Server 4.0* (Microsoft, 1998b) supports four processors while *Windows NT Server, Enterprise Edition*

(Microsoft, 1999) supports up to 32 processors. However, neither *Windows 95* (Microsoft, 1995) nor *Windows 98* (Microsoft, 1998c) will recognize multiple processors. Thus, if there are two processors and *Windows NT* is the operating system, *CrimeStat* will calculate routines in about half the time. If there are four processors and *Windows NT Server* is the operating system, *CrimeStat* will calculate routines in about a quarter of the time. The multiples are not exact since processing time must be allocated for input of data and output of tables.

For small data sets, this feature is not important as most runs will be very quick. However, for large data sets (e.g., 3000 cases or larger), the speed of calculations become important. For example, on a 266 MHZ single-processor *Pentium II* computer with 256 MB of RAM running *Windows NT*, it takes about 40 minutes to complete a nearest neighbor analysis on 19,208 cases involving the calculating of distance from every point to every other point multiple times (for different neighbors). On a dual-processor *Pentium II* computer with 256 MB of RAM running *Windows NT*, it takes about 20 minutes to complete the same task. On a single processor 133 MHZ Pentium computer with 48 MB of RAM running *Windows 95*, it takes about an hour and a half to finish this run. The larger the file that is being processed, the more critical becomes the calculating efficiency of the computer.

If a police department is expecting to run large data sets, it would benefit them to purchase fast multiple-processor computers with lots of RAM and fast hard disks to speed calculating times. The evolution of new processors is moving in this direction anyway so that a multi-processor computer will become the norm in the next couple of years.

Required Software

CrimeStat needs a Windows environment to operate. The program was designed for a *Windows NT* operating system so it is better optimized for that system. In particular, *Windows NT* has two features that allows *CrimeStat* to run more efficiently. First, it is a multi-threading operating system and can utilize multiple processors, as mentioned above. Neither *Windows 95* nor *Windows 98* can utilize multiple processors. Second, it addresses memory in a more efficient way, as a large flat block. *Windows 95* cannot handle cache memory above 64 MB. *Windows 98* can handle RAM above 64 MB, but still has poorer memory management than *NT*. Consequently, for the same machine, *CrimeStat* will run more efficiently (i.e., more quickly) in *NT* than in *98* which, in turn, will run more efficiently than *95*.

CrimeStat is a stand-alone program. Hence, it does not require any other program other than a Windows operating system. However, to be maximally useful, there should be an accompanying GIS program. While point data can be obtained from a non-GIS system (e.g., census files include lat/lon coordinates for the centroid of census units), the use of the GIS to assign the coordinates is almost necessary. Further, many of the outputs of *CrimeStat* are for GIS programs. Thus, to view an ellipse or to view a three dimensional interpolation produced by *CrimeStat* will require an appropriate GIS package.

Windows 95

While *CrimeStat* was designed for a *Windows NT* (Windows 2000) operating system, it works properly under *Windows 95.* We have run extensive tests on Windows 95 computers and have found no problems other than the routines run slightly more slowly than on a *Windows NT* system.

Windows 98

We have found problems, however, in running *CrimeStat* under *Windows 98.* While the program will run, calculating multiple routines on a single run will often crash the program. *If* it is necessary to use *Windows 98* to run *CrimeStat,* we suggest that the user run only one routine at a time.

Installing the Program

CrimeStat comes compressed in a self-installing file called *CrimeStatExtract.Exe.* To install the program:

1. Create a directory using *Windows Explorer* and copy the file to that directory.

2. Double click on the file name in *Explorer.* When the name *CrimeStatExtract* is visible in the dialog box name field, double click the name with the left mouse button. *CrimeStat* will be installed in that directory.

3. Alternatively, click on the *Start* button in *Windows* followed by *Run.* In the dialog box, click on *Browse,* point to the directory where *CrimeStatExtract* resides and click on its name followed by *Open.*

Adding an Item to the Start Menu

To add *CrimeStat* to the start menu:

1. Click on the *Start* button in Windows followed by *Settings* then *Taskbar.* Click on *Start Menu Programs* followed by *Add.*

2. In the dialog box, click on *Browse,* point to the directory where *CrimeStat* resides, and click on its name followed by *Open.* When the name *CrimeStat* is in the dialog box name field, click on the *Next* button.

3. Double-click on the folder to which *CrimeStat* is to be assigned.

4. Finally, type a name for *CrimeStat* (e.g., CrimeStat) followed by *Finish.*

Adding an Icon to the Desktop

To add *CrimeStat* to the desktop:

1. Double-click on *My Computer.*

2. Double-click on the drive in which *CrimeStat* resides followed by the directory that it is in (it may be several levels down).

3. Click once on the name *CrimeStat* with the left button and then hold down the right mouse button.

4. While holding the right mouse button, scroll to *Create Shortcut.*

5. The name *Shortcut to CrimeStat* will be placed at the end of the list of files.

6. Highlight the name by clicking on it once. Hold the left mouse button down and drag this name on to the desktop.

7. You can rename it *CrimeStat* by clicking on its icon with the right mouse button followed by *Rename.*

8. Alternatively, you can use *Windows Explorer* to create a shortcut and then drag the shortcut to the desktop.

Installing the Sample Data Set

There is a sample data set that can be used to run the program. It is also a self-extracting file (ExtractSampleData.exe). The data are simulated incident points from Baltimore City and Baltimore County in Maryland.[1] They are provided to allow a user to become familiar with the program quickly. However, ultimately, the value of the program must be tested on real data, rather than simulated data. To extract the data:

1. In *Windows Explorer*, double-click on its name and then follow the instructions.

2. Save it in the same directory in which *CrimeStat* is installed.

Step-by-Step Instructions

This manual will go through the program step-by-step to address how it can be used by a crime mapping/analysis unit within a police department. Chapter 2 provides a quick guide for all the data definition and program routines. In Part II, detailed instructions on the program are explained, including data input, data definition, and the statistical routines. The different statistics are presented and detailed examples of each technique are shown.

On-line Help

In addition, there is on-line help for the program. There is a *Help* button that can be pushed to access all the help items. In addition, the program has context-sensitive help. On any page or routine, typing *F1* will pop up an appropriate help item.

Chapter 1 Endnotes

1. The data were simulated by a random number generator following the distribution of several types of crime incidents. Because the data were selected by a random generator, the points do not necessarily fall on streets or even stay within the boundaries of Baltimore City and Baltimore County; some even fall into the Chesapeake Bay! Their purpose is to provide a simple data set so users can become familiar with the program.

Chapter 2
Quickguide to *CrimeStat*

The following are quick instructions for the use of *CrimeStat*, paralleling the online help menus in the program. Detailed instructions should be obtained from Chapters 3-7. *CrimeStat* has eight program tabs. Each tab lists routines, options and parameters. The eight tabs are:

1. Primary file
2. Secondary file
3. Reference file
4. Measurement parameters
5. Spatial distribution
6. Distance analysis
7. 'Hot Spot' analysis
8. Interpolation

Figure 2.1-2.8 show the eight tab screens with examples of data input and routine selection.

Primary File

A primary file is required for *CrimeStat*. It is a point file with X and Y coordinates. For example, the primary file could be the location of street robberies, each of which has an associated X and Y coordinate. Alternatively, the primary file could be the location of police stations, again defined by an X and Y coordinate. Also, there can be weights or intensities variables associated, although these are optional. For example, if the points are the locations of police stations, then the intensity variable could be the number of calls for service at each police station while the weighting variable could be service zones. More than one file can be selected.

Select Files

Select the primary file. *CrimeStat* can read ASCII, dBase III/IV '.dbf', and ArcView '.shp' files. Select the tab and indicate the type of file to be selected. Use the browse button to search for the file. If the file type is ASCII, select the type of data separator (comma, semicolon, space, tab) and the number of columns.

Variables

Define the file which contains the X and Y coordinates. If there are weights or intensities being used, define the file which contain these variables. Certain statistics (e.g., spatial autocorrelation, local Moran) require intensity values and most other statistics can use intensity values. Other statistics (e.g., a weighted mean center) can use weights. It is

Figure 2.1: Primary File Screen

possible to have a variable represent both intensity and a weighting; it is also possible to have separate variables for intensity and for weighting.

Column

Select the variables for the X and Y coordinates respectively (e.g., Lon, Lat, Xcoord, Ycoord). If weights or intensities are being used, select the appropriate variable names.

Directional

If the file contains directional coordinates (angles), define the file name and variable name (column) that contains the directional measurements. If directional coordinates are used, there can be an optional distance variable for the measurement. Define the file name and variable name (column) that contains the distance variable.

Type of Coordinate System and Data Units

Select the type of coordinate system. If the coordinates are in longitudes and latitudes, then a spherical system is being used and data units will automatically be decimal degrees. If the coordinate system is projected (e.g., State Plane, Universal Transverse Mercator – UTM), then data units could be either in feet (e.g., State Plane) or meters (e.g., UTM). If the coordinate system is directional, then the coordinates are angles and the data units box will be blanked out; if a distance variable is used with the directional coordinates, the data units are undefined.

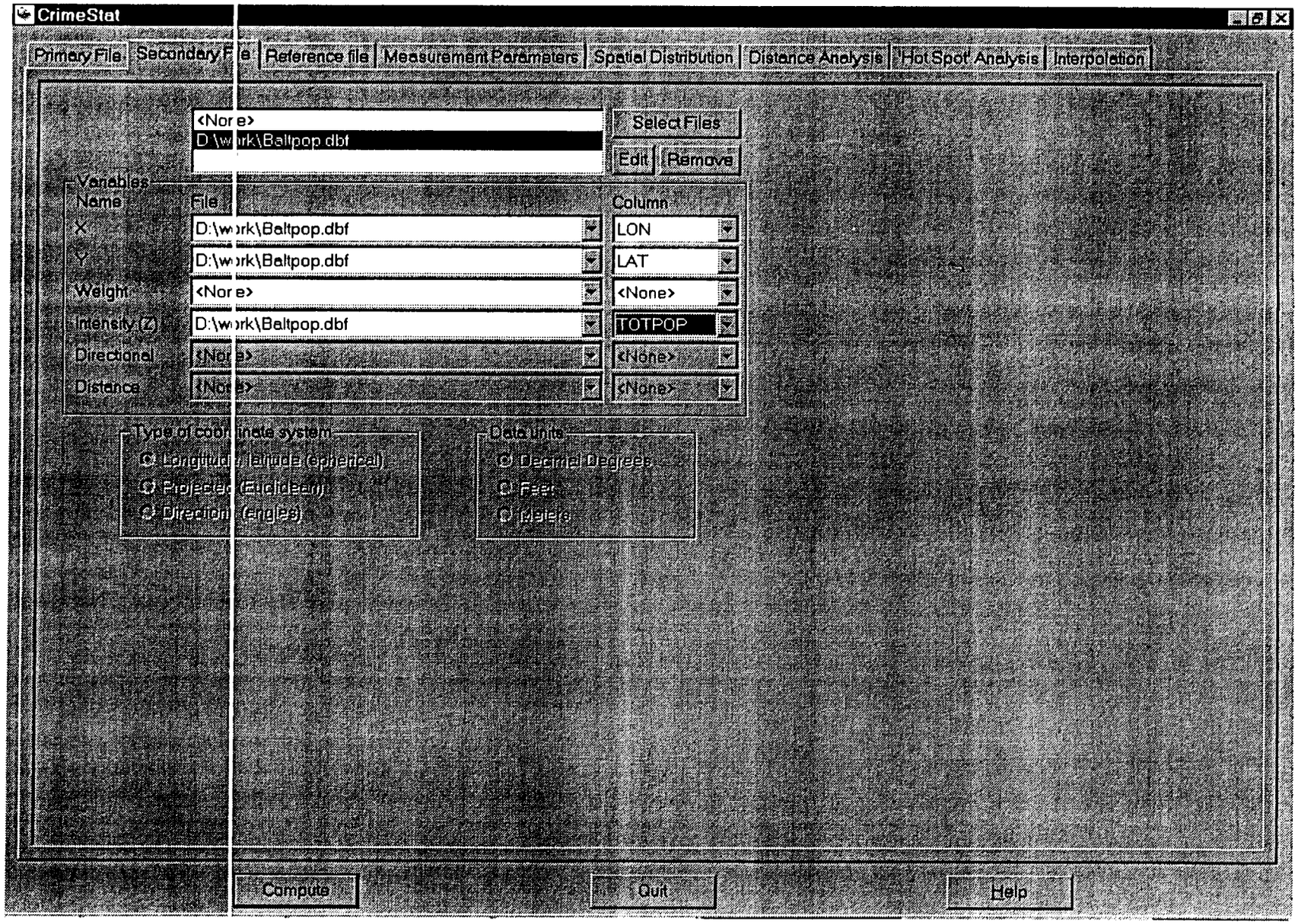
Secondary File

A secondary data file is optional. It is also a point file with X and Y coordinate and is usually used in comparison with the primary file. There can be weights or intensities variables associated, though these are optional. For example, if the primary file is the location of motor vehicle thefts, the secondary file could be the centroid of census block groups that have the population of the block group as the intensity (or weight) variable. In this case, one could compare the distribution of motor vehicle thefts with the distribution of population in, for example, the Ripley's K routine or the dual kernel density estimation routine.

Select Files

Select the secondary file. *CrimeStat* can read ASCII, dBase III/IV '.dbf', and ArcView '.shp' files. Select the tab and indicate the type of file to be selected. Use the browse button to search for the file. If the file type is ASCII, select the type of data separator (comma, semicolon, space, tab) and the number of columns.

Figure 2.2: Secondary File Screen



Variables

Define the file which contains the X and Y coordinates. If weights or intensities are being used, define the file which contain these variables. Certain statistics (e.g., spatial autocorrelation, local Moran) require intensity values and most other statistics can use intensity values. Other statistics (e.g., a weighted mean center) can use weights. It is possible to have a variable represent both an intensity and a weighting; it is also possible to have separate variables for intensity and for weighting.

Column

Select the variables for the X and Y coordinates respectively (e.g., Lon, Lat, Xcoord, Ycoord). If there are weights or intensities being used, select the appropriate variable names.

Type of Coordinate System and Data Units

The secondary file has the same coordinate system and data units as the primary file. This selection will be blanked out, indicating that the secondary file carries the same definition as the primary file.

Reference File

A reference file is used for single and dual variable kernel density estimation. The file can be an external file that is input or can be generated by *CrimeStat*. It is usually, though not always, a grid which is overlaid on the study area.

From File

Select the reference file. *CrimeStat* can read ASCII, dBase III/IV '.dbf', and ArcView '.shp' files. Select the tab and indicate the type of file to be selected. Use the browse button to search for the file. If the file type is ASCII, select the type of data separator (comma, semicolon, space, tab) and the number of columns. A file that is read into *CrimeStat* need not be a true grid (a matrix with k columns and l rows). However, if a file is read in, then the results can only be output to *Surfer for Windows*. The Grid Cells indicator will estimate the number of cells in the file as if it was a grid; this is an estimate since the imported file may not be a true grid.

Generated

CrimeStat can generate a true grid based on the inputting of the X and Y coordinates of a rectangle placed over the study area. The lower left and the upper right coordinates must be defined in the same coordinate system and data units as the primary file. Cells can be defined either by cell size in the same coordinate system and data units as the primary file or by the number of columns in the grid. If the latter is selected, *CrimeStat* will determine the number of rows to be generated based on the cell spacing.

Figure 2.3: Reference File Screen

Figure 2.4: Measurement Parameters Screen

Measurement Parameters

The measurement parameters define the measurement units of the coverage and the type of distance measurement to be used.

Area

Define the geographical area of the study area in area units (square miles, square nautical miles, square feet, square kilometers, square meters). Irrespective of the data units that are defined for the primary file, *CrimeStat* can convert to various area measurement units. These units are used in the nearest neighbor, Ripley's K, nearest neighbor hierarchical clustering, and K-means clustering routines. If no area units are defined, then *CrimeStat* will define a rectangle by the minimum and maximum X and Y coordinates.

Length of Street Network

Define the total length of the street network within the study area or an appropriate comparison network (e.g., freeway system) in distance units (miles, nautical miles, feet, kilometers, meters). The length of the street network is used in the linear nearest neighbor routine. Irrespective of the data units that are defined for the primary file, *CrimeStat* can convert to distance measurement units. The distance units should be in the same metric as the area units (e.g., miles and square miles/meters and square meters).

Type of Distance Measurement

Select the type of distance measurement to be used, direct or indirect.

Direct

If direct distances are used, each distance is calculated as the shortest distance between two points. If the coordinates are spherical (i.e., latitude, longitude), then the shortest direct distance is a 'Great Circle' arc on a sphere. If the coordinates are projected, then the shortest direct distance is a straight line on a Euclidean plane.

Indirect

If indirect distances are used, each distance is calculated as the shortest distance between two points on a grid, that is with distance being constrained to the horizontal or vertical directions (i.e., not diagonal). This is sometimes called 'Manhattan' metric. If the coordinates are spherical (i.e., latitude, longitude), then the shortest indirect distance is a modified right angle on a spherical right triangle; see the documentation for more details. If the coordinates are projected, then the shortest indirect distance is the right angle of a right triangle on a two-dimensional plane

Spatial Distribution

Spatial distribution provides statistics that describe the overall spatial distribution. These are sometimes called centrographic, global, or first-order spatial statistics. There are three routines for describing the spatial distribution and two routines for describing spatial autocorrelation. An intensity variable and a weighting variable can be used for the first three routines. An intensity variable is required for the two spatial autocorrelation routines; a weighting variable can also be used for the spatial autocorrelation indices. All outputs can be saved as text files. Some outputs can be saved as graphical objects for import into desktop GIS programs.

Mean Center and Standard Distance (Mcsd)

The mean center and standard distance define the arithmetic mean location and the degree of dispersion of the distribution. The Mcsd routine calculates 11 statistics:

1. The sample size
2. The minimum X value
3. The minimum Y value
4. The maximum X value
5. The maximum Y value
6. The mean of the X coordinates
7. The mean of the Y coordinates
8. The standard deviation of the X coordinates
9. The standard deviation of the Y coordinates
10. The standard distance deviation, in meters, feet and miles. This is the standard deviation of the distance of each point from the mean center.
11. The circle area defined by the standard distance deviation, in square meters, square feet and square miles.

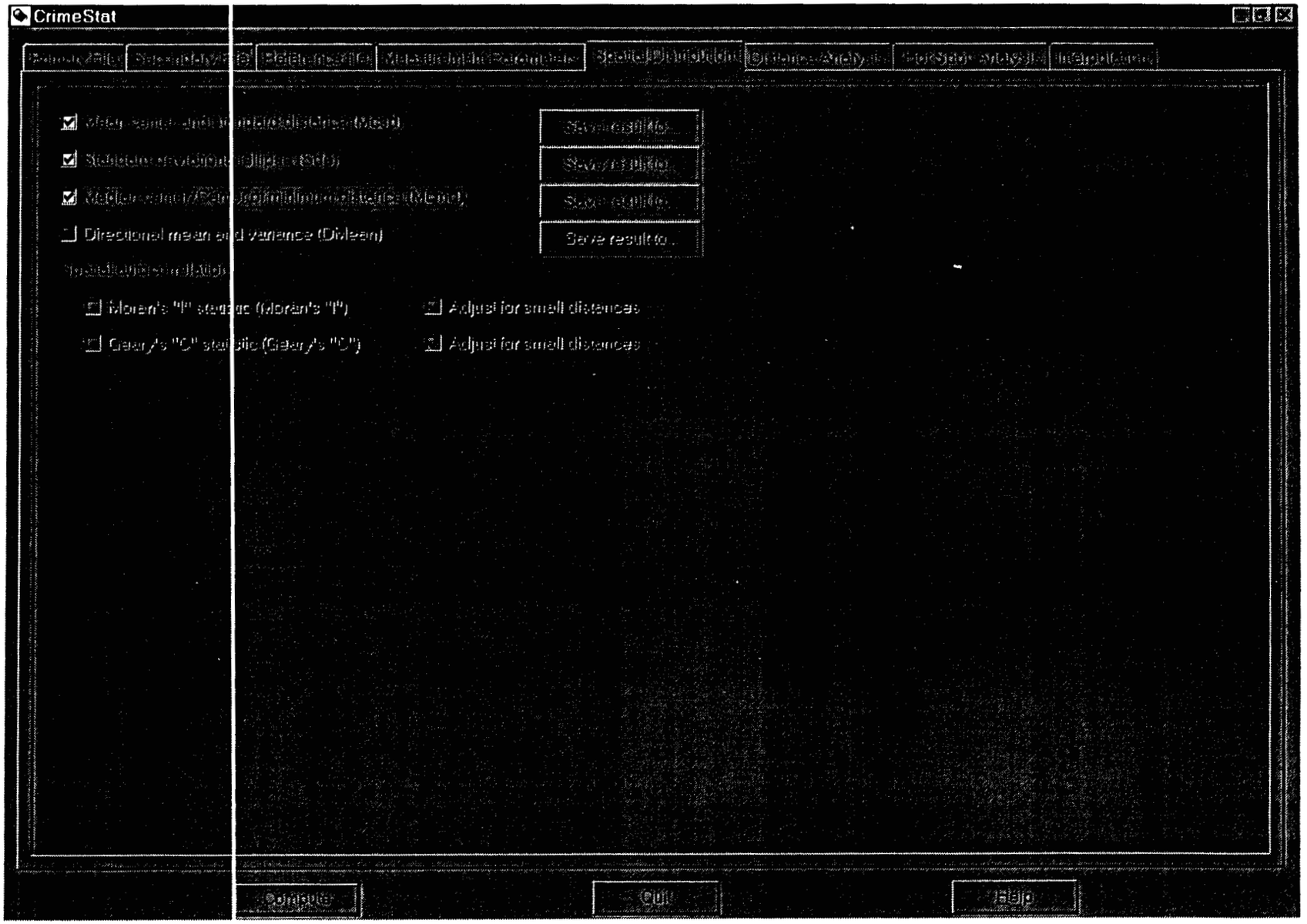
The tabular output can be printed and the mean center (mean X, mean Y), the standard deviations of the X and Y coordinates, and the standard distance deviation can be output as graphical objects to ArcView '.shp', MapInfo '.mif' and Atlas*GIS '.bna' formats. A root name should be provided. The mean center is output as a point (MC<root name>). The standard deviations of the X and Y coordinates are output as a rectangle (XYD<root name>). The standard distance deviation is output as a circle (SDD<root name>).

Standard Deviational Ellipse (Sde)

The standard deviational ellipse defines both the dispersion and the direction (orientation) of that dispersion. The Sde routine calculates 13 statistics:

1. The sample size
2. The clockwise angle of Y-axis rotation in degrees
3. The ratio of the long to the short axis after rotation
4. The standard deviation along the new Y axis in meters, feet and miles

Figure 2.5: Spatial Distribution Screen



5. The standard deviation along the new X axis in meters, feet and miles
6. The Y axis length in meters, feet and miles
7. The X axis length in meters, feet and miles
8. The area of the ellipse defined by these axes in square meters, square feet and square miles
9. The standard deviation along the Y axis in meters, feet and miles for a 2X standard deviational ellipse
10. The standard deviation along the X axis in meters, feet and miles for a 2X standard deviational ellipse
11. The Y axis length in meters, feet and miles for a 2X standard deviational ellipse
12. The X axis length in meters, feet and miles for a 2X standard deviational ellipse
13. The area of the 2X ellipse defined by these axes in square meters, square feet and square miles.

The tabular output can be printed and the 1X and 2X standard deviational ellipses can be output as graphical objects to ArcView '.shp', MapInfo '.mif' and Atlas*GIS '.bna' formats. A root name should be provided. The 1X standard deviational ellipse is output as an ellipse (SDE<root name>). The 2X standard deviational ellipse is output as an ellipse with axes that are twice as large as the 1X standard deviational ellipse (2SDE<root name>).

Median Center/Center of Minimum Distance (Mcmd)

The center of minimum distance (or median center) defines the point at which the distance to all other points is at a minimum. The Mcmd routine outputs 7 statistics:

1. The sample size
2. The mean of the Y coordinates
3. The mean of the X coordinates
4. The number of iterations required to identify a median center
5. The degree of error (tolerance) for stopping the iterations
6. The Y coordinate which defines the center of minimum distance (median center)
7. The X coordinate which defines the center of minimum distance (median center).

The tabular output can be printed and the median center can be output as a graphical object to ArcView '.shp', MapInfo '.mif' or Atlas*GIS '.bna' files. A root name should be provided. The median center is output as a point (MDN<root name>).

Directional Mean and Variance (DMean)

The directional mean and variance are calculated if the input variable is a collection of angular measures ($0^0 - 360^0$) and the coordinate system is defined as directions on the

primary file screens. The directional mean is an angle while the directional variance is a relative indicator varying from 0 (no variance) to 1 (maximal variance). The tabular output can be printed. If an additional distance variable is input, then the mean distance of the measurements is also output.

Spatial Autocorrelation Indices

Spatial autocorrelation indices identify whether point locations are spatially related, either clustered or dispersed. Two spatial autocorrelation indices are calculated. Both require an intensity variable in the primary file.

Moran's I (MoranI)

Moran's I statistic is the classic indicator of spatial autocorrelation. It is an index of covariation between different point locations and is similar to a product moment correlation coefficient, varying from –1 to +1. The MoranI routine calculates 6 statistics:

1. The sample size
2. Moran's I
3. The spatially random (expected) I
4. The standard deviation of I
5. A significance test of I under the assumption of normality (Z-test)
6. A significance test of I under the assumption of randomization (Z-test)

Values of I greater than the expected I indicate clustering while values of I less than the expected I indicate dispersion. The significance test indicates whether these differences are greater than what would be expected by chance. The tabular output can be printed.

Adjust for small distances

If checked, small distances are adjusted so that the maximum distance weighting is 1 (see documentation for details). This ensures that I will not become excessively large for points that are close together. This is the default setting.

Geary's C (GearyC)

Geary's C statistic is an alternative indicator of spatial autocorrelation. It is an index of paired comparison between different point locations and varies from 0 (similar values) to 2 (dissimilar values). The GearyC routine calculates 5 statistics:

1. The sample size
2. Geary's C
3. The spatial random (expected) C
4. The standard deviation of C
5. A significance test of I under the assumption of normality (Z-test)

Values of C less than the expected C indicate clustering while values of C greater than the expected C indicate dispersion. The significance test indicates whether the differences are greater than what would be expected by chance. The tabular output can be printed.

Adjust for small distances

If checked, small distances are adjusted so that the maximum distance weighting is 1 (see documentation for details). This ensures that C will not become excessively large or excessively small for points that are close together. This is the default setting.

Distance Analysis

Distance analysis provides statistics about the distances between point locations. It is useful for identifying the degree of clustering of points. It is sometimes called second-order analysis. There are three routines for describing properties of the distances and there are two routines that output distance matrices.

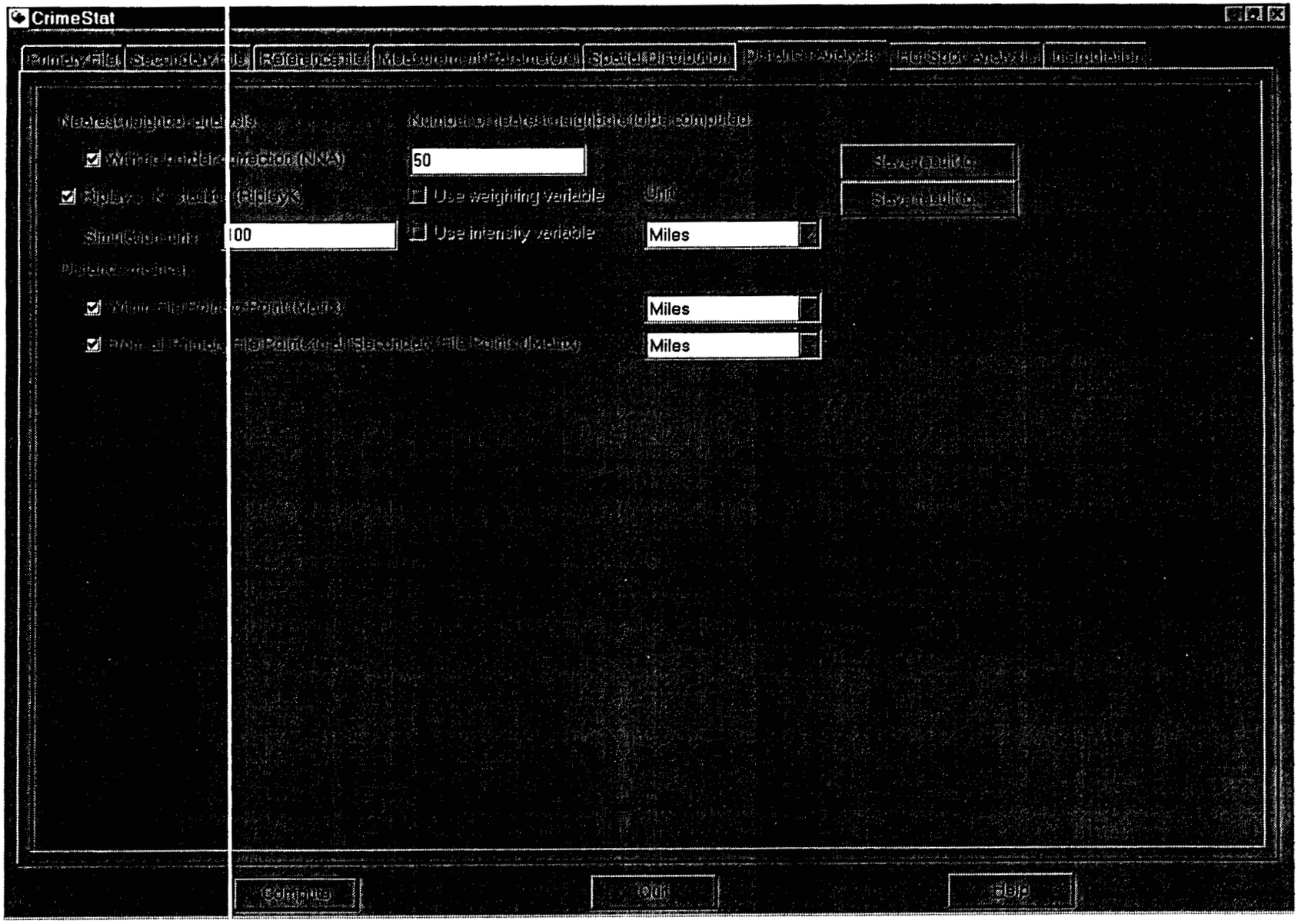
Nearest Neighbor Analysis (Nna)

The nearest neighbor index provides an approximation about whether points are more clustered or dispersed than would be expected on the basis of chance. It compares the average distance of the nearest other point (nearest neighbor) with a spatially random expected distance by dividing the empirical average nearest neighbor distance by the expected random distance (the nearest neighbor index). The nearest neighbor routine requires that the geographical area be entered on the Measurement Parameters page and that direct distances be used. The *Nna* routine calculates 10 statistics:

1. The sample size
2. The mean nearest neighbor distance in meters, feet and miles
3. The standard deviation of the nearest neighbor distance in meters, feet and miles
4. The minimum distance in meters, feet and miles
5. The maximum distance in meters, feet and miles
6. The mean random distance (for both the maximum bounding rectangle and the user input area, if provided
7. The mean dispersed distance in meters, feet and miles (for both the maximum bounding rectangle and the user input area, if provided)
8. The nearest neighbor index (for both the maximum bounding rectangle and the user input area, if provided)
9. The standard error of the nearest neighbor index (for both the maximum bounding rectangle and the user input area, if provided)
10. A significance test of the nearest neighbor index (Z-test)

The tabular results can be printed, saved to a text file, or saved as a '.dbf' file.

Figure 2.6: Distance Analysis Screen



Number of nearest neighbors

The K-nearest neighbor index compares the average distance to the K^{th} nearest other point with a spatially random expected distance. The user can indicate the number of K-nearest neighbors to be calculated, if more than one are to be calculated. *CrimeStat* will calculate 3 statistics for each order specified:

1. The mean nearest neighbor distance in meters for the order
2. The expected nearest neighbor distance in meters for the order
3. The nearest neighbor index for the order

The *Nna* routine will use the user-defined area unless none is provided in which case it will use the maximum bounding rectangle. The tabular results can be printed, saved to a text file or output as a '.dbf' file.

Linear Nearest Neighbor Analysis

The linear nearest neighbor index provides an approximation about whether points are more clustered or dispersed along road segments than would be expected on the basis of chance. It is used with indirect (Manhattan) distances and requires the input of the total length of a road network on the measurement parameters page (see Measurement Parameters). That is, if indirect distances are checked on the measurement parameters page, then the linear nearest neighbor will be calculated. The linear nearest neighbor index is the ratio of the empirical average linear nearest neighbor distance to the expected linear random distance. The *Nna* routine calculates 9 statistics for the linear nearest neighbor index:

1. The sample size
2. The mean linear nearest neighbor distance in meters, feet and miles
3. The minimum distance between points along a grid network
4. The maximum distance between points along a grid network
5. The mean random linear distance
6. The linear nearest neighbor index
7. The standard deviation of the linear nearest neighbor distance in meters, feet and miles
8. The standard error of the linear nearest neighbor index
9. A t-test of the difference between the empirical and expected linear nearest neighbor distance

Number of linear nearest neighbors

Nna can calculate K-nearest linear neighbors and compare this distance the average linear distance to the K^{th} nearest other point with a spatially random expected distance. The user can indicate the number of K-nearest linear neighbors to be calculated, if more than one are to be calculated. *CrimeStat* will calculate 3 statistics for each order specified:

1. The mean linear nearest neighbor distance in meters for the order
2. The expected linear nearest neighbor distance in meters for the order
3. The linear nearest neighbor index for the order

Ripley's K Statistic (RipleyK)

Ripley's K statistic compares the number of points within any distance to an expected number for a spatially random distribution. The empirical count is transformed into a square root function, called L (see documentation for more details). Values of L that are greater than the upper limit of the simulations indicate concentration while values of L less than the lower limit of the simulations indicate dispersion. L is calculated for each of 100 distance intervals (bins). The RipleyK routine calculates 6 statistics:

1. The sample size
2. The maximum distance in meters, feet and miles
3. 100 distance bins
4. The distance for each bin
5. The transformed statistic, L(t), for each distance bin
6. The expected random L under complete spatial randomness, L(csr)

In addition, *CrimeStat* can estimate the sampling distribution by running spatially random simulations over the study area. If one or more spatially random simulations are specified, there are 6 additional statistics:

7. The minimum L value for the spatially random simulations
8. The maximum L value for the spatially random simulations
9. The 2.5 percentile L value for the spatially random simulations
10. The 97.5 percentile L value for the spatially random simulations
11. The 0.5 percentile L value for the spatially random simulations
12. The 99.5 percentile L value for the spatially random simulations

The tabular results can be printed, saved to a text file, or saved as a '.dbf' file.

Distance Matrices

CrimeStat can calculate the distances between points for a single file or the distances between points for two different files. These matrices can be useful for examining the frequency of different distances or for providing distances for another program.

Within File Point-to-Point (Matrix)

This routine outputs the distance between each point in the primary file to every other point in a specified distance unit (miles, nautical miles, feet, kilometers, or meters). The Matrix output can be saved to a text file.

From All Primary File Points to All Secondary File Points (Imatrix)

This routine outputs the distance between each point in the primary file to each point in the secondary file in a specified distance unit (miles, nautical miles, feet, kilometers, or meters). The IMatrix output can be saved to a text file.

'Hot Spot' Analysis

'Hot spot' (or cluster) analysis identifies groups of incidents that are clustered together. It is a method of second-order analysis that identifies the cluster membership of points. There are three statistics that can be used to identify 'hot spots': 1) Nearest neighbor hierarchical spatial clustering; 2) K-means clustering; and 3) Local Moran Statistics.

Nearest Neighbor Hierarchical Spatial Clustering (Nnh)

The nearest neighbor hierarchical spatial clustering routine groups points together on the basis of spatial proximity. The user defines a significance level associated with a threshold, a minimum number of points that are required for each cluster, and an output size for displaying the clusters with ellipses.

First, the threshold distance is the lower limit of the confidence interval around a random expected distance. The default value is 0.1 (i.e., fewer than 10% of distances could be expected to be as small or smaller by chance). Pairs of points that are closer together than the threshold distance are grouped together, whereas pairs of points that are greater than the threshold distance are ignored. The smaller the significance level that is selected, the smaller the threshold distance. Move the slider bar to the desired likelihood level.

Second, the minimum number of points required for each cluster allows the user to specify a minimum number of points for each cluster. The default is 10 points. Third, the output size for the clusters can be adjusted by the second slider bar. These are the number of standard deviations defined by the ellipse, from one standard deviation (the default value) to five standard deviations. Typically, one standard deviation will cover about 65% of the cases whereas five standard deviations will cover more than 99% of the cases.

Clustering is hierarchical in that the first-order clusters are treated as separate points to be clustered into second-order clusters, and the second-order clusters are treated as separate points to be clustered into third-order clusters, and so on. Higher-order clusters will be identified only if the distance between their centers are closer than a new threshold distance. The results can be printed, saved to a text file, output as a '.dbf' file, or output as ellipses to *ArcView* '.shp', *MapInfo* '.mif' or *Atlas*GIS* '.bna' files

Significance level

The threshold distance is adjusted by a significance level on the upper slider bar that indicates the Type I error for a one-tailed lower range of a confidence interval around

Figure 2.7: 'Hot Spot' Analysis Screen

an expected spatially random distance. Distances smaller than this threshold are candidates for clustering. The range of values vary from 0.5 likelihood for a Type I error (i.e., a distance that is equal to an expected spatially random distance) to 0.0001 likelihood for a Type I error (i.e., a distance that is very unlikely to come from a spatially random process). The higher the p-level chosen, the larger the area the clusters will cover with larger ellipses. The smaller the likelihood, then clusters will cover smaller areas with smaller ellipses. However, the higher the p-level chosen, the greater the likelihood that clusters could be chance groupings. Slide the bar to choose a significance level.

Minimum points per cluster

Restrictions on the number of clusters can be placed by defining a minimum number of points that are required. The default is 10. If there are too few points allowed, then there will be many very small clusters. By increasing the number of required points, the number of clusters will be reduced.

Output size for ellipses

The output size for the clusters can be adjusted by the lower slider bar. These are the number of standard deviations defined by the ellipse, from one standard deviation (1X - the default value) to five standard deviations (5X). The default value is one standard deviation. Typically, one standard deviation will cover about 60% of the cases whereas five standard deviations will cover more than 99% of the cases, although the exact percentage will depend on the distribution. Slide the bar to select the number of standard deviations for the ellipses.

K-means Clustering (KMeans)

The K-means clustering routine is a procedure for partitioning all the points into K groups in which K is a number assigned by the user. The default K is 5. The routine finds K seed locations in which the distance between points within a cluster are small but the distances between seed locations are large. If K is small, the clusters will typically cover larger areas. Conversely, if K is large, the clusters will typically cover smaller areas. The results can be printed, saved to a text file, output as a '.dbf' file, or output as ellipses to *ArcView* '.shp', *MapInfo* '.mif' or *Atlas*GIS* '.bna' files

Local Moran Statistics (L-Moran)

The local Moran statistic applies the Moran's I statistic to individual points (or zones) to assess whether particular points/zones are spatially related to the nearby points (or zones). The statistic requires an intensity variable in the primary file. Unlike the global Moran's I statistic, the local Moran is applied to each individual point/zone. The index points to clustering or dispersion relative to the local neighborhood. Points (or zones) with high I values have an intensity value that is higher than their neighbors while points with low I values have intensity values lower than their neighbors. The output can be printed or output as a '.dbf' file.

Adjust for small distances

If checked, small distances are adjusted so that the maximum distance weighting is no higher than 1 (see documentation for details). This ensures that the local I will not become excessively large for points that are grouped together. This is the default setting.

Interpolation

The interpolation tab allows estimates of point density using the kernel density smoothing method. There are two types of kernel density smoothing: one applied to a single distribution of points and the other applied to two different distributions. Each type has variations on the method that can be selected. Both types require a reference file that is overlaid on the study area (see Reference File). The kernels are placed over each point and the distance between each reference cell and each point is evaluated by the kernel function. The individual kernel estimates for each cell are summed to produce an overall estimate of density for that cell. The intensity and weighting variables can be used in the kernel estimate. The densities can be converted into probabilities.

Single Kernel Density Estimate (KernelDensity)

The single kernel density routine estimates the density of points for a single distribution by overlaying a symmetrical surface over each point, evaluating the distance from the point to each reference cell by the kernel function, and summing the evaluations at each reference cell.

File to be interpolated

The estimate can be applied to either the primary file (see Primary file) or a secondary file (see Secondary File). Select which file is to be interpolated. The default is the Primary.

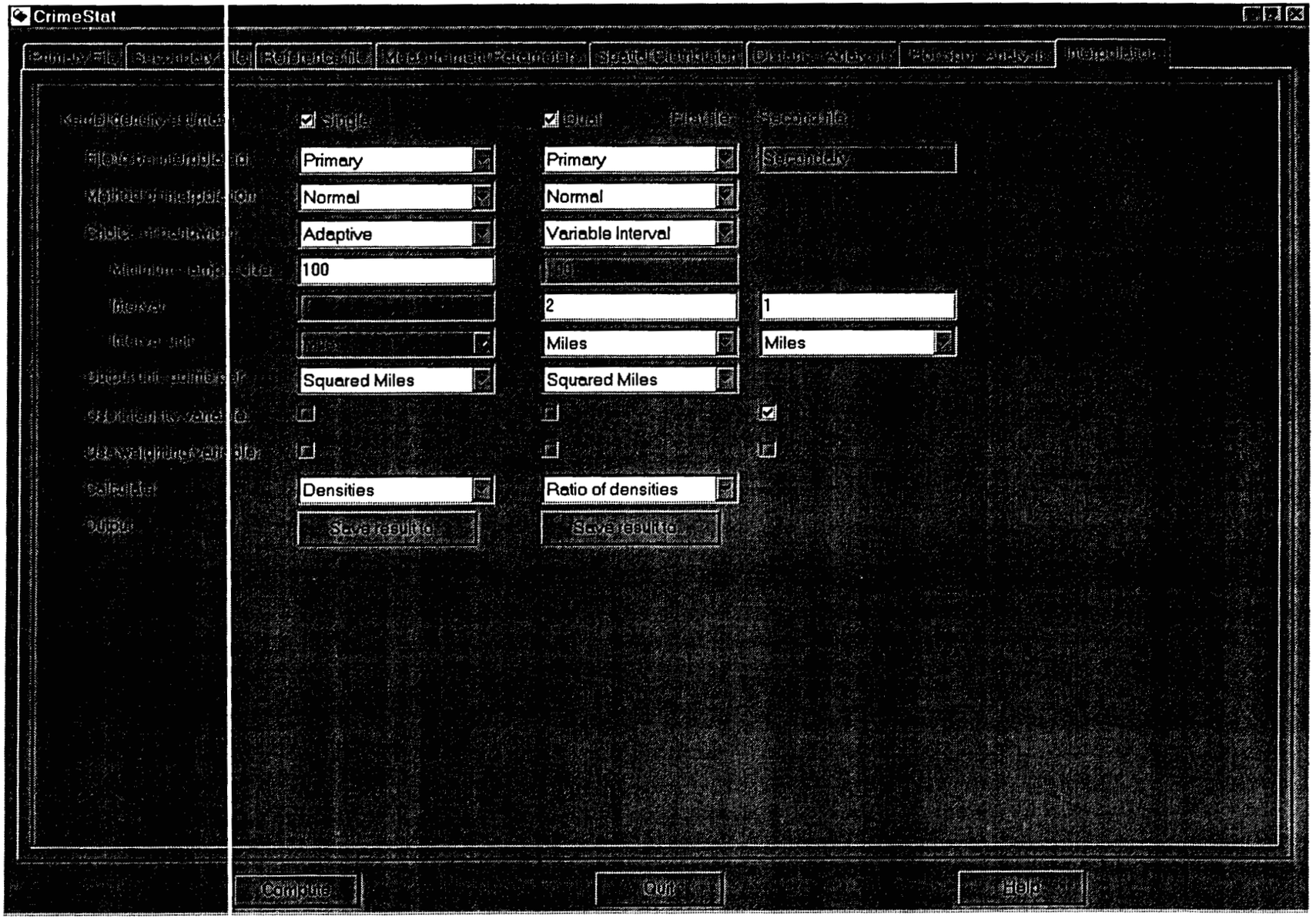
Method of interpolation

There are two types of kernel distributions that can be used to estimate the density points. The normal distribution overlays a normal distribution over each point, which then extends over the area defined by the reference file. This is the default kernel function. The quartic kernel overlays a surface that only extends for a limited distance from each point. The two methods produce similar results although the normal is generally smoother for any given bandwidth.

Choice of bandwidth

The kernels are applied over a limited search distance, called 'bandwidth'. For the normal kernel, bandwidth is the standard deviation of the normal distribution. For the quartic kernel, bandwidth is the radius of a circle defined by the surface. For both types,

Figure 2.8: Interpolation Screen



larger bandwidth will produce smoother density estimates. For each, both adaptive and fixed bandwidth intervals can be selected.

Adaptive bandwidth

An adaptive bandwidth distance is identified by the minimum number of other points found within a circle drawn around a single point. A circle is placed around each point, in turn, and the radius is increased until the minimum sample size is reached. Thus, each point has a different bandwidth interval. This is the default bandwidth setting. The user can modify the minimum sample size. The default is 100 points.

Fixed bandwidth

A fixed bandwidth distance is a fixed interval for each point. The user must define the interval and the distance units by which it is calculated (miles, nautical miles, feet, kilometers, meters).

Output units

Specify the density units as points per square mile, per squared nautical miles, per square feet, per square kilometers, or per square meters. The default is points per square mile.

Use intensity variable

If an intensity variable is being interpolated, then this box should be checked.

Use weighting variable

If a weighting variable is being used in the interpolation, then this box should be checked.

Calculate densities or probabilities

Select whether densities (points per square unit of area) or probabilities (the proportion of all points) are to be output for each cell. The default is densities.

Output

The results can be output as a *Surfer for Windows* file (for both an external or generated reference file) or as an *ArcView* '.shp', *MapInfo* '.mif', *Atlas*GIS* '.bna', or *ArcView Spatial* file (only if the reference file is generated by *CrimeStat*).

Dual Kernel Density Estimate (DuelKernel)

The dual kernel density routine compares two different distributions involving the primary and secondary files. A 'first' file and 'second' file need to be defined. The comparison allows the ratio of the first file divided by the second file, the logarithm of the ratio of the first file divided by the second file, the difference between the first file and second file (i.e., first file – second file), or the sum of the first file and the second file.

File to be interpolated

Identify which file is to be the 'first file' (primary or secondary) and which is to be the 'second file (primary or secondary). The default is Primary for the first file and Secondary for the second file.

Method of interpolation

There are two types of kernel distributions that can be used to estimate the density points. The normal distribution overlays a normal distribution over each point, which then extends over the area defined by the reference file. This is the default kernel function. The quartic kernel overlays a surface that extends only for a limited distance from each point. The two methods produce similar results although the normal is generally smoother for any given bandwidth.

Choice of bandwidth

The kernels are applied to a limited search distance, called 'bandwidth'. For the normal kernel, bandwidth is the standard deviation of the normal distribution. For the quartic kernel, bandwidth is the radius of a circle defined by the surface. For both types, larger bandwidth will produce smoother density estimates. For each, adaptive, fixed and variable bandwidth intervals can be selected.

Adaptive bandwidth

An adaptive bandwidth distance is identified by the minimum number of other points found within a circle drawn around a single point. A circle is placed around each point, in turn, and the radius is increased until the minimum sample size is reached. Thus, each point has a different bandwidth interval. This is the default bandwidth setting. The user can modify the minimum sample size. The default is 100 points.

Fixed bandwidth

A fixed bandwidth distance is a fixed interval for each point. The user must define the interval and the distance units by which it is calculated (miles, nautical miles, feet, kilometers, meters). The default is one mile.

Variable bandwidth

A variable bandwidth allows separate fixed intervals for both the first and second files. For each, the user must define the interval and the distance units by which it is calculated (miles, nautical miles, feet, kilometers, meters). The default is one mile for both the first and second files.

Output units

Specify the density units as points per square mile, per square nautical miles, per square feet, per square kilometers, or per square meters. The default is points per square mile.

Use intensity variable

For the first and second files separately, check the appropriate box if an intensity variable is being interpolated.

Use weighting variable

For the first and second files separately, check the appropriate box if a weighting variable is being used in the interpolation.

Calculate densities or probabilities

Select whether densities (points per square unit of area) or probabilities (the proportion of all points) are to be output for each cell. The default is densities.

Output

The results can be output as a *Surfer for Windows* file (for both an external or generated reference file) or as an *ArcView* '.shp', *MapInfo* '.mif', *Atlas*GIS* '.bna', or *ArcView Spatial* file (only if the reference file is generated by *CrimeStat*).

Dynamic Data Exchange (DDE) Support

CrimeStat supports Dynamic Data Exchange (DDE). See Appendix A in the documentation or the online help screens for more information.

Part II: *CrimeStat* Instructions and Statistics

Chapter 3
Entering Data into *CrimeStat*

The graphic user interface of *CrimeStat* is a tabbed form (figure 3.1). There are eight tabs, four for data definition and four for analysis:

Data Definition

Primary file	Data file of incident/point locations (Required)
Secondary file	Secondary data file of incident/point locations
Reference file	File for referencing interpolations
Measurement Parameters	Areal and linear characteristics of study area

Analysis

Spatial distribution	Basic characteristics of the incident distribution
Distance analysis	Characteristics of the distances between points
'Hot Spot' analysis	Tools for identifying 'Hot Spots'
Interpolation	Three-dimensional density analysis

This section discusses the first four tabs.

Required Data

Coordinates

CrimeStat analyzes point data, defined geographically by X and Y coordinates. These X/Y coordinates represent a single location where either an incident occurred (e.g., a burglary) or where a building or other object can be represented as a single point. A point will have X and Y coordinates in a spherical or Cartesian system. In a spherical coordinate system, each point can be defined by longitude (for X) and latitude (for Y). In a projected coordinate system, such as State Plane or UTM, each X and Y is defined by feet or meters from an arbitrary reference origin. *CrimeStat* can handle both spherical and projected points. For some uses, coordinates can be polar, that is defined as angles from an arbitrary reference vector, usually direct north.[1] One of the routines in the program calculates the angular mean and variance of a collection of angles.

Point data can be obtained from a number of sources. The most frequent would be the various incident data bases stored by a police department, which could include calls for service, crime reports, or closed cases. Other sources of incident data can include secondary data from other agencies (e.g., hospital records, emergency medical service records, locations of businesses) or even sampled data (Levine and Wachs, 1986a; 1986b). There are also point data from broadcast sources, such as radios, televisions, or microwaves.

Figure 3.1: Basic *CrimeStat* Program Layout

Intensities and weights

For some uses, points can have *intensity* values or *weights*. These are optional inputs in *CrimeStat*. An *intensity* is a value assigned to a point location aside from the X/Y coordinates. It is another variable, typically denoted as a Z-value. For example, if the point location is the location of a police station, then the intensity could be the number of calls for service over a month at that station. Or, to use census geography, if the point is the centroid of a census tract, then the intensity could be the population of that census tract. In other words, an intensity is a variable assigned to a particular location.

Some of the routines in *CrimeStat* require an intensity value (e.g., the spatial autocorrelation indices) and others can utilize a point location with an intensity value assigned (e.g., kernel density interpolation). If no intensity value is assigned, the routines which require it cannot be run while the routines which can utilize it will assume that the intensity is 1 (i.e., that all points have equal intensity).

A *weight* occurs when different point locations are to receive differential statistical treatment. For example, if a police department has designated different areas for service, for example 'urban' and 'rural', a value can be assigned for each of these areas (e.g., '1' for urban and '2' for rural). Most of the routines in *CrimeStat* will use the weights in the calculations. Weights would be useful if different zones are to be evaluated on the basis of another variable. For example, suppose a police department has divided its service area into urban and rural. In the rural part, there are twice as many patrol officers assigned per capita than in the urban areas; the higher population densities in the urban areas are assumed to compensate for the longer travel distances in the rural areas. Let's assume that all crimes occurring in the rural areas receive a weight of 2 while those in the urban area receive a weight of 1. The police department then wants to estimate the density of household burglaries relative to the population using the duel kernel density function (see Chapter 7). But, to reflect the differential assignment of police officers, the analysts use the service area as a weight. The result would be a per capita estimate of burglary density (i.e., burglaries per person), but weighted by the service area. It would provide an estimate of burglary risk adjusted for differential service in rural and urban areas. In most cases, there will no weights, in which case, all points are assumed to have an equal weight of '1'.

It is possible to have both intensities and weights, although this would be rare. For example, if the X and Y coordinates are the centroids of census tracts, a third variable - the total population of each census tract could be an intensity. There could also be an weighting based on service area. In calculating the Moran's I spatial autocorrelation index, the total population is used as an intensity while the service area is used as a weight. In this case, *CrimeStat* calculates a weighted Moran's I spatial autocorrelation.

But the use of both an intensity *and* a weight would be less common. For most of the statistics, a variable could be used as *either* a weight or an intensity, and the results will be the same. **However, one should be careful in assigning the same variable as both an intensity and a weight.** In such instances, cases may end up being weighted twice, which will produce distorted results.[2]

'Clean' Data

CrimeStat can input data in three formats - ASCII, *dbase III/IV* 'dbf', and *ArcView* 'shp'. It is essential that the files have X and Y coordinates as part of their structure. The program assumes that the assigned X and Y coordinates are correct. It reads a file - ASCII, 'dbf' or 'shp' and takes the given X and Y coordinates.

There is one other requirement of the data in order to work with *CrimeStat*. The data must be 'clean', that is that all the coordinates for the X and Y fields (or, for the angles if the data are angular) be assigned. Many police departments cannot geocode all addresses because of various errors. Different GIS programs will assign default values to unmatched records, for example a coordinate of 0,0 or -1,-1 for the X and Y coordinate. These records, however, may be left in the files. For GIS mapping, the unmatched records do not pose a problem because they cannot be seen within a particular view. However, for *CrimeStat*, they may pose a major problem since the program assumes the data are correct. Thus, if a record has a non-match code of 0,0 for the X and Y coordinate respectively, *CrimeStat* will treat the 0,0 as a real coordinate since, on a spherical system, it is a legitimate coordinate (latitude =0 is on the equator and longitude=0 is at the Greenwich Meridian. The point will be assigned to a location in the Atlantic Ocean off of western Africa.

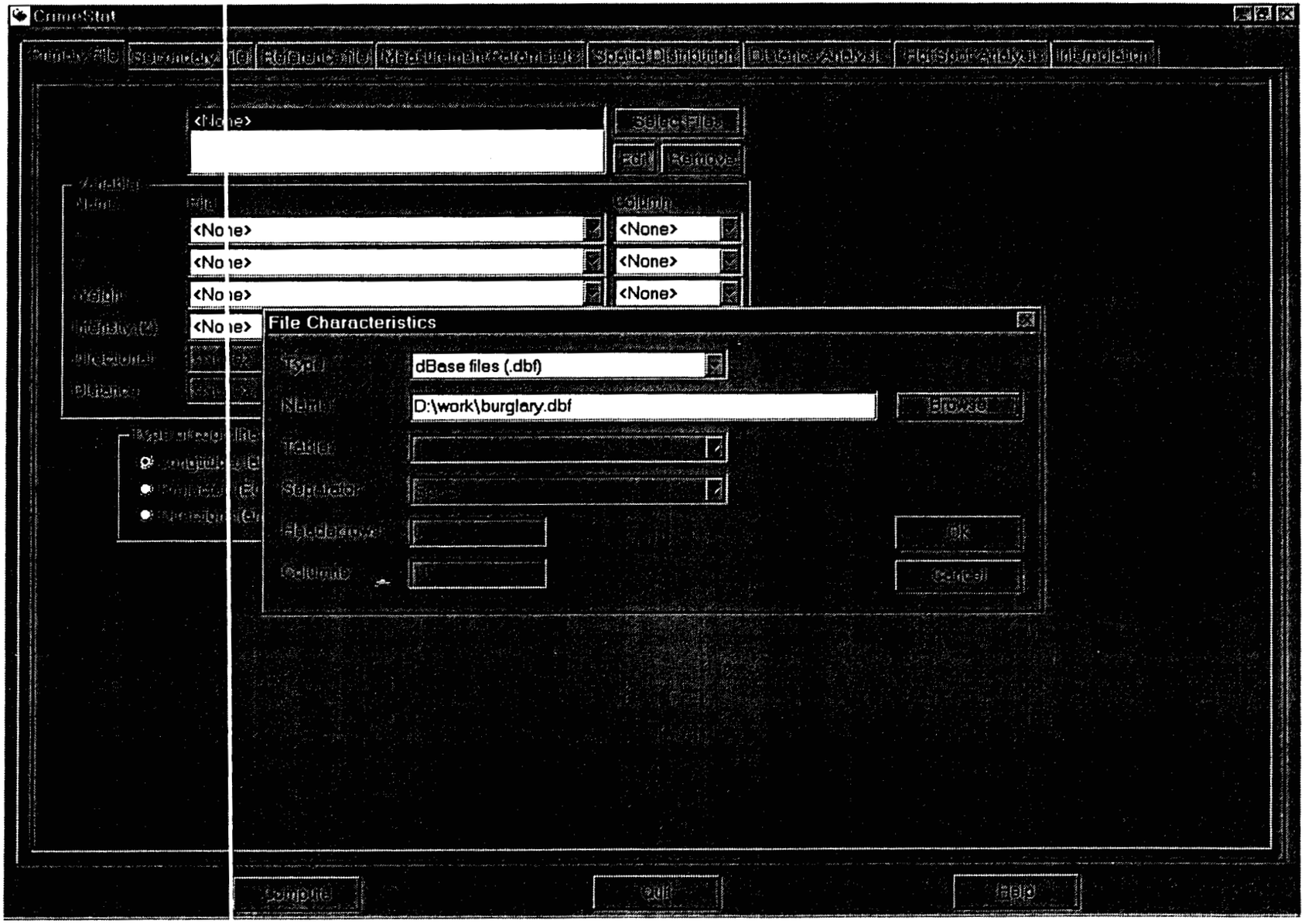
It is important that unmatched records before removed from the file before running *CrimeStat*. For most GIS packages, this is an easy operation involving the selection of legitimate cases and saving them to a separate file (e.g., SELECT IF LAT>0 AND LON<-1). Otherwise, if unmatched records are included, the calculated statistics will be meaningless (e.g., a mean center somewhere in the middle of the Atlantic Ocean). The program also assumes that the data that have been assigned are correct. Unfortunately, most geocoding routines wrongly assign some points, the consequence of which is spatial error introduced into the analysis.[3]

Similarly, if weights or intensities are used, *CrimeStat* assumes they are correct. It is essential that the users ensure that all weights and intensities values are legitimate and that blank or special values (e.g., -1) are not used. There is some error checking in the program. If *CrimeStat* finds a blank value, it will return an error code to the results output. But in many cases, the program cannot tell whether the data are correct or not.

Primary File

The *Primary File* is required and provides the coordinates of points of incidents. On the primary file tab, the user must first click on *Select Files*. A dialog box appears that allows the user to select which of the three file formats applies to the primary file (Figure 3.2). For each of the file formats, the user must define two characteristics - the type of file (ASCII, '.dbf' or '.shp') and the name of the file. There is a browse window which allows the user to find the file.

Figure 3.2: DBase File Selection



In developing this program, we have targeted it towards users of *ArcView*, *MapInfo* and *Atlas*GIS*. These GIS programs either store their attribute data in *dBase III/IV* format in a file with a 'dbf' extension (e.g., precinct1.dbf) or can read and write directly 'dbf' files. Many other GIS programs, however, also can read 'dbf' files. For *ArcView* and *MapInfo*, the X and Y coordinates which define crime incident points are not directly part of the 'dbf' file, but instead exist on the geographic file.

Input File Formats

ArcView

In *ArcView* the coordinates are stored on the 'shp' file, not the 'dbf' file. *CrimeStat* can read directly a 'shp' file so the 'dbf' file is not required to have the X and Y coordinates.

MapInfo

However, in *MapInfo*, the coordinates are stored in 'tab' files. To use *CrimeStat* with *MapInfo*, therefore, requires that the X and Y coordinates be assigned to two fields in the 'tab' file and then saved as a 'dbf' file. See the endnotes for directions on doing this.[4] Even in *ArcView*, some users may wish to export the points as a 'dbf' file because of other information that are on the records. The endnotes also list these directions.[5]

Atlas*GIS

In *Atlas*GIS*, on the other hand, a point file is already a 'dbf' file and will have fields for the X and Y coordinates.

ASCII

For an ASCII file, however, three additional attributes must be defined. The first is the type of character that is used to separate the variables in the file. There are four possibilities:[6]

Space (one or more, the default)
Comma
Semicolon
Tab

The second characteristic is the number of rows which have labels on them (*Header Rows*). Some ASCII files will have rows which label the names of the variables. The user should indicate the number if this is the case otherwise *CrimeStat* will produce an error code. The default is 0, that is the program assumes that there are no headers unless instructed otherwise. To change this, the user should insert the cursor in the appropriate cell, backspace to erase the default number and type in the correct number.

The third characteristic of an ASCII file that must be defined is the number of variables (columns or fields) in the file. With spherical or projected coordinates, there will be at least two variables (the X and Y coordinate) and there may be more if other variables are included in the file. However, with directional coordinates (see below), there may be only one. *CrimeStat* assumes that the number of columns in the ASCII file is two unless instructed otherwise. Again, the user should insert the cursor in the appropriate cell, backspace to erase the default number and type in the correct number.

After defining the file type and name, the user should click on *OK*.

Identifying Variables

After defining a file, either '.dbf', ASCII, or '.shp', it is necessary to identify the variables. Two variables are required and two are optional. The required variables are the X and Y coordinates. The user should indicate the file name that contains the coordinates by clicking on the drop down menu and highlighting the correct name. After having identified which file contains the X and Y coordinates, it is necessary to identify the variable name. Click on the drop down menu under *Column* and highlight the name of the variable for the X and Y coordinates respectively.[7] Figure 3.3 shows a correct defining of file and variable names for the primary file.

Multiple files can be entered on the primary file tab. However, only one can be utilized at a time. In theory, one can have separate files containing the X and Y coordinates, though in practice this will rarely occur.

Weight Variable

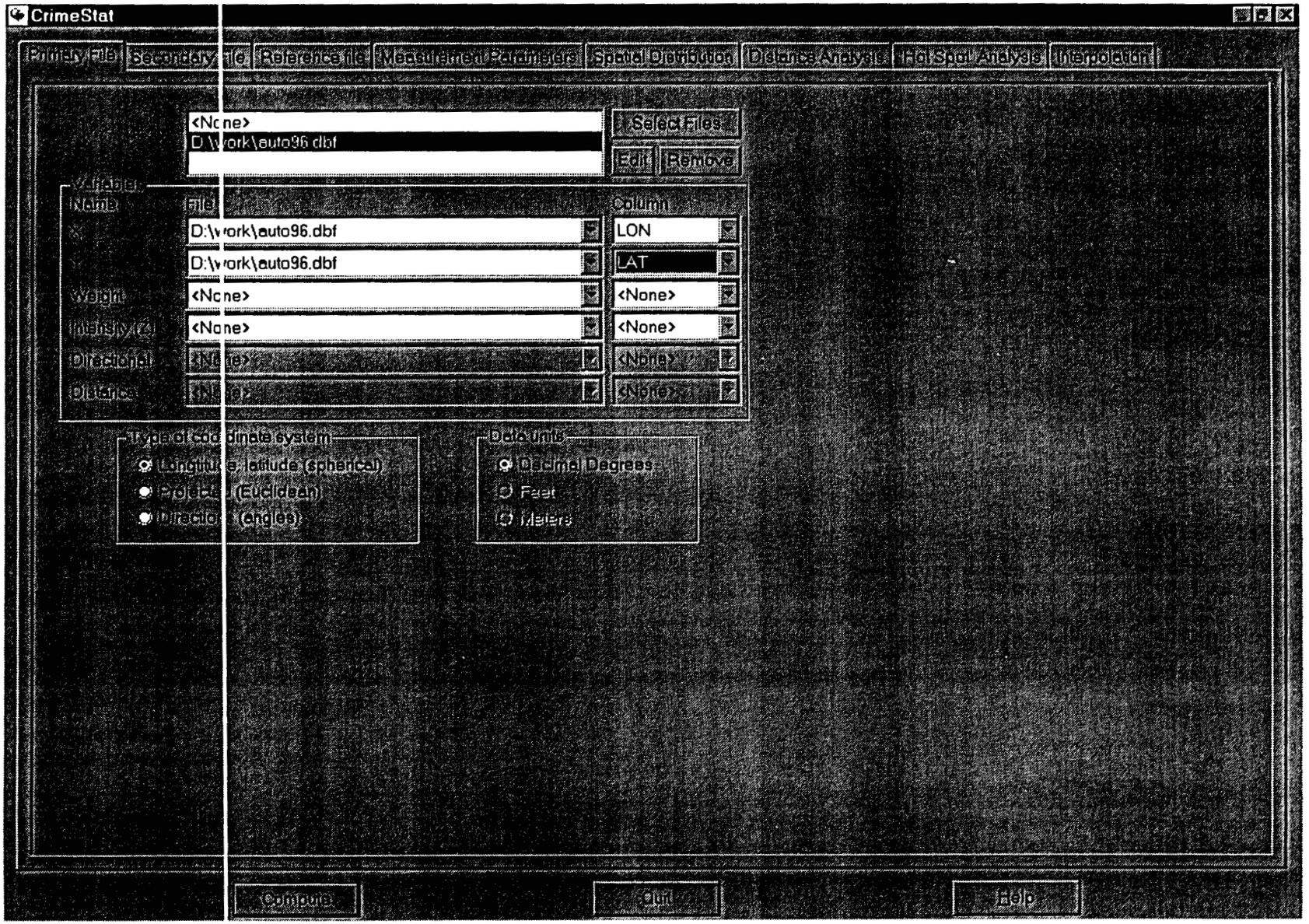
Sometimes, a point location is weighted. As mentioned above, weights are used when points represents areas and the areas are statistically treated differently. For most of the statistics, *CrimeStat* can weight the statistics during the calculation (e.g., the weighted mean center, the weighted nearest neighbor index).

By default, *CrimeStat* assigns a weight of 1 to each point. If the user does not define a weight variable, then the program assumes that each point has equal weight (i.e., 1). On the other hand, if there are weights, then the weight variable should be defined on the primary file screen and its name listed.

Intensity Variable

Similarly, a point location can have an intensity assigned to it. Most of the statistics in *CrimeStat* can use an intensity variable and some statistics require it (Moran's I, Geary's C and Local Moran). If no intensity is defined, *CrimeStat* will not calculate statistics requiring an intensity variable and, in statistics where an intensity is optional (e.g., interpolation), will assume a default intensity of 1. On the other hand, if there is an intensity variable, then this should be defined on the primary file screen and its variable name identified.

Figure 3.3: Primary File Definition



In general, be very careful about using *both* an intensity variable *and* a weighting variable. Use both only when there are separate weights and intensities. Most of the routines can use both intensities and weighting and may, consequently, double-weight cases. Figure 3.4 shows a primary file screen with an intensity variable defined.

Coordinate System

In addition to the primary file name and variable assignment, it is necessary to identify the type of coordinate system used and the units of measurement. *CrimeStat* recognizes three coordinate systems:

Spherical coordinates (longitude and latitude)

This is a universal coordinate system that measures location by angles from reference points on Earth.[8]

Projected coordinates

Projected coordinates are arbitrary coordinates based on a particular projection of the earth to a flat plane. They have an arbitrary origin (the place where X=0 and Y=0) and are almost always defined in units of feet or meters.[9]

CrimeStat can work with either spherical or projected coordinates. On the primary file tab, the user indicates which coordinate system is being used. If the coordinate system is spherical, then units are automatically assumed to be latitude and longitude in decimal degrees. If the coordinate system is projected, then it is necessary to specify whether the measurement units are feet or meters.

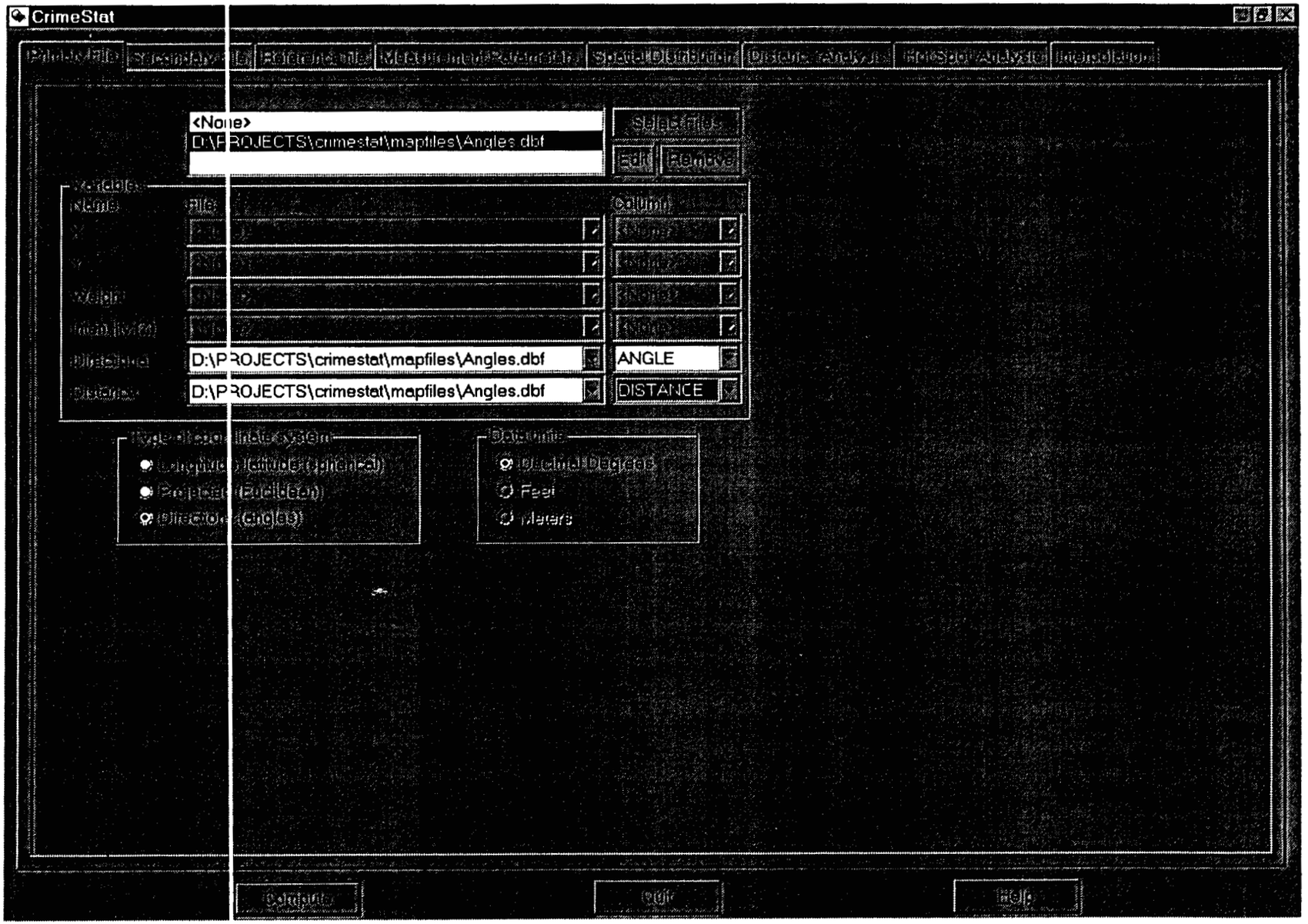
Directional coordinates

For some uses, a polar coordinate system can be used. Point locations are defined by angles from an arbitrary reference line, usually true north and vary between 0^0 and 360^0 in a clockwise rotation. All locations are measured as an angular deviation from the reference point and with distance being measured from a central location. *CrimeStat* has the ability to read in angles for use in calculating the angular mean and variance. In addition, if directional coordinates are used, an optional distance variable for each measurement can be used.

If the file contains directional coordinates (angles), define the file name and variable name (column) that contains the directional measurements. If used, define the file name and variable name (column) that contains the distance variable. Figure 3.5 shows the primary file definition using directions.

Figure 3.4: Primary File With Intensity Variable Defined

Figure 3.5: File Definition With Angles (Directions)



Secondary File

CrimeStat also allows for the inputting of a secondary file. For example, the primary file could be locations where motor vehicles were stolen while the secondary file could be the location where stolen vehicles were recovered. Alternatively, the primary file could be burglary locations while the secondary file could be police stations. *CrimeStat* can construct two different types of indices with a secondary file. First, it can calculate the distance from every primary file point to every secondary file point. For example, this might be useful in assessing where to place police cars in order to minimize travel distance in response to calls for service. Second, *CrimeStat* can utilize both primary and secondary files in estimating a three-dimensional density surface (see Chapter 7). For example, if the primary file are residential burglaries and the secondary file contains the centroids of census block groups with the population within each block group assigned as an intensity variable, then *CrimeStat* can estimate the density of burglaries relative to the density of population (i.e., burglary risk).

The secondary file can also be either a '.dbf', '.shp' or ASCII. As with a primary file, there must be an X and Y variable defined, but it must be in the same coordinate system and data units as the primary file. The secondary file can also have weights and intensities assigned. Figure 3.6 shows the inputting of an ASCII file for the secondary data set while figure 3.7 shows a correct definition of the secondary file characteristics.

Reference File

Several of the routines in *CrimeStat* generalize the point data to all locations in the study area, the one-variable and two-variable density interpolation routines. The generalization uses a reference file placed over the study area. Typically, it is a rectangular grid file (true grid), that is a rectangle with cells defined by columns and rows.; each grid cell is a rectangle and column-row combinations are used. It is possible to use a non-rectangular grid file under special circumstances (e.g., a grid with water, mountains or other jurisdictions removed), but a rectangular grid would be used in most cases. *CrimeStat* can read in a grid or can create one itself. Figure 3.8 shows a grid placed over both the County of Baltimore and the City of Baltimore in Maryland.

Existing Grid File

Many GIS programs can create uniform grids which cover a geographical area. As with the primary and secondary files, these need to be converted to either '.dbf', ASCII or '.shp' files. To use an existing grid file created in a GIS or another program, the user clicks on *From File* on the Reference File tab and selects the file.

There are three characteristics which should be identified for an existing grid file:

1. The name of the file. The user selects the file from a dialog box similar to the primary file (see figure 3.2).

Figure 3.6: ASCII File Selection of Secondary File

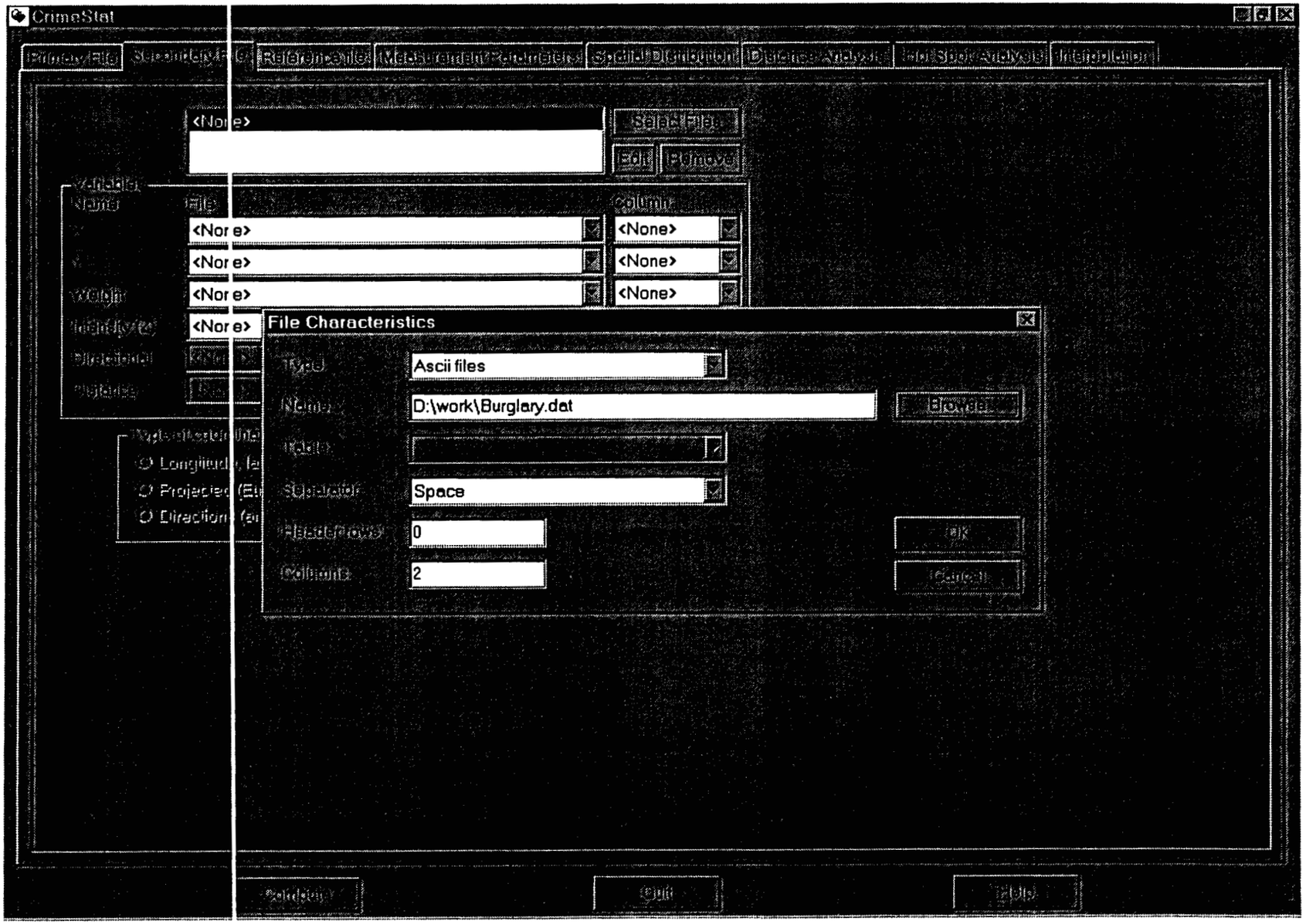
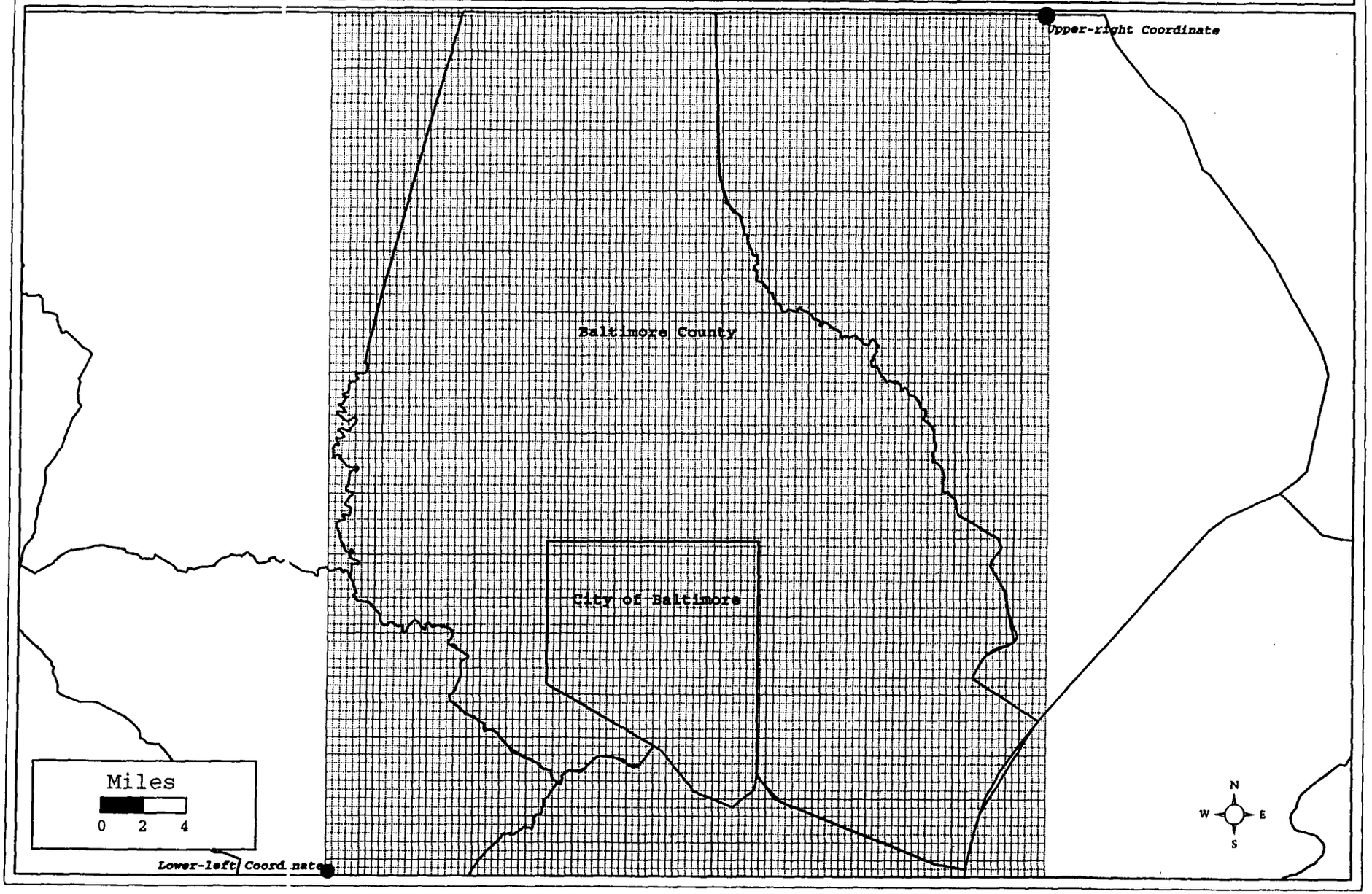


Figure 3.7: Secondary File Definition

Figure 3.8: Grid Cell Structure for Baltimore Region

108 Width x 100 Height Grid Cells



2. If the existing reference file is a true grid, the *True Grid* box should be checked.

3. If it is a true grid, the number of columns should be entered. *CrimeStat* will automatically count the number of records in the file and place it in the *Cells* box. When the number of columns is entered, *CrimeStat* will automatically calculate the number of rows.

Figure 3.9 shows a correctly defined reference file using an existing grid file. One must be careful in using a file which is not a grid. *CrimeStat* can output the results of the interpolation routines in several GIS formats - *Surfer for Windows, ArcView Spatial Analyst, ArcView, MapInfo* and *Atlas*GIS*. Of these, only the output to *Surfer for Windows* will allow the reference to be a shape other than a true grid. For the interpolation outputs of *ArcView Spatial Analyst, ArcView, MapInfo* and *Atlas*GIS*, it is essential that the reference file be a true grid.

Generating a Reference File

CrimeStat can also generate a true grid. There are two steps:

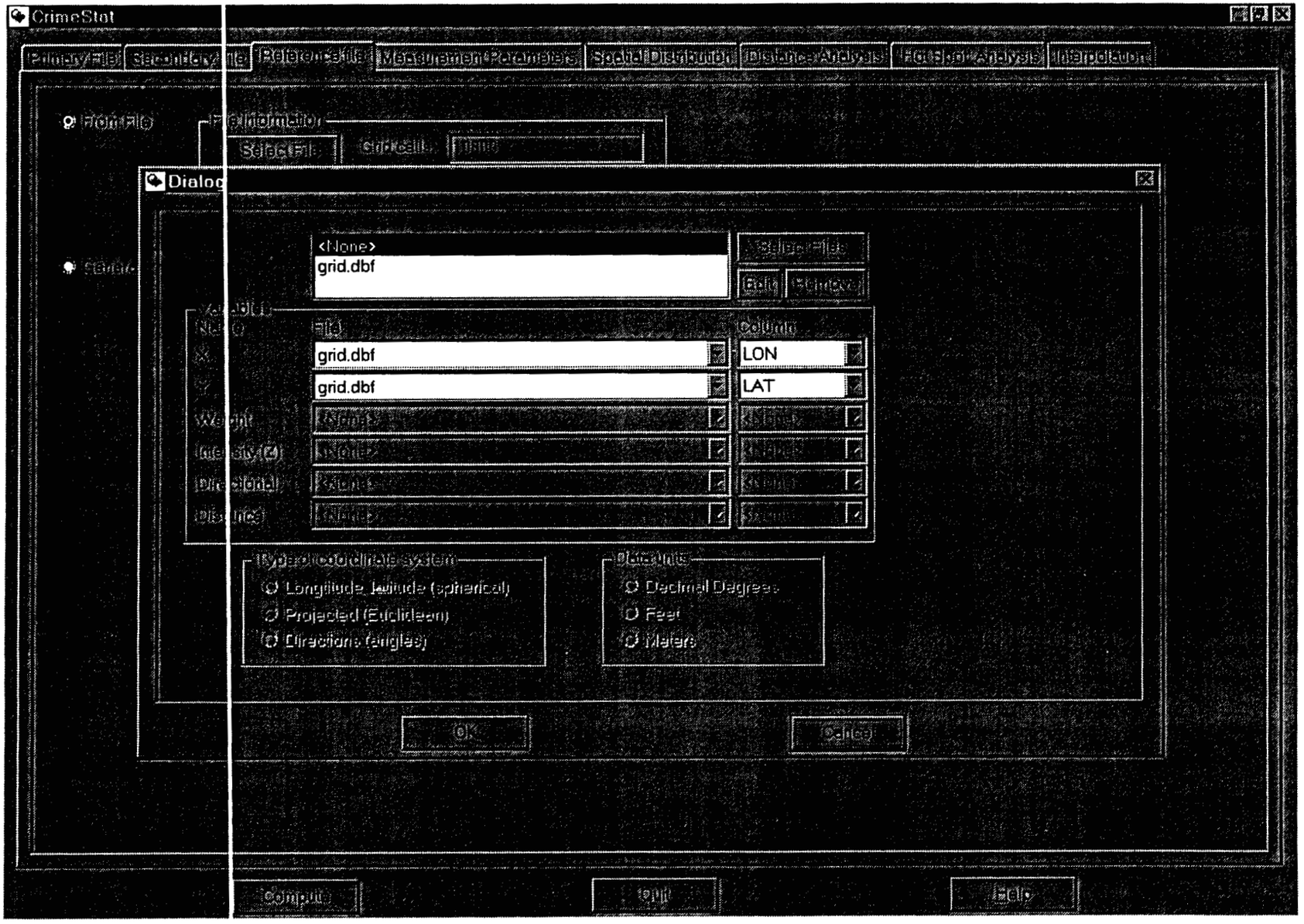
1. The user selects *Generated* from the Reference File tab and inputs the X and Y coordinates of the lower-left and upper-right coordinates of the grid. These coordinates must be the same as for the primary file.

Thus, if the primary file is using spherical lat/lon coordinates, then the grid file coordinates must also be lat/lon. Conversely, if the primary file coordinates are projected, then the grid file coordinates must also be projected, using the same measurement units (feet or meters). The lower-left and upper-right coordinates are those from a grid which covers the geographical area. A user should identify these with a GIS program or from a properly indexed map.

2. The user selects whether the grid is to be generated by cell spacing or by the number of columns.

With *By cell spacing*, the size of the cell is defined by its horizontal width, in the same units as the measurement units of the primary file. This would be used to maintain a certain size of spacing for a cell. For example, if the coordinate system is spherical and the lower-left coordinates are -76.90 and 39.20 degrees and the upper-right coordinates are -76.32 and 39.73 degrees (a grid which overlaps Baltimore City and Baltimore County), then the horizontal distance - the difference in the two longitudes (0.58 degrees) must be divided into appropriate sized intervals. At this latitude, the difference in longitudes is 34.02 miles. If a user wanted cell spacing of 0.01 degrees, then this would be entered and *CrimeStat* would calculate 59 columns (cells) in the horizontal direction, one for each interval of 0.01 and one for the fractional remainder. If the coordinate system is projected, then similar calculations would be made using the projected units (feet or meters).

Figure 3.9 Reference File Definition With An Existing File



Probably an easier way to specify the grid is to indicate the number of columns. By checking *By number of columns*, the user defines the number of columns to be calculated. *CrimeStat* will automatically calculate the cell spacing needed and will calculate the required number of rows. For example, using the same coordinates as above, if a user wanted half mile squares for the cells, then they would need approximately 68 cells in the horizontal direction since 34.02 miles divided by 0.5 mile squares equals about 68 cells. Figure 3.10 shows a correctly defined reference file where *CrimeStat* generates the reference grid with the number of columns being defined; in the example, 100 columns are requested.

Measurement Parameters

The final properties that complete data definition are the measurement parameters. On the *Measurement Parameters* tab, the user defines the geographical area and the length of street network for the study area, and indicates whether direct or indirect distances are to be used. Figure 3.11 shows the measurement parameters tab page.

Area and Length of Street Network

In calculating distances between points for two of the statistics - the nearest neighbor index and the Ripley 'K' index, the area for which the points fall within needs to be defined (the study area). The user indicates the area of the geographical coverage and the measurement units that distances are calculated (feet, meters, miles, nautical miles, kilometers). Unlike the data units for the coordinate system, which must be consistent, *CrimeStat* can calculate distances in any of these units. In some cases, analysis will be conducted on a subset of the study area, rather than the entire area. For each analysis, the user should identify the area of the subset for which distance statistics are to be calculated.

In addition, the linear nearest neighbor statistic uses the total length of the street network as a baseline for comparison (see chapter 5). If this statistic is to be used, the total length of the street network should be defined. Most GIS programs can sum the total length of the street network. Again, if subsets of the study are used, the user should indicate the appropriate length of street network for the subset so that the comparison is appropriate.

Direct and Indirect Distance

CrimeStat can calculate both direct and indirect distances. *Direct* distances are the shortest distance between two points. On a flat plane, that is with a projected coordinate system, the shortest distance between two points is a straight line. However, on a spherical coordinate system, the shortest distance between two points is a Great Circle line. Depending on the coordinate system, *CrimeStat* will calculate Great Circle distances using spherical geometry for spherical coordinates and Euclidean distances for projected coordinates. The drawings in figure 3.12 illustrate direct distances with a projected and spherical coordinate system. The shortest distance between point A and point B is either a

Figure 3.10: Reference File Definition By Generating A File

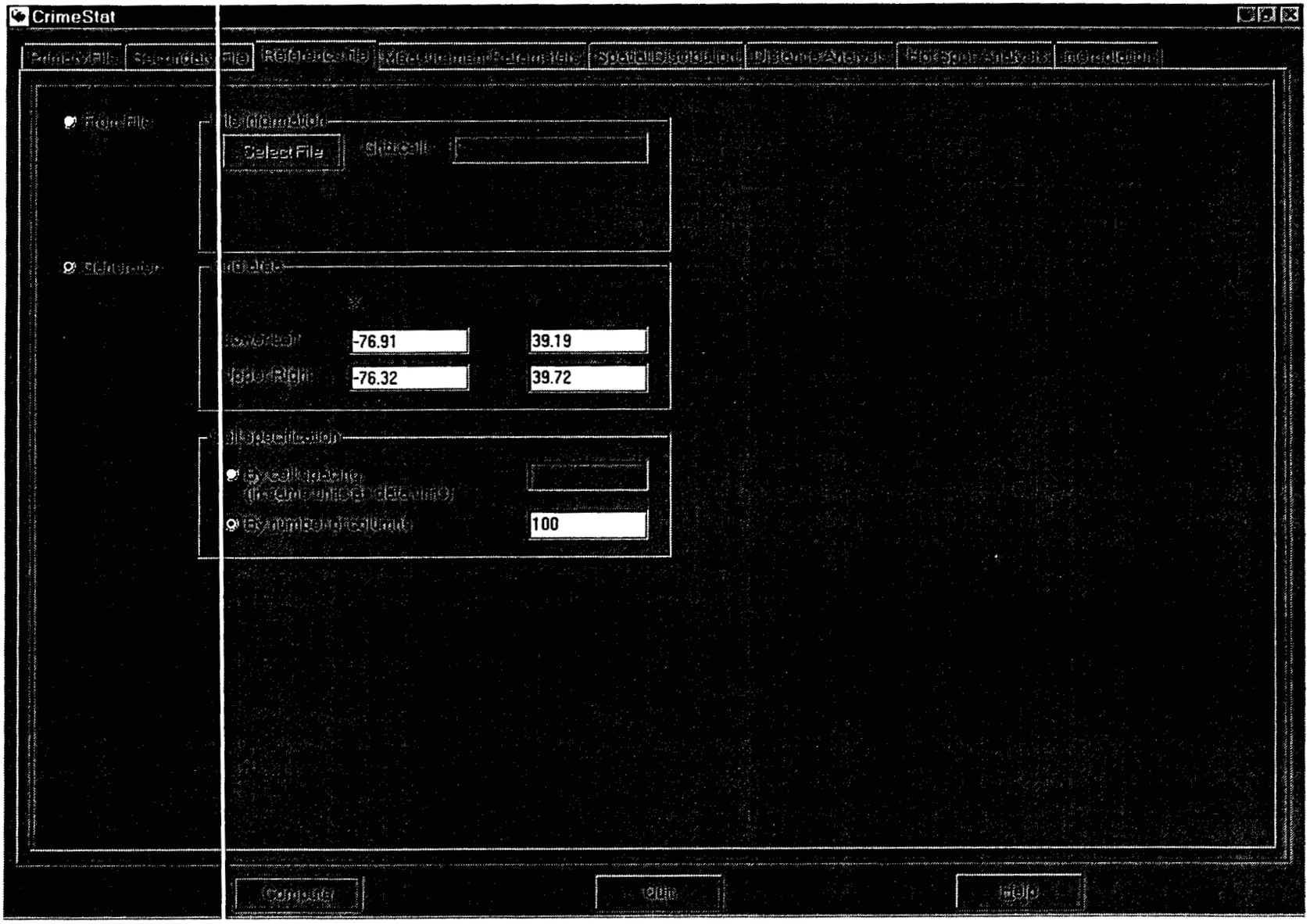


Figure 3.11: Measurement Parameters Definition

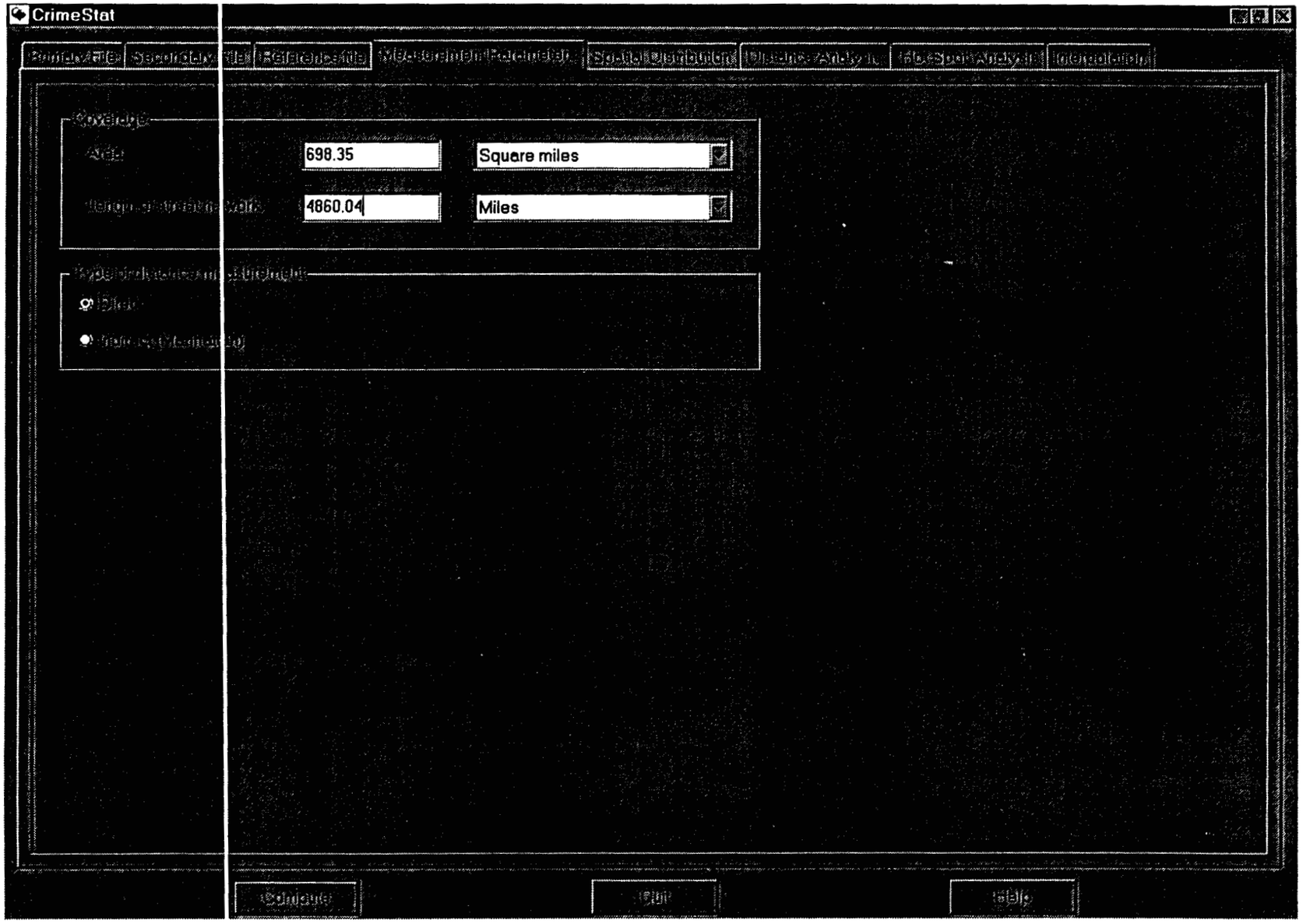
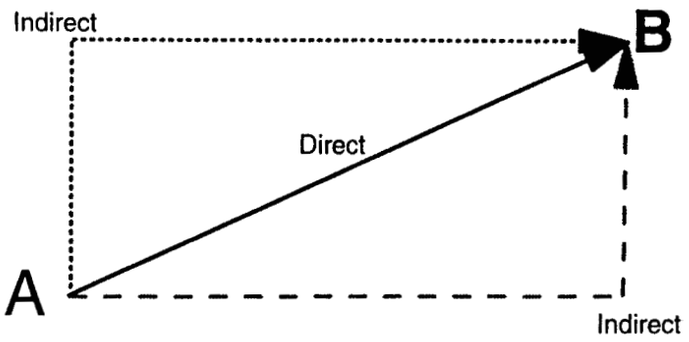


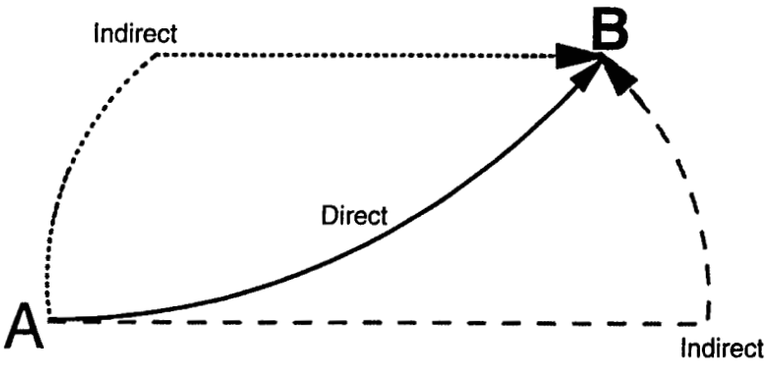
Figure 3.12: Direct and Indirect Distances

Two-dimensional Projected Geometry: Euclidean distance



A-B distance ('dotted route') = A-B distance('dashed route')

Three-dimensional Spherical Geometry: Great Circle distance



A-B distance ('dotted route') < A-B distance ('dashed route')

straight line (projected) or a Great Circle (spherical). For details see McDonnell, 1979 (chapter 1) or Snyder, 1987 (pp. 29-33).

Indirect distance is an approximation of travel on a rectangular road network. This is frequently called *Manhattan* distance, referring to the grid-like structure of Manhattan. Many cities, but certainly not all, lay out their streets in grids. The degree in which this is true varies. Older cities will not usually have grid structures whereas newer cities tend to use grid layouts more. Of course, no real city is a perfect grid, though some come close (e.g., Salt Lake City). Distances measured over a street network are always longer than a direct line or arc. In a perfect grid, travel can only occur in horizontal or vertical directions so that distances are the sum of the horizontal and vertical street lengths that have been traveled (i.e., one cannot cut diagonally across a block). Distances are measured as the sum of horizontal and vertical distances traveled between two points.

For some purposes, it may be useful to calculate distances that approximate an actual travel pattern rather than assume the shortest distance between points. In this case, indirect distances would be a more appropriate distance measurement than direct distances. Also, there is a linear nearest neighbor index which measures the distribution of point locations in relation to the street network rather than the geographical area and uses indirect distances. This will be discussed in Chapter 5. In this case, the use of indirect distances would be preferable than direct distances.[10]

Distance Calculations

Distances in *CrimeStat* are calculated with the following formulas:

Direct, projected coordinate system

Distance is measured as the hypotenuse of a right triangle using Euclidean geometry.

$$d_{AB} = \sqrt{(X_A + X_B)^2 + (Y_A + Y_B)^2} \qquad (3.1)$$

where d_{AB} is the distance between two points, A and B, X_A and X_B are the X-coordinates for points A and B in a projected coordinate system, Y_A and Y_B are the Y-coordinates for points A and B in a projected coordinate system.

Direct, spherical coordinate system

Distance is measured as the Great Circle distance between two points. All latitudes (ϕ) and longitudes (λ) are first converted into radians using:

$$\text{Radians } (\phi) = \frac{2\pi\,\phi}{360} \qquad (3.2)$$

$$\text{Radians } (\lambda) = \frac{2\pi\,\lambda}{360} \qquad (3.3)$$

Then, the distance between the two points is determined from

$$d_{AB} = 2^* \text{ Arcsin } \{ \text{Sin}^2[(\phi_B - \phi_A)/2] + \text{Cos } \phi_A * \text{Cos}\phi_B * \text{Sin}^2[(\lambda_B - \lambda_A)/2]^{1/2} \} \qquad (3.4)$$

with all angles being defined in radians where d_{AB} is the distance between two points, A and B, ϕ_A and ϕ_B are the latitudes of points A and B, and λ_A and λ_B are the longitudes of points A and B (Snyder, 1987, p. 30, 5-3a).

Indirect, projected coordinate system

Distance is measured as the sides of a right triangle using Euclidean geometry.

$$d_{AB} = (X_A - X_B) + (Y_A - Y_B) \qquad (3.5)$$

where d_{AB} is the distance between two points, A and B, X_A and X_B are the X-coordinates for points A and B in a projected coordinate system, Y_A and Y_B are the Y-coordinates for points A and B in a projected coordinate system.

Indirect, spherical coordinate system

Distance is measured by the average of summed Great Circle distances of two routes, one in the east-west direction followed by a north-south direction and the other in the north-south direction followed by an east-west direction.

$$d_{AB} = \frac{[d_{AB}(1) + d_{AB}(2)]}{2} \qquad (3.6)$$

where d_{AB} is the distance between two points, A and B, $d_{AB}(1)$ is the sum of distances between points A and B by measuring the Great Circle distance of the east or west direction from a particular latitude first, and adding this to the Great Circle distance of the north or south direction from that same latitude, and $d_{AB}(2)$ is the sum of distances between points A and B by measuring the Great Circle distance of the north or south direction from a particular longitude first, and adding this to the Great Circle distance of the east or west direction from that same longitude.

In the next chapter, the analysis of spatial distributions will be discussed.

Endnotes for Chapter 3

1. The spherical 'lat/lon' system is, of course, one type of polar coordinate system. But, it is a polar coordinate system with particular restrictions. Latitudes are angles up to 90^0, north or south of the Equator. Longitudes are angles from 0^0 to 180^0, east and west of the Greenwich Meridian. In the usual polar coordinate system, angles can vary from 0^0 to 360^0.

2. An alternative way to thinking about intensities and weights is to treat both as two different weights - weight #1 and weight #2. For example, weight #1 could be the population in a surrounding zone while weight #2 could be the employment in that same zone. Thus, incidents (e.g., burglaries) could be weighted both by the surrounding population and the surrounding employment. The analogy with double weights is not quite correct since several of the statistics (Moran's I, Geary's C and Local Moran) use only an intensity, but not a weight. The distinction between intensities and weights is historical, relating to the manner in which the statistics have been derived.

3. Many police departments use their GIS systems to assign X and Y coordinates, a process called *geocoding*. If they are using a street network which has a linear reference system (e.g., the U.S. Census Bureau's *TIGER* system; U.S. Census Bureau, 1996), then street addresses can be approximated using a geocoding algorithm in the GIS package (Huxhold, 1991; Levine and Kim, 1999). TIGER segments are generally street links between one intersection and another. Because they are defined with a direction (typically the direction in which they are digitized), they have a 'left' and 'right' side. Attributes of the street segment are attached to the database including the name of the street, the prefix (e.g., S., E., W., N.) and suffix respectively (e.g., St., Blvd., Dr.). In addition, street segments have two address ranges, for the 'left' and 'right' sides of the segment respectively. A GIS program will take an address and find the segment with the same name within which the street number falls and will select the side of the street and the approximate location along the address range. Different GIS programs have different levels of matching and accuracy, but they all approximate locations by relating the address to the street range on a street segment. In other words, the assigned X and Y coordinate is the approximate location of the event.

   Sometimes, police agencies will match the nearest intersection rather than the address. This is particularly true for motor vehicle crashes where a crash is defined by a primary street and a nearest reference street (Levine and Kim, 1998; Levine, Kim and Nitz, 1996a). Whether the incident location is identified by an address or by a nearest intersection, there is error in the assignment. For an address, the interpolated location is an approximation and could involve error. For example, most GIS geocoding algorithms assume that addresses are evenly distributed within a segment, which may not be true; they therefore assign the location on this assumption. Intersection matching also involves events do not always occur exactly at an intersection. There are more precise methods for geocoding that have been experimented with by police departments. One of these involves the use of the

Global Positioning System (GPS) in which a police officer will take a reading from a GPS receiver at a location (Kim and Parke, 1996; Harries and Canter, 1998). But even here there is error since GPS is vulnerable to nearby large objects (e.g., buildings, bridge) which can offset readings. Other methods for increasing geocoding precision involve the use of detailed parcel-based street maps which identify every address (Huxhold, 1991). In theory, these methods should be more precise than the algorithmic approach of geocoding packages. In time, as police departments adopt GPS and parcel-based street maps, the accuracy of geocoding will improve. For now, most police departments use the approximate methods.

But even with more precise methods, there is still error in the geocoding process. For example, a residential burglar will enter a building at one particular location (e.g., the back of the house). The geocoding method used will simply assign a location to the entire address, not distinguishing one part of the building with another. There have been some experiments with high-resolution overlay maps that identify buildings as polygons with particular shapes rather than as point locations (e.g., Harries and Canter, 1998). In time, this type of methodology will lead to even more accurate geocoding than current methods.

4. In *MapInfo*, point data are stored in a table. If the X and Y coordinates are not already part of the table, it will be necessary to add these fields.

   A. Click on *Table Maintenance TableStructure* <tablename>

   B. Click on *Add Field*

   C. Define the X field. If the coordinates are spherical, then an appropriate name might be Longitude or Lon. If the coordinates are projected, then X or XCoord might be appropriate names.

   D. Fill in the parameters of the new name.

      i. The type should be decimal.

      ii. The width should be sufficient to handle the longest string. With spherical coordinates, 12 would be sufficient.

      iii. Be sure to define an appropriate number of decimals places. With longitude, there should be at least 4 decimals places with 6 providing more accuracy. In a projected coordinate system, the number of decimal places would be usually 0 or 1.

   E. Click *OK* when finished.

   F. If a Map Basic Window is not already open, click on *Options ShowMapBasicWindow.*

G. Make the Map Basic Window active by click on its top border.

H. Inside the window, type

update <tablename> set <Xvariablename> = centroidX(obj)
update <tablename> set <Yariablename> = centroidY(obj)

After each line, hit <Enter>. The appropriate names would be chosen. For example, if the point table was named robberies and the coordinates were spherical, then the statements would be

update robberies set lon=centroidX(obj)
<Enter>
update robberies set lat=centroidY(obj)
<Enter>

I. The X and Y field names should be populated with the correct values for each point. To view the table, click on *Window NewBrowserWindow* <filename>.

J. Save the table as a 'dbf' with 'Save Copy As <name>'. Be sure to specify that the file is to be saved in 'dbf' format.

5. The following steps would be followed to add X and Y coordinates to a 'dbf' file of point locations in *ArcView*.

A. Make the point table active by clicking on it.

B. Open the theme table by clicking on the *Open Theme Table* button.

C. Click on *Table StartEditing*.

D. Click on *Edit AddField*.

E. In the Field Definition window, define a name for the X field (e.g., X, Longitude, Lon).

F. Define the parameters for the X field.

a. Make sure that the type is *Number*

b. Be sure that the width is large enough to handle the largest value. For spherical coordinates (i.e., longitude, latitude), 12 columns should be sufficient. For a projected coordinate system, the number of columns should be two larger than the largest value.

c. Be sure that there are sufficient decimal places. With a spherical coordinate system, the minimum should be 4 decimal places with 6 being more accurate. With a projected coordinate system, 0 or 1 decimal places would be sufficient.

G. Click *OK* when finished.

H. Repeat steps 5 through 7 for the Y field.

I. For the X and Y variable in turn, click on the field name to highlight it.

J. Click on the *Calculate* button.

K. Double-click on the *[Shape]* field name.

L. In the dialog box, type *.GetX* for the X field and *.GetY* for the Y field after *[Shape]*, that is

[Shape].GetX
[Shape].GetY

M. Click *OK* when finished. The field will be populated with the X and Y values for the points in the same units as the data (e.g., lat/lon, feet or meters for UTM or State Plane Coordinates).

6. Note that in an ASCII file, a tab *looks like* it is separated by spaces. However, the underlying ASCII code is different and *CrimeStat* will treat these characteristics differently. That is, if the separator is a tab but the user indicates that it is a space, *CrimeStat* will not properly read the data.

7. Hint: If you type the first letter of the name (e.g., 'L' for longitude), then the program will find the first name that begins with that letter). Typing the letter again will find the second name, and so forth.

8. Since the world is approximately round, all lines are actually circles that eventually come back on to themselves. These are called *Great Circles* because they divide the Earth into two equal halves (Greenhood, 1964). On a sphere, such as the Earth, the shortest distance between any two points is a Great Circle. There are an infinite number of Great Circles, but coordinates are only referenced to two Great Circles. North-south lines are called *Meridians* (and are half Great Circles) and east-west lines are called *Parallels*. The basic reference parallel is the Equator, which is a Great Circle, and the two reference meridians are the Greenwich Meridian and the International Date Line (which is actually the same Great Circle on two sides of the earth).

There are two coordinates - *Longitude* and *Latitude*. For longitude, all east-west

directions are defined as an angle from 0^0 to 180^0 with 0^0 being at the Greenwich Meridian and 180^0 being the International Date Line. All directions east of the Greenwich Meridian have a positive longitude whereas all directions west of this meridian have a negative longitude. For example, in the United States, Washington, DC, has a longitude of approximately -77.03 degrees because it is west of the Greenwich Meridian whereas New Delhi, India has a longitude of approximately +77.20 degrees because it is east of the Greenwich Meridian. These locations are approximate because cities cover areas and only a single point within the city has been classified (the center or *centroid* of the city).

For latitude, all north-south directions are defined in terms of an angle from the equator, which has a latitude of 0^0. The maximum is the North or South Poles which have latitudes of $+90^0$ and -90^0 respectively. Locations that are north of the Equator have a positive latitude while locations that are south have a negative latitude. Thus, in the United States, Los Angeles has a latitude of approximately +34.06 degrees whereas Buenos Aires in Argentina has an approximate latitude of -34.60 degrees.

To measure variations between degrees, subdivision of the angles are necessary. The traditional use of spherical coordinates divides angles into multiples of 60 and defines angles in relation to the reference Great Circles. Thus, each degree is subdivided into 60 minutes and each minute, in turn, can be divided into 60 seconds. For example, New York City has an approximate longitude of 73 degrees 58 minute 22 seconds West and an approximate latitude of 40 degrees 52 minutes 46 seconds North. However, with the advent of computers, most coordinates are now converted into decimal degrees. Thus, New York City has an approximate longitude of -77.973 degrees and an approximate latitude of +40.880 degrees. The conversion is simply

Decimal degrees = Degrees + Minutes/60 + Seconds/3600

9. Because the Earth is curved, any two dimensional representation produces distortion. The spherical latitude/longitude system (called 'lat/lon' for short) is a universal coordinate system. It is universal because it utilizes the spherical nature of the Earth and each location has a unique set of coordinates. Most other coordinate systems are projected because they are portrayed on a two-dimensional flat plane. Strictly speaking, spherical coordinates - longitudes and latitudes, are not X and Y coordinates since the world is round. However, by convention, they are often referred to as X and Y coordinates, particularly if a small section of the Earth is projected on a flat plane (a computer screen or a printed map).

Projections differ in how they 'flatten' or *project* a sphere onto a two dimensional plane. Typically, there are four properties of maps which cannot all be maintained in any two dimensional representation:

Shape - maintaining correct shape of a land body

Area - if the space represented on a map covers the same area throughout the map, it is called an equal-area map. The proportionality is maintained.

Distance - the distance between two points is in constant scale (i.e., the scale does not change)

Direction - the direction from a point towards another point is true.

Any projection creates one or more types of distortion and particular projections are chosen in order to have accuracy in one or two of these properties. Different projections portray different types of information. Most projections assume that the Earth is a sphere, a situation that is not completely true. The Earth's diameter at the equator is slightly greater than the distance between the poles (Snyder, 1987). The circumference of the Earth between the Poles is about 24,860 miles on a meridian; the circumference at the Equator is about 75 miles more.

There is an infinite number of projections. However, only a couple dozen have been used in practice (Greenhood, 1964; Snyder, 1987; Snyder and Voxland, 1989). They are based on projections of the sphere onto a cylinder, cone or flat plane. In the United States, several common coordinate systems are used. Theoretically, the projection and the coordinate system can be distinguished (i.e., a particular projection could use one of several coordinate systems, e.g. meters or feet). However, in practice, particular projections use common coordinates. Among the most common in use in the United States are:

A. Mercator - The *Mercator* is an early projection, and one of the most famous, which is used for world maps. The projection is done on a cylinder, which is vertically centered on a meridian, but touching a parallel. The globe is projected on the cylinder as if light is emanating from the center of the globe while the Earth turns. The meridians cut the equator at equal intervals. However, they maintain parallel lines, unlike the globe where they converge at the poles. The longitudes are stretched with increasing latitude (in both north and south directions) up until the 80^{th} parallel. The effect is that shape is approximately correct and direction is true. Distance, however, is distorted. For example, on a Mercator map, Greenland appears as big as the United States, which it is not. Distances can be measured in any units for a Mercator though usually they are measured in miles or kilometers.

B. Transverse Mercator - If the Mercator is rotated 90^0 so that the cylinder is centered on a parallel, rather than a meridian, it is called a *Transverse Mercator*. The cylinder is projected as being horizontal but is touching a meridian. The Transverse Mercator is divided into narrow north-south zones in order to reduce distortion. The meridian that the cylinder is touching is called the *Central Meridian* of the zone. Distances are accurate within a limited distance from the central meridian. Thus, the boundaries of zones are selected in order to maintain reasonable distance accuracy. In the U.S.,

many states use the Transverse Mercator as the basis for their state plane coordinate system including Arizona, Hawaii, Illinois, and New York.

C. Universal Transverse Mercator (UTM) - In 1936, the International Union of Geodesy and Geophysics established a standard use of the Transverse Mercator, called the *Universal Transverse Mercator* (or UTM). In order to reduce distortion, the globe is divided into 60 zones, 6 degrees of longitude wide. For latitude, each zone is divided further into strips of 8 degrees latitude, from 84° N to 80° S. Within each band, there is a central meridian which, in theory, would be geodetically true. But, to reduce distortion across the area covered by each zone, scale along the central meridian is reduced to 0.9996. This produces two parallel lines of zero distortion approximately 180 km away from the central meridian. Scale at the boundary of the zone is approximately 1.0003 at U.S. latitudes. Coordinates are expressed in meters. By convention, the origin is the lower left corner of the zone. From the origin, *Eastings* are displacements eastward and from the origin, *Northings* are displacements northward. The central meridian is given an Easting of 500,000 meters. The Northing for the equator varies depends on the hemisphere. For the northern hemisphere, the equator has a Northing of 0 meters. For the southern hemisphere, the Equator has a Northing of 10,000,000 meters. The UTM system was adopted by the U.S. Army in 1947 and has been adopted by many national and international mapping agencies. Distances are always measured in meters in UTM.

D. Oblique Mercator - There are a number of cylindrical projections which are neither centered on a meridian (as in the Mercator) or on a parallel (as in the Transverse Mercator). These are called *Oblique Mercator* projections because the cylinder is centered on a line which is oblique to parallels or meridians. In the U.S., the *Hotine Oblique Mercator* is used for Alaska.

E. Lambert Conformal Conic - The *Lambert Conformal Conic* is a projection made on a cone, rather than a cylinder. Lambert's conformal projection centers the cone over a central location (usually the North Pole) and the cone 'cuts' through the globe at parallels chosen to be standards. Within those standards, shapes are true and meridians are straight. Outside those standards, parallels are spaced at increasing intervals the further north or south they go to reduce distance distortion. The projection is the basis of many state plane coordinate systems, including California, Connecticut, Maryland, Michigan, and Virginia.

F. Alber's Equal-Area - Another projection on a cone is the *Albers Equal-Area* except that parallels are spaced at decreasing intervals the further north or south they are placed from the standard parallels. The map is an equal-area projection and scale is true in the east-west direction.

G. State Plane Coordinates - Every state in the United States has an official

coordinate system, called the *State Plane Coordinate System.* Each state is divided into one or more zones and a particular projection is used for each zone. With the exception of Alaska, which uses the Hotine Oblique Mercator for one of its eight zones, all state plane coordinate systems use either the Transverse Mercator or the Lambert Conformal Conic. Each state's shape determines which projection is chosen to represent that state. Typically, states extending in a north-south direction use Transverse Mercator projections while states extending in an east-west direction use Lambert Conformal Conic projections. But, there are exceptions, such as California which uses the Lambert. Projections are chosen to minimize distortion over the state. Several states use both projections (Florida, New York) and Alaska uses all three. Distances are measured in feet.

See Snyder (1987) and Snyder and Voxland (1989) for more details on these and other projections including the mathematical transformations used in the various projections. Other good references are Maling (1973), Robinson, Sale, Morrison and Muehrcke (1984), and the Committee on Map Projections (1986).

10. With a projected coordinate system, indirect distances can be measured by perpendicular horizontal or vertical lines on a flat plane because all direct paths between two points have equal distances. For example in figure 3.12, whether the distance is measured from point A north to the Y-coordinate of point B and then eastward until point B is reached or, alternatively, from point A eastward to the X-coordinate of point B, then northward until point B is reached, the distances will be the same. One of the advantages of a Manhattan geometry is that travel distances that are direct (i.e., that are pointed towards the final direction) are equal.

    With a spherical coordinate system, however, Manhattan distances are not equal with different routes. Because the distance between two points at the same latitude decreases with increasing latitude (north or south) from the equator, the path between two points will differ on the route with Manhattan rules. In figure 3.12, for example, it is a longer distance to travel from point A eastward to the longitude of point B, before traveling north to point B than to travel northward from point A to the same latitude as point B before traveling eastward to point B. Consequently, *CrimeStat* modifies the Manhattan rules for a spherical coordinate system by calculating both routes between two points and averaging them. This is called a *Modified Spherical Manhattan Distance.*

Chapter 4
Spatial Distribution

In this chapter, the spatial distribution of crime incidents will be discussed. The statistics that are used in describing the spatial distribution of crime incidents will be explained and will be illustrated with examples from *CrimeStat.* For the examples, crime incident data from Baltimore County and Baltimore City will be used. Figure 4.1 shows the user interface for the spatial distribution statistics in *CrimeStat.* For each of these, the statistics will first be presented followed by examples of their use in crime analysis.

Centrographic Statistics

The most basic type of descriptors for the spatial distribution of crime incidents are *centrographic statistics.* These are indices which estimate basic parameters about the distribution (Lefever, 1926; Furfey, 1927; Bachi, 1957; Neft, 1962, Hultquist, Brown and Holmes, 1971; Ebdon, 1988). They include:

1. Mean center
2. Center of minimum distance (median center)
3. Standard deviation of X and Y coordinates
4. Standard distance deviation
5. Standard deviational ellipse

They are called centrographic in that they are two dimensional correlates to the basic statistical moments of a single-variable distribution - mean, standard deviation, skewness, and kurtosis (see Bachi, 1957). They have been applied to crime analysis by Stephenson (1980) and, more recently, by Langworthy and Jefferis (1998).

Because two dimensions adds complexity not seen in one dimension, these statistical moments have been modified to be appropriate. Figure 4.2 shows how the centrographic statistics are selected in *CrimeStat.*

Mean Center

The simplest descriptor of a distribution is the *mean center.* This is merely the mean of the X and Y coordinates. It is sometimes called a *center of gravity* in that it represents the point in a distribution where all other points are balanced if they existed on a plane and the mean center was a fulcrum (Ebdon, 1988; Burt and Barber, 1996).

For a single variable, the mean is the point at which the sum of all differences between the mean and all other points is zero. Unfortunately, for two variables, such as the location of crime incidents, the mean center is not necessarily the point at which the sum of all distances to all other points is minimized. That property is attributed to the center of minimum distance (see below). However, the mean center can be thought of as a point where both the sum of all differences between the mean X coordinate and all other X

Figure 4.1: Spatial Analysis Layout

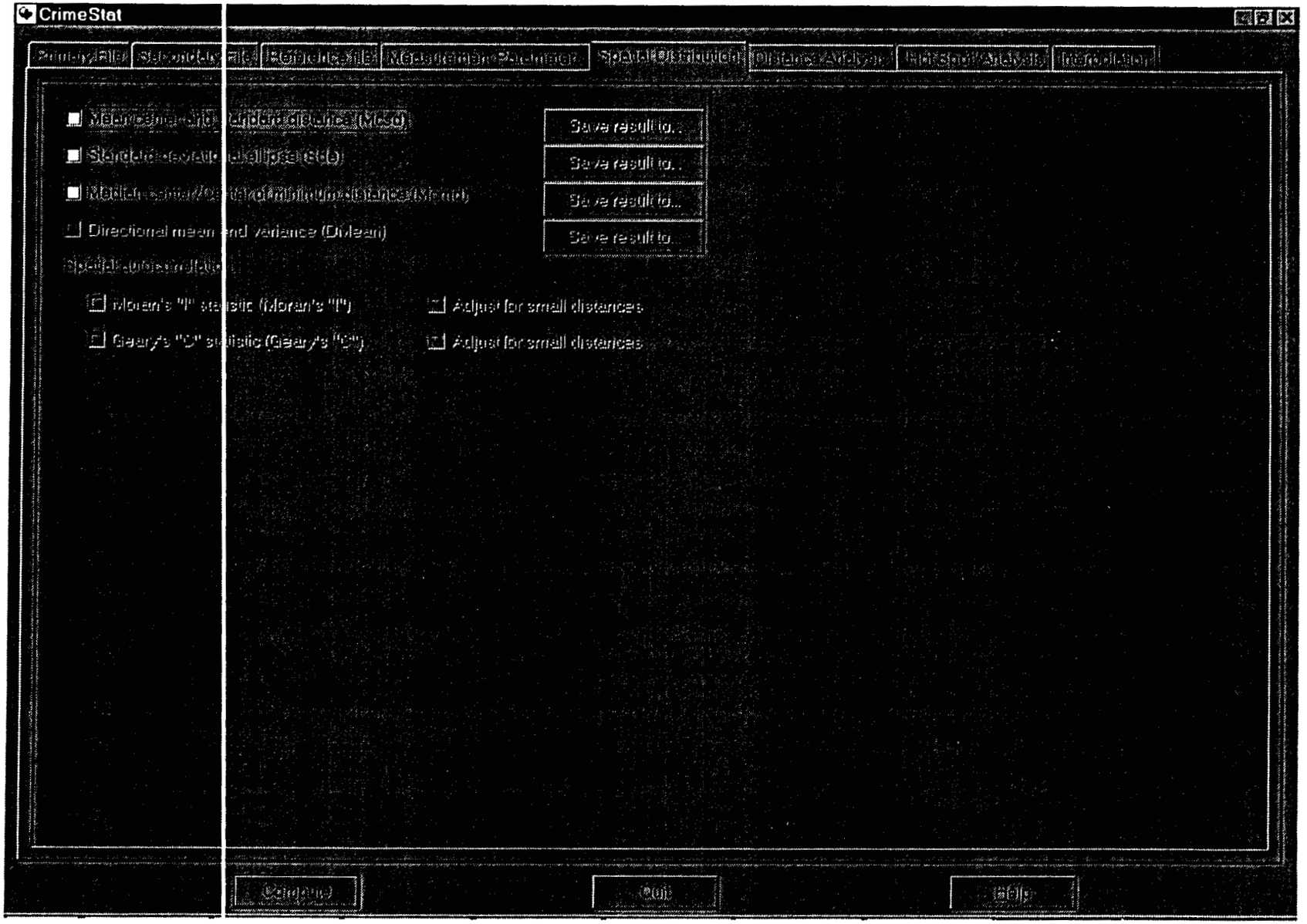
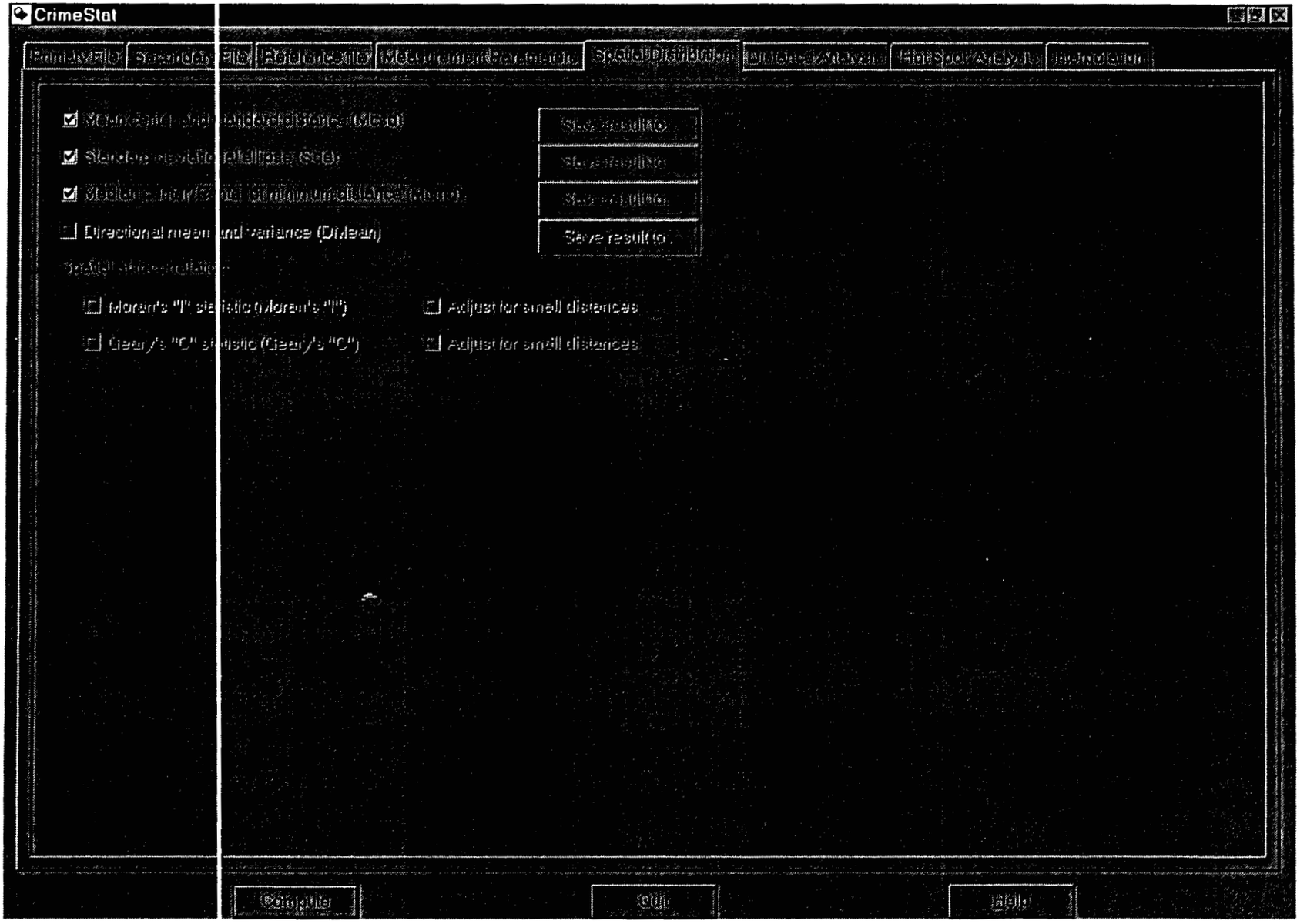


Figure 4.2: Selecting Centrographic Statistics



coordinates is zero and the sum of all differences between the mean Y coordinate and all other Y coordinates is zero.

The formula for the mean center is:

$$\bar{X} = \sum_{i=1}^{N} \frac{X_i}{N} \qquad \bar{Y} = \sum_{i=1}^{N} \frac{Y_i}{N} \tag{4.1}$$

where X_i and Y_i are the coordinates of individual locations and N is the total number of points.

To take a simple example, the mean center for burglaries in Baltimore County has spherical coordinates of longitude -76.608482, latitude 39.348368 and for robberies longitude -76.620838, latitude 39.334816. Figure 4.3 illustrates these two mean centers.

Weighted Mean Center

A weighted mean center can be produced by weighting each coordinate by another variable, W_i. For example, if the coordinates are the centroids of census tracts, then the weight of each centroid could be the population within the census tract. Formula 4.1 is extended slightly to include a weight.

$$\bar{X} = \sum_{i=1}^{N} \frac{W_i X_i}{N} \qquad \bar{Y} = \sum_{i=1}^{N} \frac{W_i Y_i}{N} \tag{4.2}$$

The advantage of a weighted mean center is that points associated with areas can have the characteristics of the areas included. For example, if the coordinates are the centroids of census tracts, then the weight of each centroid could be the population within the census tract. This will produce a different center of gravity than, say, the unweighted center of all census tracts. *CrimeStat* allows the mean to be weighted by either the weighting variable or by the intensity variable. Users should be careful, however, not to weight the mean with both the weighting and intensity variable unless there is an explicit distinction being made between weights and intensities.

To take an example, in the six jurisdictions making up the metropolitan Baltimore area (Baltimore City, and Baltimore, Carroll, Harford, Howard and Anne Arundel counties), the mean center of all census block groups is longitude -76.619121, latitude 39.304244. This would be an *unweighted* mean center of the block groups. On the other hand, the mean center of the 1990 population for the Baltimore metropolitan area had coordinates of longitude -76.625186 and latitude 39.304186, a position slightly southwest of the unweighted mean center. Weighting the block groups by median household income produces a mean center which is still more southwest. Figure 4.4 illustrates these three mean centers.

Figure 4.3: Burglary and Robbery in Baltimore County

Comparison of Mean Centers

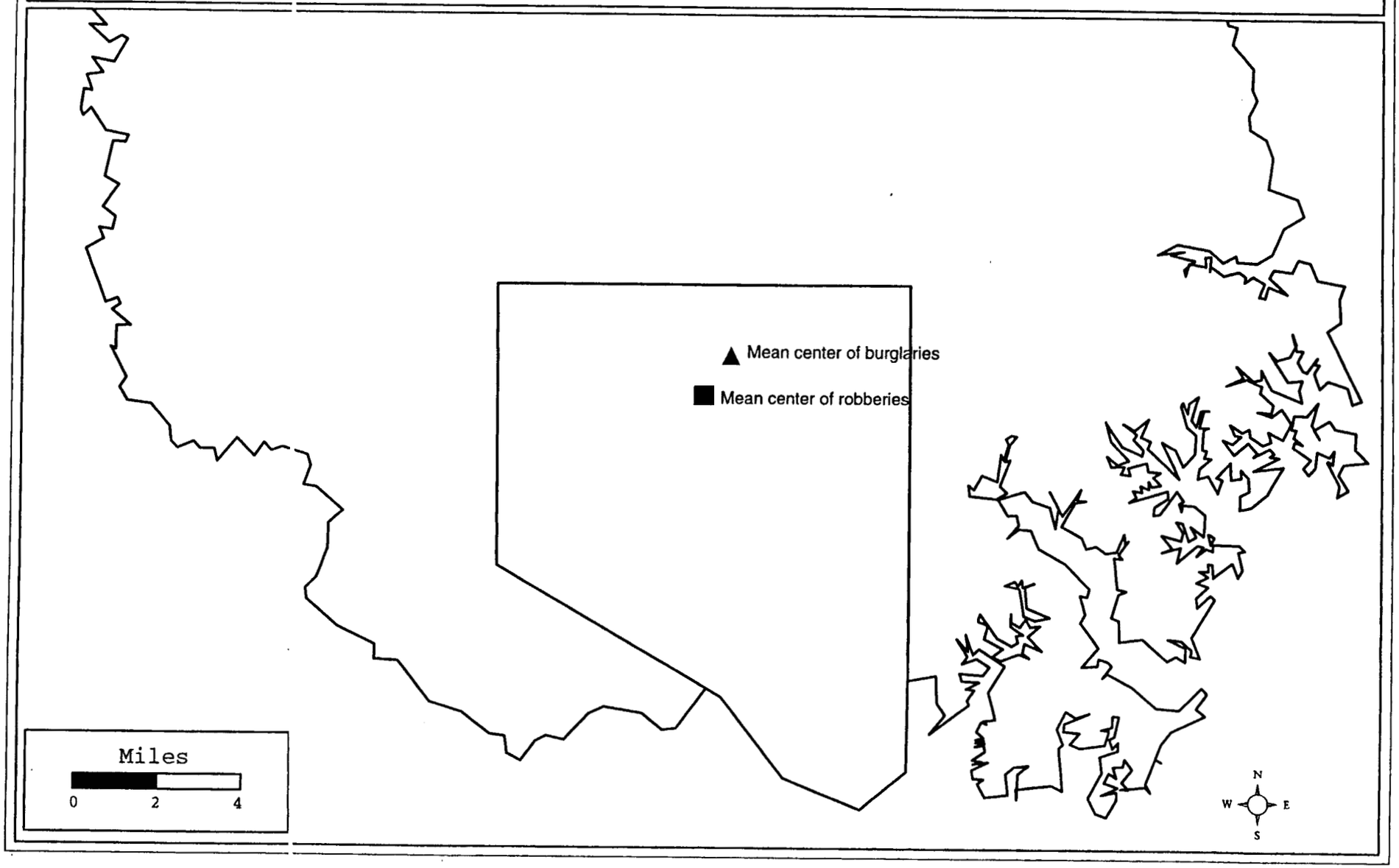
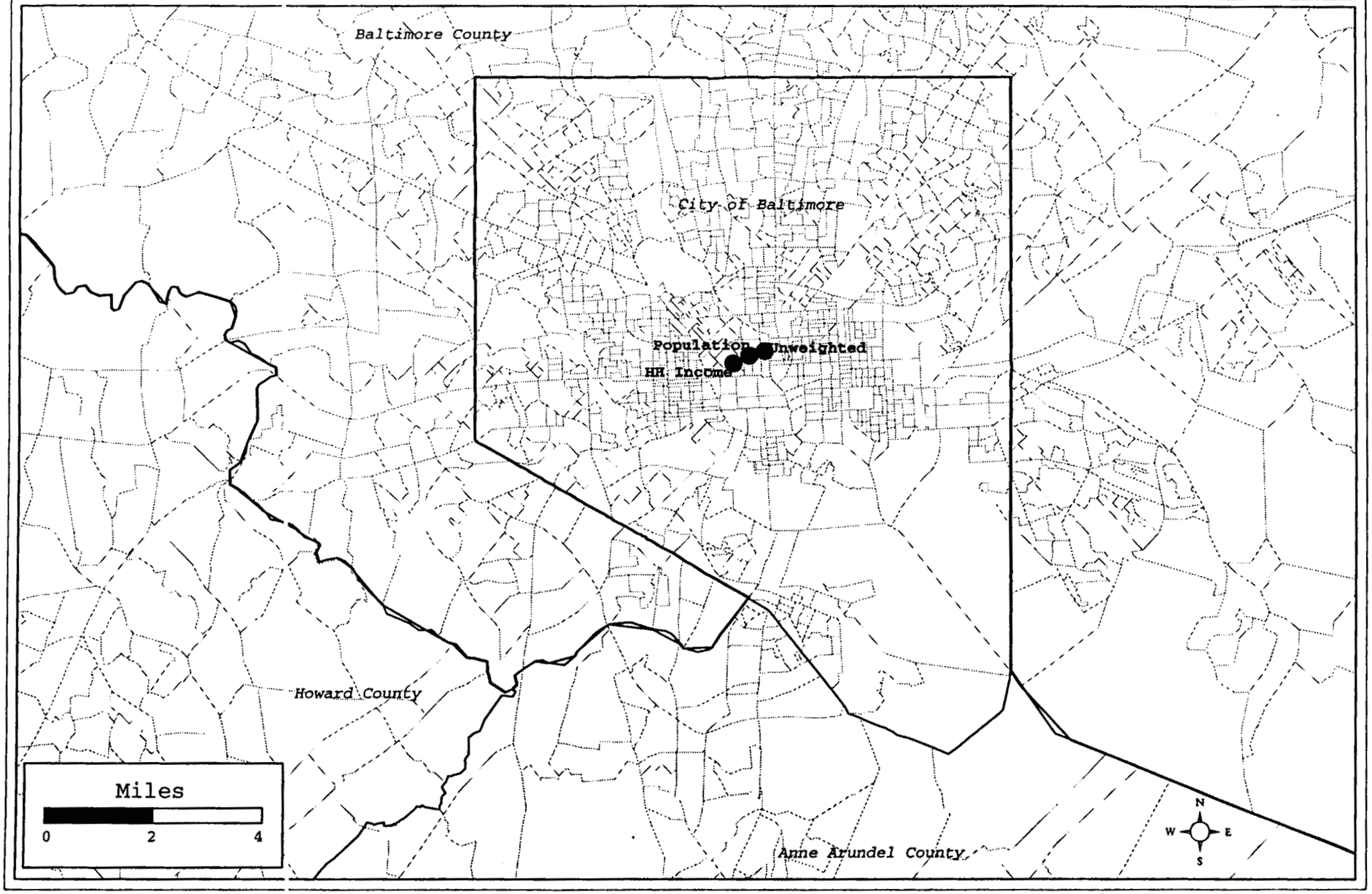


Figure 4.4: Center of Baltimore Metropolitan Population

Mean Center of Block Groups Weighted By Selected Variables



Weighted mean centers can be useful because they describe spatial differentiation in the metropolitan area and factors that may correlate with crime distributions. Another example is the weighted mean centers of different ethnic groups in the Baltimore metropolitan area (figure 4.5). The mean center of the White population is almost identical to the unweighted mean center. On the other hand, the mean center of the African-American/Black population is southwest of this and the mean center of the Hispanic/Latino population is considerably south of that for the White population. In other words, different ethnic groups tend to live in different parts of the Baltimore metropolitan area. Whether this has any impact on crime distributions is an empirical question. As we will see, there is not a simple spatial correlation between these weighted mean centers and particular crime distributions.

When the *Mcsd* box is checked, *CrimeStat* will run the routine. *CrimeStat* has a status bar that indicates how much of the routine has been run (Figure 4.6).[1] The results of these statistics are shown in the *Mcsd* output table (figure 4.7).

Center of Minimum Distance

Another centrographic statistic is the *center of minimum distance.* This is frequently called the *median center*, though it is not strictly a median. For a single variable, such as median household income, the median is that point at which 50% of the cases fall below and 50% fall above. On a two dimensional plane, however, there is not a single median because the location of a median is defined by the way that the axes are drawn. For example, in figure 4.8, there are eight incident points shown. Four lines have been drawn which divide these eight points into two groups of four each. However, the four lines do not identify an exact location for a median. Instead, there is an area of non-uniqueness in which any part of it could be considered the 'median center'. This violates one of the basic properties of a statistic is that it be a unique value.

The center of minimum distance is a unique statistic in that it defines the point at which the sum of the distance to all other points is the smallest (Burt and Barber, 1996). Sometimes called the *Euclidean median* or *center of minimum travel*, the center of minimum distance is defined as:

$$\text{Center of Minimum Distance} = C = \sum_{i=1}^{N} d_{ic} \text{ is a minimum} \qquad (4.3)$$

where d_{ic} is the distance between a single point, i, and C, the center of minimum distance (with an X and Y coordinate). Unfortunately, there is not a formula that can calculate this location. Instead, an iterative algorithm is used which approximates this location (Kuhn and Kuenne, 1962; Burt and Barber, 1996). Depending on whether the coordinates are spherical or projected, *CrimeStat* will calculate distance as either Great Circle (spherical) or Euclidean (projected), as discussed in the previous chapter.[2] The results are shown in the *Mcmd* output table (figure 4.9).

Figure 4.5: Center of Baltimore Metropolitan Population

Mean Center of Block Groups Weighted By Selected Variables

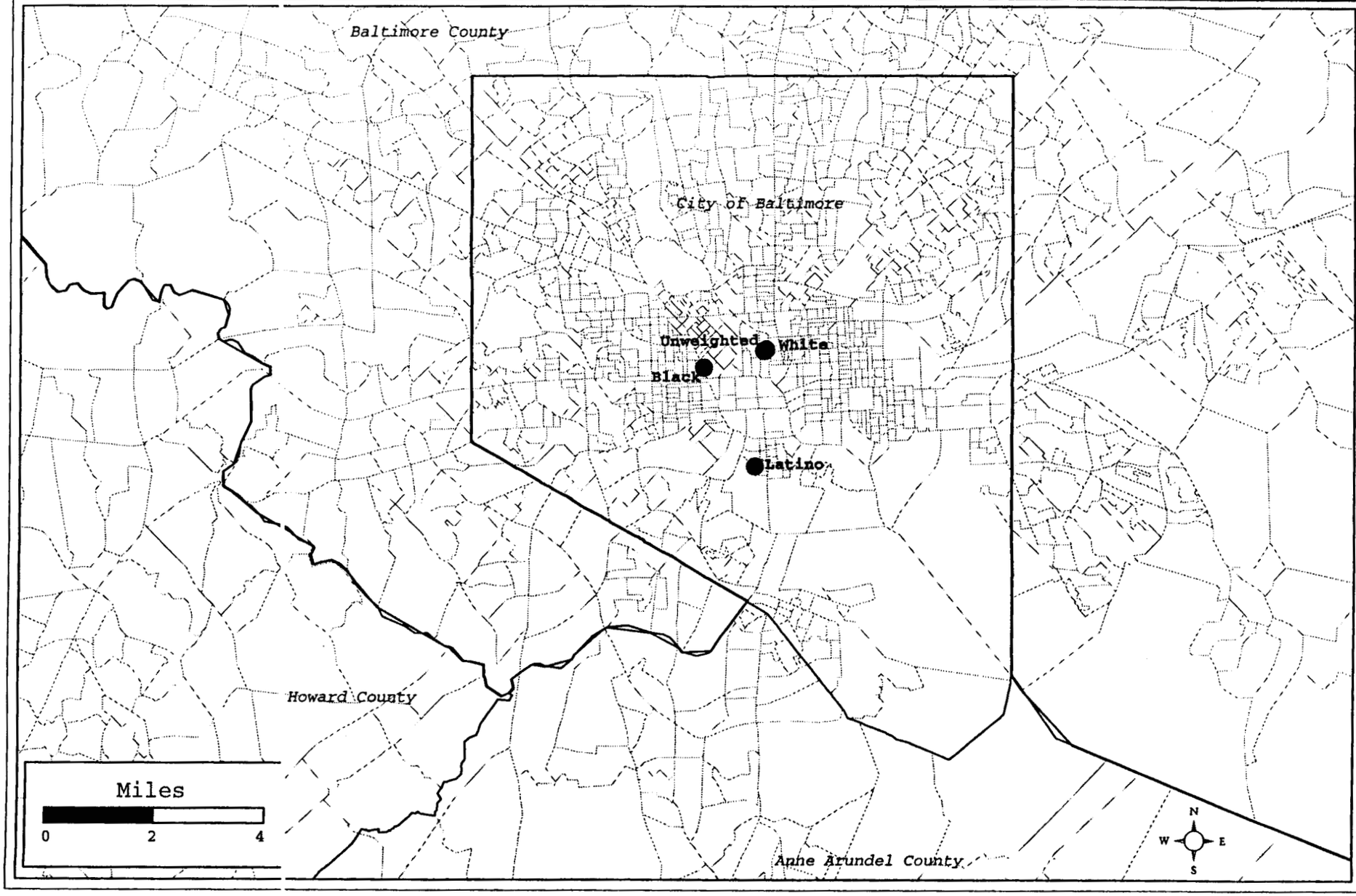


Figure 4.6: *CrimeStat* Calculating A Routine

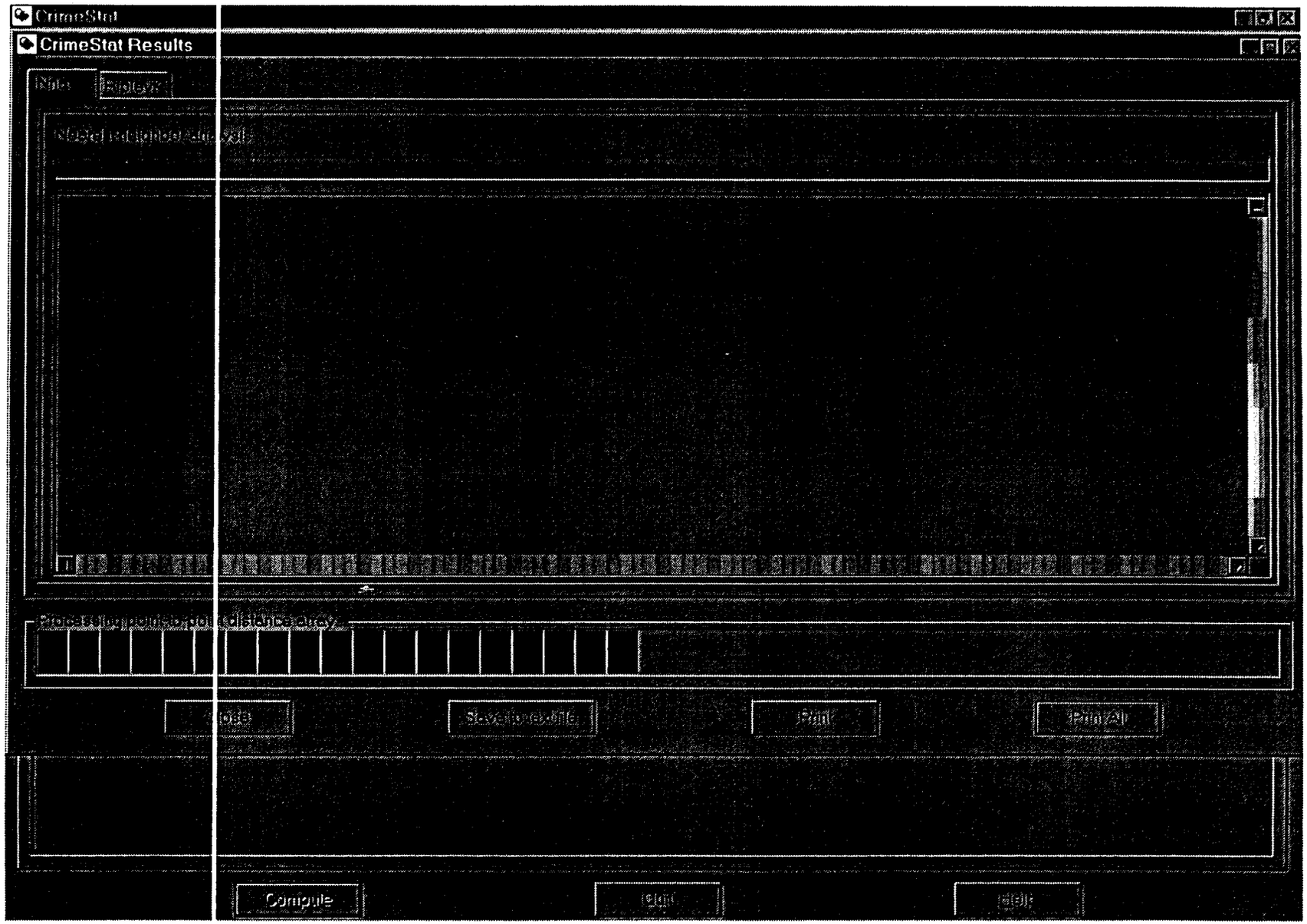


Figure 4.7: Mean Center and Standard Distance Deviation Output

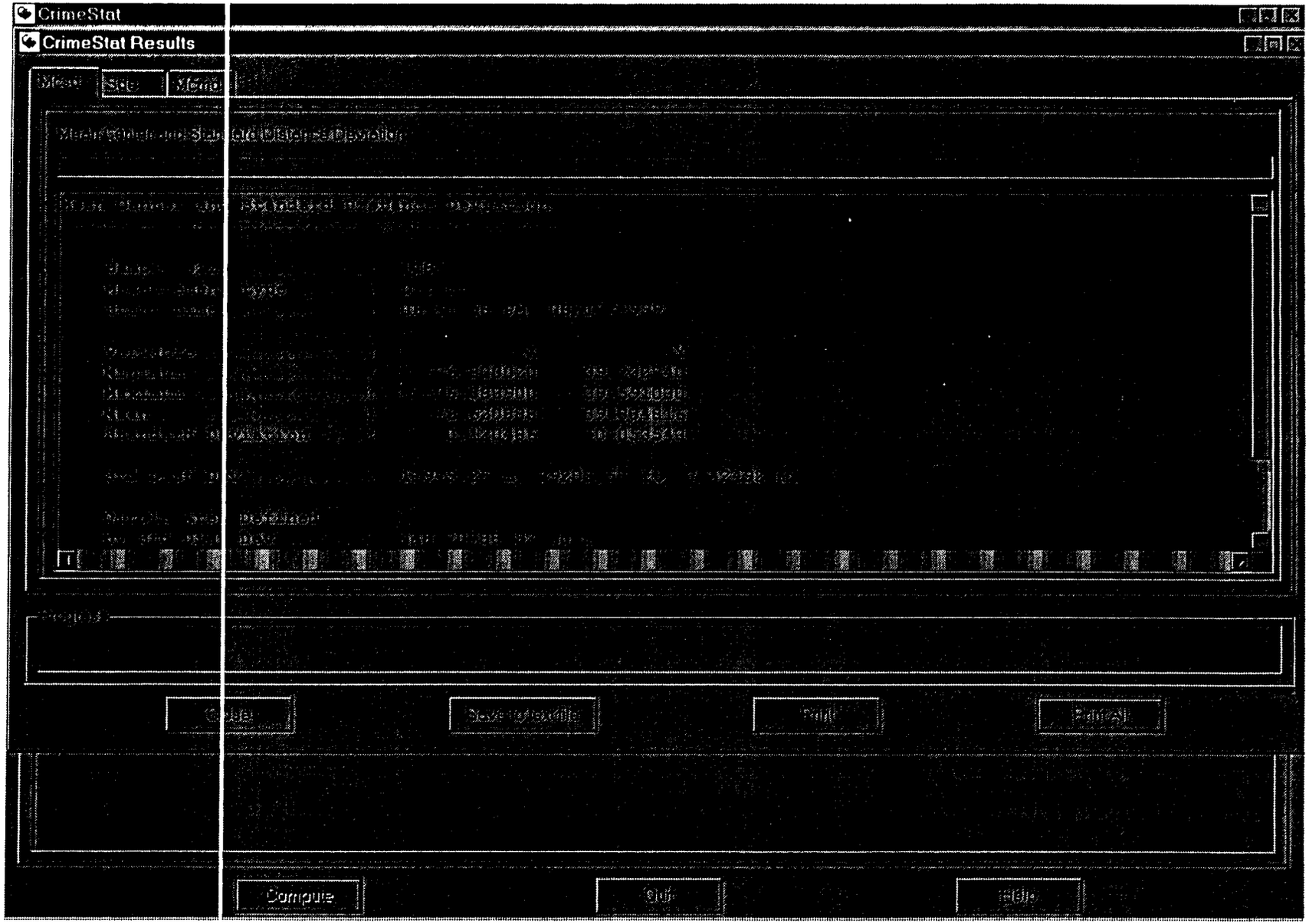


Figure 4.8: Non-Uniqueness of a Median Center

Lines Splitting Incident Locations Into Two Halves

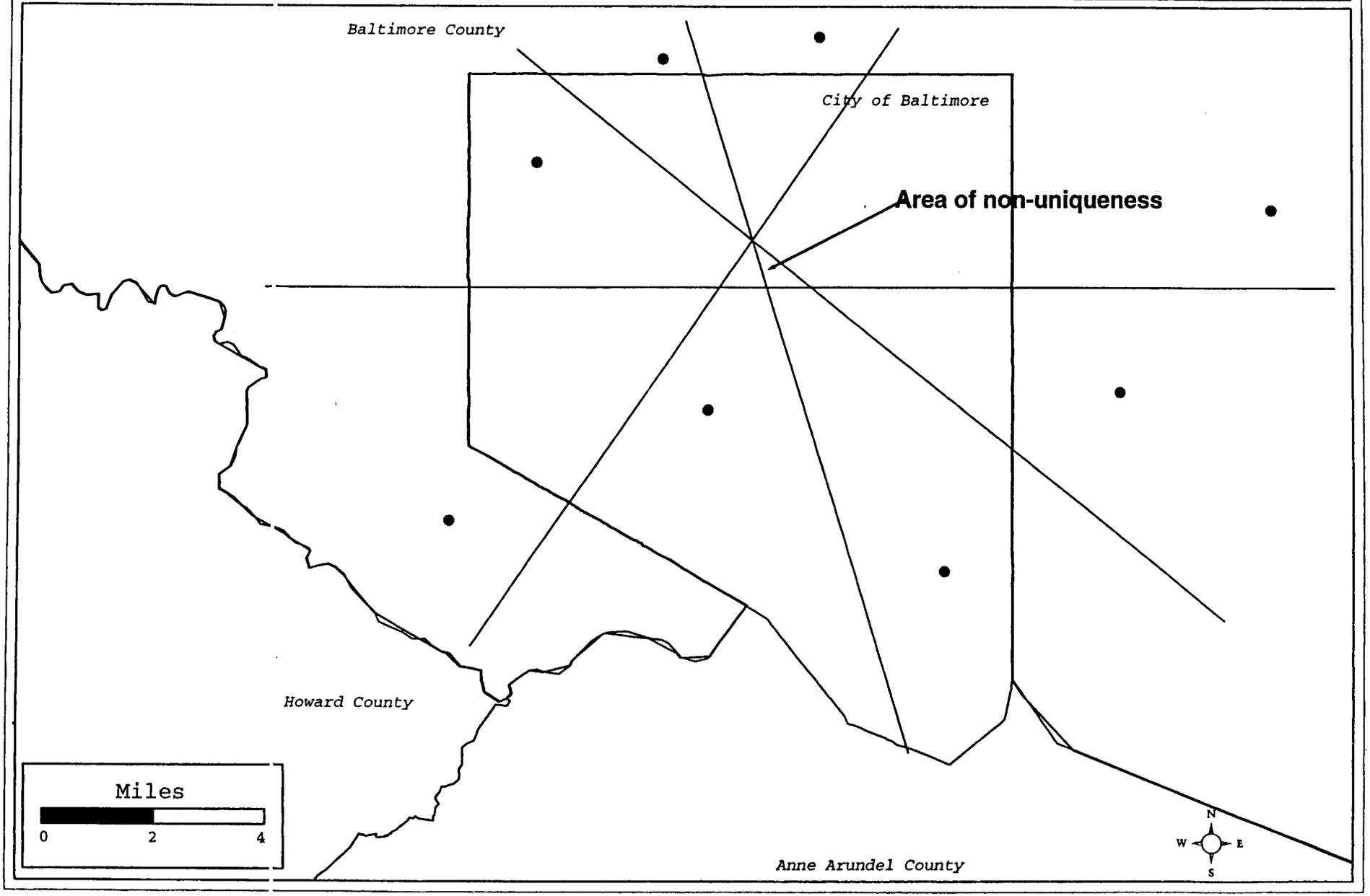
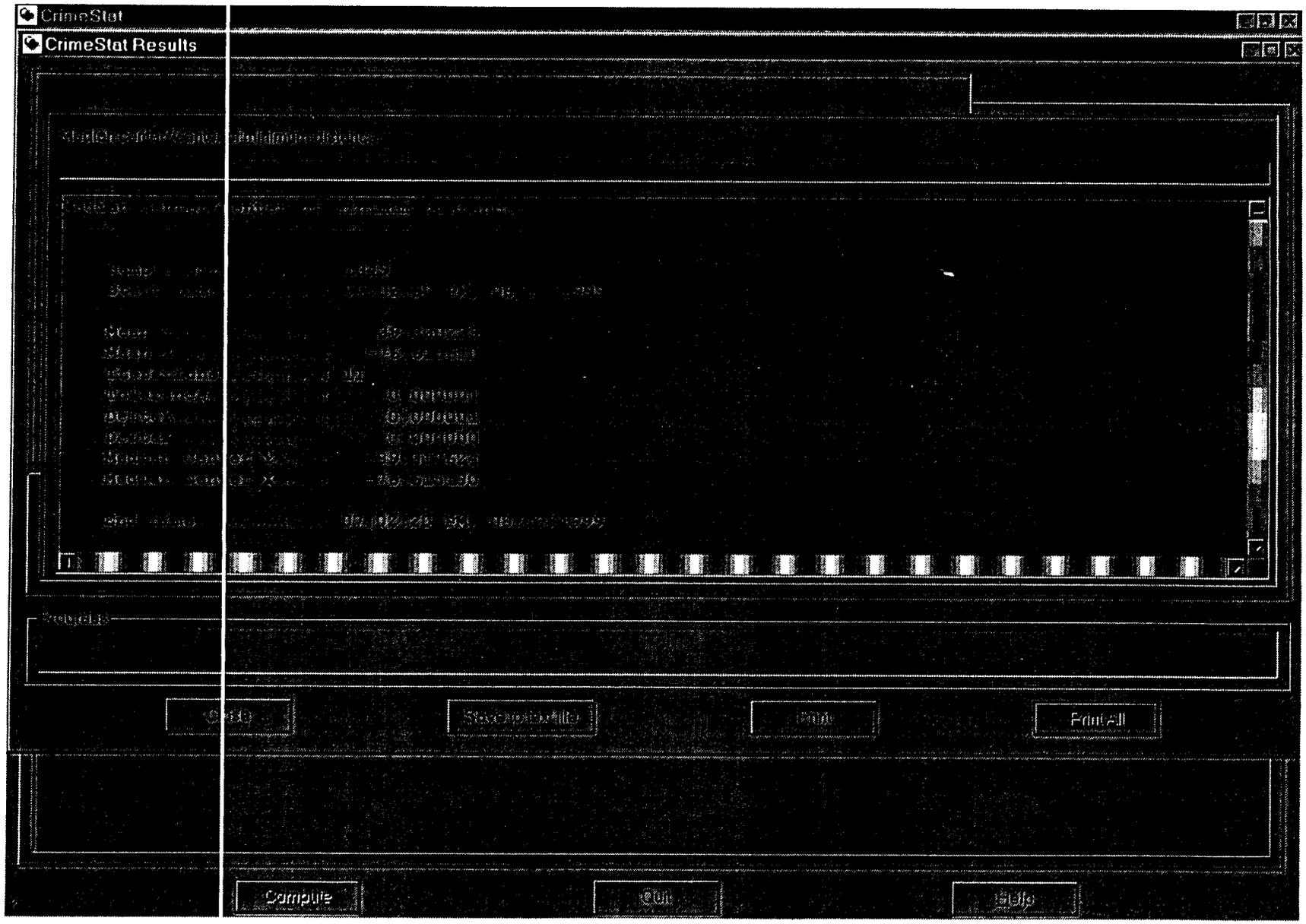


Figure 4.9: Center of Minimum Distance Output



The importance of the center of minimum distance is that it is a location where distance to all the defining incidents is the smallest. Since *CrimeStat* only measures distances as either direct or indirect, actual travel time is not being calculated. But in many jurisdictions, the minimum distance to all points is a good approximation to the point where travel distances are minimized. For example, in a police precinct, a patrol car could be stationed at the center of minimum distance to allow it to respond quickly to calls for service.

For example, figure 4.10 maps the center of minimum distance for auto thefts in both Baltimore City and Baltimore County and compares this to the mean center. As seen, the center of minimum distance is slightly south of the mean center, indicating that there are slightly more incidents in the southern part of the metropolitan area than in the northern part. However, the difference in these two statistics is very small.

Standard Deviation of the X and Y Coordinates

In addition to the mean center and center of minimum distance, *CrimeStat* will calculate various measures of spatial distribution, which describe the dispersion, orientation, and shape of the distribution of a variable (Hammond and McCullogh 1978; Ebdon 1988). The simplest of these is the raw standard deviations of the X and Y coordinates, respectively. The formulas used are the standard ones found in most elementary statistics books:

$$S_X = \text{SQRT}\left[\sum_{i=1}^{N} \frac{(X_i - \bar{X})^2}{N-1}\right] \quad (4.4)$$

$$S_Y = \text{SQRT}\left[\sum_{i=1}^{N} \frac{(Y_i - \bar{Y})^2}{N-1}\right] \quad (4.5)$$

where X_i and Y_i are the X and Y coordinates for individual points, $\bar{X}$ and $\bar{Y}$ are the mean X and mean Y, and N is the total number of points. Note that 1 is subtracted from the number of points to produce an unbiased estimate of the standard deviation.

The standard deviations of the X and Y coordinates indicate the degree of dispersion. Figure 4.11 shows the standard deviation of the coordinates for auto thefts and represents this as a rectangle. As seen, the distribution of auto thefts spreads more in an east-west direction than in a north-south direction.

Standard Distance Deviation

While the standard deviation of the X and Y coordinates provides some information about the dispersion of the incidents, there are two problems with it. First, it does not

Figure 4.10: 1996 Metropolitan Baltimore Auto Thefts

Mean Center and Center of Minimum Distance for 1996 Auto Thefts

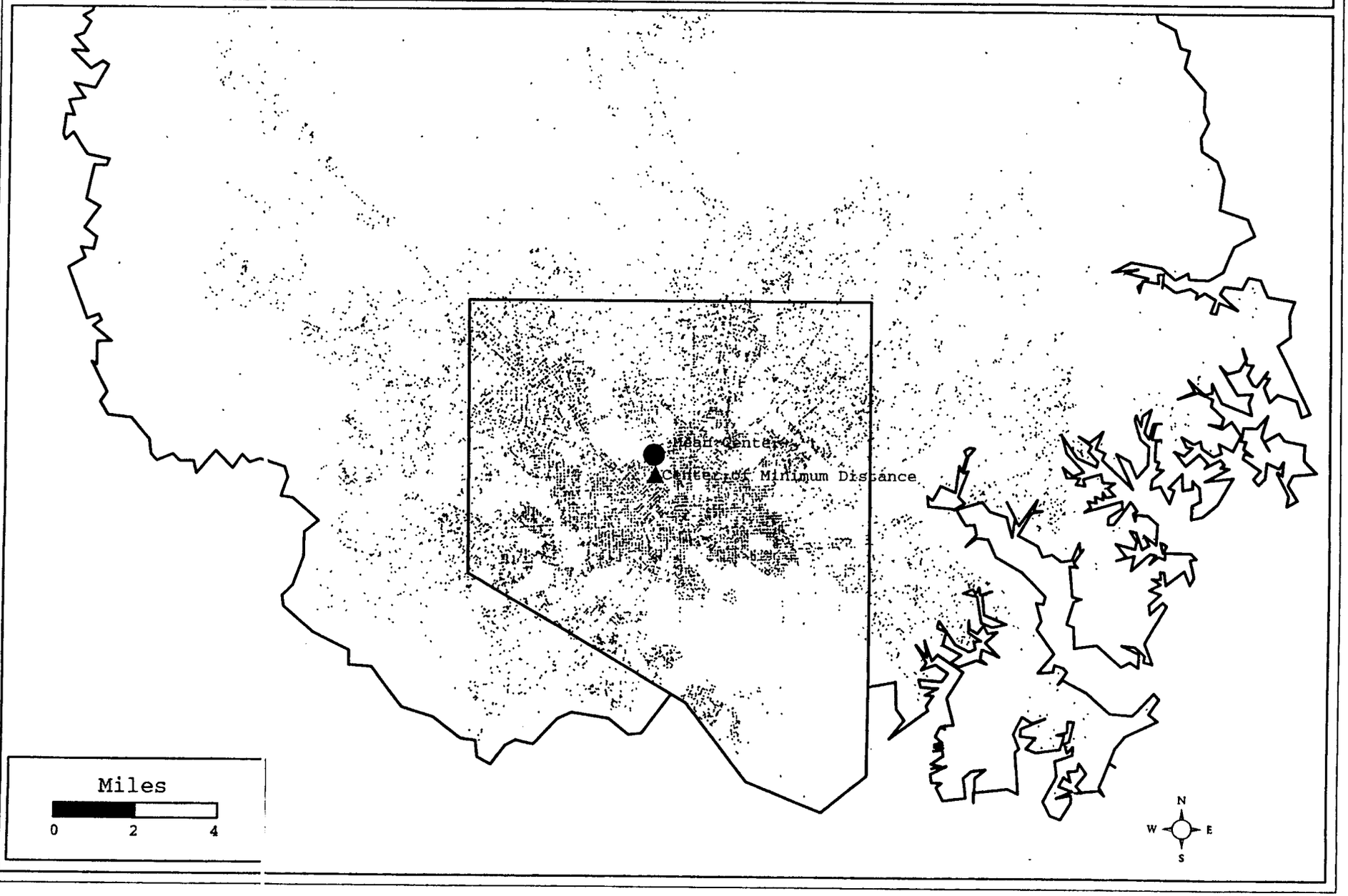
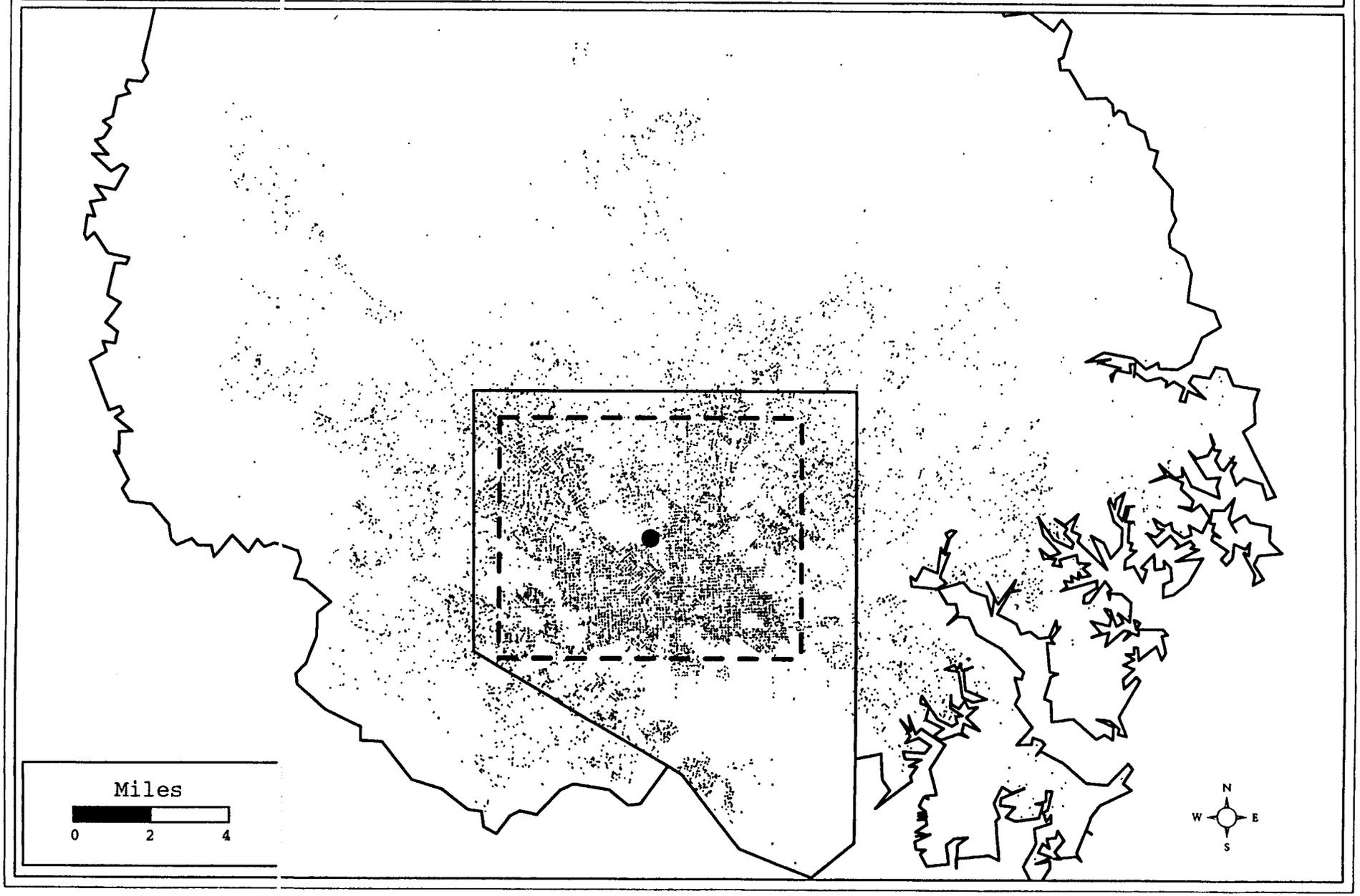


Figure 4.11: 1996 Metropolitan Baltimore Auto Thefts

Mean Center and Standard Deviations of X and Y Coordinates



provide a single summary statistic of the dispersion in the incident locations and is actually two separate statistics (i.e., dispersion in X and dispersion in Y). Second, it provides measurements in the units of the coordinate system. Thus, if spherical coordinates are being used, then the units will be decimal degrees.

A measure which overcomes these problems is the *standard distance deviation* or *standard distance*, for short. This is the standard deviation of the distance of each point from the mean center and is expressed in measurement units (feet, meters, miles). It is the two-dimensional equivalent of a standard deviation.

The formula for it is

$$S_{XY} = \sum_{i=1}^{N} \frac{(d_{iMC})^2}{N-2}] \qquad (4.6)$$

where d_{iMC} is the distance between each point, i, and the mean center and N is the total number of points. Note that 2 is subtracted from the number of points to produce an unbiased estimate of standard distance since there are two constants from which this distance is measured (mean of X, mean of Y).

Since the standard distance is an average distance (after correction for bias) from the mean center, it can be represented as a single vector rather than two vectors as with the standard deviation of the X and Y coordinates. Figure 4.12 shows the standard distance deviations of robberies and burglaries for 1996 in Baltimore County represented as circles.

Standard Deviational Ellipse

The standard distance deviation is a good single measure of the dispersion of the incidents around the mean center. However, with two dimensions, distributions are frequently skewed in one direction or another (a condition called *anisotropy*). Instead, there is another statistic which gives dispersion in two dimensions, the *standard deviational ellipse* or *ellipse*, for short (Ebdon, 1988; Cromley, 1992).

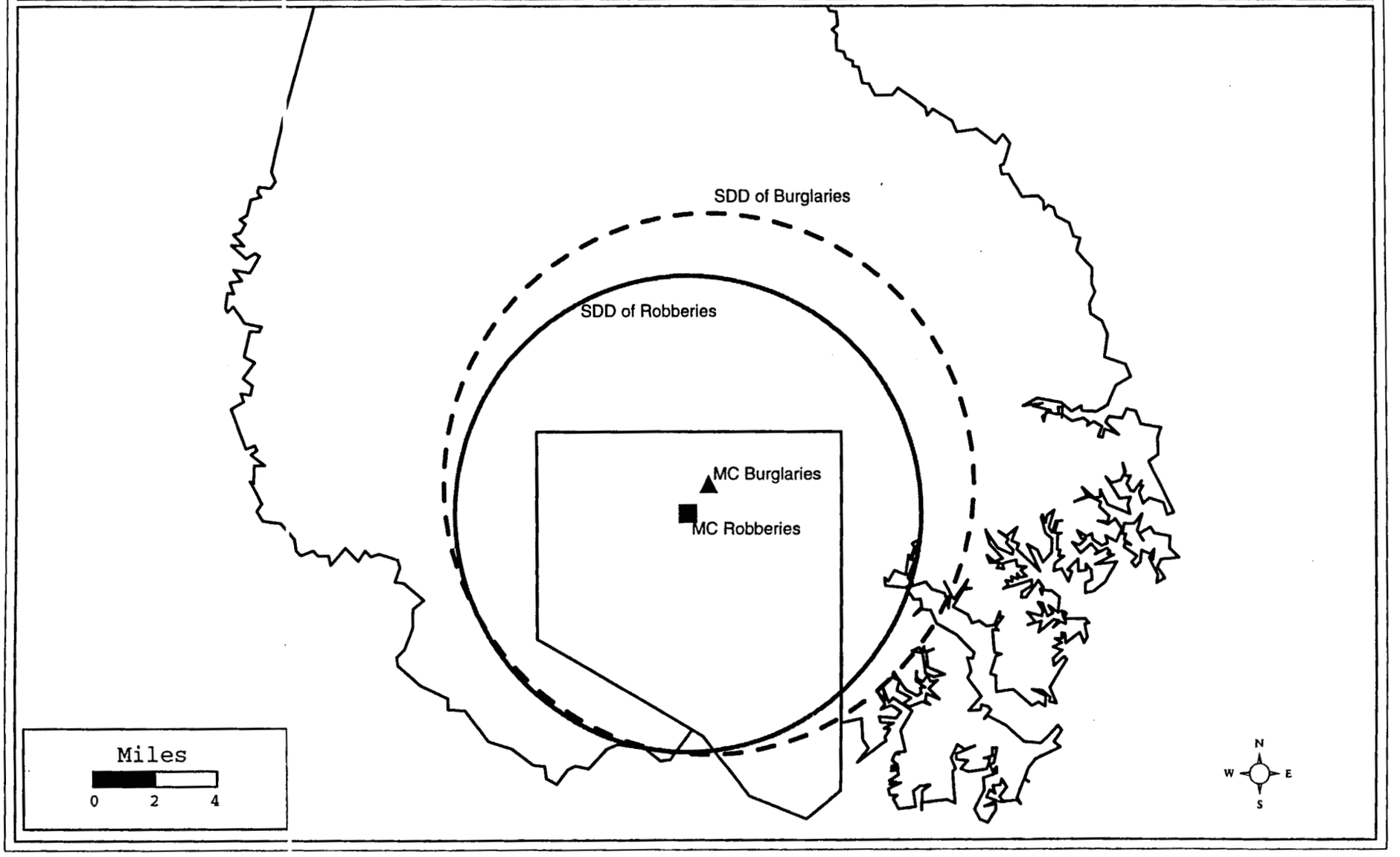
The standard deviational ellipse is derived from the bivariate distribution (Furfey, 1927; Neft, 1962; Bachhi, 1957) and is defined by

$$\text{Bivariate Distribution} = \text{SQRT} \frac{[\sigma^2_x + \sigma^2_y]}{2} \qquad (4.7)$$

The two standard deviations, in the X and Y directions, are orthogonal to each other and define an ellipse. Ebdon (1988) rotates the X and Y axis so that the sum of squares of distances between points and the axes are minimized. By convention, it is shown as an ellipse.

Figure 4.12: 1996 Baltimore County Burglaries and Robberies

Comparison of Mean Centers and Standard Distance Deviations



Aside from the mean X and mean Y, the formulas for these statistics are as follows:

1. The Y-axis is rotated *clockwise* through an angle, θ, where

$$\theta = \text{ARCTAN}\{(\Sigma(X_i-\bar{X})^2 - \Sigma(Y_i-\bar{Y})^2) + [(\Sigma(X_i-\bar{X})^2 - \Sigma(Y_i-\bar{Y})^2)^2 + 4(\Sigma(X_i-\bar{X})(Y_i-\bar{Y}))^2]^{1/2}\}/(2\Sigma(X_i-\bar{X})(Y_i-\bar{Y})) \quad (4.8)$$

where all summations are for i=1 to N (Ebdon, 1988).

2. Two standard deviations are calculated, one along the transposed X-axis and one along the transposed Y-axis.

$$S_x = \text{SQRT}(2)\{\sum_{i=1}^{N}[(X_i - \bar{X})\text{Cos}\theta - (Y_i - \bar{Y})\text{Sin}\theta]^2/(N-2)\}^{1/2} \quad (4.9a)$$

$$S_y = \text{SQRT}(2)\{\sum_{i=1}^{N}[(X_i - \bar{X})\text{Sin}\theta - (Y_i - \bar{Y})\text{Cos}\theta]^2/(N-2)\}^{1/2} \quad (4.9b)$$

where N is the number of points. Note, again, that 2 is subtracted from the number of points in both denominators to produce an unbiased estimate of the standard deviational ellipse since there are two constants from which the distance along each axis is measured (mean of X, mean of Y).[3]

3. The X-axis and Y-axis of the ellipse are defined by

$$\text{Length}_x = 2S_x \quad (4.10a)$$

$$\text{Length}_y = 2S_y \quad (4.10b)$$

4. The area of the ellipse is

$$A = \pi S_x S_y \quad (4.11)$$

Figure 4.13 shows the output of the ellipse routine and figure 4.14 maps the standard deviational ellipse of auto thefts in Baltimore City and Baltimore County for 1996.

Table Outputs

For each of these statistics, *CrimeStat* produces tabular output. In *CrimeStat*, all tables are labeled by symbols, for example Mcsd for the mean center and standard distance deviation or Mcmd for the mean center/center of minimum distance. All tables present the sample size. In the Mcsd output table, there are minimum and maximum X and Y values

Figure 4.13: Standard Deviational Ellipse Output

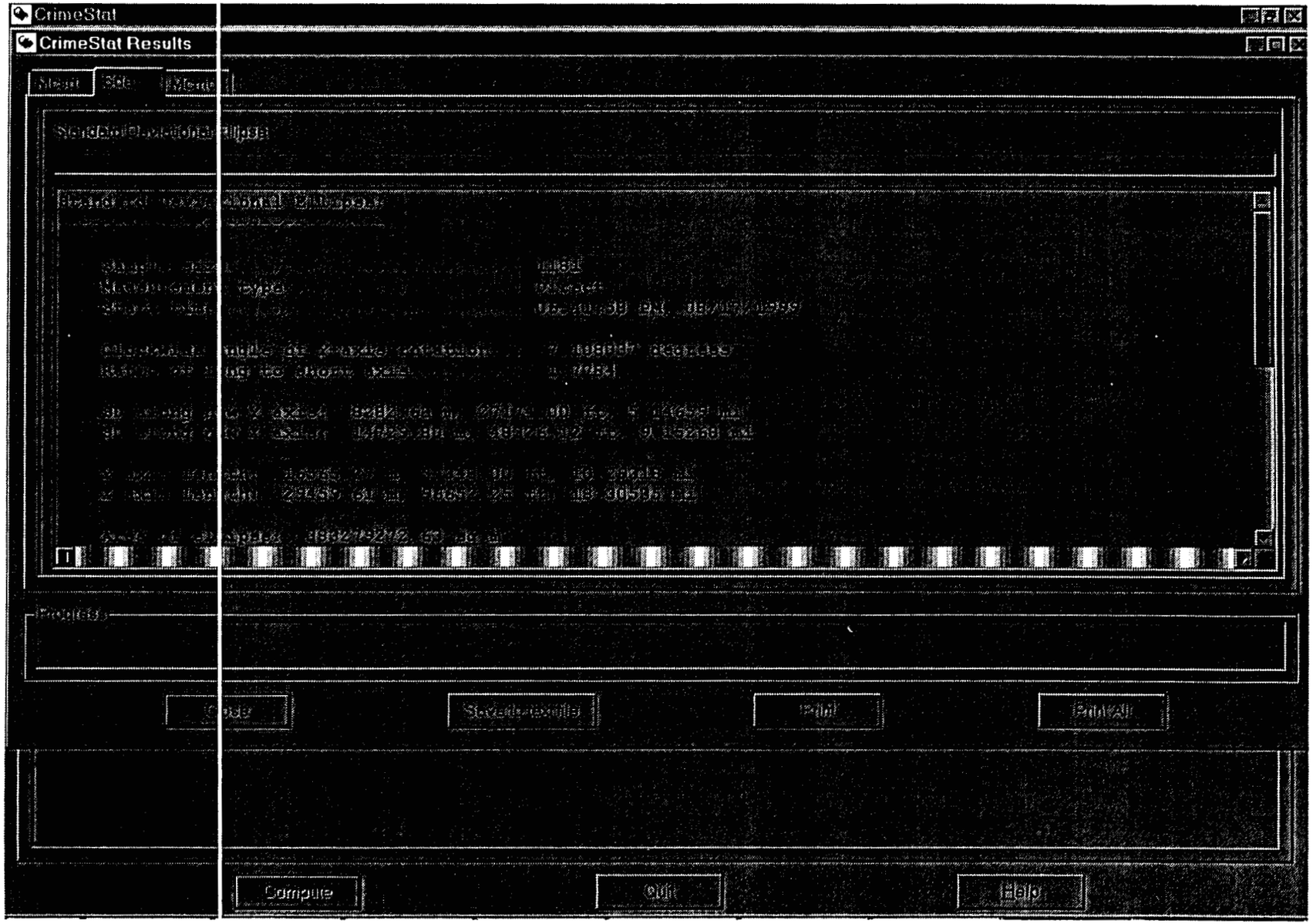
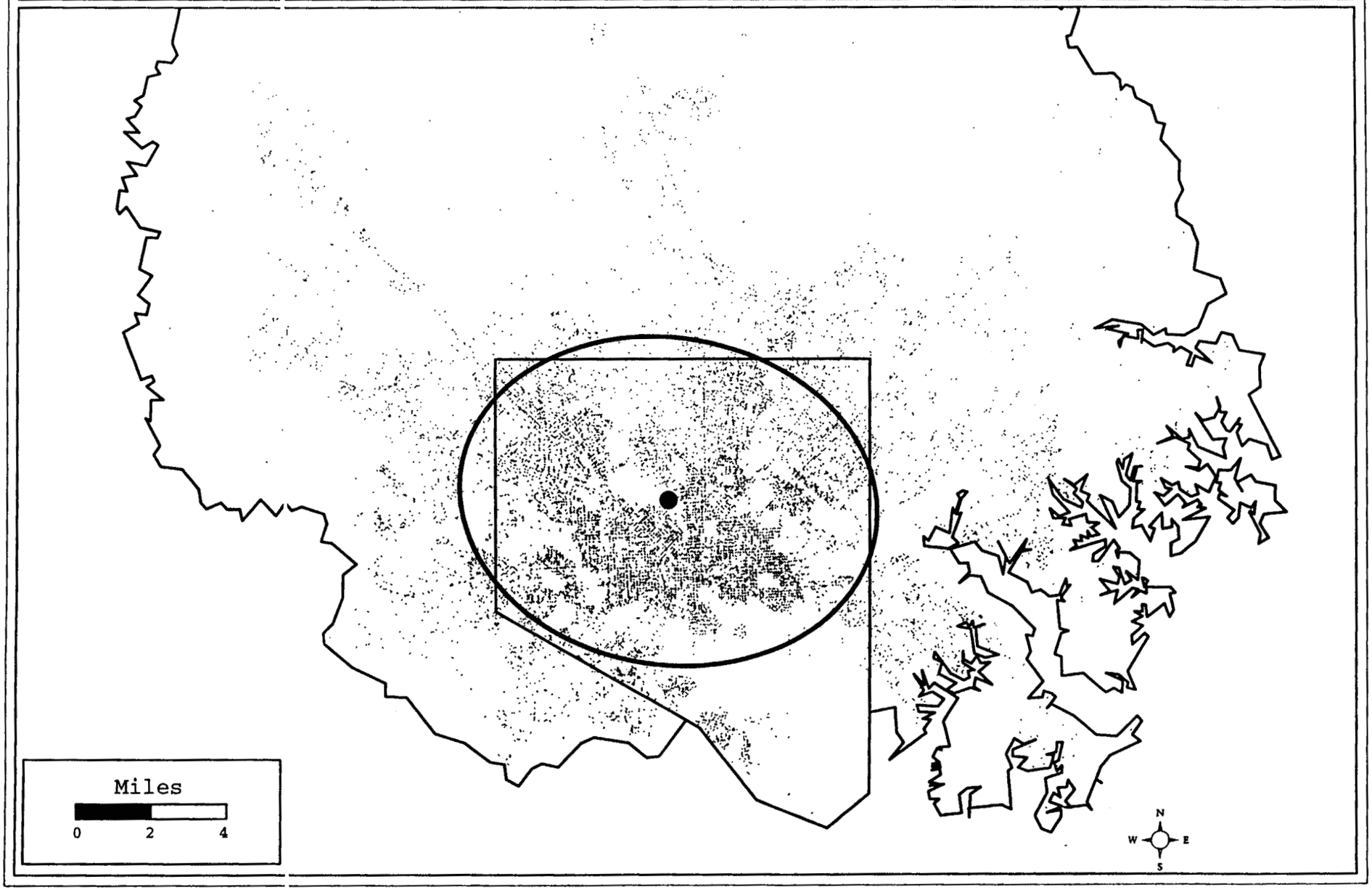


Figure 4.14: 1996 Metropolitan Baltimore Auto Thefts

Mean Center and Standard Deviational Ellipse



respectively, the mean X, the mean Y, the standard deviations of the X and Y values respectively, the standard distance deviation (in meters, feet and miles), and the area of a circle defined by the standard distance deviation (in square meters, square feet and square miles). In the *Mcmd* output table, there are the mean X, mean Y, median center X, median center Y, and the number of iterations. In the *Sde* output table (the standard deviational ellipse), there is clockwise angle of rotation, the standard deviation along the new Y axis (in meters, feet, and miles), the standard deviation along the new X axis (also in meters, feet, and miles), the ratio of the long to short axis, and the area of the ellipse defined by the transformed axes (in square meters, square feet, and square miles).

Selecting Output Objects

The five centrographic statistics can be output as graphical objects to three GIS packages. The mean center and center of minimum distance are output as single points. The standard deviation of the X and Y coordinates is output as a rectangle. The standard distance deviation is output as a circle and the standard deviational ellipse is output as an ellipse.

CrimeStat currently supports graphical outputs to *ArcView* '.shp' files, to *MapInfo* '.mif' and to *Atlas*GIS* '.bna' files. Before running the calculation, the user should select the desired output files and specify a root name (e.g., Precinct1Burglaries). Figure 4.15 shows a dialog box for selecting for the GIS program output. For *MapInfo* output only, the user has to also indicate the name of the projection, the projection number and the datum number. These can be found in the *MapInfo* users guide. By default, *CrimeStat* will use the standard parameters for a spherical coordinate system (Earth projection, projection number 1, and datum number 33). If a user requires a different coordinate system, the appropriate values should be typed into the space. Figure 4.16 shows the selection of the *MapInfo* coordinate parameters.

Calculating the Statistics

Once the statistics have been selected, the user clicks on *Compute* to run the routine. The results are shown in a results table.

Output Files

If requested, the output files are saved in the specified directory under the specified (root) name. For each statistic, *CrimeStat* will add prefix letters to the root name.

MC<*root*> for the mean center
MDN<*root*> for center of minimum distance
XYD<*root*> for the standard deviation of the X and Y coordinates
SDD<*root*> for the standard distance deviation
SDE<*root*> for the standard deviational ellipse.

Figure 4.15: Outputting Objects to A GIS Program

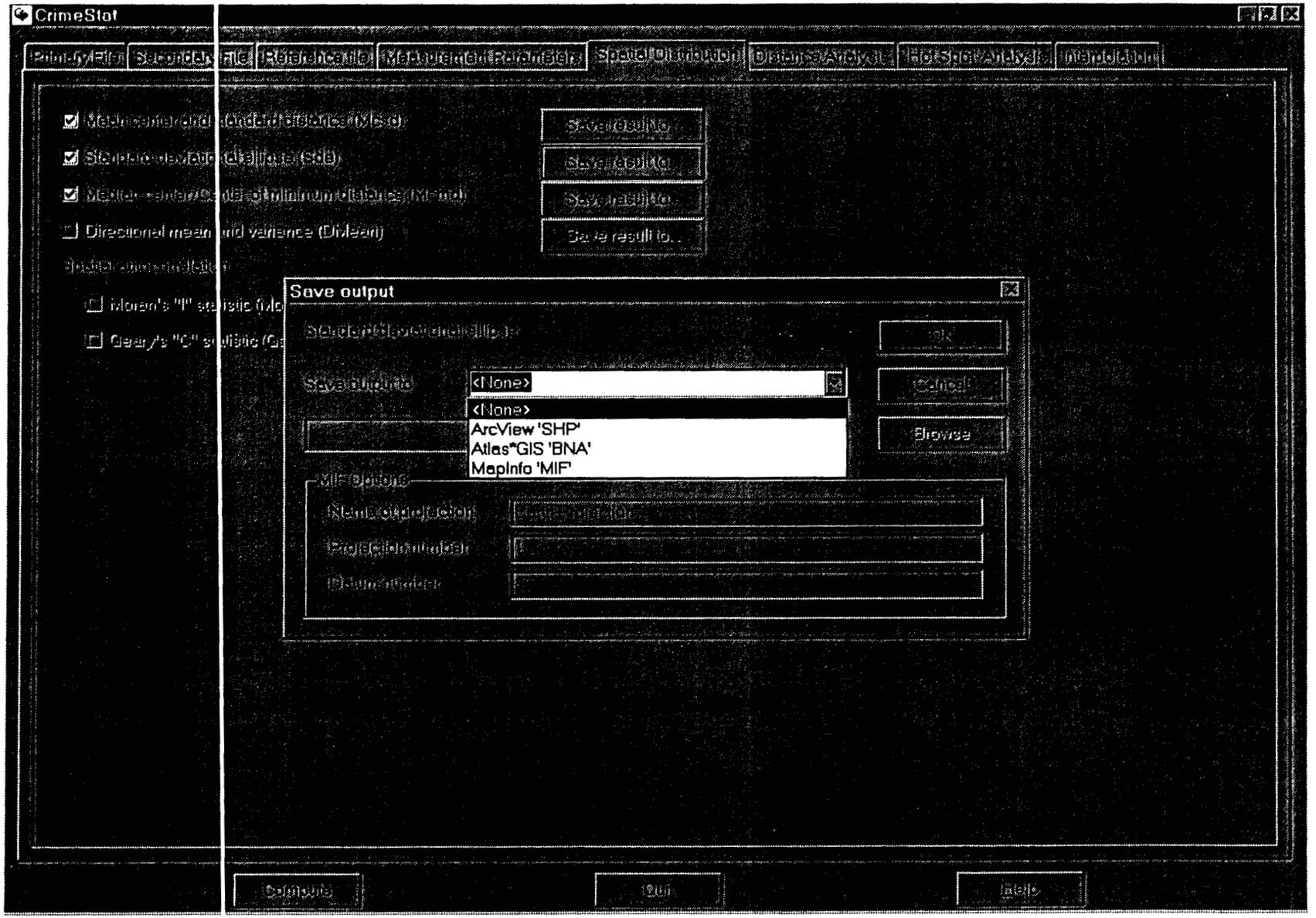
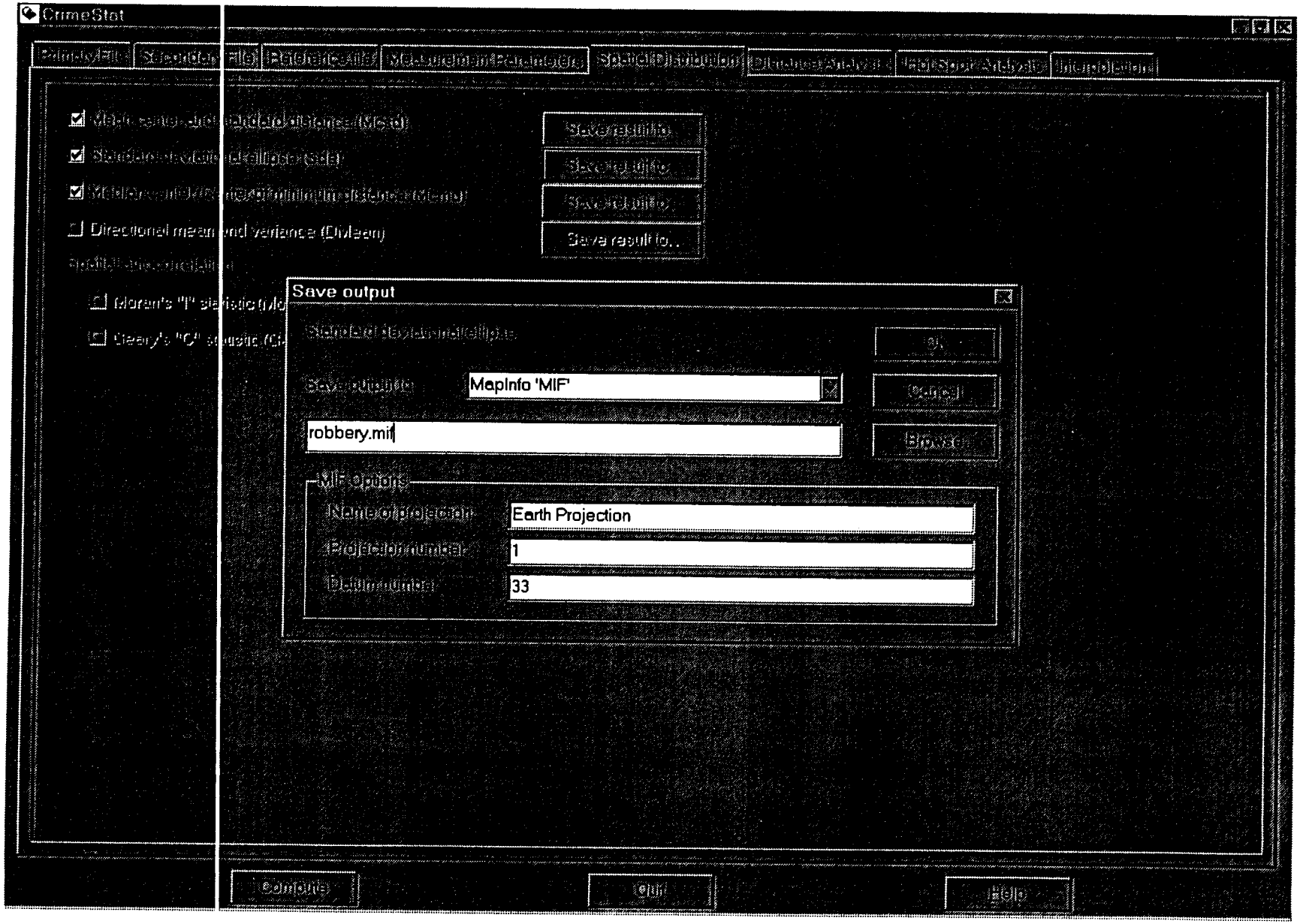


Figure 4.16: *MapInfo* Output Options



The '.shp' files can be read directly into *ArcView* as themes. The '.mif' and '.bna' files have to be imported into *MapInfo* and *Atlas*GIS,* respectively.[4]

Statistical Tests

While the current version of *CrimeStat* does not conduct statistical tests for the centrographic statistics, it is possible to conduct such tests between two groups, *A* and *B*.

Differences in the Mean Centers of Two Samples

For differences between two sample in the mean center, it is necessary to test both differences in the X coordinate and differences in the Y coordinates. Since *CrimeStat* outputs both the mean X, mean Y, standard deviation of X, and standard deviation of Y, a simple t-test can be set up. The null hypothesis is that the mean centers are equal

$$H_0: \quad \mu_{XA} = \mu_{XB}$$
$$\mu_{YA} = \mu_{YB}$$

and the alternative hypothesis is that the mean centers are not equal

$$H_1: \quad \mu_{XA} \neq \mu_{XB}$$
$$\mu_{YA} \neq \mu_{YB}$$

Because the true standard deviations of sample A, σ_{XA} and σ_{YA}, and sample B, σ_{XB} and σ_{YB}, are not known, the sample standard deviations are taken, S_{XA}, S_{YA}, S_{XB} and S_{YB}. However, since there are two different variables being tested (mean of X and mean of Y for groups 1 and 2), the alternative hypothesis has two fundamentally different interpretations:

Comparison I: That EITHER $\mu_{XA} \neq \mu_{XB}$ OR $\mu_{YA} \neq \mu_{YB}$ is true

Comparison II: That BOTH $\mu_{XA} \neq \mu_{XB}$ AND $\mu_{YA} \neq \mu_{YB}$ are true

In the first case, the mean centers will be considered not being equal if either the mean of X or the mean of Y are significantly different. In the second case, both the mean of X and the mean of Y must be significantly different for the mean centers to be considered not equal. The first case is clearly easier to fulfill than the second.

Significance levels

By tradition, significance tests for comparisons between two means are made at the $\alpha \leq .05$ or $\alpha \leq .01$ levels, though there is nothing absolute about those levels. The significance levels are selected to minimize *Type 1 Errors*, inadvertently declaring a difference in the means when in reality there is not a difference. Thus, a test establishes that the likelihood of falsely rejecting the null hypothesis be less than one-in-twenty (less strict) or one-in-one

hundred (more strict). However, with multiple comparisons, the chances increase for finding 'significance' due to the multiple tests. For example, with two tests - a difference in the means of the X coordinate and a difference in the means of the Y coordinate, the likelihood of rejecting the first null hypothesis ($\mu_{XA} \neq \mu_{XB}$) is one-in-twenty and the likelihood of rejecting the second null hypothesis ($\mu_{YA} \neq \mu_{YB}$) is also one-in-twenty, then the likelihood of rejecting either one null hypothesis or the other is actually one-in-ten.

To handle this situation, comparison I - the 'either/or' condition, a Bonferoni test is appropriate (Anselin, 1995; Systat, 1996). Because the likelihood of achieving a given significance level increases with multiple tests, a 'penalty' must be assigned in finding either the differences in means for the X coordinate or differences in means for the Y coordinates significant. The Bonferoni criteria divides the critical probability level by the number of tests. Thus, if the $\alpha \leq .05$ level is taken for rejecting the null hypothesis, the critical probability for each mean must be .025 (.05/2); that is, differences in either the mean of X or mean of Y between two groups must yield a significance level less than .025.

For comparison II - the 'both/and' condition, on the other hand, the test is more stringent since the differences between the means of X and the means of Y must both be significant. Following the logic of the Bonferoni criteria, the critical probability level is multiplied by the number of tests. Thus, if the $\alpha = .05$ level is taken for rejecting the null hypothesis, then both tests must be significant at the $\alpha \leq .10$ level (i.e., .05*2).[5]

Tests

The statistics used are the usual ones for the t-test of the difference between means (Kanji, 1993).

A. First, test for equality of variances by taking the ratio of the variances (squared sample standard deviations) of both the X and Y coordinates:

$$F_X = \frac{S_{XA}^2}{S_{XB}^2} \qquad (4.12a)$$

$$F_Y = \frac{S_{YA}^2}{S_{YB}^2} \qquad (4.12b)$$

with $(N_A - 1)$ and $(N_B - 1)$ degrees of freedom for groups *A* and *B* respectively. This test is usually done with the larger of the variances in the numerator. Since there are two variances being compared (for X and Y, respectively), the logic should follow either *I* or *II* above (i.e., if either are to be true, then the critical α will be actually $\alpha/2$ for each; if both must be true, then the critical α will be actually $2*\alpha$ for each).

B. Second, if the variances are considered equal, then a t-test for two group means with unknown, but equal, variances can be used (Kanji, 1993; 28). Let

$$S_{XAB} = \text{SQRT}\left[\frac{\sum_{i=1}^{N(A)} (X_{Ai} - \bar{X}_A)^2 + \sum_{i=1}^{N(B)} (X_{Bi} - \bar{X}_B)^2}{(N_A + N_B - 2)}\right] \quad (4.13a)$$

$$S_{YAB} = \text{SQRT}\left[\frac{\sum_{i=1}^{N(A)} (Y_{Ai} - \bar{Y}_A)^2 + \sum_{i=1}^{N(B)} (Y_{Bi} - \bar{Y}_B)^2}{(N_A + N_B - 2)}\right] \quad (4.13b)$$

where the summations are for i=1 to N within each group separately. Then the test becomes

$$t_X = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_{XA} - \mu_{XB})}{S_{XAB} * \text{SQRT}\left[\frac{1}{N_A} + \frac{1}{N_B}\right]} \quad (4.14a)$$

$$t_Y = \frac{(\bar{Y}_A - \bar{Y}_B) - (\mu_{YA} - \mu_{YB})}{S_{YAB} * \text{SQRT}\left[\frac{1}{N_A} + \frac{1}{N_B}\right]} \quad (4.14b)$$

with $(N_A + N_B - 2)$ degrees of freedom for each test. (4.14c)

C. Third, if the variances are not considered equal, then a t-test for two group means with unknown and unequal variances should be used (Kanji, 1993; 29).

$$t_X = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_{XA} - \mu_{XB})}{\text{SQRT}\left\{\left[\frac{S_{XA}^2}{N_A} + \frac{S_{XB}^2}{N_B}\right]\right\}} \quad (4.15a)$$

$$t_Y = \frac{(\bar{Y}_A - \bar{Y}_B) - (\mu_{YA} - \mu_{YB})}{\text{SQRT}\left\{ \left[\frac{S_{YA}^2}{N_A} + \frac{S_{YB}^2}{N_B} \right] \right\}} \qquad (4.15b)$$

with degrees of freedom

$$v = \left\{ \frac{\left[\frac{S_A^2}{N_A} + \frac{S_B^2}{N_B} \right]}{\left[\frac{S_A^4}{N_A^2(N_A + 1)} + \frac{S_B^4}{N_B^2(N_B + 1)} \right]} \right\} - 2 \qquad (4.15c)$$

for both the X and Y test. Even though this latter formula is cumbersome, in practice, if the sample size of each group is greater than 100, then the t-values for infinity can be taken as a reasonable approximation and the above degrees of freedom need not be tested (t=1.645 for α=.05; t=1.960 for α=.01).

D. The significance levels are those selected above. For comparison I - that either differences in the means of X or differences in the means of Y are significant, the critical probability level is $\alpha/2$ (e.g., .05/2 = .025; .01/2 = .005). For comparison II - that both differences in the means of X and differences in the means of Y are significant, the critical probability level is $\alpha*2$ (e.g., .05*2 = .10; .01*2 - .02).

E. Reject the null hypothesis if:

Comparison I:	Either tested t-value (t_x or t_y) is greater than the Critical t for $\alpha/2$
Comparison II:	Both tested t-values (t_x and t_y) are greater than the critical t for $\alpha*2$

Example 1: Burglaries and robberies in Baltimore County

To illustrate, compare the distribution of burglaries in Baltimore County with those of robberies, both for 1996. Figure 4.17 shows the distribution of all burglaries in Baltimore County with the location of the mean center (triangle) while figure 4.18 shows the distribution of all robberies in Baltimore county with the location of the mean center (square); the data do not include incidents in Baltimore City. As can be seen, the mean centers are located within Baltimore City, a property of the unusual shape of the county (which surrounds the city on three sides). Thus, these mean centers cannot be considered

Figure 4.17: 1996 Baltimore County Burglaries

Location of Incidents and Mean Center

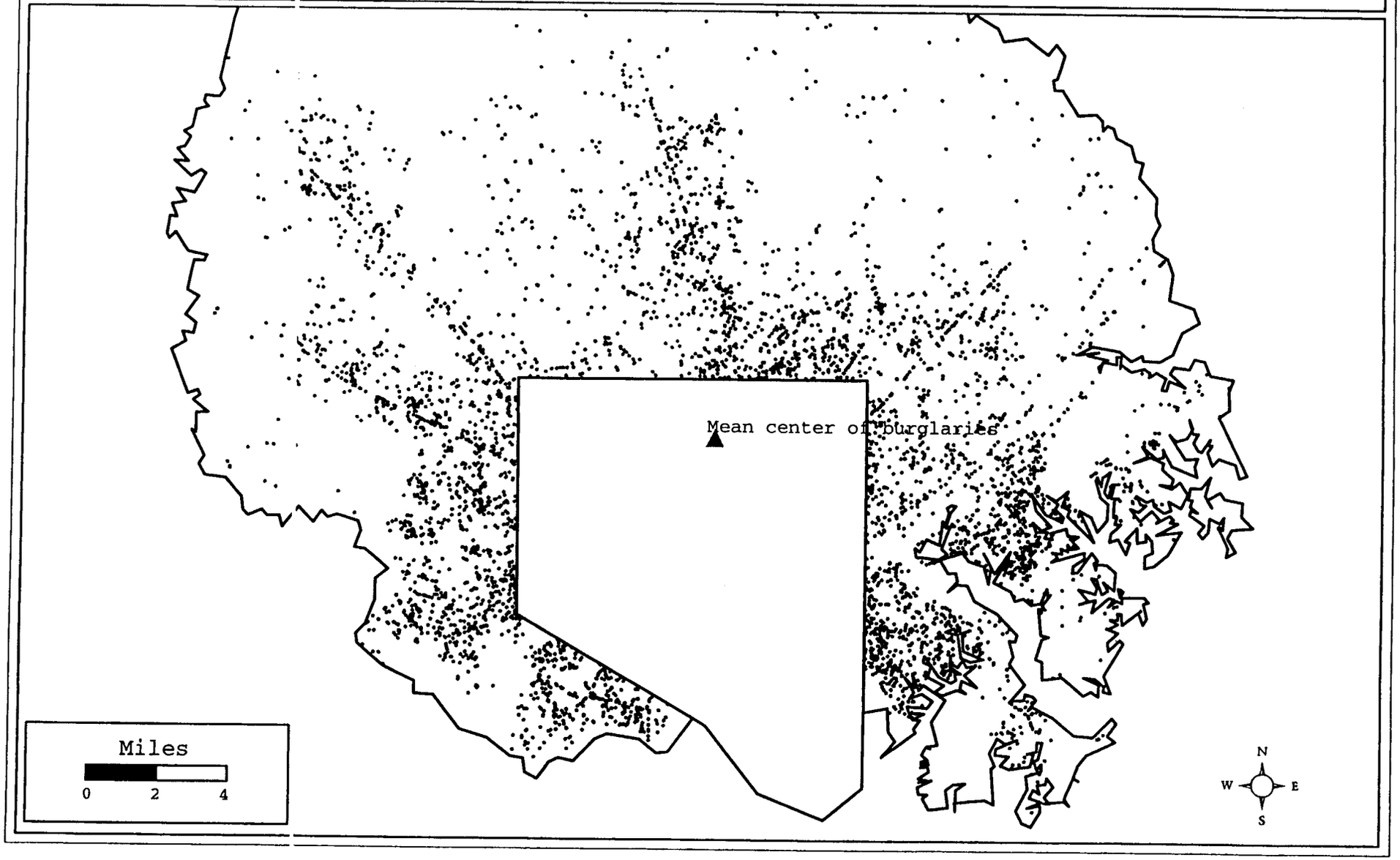
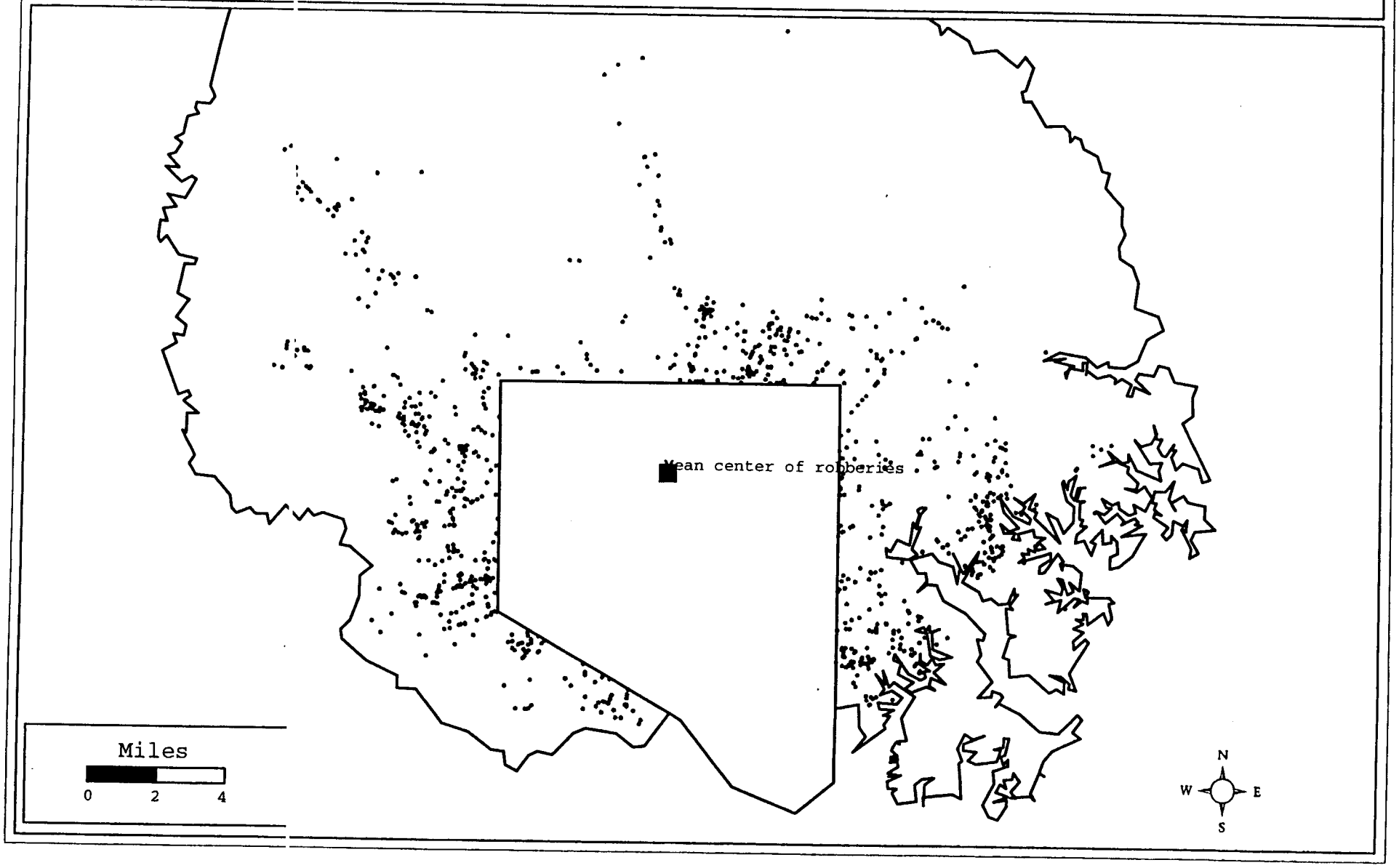


Figure 4.18: 1996 Baltimore County Robberies

Location of Incidents and Mean Center



an unbiased estimate of the metropolitan area, but unbiased estimates for the County only. When the relative positions of the two mean centers are compared (see figure 4.12), the center of robberies is south and west of the center for burglaries. Is this difference significant or not?

To test this, the standard deviations of the two distributions are first compared and the F-test of the larger to the smaller variance is used (equations 4.10a and 4.10b). *CrimeStat* provides the standard deviation of both the X and Y coordinates; the variance is the square of the standard deviation. In this case, the variance for burglaries is slightly larger than for robberies for both the X and Y coordinates.

$$F_X = \frac{S_{XA}^2}{S_{XB}^2} = \frac{0.0154}{0.0145} = 1.058$$

$$F_Y = \frac{S_{YA}^2}{S_{YB}^2} = \frac{0.0058}{0.0029} = 2.007$$

Because both samples are fairly large (1180 robberies and 6051 burglaries), the degrees of freedom are also very large. The F-tables are a little indeterminate with large samples, but the variance ratio approaches 1.00 as the sample reaches infinity. An approximate critical F-ratio can be obtained by the next largest pair of values in the table (1.22 for $p \leq .05$ and 1.32 for $p \leq .01$). Using this criteria, differences in the variances for the X coordinate are probably not significant while that for the Y coordinates definitely are significant. Consequently, the test for a difference in means with unequal variances is used (equations 4.13a and 4.13b).

$$t_X = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_{XA} - \mu_{XB})}{\text{SQRT}\left\{\left[\frac{S_{XA}^2}{N_A} + \frac{S_{XB}^2}{N_B}\right]\right\}} = \frac{-76.608482 - (-76.620838)}{\text{SQRT}\left\{\left[\frac{0.0154}{6051} + \frac{0.0145}{1180}\right]\right\}}$$

$$= \frac{0.0124}{0.0039} = 3.21 \ (p \leq .005)$$

$$t_Y = \frac{(\bar{Y}_A - \bar{Y}_B) - (\mu_{YA} - \mu_{YB})}{\text{SQRT}\left\{\left[\frac{S_{YA}^2}{N_A} + \frac{S_{YB}^2}{N_B}\right]\right\}} = \frac{39.348368 - 39.334816}{\text{SQRT}\left\{\left[\frac{0.0058}{6051} + \frac{0.0029}{1180}\right]\right\}}$$

$$= \frac{0.0136}{0.0018} = 7.36 \quad (p \leq .005)$$

Therefore, whether we use the 'either/or' test (critical $\alpha \leq .025$) or the 'both/and' test (critical $\alpha \leq .1$), we find that the difference in the mean centers is highly significant. Burglaries have a different center of gravity than robberies in Baltimore County.

Differences in the Standard Distance Deviations of Two Samples

Since the standard distance deviation, S_{XY} (equation 4.6) is a standard deviation, differences in the standard distances of two groups can be compared with an equality of variance test (Kanji, 1993, 37).

$$F = \frac{S_{XYA}^2}{S_{XYB}^2} \tag{4.16}$$

with $(N_A - 1)$ and $(N_B - 1)$ degrees of freedom for groups A and B, respectively. This test is usually done with the larger of the variances in the numerator. Since there is only one variance being compared, the critical α are as listed in the tables.

From *CrimeStat*, we find that the standard distance deviation of burglaries is 8.44 miles while that for robberies is 7.42 miles. Earlier, figure 4.12 displayed these two standard distance deviations. As can be seen, the dispersion of incidents, as defined by the standard distance deviation, is greater for burglaries than for robberies. The F-test of the difference is calculated by

$$F = \frac{S_{XYA}^2}{S_{XYB}^2} = \frac{8.44^2}{7.42^2} = 1.29$$

with 6050 and 1180 degrees of freedom respectively. Again, the F-tables are slightly indeterminate with respect to large samples, but the next largest F beyond infinity is 1.25 for $p \leq .05$ and 1.38 for $p \leq .01$. Thus, it appears that burglaries have a significantly greater dispersion than robberies, at least at the $p \leq .05$ level.

Differences in the Standard Deviational Ellipses of Two Samples

In a standard deviational ellipse, there are actually six variables being compared:

Mean of X
Mean of Y
Angle of rotation
Standard deviation along the transformed X axis

Standard deviation along the transformed Y axis
Area of the ellipse

Differences in the mean centers

Comparisons between the two mean centers can be tested with the above statistics.

Differences in the angle of rotation

Unfortunately, to our knowledge, there is not a formal test for the difference in the angle of rotation. Until this test is developed, we have to rely on subjective judgements.

Differences in the standard deviations along the transformed axes

The differences in the standard deviations along the transformed axes (X and Y) can be tested with an equality of variance test (Kanji, 1993, 37).

$$F_{Sx} = \frac{S_{X1}^2}{S_{X2}^2} \qquad (4.17a)$$

$$F_{SY} = \frac{S_{Y1}^2}{S_{Y2}^2} \qquad (4.17b)$$

with $(N_A - 1)$ and $(N_B - 1)$ degrees of freedom for groups A and B respectively. This test is usually done with the larger of the variances in the numerator. The example above for comparing the mean centers of Baltimore County burglaries and robberies illustrated the use of this test.

Differences in the areas of the two ellipses

Since an area is a variance, the differences in the areas of the two ellipses can be compared with an equality of variance test (Kanji, 1993, 37).

$$F = \frac{Area_A}{Area_B} \qquad (4.18)$$

with $(N_A - 1)$ and $(N_B - 1)$ degrees of freedom for groups 1 and 2 respectively. This test is done with the larger of the variances in the numerator.

Significance levels

The testing of each of these parameters for the difference between two ellipses is even more complicated than the difference between two mean centers since there are up to six parameters which must be tested (differences in mean X, mean Y, angle of rotation, standard deviation along transformed X axis, standard deviation along transformed Y axis, and area of ellipse). However, as with differences in mean center of two groups, there are two different interpretations of differences.

Comparison I: That the two ellipses differ on ANY of the parameters

Comparison II: That the two ellipses differ on ALL parameters.

In the first case, the critical probability level, α, must be divided by the number of parameters being tested, α/p. In theory, this could involve up to six tests, though in practice some of these may not be tested (e.g., the angle of rotation). For example, if five of the parameters are being estimated, then the critical probability level at $\alpha \leq .05$ is actually $\alpha \leq .01$ (.05/5).

In the second case, the critical probability level, α, is multiplied by the number of parameters being tested, $\alpha * p$, since *all* tests must be significant for the two ellipses to be considered as different. For example, if five of the parameters are being estimated, then the critical probability level, say, at $\alpha \leq .05$ is actually $\alpha \leq .25$ (.05*5).

Decision-making Without Formal Tests

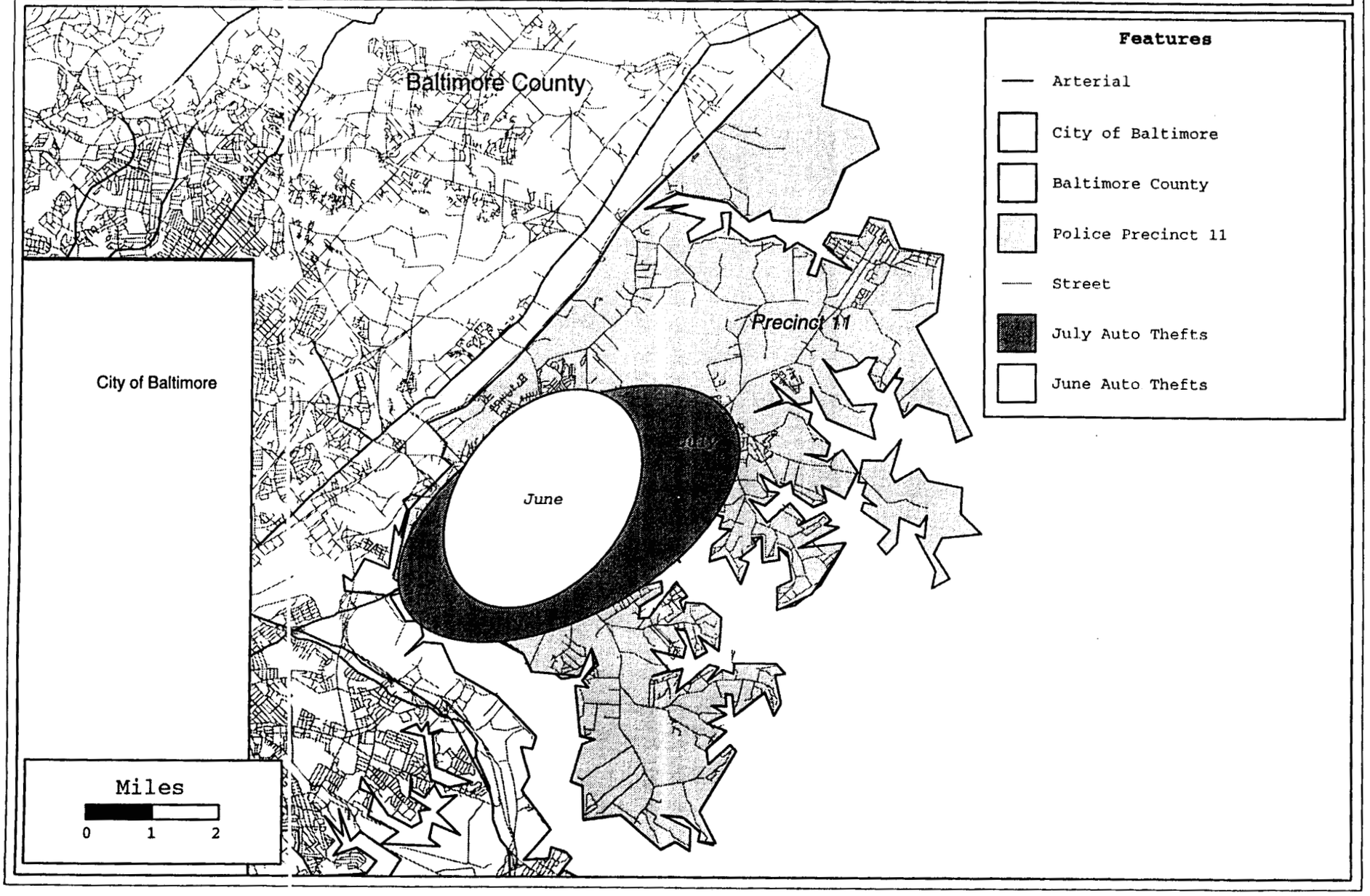
Formal significance testing has the advantage of providing a consistent inference about whether the difference in two distributions is likely or unlikely to be due to chance. Almost all formal tests compare the distribution of a statistic with that of a random distribution. However, police departments frequently have to make decisions based on small samples, in which case the formal tests are less useful than they would with larger samples. Still, the centrographic statistics calculated in *CrimeStat* can be useful and can help a police department make decision even in the absence of formal tests.

Example 2: June and July auto thefts in Precinct 11

We want to illustrate the use of these statistics to make decisions with two more examples. The first is a comparison of crimes in small geographical areas. In most metropolitan areas, most analysts will concentrate on particular sub-areas of the jurisdiction, rather than on the jurisdiction itself. In Baltimore County, for instance, analysis is done both for the jurisdiction as a whole as well as by individual precincts. Below in Figure 4.19 are the standard deviational ellipses for 1996 auto thefts for June and July in Precinct 11 of Baltimore County. As can be seen, there was a spatial shift that occurred between June and July of that year, the result most probably of increased vacation travel to the Chesapeake Bay. While the comparison is very simple, involving looking at the graphical object created by *CrimeStat*, such a month to month comparison

Figure 4.19: Auto Theft Change in Precinct 11

Ellipses of June and July 1996



can be useful for police departments because it points to a shift in incident patterns, allowing the police department to reorient their patrol units.

Example 3: Serial burglaries in Baltimore City and Baltimore County

The second example illustrates a rash of burglaries that occurred on both sides of the border of Baltimore City and Baltimore County. On one hand there were ten residential burglaries that occurred on the western edge of the City/County border within a short time period of each other and, on the other hand, there were 13 commercial burglaries that occurred in the central part of the metropolitan areas. Both police departments suspected that these two sets were the work of a serial burglar (or group of burglars). What they were not sure about was whether the two sets of burglaries were done by the same individuals or by different individuals.

The number of incidents involved are too small for significance testing; only one of the parameters tested was significant and that could easily be due to chance. However, the police do have to make a guess about the possible perpetrator even with limited information. Let's use *CrimeStat* to try and make a decision about the distributions.

Figure 4.20 illustrates these distributions. The thirteen commercial burglaries are shown as squares while the ten residential burglaries are shown as triangles. Figure 4.21 plots the mean centers of the two distributions. They are close to each other, but not identical. An initial hunch would suggest that the robberies are committed by two perpetrators (or groups of perpetrators), but the mean centers are not different enough to truly confirm this expectation. Similarly, figure 4.22 plots the center of minimum distance (or median center). Again, there is a difference in the distribution, but it is not great enough to truly rule out the single perpetrator theory.

Figure 4.23 plots the raw standard deviations, expressed as a rectangle by *CrimeStat*. The dispersion of incidents overlaps to a sizeable extent and the area defined by the rectangle is approximately the same. In other words, the search area of the perpetrator or perpetrators is approximately the same. This might argue for a single perpetrator, rather than two. Figure 4.24 shows the standard distance deviation of the two sets of incidents. Again, there is sizeable overlap and the search radiuses are approximately the same.

Only when we get to the standard deviational ellipse, however, do we see a fundamental difference between the two distributions (figure 4.25). The pattern of commercial robberies is falling along a northeast-southwest orientation while that for residential robberies along a northwest-southeast axis. In other words, when the orientation of the incidents is examined, as defined by the standard deviational ellipse, there are two completely opposite patterns. Unless this difference can be explained by an obvious factor (e.g., the distribution of commercial establishments), it is probable that the two sets of robberies were committed by two different perpetrators (or groups of perpetrators). In the actual case, the police actually did conclude that the burglaries were committed by two sets of offenders.

Figure 4.20: Profiling Serial Burglaries

Incident Distribution of Two Serial Offenders

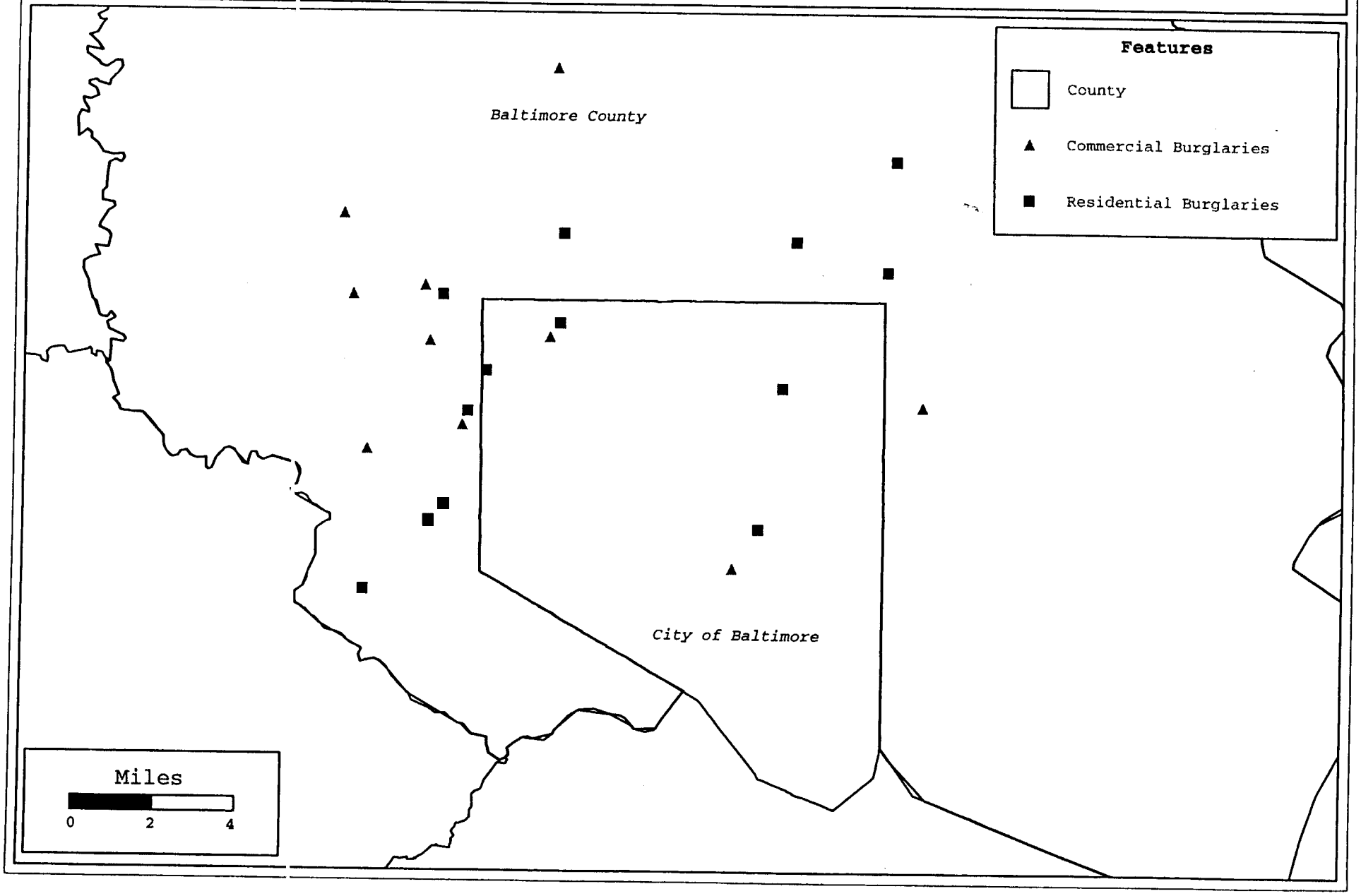


Figure 4.21: Profiling Serial Burglaries

Mean Centers of Incidents for Two Serial Offenders

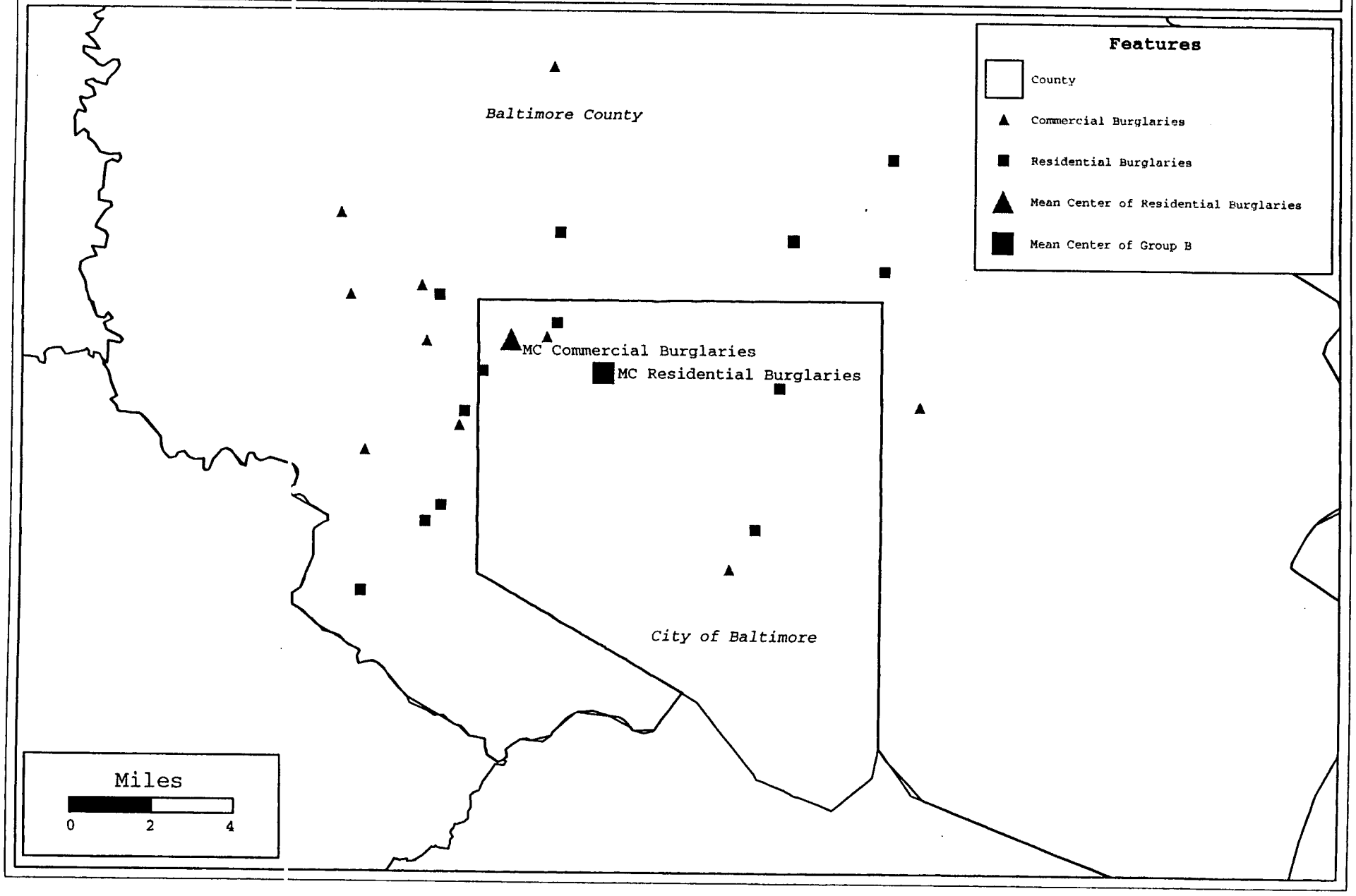


Figure 4.22: Profiling Serial Burglaries

Median Centers of Incidents for Two Serial Offenders

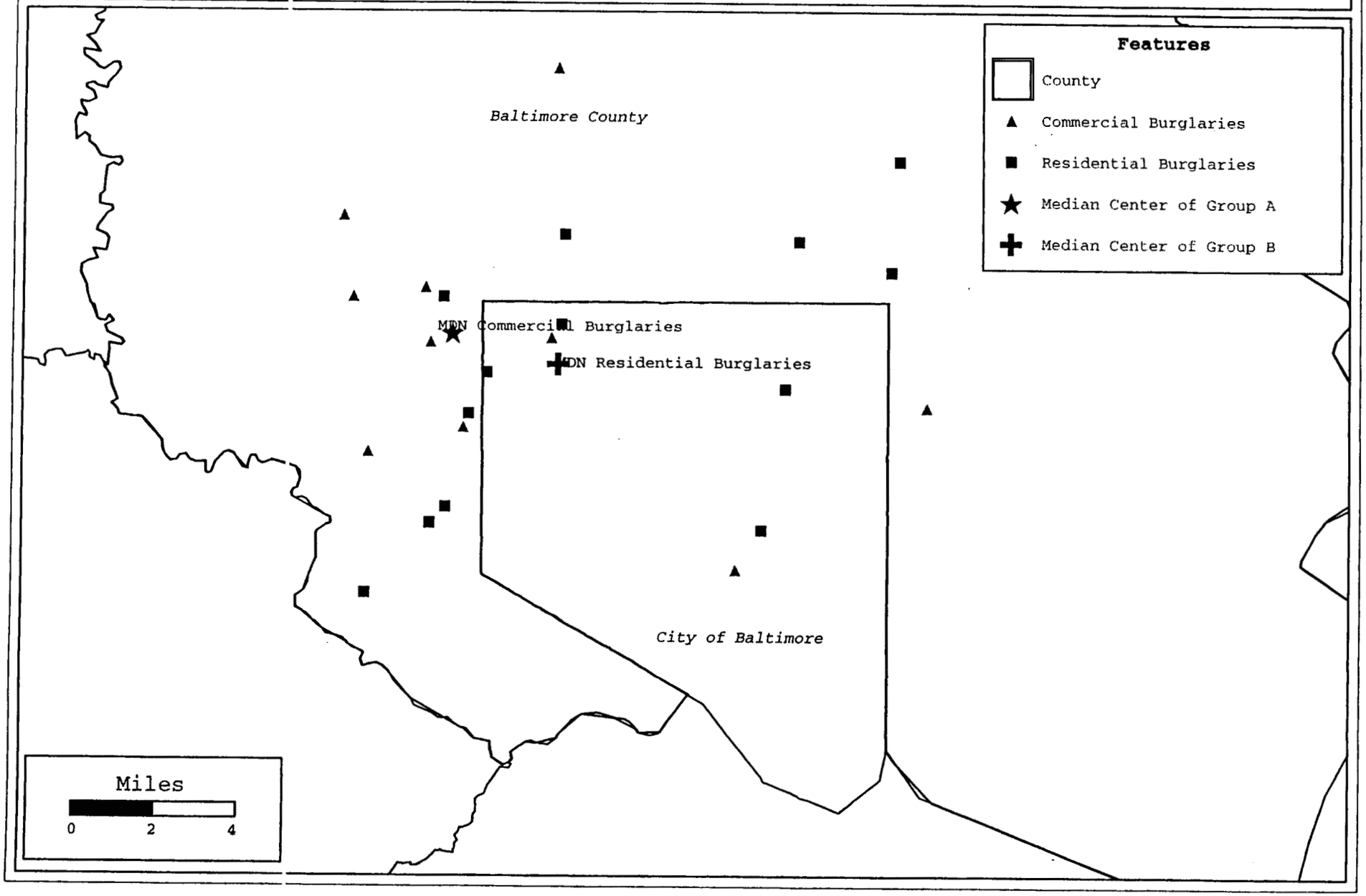


Figure 4.23: Profiling Serial Burglaries

Standard Deviations of Incidents for Two Serial Offenders

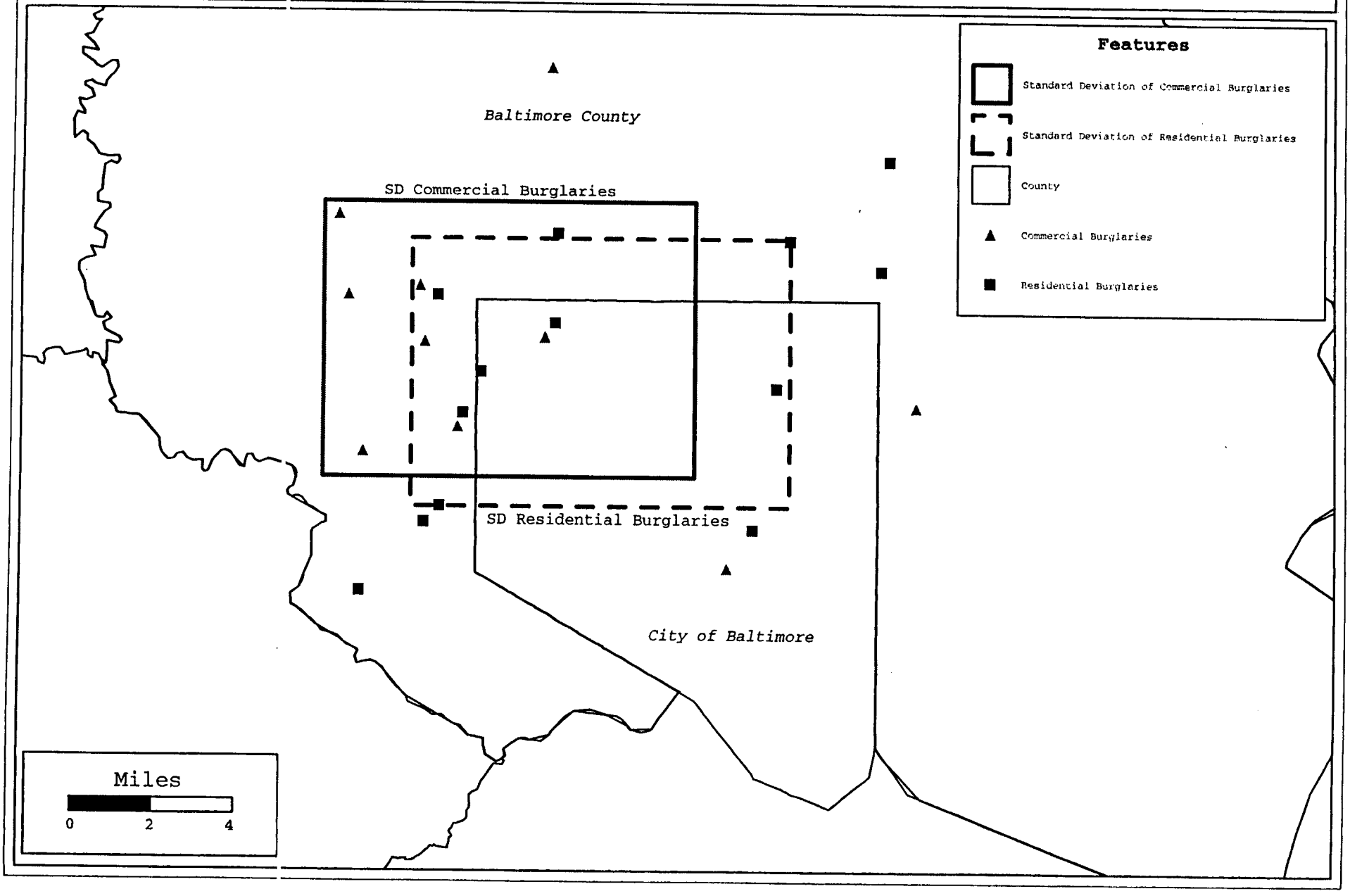


Figure 4.24: Profiling Serial Burglaries

Standard Distance Deviation of Incidents for Two Serial Offenders

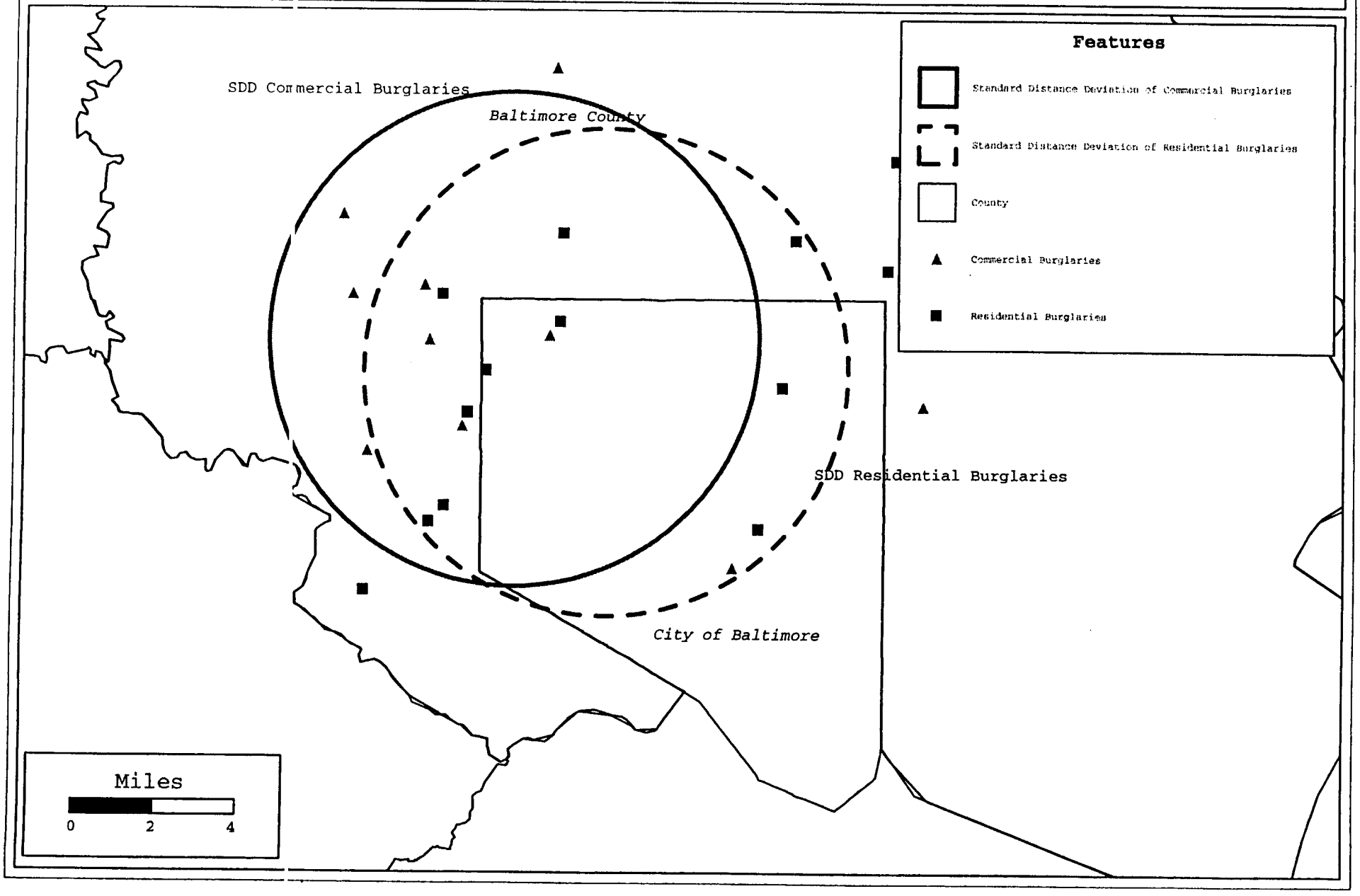
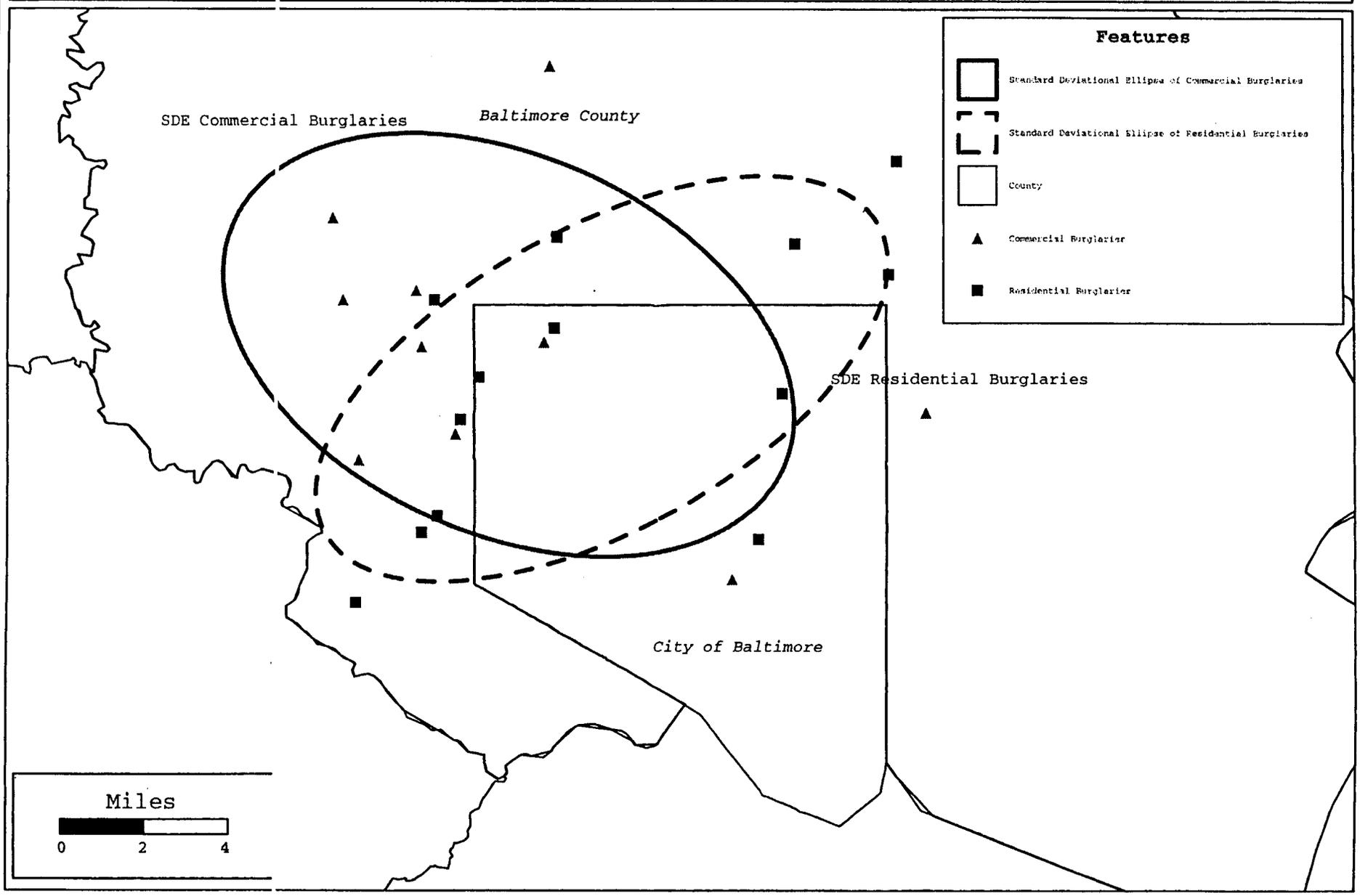


Figure 4.25: Profiling Serial Burglaries

Standard Deviational Ellipse of Incidents for Two Serial Offenders



Centrographic descriptors can be very powerful tools for examining spatial patterns. They are a first step in any spatial analysis, but an important one. The above example illustrates how they can be a basis for decision-making, even with small samples. A couple of other examples can be illustrated.

Example 4: Auto thefts over time in Baltimore County

These statistics are useful for comparing different types of crimes (e.g., burglaries versus robberies as in the above examples) and for the same crime at different time periods. Figure 4.26 shows a comparison of standard deviational ellipses for motor vehicle thefts in both Baltimore County and Baltimore City, broken out by time period. Incidents were categorized into one of four time periods: nighttime (Midnight-6 am); morning (6 am-Noon); afternoon (Noon-6 PM); and evening (6 PM - Midnight). The widest ellipse is for morning thefts while the tightest ellipse is for evening thefts. Afternoon thefts and nighttime thefts are intermediate. This pattern can generate hypotheses about the behavior patterns of auto thieves. For example, in the evening, many of the thefts are concentrated because restaurant use, entertainment and evening shopping tend to be concentrated more in the central Baltimore core. The widening dispersion in nighttime auto thefts and the very wide dispersion in morning auto thefts could indicate a pattern of thefts occurring at people's residences, rather than at employment centers. More research will be needed to determine some of the causes of this distribution, but a comparison of auto thefts by time period reveals a dynamic pattern.

Directional Mean and Variance

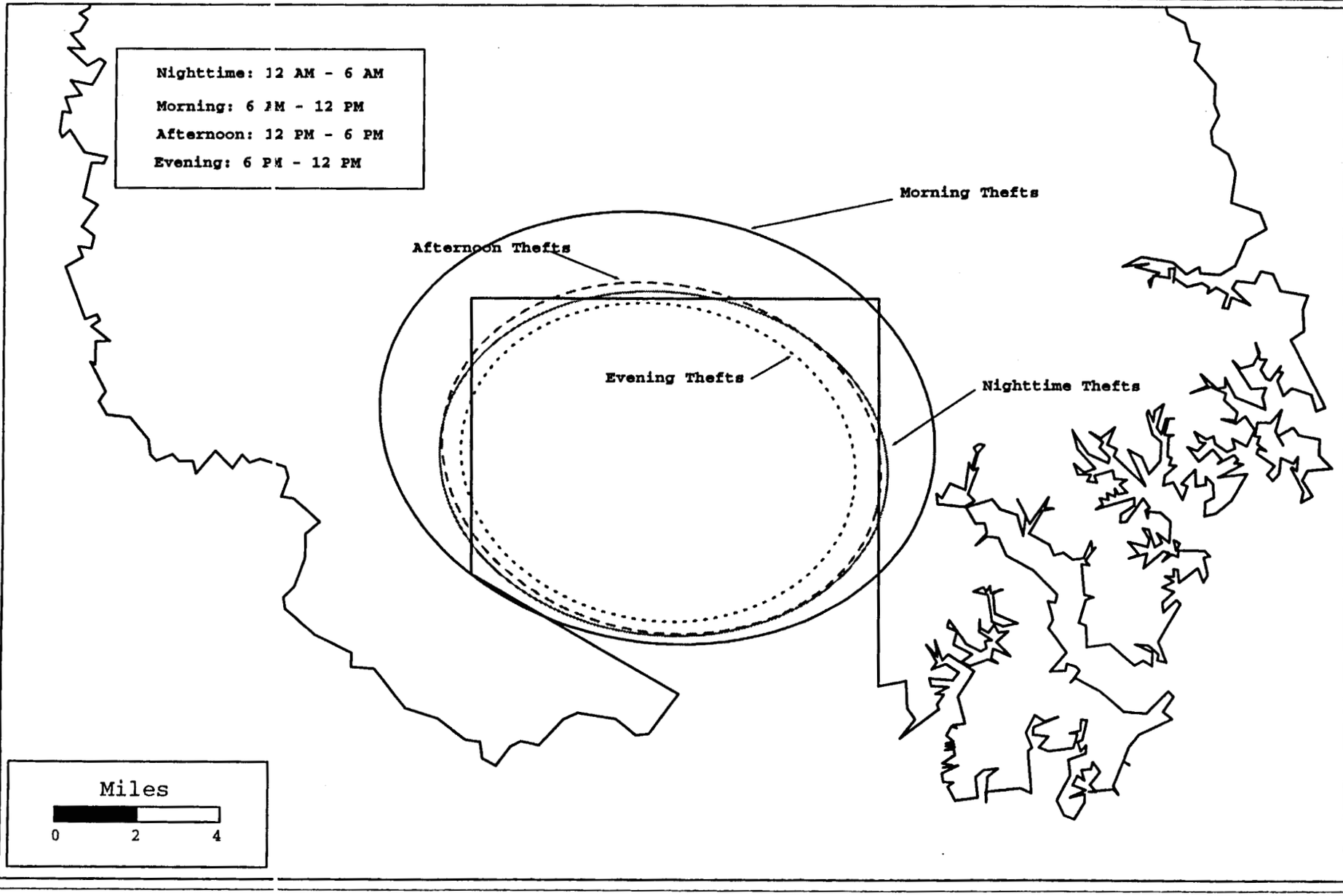
Centrographic statistics utilize the coordinates of a point, defined as an X and Y value on either a spherical or projected/Cartesian coordinate system. There is another type of metric that can be used for identifying incident locations, namely a *polar coordinate* system. A *vector* is a line with direction and length. In this system, there is a reference vector (usually 0^0 due North) and all locations are defined by angular deviations from this reference vector. By convention, angles are defined as deviations from 0^0, clockwise through 360^0. Note the measurement scale is a circle which returns back on itself (i.e. 0^0 is also 360^0). Point locations can be represented as vectors on a polar coordinate system.

With such a system, ordinary statistics cannot be used. For example, if there are five points which on the northern side of the polar coordinate system and are defined by their angular deviations as 0^0, 10^0, 15^0, 345^0, and 350^0 from the reference vector (moving clockwise from due North), the statistical mean will produce an erroneous estimate of 144^0. This vector would be southeast and will lie in an opposite direction from the distribution of points.

Instead, statistics have to be calculated by trigonometric functions. The input for such a system is a set of vectors, defined as angular deviations from the reference vector. As mentioned above, for trigonometric calculations, all decimal degree angles have to be first converted into radians. It is also possible to define the length of each vector, though *CrimeStat* does not currently calculate length statistics. Instead, it is assumed that all

Figure 4.26: Temporal Changes in 1996 Baltimore Auto Thefts

Ellipses of Four Time Periods



vectors have equal (or unknown) lengths. There are, however, several directional statistics that use information on the angular displacement from the reference vector.

The *Mean Direction* is the resultant of all individual vectors (i.e., points defined by their angles from the reference vector). It is an angle which summarizes the mean direction. Graphically, a *resultant* is the sum of all vectors and can be shown by laying end-to-end each vector. Statistically, it is defined as

$$\text{Mean direction} = \bar{\theta} = \text{Arctan}\left[\frac{\Sigma \sin \theta_i}{\Sigma \cos \theta_i}\right] \qquad (4.19)$$

where the summation of sines and cosines is over the total number of points, i, defined by their angles, θ_i. In determining the mean direction, the quadrant of the resultant must be identified:

A. If $\Sigma \sin \theta_i > 0$ and $\Sigma \cos \theta_i > 0$, then $\bar{\theta}$ can be used directly as the mean direction.

B. If $\Sigma \sin \theta_i > 0$ and $\Sigma \cos \theta_i < 0$, then the mean direction is $180-\bar{\theta}$.

C. If $\Sigma \sin \theta_i < 0$ and $\Sigma \cos \theta_i < 0$, then the mean direction is $180+\bar{\theta}$.

D. If $\Sigma \sin \theta_i < 0$ and $\Sigma \cos \theta_i > 0$, then the mean direction is $360-\bar{\theta}$.

Conceptually, the mean direction can be thought of as a vector from the origin to the resultant of all the points. On a two-dimensional 'Cartesian' plane, this can be represented as the hypotenuse of a right triangle where the X value is the length of the near side of the triangle and the Y value is the length of the far side of the triangle (from the origin).[6]

The numerator and denominator can be treated as separate terms.

$$\text{Mean of Sines } (\bar{S}) = \Sigma \sin \theta_i / N \qquad (4.20a)$$

$$\text{Mean of Cosines } (\bar{C}) = \Sigma \cos \theta_i / N \qquad (4.20b)$$

where both summations are over the angles for all points, i.

The dispersion (or variance) of the angles are also defined by trigonometric functions. The unstandardized variance, R, is sometimes called the *sample resultant length* since it is the resultant of all vectors (angles).

$$R = \sqrt{[(\Sigma \sin \theta_i)^2 + (\Sigma \cos \theta_i)^2]} \qquad (4.21)$$

Because R is generally higher with larger samples, it is standardized by dividing by N to produce a *mean resultant length.*

$$\bar{R} = \frac{R}{N} \tag{4.22}$$

where N is the number points (sample size). Finally, the *circular variance* is calculated

$$\text{Circular variance} = 1 - \frac{R}{N} = 1 - \bar{R} \tag{4.23}$$

This is a standardized variance which varies from 0 (no variability) to 1 (maximum variability). The details of the derivations can be found in Burt and Barber (1996) and Gaile and Barber (1980).

CrimeStat Input and Output for Directional Mean and Variance

The required input to use the directional mean and variance routine in *CrimeStat* is a list of angles, one for each case. The angles should be in decimal degrees. *CrimeStat* will convert the angles automatically to radians during the calculations, and will convert the results back to angles.

CrimeStat calculates the mean direction and the circular variance of a series of points defined by their angles. On the primary file screen, the user must select Direction (angles) as the coordinate system. On the spatial distribution screen, the user selects *Directional mean and variance* (figure 4.27). The results screen prints out the directional values for the primary file (figure 4.28).

What use is this to police departments? It is conceivable that in special circumstances, a police department would only be able to identify incidents by their angular dispersion rather than by their actual X/Y coordinates. For example, a stolen vehicle with a radio broadcast unit on board may produce intermittent signals which can be detected by a central receiver. Each signal would be identified as an angular deviation from due North, but the exact location could not be easily determined. If the signal was continuous, then police cars could converge on it from several directions, but if the signal only is picked up periodically, then the only information is a series of angular deviations. The method gives an approximate solution and is, therefore, less precise than X/Y coordinate locations. But, if angular directions are the only information that is available, the method can be used to produce an approximate center to the distribution.

The *Los Angeles Times* recently published an article on the Federal Communications Commission Compliance and Information Bureau that investigate thousands of minor incidents and over 500 serious incidents of airwave interference a year (Los Angeles Times, 1998). These vary from pirate radio stations to news walkie-talkies

Figure 4.27: Selecting Directional Mean And Variance

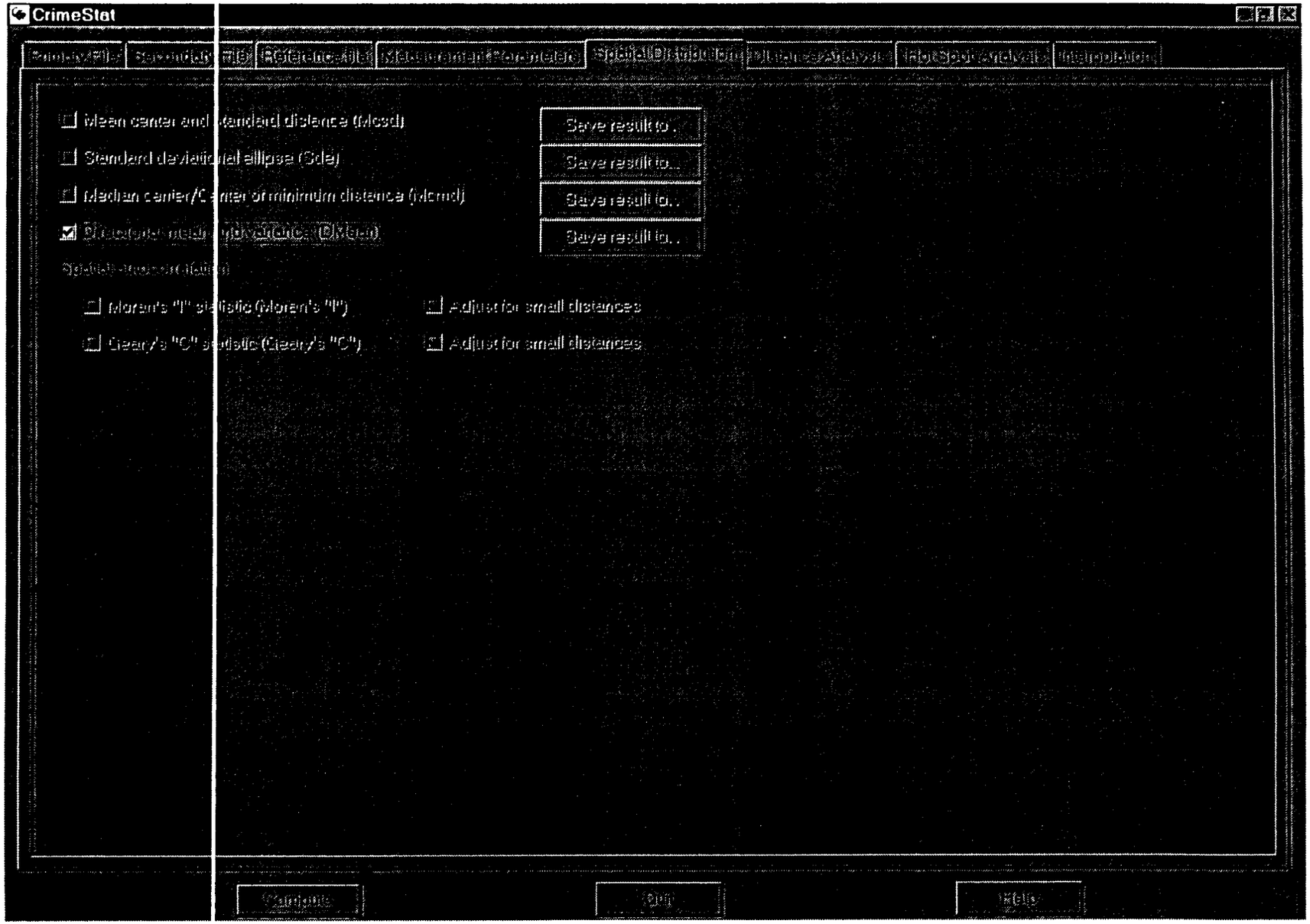
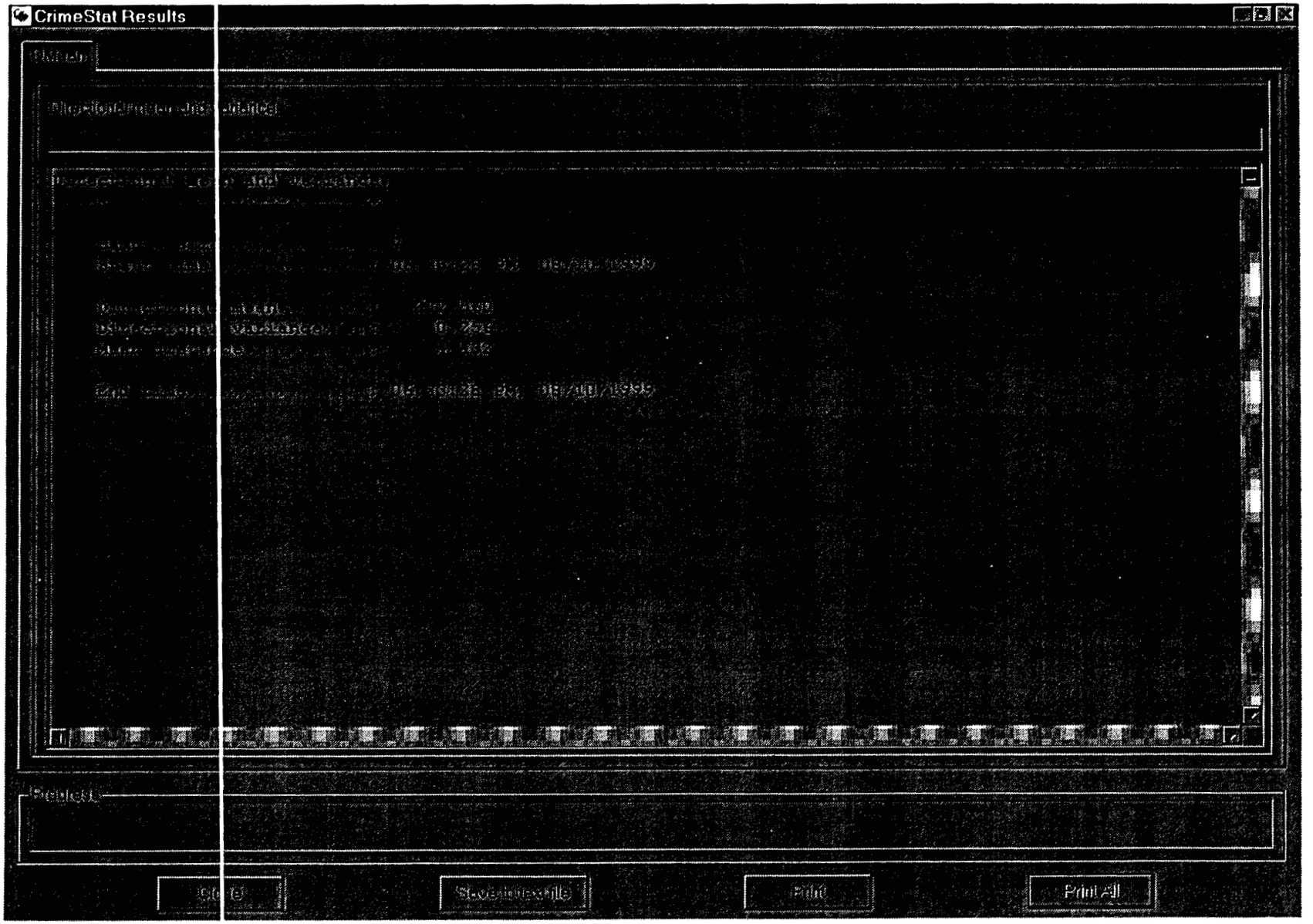


Figure 4.28: Directional Mean And Variance Output



that have interfered with Secret Service communications to small community broadcasters to unauthorized disruptions of regular television stations. For continuous signals, the Bureau can hone in on the location with multiple detectors (triangulation). But for periodic signals, a statistical system such as this could help them focus their search.

Example 5: Directional mean and variance

An example will be given of the use of the directional mean and variance. Table 4.1 presents a set of seven angular measurements taken by a police station of an unauthorized mobile unit emitting periodic signals and an estimate of the approximate distance (in miles) of the mobile source from the station. The reference vector is due North. What is the directional mean and variance? Figure 4.29 displays the location of the police station and the location of the seven incidents. The angular measurements have been taken as clockwise rotations from the reference vector, which is true North; this is displayed with a dotted reference circle. The approximate distance of each incident from the police station is shown by solid lines.

Using these data, *CrimeStat* calculates that the mean direction is 272.56^0 and the circular variance is 0.256. Since the circular variance is a standardized measure varying from 0 to 1, this value indicates a relatively small angular variance. The absolute range of measurements varies by 140^0 (335^0 - 195^0), which is a little over one-third of a complete circle (360^0). The variance of 0.256 represents about a quarter of a complete circle. The average distance is 4.39 miles. Figure 4.30 shows the intersection of the mean angle and the mean distance, and provides an approximate center to the distribution.

Statistical Test of Differences in Mean Direction Between Two Groups

Statistical tests of different angular distributions can be made with these statistics. To test the difference in the angle of rotation between two groups, a Watson-Williams test can be used (Kanji, 1993; 153-54). The steps in the test are as follows:

1. All angles, θ_i, are converted into radians

$$\text{Radian}_i = \text{Angle}_i * \pi/180 \qquad (4.24)$$

2. For each sample separately, *A* and *B*, the following measures are calculated

$$C_j = \sum_{A=1}^{N_1} \cos \theta_j \qquad S_j = \sum_{A=1}^{N_1} \sin \theta_j \qquad (4.25a)$$

$$C_k = \sum_{B=1}^{N_2} \cos \theta_k \qquad S_k = \sum_{B=1}^{N_2} \sin \theta_k \qquad (4.25b)$$

where θ_j and θ_k are the individual angles for the respective groups, *A* and *B*.

Table 4.1

Unauthorized Mobile Signals
Angle of Deviation From Due North

Incident	Measured Angle	Estimated Distance (miles)
1	270^0	3.5
2	285^0	5.0
3	240^0	4.0
4	315^0	0.75
5	335^0	2.5
6	195^0	8
7	260^0	7

3. Calculate the resultant lengths of each group

$$R_A = \sqrt{[\, C_A^2 + S_A^2 \,]} \qquad (4.26a)$$

$$R_B = \sqrt{[\, C_B^2 + S_B^2 \,]} \qquad (4.26b)$$

4. Resultant lengths for the combined sample are calculated as well as the length of the resultant vector.

$$C = C_A + C_B \qquad (4.27a)$$

$$S = S_A + S_B \qquad (4.27b)$$

$$R = \sqrt{[\, C^2 + S^2 \,]} \qquad (4.27c)$$

$$N = N_A + N_B \qquad (4.27d)$$

$$R^* = \frac{(R_A + R_B)}{N} \qquad (4.27e)$$

5. An F-test of the two angular means is calculated with

$$F = g\,(N - 2)\,\frac{R_A + R_B - R}{N - (R_A + R_B)} \qquad (4.28a)$$

Figure 4.29: Unauthorized Mobile Signals

Angle from Reference Vector and Approximate Distance

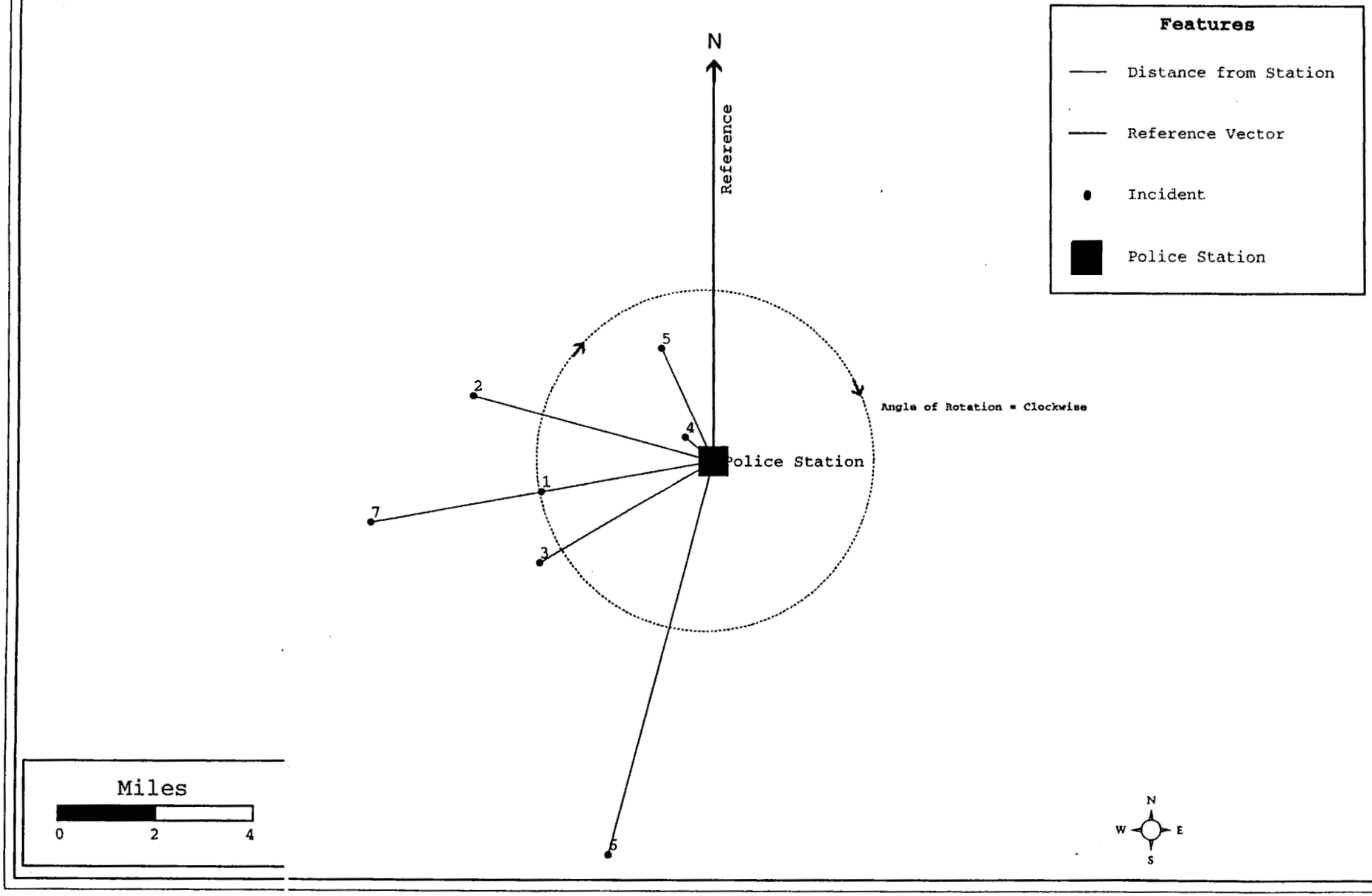
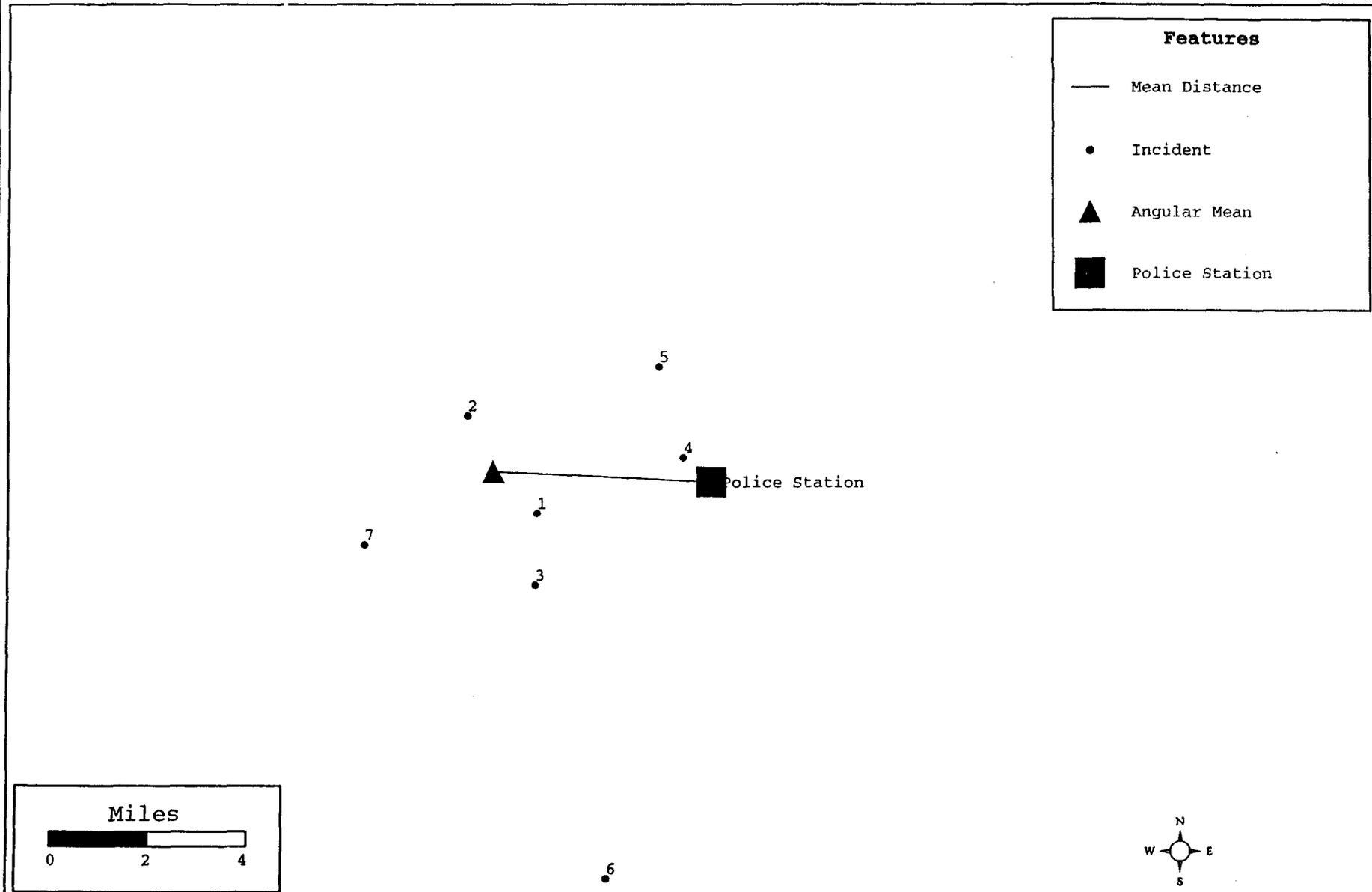


Figure 4.30: Unauthorized Mobile Signals

Intersection of Angular Mean and Mean Distance



where

$$g = 1 - \frac{3}{8k} \qquad (4.28b)$$

with k being identified from a maximum likelihood Von Mises distribution by referencing R^* with 1 and N-2 degrees of freedom (Mardia, 1972; Gaile and Burt, 1980). Some of the reference k's are given in table 4.2 (from Mardia, 1972; Kanji, 1993, table 38).

6. Reject the null hypothesis of no angular difference if the calculated F is greater than the critical value $F_{1, N-2}$.

Example 6: Angular comparisons between two groups

A second example is that of sets of angular measurements from two different groups, A and B. Table 4.3 provides the data for the two sets. The angular mean for Group A is 144.83^0 with a directional variance of 0.35 while the angular mean for Group B is 258.95^0 with a directional variance of 0.47. The higher directional variance for Group B suggests that there is more angular variability than for Group A.

Using the Watson-Wheeler test, we compare these two distributions.

1. All angles are converted into radians (equation 4.24)

2. The cosines and sines of each angle are taken and are summed within groups (equations 4.31a and 4.31b)

$C_A = -3.1981$ $\quad S_A = 2.2533$
$C_B = -.8078$ $\quad S_B = -4.1381$

3. The resultants are calculated (equations 4.32a and 4.32b)

$R_A = 3.9121$
$R_B = 4.2162$

4. Combined sample characteristics are defined (equations 4.33a through 4.33e)

$C = -4.0059$
$S = -1.8848$
$R = 4.4271$
$N = 14$
$R^* = 0.5806$

Table 4.2

Maximum Likelihood Estimates for Given R^* in the Von Mises Case
(from Mardia, 1972; Kanji, 1993, table 38)

R^*	k
0.00	0.00000
0.05	0.10013
0.10	0.20101
0.15	0.30344
0.20	0.40828
0.25	0.51649
0.30	0.62922
0.35	0.74783
0.40	0.87408
0.45	1.01022
0.50	1.15932
0.55	1.32570
0.60	1.51574
0.65	1.73945
0.70	2.01363
0.75	2.36930
0.80	2.87129
0.85	3.68041
0.90	5.3047
0.95	10.2716
1.00	infinity

Table 4.3

Comparison of Two Groups for Angular Measurements
Angle of Deviation From Due North

Group A		Group B	
Incident	Measured Angle	Incident	Measured Angle
1	160	1	196
2	164	2	212
3	240	3	297
4	100	4	280
5	95	5	235
6	120	6	353
		7	190
		8	340

5. Once the parameter, k, is obtained (approximated from table 4.2 or obtained from Mardia, 1972 or Kanji, 1993), g is calculated, and an F-test is constructed (equations 4.34a and 4.34b).

   k = 1.44
   g = 0.7396
   F = 5.59

6. The critical F for 1 and 12 degrees of freedom is 4.75 ($p \leq .05$) and 9.33 ($p \leq .01$). The test is significant at the $p \leq .05$ level and we reject the null hypothesis of no angular differences between the two groups. Group A has a different angular distribution than Group B.

Spatial Autocorrelation

The concept of *spatial autocorrelation* is one of the most important in spatial statistics. Spatial *independence* is an arrangement of incident locations such that there are no spatial relationships between any of the incidents. The intuitive concept is that the location of an incident (e.g., a street robbery, a burglary) is unrelated to the location of any other incident. The opposite condition - spatial autocorrelation, is an arrangement of incident locations where the location of points are related to each other, that is they are not statistically independent of one another. In other words, spatial autocorrelation is a spatial arrangement where spatial independence has been violated.

When events or people or facilities are clustered together, we refer to this arrangement as *positive* spatial autocorrelation. Conversely, an arrangement where people, events or facilities are dispersed is referred to as *negative* spatial autocorrelation; it is a rarer arrangement, but does exist (Levine, 1999).

Many, if not most, social phenomena are spatially autocorrelated. In any large metropolitan area, most social characteristics and indicators, such as the number of persons, income levels, ethnicity, education, employment, and the location of facilities are not spatially independent, but tend to be concentrated.

There are practical consequences. Police and crime analysts know from experience that incidents frequently cluster together in what are called 'hot spots'. This non-random arrangement allows police to target certain areas or zones where there are high concentrations as well as prioritize areas by the intensity of incidents. Many of the incidents are committed by the same individuals. For example, if a particular neighborhood had a concentration of street robberies over a time period (e.g., a year), many of these robberies will have been committed by the same perpetrators. Statistical dependence between events often has common causes.

Statistically, however, non-spatial independence suggests many statistical tools and inferences are inappropriate. For example, the use of correlation coefficients or Ordinary Least Squares regression (OLS) to predict a consequence (e.g., the correlates or predictors

of burglaries) assumes that the observations have been selected randomly. If the observations, however, are spatially clustered in some way, the estimates obtained from the correlation coefficient or OLS estimator will be biased and overly precise. They will be biased because the areas with higher concentration of events will have a greater impact on the model estimate and they will overestimate precision because, since events tend to be concentrated, there are actually fewer number of independent observations than are being assumed. This concept of spatial autocorrelation underlies almost all the spatial statistics tools which are included in *CrimeStat*. We will return to the concept in each of the next three chapters because the concept is implicit in all the tools that will be discussed.

Indices of Spatial Autocorrelation

There are a number of formal statistics which attempt to measure spatial autocorrelation. This include simple indices, such as the Moran's I" or Geary's C statistic; derivatives indices, such as Ripley's K statistic (Ripley, 1976) or the application of Moran's I to individual zones (Anselin, 1995); and multivariate indices, such as the use of a spatial autocorrelation parameter in a bivariate regression model (Cliff and Ord, 1973; Griffith, 1987) or the use of a spatially-lagged dependent variable in a multiple variable regression model (Anselin, 1992). The simple indices attempt to identify whether spatial autocorrelation exists for a single variable, while the more complicated indices attempt to estimate the effect of spatial autocorrelation on other variables.

CrimeStat includes two simple indices: Moran's I statistic and Geary's C statistic. They are very similar indices and are often used in conjunction. The Moran statistic is slightly more robust than the Geary, but the Geary is often used as well.

Moran's I

Moran's I statistic (Moran, 1950) is one of the oldest indicators of spatial autocorrelation. It is applied to zones or points which have continuous variables associated with them (intensities). For any continuous variable, X_i, a mean can be calculated and the deviation of any one observation from that mean can also be calculated. The statistic then compares the value of the variable at any one location with the value at all other locations (Ebdon, 1985; Griffith, 1987; Anselin, 1992). Formally, it is defined as

$$I = \frac{N \sum_i \sum_j W_{ij} (X_i - \bar{X})(X_j - \bar{X})}{(\sum_i \sum_j W_{ij}) \sum_i (X_i - \bar{X})^2} \qquad (4.29)$$

where N is the number of cases, X_i is the variable value at a particular location, i, X_j is the variable value at another location (where $i \neq j$), X is the mean of the variable and W_{ij} is a weight applied to the comparison between location i and location j.

In Moran's initial formulation, the weight variable, W_{ij}, is a contiguity matrix. If zone j is adjacent to zone i, the interaction receives a weight of 1. Otherwise, the

interaction receives a weight of 0. Cliff and Ord (1973) generalized these definitions to include any type of weight. In more current use, W_{ij} is a distance-based weight which is the inverse distance between locations i and j ($1/d_{ij}$). *CrimeStat* uses this interpretation.

Moran's I is similar to a correlation coefficient in that it compares the sum of the cross-products of values at different locations, two at a time weighted by the inverse of the distance between the locations, with the variance of the variable. Like the correlation coefficient, it varies between -1.0 and + 1.0. When nearby points have similar values, the cross-product is high. Conversely, when nearby points have dissimilar values, the cross-product is low. Consequently, an I value which is high indicates more spatial autocorrelation than an I which is low.

However, unlike the correlation coefficient, the theoretical value of the index does not equal 0 for lack of spatial dependence, but instead a number which is negative but very close to 0.

$$E(I) = -\frac{1}{N-1} \tag{4.30}$$

Values of I above the theoretical mean, E(I), indicate positive spatial autocorrelation while values of I below the theoretical mean indicate negative spatial autocorrelation.

Adjustment for small distances

CrimeStat calculates the traditional Moran's I formula using equation 4.29. However, there is one problem with this formula which can lead to unreliable results. The distance weights between two locations, W_{ij}, is defined as the reciprocal of the distance between the two points:

$$W_{ij} = \frac{1}{d_{ij}} \tag{4.31}$$

Unfortunately, as d_{ij} becomes small, then W_{ij} becomes very large, approaching infinity as the distance between the points approaches 0. If the two zones were next to each other, which would be true for two adjacent blocks for example, then the pair of observations would have a very high weight, sufficient to distort the I value for the entire sample. Further, there is a scale problem which alters the value of the weight. If the zones are police precincts, for example, then the minimum distance between precincts will be a lot larger than the minimum distance between a smaller type of geographical unit, such as blocks. We need to take into account these different scales.

CrimeStat includes an adjustment for small distances so that the maximum weight can never be greater than 1.0. The adjustment scales distances to one mile, which is a typical distance unit in the measurement of crime incidents. When the small distance

adjustment is turned on, the minimal distance is automatically scaled to be one mile. The formula used is

$$W_{ij} = \frac{\text{one mile}}{\text{one mile} + d_{ij}} \quad (4.32)$$

in whatever units are specified. For example, if the distance units, d_{ij}, are being calculated as feet, then

$$W_{ij} = \frac{5{,}280}{5{,}280 + d_{ij}}$$

where 5,280 is the number of feet in a mile. This has the effect of insuring that the weight of a particular pair of point locations will not have an undue influence on the overall statistic. This is the default condition in *CrimeStat*, but the user can turn it off to obtain a more traditional measure of I (figure 4.31).

Testing the significance of Moran's I

The empirical distribution can be compared with the theoretical distribution by dividing by an estimate of the theoretical standard deviation

$$Z(I) = \frac{I - E(I)}{S_{E(I)}} \quad (4.33)$$

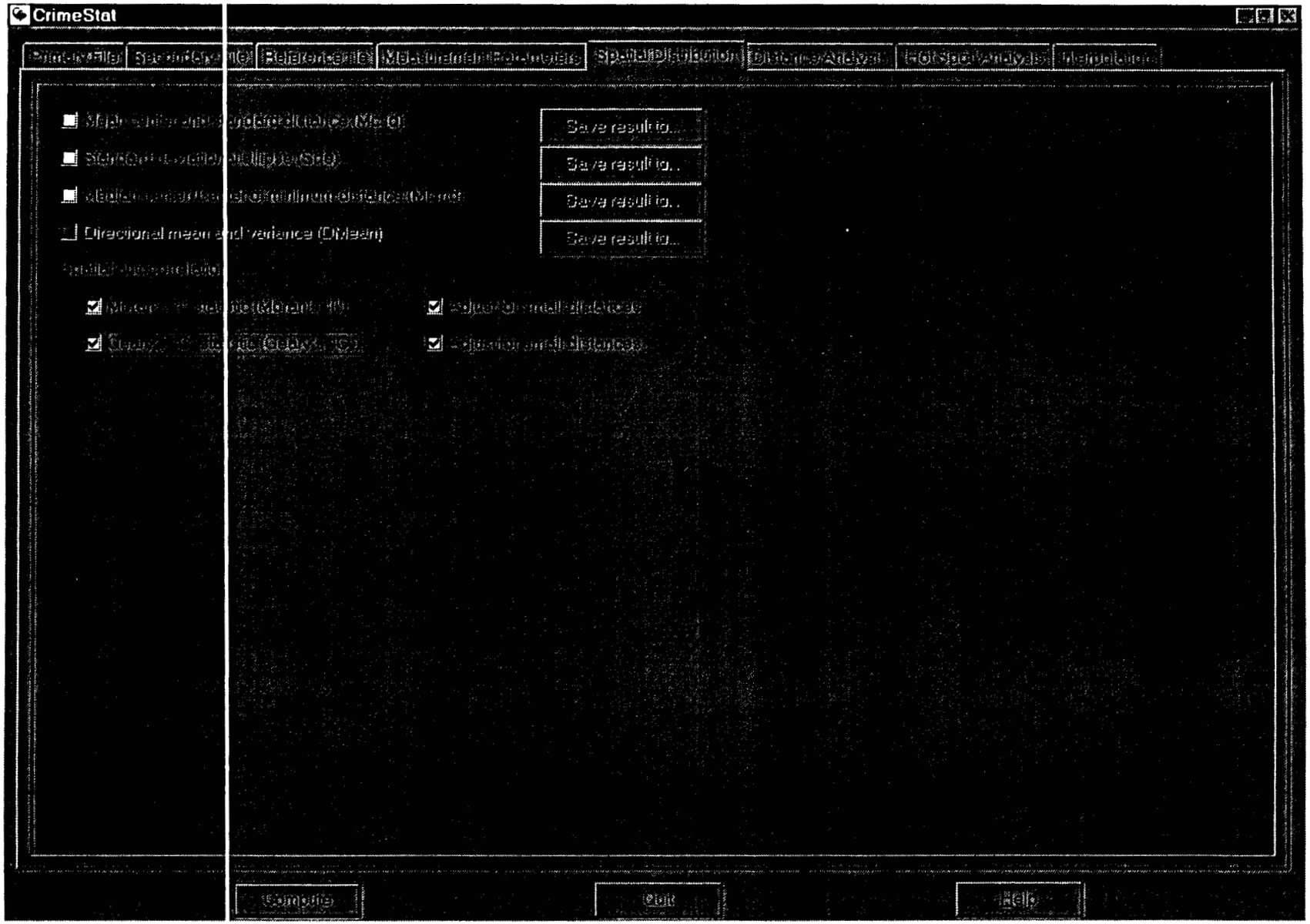
where I is the empirical value calculated from a sample, E(I) is the theoretical mean of a random distribution and $S_{E(I)}$ is the theoretical standard deviation of E(I).

There are several interpretations of the theoretical standard deviation which affect the particular statistic used for the denominator as well as the interpretation of the significance of the statistic (Anselin, 1992). The most common assumption is to assume that the standardized variable, Z(I), has a sampling distribution which follows a standard normal distribution, that is with a mean of 0 and a variance of 1. This is called the *normality* assumption.[7] A second interpretation assumes that each observed value could have occurred at any location, that is the location of the values and their spatial arrangement is assumed to be unrelated. This is called the *randomization* assumption and has a slightly different formula for the theoretical standard deviation of I.[8] *CrimeStat* outputs the Z-values for both the normality and randomization assumptions (figure 4.32).

Example 7: Testing auto thefts with Moran's I

To illustrate the use of Moran's I with point locations requires data to have intensity values associated with each point. Since most crime incidents are represented as

Figure 4.31: Selecting Spatial Autocorrelation Statistics



a single point, they do not naturally have associated intensities. It is necessary, therefore, to adapt crime data to fit the form required by Moran's I. One way to do this is assign crime incidents to geographical areas and count the number of incidents per area. Figure 4.33 shows 1996 motor vehicle thefts in both Baltimore County and Baltimore City by individual blocks. With a GIS program, 14,853 vehicle theft locations were overlaid on top of a map of 13,101 census blocks and the number of motor vehicle thefts within each block were counted and then assigned to the block as a variable. The numbers varied from 0 incidents (for 7,675 blocks) up to 46 incidents (for 1 block). The map shows the plot of the number of auto thefts per block.

Clearly, aggregating incident locations to zones, such as blocks, eliminates some information since all incidents within a block are assigned to a single location (the centroid of the block). The use of Moran's I, however, requires the data to be in this format. Using data in this form, Moran's I was calculated using the default small distance adjustment because many blocks are very close together. *CrimeStat* calculated I as 0.012464 and the theoretical value of I as -0.000076 (minus 1 divided by 13,100). The test of significance using the normality assumption gave a Z-value of 125.13, a highly significant value. Below are the calculations.

$$Z(I) = \frac{I - E(I)}{S_{E(I)}} = \frac{0.012464 - (-0.00076)}{0.000100} = 125.13 \ (p \leq .001)$$

In other words, motor thefts are highly and positively spatially autocorrelated. Blocks with many incidents tend to be located close to blocks which also have many incidents and, conversely, blocks with few or no incidents tend to be located close to blocks which also have few or no incidents.

How does this compare with other distributions? As was argued above, finding positive spatial autocorrelation for auto thefts is not surprising given that there is such a high concentration of population (and, hence, motor vehicles) towards the metropolitan center. To put this in perspective, we ran Moran's I for the population of the blocks (Figure 4.34).[9] With these data, Moran's I for population is 0.001659 with a Z-value of 17.32; the theoretical I is the same since the same number of blocks is being used for the statistic (n=13,101).

Comparing the I value for motor vehicle thefts (0.012464) with that of population (0.00166) suggests that motor vehicle thefts are slightly more concentrated than would be expected on the basis of the population distribution. We can set up an approximate test of this hypothesis. The joint sampling distribution for two variables, such as motor vehicle thefts and population, is not known. However, if we assume that the standard error of the distribution follows a spatially random distribution under the assumption of normality, then equation 4.32 can be applied:

Figure 4.32: Moran's I Statistic Output

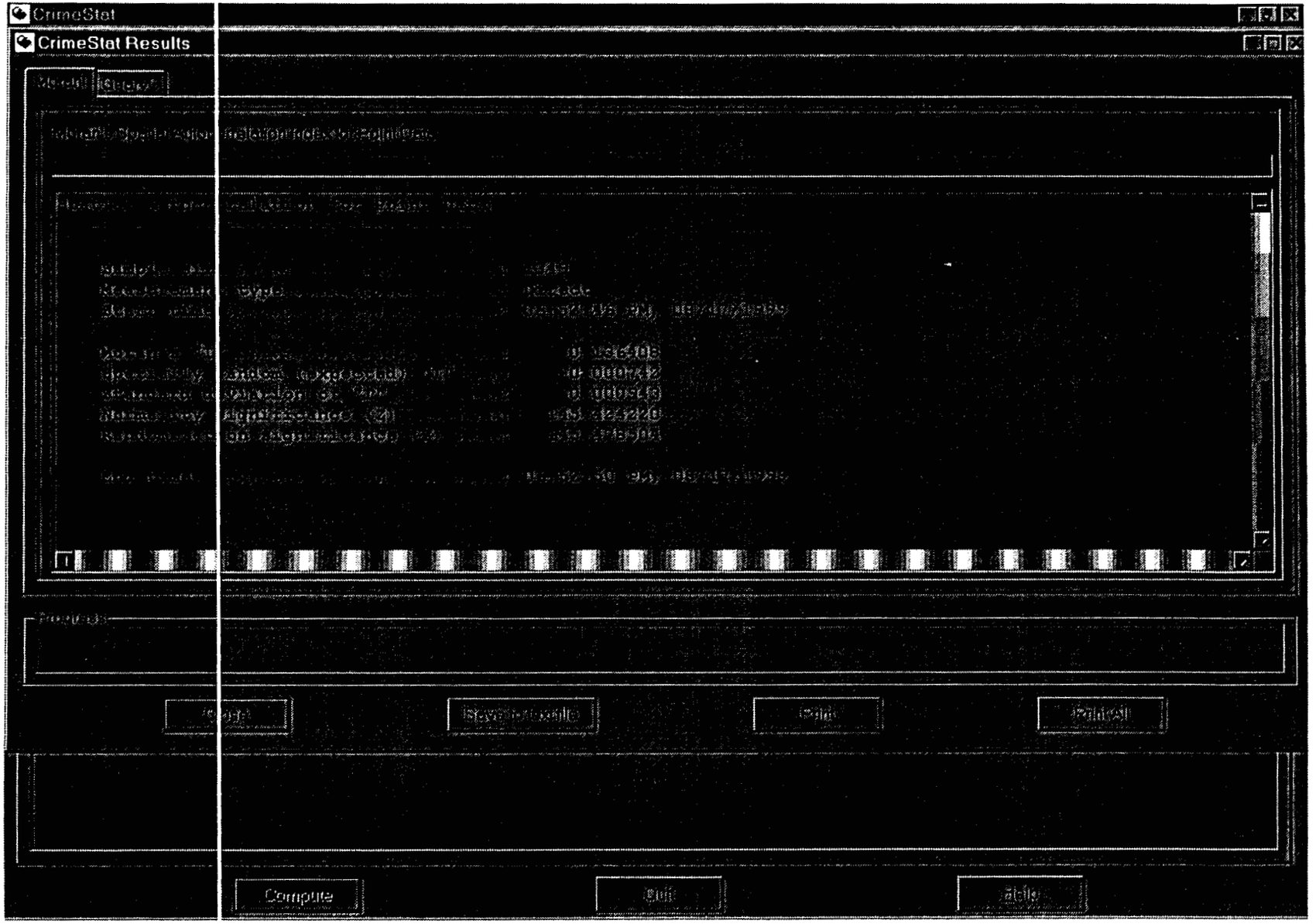
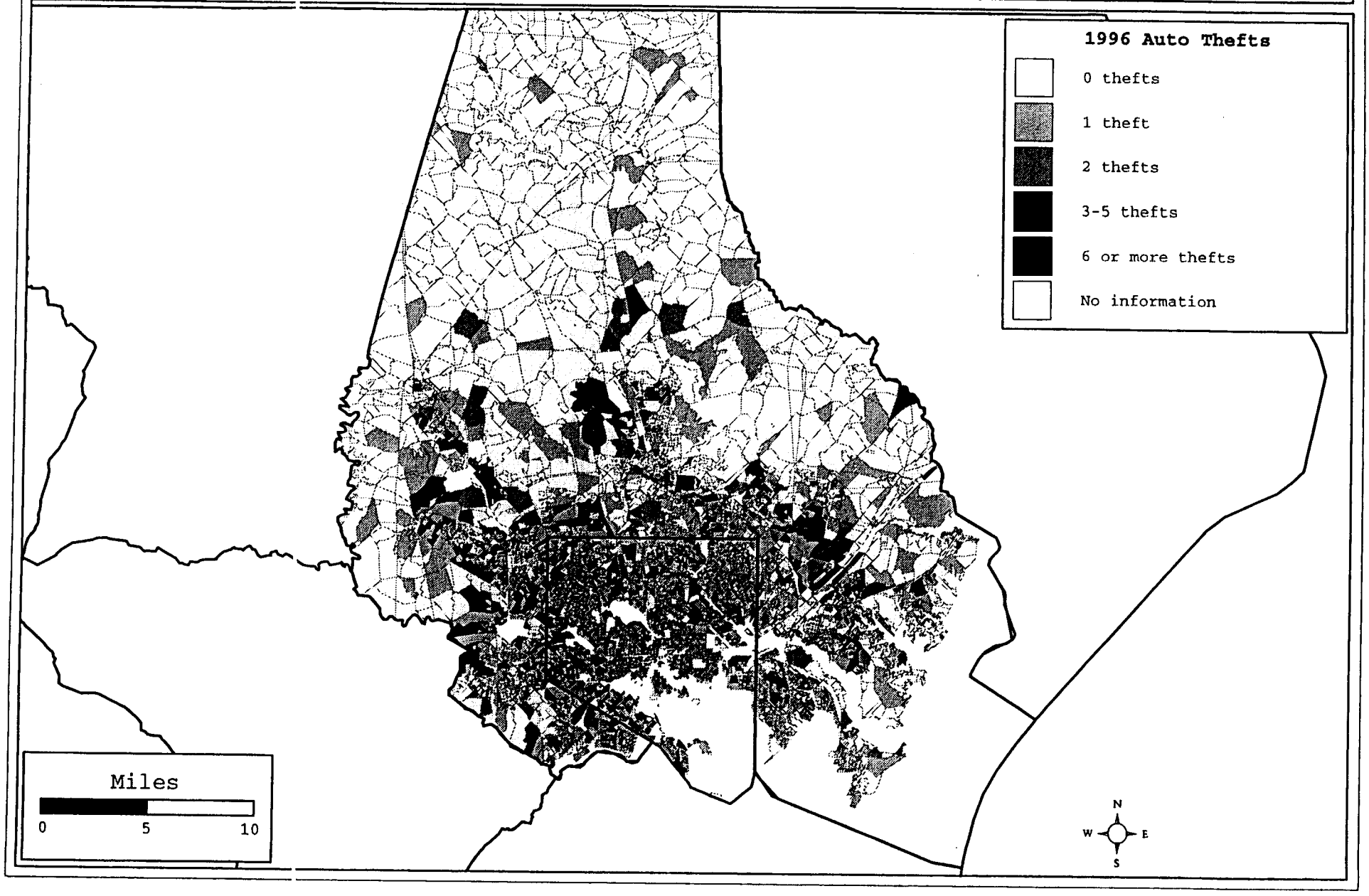


Figure 4.33: 1996 Baltimore Metropolitan Auto Thefts

Number of Auto Thefts Per Block



$$Z(I) = \frac{I_{MV} - I_P}{S_{E(I)}} = \frac{0.012464 - 0.001659}{0.000100} = 108.05 \ (p \leq .001)$$

where I_{MV} is the I value for motor vehicle thefts, I_P is the I value for population, and $S_{E(I)}$ is the standard deviation of I under the assumption of normality. The high Z-value suggests that motor vehicle thefts are much more clustered than the clustering of population. To put it another way, they are more clustered than would be expected from the population distribution. As mentioned, this is an approximate test since the joint distribution of I for two empirical distributions of I is not known.

Geary's C Statistic

Geary's C statistic is similar to Moran's I (Geary, 1954). In this case, however, the interaction is not the cross-product of the deviations from the mean, but the deviations in intensities of each observation location with one another. It is defined as

$$C = \frac{(N-1)\left[\Sigma_i \Sigma_j W_{ij} (X_i - X_j)^2\right]}{2(\Sigma_i \Sigma_j W_{ij}) \Sigma_i (X_i - \bar{X})^2} \qquad (4.34)$$

The values of C typically vary between 0 and 2 although 2 is not a strict upper limit (Griffith, 1987). The theoretical value of C is 1; that is, if values of any one zone are spatially unrelated to any other zone, then the expected value of C would be 1. Values less than 1 (i.e., between 0 and 1) typically indicate positive spatial autocorrelation while values greater than 1 (i.e., between 1 and 2) indicate negative spatial autocorrelation. Thus, this index is inversely related to Moran's I. It will not provide identical inference because it emphasizes the differences in values between pairs of observations comparisons rather than the covariation between the pairs (i.e., product of the deviations from the mean). The Moran coefficient gives a more global indicator whereas the Geary coefficient is more sensitive to differences in small neighborhoods.

Adjustment for small distances

Like Moran's I, the weights are defined as the inverse of the distance between the paired points

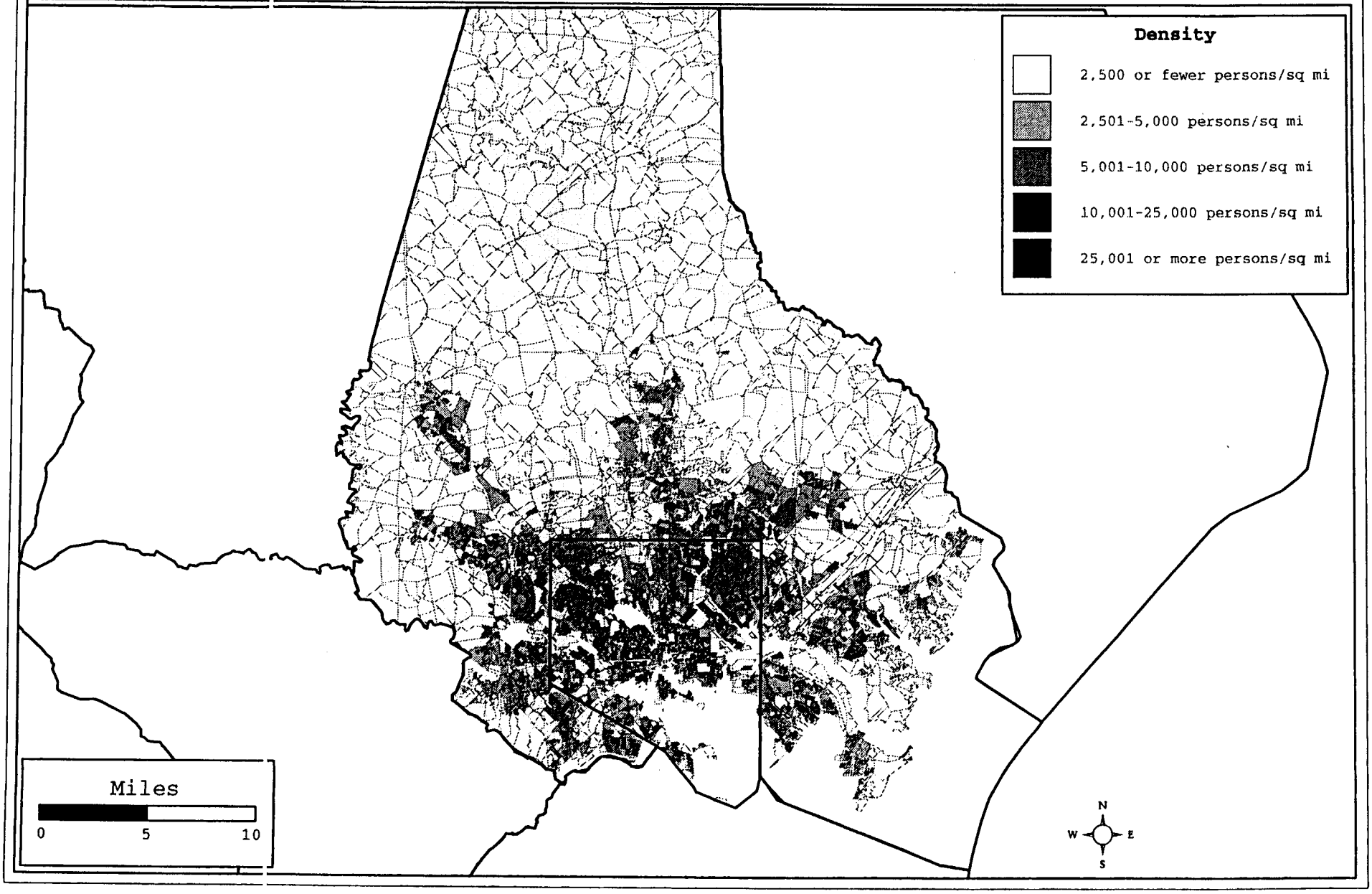
$$W_{ij} = \frac{1}{d_{ij}} \qquad (4.31)$$

repeat

However, the weights will tend to increase substantially as the distance between points decreases. Consequently, a small distance adjustment is allowed which ensures that no weight is greater than 1.0. The adjustment scales the distances to one mile

Figure 4.34: 1990 Baltimore Population Density

Number of Persons Per Square Mile by Block



$$W_{ij} = \frac{\text{one mile}}{\text{one mile} + d_{ij}} \quad \text{repeat} \tag{4.32}$$

in whatever units are specified. This is the default condition although the user can calculate all weights as the reciprocal distance by turning off the small distance adjustment.

Testing the significance of Geary's C

The empirical C distribution can be compared with the theoretical distribution by dividing by an estimate of the theoretical standard deviation

$$Z(C) = \frac{C - E(C)}{S_{E(C)}} \tag{4.35}$$

where C is the empirical value calculated from a sample, E(C) is the theoretical mean of a random distribution and $S_{E(C)}$ is the theoretical standard deviation of E(C). The usual test for C is to assume that the sample Z follows a standard normal distribution with mean of 0 and variance of 1 (normality assumption). *CrimeStat* only calculates the normality assumption though it is possible to calculate the standard error under a randomization assumption (Ripley, 1981).[10] Figure 4.35 illustrates the output.

Example 8: Testing auto thefts with Geary's C

Using the same data on auto thefts for Baltimore County and Baltimore City, the C value for auto thefts was 1.0355 with a Z-value of 10.68 ($p \leq .001$) while that for population was 0.924811 with a Z-value of 122.61 ($p \leq .001$). The C value of motor vehicle thefts is greater than the theoretical C of 1 and suggests *negative* spatial autocorrelation, rather than positive spatial autocorrelation. That is, the index suggests that blocks with a high number of auto thefts are adjacent to blocks with a low number of auto thefts or with low population density. The C value of population, on the other hand, is below the theoretical C of 1 and points to positive spatial autocorrelation.

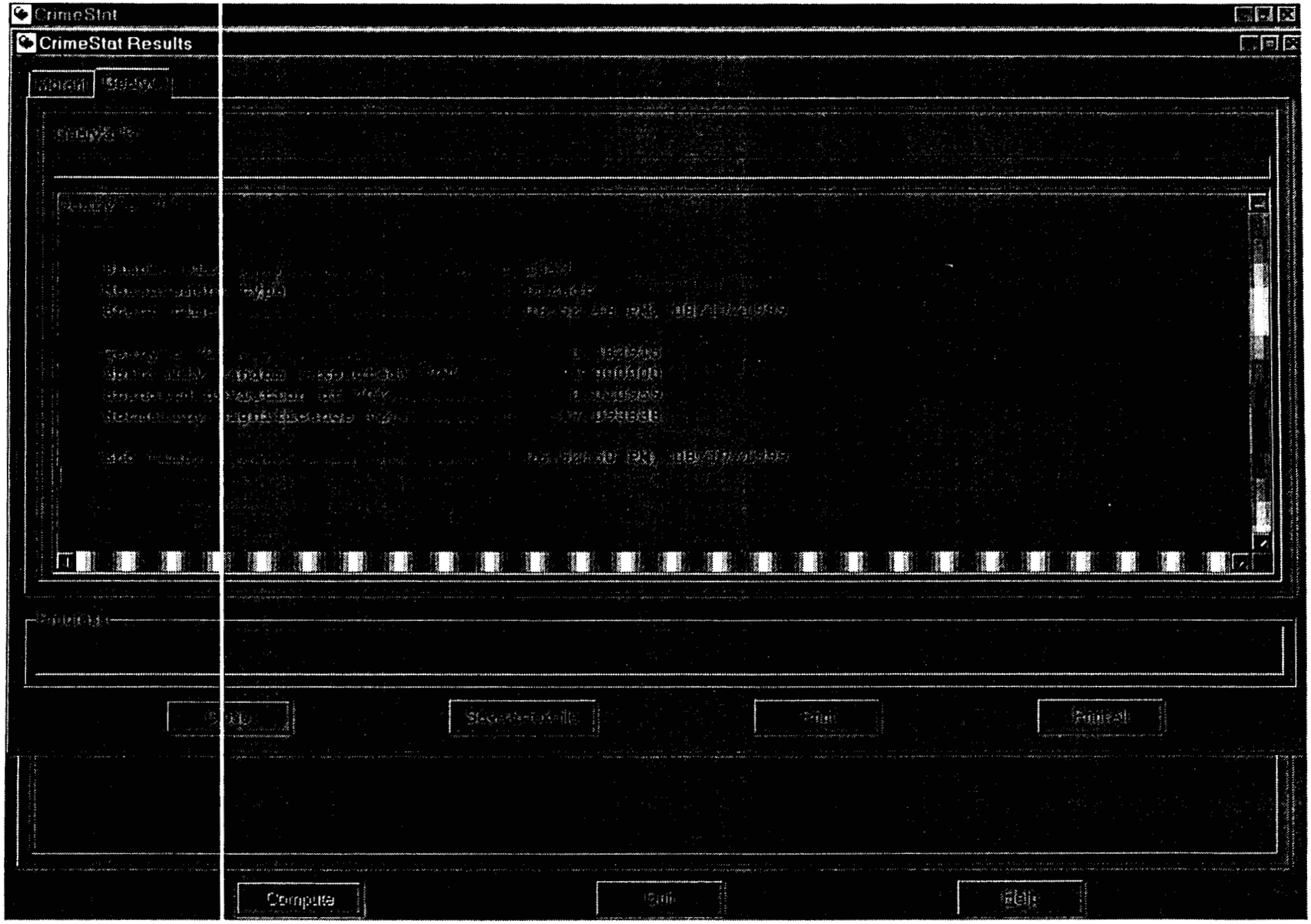
Thus, Geary's C provides a different inference from Moran's I regarding the spatial distribution of the blocks. From the example above, Moran's I indicated positive spatial autocorrelation for both auto thefts and population density. An inspection of figure 4.34 show however, that there are little 'peaks' and 'valleys' among the blocks. Several blocks have a high number of auto thefts, but are surrounded by blocks with a low number of auto thefts.

In other words, the Moran coefficient has indicated that there is more positive spatial autocorrelation for motor vehicle thefts among the 13,101 blocks while the Geary coefficient has emphasized the irregular patterning among the blocks. The Geary index is more sensitive to local clustering (second-order effects) than the Moran index, which is

better seen as measuring first-order spatial autocorrelation. This illustrates how these indices have to be used with care and cannot be generalized by themselves. Each of them emphasizes slightly different information regarding spatial autocorrelation, yet neither is sufficient by itself. They should be used as part of a larger analysis of spatial patterning.[11]

The next chapter will examine tools for measuring *second-order* effects using properties of the distances between incident locations.

Figure 4.35: Geary's C Statistic Output



Endnotes for Chapter 4

1. Hint. There are 40 bars indicated in the status bar while a routine is running. For long runs, users can estimate the calculation time by timing how long it takes for two bars to be displayed and then multiply by 20.

2. *CrimeStat*'s implementation of the Kuhn and Kuenne algorithm is as follows (from Burt and Barber, 1996, 112-113):

   F. Let t be the number of the iteration. For the first iteration only (i.e., t=1) the weighted mean center is taken as the initial estimate of the median location, X_t and Y_t.

   G. Calculate the distance from each point, i, to the current estimate of the median location, d_{ict}, where i is a single point and ct is the current estimate of the median location during iteration t.

      a. If the coordinates are spherical, then Great Circle distances are used.

      b. If the coordinates are projected, then Euclidean distances are used.

   H. Weight each case by a weight, W_i, and calculate

$$K_{it} = W_i \, e^{-d(ict)}$$

   where **e** is the base of the natural logarithm(2.7183..) and $d_{(ict)}$ is an alternative way to write d_{ict}.

      a. If no weights are defined in the primary file, W_i is assumed to be 1.

      b. If weights are defined in the primary file, W_i takes their values.

   Note that as the distance, d_{ict}, approaches 0, then $e^{-d(ict)}$ becomes 1.

   I. Calculate a new estimate of the center of minimum distance from

$$X^{t+1} = \frac{\Sigma K_{it} X_i}{\Sigma K_{it}} \quad \text{for i=1...n}$$

$$Y^{t+1} = \frac{\Sigma K_{it} Y_i}{\Sigma K_{it}} \quad \text{for i=1...n}$$

where X_i and Y_i are the coordinates of point i (either lat/lon for spherical or feet or meters for projected).

J. Check to see how much change has occurred since the last iteration

$$ABS\,|\,X^{t+1} - X^t\,| \leq 0.000001$$

$$ABS\,|\,Y^{t+1} - Y^t\,| \leq 0.000001$$

a. If either the X or Y coordinates have changed by greater than 0.000001 between iterations, substitute X^{t+1} for X^t and Y^{t+1} for Y^t and repeat steps B through D.

b. If *both* the change in X and the change in Y is less than or equal to 0.000001, then the estimated X_t and Y_t coordinates are taken as the center of median distance.

3. Formulas for the new axes provided by Ebdon (1988) and Cromley (1992) yield standard deviational ellipses that are too small, for two different reasons. First, they produce transformed axes that are too small. If the distribution of points is random and even in all directions, ideally the standard deviational ellipse should be equal to the standard distance deviation, since $S_x = S_y$. The formula used here has this property. Since the formula for the standard distance deviation is (4.5):

$$SDD = SQRT\left[\frac{\Sigma(X_i - \bar{X})^2 + \Sigma(Y_i - \bar{Y})^2}{N-2}\right]$$

If $S_x = S_y$, then $\Sigma(X_i - \bar{X})^2 = \Sigma(Y_i - \bar{Y})^2$, therefore

$$SDD = SQRT\left[2^* \frac{\Sigma(X_i - \bar{X})^2}{N-2}\right]$$

Similarly, the formula for the transformed axes are (4.8a, 4.8b):

$$S_x = SQRT\left[2^* \frac{\Sigma\{(X_i - \bar{X})\,Cos\theta - \Sigma(Y_i - \bar{Y})\,Sin\theta\}^2}{N-2}\right]$$

$$S_y = SQRT\left[2^* \frac{\Sigma\{(X_i - \bar{X})\,Sin\theta - \Sigma(Y_i - \bar{Y})\,Cos\theta\}^2}{N-2}\right]$$

However, if $S_x = S_y$, then $\theta = 0$, $Cos0 = 1$, $Sin0 = 0$ and, therefore,

$$S_x = S_y = \text{SQRT}\left[2^* \frac{\Sigma(X_i - \bar{X})}{N-2} \right]$$

which is the same as for the standard distance deviation (SDD) under the same conditions. The formulas used by Ebdon (1988) and Cromley (1992) produce axes which are SQRT(2) times too small.

The second problem with the Ebdon and Cromley formulas is that they do not correct for degrees of freedom and, hence, produce too small a standard deviational ellipse. Since there are two constants in each equation, MeanX and MeanY, then there are only N-2 degrees of freedom. The cumulative effect of using transformed axes that are too small and not correcting for degrees of freedom yields a much smaller ellipse than that used here.

4. In *MapInfo*, the command is *T*able *I*mport *<Mapinfo interchange file>*. With *Atlas*GIS*, the command is *F*ile *O*pen *<boundary (*.bna) file>*. With the DOS version of *Atlas*GIS*, the *Atlas Import-Export* program has to be used to convert the '.bna' output file to an *Atlas*GIS* '.agf' file.

5. There are limits to the Bonferoni logic. For example, if there were 10 tests, having a threshold significance level of .005 (.05 / 10) for the 'either/or' conditions and a threshold significance level of .50 (.05 * 10) for the 'both/and' would lead to an excessively difficult test in the first case and a much too easy test in the second. Thus, the Bonferoni logic should be applied to only a few tests (e.g., 5 or fewer).

6. From trigonometry,

$$\text{Tan}\bar{\theta} = \frac{\text{Resultant of all Y values}}{\text{Resultant of all X values}} = \frac{\Sigma \sin \theta_i}{\Sigma \cos \theta_i}$$

Hence, the Arctangent of the ratio of the resultant of Y divided by the resultant of X is the direction from the origin to the mean center (Gaile and Burt, 1980).

7. The theoretical standard deviation of I under the assumption of normality is (from Ebdon, 1985):

$$S_{E(I)} = \text{SQRT}\left[\frac{N^2 \Sigma_{ij} w_{ij}^2 + 3(\Sigma_{ij} w_{ij})^2 - N \Sigma_i (\Sigma_j w_{ij})^2}{(N^2 - 1)(\Sigma_{ij} w_{ij})^2} \right]$$

8. The formula for the theoretical standard deviation of I under the randomization assumption is (from Ebdon, 1985):

$$S_{E(I)} = \mathrm{SQRT}\left[\frac{N\{(N^2+3-3N)\Sigma_{ij} w_{ij}^2+3(\Sigma_{ij} w_{ij})^2-N\Sigma_i(\Sigma_j w_{ij})^2\}-k((N^2-N)\Sigma_{ij} w_{ij}^2+6(\Sigma_{ij} w_{ij})^2-2N(\Sigma_i(\Sigma_j w_{ij})^2)}{(N-1)(N-2)(N-3)(\Sigma_{ij} w_{ij})^2}\right]$$

9. We could have compared Moran's I for auto thefts with that of population, rather than population density. However, since the areas of blocks tend to get larger the farther the distance from the metropolitan center, the effect of testing only population is partly being minimized by the changing sizes of the blocks. Consequently, population density was used to provide a more accurate measure of population concentration. In any case, Moran's I for population is also highly significant: I = 0.00166 (Z=17.32).

10. The theoretical standard deviation for C under the normality assumption is (from Ripley, 1981):

$$S_{E(I)} = \mathrm{SQRT}\left[\frac{(2\,\Sigma_{ij} w_{ij}^2+\Sigma_i(\Sigma_j w_{ij})^2)(N-1) - 4(\Sigma_{ij} w_{ij})^2}{2(N+1)(\Sigma_{ij} w_{ij})^2}\right]$$

11. Anselin (1992) points out that the results of the two indices are determined to a large extent by the type of weighting used. In the original formulation, where adjacent weights of 1 and 0 are used, the two indices are linearly related, though moving in opposite directions (Griffith, 1987). Thus, only adjacent zones have any impact on the index. With inverse distance weights, however, zones farther removed can influence the overall index so it is possible to have a situation whereby adjacent zones have similar values (hence, are positively autocorrelated) whereas zones farther away could have dissimilar values (hence, are negatively autocorrelated).

Chapter 5
Distance Analysis

In this chapter, tools that identify characteristics of the distances between points will be described. The previous chapter provided tools for describing the general spatial distribution of crime incidents or *first-order* properties of the incident distribution (Bailey and Gattrell, 1995). First-order properties are global because they represent the dominant pattern of distribution - where it is centered, how far it spreads out, and whether there is any orientation or direction to its dispersion. *Second-order* (or *local*) properties, on the other hand, refer to sub-regional patterns or 'neighborhood' patterns within the overall distribution. If there are distinct 'hot spots' where many crime incidents cluster together, their distribution is spatially related not so much to the overall global pattern as to something unique in the sub-region or neighborhood. Thus, second-order characteristics tell something about particular environments that may concentrate crime incidents. Figure 5.1 shows the distance analysis tab and distance statistics that are calculated by *CrimeStat.*

Nearest Neighbor Index (Nna)

One of the oldest distance statistics is the *nearest neighbor index.* It is particularly useful because it is a simple tool to understand and to calculate. It was developed by two botanists in the 1950s (Clark and Evans, 1954), primarily for field work, but it has been used in many different fields for a wide variety of problems (Cressie, 1991). It has also become the basis of many other types of distance statistics, some of which are implemented in *CrimeStat.*

The nearest neighbor index compares the distances between nearest points and distances that would be expected on the basis of chance. It is an index that is the ratio of two summary measures. First, there is the *nearest neighbor distance.* For each point (or incident location) in turn, the distance to the closest other point (nearest neighbor) is calculated and averaged over all points.

$$\text{Nearest Neighbor Distance} = d(NN) = \sum_{i=1}^{N} \left[\frac{\text{Min}(d_{ij})}{N} \right] \qquad (5.1)$$

where $\text{Min}(d_{ij})$ is the distance between each point and its nearest neighbor and N is the number of points in the distribution. Thus, in *CrimeStat*, the distance from a single point to every other point is calculated and the smallest distance (the minimum) is selected. Then, the next point is taken and the distance to all other points (including the first point measured) is calculated with the nearest being selected and added to the first minimum distance. This process is repeated until all points have had their nearest neighbor selected. The total sum of the minimum distances is then divided by N, the sample size, to produce an average minimum distance.

Figure 5.1: Distance Analysis Layout

The second summary measure is the expected nearest neighbor distance if the distribution of points is completely random. This is the *mean random distance* (or the mean random nearest neighbor distance). It is defined as

$$\text{Mean Random Distance} = d(ran) = 0.5 \text{ SQRT } [\frac{A}{N}] \quad (5.2)$$

where A is the area of the region and N is the number of incidents. Since A is defined by the square of the unit of measurement (e.g., square mile, square meters, etc.), it yields a random distance measure in the same units (i.e., miles, meters, etc.).[1] If defined on the measurement parameters page by the user, *CrimeStat* will use the specified area in calculating the mean random distance. If no area measurement is provided, *CrimeStat* will take the rectangle defined by the minimum and maximum X and Y points.

The nearest neighbor index is the ratio of the observed nearest neighbor distance to the mean random distance

$$\text{Nearest Neighbor Index} = NNI = \frac{d(NN)}{d(ran)} \quad (5.3)$$

Thus, the index compares the average distance from the closest neighbor to each point with a distance that would be expected on the basis of chance. If the observed average distance is about the same as the mean random distance, then the ratio will be about 1.0. On the other hand, if the observed average distance is smaller than the mean random distance, that is, points are actually closer together than would be expected on the basis of chance, then the nearest neighbor index will be less than 1.0. This is evidence for clustering. Conversely, if the observed average distance is greater than the mean random distance, then the index will be greater than 1.0. This would be evidence for dispersion, that points are more widely dispersed than would be expected on the basis of chance.

Testing the Significance of the Nearest Neighbor Index

Some differences from 1.0 in the nearest neighbor index would be expected by chance. Clark and Evans (1954) proposed a Z-test to indicate whether the observed average nearest neighbor distance was significantly different from the mean random distance (Hammond and McCullagh, 1978; Ripley, 1981). The test is between the observed nearest neighbor distance and that expected from a random distribution and is given by

$$Z = \frac{d(NN) - d(ran)}{SE_{d(ran)}} \quad (5.4)$$

where the standard error of the mean random distance is approximately given by:

$$SE_{d(ran)} \approx SQRT\left[\frac{(4-\pi)\,A}{4\pi\,N^2}\right] \approx \frac{0.26136}{SQRT[\,N^2/A\,]} \qquad (5.5)$$

with A being the area of region and N the number of points. There have been other suggested tests for the nearest neighbor distance as well as corrections for edge effects (see below). However, equations 5.4 and 5.5 are used most frequently to test the average nearest neighbor distance. See Cressie (1991) for details of other tests.

Calculating the statistics

Once nearest neighbor analysis has been selected, the user clicks on *Compute* to run the routine. The program outputs 10 statistics:

1. The sample size
2. The mean nearest neighbor distance in meters, feet and miles
3. The standard deviation of the nearest neighbor distance in meters, feet and miles
4. The minimum distance in meters, feet and miles
5. The maximum distance in meters, feet and miles
6. The mean random distance (for both the bounding rectangle and the user input area, if provided)
7. The mean dispersed distance in meters, feet and miles (for both the bounding rectangle and the user input area, if provided)
8. The nearest neighbor index (for both the bounding rectangle and the user input area, if provided)
9. The standard error of the nearest neighbor index (for both the maximum bounding rectangle and the user input area, if provided)
10. A significance test of the nearest neighbor index (Z-test)

In addition, the output can be saved to a '.dbf' file, which can then be imported into spreadsheet or graphics programs.

Example 1: The nearest neighbor index for street robberies

In 1996, there were 1181 street robberies in Baltimore County. The area of the County is about 607 square miles and is specified on the measurement parameters page. *CrimeStat* returns the statistics shown in Table 5.1 with the NNA routine.

CrimeStat does not provide the significance level of the test, but only the Z-value. However, the significance level of the Z-value can be found in any table of standard normal deviants. In this case, a Z-value of -44.4672 is highly significant ($p \leq .001$). In other words, the distribution of the nearest neighbors of street robberies in Baltimore County is significantly smaller than the expected distribution of nearest neighbors.

Table 5.1
Nearest Neighbor Statistics for
1996 Street Robberies in Baltimore County
N=1181

Mean nearest neighbor distance:	0.11598 mi
Mean random distance based on user input area:	0.35837 mi
Nearest neighbor index:	0.3236
Standard error:	0.00545 mi
Test Statistic (Z):	-44.4672

Nearest Neighbors Analysis is Not a Test for Complete Spatial Randomness

It should be noted that the significance test for the nearest neighbor index is not a test for complete spatial randomness, for which it is sometimes mistaken. It is only a test whether the average nearest neighbor distance is significantly different than what would be expected on the basis of chance. In other words, it is a test of *first-order* nearest neighbor randomness.[2] There are also second-order, third-order, and so forth distributions that may or may not be significantly different from complete spatial randomness. A complete test would have to test for all those effects, what are called *K-order* effects.

Edge Effects

There are also edge effects that can bias the nearest neighbor index. An incident occurring near the border of the study area may actually have its nearest neighbor on the other side of the border. However, since there are usually no data on the distribution of incidents outside the study area, the program selects another point within the study area as the nearest neighbor of the border point. Thus, there is the potential for exaggerating the nearest neighbor distance, that is, the observed nearest neighbor distance is probably greater than what it should be. This potential bias has been indicated on the distance analysis page by specifying 'With no border correction' next to the selection. There have been various suggestions for correcting the border or edge effect of the nearest neighbor index (see Cressie, 1991 for details). However, a consensus has not been established for handling this and most researchers test the index with all its potential biases. In this version of *CrimeStat*, all nearest neighbor tests are uncorrected for edge effects.[3]

Nevertheless, since the effect of the edge is to exaggerate the nearest neighbor distance and since *most* nearest neighbor tests of social characteristics show clustering, rather than dispersion (i.e., the nearest neighbor index is usually less than 1.0), the test can be considered conservative if significance is obtained. That is, in spite of a potential bias which would make the nearest neighbor distance larger than it should be, the test has revealed that it is significantly smaller than what would be expected on the basis of chance. This is important if decisions have to be made on the basis of the statistics, for example in placing a police car at a particular location to minimize response time.

Example 2: The nearest neighbor index for residential burglaries

The nearest neighbor index and test can be very useful for understanding the degree of clustering of crime incidents in spite of its limitations. For example, in Baltimore County, the distribution of 6051 residential burglaries in 1996 yields the following nearest neighbor statistics (Table 5.2):

Table 5.2
Nearest Neighbor Statistics for
1996 Residential Burglaries in Baltimore County
N=6051

Mean nearest neighbor distance:	0.07134 mi
Mean random distance based on user input area:	0.16761 mi
Nearest neighbor index:	0.4256
Standard error:	0.00113 mi
Test Statistic (Z):	-85.4750

The distribution of residential burglaries is also highly significant. Now, suppose we want to compare the distribution of street robberies (table 5.1) with that residential burglaries (table 5.2). The significance test is not very useful for the comparison because the sample sizes are so large (1181 v. 6051); the much higher Z-value for residential burglaries indicates primarily that there was a larger sample size to test it. However, comparing the relative nearest neighbor indices can be meaningful.

$$\text{Relative Nearest Neighbor Comparison} = \frac{NNI(A)}{NNI(B)} \qquad (5.6)$$

where NNI(A) is the nearest neighbor index for one group (A) and NNI(B) is the nearest neighbor index for another group (B). Thus, comparing street robberies with residential burglaries, we have

$$\frac{NNI\ (A)}{NNI\ (B)} = \frac{NNI\ (robberies)}{NNI\ (burglaries)} = \frac{0.3057}{0.4256} = 0.7182$$

In other words, the distribution of street robberies relative to an expected random distribution appears to be more concentrated than that of burglaries relative to an expected random distribution. There is no simple significance test of this comparison since the standard error of the joint distributions is not known. But the relative index suggests that robberies are more concentrated than burglaries and, hence, are more likely to have 'hot spot' or 'hot zones' where they are particularly concentrated. This index, of course, does not prove that there are 'hot spots', but only points us towards the higher

concentration of robberies relative to burglaries. In the previous chapter, it was shown that robberies had a smaller dispersion than burglaries. Here, however, the analysis is taken a step further to suggest that robberies are more concentrated than burglaries.

K-Order Nearest Neighbors

As mentioned above, the nearest neighbor index is only an indicator of first-order randomness. It compares the average distance for the nearest neighbor to an expected random distance. But what about the second nearest neighbor? Or the third nearest neighbor? Or the K^{th} nearest neighbor? *CrimeStat* constructs K-order nearest neighbor indices. On the distance analysis page, the user can specify the number of nearest neighbor indices to be calculated.

The K-order nearest neighbor routine returns four columns:

1. The order, starting from 1
2. The mean nearest neighbor distance for each order (in meters)
3. The expected nearest neighbor distance for each order (in meters)
4. The nearest neighbor index for each order

For each order, *CrimeStat* calculates the K^{th} nearest neighbor distance for each observation and then takes the average. The expected nearest neighbor distance for each order is calculated by:

$$\text{Mean Random Distance to } K^{th} \text{ nearest neighbor} = d(K_{ran}) = \frac{K\,(2K)!}{(2^K K!)^2\ \text{SQRT}\ [N/A]} \qquad (5.7)$$

where K is the order and ! is the factorial operation (e.g., 4! = 4 x 3 x 2 x 1; Thompson, 1956). The K^{th} nearest neighbor index is the ratio of the observed K^{th} nearest neighbor distance to the K^{th} mean random distance. There is not a good significance test for the K^{th} nearest neighbor index due to the non-independence of the different orders, though there have been attempts (see examples in Getis and Boots, 1978; Aplin, 1983). Consequently, *CrimeStat* does not provide a test of significance.

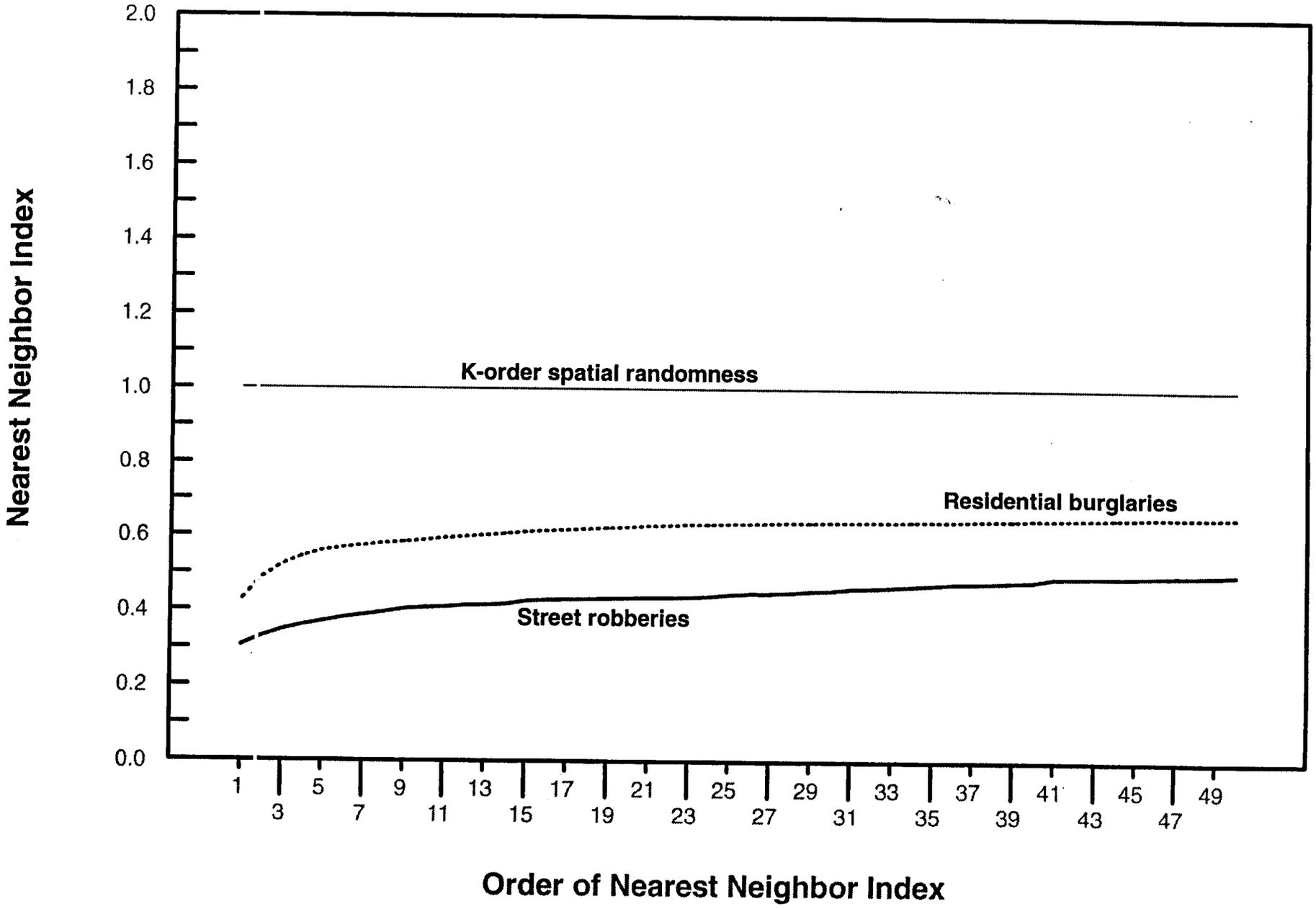
There are no restrictions on the number of nearest neighbors that can be calculated. However, since the average distance increases with higher-order nearest neighbors, the potential for bias from edge effects will also increase. It is suggested that not more than 100 nearest neighbors be calculated.[4]

Nevertheless, the K-order nearest neighbor distance and index can be useful for understanding the overall spatial distributions. Figure 5.2 compares the K-order nearest neighbor index for street robberies with that of residential burglaries. The output was saved as a '.dbf' and was then imported into a graphics program. The graph shows the nearest neighbor indices for both robberies and burglaries up to the 50^{th} order (i.e., the 50^{th} nearest neighbor). The nearest neighbor index is scaled from 0 (extreme clustering) up to 1

Figure 5.2

K-Order Nearest Neighbor Indices

1996 Street Robberies and Residential Burglaries



(extreme dispersion). Since a nearest neighbor index of 1 is expected under randomness, the thin straight line at 1.0 indicates the expected K-order index. As can be seen, both street robberies and residential burglaries are much more concentrated than K-order spatial randomness. Further, robberies are more concentrated than even burglaries for each of the 50 nearest neighbors. Thus, the graph reinforces the analysis above that robberies are more concentrated than burglaries, and both are more concentrated than a random distribution.

In other words, even though there is not a good significance test for the K-order nearest neighbor index, a graph of the K-order indices (or the K-order distances) can give a picture of how clustered the distribution is as well as allow comparisons in clustering between the different types of crimes (or the same crime at two different time periods).

Linear Nearest Neighbor Index (Lnna)

The *linear nearest neighbor index* is a variation on the nearest neighbor routine, but one applied to a street network. All distances along this network are assumed to travel along a grid, hence indirect distances are used. Whereas the nearest neighbor routine calculates the distance between each point and its nearest neighbor using direct distances, the linear nearest neighbor routine uses indirect ('Manhattan') distances (see chapter 3). Similarly, whereas the nearest neighbor routine calculates the expected distance between neighbors in a random distribution of N points using the geographical area of the study region, the linear nearest neighbor routine uses the total length of the street network.

The theory of linear nearest neighbors comes from Hammond and McCullagh (1978). The observed linear nearest neighbor distance, Ld(NN), is calculated by *CrimeStat* as the average of indirect distances between each point and its nearest neighbor. The expected linear nearest neighbor distance is given by

$$Ld(ran) = 0.5 \left[\frac{L}{N - 1}\right] \tag{5.8}$$

where L is the total length of street network and N is the sample size (Hammond and McCullagh, 1978, 279). Consequently, the linear nearest neighbor index is defined as

$$\text{Linear Nearest Neighbor Index} = LNNI = \frac{Ld(NN)}{Ld(ran)} \tag{5.9}$$

Testing the Significance of the Linear Nearest Neighbor Index

Since the theoretical standard error for the random linear nearest neighbor distance is not known, the author has constructed an approximate standard deviation for the observed linear nearest neighbor distance:

$$S_{Ld(NN)} \approx SQRT\left[\frac{\Sigma\,(\,Min(d_{ij}) - Ld(NN)\,)^2}{N - 1}\right] \quad (5.10)$$

where $Min(d_{ij})$ is the nearest neighbor distance for point i and Ld(NN) is the average linear nearest neighbor distance. This is the standard deviation of the linear nearest neighbor distances. The standard error is calculated by

$$SE_{Ld(NN)} = \frac{S_{Ld(NN)}}{SQRT[N]} \quad (5.11)$$

An approximate significance test can be obtained by

$$t = \frac{Ld(NN) - Ld(ran)}{SE_{Ld(NN)}} \quad (5.12)$$

where Ld(NN) is the average linear nearest neighbor distance, Ld(ran) is the expected linear nearest neighbor distance (equation 5.8), and $SE_{Ld(NN)}$ is the approximate standard error of the linear nearest neighbor distance (equation 5.11). Since the empirical standard deviation of the linear nearest neighbor is being used instead of a theoretical value, the test is a *t-test* rather than a Z-test.

Calculating the statistics

On the measurements parameters page, there are two parameters that are input, the geographical area of the study region and the length of street network. At the bottom of the page, the user must select which type of distance measurement to use, direct or indirect. If the measurement type is direct, then the nearest neighbor routine returns the standard nearest neighbor analysis (sometimes called *areal* nearest neighbor). On the other hand, if the measurement type is indirect, then the routine returns the linear nearest neighbor analysis. To calculate the linear nearest neighbor index, therefore, distance measurement must be specified as indirect and the length of the street network must be defined.

Once nearest neighbor analysis has been selected, the user clicks on *Compute* to run the routine. The *Lnna* routine outputs 9 statistics:

1. The sample size
2. The mean linear nearest neighbor distance in meters, feet and miles
3. The minimum linear distance between nearest neighbors in meters, feet and miles
4. The maximum linear distance between nearest neighbors in meters, feet and miles
5. The mean linear random distance

6. The linear nearest neighbor index
7. The standard deviation of the linear nearest neighbor distance
8. The standard error of the linear nearest neighbor distance
9. A significance test of the nearest neighbor index (t-test)

Example 3: Auto thefts along two highways

The linear nearest neighbor index is useful for analyzing the distribution of crime incidents along particular streets. For example, in Baltimore County, state highway 26 in the western part and state highway 150 in the eastern part have particularly high concentrations of motor vehicle thefts (figure 5.3). In 1996, there were 87 vehicle thefts on highway 26 and 47 on highway 150. A GIS can be used with the linear nearest neighbor index to indicate whether these incidents are greater than what would be expected on the basis of chance.

Table 5.3 presents the data. Using the GIS, we estimate that there are 3,333.54 miles of roadway segments; this number was estimated by adding up the total length of the street network in the GIS. Of all the road segments in Baltimore County, there are 241.04 miles of major arterial roads of which state highway 26 has a total length of 10.42 miles and state highway 150 has a total road length of 7.79 miles.

In 1996, there were 3,774 motor vehicle thefts in the county. If these thefts were distributed randomly, then the random expected distance between incidents would be 0.44 miles (equation 5.8). Using this estimate, table 5.3 shows the number of incidents that would be expected on each of the two state highways if the distribution were random and the ratio of the actual number of motor vehicle thefts to the expected number. As can be seen, the distribution of motor vehicle thefts is not random. On all major arterial roads, there are 2.2 times as many thefts as would be expected by a random spatial distribution. In fact, in 1996, of 28,551 road segments in Baltimore County, only 7791 (27%) had one or more motor vehicle thefts occur on them; most of these are major roads. Further, on highway 26 there were 7.4 times as much and on highway 150 there were 5.3 times as much as would be expected if the distribution was random. Clearly, these two highways had more than their share of auto thefts in 1996.

But what about the distribution of the incidents *along* each of these highways? If there were any pattern, for example, most of the incidents clustering on the western edge or in the center, then police could use that information to more efficiently deploy vehicles to respond quickly to events. On the other hand, if the distribution along these highways were no different than a random distribution, then police vehicles must be positioned in the middle, since that would minimize the distance to all occurring incidents.

Unfortunately, the results appear to be close to a random distribution. *CrimeStat* calculates that for highway 26, the average linear nearest neighbor distance is 0.05 miles which is close to the average random linear nearest neighbor distance (0.06 miles). The ratio - the linear nearest neighbor index, is 0.96 with a t-value of -0.16, which is not significantly different from chance. Similarly, for highway 150, the average linear nearest

Figure 5.3: 1996 Auto Thefts in Baltimore County

Incident Distribution on State Highways 26 and 150

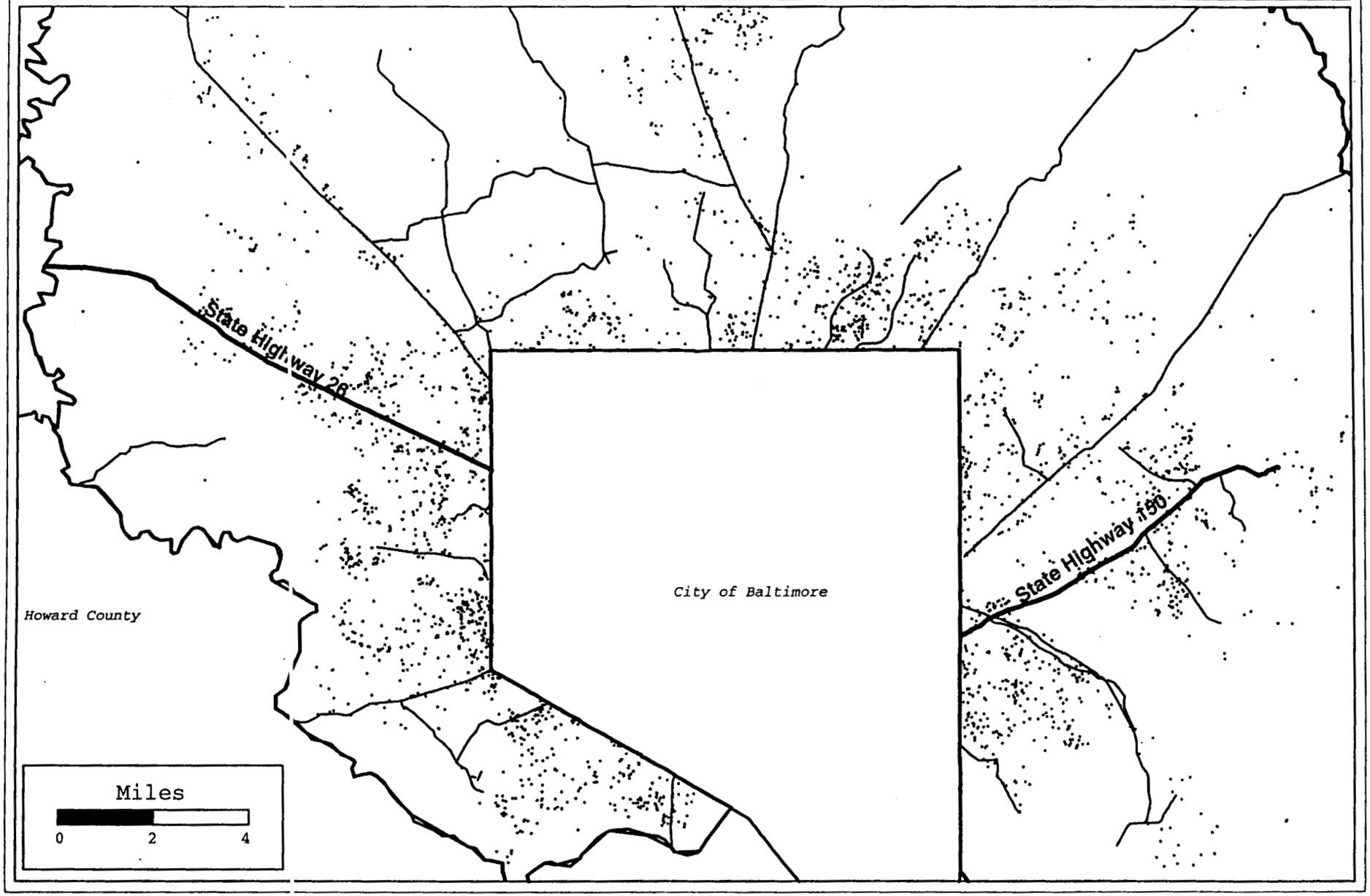


Table 5.3

Comparison of 1996 Baltimore County Auto Thefts for Different Types of Roads
(N = 3774 Incidents)

Length of Road Segments:

Highway 26	10.42 mi
Highway 150	7.79 mi
All Major Arterials	241.04 mi
All Roads	3333.54 mi

Random Expected Distance Between Incidents = 0.44 miles

		Proportional To Network		Proportional to Same Road		
Where Incidents Occurred	Number of Incidents	Expected Number *If* Random	*"Relative to Random"* Ratio of Frequency	Average Linear Nearest Neighbor Distance	Average Random Linear Nearest Neighbor Distance	*"Relative to Itself"* Linear Nearest Neighbor Index
Highway 26	87	11.8	7.4	0.05 mi	0.06	0.96
Highway 150	47	8.8	5.3	0.08 mi	0.08	0.94
All Major Arterials	607	272.8	2.2	0.13 mi	0.20	0.64 ($p \leq .001$)
All Roads	3774	3774.0	1.0	0.09 mi	044	0.21 ($p \leq .001$)

neighbor distance is 0.079 miles which, again, is almost identical to the average random linear nearest neighbor distance (0.084 miles); the nearest neighbor index is 0.94 and the t-value is -0.41 (not significant). In short, even though there was a higher concentration of vehicle thefts on these two state highways than would be expected on the basis of chance, the distribution *along* each highway is not very different than what would be expected on the basis of chance.[5]

K-Order Linear Nearest Neighbors

There is also a K-order linear nearest neighbor analysis, as with the areal nearest neighbors. The user can specify how many additional nearest neighbors are to be calculated. The linear K-order nearest neighbor routine returns four columns:

1. The order, starting from 1
2. The mean linear nearest neighbor distance for each order (in meters)
3. The expected linear nearest neighbor distance for each order (in meters)
4. The linear nearest neighbor index for each order

Since the expected linear nearest neighbor distance has not been worked out for orders higher than one, the calculation produced here is a rough approximation. It applies equation 5.8 only adjusting for the decreasing sample size, N_k, which occurs as degrees of freedom are lost for each successive order. In this sense, the index is really the k-order linear nearest neighbor distance relative to the expected linear neighbor distance for the first order. It is not a strict nearest neighbor index for orders above one.

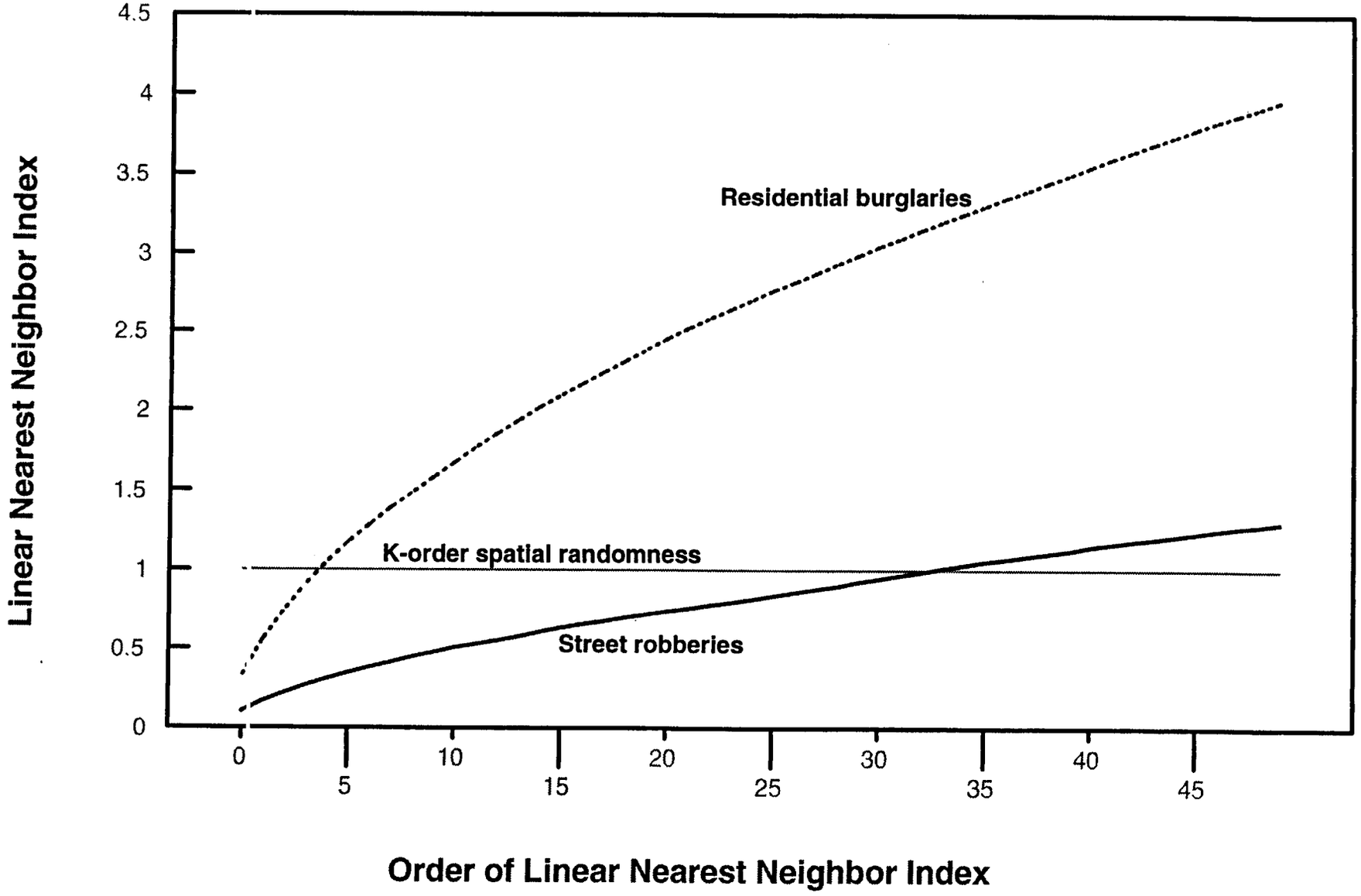
Nevertheless, like the areal k-order nearest neighbor index, the k-order linear nearest neighbor index can provide insights into the distribution of the points, even if the first-order is random. Figure 5.4 shows a graph of 50 linear nearest neighbors for 1996 residential burglaries and street robberies for Baltimore County. As with the areal k-order nearest neighbors (see figure 5.2) both burglaries and robberies show evidence of clustering. For both, the first nearest neighbors are closer together than a random distribution. Similarly, over the 50 orders, street robberies are more clustered than burglaries. However, measuring distance on a grid shows that for burglaries, there is only a small amount of clustering. After the fourth order neighbor, the distribution for burglaries is more dispersed than a random distribution. An interpretation of this is that there are small number of burglaries which are clustered, but the clusters are relatively dispersed. Street robberies, on the other hand, are highly clustered, up to over 30 nearest neighbors.

The linear k-order nearest neighbor distribution gives a slightly different perspective on the distribution than the areal. For one thing, the index is slightly biased as the denominator - the K-order expected linear neighbor distance, is only approximated. For another thing, the index measures distance *as if* the street follow a true grid, oriented in an east-west and north-south direction. In this sense, it may be unrealistic for many places, especially if streets traverse in diagonal patterns; in these cases, the use of indirect distance measurement will produce greater distances than what actually occur on the network. Still, the linear nearest neighbor index is an attempt to approximate travel along the street

Figure 5.4

K-Order Linear Nearest Neighbor Indices

1996 Street Robberies and Residential Burglaries



network. To the extent that a particular jurisdiction's street pattern fall in this manner, it can provide useful information.

Ripley's K Statistic

Ripley's K statistic is an index of non-randomness for different scale values (Ripley, 1976; Ripley, 1981; Bailey and Gattrell, 1995; Venables and Ripley, 1997) . In this sense, it is a 'super-order' nearest neighbor statistic, providing a test of randomness for every distance from the smallest up to the size of the study area. It is sometimes called the *reduced second moment measure*, implying that it is designed to measure second-order trends (i.e., local clustering as opposed to a general pattern over the region). However, it is also subject to first-order effects so that it is not strictly a second-order measure.

Consider a *spatially random* distribution of N points. If circles of radius, d_s, are drawn around each point, where s is the order of radii from the smallest to the largest, and the number of other points that are found within the circle are counted and then summed over all points (allowing for duplication), then the expected number of points within that radius are

$$E(\text{\# of points within distance } d_i) = \frac{N}{A} K(d_s) \qquad (5.13)$$

where N is the sample size, A is the total study area, and $K(d_s)$ is the area of a circle defined by radius, d_s. For example, if the area defined by a particular radius is one-fourth the total study area and *if* there is a spatially random distribution, on average approximately one-fourth of the cases will fall within any one circle (plus or minus a sampling error). More formally, with *complete spatial randomness* (csr), the expected number of points within distance, d_s, is

$$E(\text{\# under csr}) = \frac{N}{A} \pi d_s^2 \qquad (5.14)$$

On the other hand, if the average number of points found within a circle for a particular radius placed over each point, in turn, is greater than that found in equation 5.14, this points to clustering, that is points are, on average, closer than would be expected on the basis of chance for that radius. Conversely, if the average number of points found within a circle for a particular radius placed over each point, in turn, is less than that found in equation 5.14, this points to dispersion; that is points are, on average, farther apart than would be expected on the basis of chance for that radius. By counting the number of total numbers within a particular radius and comparing it to the number expected on the basis of complete spatial randomness, the statistic is an indicator of non-randomness.

In this sense, the K statistic is similar to the nearest neighbor distance in that it provides information about the average distance between points. However, it is more comprehensive than the nearest neighbor statistic for two reasons. First, it applies to all

orders cumulatively, not just a single order. Second, it applies to all distances up to the limit of the study area because the count is conducted over successively increasing radii.

Under unconstrained conditions, K is defined as

$$K(d_s) = \frac{A}{N^2} \sum_i \sum_j I(d_{ij}) \quad (5.15)$$

where $I(d_{ij})$ is the number of other points, j, found within distance, d_s, summed over all points, i. That is, a circle of radius, d_s, is placed over each point, i. Then, the number of other points, ij, are counted. The circle is moved to the next i and the process is repeated. Thus, the double summation points to the count of all j's for each i, over all i's. After this process is completed, the radius of the circle is increased, and the entire process is repeated. Typically, the radii of circles are increased in small increments so that there are 50-100 intervals by which the statistic can be counted. In *CrimeStat*, 100 intervals (radii) are used, based on

$$d_s = \frac{R}{100} \quad (5.16)$$

where R is the radius of a circle for whose area is equal to the study area (i.e., the area entered on the measurement parameters page).

One can graph $K(d_s)$ against the distance, d_s, to reveal whether there is any clustering at certain distances or any dispersion at others (if there is clustering at some scales, then there must be dispersion at others). Such a plot is non-linear, however, typically increasing exponentially (Kaluzny et al, 1998. Consequently, $K(d_s)$ is transformed into a square root function, $L(d_s)$, to make it more linear. $L(d_s)$ is defined as:

$$L(d_s) = SQRT\left[\frac{K(d_s)}{\pi}\right] - d_s \quad (5.17)$$

That is, $K(d_s)$ is divided by π and then the square root is taken. Then the distance interval (the particular radius), d_s, is subtracted from this.[6] In practice, only the L statistic is used even though the name of the statistic K is based on the K derivation. Figure 5.5 shows a graph of L against distance for 1996 robberies in Baltimore County. As can be seen, L increases up to a distance of about 3 miles whereupon it decreases again.

Potential Bias in the Statistic

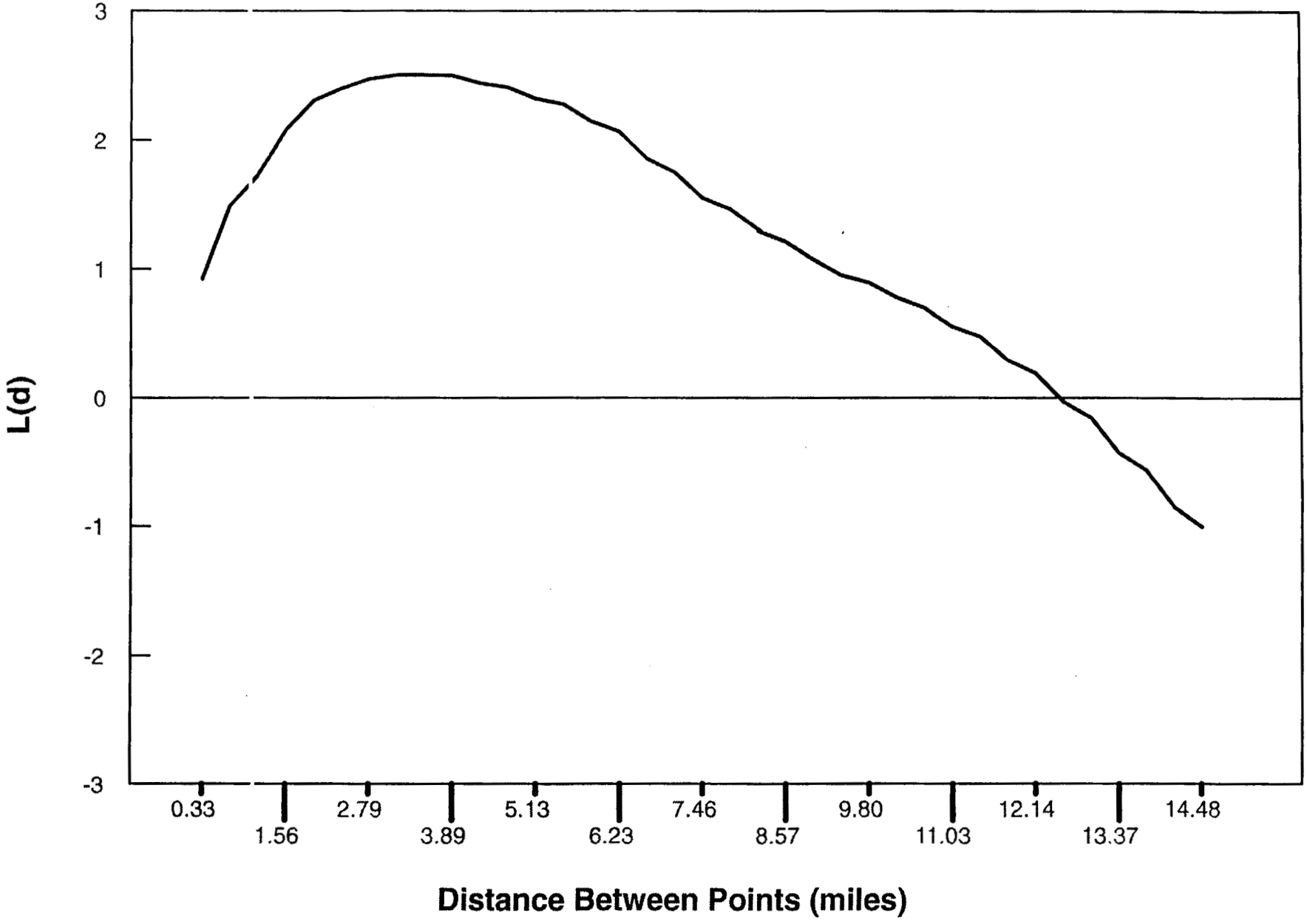
The L statistic is prone to edge effects just like the nearest neighbor statistic. That is, for points located near the boundary of the study area, the number enumerated by any circle for those points will, all other things being equal, necessarily be less than points in the center of the study area because points outside the boundary are not counted. Further, the greater

F' 're 5.5:

K Statistic For 1996 Robberies

As a Function of Distance between Points

L(d) = Sqrt[K(d)/pi] - d



the distance between points that are being tested (i.e., the greater the radius of the circle placed over each point), the greater the bias. Thus, a plot of L against distance will show a declining curve as distance increases.

Ripley has proposed various adjustments to the function to correct the bias. One is a 'guard rail' within the study area so that points outside the guard rail, but inside the study area can only be counted for points inside the guard rail, but cannot be used for enumerating other points within a circle placed over them (that is, they can only be *j*'s and not i's, to use the language of equation 5.15). Such an operation, however, requires manually constructing these guard rails and enumerating whether each point can be both an enumerator and a recipient or a recipient only. For complex boundaries, such as are found in most police departments, this type of operation is extremely tedious and difficult.

Similarly, Ripley has proposed a simple weighting to account for the proportion of the circle placed over each point that is within the study area (Venables and Ripley, 1997). Thus, equation 5.15 is re-written as:

$$K(d_s) = \frac{A}{N^2} \sum_i \sum_j W_{ij}^{-1} I(d_{ij}) \qquad (5.18)$$

where W_{ij}^{-1} is the inverse of the proportion of a circle of radius, d_s, placed over each point which is within the total study area. Thus, if a point is near the study area border, it will receive a greater weight because a smaller proportion of the circle placed over it will be within the study area. Again, however, such a weighting can only be applied if simple geographical shapes are used, such as circles or rectangles (Bailey and Gattrell, 1995). However, most social boundaries, such as those used by police departments, have irregular shapes and there are not simple formulas that can be used to correct the edge bias. Again, one has to enumerate it mechanically which is a very time consuming process.[7]

In practice, therefore, one does not do the adjustment and, instead, learns to anticipate the bias. In *CrimeStat*, only the unadjusted L is calculated. Even though the program will calculate 100 distance intervals, the L statistic should only be examined for small distances, where the biases will be the smallest. Any conclusions for longer distances will be subject to more bias.

Comparison to A Spatially Random Distribution

To understand whether an observed K distribution is different from chance, one typically uses a random distribution. Because the sampling distribution of $L(d_s)$ is not known, a simulation can be conducted by randomly assigning points to the study area. Because any one simulation might produce a clustered or dispersed pattern strictly by chance, the simulation is repeated many times, typically 100 or more. Then, for each random simulation, the L statistic is calculated for each distance interval. Finally, after all simulations have been conducted, the highest and lowest L-values are taken for each distance interval. This is called

an *envelope*. Thus, by comparing the distribution of L to the random envelope, one can assess whether the particular observed pattern is likely to be different from chance.[8]

Specifying simulations

Because simulations can take a long time, particularly if the data sets are large, the default number of simulations is 0. However, a user can conduct simulations by writing a positive number (e.g., 10, 100, 300). If simulations are selected, *CrimeStat* will conduct the number of simulations specified by the user and will calculate the upper and lower limits for each distance interval, as well as the 0.5%, 2.5%, 5%, 95%, 97.5% and 99% intervals; these latter statistics only make sense if many simulation runs are conducted (e.g. 1000).

The way *CrimeStat* conducts the simulation is as follows. It takes the maximum bounding rectangle of the distribution, that is the rectangle formed by the maximum and minimum X and Y coordinates respectively and re-scales this (up or down) until the rectangle has an area equal to the study area (defined on the measurement parameters page). It then assigns N points, where N is the same number of points as in the incident distribution, using a uniform random number generator to this rectangle and calculates the L statistic. It then repeats the experiment for the number of specified simulations, and calculates the above statistics. For example, with 1181 robberies for 1996, the Ripley's K function calculates the empirical L statistics for 100 distance intervals and compares this to a simulation of 1181 points randomly distributed over a rectangle k times, where k is a user-defined number.

In practice, the simulation test also has biases associated with edges. Unlike the theoretical L under uniform conditions of complete spatial randomness (i.e., stretching in all directions well beyond the study area) where L is a straight horizontal line, the simulated L also declines with increasing distance separation between points. This is a function of the same type of edge bias. Consequently, it is possible to compare the empirical L with the random L for even longer distance separations since both have edge biases. There are some subtle differences between the two, however, so some care should be used. The empirical L is obtained from the points within the study area, the geography of which is usually irregular. The random L, however, is calculated from a rectangle. Thus, the differences in the shape comparisons may account for some variations.

Comparison to Baseline Populations

For most social distributions, such as crime incidents, randomness is not a very meaningful baseline. Most social characteristics are non-random. Consequently, to find that the amount of clustering that is occurring is greater than what would be expected on the basis of chance is not very useful for crime analysts. However, it is possible to compare the distribution of L for crime incidents with the distribution of L for various baseline characteristics, for example, for the population distribution or the distribution of employment. In almost all metropolitan areas, population is more concentrated towards the center than at the periphery; the drop-off in population density is very sharp as was shown in the last chapter. All other things being equal, one would expect more incidents towards the metropolitan center than at the periphery; consequently, the average distance between

incidents will be shorter in the center than farther out. This is nothing more than a consequence of the distribution of people. However, to say something about concentrations of incidents above-and-beyond that expected by population requires us to examine the pattern of population as well as of crime incidents.

CrimeStat allows the use of intensity and weighting variables in the calculation of the K statistic. The user must define an intensity or a weight (or both in special circumstances) on the primary file page. The K routine will then use the intensity (or weight) in the calculation of L. Figure 5.6 shows a graph of L against distance for 1996 street robberies in Baltimore County and compares this to both the envelope produced from 100 random simulations as well as the L distribution from the 1990 population; the latter variable was obtained by taking the centroid of census block groups from the 1990 census and using population as the intensity variable. As can be seen, the amount of clustering for robberies is much greater than both the random envelope as well as the distribution of population. In other words, robberies are more clustered together than even what would be expected on the basis of the population distribution and this holds for distances up to about 7 miles, whereupon the distribution of robberies is indistinguishable from a random distribution. For comparison, figure 5.7 shows the distribution of 1996 burglaries, again compared to a random envelope and the distribution of population. We find that burglaries are more clustered than even population, but less so than for robberies; the L value is higher for robberies than for burglaries for near distances. Thus, the distribution of L confirms the result that burglaries tend to be spread over a much larger geographical area in smaller clusters than street robberies, which tend to be more concentrated in large clusters. In terms of looking for 'hot spots', one would expect to find more with robberies than with burglaries.

Distance Matrices

CrimeStat has the capability for outputting distance matrices. There are two types of matrices that can be output. First, the distance between every point in the primary file and every other point can be calculated in miles, nautical miles, feet, kilometers or meters. This is called the *within file point-to-point matrix* (Matrix). Second, if there is also a secondary file, *CrimeStat* can calculate the distance from every point in the primary file to every point in the secondary file, again in miles, nautical miles, feet, kilometers or meters. This is called the *From all primary file points to all secondary file points matrix* (Imatrix).

Both types of matrices can be displayed or saved to a text file for import into another program. Each matrix defines incidents by the order in which they occur in the files (i.e., Record number 1 is listed as '1'; record number 2 is listed '2'; and so forth). Only a subset of each matrix is displayed on the results tab. However, there are horizontal and vertical slider bars that allow the user to scroll through the matrix. The user should move the vertical slide bar first to an approximate proportion of the matrix and click the *Go* button. The matrix will scroll through the rows of the matrix to a place which represents that proportion indicated in the slide bar. The user can then scroll across the rows with the upper slide bar.

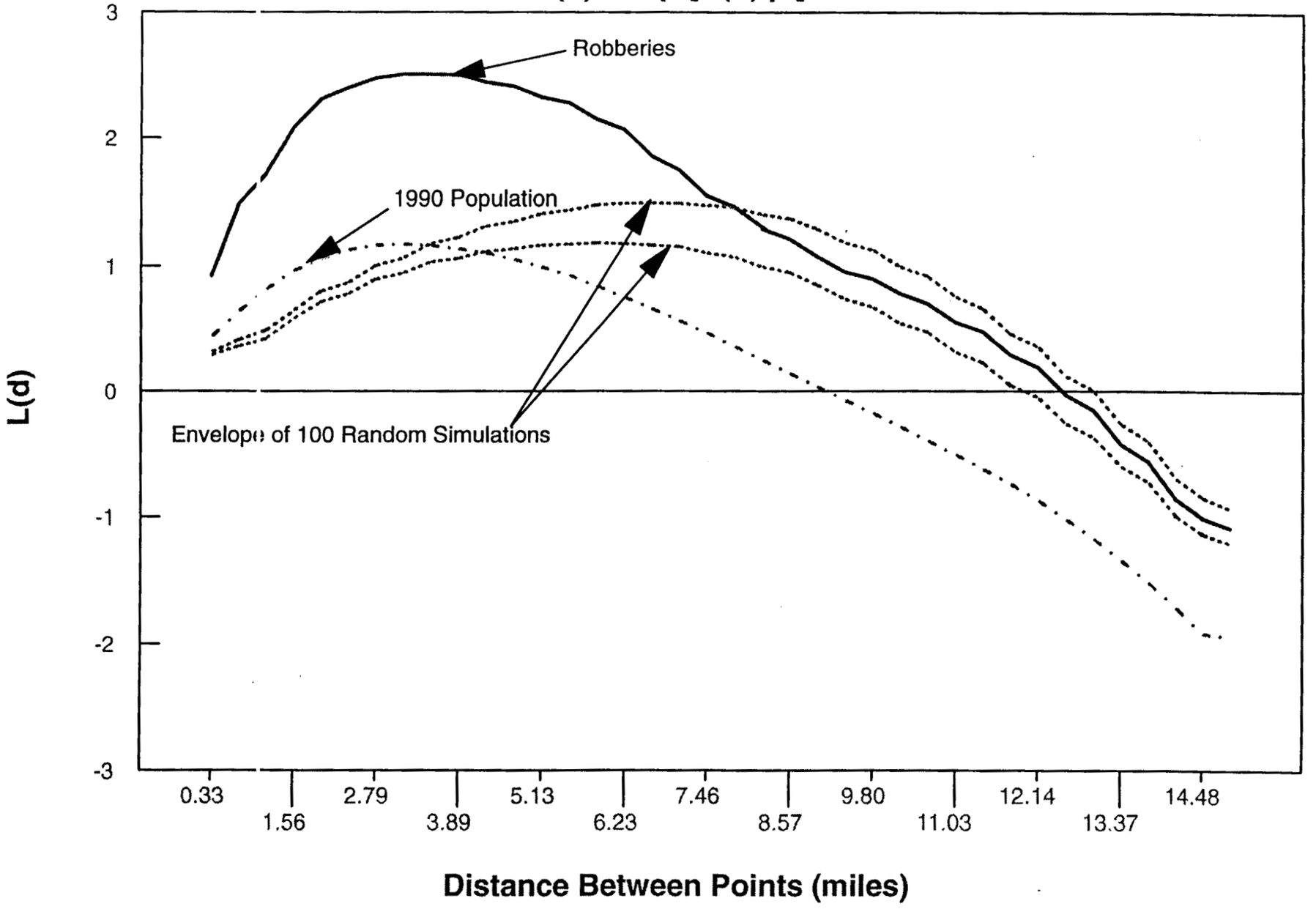
The matrices can be used for various purposes. The *within file point-to-point matrix* can be used to examine distances between particular incidents. The *saved '.txt' matrix* can

Fig 5.6:

K Statistic For 1996 Robberies

Compared to Random and Population Distributions

L(d) = Sqrt[K(d)/pi] - d

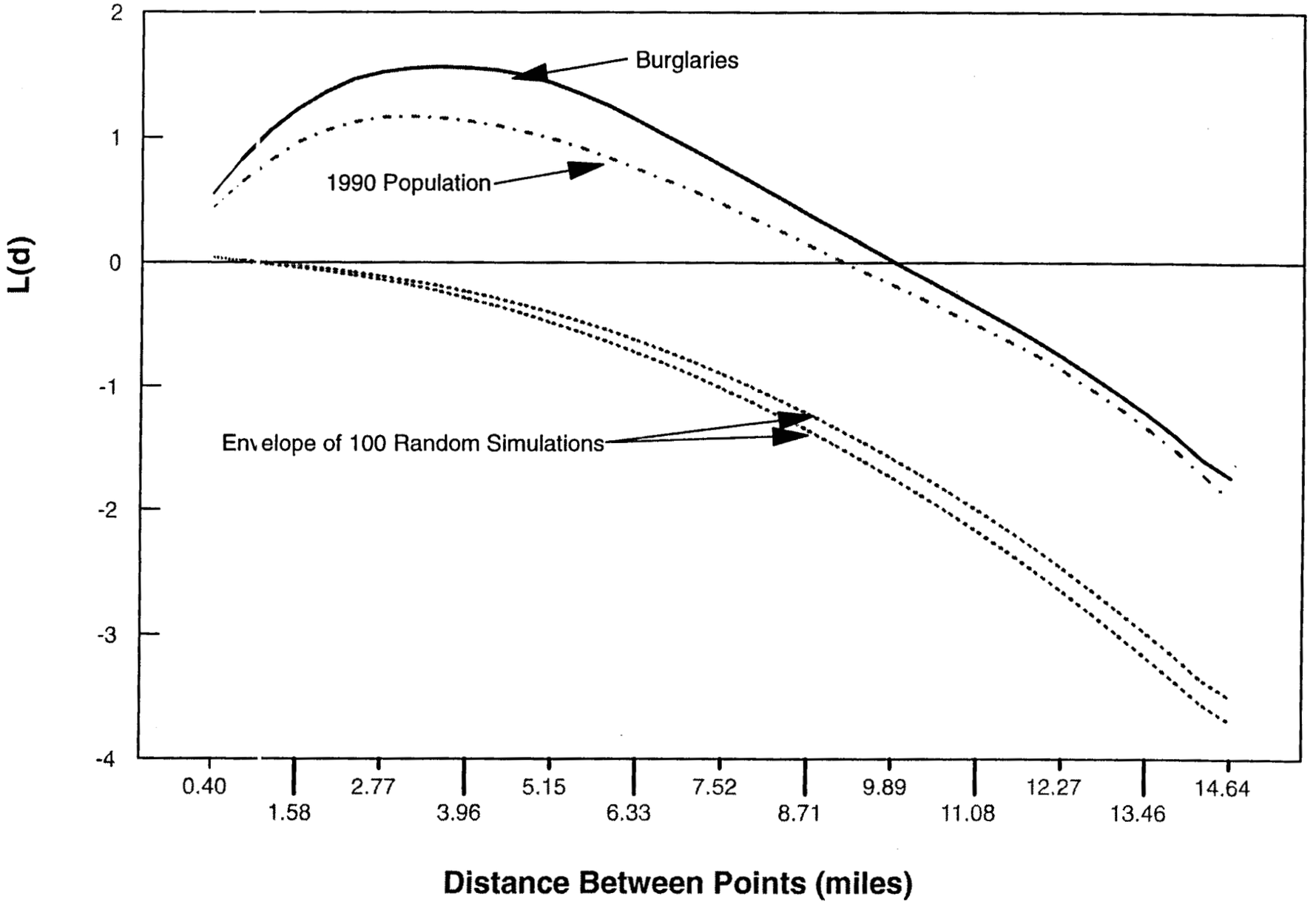


F re 5.7:

K Statistic For 1996 Burglaries

Compared to Random and Population Distributions

L(d) = Sqrt[K(d)/pi] - d



also be imported into a network program for estimating transportation routes. The *primary-to-secondary file matrix* can be used in optimization routines, for example in trying to assess optimal allocation of police cars in order to minimize response time in a police district.

The next chapter will discuss how to identify 'hot spots' with *CrimeStat*.

Endnotes for Chapter 5

1. There is also a mean random distance for a dispersed pattern, called the *mean dispersed distance* (Ebdon, 1988). It is defined as

$$d(dis) = \frac{SQRT[2]}{3^{1/4}\ SQRT[\ N/A\ ]}$$

A nearest neighbor index can be set up comparing the observed mean neighbor distance with that expected for a dispersed pattern. *CrimeStat* only provides the traditional nearest neighbor index, but it does output the mean dispersed distance.

2. Unfortunately, the term *order* when used in the context of nearest neighbor analysis has a slightly different meaning than when used as *first-order* compared to *second-order* statistics. In the nearest neighbor context, *order* really means *neighbor* whereas in the type of statistics context, *order* means the scale of the statistics, global or local. The use of the terms is historical.

3. For geographic areas with simple shapes, such as rectangles or circles, meaningful edge corrections can be made. For example, with a rectangle or a circle, an inner 'guard rail' can be constructed so that points outside the 'guard rail' can only be the nearest neighbor to points inside, but cannot have their nearest neighbor selected. Alternatively, Ripley has proposed a simple weighting for each point of the proportion of a circle drawn from that point to the boundary which falls within the study area (Ripley, 1976; Venables and Ripley, 1997). However, for complex boundaries such as used for census geography or by police departments, such principals are hard to implement.

4. There is not a hard-and-fast rule about how many K-order nearest neighbor distances may be calculated. Cressie (1991, p. 613) shows that error increases with increasing order and the degree of divergence from an edge-corrected measure increases over time. In a test case of 584 point locations, he shows that even after only 25 nearest neighbors, the uncorrected measure yields opposite conclusions about clustering from the corrected measures. So, as a rough approximation, orders no greater than 2.5% of the cases should be calculated.

5. Because *CrimeStat* is using indirect distance for the linear nearest neighbor index (i.e. measurement only in an horizontal or vertical direction), there is a slight distortion that can occur if the incidents are distributed in a diagonal manner, such as with State Highways 26 and 150 in Figure 4.3. The distortion is very small, however. For example, with the incidents along State Highway 26, after rotating the incident points so that they fell approximately in a horizontal orientation, the observed average linear nearest neighbor distance decreased slightly from 0.05843 miles to 0.05061 miles and the linear nearest neighbor index became 0.8354 (t=-.91; not significant). In other words, the effects of the diagonal distribution lengthened

the estimate for the average linear nearest neighbor distance by about 41 feet compared to the actual distances between incidents. For a small sample size, this could be relevant, but for a larger sample it generally will be a small distortion. However, if a more precise measure is required, then the user should rotation the distribution so that the incidents have as closely as possible a horizontal or vertical orientation.

6. This form of the $L(d_s)$ is taken from Cressie (1991). In Ripley's original formulation (Ripley, 1976), distance is not subtracted from the square root function. The advantage of the Cressie formulation is that a complete random distribution will be a straight line that is parallel to the X-axis.

7. Theoretically, it would be possible to write a program which could query the geographic boundary within a GIS to see whether what proportion of a circle placed over each incident point falls within the study area. This program would have to test the location of each point and compare it with the location of the nearest boundary, a process that would require enumerating all locations defined by the geographical layer. With current technology, such calculations would require very extensive programming and would be extremely slow. It's not really very practical. However, in the future, as computers systems become increasingly fast, the feasibility of such 'edge testing' algorithms may become more practical.

8. Note, that since there is not a formal test of significance, the comparison with an envelope produced from a number of simulations provides only approximate confidence about whether the distribution differs from chance or not. That is, one cannot say that the likelihood of obtaining this result by chance is less than 5%, for example.

Chapter 6
'Hot Spot' Analysis

In this chapter, we describe three tools for identifying clusters of crime incidents. Typically called *hot spots* or *hot spot areas*, these are concentrations of incidents within a limited geographical area that appear over time. Police have learned from experience that there are particular environments that attract crimes in larger-than-expected concentrations, so-called *crime generators*. Sometimes these hot spot areas are defined by particular activities (e.g., drug trading; Maltz, Gordon, and Friedman, 1989) and other times by specific concentrations of land uses (e.g., skid row areas, bars, adult bookshops, itinerant hotels; Block and Block, 1995; Levine, Wachs and Shirazi, 1986). Whatever the reasons for the concentration, they are real and are known by most police departments.

While there are some theoretical concerns about what links disparate crime incidents together into a cluster, nonetheless, the concept is very useful. Police officers patrolling a precinct can focus their attention on particular environments because they know that crime incidents will continually reappear in these places. Crime prevention units can target their efforts knowing that they will achieve a positive effect in reducing crime with limited resources. In short, the concept is very useful. Nevertheless, the concept is a perceptual construct. 'Hot spots' do not exist in reality, but are areas where there is sufficient concentration of certain activities (in this case, crime incidents) such that they get labeled as being an area of high concentration. There is not a boundary around these incidents, but a gradient where people draw an imaginary line to indicate the location at which the 'hot spot' *starts*. In reality, any variable that is measured, such as the density of crime incidents, will be continuous over an area, being higher in some parts and lower in others. Where a line is drawn in order to define a 'hot spot' is somewhat arbitrary. The end of this chapter raises some questions about the interpretation of 'hot spots'.

Statistical Approaches to the Measurement of 'Hot Spots'

Unfortunately, measuring a 'hot spot' is also a complicated problem. There are literally dozens of different statistical techniques designed to identify 'hot spots' (Everitt, 1974). For example, the *Spatial and Temporal Analysis of Crime* (STAC) program is a well known 'hot spot' identifier technique used in crime analysis (Block, 1994; Block and Green, 1994). But it is not the only technique. There are many others which are typically known under the general statistical label of *cluster analysis*. These are statistical techniques aimed at grouping cases together into relatively coherent clusters. All of the techniques depend on optimizing various statistical criteria, but the techniques differ among themselves in their methodology as well as in the criteria used for identification. Because 'hot spots' do not really exist, as they are perceptual constructs, any technique that is used must approximate how someone would perceive an area. The techniques do this through various mathematical criteria.

Types of Cluster Analysis ('Hot spot') Methods

Several typologies of cluster analysis have been developed as cluster routines typically fall into several general categories (Everitt, 1974; Çan and Megbolugbe, 1996):

1. *Hierarchical* techniques (Sneath, 1957; McQuitty, 1960; Sokal and Sneath, 1963; King, 1967; Sokal and Michener, 1958; Ward, 1963; Hartigan, 1975) are like an inverted tree diagram in which two or more incidents are first grouped on the basis of some criteria (e.g., nearest neighbor). Then, the pairs are grouped into second-order clusters. The second-order clusters are then grouped into third-order clusters, and this process is repeated until either all incidents fall into a single cluster or else the grouping criteria fails. Thus, there is a hierarchy of clusters that can be displayed with a dendogram (an inverted tree diagram).

   Figure 6.1 shows an example of a hierarchical clustering where there are four orders (levels) of clustering; the visualization is non-spatial in order to show the linkages. In this example, all individual incidents are grouped into first-order clusters which, in turn, are grouped into second-order clusters which, in turn, are grouped into third-order clusters which all converge into a single fourth-order cluster. Many hierarchical techniques, however, do not group all incidents or all clusters into the next highest level.

2. *Partitioning* techniques, frequently called the K-means technique, partition the incidents into a specified number of groupings, usually defined by the user (Thorndike, 1953; MacQueen, 1967; Ball and Hall, 1970; Beale, 1969). Thus, all points are assigned to one, and only one, group. Figure 6.2 shows a partitioning technique where all points are assigned to clusters and are displayed as ellipses.

3. *Density* techniques identify clusters by searching for dense concentrations of incidents (Carmichael et al, 1968; Gitman and Levine, 1970; Cattell and Coulter, 1966; Wishart, 1969). In the next chapter, one type of density search algorithm using the kernel density method will be discussed.

4. *Clumping* techniques involve the partitioning of incidents into groups or clusters, but allow overlapping membership (Jones and Jackson, 1967; Needham, 1967; Jardine and Sibson, 1968; Cole and Wishart, 1970).

5. *Miscellaneous* techniques are other methods that are less commonly used including techniques applied to zones, not incidents. In this chapter, we discuss the *Local Moran* technique for identifying neighborhood discrepancies (Anselin, 1995).

There are also hybrids between these methods. For example, *STAC* is primarily a partitioning method but with elements of hierarchical grouping (Block and Green, 1994).

Figure 6.1:

Hierarchical Clustering Technique

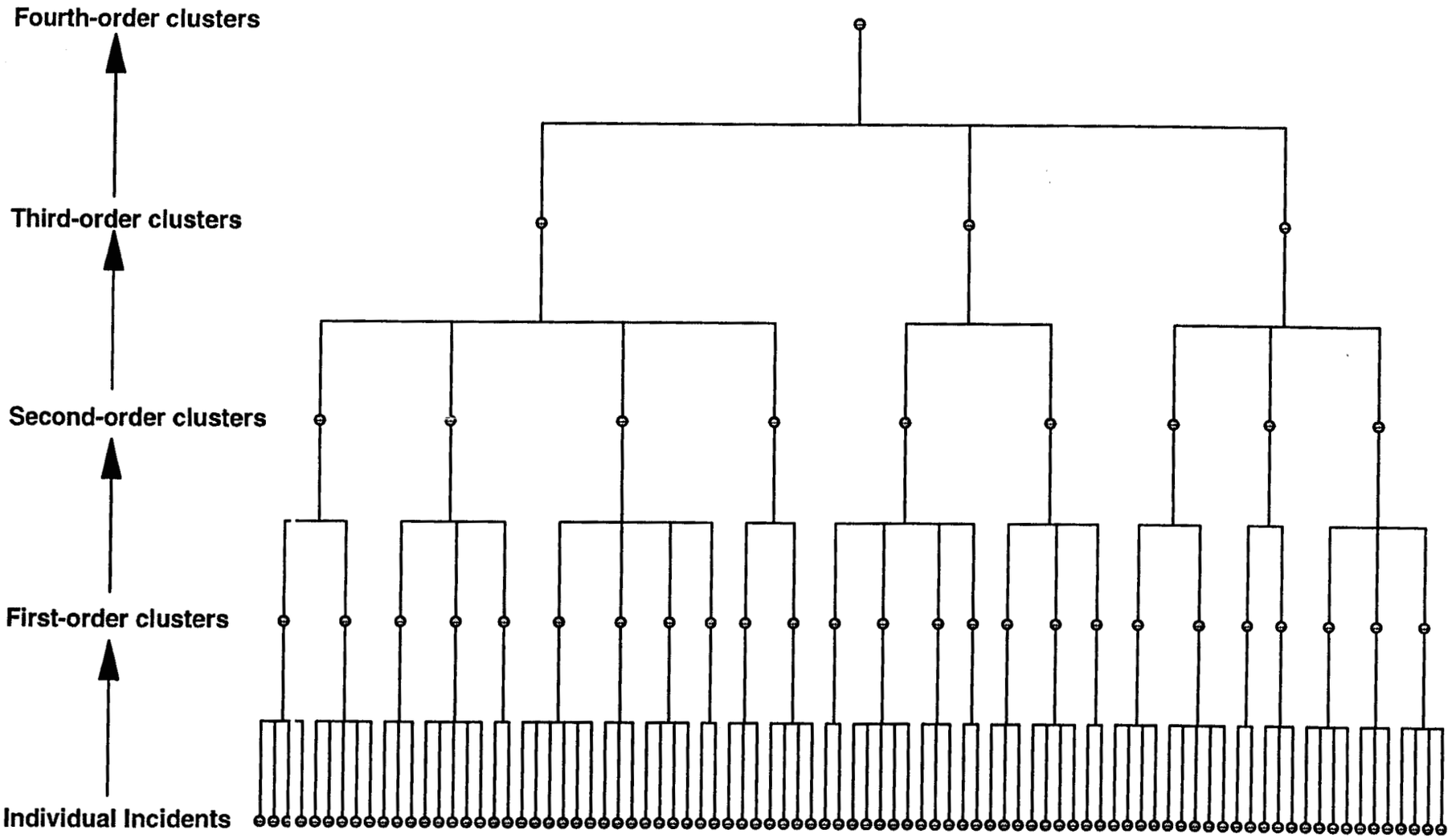
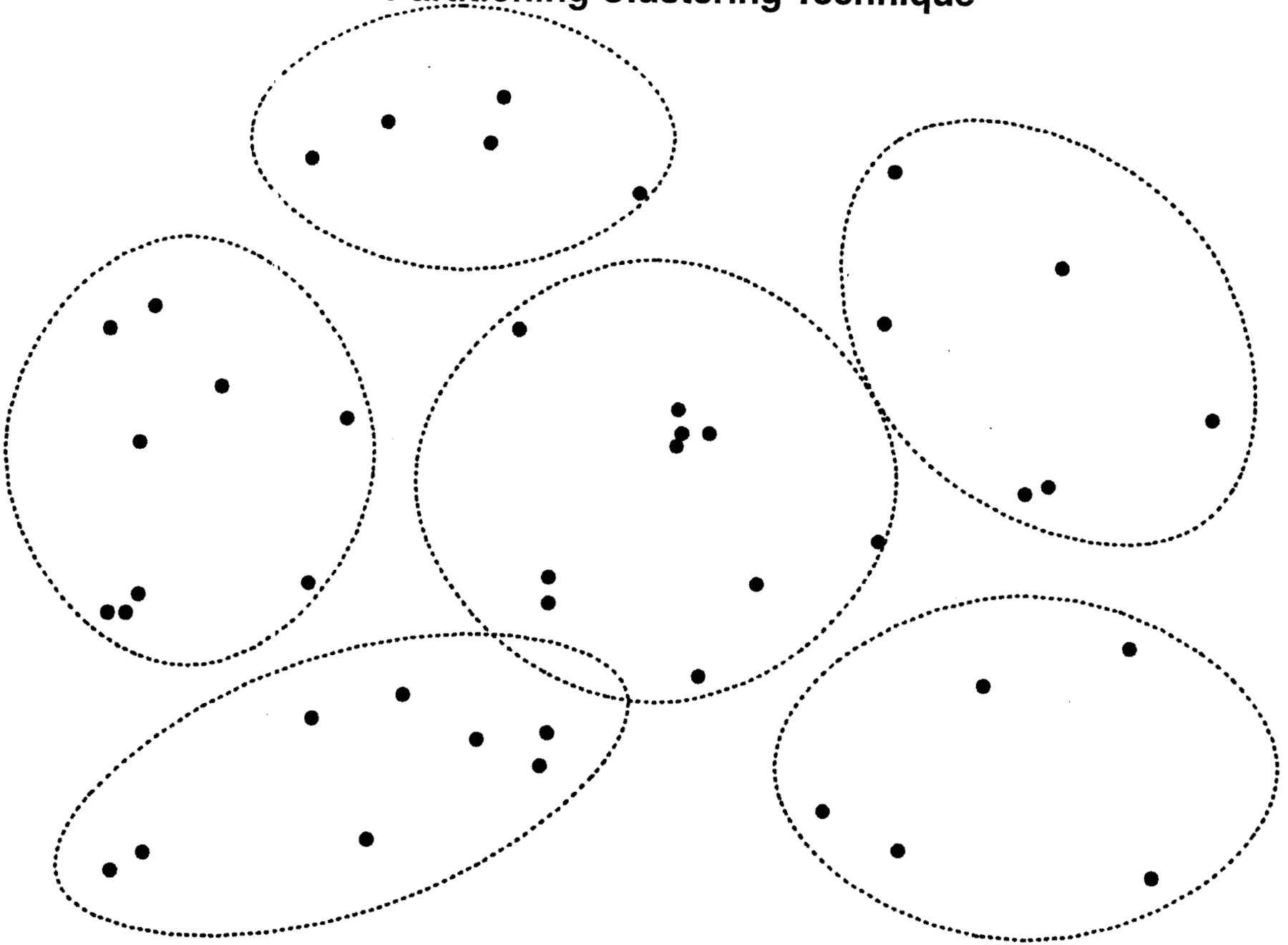


Figure 6.2:

Partitioning Clustering Technique



Optimization Criteria

In addition to the different types of cluster analysis, there are different criteria that distinguish techniques applied to space. Among these are:

1. The *definition* of a cluster - whether it is a discrete grouping or a continuous variable; whether points must belong to a cluster or whether they can be isolated; whether points can belong to multiple clusters.

2. The *choice of variables* in addition to the X and Y coordinates - whether weighting or intensity values are used to define similarities.

3. The measurement of *similarity and distance* - the type of geometry being used; whether clusters are defined by closeness or not; the types of similarity measures used.

4. The *number* of clusters - whether there are a fixed or variable number of clusters; whether users can define the number or not.

5. The geographical *scale* of the clusters - whether clusters are defined by small or larger areas; for hierarchical techniques, what level of abstraction is considered optimal.

6. The *initial selection* of cluster locations ('seeds') - whether they are mathematically or user defined; the specific rules used to define the initial seeds.

7. The *optimization routines* used to adjust the initial seeds into final locations - whether distance is being minimized or maximized; the specific algorithms used to readjust seed locations.

8. The *visual display* of the clusters, once extracted - whether drawn by hand or by a geometrical object (e.g., an ellipse); the proportion of cases represented in the visualization.

This is not the place to provide a comprehensive review of cluster techniques. Nevertheless, it should be clear that with the several types of cluster analysis and the many criteria that can be used for any particular technique, there is a large number of different cluster techniques that could be applied to an incident data base. It should be realized that there is not a single solution to the identification of 'hot spots', but that different techniques will reveal different groupings and patterns among the groups. A user must be aware of this variability and must choose techniques that can complement other types of analysis. It would be very naive to expect that a single technique can reveal the existence of 'hot spots' in a jurisdiction which are unequivocally clear. In most cases

analysts are not even sure why there are 'hot spots' in the first place and, until that is solved, it would be unreasonable to expect a mathematical or statistical routine to solve that problem. I will return to this point at the end of the chapter.

Cluster Routines in *CrimeStat*

Because of the variety of cluster techniques, *CrimeStat* includes three techniques that cover the range of techniques that have been used. One of these is a hierarchical clustering method based on nearest neighbor analysis . A second is a partitioning technique based on the *K-means* algorithm. The third is a zonal technique designed to identify zones which are different from their nearby environment, whether they are 'peaks' or 'troughs'. These are not the only techniques, of course, and analysts should use them as complements to other types of analysis. Figure 6.3 shows the 'hot spot' analysis page in *CrimeStat* and the three routines.

Nearest Neighbor Hierarchical Clustering (Nnh)

The *nearest neighbor hierarchical clustering* (Nnh) routine in *CrimeStat* identifies groups of incidents that are spatially closer than would be expected on the basis of chance. Only points which fit this criteria are clustered at the first level (first-order clusters). Subsequent clustering produces a hierarchy of clusters. The first-order clusters are themselves clustered into second-order clusters. Again, only clusters that are spatially closer than would be expected on the basis of chance are included. The second-order clusters, in turn, are clustered into third-order clusters, and this re-clustering process is continued until no more clustering is possible, either all clusters converge into a single cluster or, more likely, no two clusters are closer together than would be expected on the basis of chance.

Nearest Neighbor Criteria

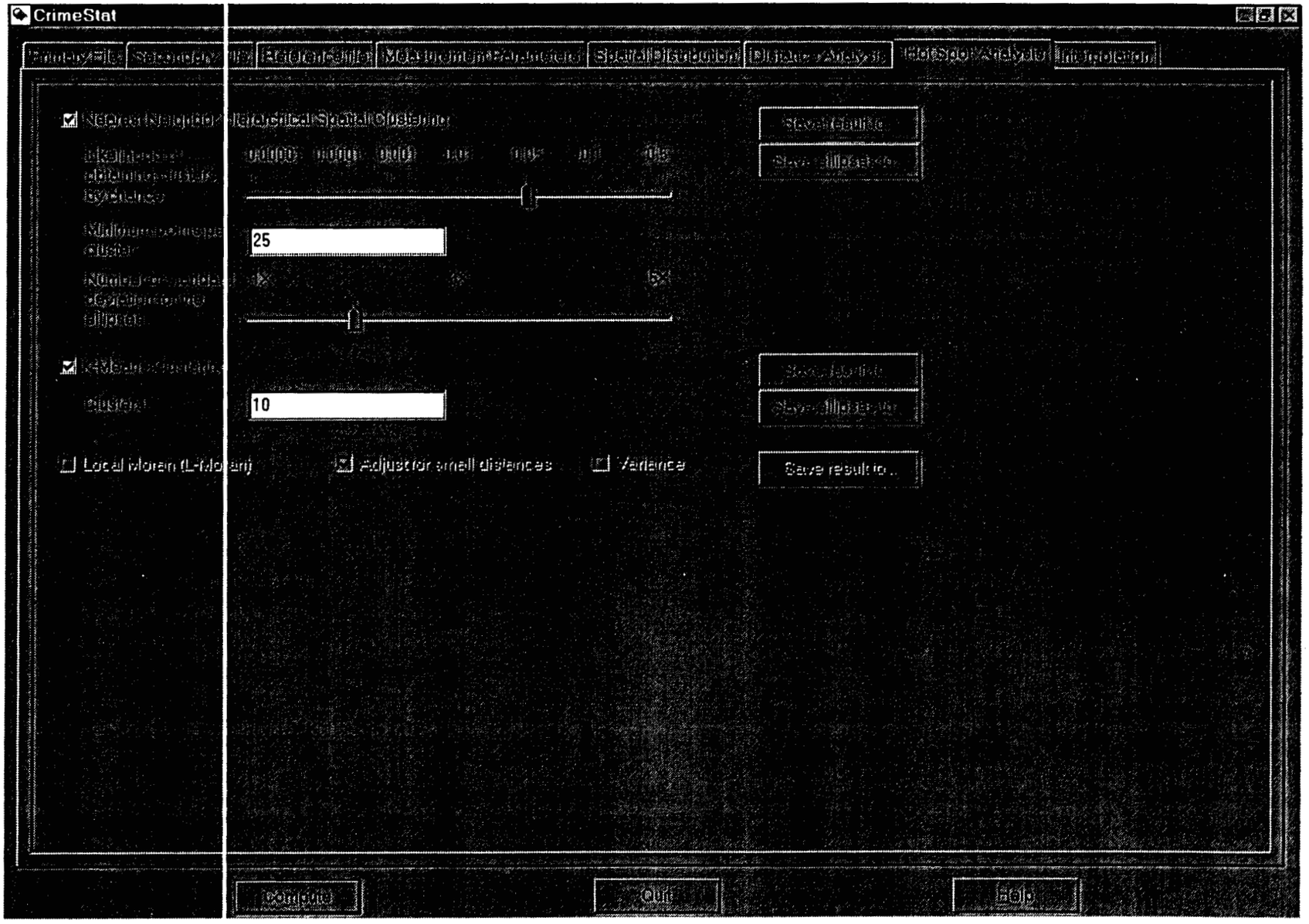
The criteria that is used for clustering points together is the lower confidence interval for the random expected nearest neighbor. From chapter 5, the mean random distance was defined as

$$\text{Mean Random Distance} = d(ran) = 0.5 \text{ SQRT } \left[\frac{A}{N} \right] \quad \text{repeat} \tag{5.2}$$

where A is the area of the region and N is the number of incidents. The confidence interval around that distance is defined as

$$\text{Confidence Interval for Mean Random Distance} = \text{Mean Random Distance} \pm t^* SE_{d(ran)}$$

Figure 6.3: 'Hot Spot' Analysis Layout



$$= \quad 0.5\ \mathrm{SQRT}\left[\frac{A}{N}\right] \pm t\left[\frac{0.26136}{\mathrm{SQRT}[N^2/A]}\right] \qquad (6.1)$$

where A is the area of the region, N is the number of incidents, t is the t-value associated with a probability level in the Student's t-distribution.

The lower limit of this confidence interval is

$$\text{Lower Limit of Confidence Interval for Mean Random Distance} = 0.5\ \mathrm{SQRT}\left[\frac{A}{N}\right] - t\left[\frac{0.26136}{\mathrm{SQRT}[N^2/A]}\right] \qquad (6.2)$$

That is, for a specific *one-tailed* probability, p, fewer than p% of the incidents would have nearest neighbor distances smaller than this lower limit *if* distribution was spatially random. This is the *threshold distance* for the routine. Taking a broader conception of this, if there is a spatially random distribution, then for all distances between points, of which there are

$$\frac{N(N-1)}{2}$$

combinations, fewer than p% of the distances will be smaller than this threshold distance. The lower limit of the confidence interval for the mean random distance is the starting point for the Nnh routine. First-order clustering is conducted as follows:

First-order clustering

In order to conduct clustering, the user specifies two parameters:

1. First, a *one-tailed* probability level for defining the threshold distance.[1] The t-value corresponding to this probability level, t, is selected from the Student's t-distribution under the assumption that the degrees of freedom are at least 120.[2] The range is from a probability level of 0.5 (or a 50% likelihood that the distance could be due to chance) to as small a probability level as 0.00001 (or a 0.001% likelihood that the distance could be due to chance). The choice is made with a slider bar. The default setting is 0.1 (or a 10% likelihood that the distance could be due to chance).

2. Second, the minimum number of points that are required for each cluster. This criteria is used to reduce very small clusters. The default is 10. By decreasing this number, more clusters are produced; conversely, by increasing this number, fewer clusters are produced.

Using these criteria, CrimeStat constructs a first-order clustering of the points and outputs the clusters as ellipses.[3] For each first-order cluster, the center of minimum distance is output as the cluster center, which can be saved as a '.dbf' file. A standard deviational ellipse is calculated for each cluster in turn (see chapter 4 for definition) and the number of standard deviations can be varied from one (1X - the default) up to five (5X). The user specifies the number of standard deviations to save as ellipses in *ArcView* '.shp', *MapInfo* '.mif' or *Atlas*GIS* '.bna' formats.

Second and higher-order clusters

The first-order clusters are then tested for second-order clustering. The procedure is similar to first-order clustering except that the cluster centers are now treated as 'points' which themselves are clustered.[4] The process is repeated until no further clustering can be conducted, either all sub-clusters converge into a single cluster, or the threshold distance criteria fails, or there are fewer than four seeds in the higher-order cluster.

Guidelines for Selecting Parameters

In the Nnh routine, the user has to define three parameters - the likelihood (or p-value) that the threshold distance is obtained by chance, the number of standard deviations for the ellipses that are output, and the minimum number of points. The p-value is selected with a likelihood slider bar (see figure 6.3). This bar indicates a range of p-values from 0.5 (i.e., the likelihood of obtaining the threshold distance by chance is 50%) down to 0.00001 (i.e., the likelihood of obtaining the threshold distance by chance is 0.001%). The slider bar actually controls the value of t in equation 6.3, which varies from 0.0 to 4.4416. The larger the t-value, the smaller the threshold distance. With smaller threshold distances, fewer clusters are extracted which are typically smaller (although not always).

In the case of a p-value of 0.5 (or t=0), then the threshold distance is equal to the random expected distance between two points. Any cluster that is identified with this p-value will be large, but there is a 50% likelihood that the grouping of points is due to chance. Or, put another way, if clusters are identified on any particular run, then most likely about half of them are grouped together by chance. At the other end of the spectrum is a p-value which is very, very small - 0.00001. Any cluster that is identified with this p-value will be small, but there is a very small likelihood that the grouping of incidents is due to chance. On any particular run, there is little chance that even one of them is grouped because of chance. Thus, a user must trade off the number of clusters and the size of an area that defines a cluster with the likelihood that the result could be due to chance. Statistically, there is more certainty with small threshold distances than with larger ones using this technique.

This choice will depend on the needs of the user. For interventions around particular locations, the use of a small threshold distances may actually be appropriate; some of the ellipses seen in figure 6.4 below cover only a couple of street segments. These define micro-neighborhoods or almost pure 'hot spot' locations. On the other hand, for a patrol route, for example, a cluster the size of several neighborhoods might be more appropriate. A patrol car would need to cover a sizeable area and having a larger area to target might be more appropriate than a 'micro' environment. However, there will be less precision with a larger cluster size covering this type of area.

A second criterion is the output size of the clusters. For each cluster in turn, a standard deviational ellipse is calculated (see chapter 4). The user specifies the size of the ellipse in terms of standard deviations. The range is from one standard deviation (1X - the default) up to five standard deviations. Typically, one standard deviation will cover more than 60% of the cases whereas five standard deviations will cover more than 99.99% of the cases, although the exact percentage will depend on the distribution.

The third criterion is the minimum number of points that are required to define a cluster. If a cluster does not have this minimum number, *CrimeStat* will ignore the seed location. Without this criteria, the Nnh routine could identify clusters of two or three incidents each. A 'hot spot' of this size is usually not very useful. Consequently, the user should increase the number to ensure that the identified cluster represents a meaningful number of cases. The default value is 10, but the user can type in any other value.

Nnh Output Files

The Nnh routine has three outputs. First, final seed locations of each cluster and the parameters of a 1X standard deviational ellipse are calculated for each cluster. These can be output to a '.dbf' file or saved as a text ('.txt') file. Only 45 of the seed locations are displayed on the screen. The user can scroll down or across by adjusting the horizontal and vertical slider bars and clicking on the *Go* button.

Second, for each order that is calculated, *CrimeStat* calculates the mean center of the cluster. This can be saved as a '.dbf' file. Third, the standard deviational ellipses of the clusters can be saved in *ArcView* '.shp', *MapInfo* '.mif' or *Atlas*GIS* '.bna' formats. The size of the ellipses are determined by the number of standard deviations to be calculated (see above). Because there are also orders of clusters (i.e., first-order, second-order, etc.), there is a naming convention that distinguishes the order. The convention is

Nnh<O><*username*>

where *O* is the order number and *username* is a name provide by the user. Thus,

Nnh1robbery

are the first-order clusters for a file called 'robbery' and

Nnh2burglary

are the second-order clusters for a file called 'burglary'. Within files, clusters are named

Nnh<O>Ell<N><*username*>

where*O* is the order number, *N* is the ellipse number and *username* is the user-defined name of the file. Thus,

Nnh1Ell10robbery

is the tenth ellipse within the first-order clusters for the file 'robbery' while

Nnh2Ell1burglary

is the first ellipse within the second-order clusters for the file 'burglary'.

In other words, names of files and features can get complicated. The easiest way to understand this, therefore, is to import the file into one of the GIS packages and display it.

Example 1: Nearest neighbor hierarchical clustering of burglaries

The Nnh routine was applied to the Baltimore County 1996 burglary data (n=6,051 incidents). A one-tailed probability level of .05 (or 5%) was selected and each cluster was required to contain a minimum of 10 points (the default). *CrimeStat* returned 153 first-order clusters, 20 second-order clusters and two third–order clusters. Figure 6.4 shows the first-order clusters displayed as 1x standard deviational ellipses. Since the criteria for clustering is the lower limit of the mean random distance, the distances involved are very small, as can be seen. Note, the standard deviational ellipse is defined by the points in the cluster and includes approximately two-thirds of the points. Thus, the clusters actually extend a little beyond the ellipses. Figure 6.5 shows the 20 second-order clusters (dashed lines) and the two third-order clusters (double lines). As seen, they cover much larger areas than the first-order clusters. Finally, figure 6.6 shows a part of east Baltimore County where there are 29 first-order clusters (solid line), five second-order clusters (dashed lines), and one third-order cluster (double line). The street network is presented to indicate the scale. Most first-order clusters cover an area the size of a small neighborhood while the second-order clusters cover larger neighborhoods.

Advantages of Hierarchical Clustering

There are four advantages to this technique. First, it can identify small geographical environments where there are concentrated incidents. This can be useful for specific targeting, either by police deployment or community intervention. There are clearly micro-environments which generate crime incidents (Levine, Wachs and Shirazi, 1986; Maltz, Gordon and Friedman, 1989). The technique tends to identify these small environments because the lower limit of the mean random distance is used to group the

Figure 6.4: First-Order Baltimore County Burglary 'Hot Spots'

Using Nearest Neighbor Hierarchical Clustering Method

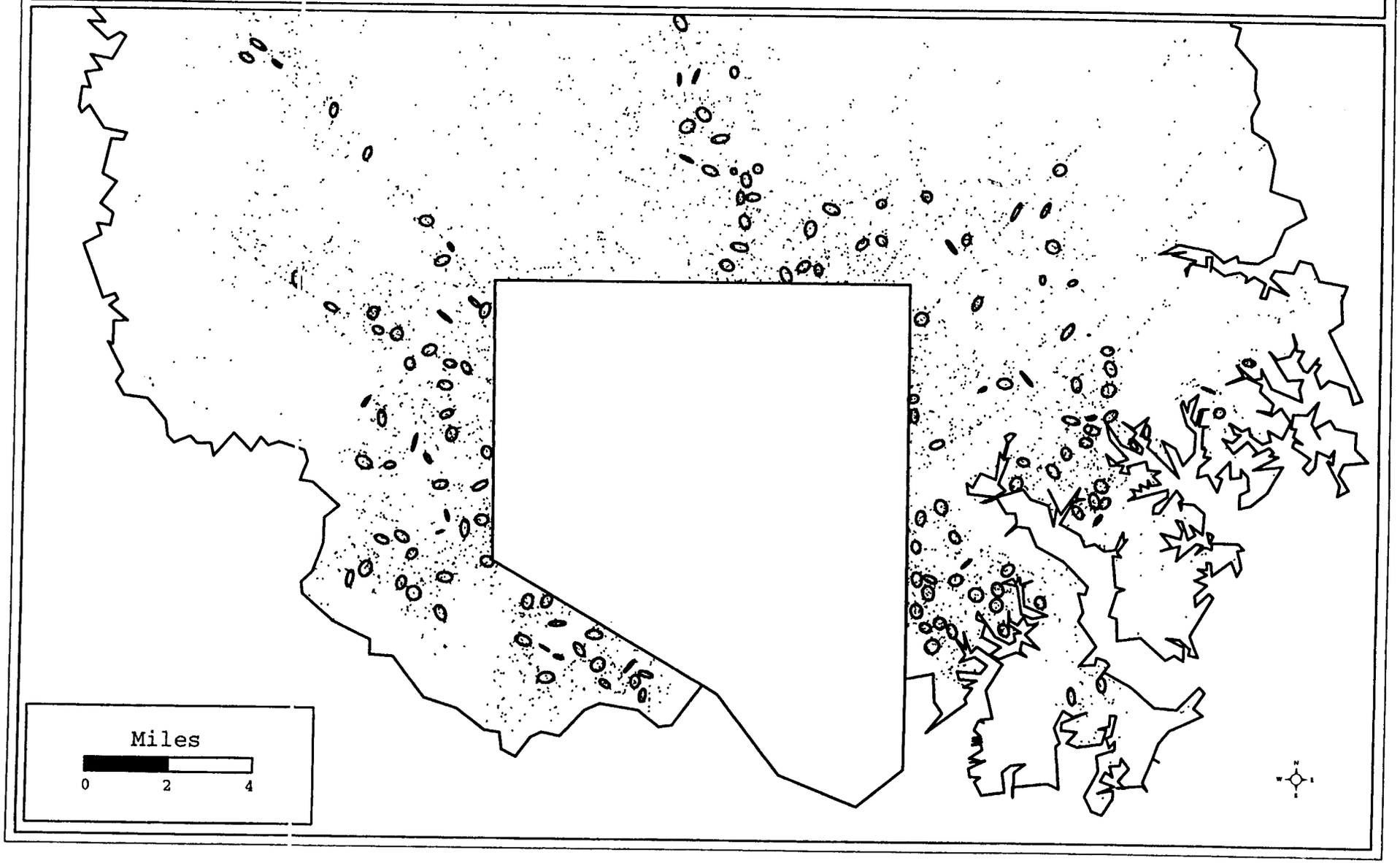


Figure 6.5: Second- and Third-Order Burglary 'Hot Spots'

Using Nearest Neighbor Hierarchical Clustering Method

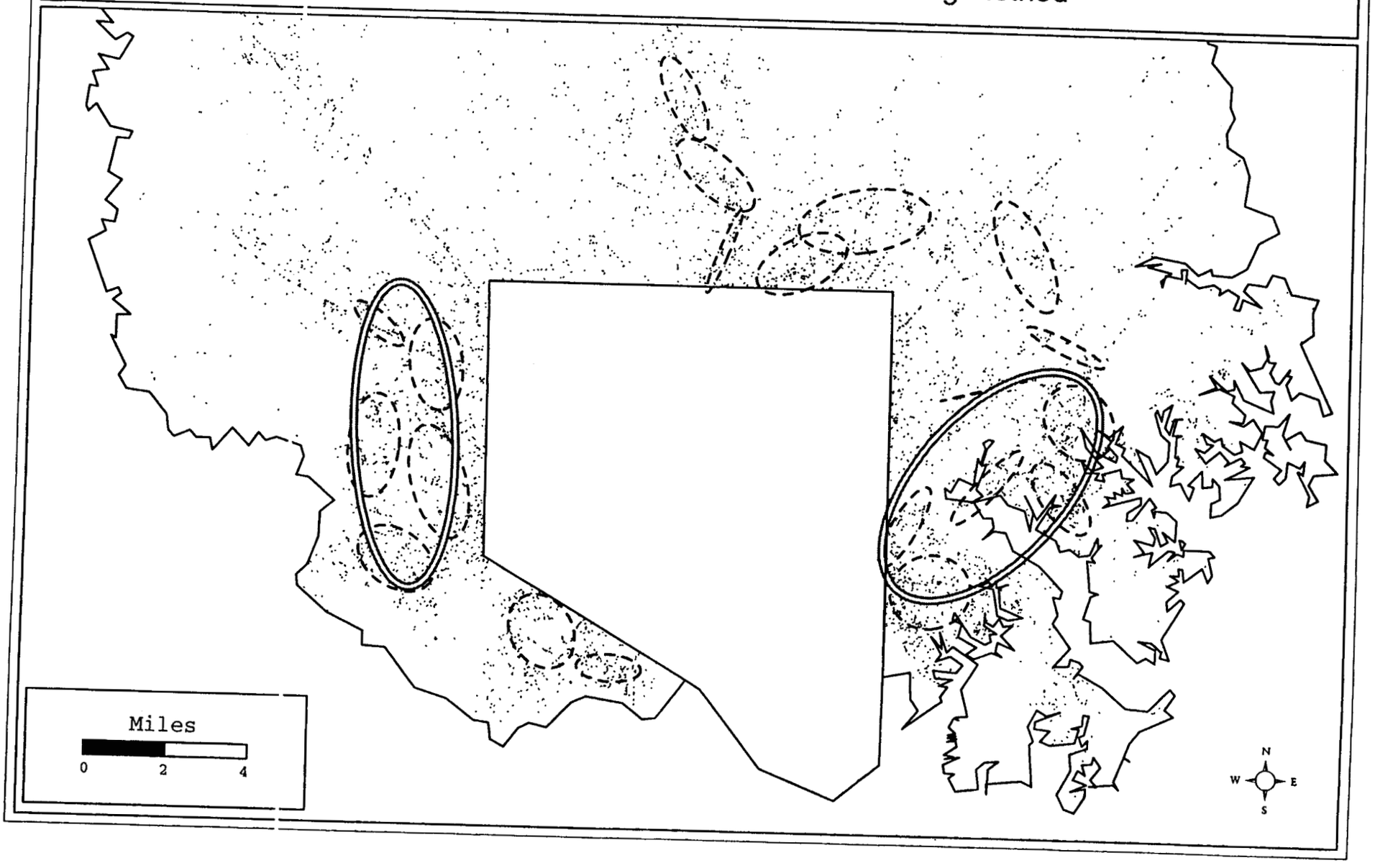


Figure 6.6: East Baltimore County Burglary 'Hot Spots'

Using Nearest Neighbor Hierarchical Clustering Method

clusters. The user can, of course, control the size of the grouping area by loosening or tightening either the p-value or the minimum number of required points. Thus, the sizes of the clusters can be adjusted to fit particular groupings of points.

Second, the technique can be applied to any entire data set, such as for Baltimore County and Baltimore City, and need not only be applied to smaller geographical areas, such as precincts. This increases the ease of use for analysts and can facilitate comparisons between different areas without having to limit arbitrarily the data set prior to the analysis.

Third, the linkages between several small clusters can be seen through the second- and higher-order clusters. Frequently, 'hot spots' are located near other 'hot spots' which, in turn, are located near other 'hot spots'. As we've seen from the maps of robbery, burglary and motor vehicle thefts in Baltimore County, there are large areas within the County that have a lot of incidents. Within these large areas, there are smaller 'hot spots' and within some of those 'hot spots', there are even small ones. In other words, there are different scales to the clustering of points - different geographical levels, if you will, and the hierarchical clustering technique can identify these levels.

Fourth, each of the levels imply different management strategies. For the smallest level, officers can intervene effectively in small neighborhoods, as discussed above. Second-order clusters, on the other hand, are more appropriate as patrol areas; these areas are larger than first-order clusters, but include several first-order clusters within them. If third- or higher-order clusters are identified, these are generally areas with very high concentrations of crime incidents over a fairly large section of the jurisdiction. The areas start to approximate precinct sizes and need to be thought of in terms of an integrated management strategy - police deployment, crime prevention, community involvement, and long-range planning. Thus, the hierarchical technique allows different security strategies to be adopted and provides a coherent way of approaching these communities.

Limitation to Hierarchical Clustering

At the same time, there are limitations to the technique, some technical and others theoretical. First, the size of the grouping area is dependent on the sample size since the lower limit of the mean random distance is used as the criteria (see equation. 4.2). For crime distributions which have many incidents (e.g., burglary), the threshold distance will be a lot smaller than distribution which have fewer incidents (e.g., robbery). In theory, a 'hot spot' is dependent on an environment, not the number of incidents. Thus, the technique does not produce a consistent definition of a 'hot spot' area.

Second, there is a certain arbitrariness in the technique due to the minimum points rule. This implicitly requires the user to define a meaningful cluster size, whether the number of points are 5, 10, 15 or whatever. To some extent, this is how patterns are defined by human beings; with one or two incidents in a small area, people don't perceive any pattern. As soon as the number of incidents increases, say to 10 or more, people perceive the pattern. This is not a statistical way for defining regularity, but it is a human

way. However, it can lead to arbitrariness since two different users may interpret the size of a 'hot spot' differently. Similarly, the selectivity of the p-value, vis-a-via the Student's t-distribution, can allow variability between users. In short, the technique does produce a constant result, but one subject to manipulation by users. Hierarchical techniques are, of course, not the only clustering procedure to allow users to adjust the parameters; in fact, almost all the cluster techniques have this property. But it is a statistical weakness in that it involves subjectivity and is not necessarily consistently applied across users.

Finally, there is no theory or rationale behind the clusters. They are empirical derivatives of a procedures. Again, many clustering techniques are empirical groupings and also do not have any explanatory theory. However, if one is looking for a substantive 'hot spot' defined by a unique constellation of land uses, activities, and targets, the technique does not provide any insight into why the clusters are occurring or why they could be related. I will return to this point at the end, but it should be remembered that these are empirical groupings, not necessarily substantive ones.

K-Means Partitioning Clustering

The *K-means* clustering routine (Kmeans) is a partitioning procedure where the data are grouped into K groups defined by the user. A specified number of seed locations, K, are defined by the user. The routine tries to find the best positioning of the K centers and then assigns each point to the center that is nearest. Like the Nnh routine, Kmeans assigns points to one, and only one, cluster. However, unlike the nearest neighbor hierarchical (Nnh) procedure, all points are assigned to clusters. Thus, there is no hierarchy in the routine, that is there are no second- and higher-order clusters.

The technique is useful when a user want to control the grouping. For example, if there are 10 precincts in a jurisdiction, an analyst might want to identify the 10 most compact clusters, one for precinct. Alternatively, if a previous analysis has shown there were 24 clusters, then an analyst could check whether the clusters have shifted over time by also asking for 24 clusters. By definition, the technique is somewhat arbitrary since the user defines how many clusters are to be expected. Whether a cluster could be a 'hot spot' or not would depend on the extent to which a user wanted to replicate 'hot spots' or not.

The theory of the K-means procedure is relatively straightforward. The implementation is more complicated. K-means represents an attempt to define an optimal number of K locations where the sum of the distance from every point to each of the K centers is minimized. It is a variation of the old location theory paradigm of how to locate K facilities (e.g., police stations, hospitals, shopping centers) given the distribution of population (Haggett, Cliff, and Frey, 1977). That is, how does one identify *supply* locations in relation to *demand* locations. In theory, solving this question is an empirical solution, what is frequently called *global optimization*. One tries every combination of K objects where K is a subset of the total population of incidents (or people), N, and measures the distance from every incident point to every one of the K locations. The particular combination which gives the minimal sum of all distances is considered the best solution. In practice, however, solving this is computationally almost impossible, particularly if N is

large. For example, with 6000 incidents grouped into 20 partitions (clusters), one cannot solve this with any normal computer since there are

$$\frac{6000!}{20!\ 5980!} = 1.456 \times 10^{57}$$

combinations. No computer can solve that number and few spreadsheets can calculate the factorial of N greater than about 127.[5] In other words, it is almost impossible to solve computationally.

Practically, therefore, the different implementations of the K-means routine all make initial guesses about the K locations and then optimize the seating of this location in relation to the nearby points. This is called *local optimization.* Unfortunately, each K-means routine has a different way to define the initial locations so that two K-means procedures will usually not produce the same results, even if K is identical (Everitt, 1974; Systat, Inc., 1994).

CrimeStat K-means Routine

The K-means routine in *CrimeStat* also makes an initial guess about the K locations and then optimizes the distribution locally. The procedure that is adopted makes initial estimates about location of the K clusters (seeds), assigns all points to its nearest seed location, re-calculates a center for each cluster which becomes a new seed, and then repeats the procedure all over again. The procedure stops when there are very few changes to the cluster composition.[6]

K-means Output Files

The naming system for the K-means outputs is slightly different from the Nnh routine since there are no higher-order clusters. The final seed locations are displayed in the output table and can be saved as a '.dbf' file. For each of the K cluster groupings, the 1x and 2x standard deviational ellipses are calculated and can be output in *ArcView* '.shp', *MapInfo* '.mif' or *Atlas*GIS* '.bna' formats. Each file is named

<S>Km<*username*>

where S is the size of the ellipse (1x or 2x) and *username* is the name of the file provided by the user. For example,

1Kmrobbery

is the 1x standard deviational ellipse for a file called 'robbery' and

2Kmburglary

is the 2x standard deviational ellipse for a file called 'burglary'. Within the file, each ellipse is named

<S>KmEll<N><*username*>

where *S* is the size of the ellipse (1x or 2x), *N* is the ellipse number and *username* is the name of the file provided by the user. For example,

1KmEll3robbery

is the third ellipse for the 1x standard deviational ellipse of a file called 'robbery' and

2KmEll12burglary

is the 12th ellipse for the 2x standard deviational ellipse of a file called 'burglary'.

Example 2: K-means clustering of street robberies

In *CrimeStat*, the user specifies the number of groups to sub-divide the data. Using the 1996 robbery incidents for Baltimore County, the data were partitioned into 10 groups with the K-means routine (figure 6.7). As can be seen, the clusters tend to fall along the border with Baltimore City. But there are three more dispersed clusters, one concentrated in the central eastern part of the county and two north of the border with the City. Because these clusters are very large, a finer mesh clustering was conducting by partitioning the data into 35 clusters (figure 6.8). Though the ellipses are still larger than those produced by the nearest neighbor hierarchical procedure (see figure 6.4), there is some congruency; clusters identified by the nearest neighbor procedure have corresponding ellipses using the K-means procedure.

Figure 6.9 shows a section of southwest Baltimore County with 11 full ellipses and three partial ellipses that fall within this area. Looking at the distribution, several ellipses make intuitive sense while a couple of others do not. For example, the two ellipses that almost touch in the lower part of the figure highlight a concentration along a major arterial (U.S. Highway 40). Similarly, the three ellipses in the middle of the map which fall along a northeast-southwest axis identify incidents occurring along State Highway 26, a major arterial that also had many motor vehicle thefts (see chapter 5). On the other hand, a couple of the ellipses at the top of the view seem somewhat arbitrary; there are neither a high concentration of incidents together nor a defining roadway that would make such a concentration meaningful.

Advantages and Disadvantages of the K-means Procedure

In short, the K-means procedure will divide the data into the number of groups specified by the user. Whether these groups make any sense or not will depend on how

Figure 6.7: Baltimore County Robbery 'Hot Spots'

Using K-Means Routine with K=10 Clusters

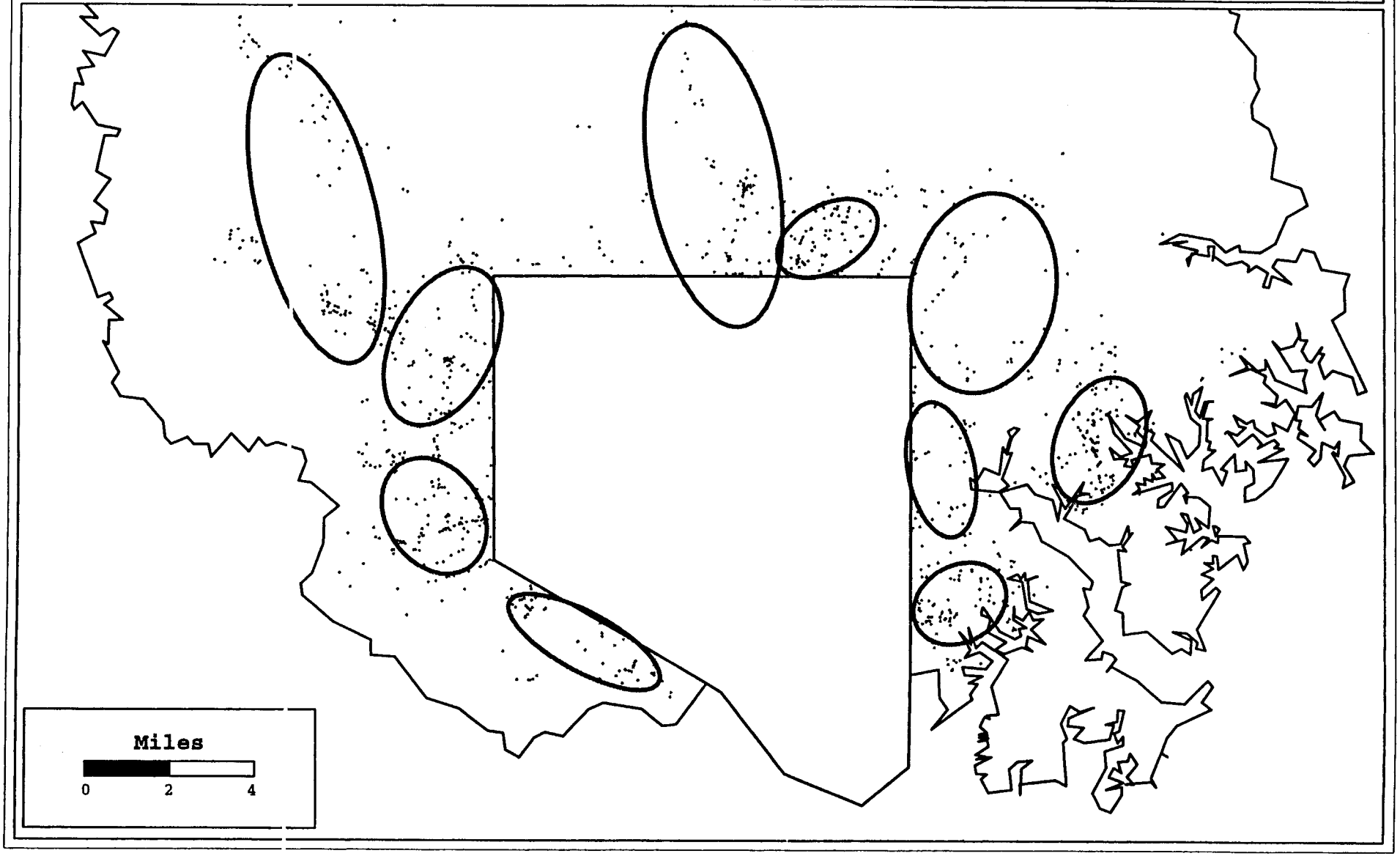


Figure 6.8: Baltimore County Robbery 'Hot Spots'

Using K-Means Routine with K=35 Clusters

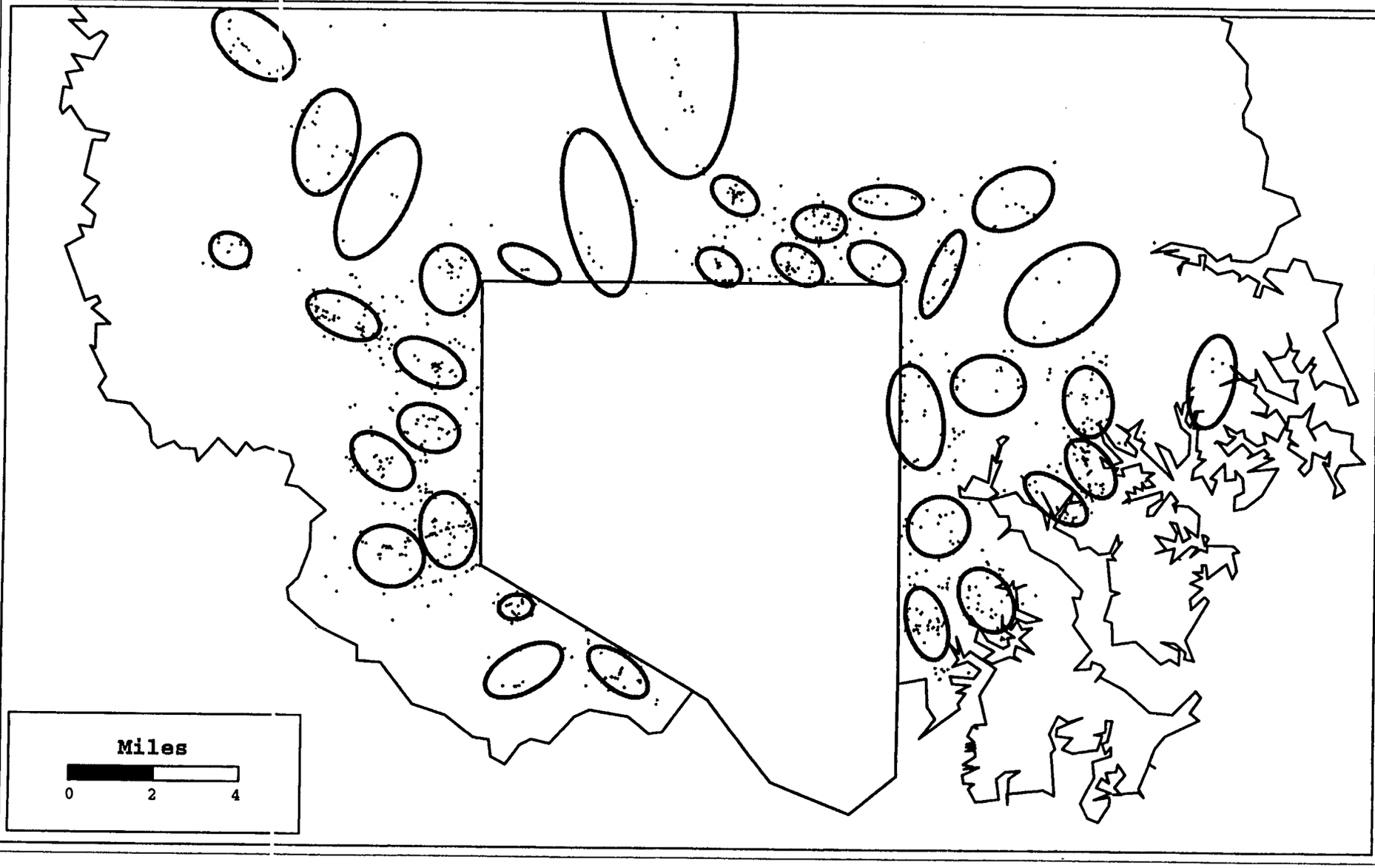
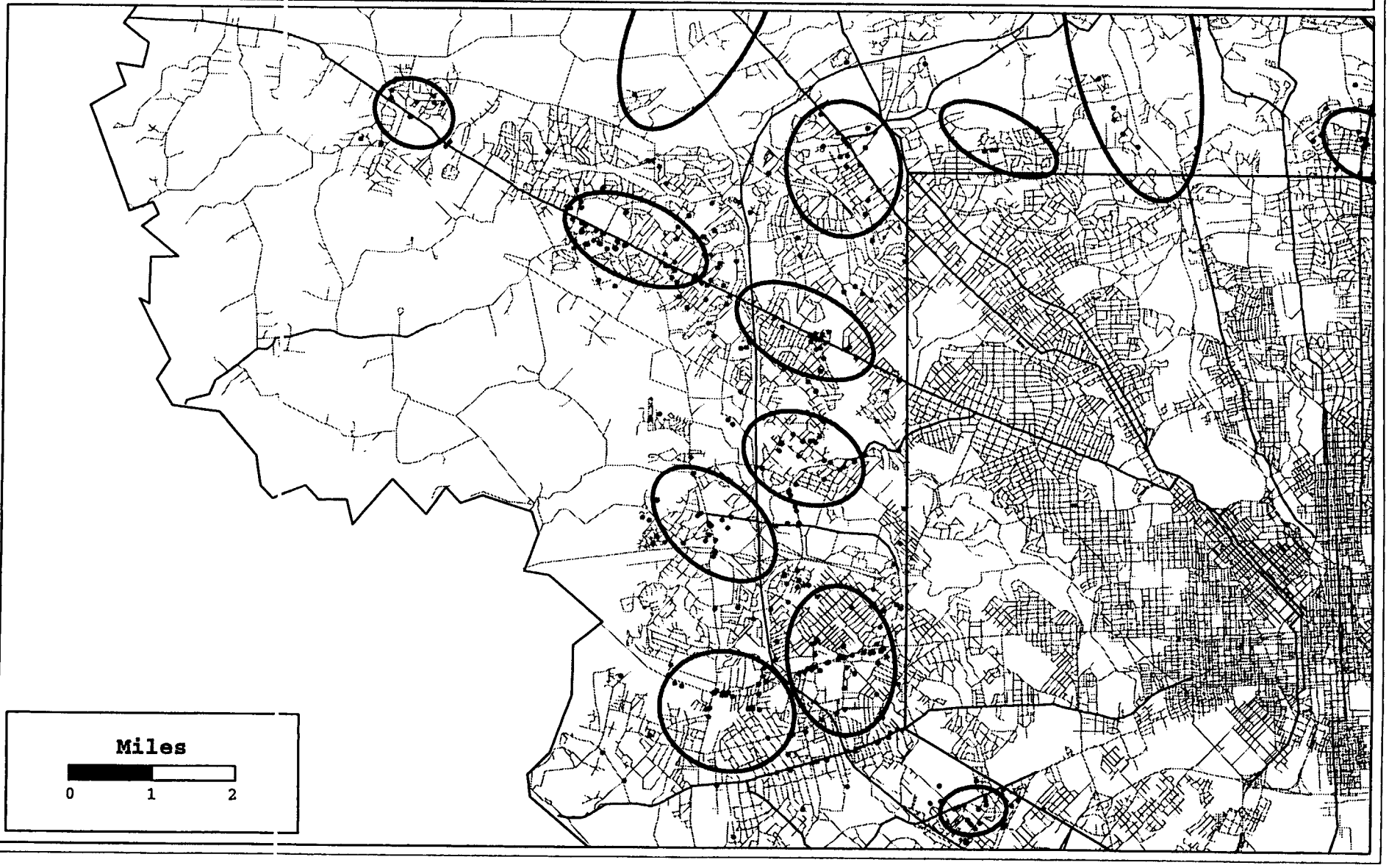


Figure 6.9: Southwest Baltimore County Robbery 'Hot Spots'

Using K-Means Routine with K=35 Clusters



carefully the user has selected clusters. Choosing too many will lead to defining patterns that don't really exist whereas choosing too few will lead to poor differentiation among neighborhoods that are distinctly different.

It is this choice that is both a strength of the technique as well as a weakness. The K-means procedure provides a great deal of control for the user and can be used as an exploratory tool to identify possible 'hot spots'. Whereas the nearest neighbor hierarchical method produces a solution based on geographical proximity with most clusters being very small, the K-means can allow the user to control the size of the clusters. In terms of policing, the K-means is better suited for defining larger geographical areas than the nearest neighbor method, perhaps more appropriate for a patrol area than for a particular 'hot spot'. Again, if carefully used, the K-means gives the user the ability to 'fine tune' a particular model of 'hot spots', adjusting the size of the clusters (vis-a-via the number of clusters selected) in order to fit a particular pattern which is known.

Yet it is this same flexible characteristic that makes the technique potentially difficult to use and prone to misuse. Since the technique will divide the data set into K groups, there is no assumption that these K groups represent real 'hot spots' or not. A user cannot just arbitrarily put in a number and expect it to produce meaningful results.

The technique is, therefore, better seen as both an exploratory tool as well as a tool for refining a 'hot spot' search. If the user has a good idea of where there should be 'hot spots', based on community experience and the reports of beat officers, then the technique can be used to see if the incidents actually correspond to the perception. It also can help identify 'hot spots' which have not been perceived or identified by officers. Alternatively, it can identify 'hot spots' that don't really exist and which are merely by-products of the statistical procedure. Experience and sensitivity are needed to know whether an identified 'hot spot' is real or not.

Local Moran Statistics (LMoran)

The third 'hot spot' technique in *CrimeStat* is a zonal technique called the *local Moran* statistics and was developed by Anselin (1995). Unlike the nearest neighbor hierarchical and K-means procedures, the local Moran statistics require data to be aggregated by zones, such as census block groups, zip codes, police reporting areas or other aggregations. The procedure applies Moran's I statistic to individual zones, allowing them to be identified as similar or different to their nearby pattern.

The basic concept is that of a *local indicator of spatial association* (*LISA*) and has been discussed by a number of researchers (Mantel, 1967; Getis, 1991; Anselin, 1995). For example, Anselin (1995) defines this as any statistic that satisfies two requirements:

1. The *LISA* for each observation indicates the extent to which there is significant spatial clustering of similar values around that observation; and

2. The sum of the *LISA*s for all observations is proportional to the global indicator of spatial association.

$$L_i = f(Y_i, Y_{Ji}) \quad (6.4)$$

where L_i is the local indicator, Y_i is the value of an intensity variable at location i, and Y_{Ji} are the values observed in the neighborhood J_i of i.

In other words, a *LISA* is an indicator of the extent to which the value of an observation is similar or different from its neighboring observations. This requires two conditions. First, that each observation has a variable value that can be assigned to it (i.e., an intensity or a weight) in addition to its X and Y coordinates. For crime incidents, this means data that are aggregated into zones (e.g., number of incidents by census tracts, zip codes, or police reporting districts). Second, the *neighborhood* has to be defined. This could be either adjacent zones or all other zones negatively weighted by the distance from the observation zone.

Once these are defined, the *LISA* indicates the value of the observation zone in relation to its neighborhood. Thus, in neighborhoods where there are 'high' intensity values, the *LISA* indicates whether a particular observation is similar (i.e., also 'high') or different (i.e., low) and, conversely, in neighborhoods where there are 'low' intensity values, the *LISA* indicates whether a particular observation is similar (i.e., also 'low') or different (i.e., 'high'). That is, the *LISA* is an indicator of similarity, not absolute value of the intensity variable.

Formal Definition of Local Moran Statistic

The I_i statistic

Anselin (1995) has applied the concept to a number of spatial autocorrelation statistics. The most commonly used, which is included in *CrimeStat*, is the Local Moran statistic, I_i, the use of Moran's I statistic as a *LISA*. The definition of I_i is (from Getis and Ord, 1996):

$$I_i = \frac{(Z_i - \bar{Z})}{S_Z^2} * \sum_{j=1}^{N} [W_{ij} * (Z_j - \bar{Z})] \quad (6.5)$$

where $\bar{Z}$ is the mean intensity over all observations, Z_i is the intensity of observation i, Z_j is intensity for all other observations, j (where $j \neq i$), S_Z^2 is the variance over all observations, and W_{ij} is a distance weight for the interaction between observations i and j. Note, the first term refers only to observation i, while the second term is the sum of the weighted values for all other observations (but not including i itself).

Distance weights

The weights, W_{ij}, can be either an indicator of the adjacency of a zone to the observation zone (i.e., '1' if adjacent; 0 if not adjacent) or a distance-based weight which decreases with distance between zones i and j. Adjacency indices are useful for defining near neighborhoods; the adjacent zones have full weight while all other zones have no weight. Distance weights, on the other hand, are useful for defining spatial interaction; zones which are farther away can have an influence on an observation zone, although one that is much less. *CrimeStat* uses distance weights, in two forms.

First, there is a traditional distance decay function:

$$W_{ij} = \frac{1}{d_{ij}} \quad \text{repeat} \tag{4.38}$$

where d_{ij} is the distance between the observation zone, i, and another zone, j. Thus, a zone which is two miles away has half the weight of a zone that is one mile away.

Small distance adjustment

Second, there is an adjustment for small distances. Depending on the distance scale used (miles, kilometers, meters), the weight index becomes problematic when the distance falls below 1 (i.e., below 1 mile, 1 kilometer); the weight then increases as the distance decreases, going to infinity for $d_{ij} = 0$. To correct for this, *CrimeStat* includes an adjustment for small distances so that the maximum weight can be never be greater than 1.0 (see chapter 4). The adjustment scales distances to one mile. When the small distance adjustment is turned on, the minimal distance is scaled automatically to be one mile. The formula used is

$$W_{ij} = \frac{\text{one mile}}{\text{one mile} + d_{ij}} \quad \text{repeat} \tag{4.39}$$

in whichever units are specified.

Similarity or dissimilarity

An exact test of significance has not been worked out because the distribution of the statistic is not known. The expected value of I_i and the variance of I_i are somewhat complicated (see endnote 7 for the formulas).[7] Instead, high positive or high negative standardized scores of I_i, $Z(I_i)$, are taken as indicators of similarity or dissimilarity. A high *positive* standardized score indicates the spatial clustering of similar values (either high or low) while a high *negative* standardized score indicates a clustering of dissimilar values

(high relative to a neighborhood that is low or, conversely, low relative to a neighborhood that is high). The higher the standardized score, the more the observation is similar (positive) or dissimilar (negative) to its neighbors.

In other words, the Local Moran statistic is a good indicator of either 'hot spots' or 'cold spots', zones which are different from their neighborhood. 'Hot spots' would be seen where the number of incidents in a zone is much higher than in the nearby zones. 'Cold spots' would be seen where the number of incidents in a zone is much lower than in the nearby zones. The Local Moran statistic indicates whether the zone is similar or dissimilar to its neighbors. A user must then look at the absolute value of the zone (i.e., the number of incidents in the zone) to see whether it is a 'hot spot' or a 'cold spot'.

For each observation, *CrimeStat* calculates the Local Moran statistic and the expected value of the Local Moran. *If* the *variance* box is checked, the program will also calculate the variance and the standardized Z-value of the Local Moran. The default is for the variance not to be calculated because the calculations are very intense and may take a long time. Therefore, a user should test how long it takes to calculate variances for a small sample on a particular computer before running the variance routine on a large sample.

Example 3: Local Moran statistics for auto thefts

Using data on 14,853 motor vehicle thefts for 1996 in both Baltimore County and Baltimore City, the number of incidents occurring in each of 1,349 census block groups was calculated with a GIS (Figure 6.10). As seen, the pattern shows a higher concentration towards the center of the metropolitan area, as would be expected, but that the pattern is not completely uniform. There are many block groups within the City of Baltimore with very low number of auto thefts and there are a number of block groups within the County with a very high number.

Using these data, *CrimeStat* calculated the Local Moran statistic with the variance box being checked and the small distance adjustment being used. The range of I_i values varied from -37.26 to +180.14 with a mean of 5.20. The pseudo-standardized Local Moran 'Z' varied from -12.71 to 50.12 with a mean of 1.61. Figure 6.11 maps the distribution. Because a negative I_i value indicates dissimilarity, these value have been drawn with a darker shade. As seen, in both the City of Baltimore and the County of Baltimore, there are block groups with large negative I_i values, indicating that they differ from their surrounding block groups. For example, in the central part of Baltimore City, there is a small area of about eight block groups with low numbers of auto thefts, compared to the surrounding block groups. These form a 'cold spot'. Consequently, they appear in dark tones in figure 6.11 indicating that they have high I_i values (i.e., negative autocorrelation). Similarly, there are several block groups on the western side of the County which have relatively high numbers of auto thefts compared to the surrounding block groups. They form a 'hot spot'. Consequently, they also appear in dark tones in figure 6.11 because this indicates negative spatial autocorrelation, having values that are dissimilar to the surrounding blocks.

Figure 6.10: 1996 Motor Vehicle Thefts

Number of Auto Thefts Per Block Group: Baltimore County and Baltimore City

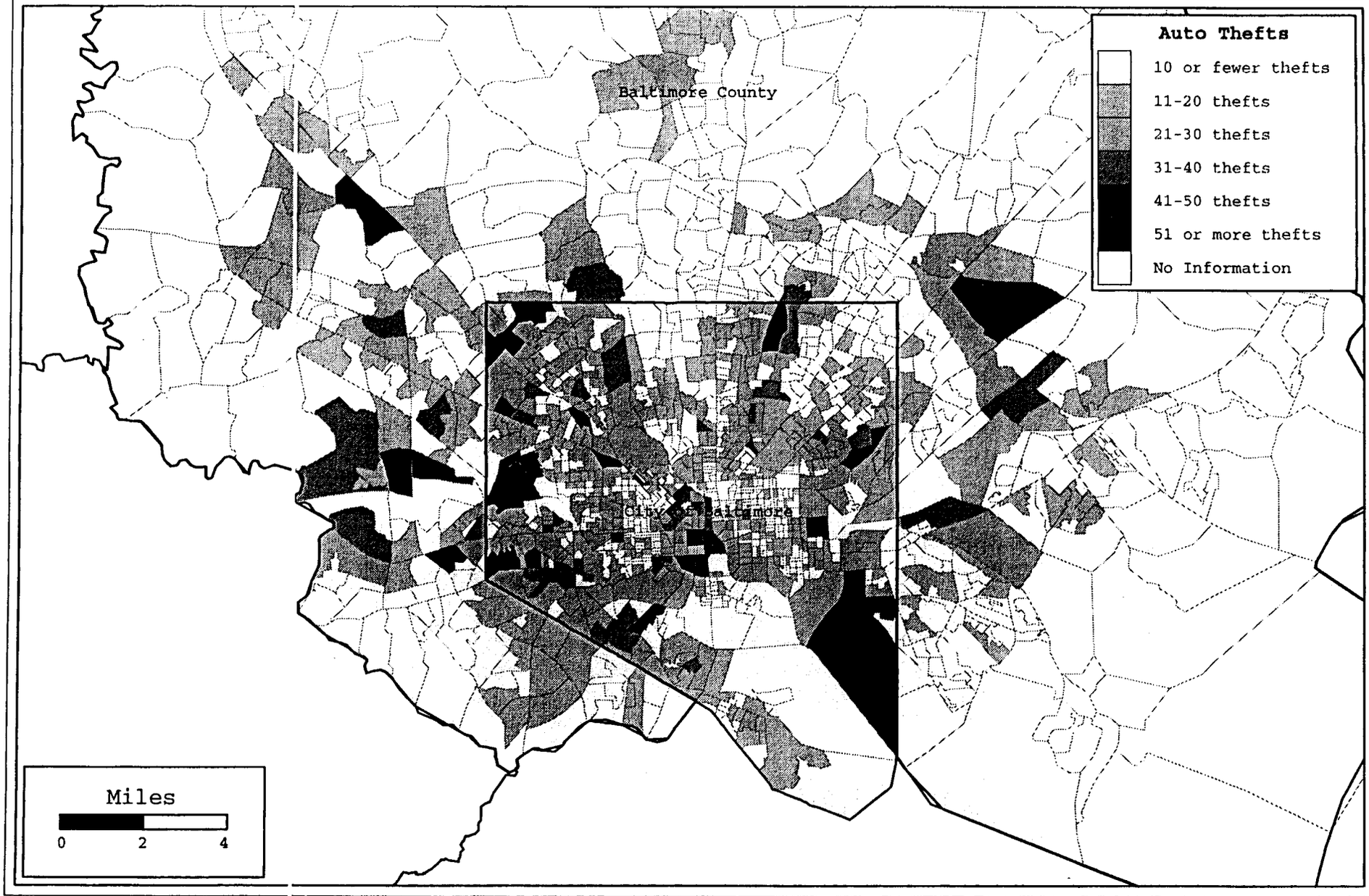
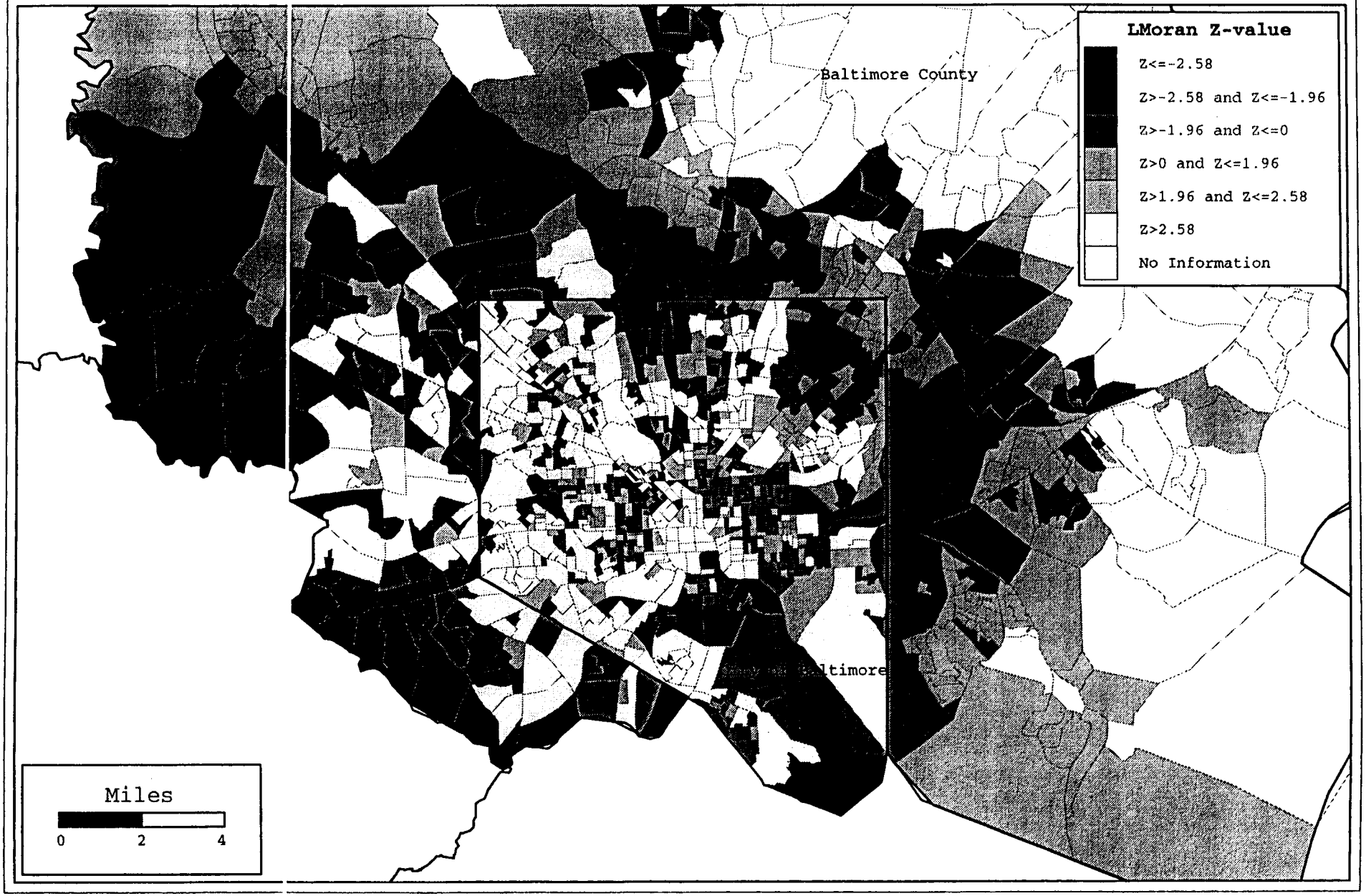


Figure 6.11: Spatial Autocorrelation of 1996 Auto Thefts

Local Moran Z-Value of Block Groups



In short, the Local Moran statistic can be a useful tool for identifying zones which are dissimilar from their neighborhood. It is the only statistic that is in *CrimeStat* that demonstrates dissimilarity. The other 'hot spot' tools will only identify areas with high concentrations. To use the Local Moran statistic, however, requires that the data be summarized into zones in order to produce the necessary intensity value. Given that most crime incident databases will list individual events without intensities, this will entail additional work by a law enforcement agency.

Some Thoughts on the Concept of 'Hot Spots'

Advantages

The three techniques discussed in this chapter have both advantages and disadvantages. Among the advantages are that they attempt to isolate areas of high concentration (or low concentration in the case of the Local Moran statistic) of incidents and can, therefore, help law enforcement agencies focus their resources on these areas. One of the powerful uses of a 'hot spot' concept is that it is focused. Given that most police departments are understaffed with limitations on resources, a strategy that prioritizes intervention is very appealing. The 'hot spot' concept is imminently practical.

Another advantage to the identification of 'hot spots' is that the techniques systematically implement an algorithm. In this sense, they minimize bias on the part of officers and analysts since the technique operates somewhat independently of preconceptions. As has been mentioned, however, these techniques are not totally without human judgement since the user must make decisions on the number of 'hot spots' and the size of the search radius, choices that can allow different users to come to different conclusions.

A third advantage is that these techniques are visual, particularly when used with a GIS. The two cluster analysis routines output ellipses that can be displayed in a GIS while the Local Moran technique can be adapted for thematic mapping (as Figure 6.11 demonstrates). Visual information can help crime analysts and officers to understand the distribution of crime in an areas, a necessary step in planning a successful intervention. We should never underestimate the importance of visualization in any analysis.

Disadvantages

However, there are also some distinct disadvantages to the concept of a 'hot spot', some technical and some theoretical. The choice involved in a user making a decision on how strict or how loose to create clusters allows the potential for subjectivity, as has been mentioned. In this sense, isolating clusters (or 'hot spots') can be as much an art as it is a science. There are limits to this, however. As the sample size goes up, there is less difference in the result that can be produced by adjusting the parameters. For example, with 6,000 or more cases, there is very little difference between using the 0.1 significance level in the nearest neighbor clustering routine and the 0.001 significance level.[8] Thus, the subjectivity of the user is more important for smaller samples than larger ones.

A second problem with the 'hot spot' concept is that it is applied almost exclusively to the volume of incidents and not to the underlying risk. Clusters (or 'hot spots') are defined by a high concentration of incidents within a small geographical area, that is on the volume of incidents within an area. This is an implicit *density* measure - the number of incidents per unit of area (e.g., incidents per square mile). But higher density can also be a function of a higher population at risk. In the central parts of all metropolitan areas, there are higher population densities; consequently, there will be more incidents per unit of area. For example, with burglaries, areas which have a higher density of housing units will usually have a higher number of incidents because of the high volume of 'targets' in the area, rather than because of any fundamental lack of safety. *Risk* which is defined as the number of incidents relative to the number of potential victims/targets is somewhat uncorrelated with the volume of incidents. Yet, 'hot spots' are almost always defined by volume, rather than risk.

For some policing policies, this is fine. For example, beat officers will necessarily concentrate on high incident density neighborhoods because so much of their activity revolves around those neighborhoods. From a viewpoint of providing concentrated policing, the density or volume of incidents is a good index for assigning police officers. From the viewpoint of ancillary security services, such as access to emergency medical services, neighborhood watch organizations, or residential burglar alarm retail outlets, areas with higher concentrations of incidents may be a good focal point for organizing these services.

But for other law enforcement policies, a density index is not a good one. From the viewpoint of crime prevention, for example, high incident volume areas are not necessarily unsafe and that effective preventive intervention will not necessarily lead to reduction in crime. It may be far more effective to target high risk areas rather than high volume areas. In high risk areas, there are special circumstances which expose the population to higher-than-expected levels of crime, perhaps particular concentrations of activities (e.g., drug trading) or particular land uses that encourage crime (e.g., skid row areas) or particular concentrations of criminal activities (e.g., gangs). A prevention strategy will want to focus on those special factors and try to reduce them.

Another law enforcement policy for which the volume or density of incidents is not the critical index is insurance, such as automobile insurance or property insurance. Theoretically, assuming there is not *redlining*[9] or other discriminatory practices, the cost of insurance should be proportional to the risk, not the volume. Thus, even though central city areas usually have a higher volume of crime incidents due to higher population concentration, the risk of auto theft or burglary is not necessarily higher. Again, in theory, insurance costs should be weighted in proportion to the risk of incidents, not the volume of incidents.

In short, a 'hot spot' is defined usually as an area with a high concentration of incidents, rather than as having higher risk, and is subject to the limitations of a volume measure. Of course, this is not only true for cluster ('hot spot') methods, but for most of the methods that are used in *CrimeStat*, such as the standard deviational ellipse (chapter 4), the nearest neighbor index (chapter 5) or the single density kernel estimate (chapter 7).

There are only a few measures that link the volume of incidents to the underlying population at risk (e.g., Ripley's K in chapter 5 and the dual density kernel estimate in chapter 7). Thus, the user needs to be aware of the distinction between volume and risk in applying these tools.

The final problem with the 'hot spot' concept is more theoretical. Namely, given a concentration of incidents, how do we explain it? To identify a concentration is one thing. To know how to intervene is another. It is imperative that the analyst discover some of the underlying causes that link the events together in a systematic way. Otherwise, all that is left is an empirical description without any concept of the underlying causes. For one thing, the concentration could be random or haphazard; it could have happened one time, but never again. For another, it could be due to the concentration of the population *at risk*, as discussed above. Finally, the concentration could be circumstantial and not be related to anything inherent about the location.

A public health example can illustrate this. Suppose that in a neighborhood five residents came down with a rare infectious disease, such as typhoid fever. Aside from ensuring treatment for the infected individuals, a public health worker will seek to find an underlying cause to the disease, what is known as a *vector* in public health circles. Did the individuals know each other and, hence, infect each other? Did the individuals shop at the same grocery store and, hence, purchase contaminated food? Did the individuals eat at the same restaurant and, hence, eat from the same infected pool of prepared food? Unless the vector that links these individuals is discovered, the five victims form only a potential 'hot spot', not necessarily a real one. Suppose that the five individuals had all been abroad in different countries and had, quite by accident, all contracted typhoid fever in completely different ways. In this case, the 'hot spot' is not real.

The point here is that an empirical description of a location where crime incidents are concentrated is only a first step in defining a real 'hot spot'. It is an *apparent* 'hot spot'. Unless the underlying vector (cause) is discovered, it will be difficult to provide adequate intervention. The causes could be environmental (e.g., concentrations of land uses that attract attackers and victims) or behavior (e.g., concentrations of gangs). The most one can do is try to increase the concentration of police officers. This is expensive, of course, and can only be done for limited periods. Eventually, if the underlying vector is not dealt with, incidents will continue and will overwhelm the additional police enforcement. In other words, ultimately, reducing crime around a 'hot spot' will need to involve many other policies than simply police enforcement, such as community involvement, gang intervention, land use modification, job creation, the expansion of services, and other community-based interventions. In this sense, the identification of an empirical 'hot spot' is frequently only a window into a much deeper problem that will involve more than targeted enforcement.

Endnotes for Chapter 6

1. Since we are only interested in the lower limit of the confidence interval, a one-tailed t-test is appropriate. Thus, if p is the probability, then p% of the distances from a spatially random sample will be greater than the threshold distance.

2. This is the next highest degree of freedom in the Student's t-table below infinity.

3. The particular steps are as follows:

   A. All distances between pairs of points are calculated, using either direct or indirect distance as defined on the measurements parameters page. The matrix is assumed to be symmetrical, that is the distance between A and B is assumed to be identical to the distance between B and A.

   B. The mean expected random distance is calculated using formula 5.2 and the threshold distance (the lower limit of the confidence interval for the corresponding t) is calculated using formula 6.2.

   C. All distance pairs smaller than the threshold distance are selected for clustering.

   D. For each incident point, the number of distances to other points which are smaller than the threshold distance are counted and placed in a *reduced matrix.* Any incident point which does not have another point within the threshold distance is not clustered. Any distance which is greater than the threshold distance is not considered for clustering.

   E. All points in the reduced matrix are sorted in descending order of the number of distances to other points shorter than the threshold distance, and the incident point with the largest number of below threshold distances is selected for the initial seed of the first cluster.

   F. All other incidents that are within the threshold distance of the initial seed point are selected for cluster 1.

   G. The number of points within the cluster are counted. If the number is equal to or greater than the minimum specified, then the cluster is kept. If the number is less than the minimum specified, then the cluster is dropped.

   H. For those clusters that are kept, the center of minimum distance (median center) is calculated for each to identify the cluster center.

   I. These points are removed from further clustering.

   J. Of the remaining points, the incident point with the largest number of

distances to other points shorter than the threshold distance is selected for the initial seed the second cluster.

K. All other points which are within the threshold distance of the first cluster seed point are selected for cluster 2.

L. The mean center of these selected points is calculated to identify the cluster center.

M. These points are removed from further clustering.

N. Steps J through M are repeated for all remaining points in the reduced matrix until no more points are remaining in the reduced matrix or until there are fewer than the specified minimum number of points for those remaining in the reduced matrix.

4. The steps are as follows:

A. Using the same p-values selected in the first-order, the mean random expected distance is calculated. However, the sample size is the number of first-order clusters identified, not the original number of points. Thus, the threshold distance is calculated by

$$\text{Lower Limit of Confidence Interval for Second-order Mean Random Distance} = 0.5\,\mathrm{SQRT}\left[\frac{A}{M}\right] - t\left[\frac{0.26136}{\mathrm{SQRT}\left[M^2/A\right]}\right] \quad (6.3)$$

where A is the area of the region and M is the number of first-order clusters identified during first-order clustering (i.e., not N). Thus, there is a different threshold distance for the second-order clustering. The t-value specified in the first-order clustering is maintained for second- and higher-order clustering.

B. All distances between first-order cluster centers are calculated and only those that are smaller than the second-order threshold distance are selected for second-order clustering.

C. If there are no distances between first-order cluster centers that are smaller than the second-order threshold distance, then the clustering process ends.

D. If there are distances between first-order cluster centers that are smaller than the second-order threshold distance, then the steps specified in endnote 3 are repeated to produce second-order clusters. A minimum of four first-order clusters is required to allow a second- or higher-order cluster.

E. If there are second-order clusters, then this process is repeated to either extract third-order clusters or to end the clustering process if no distances between second-order cluster centers are smaller than the (new) third-order threshold distance or if there are fewer than four new seeds in the cluster.

F. The process is repeated until no further clustering can be conducted, either all sub-clusters converge into a single cluster or the threshold distance criteria fails or there are fewer than four seeds in the higher-order cluster

5. The total number of ways for selecting K distinct combinations of *N* incidents, irrespective of order, is

$$\frac{N!}{K!\,(N-K)!}$$

From Burt and Barber, 1996, 155.

6. The steps are as follows:

Global Selection of Initial Seed Locations

A. A 100 x 100 grid is overlaid on the point distribution; the dimensions of the grid are defined by the minimum and maximum X and Y coordinates.

B. A separation distance is defined, which is

$$\text{Separation} = t * 0.5\ \text{SQRT}\left[\frac{A}{N}\right]$$

where *t* is the Student's t-value for the .01 significance level (2.358), *A* is the area of the region, and *N* is the sample size. The separation distance was calculated to prevent adjacent cells from being selected as seeds.

C. For each grid cell, the number of incidents found are counted and then sorted in descending order.

D. The cell with the highest number of incidents found is the initial seed for cluster 1.

E. The cell with the next highest number of incidents is temporarily selected. If the distance between that cell and the seed 1 location is *equal to or greater than* the separation distance, this cell becomes initial seed 2.

F. If the distance is less than the separation distance, the cell is dropped and the routine proceeds to the cell with the next highest number of incidents.

G. This procedure is repeated until *K initial seeds* have been located thereby selecting the remaining cell with the highest number of incidents and calculating its distance to all prior seeds. If the distance is equal to or greater than the separation distance, then the cell is selected as a seed. If the distance is less than the separation distance, then the cell is dropped as a seed candidate. Thus, it is possible that *K* initial seeds cannot be identified because of the inability to locate *K* locations greater than the threshold distance. In this case, *CrimeStat* keeps the number it has located and prints out a message to this effect.

Local Optimization of Seed Locations

H. After the *K* initial seeds have been selected, all points are assigned to the nearest initial seed location. These are the initial cluster groupings.

I. For each initial cluster grouping in turn, the center of minimum distance (median center) is calculated. These are the second seed locations.

J. All points are assigned to the nearest second seed location.

K. For each new cluster grouping in turn, the center of minimum distance is calculated. These are third seed locations.

L. Steps J and K are repeated until no more points change cluster groupings. These are the final seed locations and cluster groupings.

7. The formulas are as follows as follows. The expected value of the Local Moran is:

$$E(I_i) = \frac{-\sum_{j=1}^{N} W_{ij}}{N - 1}$$

where W_{ij} is a distance weight for the interaction between observations i and j (either an adjacency index or a weight decreasing with distance). The variance of the Local Moran is defined in three steps:

A. First, define b_2.

$$b_2 = \sum \left\{ \frac{(X_i - \bar{X})^4}{N} \right\} \Big/ \left[\sum \left\{ \frac{(X_i - \bar{X})^2}{N} \right\} \right]^2$$

This is the fourth moment around the mean divided by the squared second moment around the mean.

B. Second, define $2w_{i(kh)}$:

$$2w_{i(kh)} = \sum \sum W_{ik} W_{ih} \qquad \text{where } k \neq i \text{ and } h \neq i$$

This term is twice the sum of the cross-products of all weights for i with themselves, using k and h to avoid the use of identical subscripts. Since each pair of observations, i and j, has its own specific weight, a cross-product of weights are two weights multiplied by each other (where $i \neq j$) and the sum of these cross-products is twice the sum of all possible interactions irrespective of order (i.e., $W_{ij} = W_{ji}$). Because the weight of an observation with itself is zero (i.e., $W_{ii} = 0$), all terms can be included in the summation.

C. Third, define the variance, standard deviation, and an approximate (pseudo) standardized score of I_i:

$$\text{Var}(I_i) = \frac{(\sum w_{ij}^2)*(n - b_2)}{(n-1)} + \frac{2w_{i(kh)}(2b_2 - n)}{(n-1)(n-2)} + \frac{(\sum w_{ij})^2}{(n-1)^2}$$

$$S(I_i) = \sqrt{[\text{Var}(I_i)]}$$

$$Z(I_i) = [I_i - E(I_i)] / S(I_i)$$

8. On one test of 6,051 burglaries with a minimum cluster size requirement of 10 incidents, for example, we obtained 100 first-order clusters, 9 second-order clusters, and no third-order clusters by using a 0.1 significance level for the nearest neighbor hierarchical clustering routine. When the significance level was reduced to 0.001, the number of clusters extracted was 97 first-order clusters, 8 second-order clusters, and no third-order clusters.

9. A practice by which financial institutions limit the number of loans, either residential or commercial, that are made in low income neighborhoods in order to minimize the risk of the borrower defaulting on the loan or insurance companies limiting the number of policies issued in low income neighborhoods to minimize payment risks. However, this practice implicitly is discriminatory because of the higher concentration of minorities in low income areas.

Chapter 7
Kernel Density Interpolation

In this last chapter, we discuss tools aimed at interpolating incidents, using the kernel density approach. *Interpolation* is a technique for generalizing incident locations to an entire area. Whereas the other statistics that are used in *CrimeStat* provide statistical summaries for the data incidents themselves, interpolation techniques generalize those data incidents to the entire region. In particular, they provide *density* estimates for all parts of a region (i.e., at any location). The density estimate is an intensity variable, a Z-value, that is estimated at a particular location. Consequently, it can be displayed by either surface maps or contour maps that show the intensity at all locations.

There are many interpolation techniques, such as Kriging, trend surfaces, local regression models (e.g., Loess, splines), and Dirichlet tessellations (Anselin, 1992; Cleveland, Grosse and Shyu, 1993; Venables and Ripley, 1997). Most of these require a variable that is being estimated as a function of location. However, *kernel density estimation* is an interpolation technique that is appropriate for individual point locations (Silverman, 1986; Härdle, 1991; Bailey and Gatrell, 1995; Burt and Barber, 1996; Bowman and Azalini, 1997).

Kernel Density Estimation

Kernel density estimation involves placing a symmetrical surface over each point, evaluating the distance from the point to a reference location based on a mathematical function, and summing the value of all the surfaces for that reference location. This procedure is repeated for all reference locations. It is a technique that was developed in the late 1950s as an alternative method for estimating the density of a histogram (Rosenblatt, 1956; Whittle, 1958; Parzen, 1962). A histogram is a graphic representation of a frequency distribution. A continuous variable is divided into intervals of size, s (the interval or bin width), and the number of cases in each interval (bin) are counted and displayed as block diagrams. The histogram is assumed to represent a smooth, underlying distribution (a density function). However, in order to estimate a smooth density function from the histogram, traditionally researchers have linked adjacent variable intervals by connecting the midpoints of the intervals with a series of lines (Figure 7.1).

Unfortunately, doing this causes three statistical problems (Bowman and Azalini, 1997):

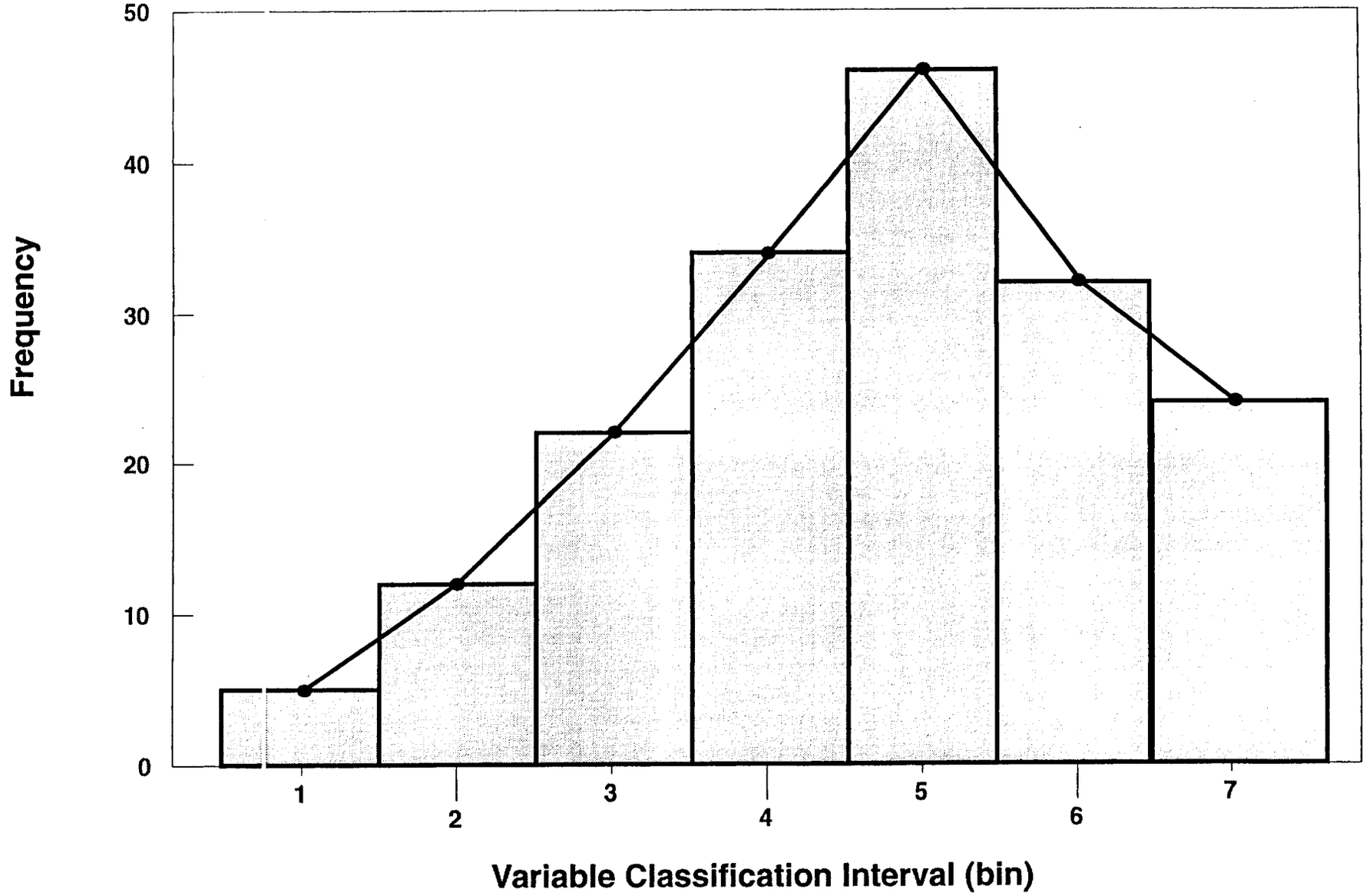
1. Information is discarded because all cases within an interval are assigned to the midpoint. The wider the interval, the greater the information loss.

2. The technique of connecting the midpoints leads to a discontinuous and not smooth density function even though the underlying density function is assumed to be smooth. To compensate for this, researchers will reduce the width of the interval. Thus, the density function becomes smoother with

Figure 7.1:

Constructing A Density Estimate From A Histogram

Method of Connecting Midpoints



smaller interval widths, although still not very smooth. Further, there are limits to this technique as the sample size decreases when the bin width gets smaller, eventually becoming too small to produce reliable estimates.

3. The technique is dependent on an arbitrarily defined interval size (bin width). By making the interval wider, the estimator becomes cruder and, conversely, by making the interval narrower, the estimator becomes finer. However, the underlying density distribution is assumed to be smooth and continuous and not dependent on the interval size of a histogram.

To handle this problem, Rosenblatt (1956), Whittle (1958) and Parzen (1962) developed the kernel density method in order to avoid the first two of these difficulties; the bin width issue still remains. What they did was to place a smooth *kernel function*, rather than a block, over each point and sum the functions for each location on the scale. Figure 7.2 illustrates the process with five point locations. As seen, over each location, a symmetrical kernel function is placed; by symmetrical is meant that is falls off with distance from each point at an equal rate in both directions around each point. In this case, it is a normal distribution, but other types of symmetrical distribution have been used. The underlying density distribution is estimated by summing the individual kernel functions at *all* locations to produce a smooth cumulative density function. Notice that the functions are summed at every point along the scale and not just at the point locations. The advantages of this are that, first, each point contributes equally to the density surface and, second, the resulting density function is continuous at all points along the scale.

The third problem mentioned above, interval size, still remains since the width of the kernel function can be varied. In the kernel density literature, this is called *bandwidth* and refers essentially to the width of the kernel. Figure 7.3 shows a kernel with a narrow bandwidth placed over the same five points while figure 7.4 shows a kernel with a wider bandwidth placed over the points. Clearly, the smoothness of the resulting density function is a consequence of the bandwidth size.

There are a number of different kernel functions that have been used, aside from the normal distribution, such as a triangular function (Burt and Barber, 1996) or a quartic function (Bailey and Gatrell, 1995). Figure 7.5 illustrates a quartic function. But the normal is the most commonly used (Kelsall and Diggle, 1995a). The normal distribution function has the following functional form:

$$g(x_j) = \sum \left\{ [W_i * I_i] * \frac{1}{h^2 * 2\pi} * e^{-\left[\frac{d_{ij}^2}{2*h^2}\right]} \right\} \qquad (7.1)$$

where d_{ij} is the distance between an incident location and any reference point in the region, h is the standard deviation of the normal distribution (the bandwidth), W_i is a weight at the point location and I_i is an intensity at the point location. This function extends to infinity in all directions and, thus, will be applied to any location in the region.

Figure: 7.2

Kernel Density Estimates

Summing of Normal Kernel Functions for 5 Points

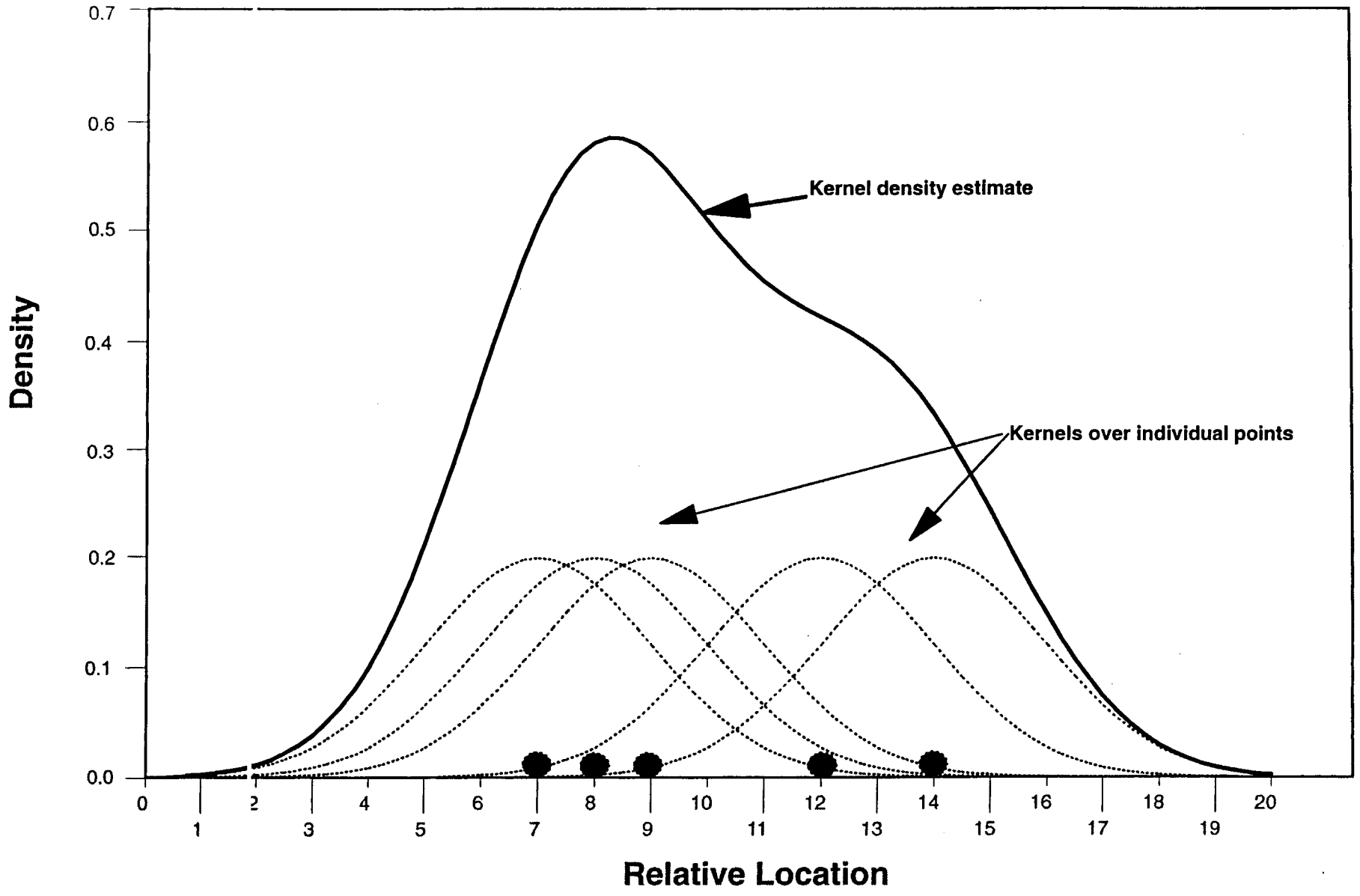


Figure 7.3:

Kernel Density Estimates

Smaller Bandwidth

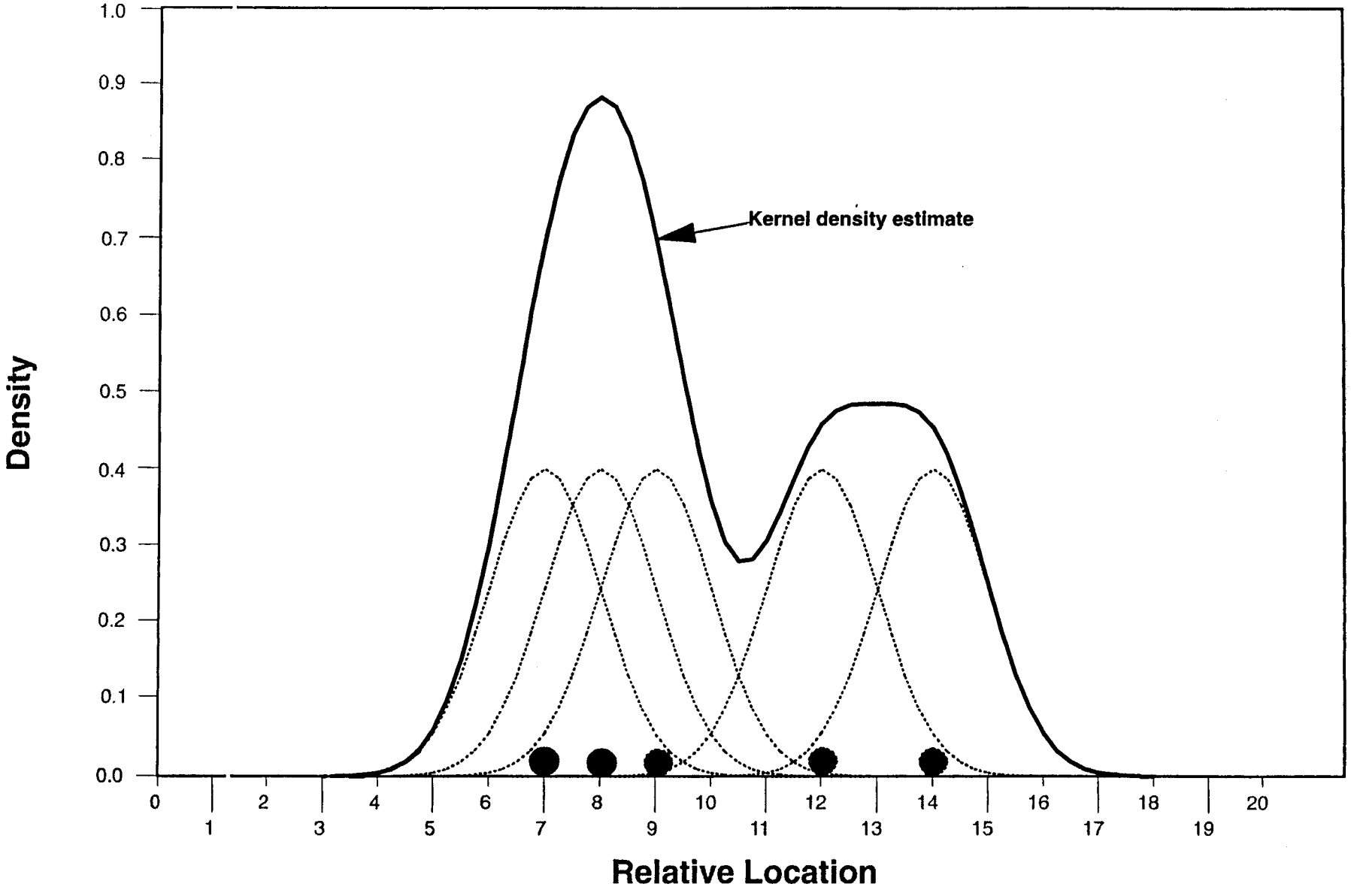


Figure 7.4:

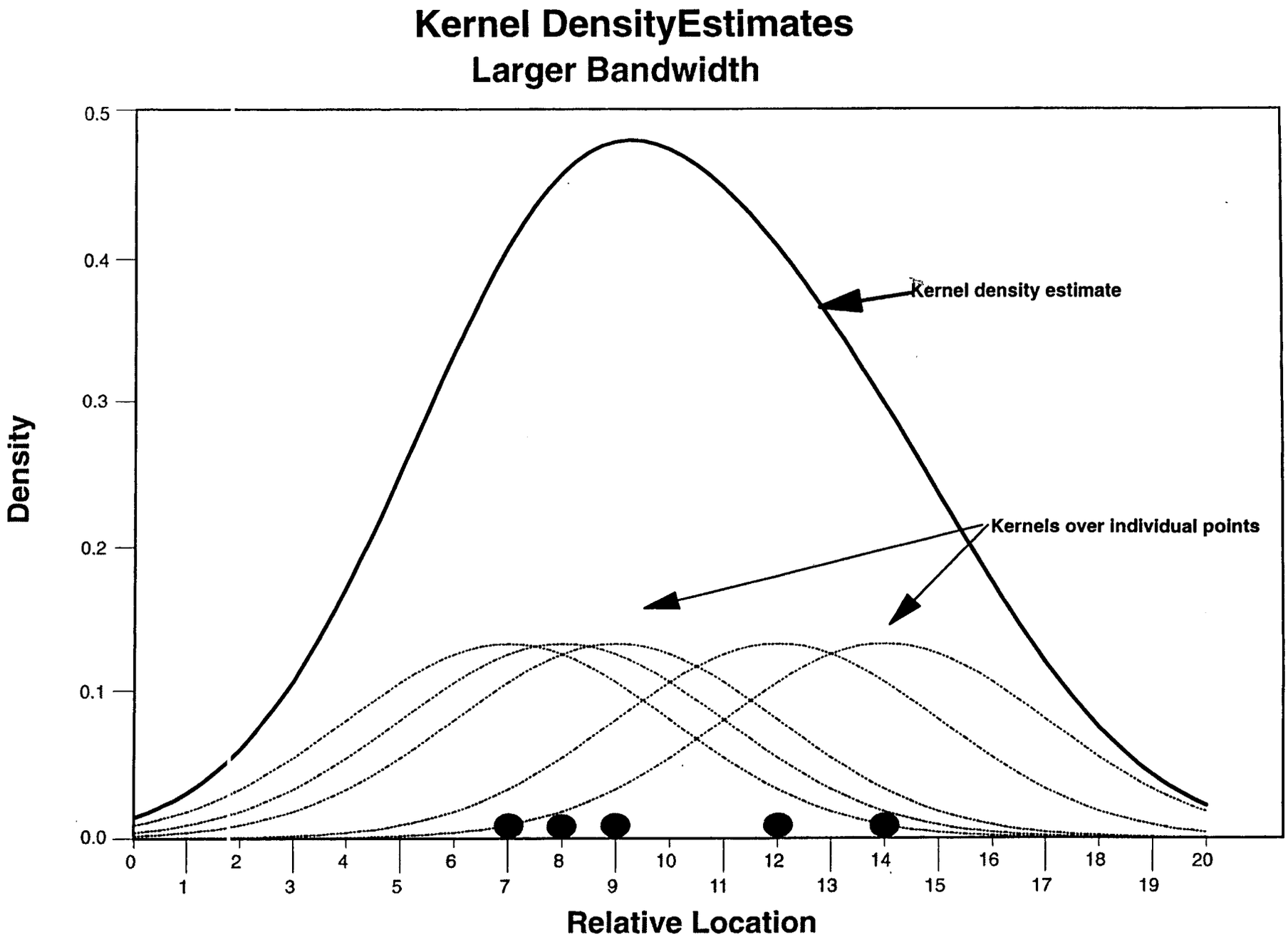
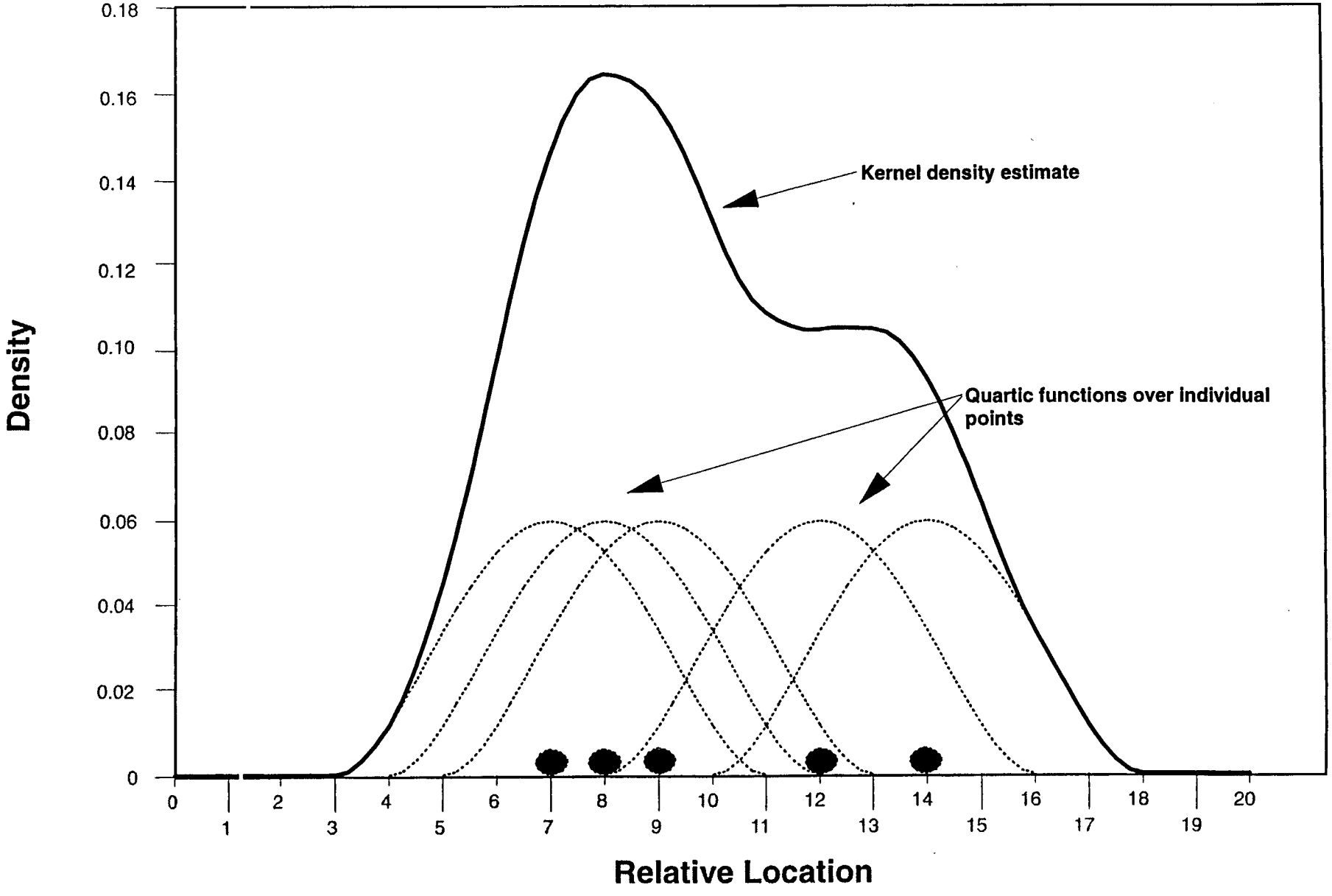


Figure 7.5

Kernel Density Estimates
Summing of Quartic Kernel Function



On the other hand, a quartic function has a circumscribed radius and is, therefore, applied to a limited area around each incident point, h. Its functional form is:

I. Outside the specified radius, h:

$$g(x_j) = 0 \tag{7.2}$$

II. Within the specified radius, h:

$$g(x_j) = \sum\left\{ [W_i * I_i] * \left[\frac{3}{h^2 * \pi}\right] * \left[1 - \frac{d_{ij}^2}{h^2}\right]^2 \right\} \tag{7.3}$$

where d_{ij} is the distance between an incident location and any reference point in the region, h is the radius of the search area (the bandwidth), W_i is a weight at the point location and I_i is an intensity at the point location. Other functions have also been used. However, Silverman (1986) has argued that it does not make that much difference as long as the kernel is symmetrical. There are also edge effects that can occur and there have been different proposed solutions to this problem (Venables and Ripley, 1997).

There have also been variations of the size of the of bandwidth with various formulas and criteria (Silverman, 1986; Härdle, 1991; Venables and Ripley, 1997). Generally, bandwidth choice fall into either fixed or adaptive (variable) choices (Kelsall and Diggle, 1995a; Bailey and Gatrell, 1995). *CrimeStat* follows this distinction, which will be explained below.

The kernel function can be expanded to more than two dimensions (Härdle, 1991; Bailey and Gatrell, 1995; Burt and Barber, 1996; Bowman and Azalini, 1997). Figure 7.6 shows a three-dimensional normal distribution placed over each of five points with the resulting density surface being a sum of all five individual surfaces. Thus, the method is particularly appropriate for geographical data, such as crime incident locations. The method has also been developed to relate two or more variables together by applying a kernel estimate to each variable in turn and then dividing one by the other to produce a three-dimensional estimate of *risk* (Kelsall and Diggle, 1995a; Bowman and Azalini, 1997).

Significance testing of density estimates is more complicated. Current techniques tend to focus on simulating surfaces under spatially random assumptions (Bowman and Azaline, 1997; Kelsall and Diggle, 1995b). Because of the still experimental nature of the testing, *CrimeStat* does not include any testing of density estimates in this version.

CrimeStat Kernel Density Methods

CrimeStat has two interpolation techniques, both based on the kernel density technique. The first applies to a single variable, while the second to the relationship between two variables. Both routines have a number of options. Figure 7.7 shows the interpolation page in *CrimeStat*. Users indicate their choices by clicking on the tab and

Figure 7.6:

Kernel Density Surfaces

Summing of Normal Kernel Surfaces for 5 Points

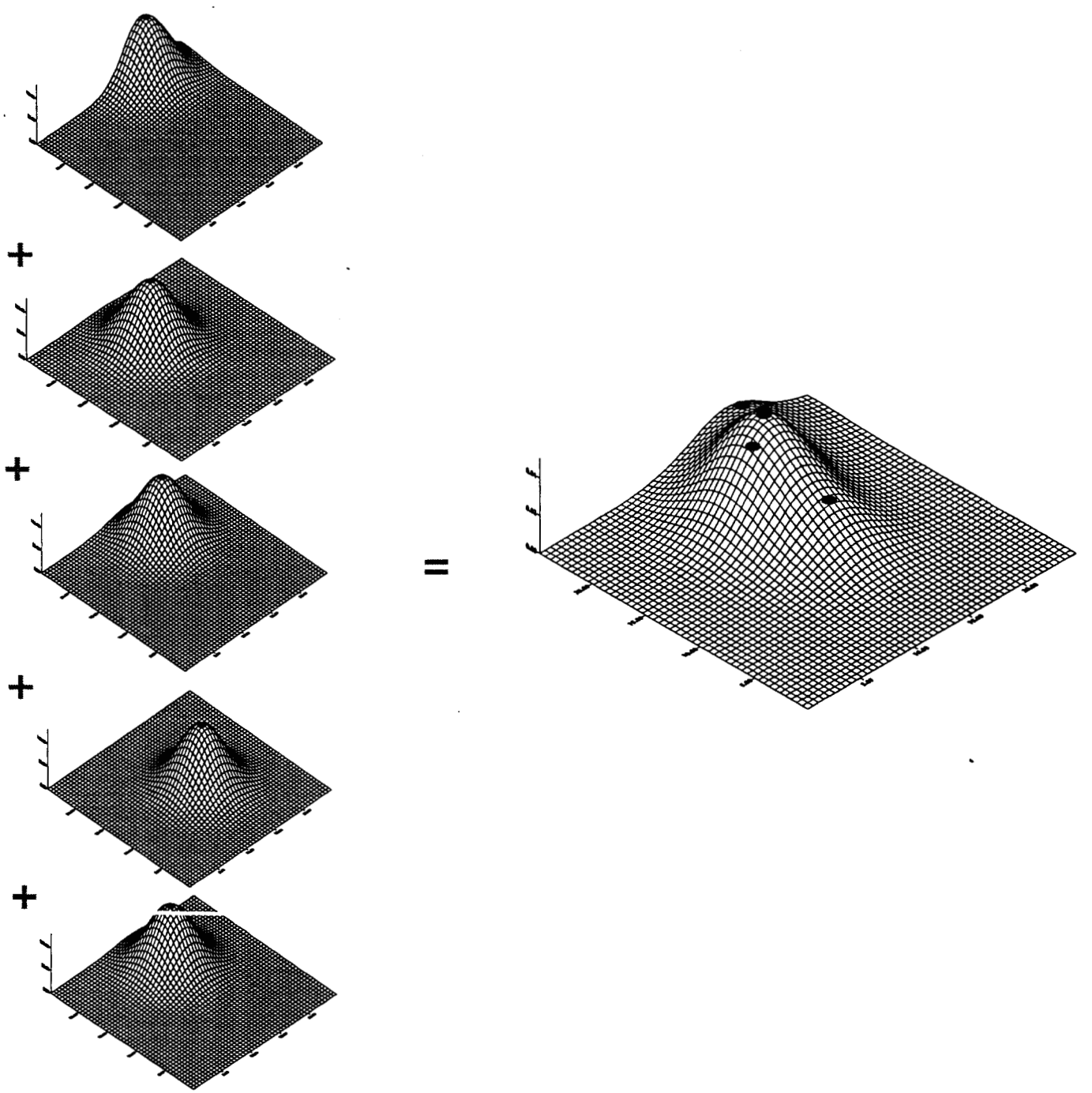


Figure 7.7: Kernel Density Interpolation Layout

menu items. For either technique, it is necessary to have a reference file, which is usually a grid placed over the study region (see chapter 3). The reference file represents the region to which the kernel estimate will be generalized (figure 7.8).

Single Density Estimates

The single kernel density routine in *CrimeStat* is applied to a distribution of point locations, such as crime incidents. It can be used with either a primary file or a secondary file; the primary file is the default. For example, the primary file can be the location of motor vehicle thefts. The points can also have a weighting or an associated intensity variable (or both). For example, the points could represent the location of police stations while the weights (or intensities) represent the number of calls for service. Again, the user must be careful in having both a weighting variable and an intensity variable as the routine will use both variables in calculating densities; this could lead to double weighting.

Having defined the file on the primary (or secondary) file tabs, the user indicates the routine by checking the 'Single' box. Also, it is necessary to define a reference file, either an existing file or one generated by *CrimeStat* (see chapter 3). There are other parameters that must be defined.

File to be Interpolated

The user must indicate whether the primary file or the secondary file (if used) is to be interpolated.

Method of Interpolation

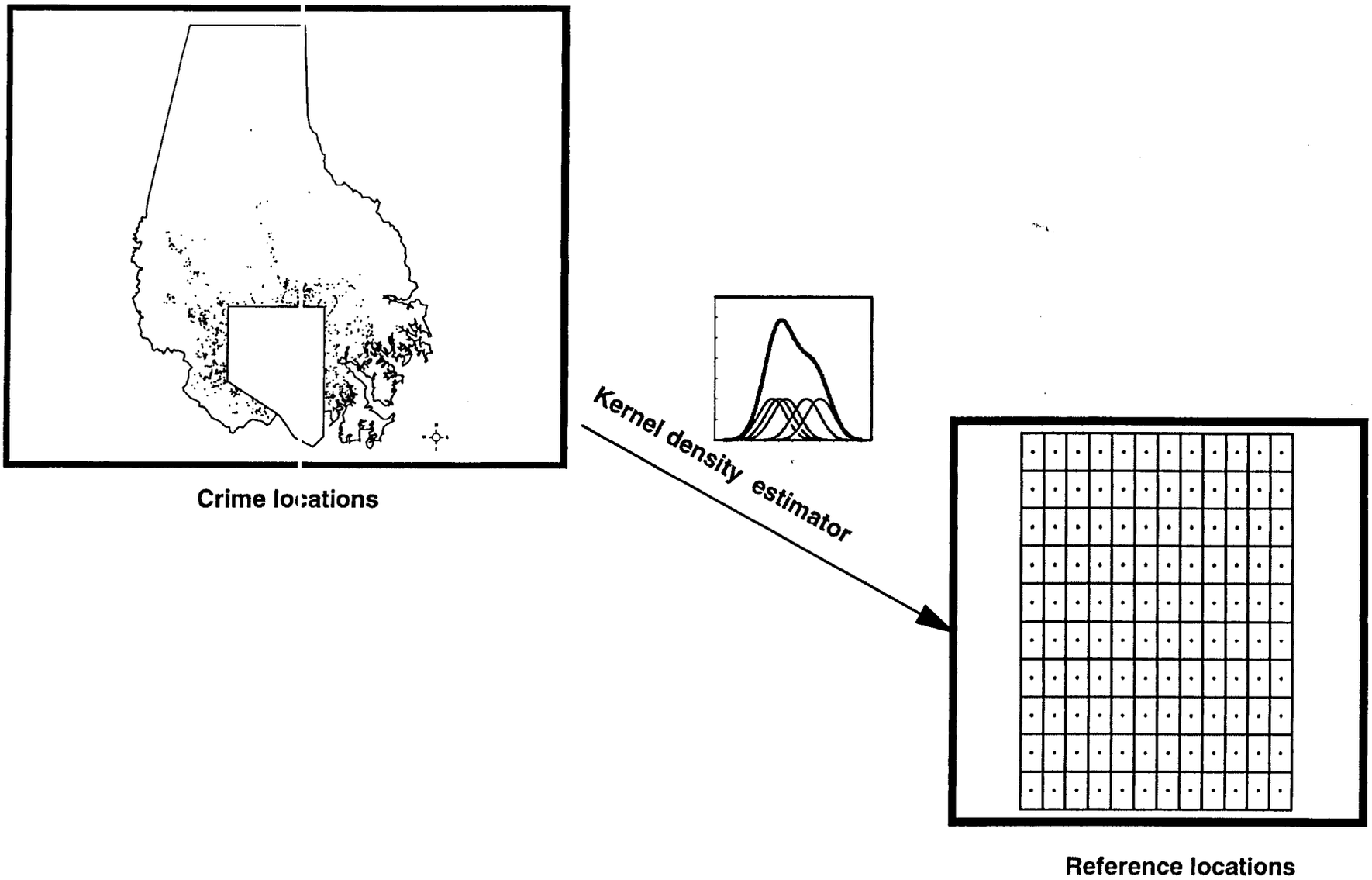
The user must indicate the method of interpolation. Two types of kernel density estimators are used, a normal distribution and a quartic distribution with the normal being the default. In our experience, there are advantages to each. The normal distribution produces an estimate over the entire region whereas the quartic produces estimates only for the circumscribed bandwidth radius. If the distribution of points is sparse towards the outer parts of the region, then the quartic will not produce estimates for those areas, whereas the normal will. Conversely, the normal distribution can cause some edge effects to occur (e.g., spikes at the edge of the reference grid), particularly if there are many points near one of the boundaries of the study area. The quartic will produce less of a problem at the edges, although it still can produce some spikes.

Choice of Bandwidth

The user must indicate how bandwidths are to be defined. There are two types of bandwidth for the single kernel density routine, fixed interval or adaptive interval.

Figure 7.8:

Kernel Density Estimation



Fixed interval

With a fixed bandwidth, the user must specify the interval to be used and the units of measurement (squared miles, squared nautical miles, squared feet, squared kilometers, or squared meters). Depending on the type of kernel estimate used, this interval has a slightly different meaning. For the normal kernel function, the bandwidth is the standard deviation of the normal distribution. For the quartic kernel, the bandwidth is the radius of the search area to be interpolated.

There are few guidelines for choosing a particular bandwidth other than by visual inspection (Venables and Ripley, 1997). A narrower bandwidth interval will lead to a finer mesh density estimate, with all the little peaks and valleys. A larger bandwidth interval, on the other hand, will lead to a smoother distribution and, therefore, less variability between areas. While the more variable estimate shows greater differentiation among areas (e.g., between 'hot spot' and 'low spot' zones), one has to keep in mind the statistical precision of the estimate. If the sample size is not very large, then a smaller bandwidth will lead to more imprecision in the estimates; the peaks and valleys may be nothing more than random variation. On the other hand, if the sample size is large, then a finer density estimate can be produced. In general, it is a good idea to experiment with different fixed intervals to see which results make the most sense.

Adaptive interval

An adaptive bandwidth adjusts the bandwidth interval so that a minimum number of points are found. This has the advantage of providing constant precision of the estimate over the entire region. Thus, in areas that have a high concentration of points, the bandwidth is narrow whereas in areas where the concentration of points is more sparse, the bandwidth will be larger. This is the default bandwidth choice in *CrimeStat* since we believe that consistency in statistical precision is paramount. The degree of precision is generally dependent on the sample size of the bandwidth interval. The default is a minimum of 100 points within the bandwidth radius. The user can make the estimate more fine grained by choosing a smaller number of points (e.g., 25) or more smooth by choosing a larger number of points (e.g., 200). Again, experimentation is necessary to see which results make the most sense.

Output Units

The user must indicate the measurement units for the density estimate in points per squared miles, squared nautical miles, squared feet, squared kilometers, or squared meters. The default is points per square mile.

Intensity or Weighting Variables

If an intensity or weighting variable is used, these boxes must be checked. Be careful about using both an intensity and a weighting variable to avoid 'double weighting'.

Calculations

The user must indicate the type of output for the density estimates. There are two types of calculation that can be conducted with the kernel density routine. The calculations are applied to each reference cell. First, the kernel estimates can be calculated as *density* estimates using formulas 7.1 or 7.2/7.3, depending on what type of kernel function is used. The estimates at each reference cell are re-scaled so that the sum of the densities over all reference grids equals the total number of incidents; this is the default value. Second, the densities can be converted into *probabilities* by dividing the density at any one cell by the total number of incidents.

Since the two types of calculation are directly interrelated, the output surface will not differ in its variability. The choice would depend on whether the calculations are used to estimate densities or probabilities. For comparisons between different types of crime or between the same type of crime and different time periods, usually densities are the unit of choice (i.e., incidents per unit of area). However, to express the output as a probability, that is, the likelihood that an incident would occur at any one location, then outputing the results as probabilities would make more sense. For display purposes, however, it makes no difference as both look the same.

Output Files

Finally, the results can be displayed in an output table or can be output into two formats: 1) Raster grid formats for display in a surface mapping program- *Surfer for Windows* '.dat' format (Golden Software, 1994) or *ArcView Spatial Analyst* 'asc' format (ESRI, 1998); or 2) Polygon grids in *ArcView* '.shp', *MapInfo* '.mif' or *Atlas*GIS* '.bna' formats.[1] However, all but *Surfer for Windows* require that the reference grid be created by *CrimeStat.*

Example 1: Kernel density estimate of street robberies

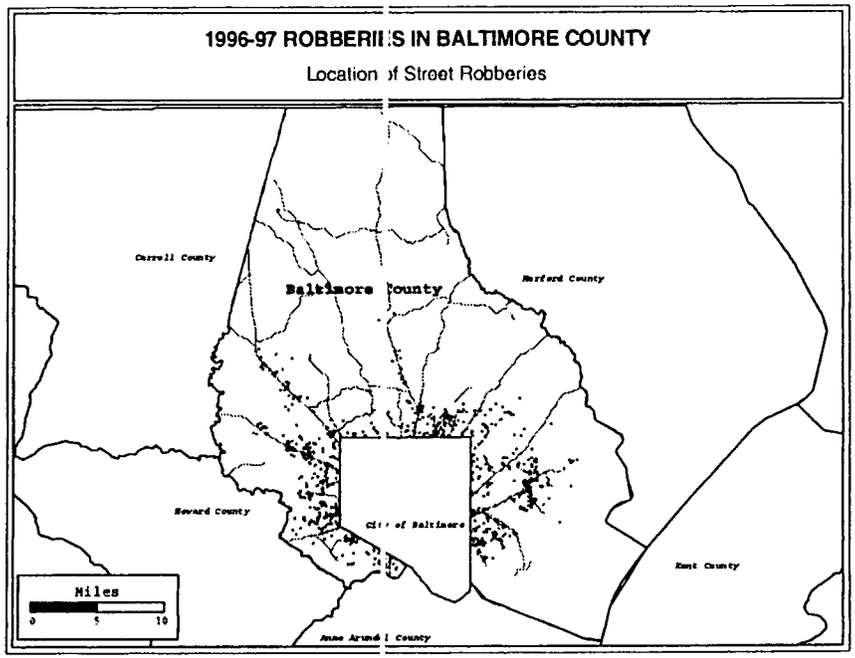
An example can illustrate the use of the single kernel density routine. Figure 7.9 shows a *Surfer for Windows* output of the 1180 street robberies for 1996 in Baltimore County. The reference grid was generated by *CrimeStat* and had 100 columns and 108 rows. Thus, the routine calculated the distance between each of the 10,800 reference cells and the 1180 robbery incident locations, evaluated the kernel function for each measured distance, and summed the results for each reference cell. The normal distribution kernel function was selected for the kernel estimator and an adaptive bandwidth with a minimum sample size of 100 was chosen as the parameters.

There are three views in the figure: 1) a map view showing the location of the incidents; 2) a surface view showing a three-dimensional interpolation of robbery density; and 3) a contour view showing contours of high robbery density. The surface and contour views provide different perspectives. The surface shows the peaks very clearly and the relative density of the peaks. As can be seen, the peak for robberies on the eastern part of the County is much higher than the two peaks in the central and western parts of the

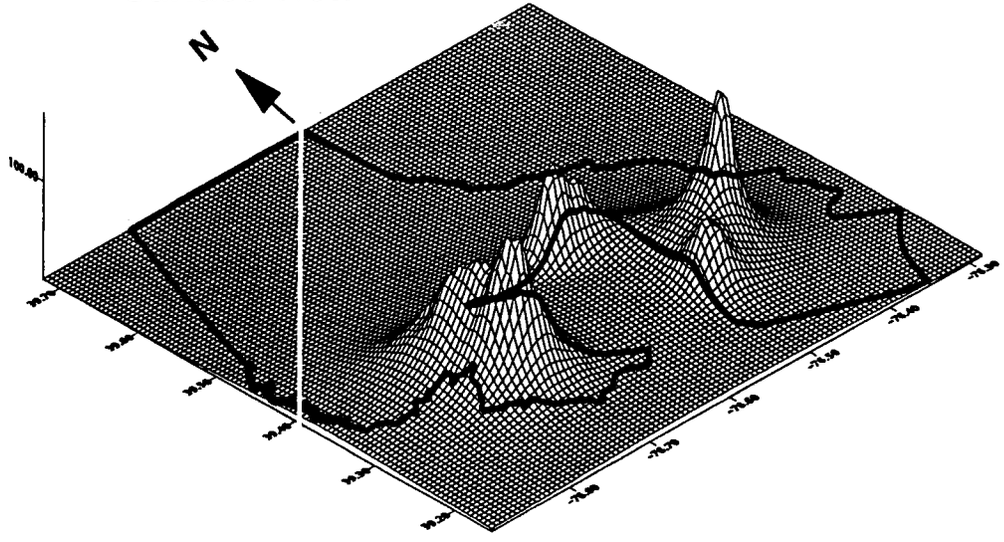
Figure 7.9:

Baltimore County Robberies: 1996-97

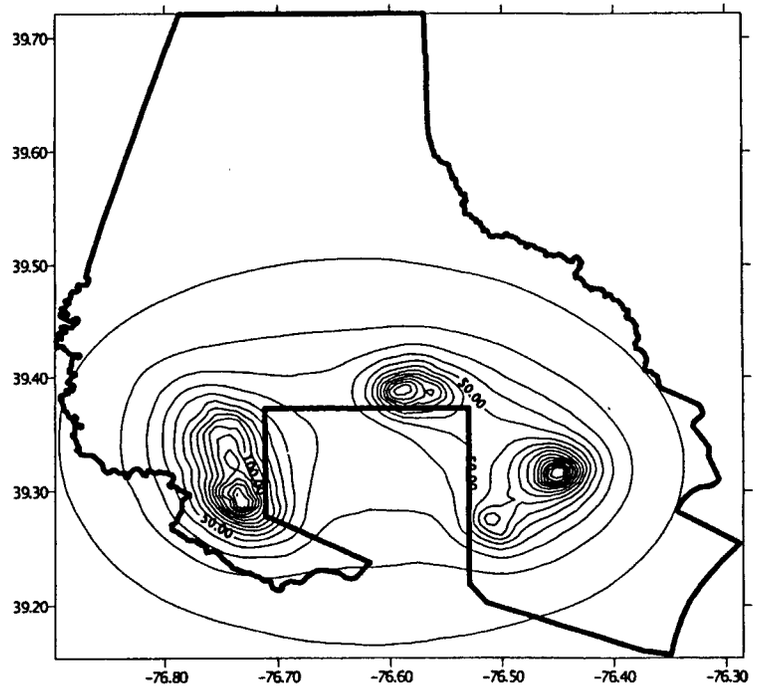
Kernel Density Interpolation



Surface View



Contour View



County. The contour view can show where these peaks are located; it is difficult to identify location clearly from a three-dimensional surface map. Highways and streets could be overlaid on top of the contour view to identify more precisely where these peaks are located.

Figure 7.10 shows an *ArcViewSpatial Analyst* map of the robbery density with the robbery incident locations overlaid on top of the density contours. Here, we can see quite clearly that there are three strong concentrations of incidents, one spreading over a distance of several miles on the west side, one on northern border between Baltimore City and Baltimore County, and one on the east side; there is also one smaller peak in the southeast corner of the County.

From a statistical perspective, the kernel estimate is a better 'hot spot' identifier than the cluster analysis routines discussed in chapter 6. Cluster routines group incidents into clusters and distinguish between incidents which belong to the cluster and those which do not belong. Depending on which mathematical algorithms are used, different clustering routines will return differing allocations of incidents to clusters. The kernel estimate, on the other hand, is a continuous surface; the densities are calculated at *all* locations; thus, the user can visually inspect the variability in density and decide what to call a 'hot spot' without having to define arbitrarily where to cut-off the 'hot spot' zone.

Going back to the *Surfer for Windows* output, figure 7.11 shows the effects of varying the bandwidth parameters. There are three fixed bandwidth intervals (0.5, 1, and 2 miles respectively) and there are two adaptive bandwidth intervals (a minimum of 25 and 100 points respectively). As can be seen, the fineness of the interpolation is affected by the bandwidth choice. For the three fixed intervals, an interval of 0.5 miles produces a finer mesh interpolation than an interval of 2 miles, which tends to 'oversmooth' the distribution. Perhaps, the intermediate interval of 1 mile gives the best balance between fineness and generality. For the two adaptive intervals, the minimum sample size of 25 gives some very specific peak locations whereas the adaptive interval with a minimum sample size of 100 gives a smoother distribution. Which of these should be used as the *best* choice would depend on how much confidence the analyst has in the results. A key question is whether the 'peaks' are real or merely byproducts of small sample sizes. The best choice would be to produce an interpolation which fits the experience of the department and officers who travel an area. Again, experimentation and discussions with beat officers will be necessary to establish which bandwidth choice should be used in future interpolations.

Note in all five of the interpolations, there is some bias at the edges with the City of Baltimore (the three-sided area in the central southern part of the map). Since the primary file only included incidents for the County, the interpolation nevertheless has estimated some likelihood at the edges; these are *edge biases* and need to be ignored or removed with an ASCII editor.[2] Further, the wider the interval chosen, the more bias is produced at the edge.

Figure 7.10:

Baltimore County Robberies: 1996-97

Robberies Per Square Mile

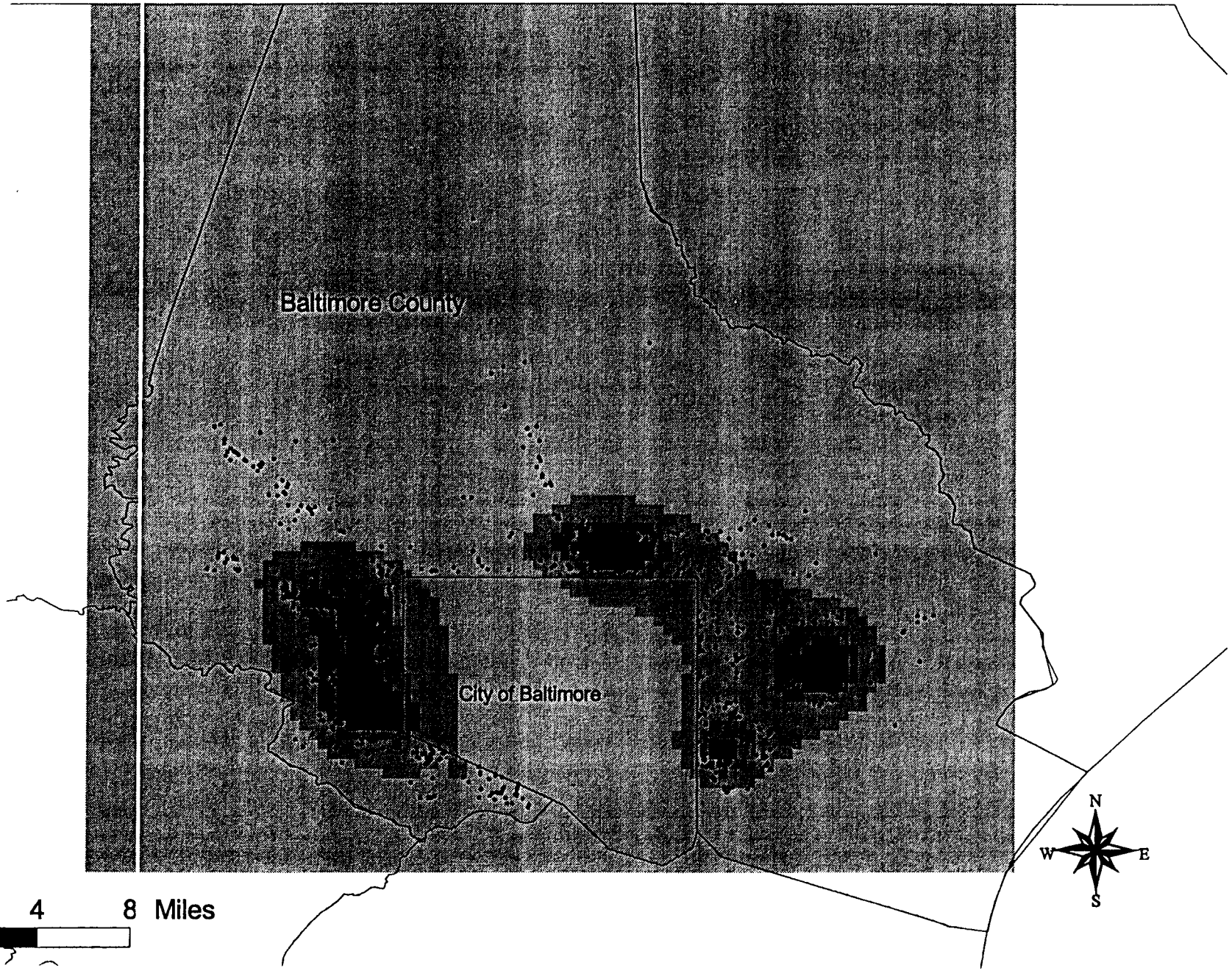
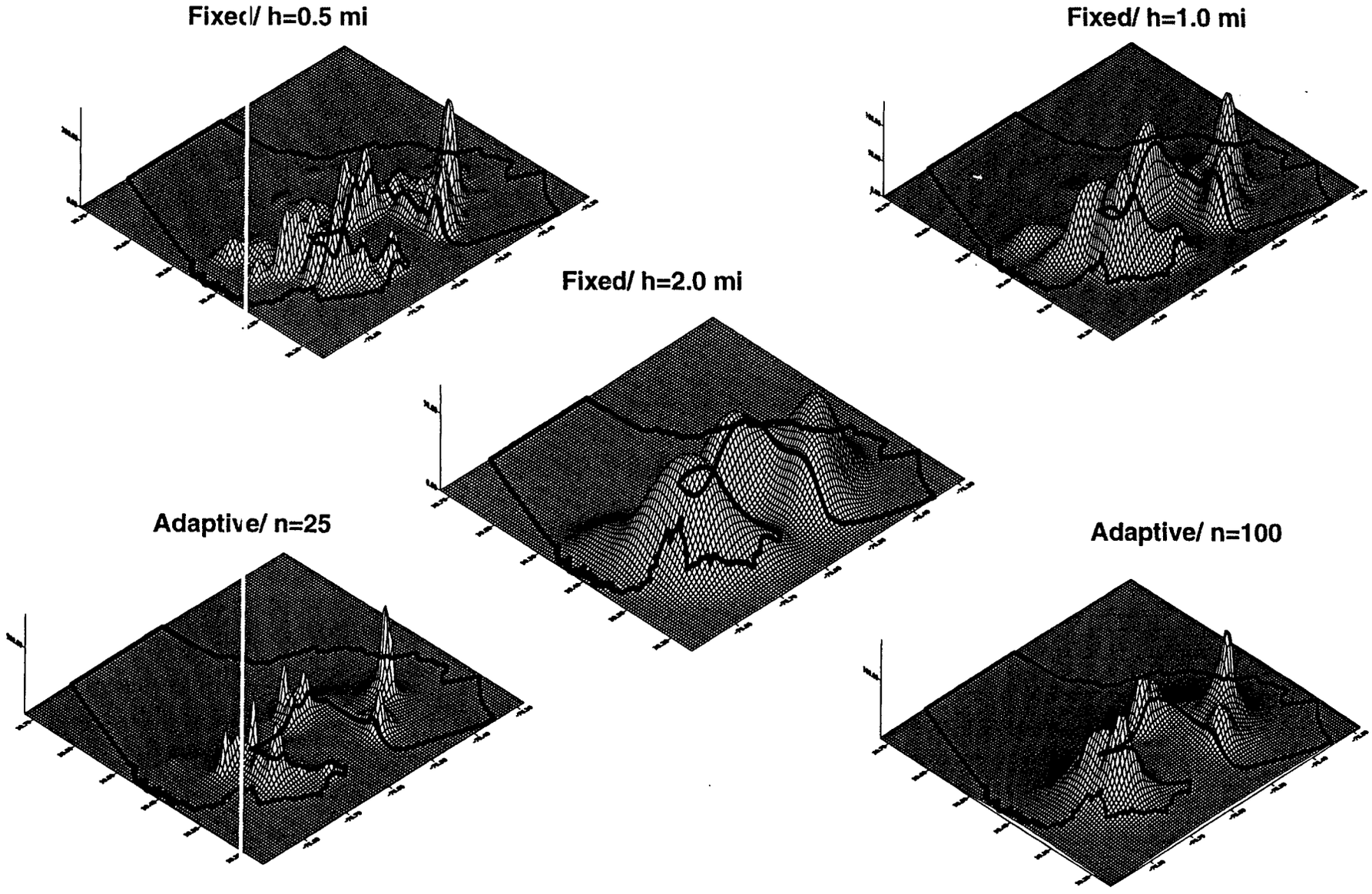


Figure 7.11:

Interpolation of Baltimore County Robberies: 1996
Different Bandwidths



Dual Density Estimates

The dual kernel density routine in *CrimeStat* is applied to *two* distributions of point locations. For example, the primary file could be the location of auto thefts while the secondary file could be the centroids of census tracts, with the population of the census tract being an intensity variable. The dual routine must be used with *both* a primary file *and* a secondary file. Also, it is necessary to define a reference file, either an existing file or one generated by *CrimeStat* (see chapter 3). Several parameters need to be defined.

File to be Interpolated

The user must indicate the order of the interpolation. The routine uses the language *first* file and *second* file in making the comparison (e.g., dividing the first file by the second; adding the first file to the second). The user must indicate which is the first file, the primary or the secondary. The default is that the primary file is the first file.

Method of Interpolation

The user must indicate the type of kernel estimator. As with the single kernel density routine, two types of kernel density estimators are used, a normal distribution and a quartic distribution with the normal being the default.

Choice of Bandwidth

The user must define the bandwidth parameter. There are three types of bandwidths for the single kernel density routine - fixed interval, variable interval, or adaptive interval.

Fixed interval

With a fixed bandwidth, the user must specify the interval to be used and the units of measurement (squared miles, squared nautical miles, squared feet, squared kilometers, or squared meters). Depending on the type of kernel estimate used, this interval has a slightly different meaning. For the normal kernel function, the bandwidth is the standard deviation of the normal distribution. For the quartic kernel, the bandwidth is the radius of the search area to be interpolated. Since there are two files being compared, the fixed interval is applied both to the first file and the second file.

Variable interval

With a variable interval, each file (the first and the second) have different intervals. For both, the units of measurements must be specified (squared miles, squared nautical miles, squared feet, squared kilometers, or squared meters). There is a good reason why a user might want variable intervals. In comparing two kernel estimates, the most common comparison is to divide one by the other. However, if the density estimate for a particular cell in the denominator approaches zero, then the ratio will blow up and become a very

large number. Visually, this will be seen as spikes in the distribution, the result, usually, of too few cases. In this case, the user might decide to smooth the denominator more than numerator in order to reduce these spikes. For example, the interval for the first file (the numerator) could be 1 mile whereas the interval for the second file (the denominator) could be 3 miles. Experimentation will be necessary to see whether this is warranted. But, in our experience, it frequently happens when either there are two few cases or there is an irregular boundary to the region with a number of incidents grouped at one of the edges.

Adaptive interval

An adaptive bandwidth adjusts the bandwidth interval so that a minimum number of points (sample size) is found. This sample size is applied to both the first file and the second file. It has the advantage of providing constant precision of the kernel estimate over the entire region. Thus, in areas that have a high concentration of points, the bandwidth is narrow whereas in areas where the concentration of points is more sparse, the bandwidth will be larger. This is the default bandwidth choice in *CrimeStat* since consistency in statistical precision is important. The degree of precision is generally dependent on the sample size of the bandwidth interval. The default is a minimum of 100 points. The user can make the estimate finer by choosing a smaller number of points (e.g., 25) or smoother by choosing a larger number of points (e.g., 200).

Output Units

The user must indicate the measurement units for the density estimate in points per squared miles, squared nautical miles, squared feet, squared kilometers, or squared meters.

Intensity or Weighting Variables

If an intensity or weighting variable is used for either the first file or the second file, these boxes must be checked. Be careful about using both an intensity and a weighting variable to avoid 'double weighting'.

Calculations

The user must indicate the type of density output. There are four types of calculations that can be conducted with the dual kernel density routine. The calculations are applied to each reference cell. There is the *ratio of densities*, that is the first file divided by the second file. This is the default choice. For example, if the first file is the location of auto thefts incidents and the second file is the location of census tract centroids with the population assigned as an intensity variable, then ratio of densities would divide the kernel estimate for auto thefts by the kernel estimate for population and would be an estimate of auto thefts risk.

There is also the *log ratio of densities*. This is the natural logarithm of the density ratio, that is

$$\text{Log ratio of densities} = \text{Ln}\,[\, g(x_j) / g(y_j)\,] \qquad (7.4)$$

where $g(x_j)$ is the density estimate for the first file and $g(y_j)$ is the density estimate for the second file. For a variable that has a spatially skewed distribution, such that most reference cells have very low density estimates, but a few have very high density estimates, converting the ratio into a log function will tend to mute the spikes that occur. This measure has been used in studies of risk (Kelsall and Diggle, 1995b).

There is the *difference in densities*, that is the first file minus the second file. This can be a useful output for examining differential effects. For example, by using the centroids of census block groups (see example 2 below) with the population of the census block group assigned as an intensity or weighting variable, there is a slight bias produced by the spatial arrangements of the block groups. The U. S. Census Bureau suggests that census units (e.g., census tracts, census block groups) be drawn so that there are approximately equal populations in each unit. Thus, block groups towards the center of the metropolitan area tend to be smaller because there is a higher population density at those locations. Thus, the spatial arrangement of the block groups will tend to produce a kernel estimate which has a higher value towards the center independent of the actual population of the block group; the bias is very small, less than 0.1%, but it does exist. A more precise estimate could be produced by subtracting the kernel estimate for the block group centroids *without* using population as the intensity variable from the kernel estimate for the block group centroids *with* population as the intensity variable. The resulting output could then be read back into *CrimeStat* and used as a more precise measure of population distribution. There are other uses of the difference function, such as subtracting the estimate for the population-at-risk from the incident distribution rather than taking the ratio.

There is the *sum of the densities*, that is, the density estimate for the first file plus the density estimate for the second file. Again, this is applied to each reference cell at a time. A possible use of the sum operation is to combine two different density surfaces, for example the density of robberies plus the density of assaults.

Output Files

Finally, the user must specify the file formats for the output. The results can be output in three forms. First, the results are displayed in an output table. Second, the results can be output into two raster grid formats for display in a surface mapping program: *Surfer for Windows* format as a '.dat' file (Golden Software, 1994) and *ArcView Spatial Analyst* format as a '.asc' file (ESRI, 1998). Third, the results can be output as polygon grids into *ArcView* '.shp', *MapInfo* '.mif' and *Atlas*GIS* '.bna' format (see footnote 1). All but *Surfer for Windows* require that the reference grid be created by *CrimeStat.*

interpolating a second variable to the same reference grid, the two variables have been interpolated to the same geographical units. The two interpolations can then be related, by dividing, subtracting, or summing. As has been mentioned throughout this manual, one of the problems with techniques that depend on the concentration of incidents is that they ignore the underlying population-at-risk. With the dual routine, however, we can start to examine the risk and not just the concentration. The comparison does not have to be between an incident distribution and a population-at-risk. It can be between two different types of crime incidents, for example, figure 7.15 shows the ratio of 1996 motor vehicle thefts to 1996 street robberies. 'Peaks', in this example, represent locations where there is a predominance of auto thefts relative to robberies whereas 'troughs' represent locations that have a predominance of robberies relative to auto thefts. Comparing two different crime distributions can allow an identification of areas where there is a concentration of one type of crime but not others.

Conclusion

Kernel density estimation is one of the 'cutting edge' spatial statistical techniques. There is currently research on the use of this technique in both the statistical theory and in developing applications. For crime analysis, the technique represents a powerful way of conducting both 'hot spot' analysis as well as being able to link the 'hot spots' to an underlying population-at-risk. It can be used both for police deployment by targeting areas of high concentration of incidents as well as for prevention by targeting areas with high risk. It can also be used as a research tool for analyzing two or more distributions. As was mentioned, many of the statistical properties are still being developed by statisticians, particularly significance testing. But, over the next few years, these will become widely used tools for crime analysis and crime research.

Example 2: Kernel density estimates of auto thefts relative to population

As an example of the use of the dual kernel density routine, the duel routine is applied in both the City of Baltimore and the County of Baltimore to 14,853 motor vehicle theft locations for 1996 relative to the 1990 population of census block groups. Again, a reference grid of 100 columns by 108 rows was generated by *CrimeStat.*

Figure 7.12 shows the resulting single kernel density estimate as a *Surfer for Windows* output; again, there is a map view, a surface view, and a contour view. The normal kernel function was used and an adaptive bandwidth of 100 points was selected. As can be seen, there is a very high concentration of auto theft incidents within the central part of the metropolitan area. The contour view suggest five or six peak areas which are very close to each other.

Much of this concentration, however, is produced by high population density in the metropolitan center. Figure 7.13, for example, shows the kernel estimate for 1349 census block groups for both the City of Baltimore and the County of Baltimore with the 1990 population assigned as the intensity variable. Again, the normal kernel function was used with an adaptive bandwidth of 100 points being selected. The map shows three views: 1) a surface view; 2) a contour view; and 3) a ground level view looking directly north. The distribution of population is, of course, also highly concentrated in the metropolitan center with two peaks, quite close to each other with several smaller peaks.

When these two kernel estimates are compared using the dual kernel density routine, a more complicated picture emerges (figure 7.14). This routine has conducted three operations: 1) it calculated the distance between each of the 10,800 reference cells and the 14,853 auto theft locations, evaluated the kernel function for each measured distance, and summed the results for each reference cell; 2) it calculated the distance between each of the 10,800 reference cells and the 1349 census block groups with population as an intensity variable, evaluated the kernel function for each intensity-weighted distance, and summed the results for each reference cell; and 3) divided the kernel density estimate for auto thefts by the kernel density estimate for population for each reference cell location.

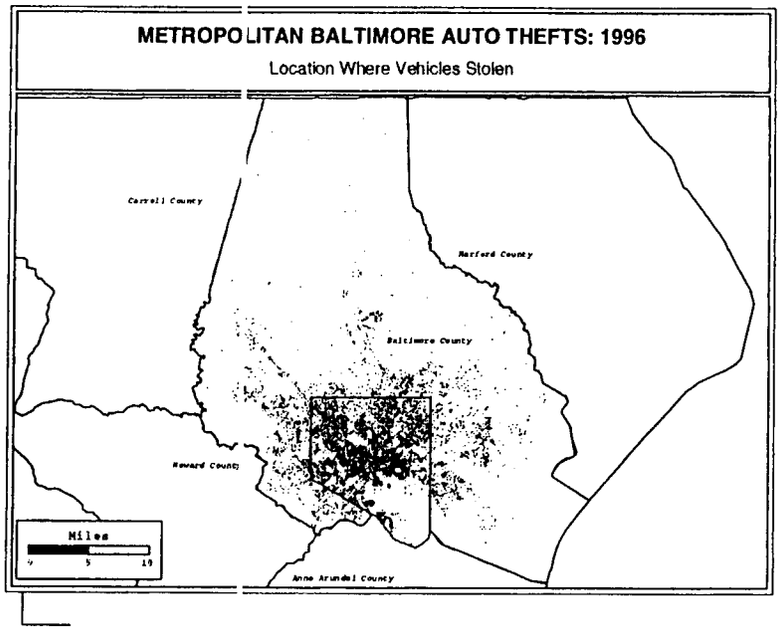
While the concentration of motor vehicle thefts relative to population ('motor vehicle theft risk") is still high in the metropolitan center, there are bands of high risk that spread outward, particularly along major arterials. There are now many 'hot spot' areas which have a high distribution of motor vehicle thefts relative to the residential population. We could, of course, refine this analysis further by taking, for example, employment as a baseline variable rather than population; employment is a better indicator for the daytime population distribution whereas the residential population is a better indicator for nighttime population distribution (Levine, Kim, and Nitz, 1995a; 1995b).

The advantage of a dual kernel density interpolation routine is that two variables can be related together. By interpolating one variable to a reference grid and then

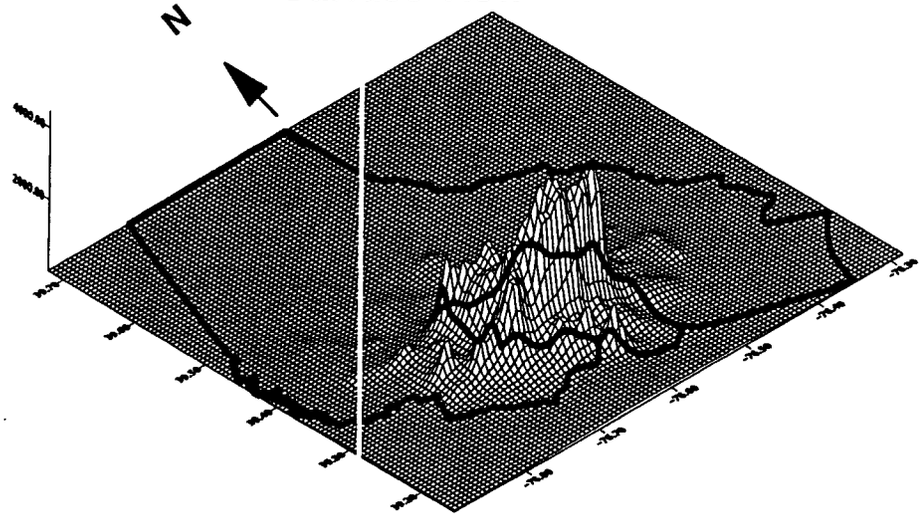
Figure 7.12:

Baltimore County Auto Thefts: 1996

Kernel Density Interpolation



Surface View



Contour View

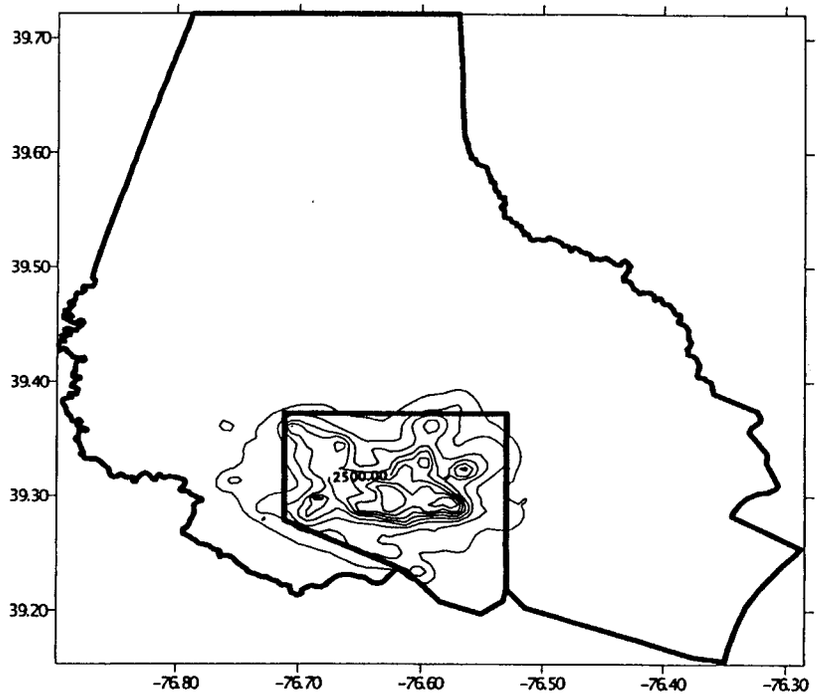
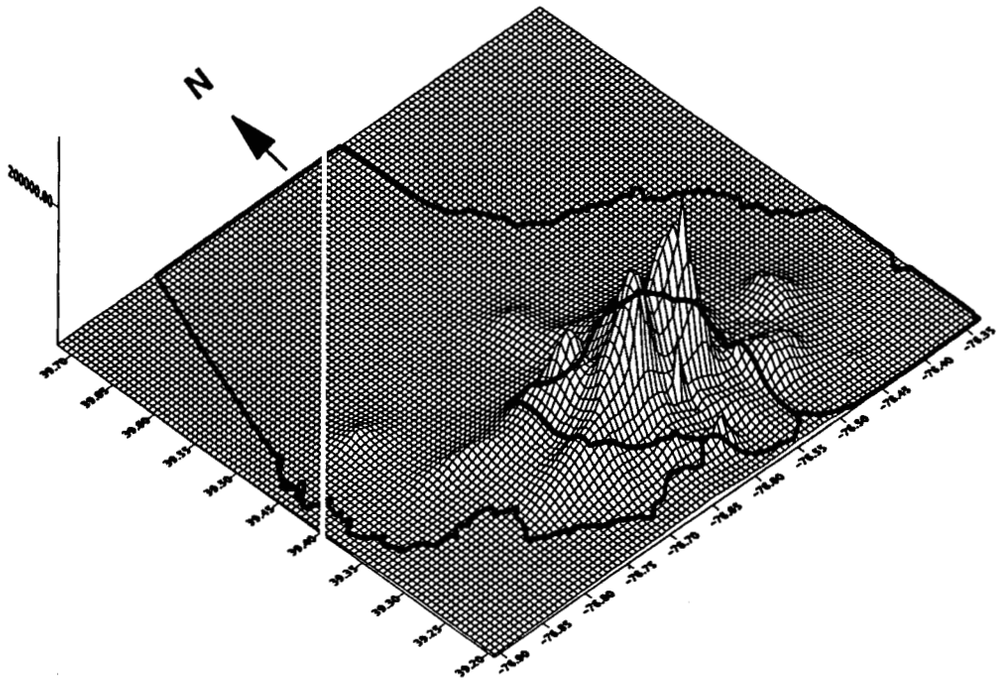


Figure 7.13:

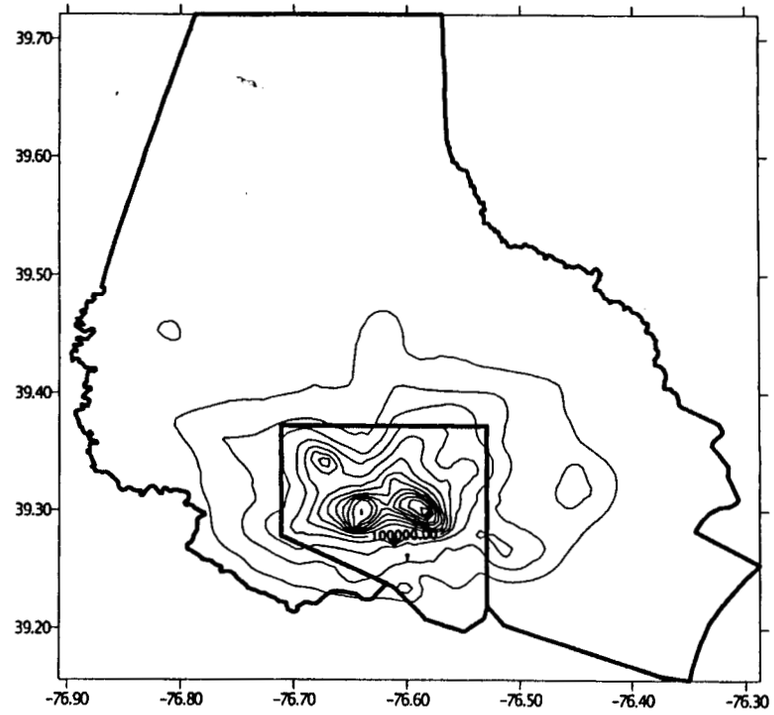
Baltimore Metropolitan Population: 1990

Block Group Population Interpolated by Adaptive Bandwidth of 100 Points

Surface View



Contour View



Ground Level View

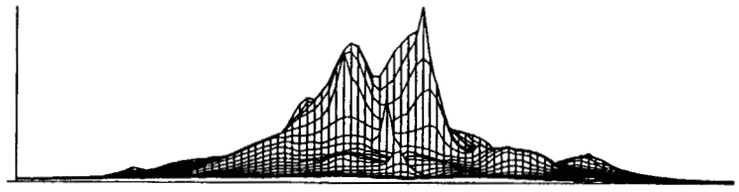
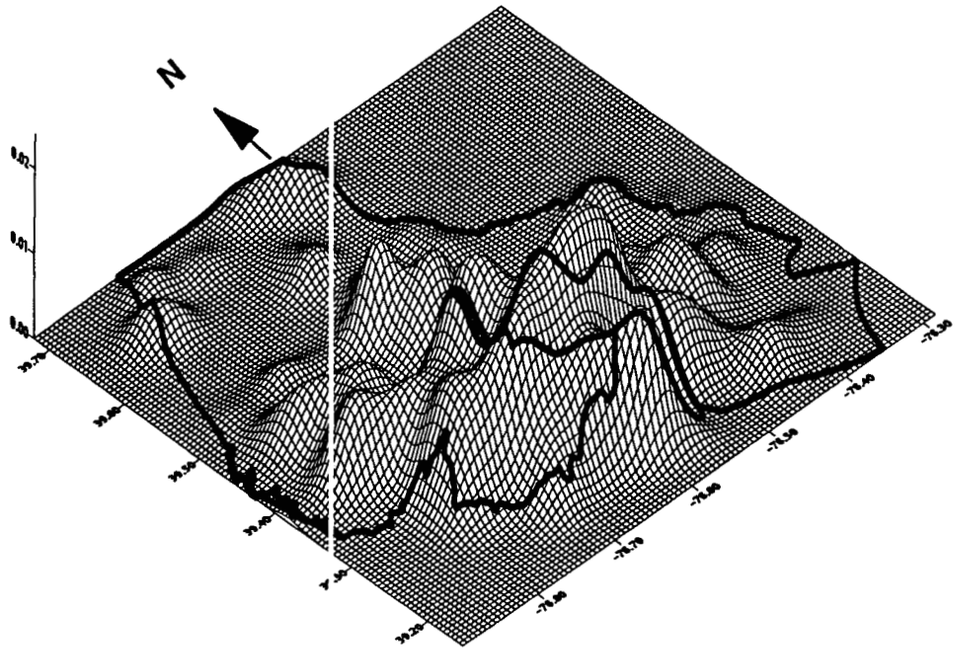


Figure 7.14:

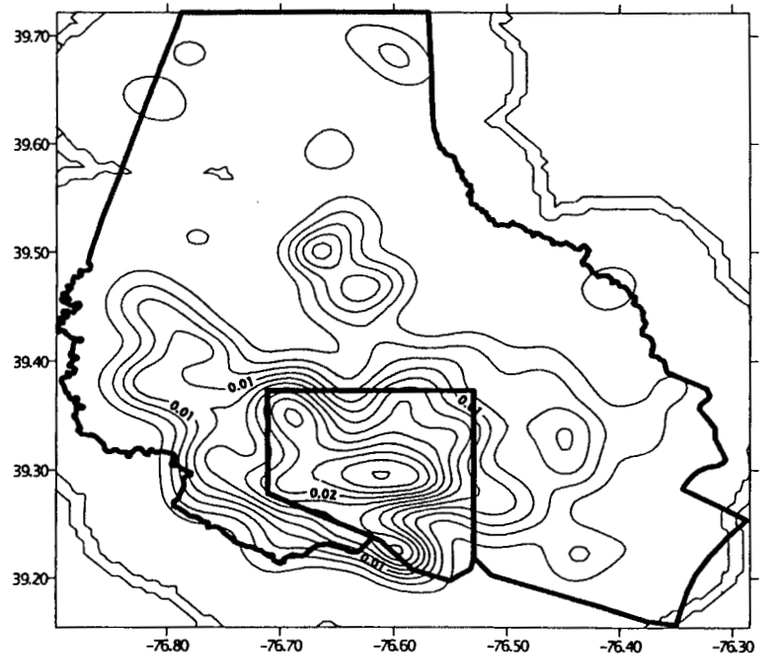
Baltimore County Auto Theft Risk

Ratio of Interpolation of 1996 Auto Thefts to 1990 Population

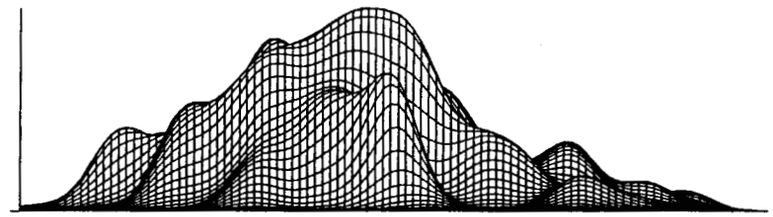
Surface View



Contour View



Ground Level View



interpolating a second variable to the same reference grid, the two variables have been interpolated to the same geographical units. The two interpolations can then be related, by dividing, subtracting, or summing. As has been mentioned throughout this manual, one of the problems with techniques that depend on the concentration of incidents is that they ignore the underlying population-at-risk. With the dual routine, however, we can start to examine the risk and not just the concentration. The comparison does not have to be between an incident distribution and a population-at-risk. It can be between two different types of crime incidents, for example, figure 7.15 shows the ratio of 1996 motor vehicle thefts to 1996 street robberies. 'Peaks', in this example, represent locations where there is a predominance of auto thefts relative to robberies whereas 'troughs' represent locations that have a predominance of robberies relative to auto thefts. Comparing two different crime distributions can allow an identification of areas where there is a concentration of one type of crime but not others.

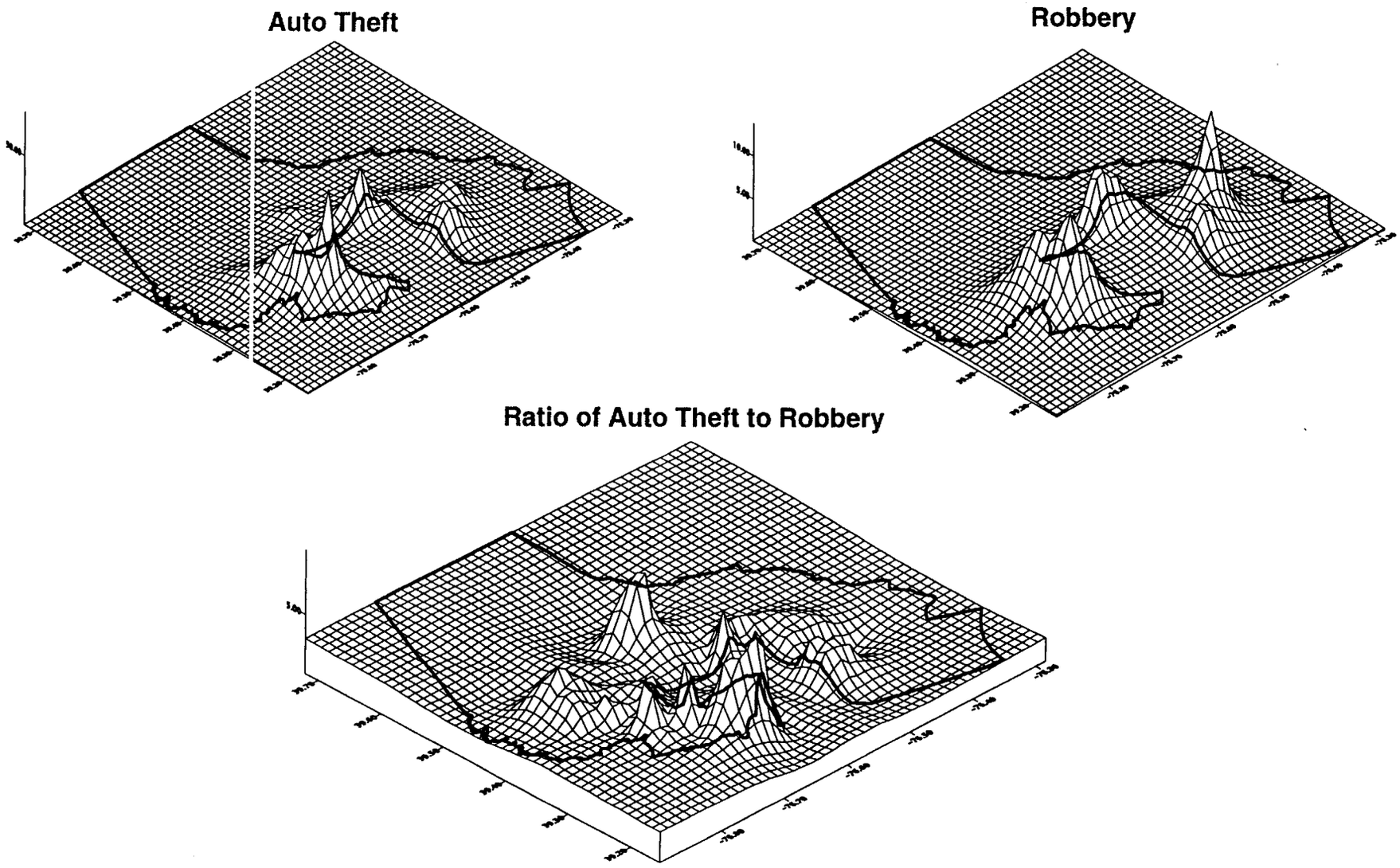
Conclusion

Kernel density estimation is one of the 'cutting edge' spatial statistical techniques. There is currently research on the use of this technique in both the statistical theory and in developing applications. For crime analysis, the technique represents a powerful way of conducting both 'hot spot' analysis as well as being able to link the 'hot spots' to an underlying population-at-risk. It can be used both for police deployment by targeting areas of high concentration of incidents as well as for prevention by targeting areas with high risk. It can also be used as a research tool for analyzing two or more distributions. As was mentioned, many of the statistical properties are still being developed by statisticians, particularly significance testing. But, over the next few years, these will become widely used tools for crime analysis and crime research.

Figure 7.15:

Comparison of Two Incident Interpolations

Ratio of Auto Theft to Street Robbery: 1996



Endnotes to Chapter 7

1. *CrimeStat* will output the geographical boundaries of the reference grid (a polygon grid) and will assign a third-variable (called *Z*) as the density estimate. Of the three polygon grid outputs, *ArcView* '.shp' files can be read directly into the program. For *MapInfo*, on the other hand, the output is in MapInfo Interchange Format (a '.mif' and a '.mid' file); the density estimate (also called *Z*) is assigned to the '.mid' file. The files must be imported to convert it to a *MapInfo* '.tab' file. For *Atlas*GIS* '.bna' format, however, there are two files that are output - a '.bna' file which includes the boundaries of the polygon grid and a '.dbf' file which includes the grid cell names (called *gridcell*) and the density estimate (also called *Z*). The '.bna' file must be read in first and then the '.dbf' file must be read in and matched to the value of *gridcell*. For all three output formats, the values of Z can be shown as a thematic map but the ranges must be adjusted to illustrate the likely locations for the offender's residence (i.e., the default values in the GIS programs will not display the densities very well). On the other hand, the default interval values for *Surfer for Windows* and *ArcView Spatial Analyst* provide a reasonably good visualization of the densities.

2. All the *CrimeStat* outputs except for *ArcView* 'shp' files are in ASCII. Edge effects can be removed with an ASCII editor by substituting '0' for the values at the edges, which are usually a very large number (i.e., spikes at the edges). Further, for 'shp' files, the values at the edges can be edited within the *ArcView* program. Care must be taken, however, to not edit an output file too much otherwise it will bear little relationship to the calculated kernel estimate.

References Used in *CrimeStat* Manual

Anselin, Luc (1995). "Local indicators of spatial association - LISA". *Geographical Analysis.* 27, No. 2 (April), 93-115.

Anselin, Luc. (1992). *SpaceStat: A Program for the Statistical Analysis of Spatial Data.* Santa Barbara, CA: National Center for Geographic Information and Analysis, University of California.

Aplin, Graeme (1983). *Order-Neighbour Analysis.* Concepts and Techniques in Modern Geography No. 36. Institute of British Geographers, Norwich, England: Geo Books.

Bachi, R. (1957). *Statistical Analysis of Geographical Series.* Central Bureau of Statistics, Kaplan School, Hebrew University: Jerusalem.

Bailey, Trevor C. and Anthony C. Gatrell (1995). *Interactive Spatial Data Analysis.* Longman Scientific & Technical: Burnt Mill, Essex, England.

Ball, G. H. and D. J. Hall (1970). "A clustering technique for summarizing multivariate data". *Behavioral Science,* 12, 153-155.

Beale, E. M. L. (1969). *Cluster Analysis.* Scientific Control Systems: London.

Block, Carolyn R. (1994). "STAC hot spot areas: a statistical tool for law enforcement decisions". In *Proceedings of the Workshop on Crime Analysis Through Computer Mapping.* Criminal Justice Information Authority: Chicago, IL.

Block, Richard and Carolyn R. Block (1995). "Space, place and crime: hot spot areas and hot places of liquor-related crime". In John E. Eck and David Weisburd, *Crime Places in Crime Theory,* Rutgers Crime Prevention Studies Series, Criminal Justice Press, Newark, NJ.

Block, Carolyn R. and Lynn A. Green (1994). *The GeoArchive Handbook: A Guide for Developing a Geographic Database an Information Foundation for Community Policing.* Illinois Criminal Justice Information Authority: Chicago, IL.

Borland.Com (1998). *dBase IV 2.0.* Inprise Corporation: Scotts Valley, CA.

Bowman, Adrian W. and Adelchi Azzalini (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations.* Oxford Science Publications, Oxford University Press: Oxford, England.

Burt, James E. and Gerald M. Barber (1996). *Elementary Statistics for Geographers* (second edition). The Guilford Press: New York.

Çan, Ayşe and Issac Megbolugbe (1996). "The geography of underserved mortgage markets". Paper presented at the American Real Estate and Urban Economics Association meeting. May.

Chrisman, Nicholas (1997) *Exploring Geographic Information Systems.* John Wiley and Sons, Inc.: New York.

Clark, P. J. and F. C. Evans (1954). "Distance to nearest neighbor as a measure of spatial relationships in populations". *Ecology*, 35, 445-453.

Cleveland, William S., Eric Grosse, and William M. Shyu (1993). "Local regression models". In John M. Chambers and Trevor J. Hastie, *Statistical Models in S.* Chapman & Hall: London.

Cliff, A. and J. Ord (1973). *Spatial Autocorrelation.* Pion: London.

Cole, A. J. and D. Wishart (1970). "An improved algorithm for the Jardine-Sibson method of generating overlapping clusters". *Comparative Journal*, 13, 156-163.

Committee on Map Projections (1986). *Which Map is Best*, American Congress on Surveying and Mapping, Falls Church, VA., 1986.

Cressie, Noel (1991). *Statistics for Spatial Data.* New York: J. Wiley & Sons, Inc.

Cromley, Robert G. (1992). *Digital Cartography.* Prentice Hall: Englewood Cliffs, NJ.

Ebdon, David (1988). *Statistics in Geography* (second edition with corrections). Blackwell: Oxford.

ESRI (1998a). *ArcView GIS 3.1.* Environmental Systems Research Institute: Redland, CA.

ESRI (1998b). *ArcInfo 7.2.1.* Environmental Systems Research Institute: Redland, CA.

ESRI (1998c). *Atlas*GIS 4.0.* Environmental Systems Research Institute: Redland, CA.

ESRI (1997). *ArcView Spatial Analyst.* Environmental Systems Research Institute: Redland, CA.

Everett, Brian (1974). *Cluster Analysis.* Heinemann Educational books, Ltd: London

Fotheringham, A. S. and M. E. O'Kelly (1989). *Spatial Interaction Models: Formulations and Applications.* Kluwer Academic Publishers: Boston.

Furfey, P. H. (1927). "A note on Lefever's 'Standard deviational ellipse'". *American Journal of Sociology.* XXIII, 94-98.

Gaile, Gary L. and James E. Burt (1980). *Directional Statistics.* Concepts and Techniques in Modern Geography No. 25. Institute of British Geographers, Norwich, England: Geo Books.

Geary, R. (1954). "The contiguity ratio and statistical mapping". *The Incorporated Statistician*, 5, 115-145.

Getis, Arthur (1991). "Spatial interaction and spatial auto-correlation: a cross-product approach". *Environment and Planning A*, 23, 1269-1277.

Getis, Arthur and J. Keith Ord (1996). "Local spatial statistics: an overview". In Paul Longley and Michael Batty (eds), *Spatial Analysis: Modelling in a GIS Environment.* GeoInformation International: Cambridge, England, 261-277.

Getis, Arthur and Barry Boots (1978). *Models of Spatial Processes: An Approach to the Study of Point, Line and Area Patterns.* London: Cambridge University Press.

Golden Software. 1994. *Surfer® for Windows (Version 6).* Golden Software, Inc.: Golden, CO.

Greenhood, David (1964). *Mapping.* The University of Chicago Press: Chicago.

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

Hammond, Robert, and Patrick McCullagh (1978). *Quantitative Techniques in Geography: An Introduction.* Second Edition. Clarendon Press: Oxford, England.

Härdle, Wolfgang (1991). *Smoothing Techniques with Implementation in S.* Springer-Verlag: New York.

Harries, Keith and Phil Canter (1998). "The use of GPS in geocoding crime incidents". Personal Communication.

Hartigan, J. A. (1975). *Clustering Algorithms.* John Wiley & Sons, Inc.: New York.

Hultquist, J., L. Brown and J. Holmes (1971). "Centro: a program for centrographic measures". Discussion paper no. 21, Department of Geography, Ohio State University: Columbus, OH.

Huxhold, William E. (1991). *An Introduction to Geographic Information Systems.* Oxford University Press: Oxford, New York, 147-184.

Jardine, N. and R. Sibson (1968). "The construction of hierarchic and non-hierarchic classifications". *Comparative Journal*, 11, 117-184.

Jones, K. S. and D. M. Jackson (1967). "Current approaches to classification and clump finding at the Cambridge Language Research Unit". *Comparative Journal,* 10, 29-37.

Kaluzny, Stephen P., Silvia C. Vega, Tamre P. Cardoso, and Alice A. Shelly (1998). *S+ Spatial Stats: User Manual for Windows and Unix.* Springer: New York.

Kanji, Gopal K. (1993). *100 Statistical Tests.* Sage Publications: Thousand Oaks, CA.

Kelsall, J. E. and P. J. Diggle (1995a). "Kernel estimation of relative risk", *Bernoulli,* 1, 3-16.

Kelsall, J. E. and P. J. Diggle (1995b). "Non-parametric estimation of spatial variation in relative risk". *Statistical Medicine,* 14, 2335-2342.

Kim, Karl E. and Michael Parke (1996). The use of GPS and GIS in traffic safety. Report to Motor Vehicle Safety Office, State of Hawaii Department of Transportation: Honolulu.

King, B. F. (1967). "Step wise clustering procedures". *Journal of the American Statistical Association.* 62, 86-101.

Kuhn, H.W. and R. E. Kuenne (1962). "An efficient algorithm for the numerical solution of the generalized Weber problem in spatial economics", *Journal of Regional Science* 4, 21-33.

Langworthy, Robert H. and Eric Jefferis (1998). "The utility of standard deviational ellipses for project evaluation". Discussion paper, National Institute of Justice: Washington, DC.

Lefever, D. (1926). "Measuring geographic concentration by means of the standard deviational ellipse". *American Journal of Sociology,* 32(1): 88-94.

Levine, Ned (1999). "The effects of local growth management on regional housing production and population redistribution in California", In press, *Urban Studies.*

Levine, Ned (1996). "Spatial statistics and GIS: software tools to quantify spatial patterns". *Journal of the American Planning Association.* 62 (3), 381-392.

Levine, Ned and Karl E. Kim (1999). "The spatial location of motor vehicle accidents: A methodology for geocoding intersections". *Computers, Environment, and Urban Systems.* 22 (6), 557-576.

Levine, Levine, Karl E. Kim, and Lawrence H. Nitz (1995a). "Spatial analysis of Honolulu motor vehicle crashes: I. Spatial patterns". *Accident Analysis & Prevention,* 27 (5), 663-674.

Levine, Ned, Karl E. Kim, and Lawrence H. Nitz (1995b). "Spatial analysis of Honolulu motor vehicle crashes: II. Generators of crashes". *Accident Analysis & Prevention*, 27 (5), 675-685.

Levine, Ned and Martin Wachs (1986a). "Bus Crime in Los Angeles: I - Measuring The Incidence". *Transportation Research.* 20 (4), 273-284.

Levine, Ned and Martin Wachs (1986b). "Bus Crime in Los Angeles: II - Victims and Public Impact". *Transportation Research.* 20 (4), 285-293.

Levine, Ned, Martin Wachs and Elham Shirazi (1986). "Crime at Bus Stops: A Study of Environmental Factors". *Journal of Architectural and Planning Research.* 3 (4), 339-361.

Los Angeles Times (1998). *Eye on the Sky.* Business section, July 20.

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations". *5th Berkeley Symposium on Mathematics, Statistics and Probability.* Vol 1, 281-298.

McDonnell, Porter W. Jr. (1979). *Introduction to Map Projections.* New York: Marcel Dekker, Inc.

McQuitty, L. L. (1960). "Hierarchical syndrome analysis". *Educational and Psychological Measurement*, 20, 293-304.

Maling, D. H. (1973). *Coordinate Systems and Map Projections* (1973). George Philip and Sons, London.

Maltz, Michael D., Andrew C. Gordon, and Warren Friedman (1989). *Mapping Crime in Its Community Setting: A Study of Event Geography Analysis.*

Mantel, N. (1967). "The detection of disease clustering and a generalized regression approach". *Cancer Research*, 27, 209-220.

MapInfo (1998). *MapInfo Professional 5.0.1.* MapInfo Corporation: Troy, NY.

Mardia, K.V. (1972). *Statistics of Directional Data.* Academic Press: New York.

Mathsoft, Inc. (1998). *S-PLUS 4.5 Professional for Windows.* MathSoft, Inc: Seattle.

Microsoft (1999). *Windows NT Server, Enterprise Edition.*Microsoft: Redmond, WA.

Microsoft (1998a). *Windows NT Workstation 4.0.* Microsoft: Redmond, WA.

Microsoft (1998b). *Windows NT Server 4.0.* Microsoft: Redmond, WA.

Microsoft (1998c). *Windows 98.* Microsoft: Redmond, WA.

Microsoft (1995). *Windows 95.* Microsoft: Redmond, WA.

Moran, P. A. P. The interpretation of statistical maps. *Journal of the Royal Statistical Society B,* 10, 1948; 243-251.

Moran, P. A. P. Notes on continuous stochastic phenomena. *Biometrika,* 37, 1950; 17-23.

Needham, R. M. (1967). "Automatic classification in linguistics". *The Statistician,* 17, 45-54.

Neft, David Samuel (1962). *Statistical Analysis for Areal Distributions.* Ph.D. dissertation, Columbia University: New York.

Parzen, E. (1962). "On the estimation of a probability density and mode". *Annuals of Mathematical Statistics,* 33, 1065-1076.

Ripley, Brian D (1981). *Spatial Statistics.* John Wiley & Sons: New York.

Ripley, Brian D. (1976). "The second-order analysis of stationary point processes". *Journal of Applied Probability* 13: 255-66.

Robinson, A. H., R. D. Sale, J. L. Morrison and P. C. Muehrcke (1984). *Elements of Cartography* (5th edition). J. Wiley and Sons: New York.

Rosenblatt, M. (1956). "Remarks on some non-parametric estimates of a density function". *Annuals of Mathematical Statistics,* 27, 832-837.

SPSS, Inc. (1999). *SPSS 9.0 for Windows.* SPSS, Inc.: Chicago.

SAS Institute Inc. (1998). *Statistical Analysis System, Version 7.* Cary, NC.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall: London.

Sneath, P. H. A. (1957). "The application of computers to taxonomy". *Journal of General Microbiology,* 17, 201-226.

Snyder, John P. (1987). *Map Projections - A Working Manual.* U.S. Geological Survey Professional Paper 1395. U. S. Government Printing Office: Washington, DC.

Snyder, John P. and Philip M. Voxland (1989). *An Album of Map Projections.* U.S. Geological Survey Professional Paper 1453. U. S. Government Printing Office: Washington, DC.

Sokal, R. R. and P. H. A. Sneath (1963). *Principles of Numerical Taxonomy.* W. H. Freeman and Co.: San Francisco.

Sokal, R. R. and C. D. Michener (1958). "A statistical method for evaluating systematic relationships". *University of Kansas Science Bulletin,* 38, 1409-1438.

Stephenson, L. (1980). "Centrographic analysis of crime". In D. George-Abeyie and K. Harries (eds), *Crime, A Spatial Perspective,* Columbia University Press: New York.

Systat, Inc. (1996). *Systat 6.0 for Windows.* SPSS, Inc.: Chicago.

Systat, Inc. (1994). Advanced Applications: Comprehensive Statistics and Graphics for DOS. Systat, Inc.: Evanston, IL.

Thompson, H. R. (1956). "Distribution of distance to nth neighbour in a population of randomly distributed individuals". *Ecology,* 37, 391-394.

Thorndike, R. L. (1953). "Who belongs in a family?". *Psychometrika,* 18, 267-276.

U.S. Census Bureau (1998). *TIGER/Line 1997.* Bureau of the Census, U. S. Department of Commerce: Washington, DC.

Venables, W.N. and B.D. Ripley (1997). *Modern Applied Statistics with S-Plus (second edition).* Springer-Verlag: New York.

Ward, J. H. (1963). "Hierarchical grouping to optimize an objective function". *Journal of the American Statistical Association.* 58, 236-244.

Whittle, P. (1958). "On the smoothing of probability density functions". *Journal of the Royal Statistical Society, Series B,* 55, 549-557.

Appendix A

Dynamic Data Exchange (DDE) Support

CrimeStat supports Dynamic Data Exchange (DDE). This allows the program to be linked to another program, which can call up *CrimeStat* as a routine. The following are the programming codes used to support DDE commands.

CrimeStat's DDE Topics That Support the DDE "poke" Command

Topic	Item	Data format
Primary File	File	RemoveAll
		AddTextFile | <name> | <columns> | <separator> | <header rows>
		AddDbfFile | <name>
		AddShpFile | <name>
	X	<file name> | <column name>
	Y	<file name> | <column name>
	Weight	<file name> | <column name>
	Intensity	<file name> | <column name>
	Direction	<file name> | <column name>
	Coordinate	<coordinate> | <unit> Valid coordinates: Longtitude, latitude Projected Direction Valid units: Decimal degrees Feet Meters
Secondary File	File	See Primary File
	X	See Primary File
	Y	See Primary File
	Weight	See Primary File
	Intensity	See Primary File
	Direction	See Primary File
Reference File	Source	<source> Valid sources: From File Generated
	File	See Primary File

	X	See Primary File
	Y	See Primary File
	True Grid	<number of columns> a value of zero (0) will uncheck the check box.
	Bound	<lower-left x>|<lower-left y>|<upper-right x>|<upper-right y>
	Cell specification	<source>|<value> Valid sources: By cell-spacing By number of columns
Measurement Parameters	Measurement Type	<type> Valid types: Direct Indirect
	Area	<area>|<area unit> Valid units: See Dialog
	Length	<length>|<length unit> Valid units: See Dialog

***CrimeStat*'s DDE Topics That Support the DDE "request" Command**

Topic	Item	Return value
System	SysItems	Name of the supported items of the "system" topic.
	ReturnMessage	Detailed information (if any) of the last message.
	Status	Server status, either "Ready" or "Busy".
	Formats	Supported data formats.
	Help	Detailed help on CrimeStat's DDE support.
	TopicItemList	Name of the supported items of the current topic.

***CrimeStat*'s DDE Topics That Support the DDE "execute" Command**

Topic	Command	Description
System	Quit	Close CrimeStat.
Primary File	Select	Select the Primary file tab.
Secondary File	Select	Select the Secondary file tab.
Reference File	Select	Select the Reference file tab.

Measurement Parameters	Select	Select the Measurement parameters tab.
Spatial Distribution	Select	Select the Spatial distribution tab.
Distance Analysis	Select	Select the Distance analysis tab.
'Hot spot' Analysis	Select	Select the 'Hot spot' analysis tab.
Interpolation	Select	Select the Interpolation tab.

Example: Controlling *CrimeStat* from within Visual basic

```
Public Function OpenCrimeStat(topic As String) As Variant
   On Error Resume Next
   Dim channel, I
   Dim file As String
   file = "CrimeStat.exe"
   channel = DDEInitiate("CrimeStat", topic)
   If Err Then
      Err = 0
      I = Shell(file, 1)
      If Err Then
         Return
      End If
      channel = DDEInitiate("CrimeStat", topic)
   End If
   OpenCrimeStat = channel
End Function

Public Sub TestCrimeStatDde(foo As String)
   On Error Resume Next
   Dim file As String
   Dim channel

   file = "SampleData.dbf"
   channel = OpenCrimeStat("Primary File")
   DDEPoke channel, "Coordinate", "Projected | Feet"
   DDEPoke channel, "File", "RemoveAll"
   DDEPoke channel, "File", "AddDbfFile | " & file
   DDEPoke channel, "X", file & " | LON"
   DDEPoke channel, "Y", file & " | LAT"
   DDEPoke channel, "Coordinate", "Longtitude, latitude | Decimal degrees"
   DDETerminate channel

   file = "Grid.dbf"
   channel = OpenCrimeStat("Reference File")
   DDEPoke channel, "Source", "From File"
```

```
    DDEPoke channel, "True Grid", "0"
    DDEPoke channel, "File", "RemoveAll"
    DDEPoke channel, "File", "AddDbfFile|" & file
    DDEPoke channel, "X", file & "|LON"
    DDEPoke channel, "Y", file & "|LAT"
    DDEPoke channel, "True Grid", "108"
    DDEPoke channel, "Source", "Generated"
    DDEPoke channel, "Bound", "-78.5|22.4|-75.3|24.2"
    DDEPoke channel, "Cell Specification", "By cell-spacing|0.5"
    DDEPoke channel, "Cell Specification", "By number of columns|20"
    DDETerminate channel

    channel = OpenCrimeStat("Measurement Parameters")
    DDEPoke channel, "Measurement Type", "Direct"
    DDEPoke channel, "Measurement Type", "Indirect"
    DDEPoke channel, "Area", "734.12|Square meters"
    DDEPoke channel, "Length", "1734.12|meters"
    DDETerminate channel

    channel = OpenCrimeStat("Interpolation")
    DDEExecute channel, "select"
    DDETerminate channel
End Sub

Private Sub CrimeStatQuit_Click()
    On Error Resume Next
    Dim channel
    channel = OpenCrimeStat("System")
    DDEExecute channel, "quit"
    DDETerminate channel
End Sub
Private Sub TestCrimeStat_Click()
    TestCrimeStatDde "bar"
End Sub
```

August, 1999

Quickguide to *CrimeStat*

by
Ned Levine, PhD
Ned Levine & Associates
Annandale, VA

The following are quick instructions for the use of *CrimeStat*, paralleling the online help menus in the program. Detailed instructions should be obtained from chapters 3-7 in the documentation. *CrimeStat* has eight program tabs. Each tab lists routines, options and parameters. The eight tabs are:

1. Primary file
2. Secondary file
3. Reference file
4. Measurement parameters
5. Spatial distribution
6. Distance analysis
7. 'Hot Spot' analysis
8. Interpolation

Figure 1-8 show the eight tab screens with examples of data input and routine selection.

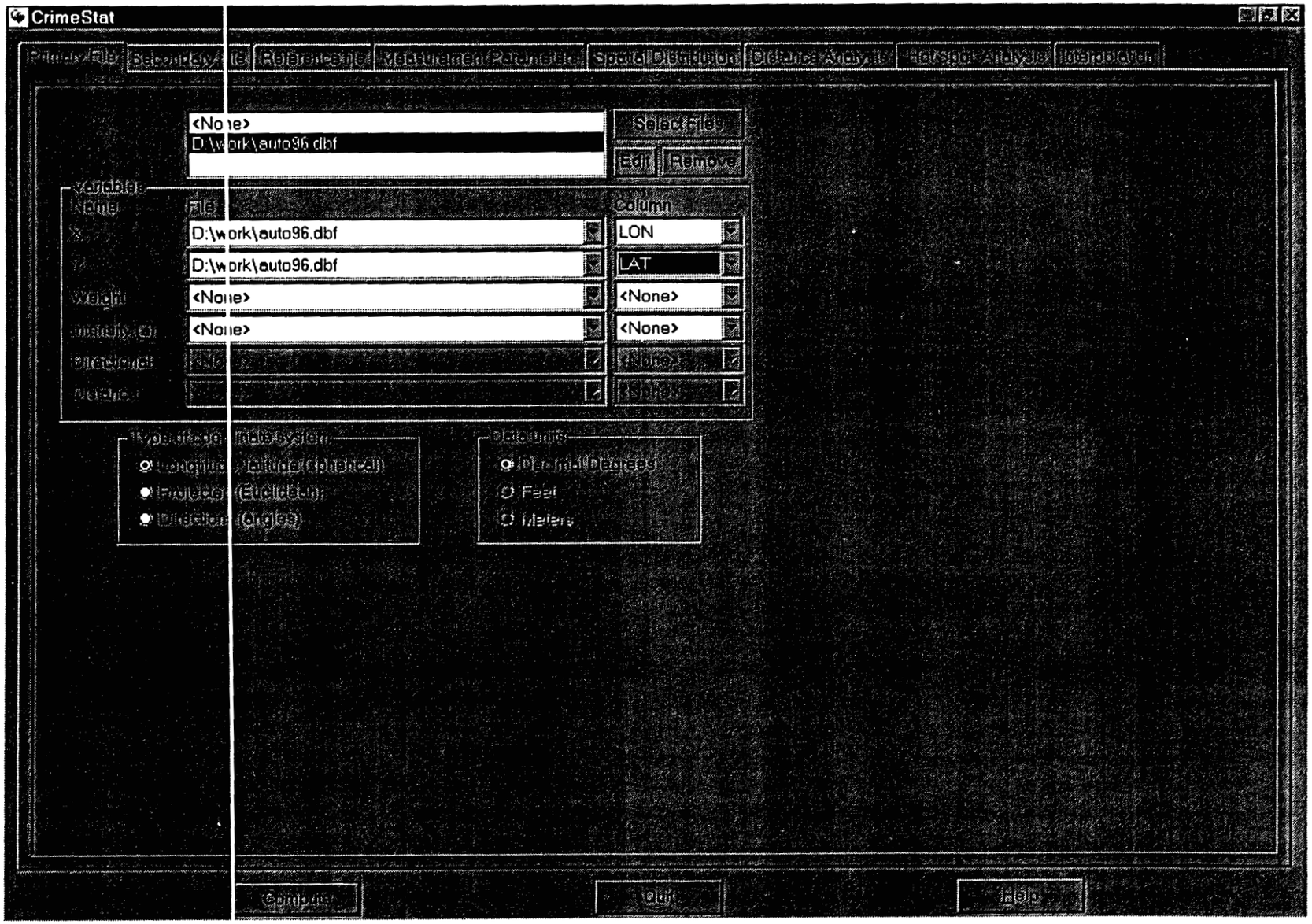
Primary File

A primary file is required for *CrimeStat*. It is a point file with X and Y coordinates. For example, the primary file could be the location of street robberies, each of which has an associated X and Y coordinate. Alternatively, the primary file could be the location of police stations, again defined by an X and Y coordinate. Also, there can be weights or intensities variables associated, although these are optional. For example, if the points are the locations of police stations, then the intensity variable could be the number of calls for service at each police station while the weighting variable could be service zones. More than one file can be selected.

Select Files

Select the primary file. *CrimeStat* can read ASCII, dBase®III/IV '.dbf', and ArcView® '.shp' files. Select the tab and indicate the type of file to be selected. Use the browse button to search for the file. If the file type is ASCII, select the type of data separator (comma, semicolon, space, tab) and the number of columns.

Figure 1: Primary File Screen



Variables

Define the file which contains the X and Y coordinates. If there are weights or intensities being used, define the file which contain these variables. Certain statistics (e.g., spatial autocorrelation, local Moran) require intensity values and most other statistics can use intensity values. Other statistics (e.g., a weighted mean center) can use weights. It is possible to have a variable represent both intensity and a weighting; it is also possible to have separate variables for intensity and for weighting.

Column

Select the variables for the X and Y coordinates respectively (e.g., Lon, Lat, Xcoord, Ycoord). If weights or intensities are being used, select the appropriate variable names.

Directional

If the file contains directional coordinates (angles), define the file name and variable name (column) that contains the directional measurements. If directional coordinates are used, there can be an optional distance variable for the measurement. Define the file name and variable name (column) that contains the distance variable.

Type of Coordinate System and Data Units

Select the type of coordinate system. If the coordinates are in longitudes and latitudes, then a spherical system is being used and data units will automatically be decimal degrees. If the coordinate system is projected (e.g., State Plane, Universal Transverse Mercator – UTM), then data units could be either in feet (e.g., State Plane) or meters (e.g., UTM). If the coordinate system is directional, then the coordinates are angles and the data units box will be blanked out; if a distance variable is used with the directional coordinates, the data units are undefined.

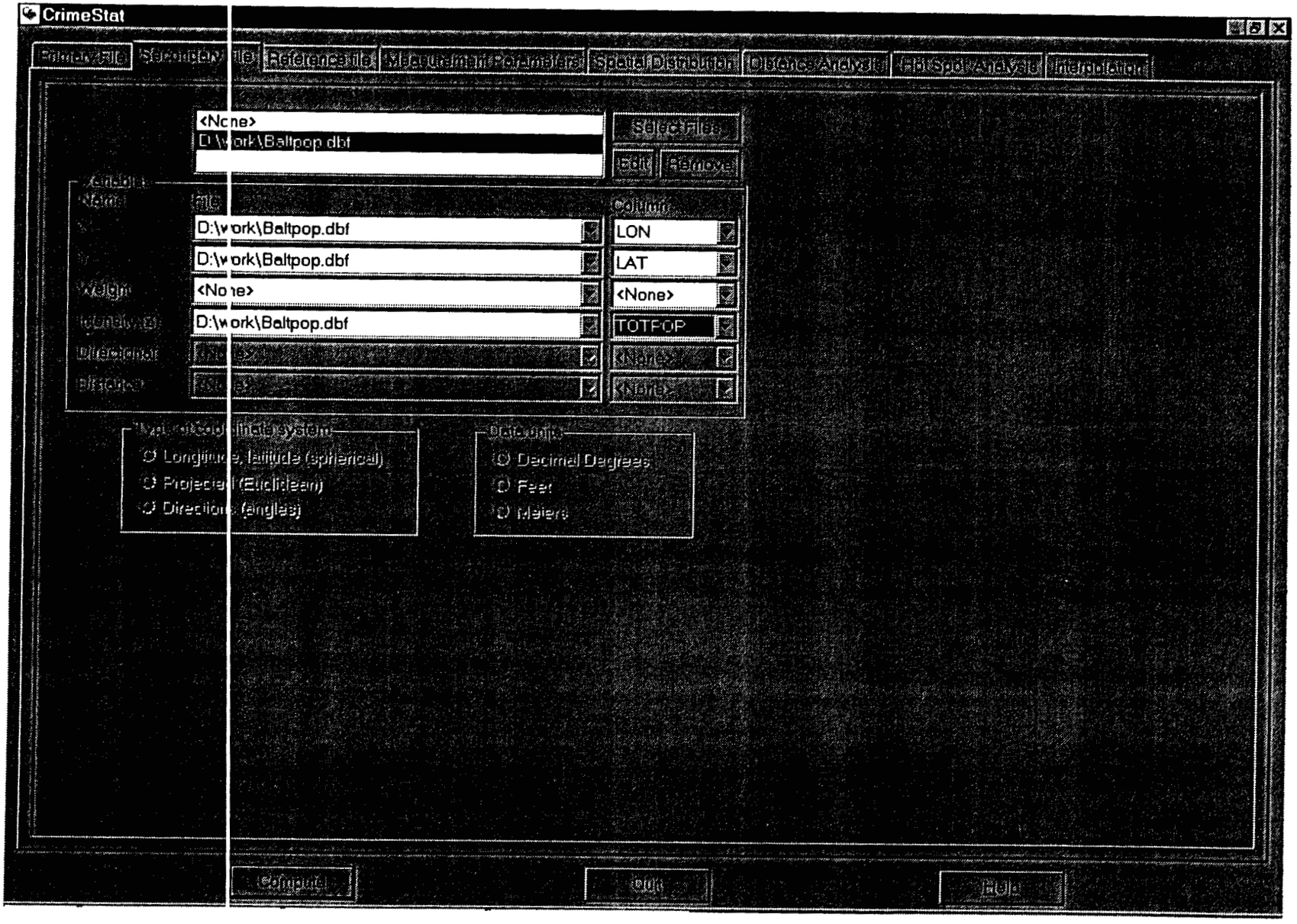
Secondary File

A secondary data file is optional. It is also a point file with X and Y coordinate and is usually used in comparison with the primary file. There can be weights or intensities variables associated, though these are optional. For example, if the primary file is the location of motor vehicle thefts, the secondary file could be the centroid of census block groups that have the population of the block group as the intensity (or weight) variable. In this case, one could compare the distribution of motor vehicle thefts with the distribution of population in, for example, the Ripley's "K" routine or the dual kernel density estimation routine.

Select Files

Select the secondary file. *CrimeStat* can read ASCII, dBase III/IV '.dbf', and ArcView '.shp' files. Select the tab and indicate the type of file to be selected. Use the

Figure 2: Secondary File Screen



browse button to search for the file. If the file type is ASCII, select the type of data separator (comma, semicolon, space, tab) and the number of columns.

Variables

Define the file which contains the X and Y coordinates. If weights or intensities are being used, define the file which contain these variables. Certain statistics (e.g., spatial autocorrelation, local Moran) require intensity values and most other statistics can use intensity values. Other statistics (e.g., a weighted mean center) can use weights. It is possible to have a variable represent both an intensity and a weighting; it is also possible to have separate variables for intensity and for weighting.

Column

Select the variables for the X and Y coordinates respectively (e.g., Lon, Lat, Xcoord, Ycoord). If there are weights or intensities being used, select the appropriate variable names.

Type of Coordinate System and Data Units

The secondary file has the same coordinate system and data units as the primary file. This selection will be blanked out, indicating that the secondary file carries the same definition as the primary file.

Reference File

A reference file is used for single and dual variable kernel density estimation. The file can be an external file that is input or can be generated by *CrimeStat*. It is usually, though not always, a grid which is overlaid on the study area.

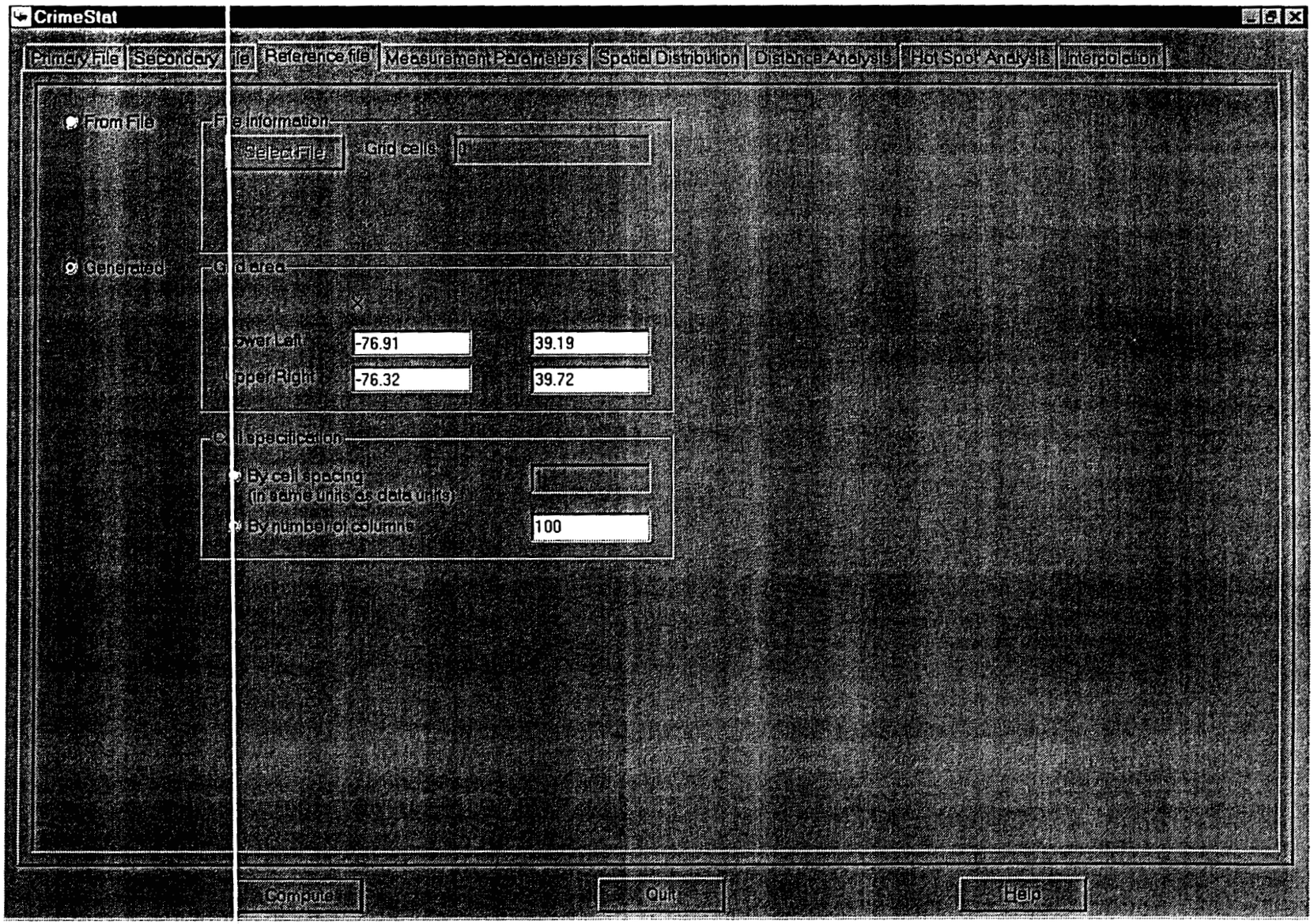
From File

Select the reference file. *CrimeStat* can read ASCII, dBase III/IV '.dbf', and ArcView '.shp' files. Select the tab and indicate the type of file to be selected. Use the browse button to search for the file. If the file type is ASCII, select the type of data separator (comma, semicolon, space, tab) and the number of columns. A file that is read into *CrimeStat* need not be a true grid (a matrix with k columns and l rows). However, if a file is read in, then the results can only be output to *Surfer® for Windows*. The Grid Cells indicator will estimate the number of cells in the file as if it was a grid; this is an estimate since the imported file may not be a true grid.

Generated

CrimeStat can generate a true grid based on the inputting of the X and Y coordinates of a rectangle placed over the study area. The lower left and the upper right coordinates must be defined in the same coordinate system and data units as the primary

Figure 3: Reference File Screen



file. Cells can be defined either by cell size in the same coordinate system and data units as the primary file or by the number of columns in the grid. If the latter is selected, *CrimeStat* will determine the number of rows to be generated based on the cell spacing.

Measurement Parameters

The measurement parameters define the measurement units of the coverage and the type of distance measurement to be used.

Area

Define the geographical area of the study area in area units (square miles, square nautical miles, square feet, square kilometers, square meters). Irrespective of the data units that are defined for the primary file, *CrimeStat* can convert to various area measurement units. These units are used in the nearest neighbor, Ripley's "K", nearest neighbor hierarchical clustering, and K-means clustering routines. If no area units are defined, then *CrimeStat* will define a rectangle by the minimum and maximum X and Y coordinates.

Length of Street Network

Define the total length of the street network within the study area or an appropriate comparison network (e.g., freeway system) in distance units (miles, nautical miles, feet, kilometers, meters). The length of the street network is used in the linear nearest neighbor routine. Irrespective of the data units that are defined for the primary file, *CrimeStat* can convert to distance measurement units. The distance units should be in the same metric as the area units (e.g., miles and square miles/meters and square meters).

Type of Distance Measurement

Select the type of distance measurement to be used, direct or indirect.

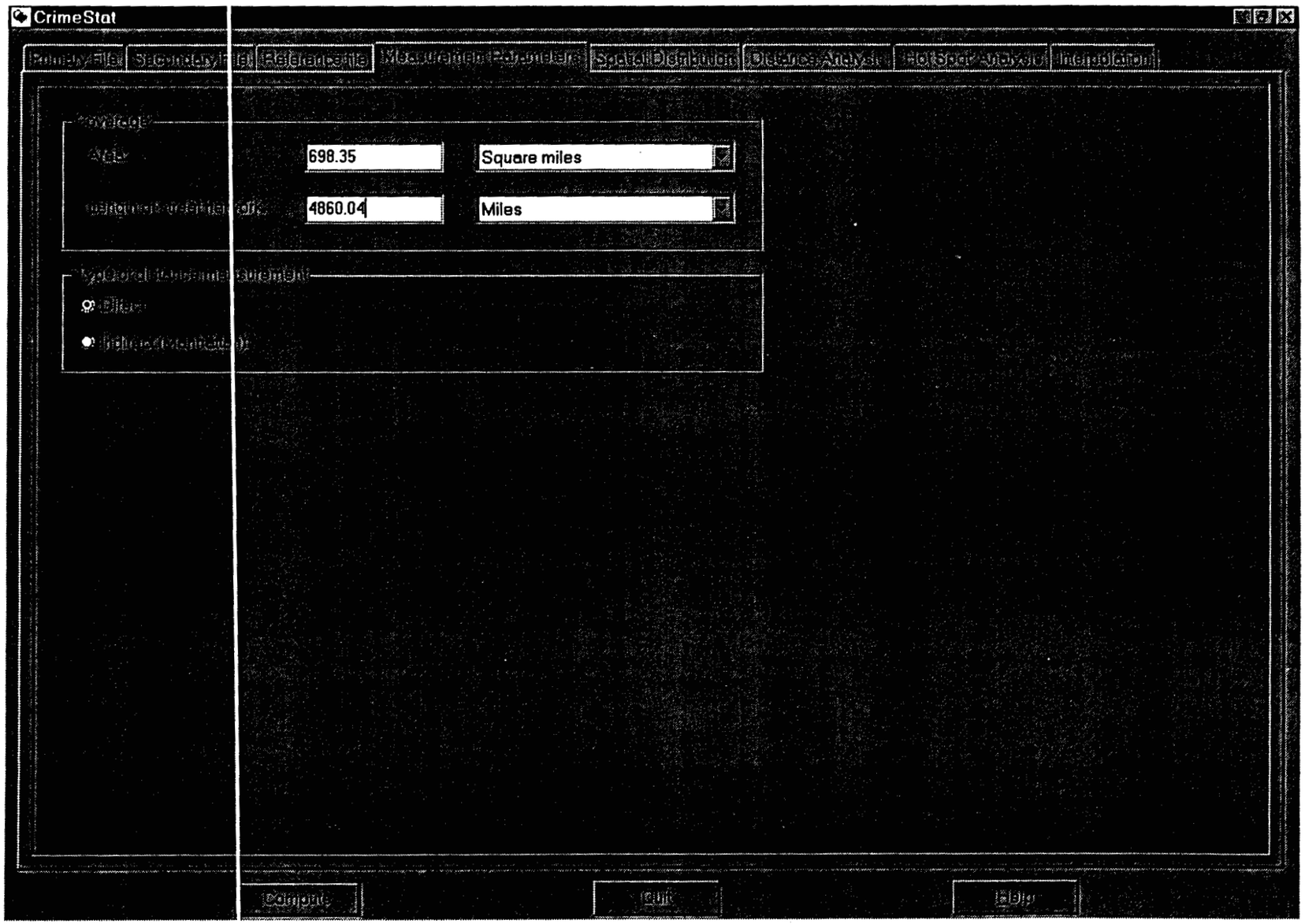
Direct

If direct distances are used, each distance is calculated as the shortest distance between two points. If the coordinates are spherical (i.e., latitude, longitude), then the shortest direct distance is a 'Great Circle' arc on a sphere. If the coordinates are projected, then the shortest direct distance is a straight line on a Euclidean plane.

Indirect

If indirect distances are used, each distance is calculated as the shortest distance between two points on a grid, that is with distance being constrained to the horizontal or vertical directions (i.e., not diagonal). This is sometimes called 'Manhattan' metric. If the coordinates are spherical (i.e., latitude, longitude), then the shortest indirect distance is a modified right angle on a spherical right triangle; see the documentation for more details.

Figure 4: Measurement Parameters Screen



If the coordinates are projected, then the shortest indirect distance is the right angle of a right triangle on a two-dimensional plane.

Spatial Distribution

Spatial distribution provides statistics that describe the overall spatial distribution. These are sometimes called centrographic, global, or first-order spatial statistics. There are three routines for describing the spatial distribution and two routines for describing spatial autocorrelation. An intensity variable and a weighting variable can be used for the first three routines. An intensity variable is required for the two spatial autocorrelation routines; a weighting variable can also be used for the spatial autocorrelation indices. All outputs can be saved as text files. Some outputs can be saved as graphical objects for import into desktop GIS programs.

Mean Center and Standard Distance (Mcsd)

The mean center and standard distance define the arithmetic mean location and the degree of dispersion of the distribution. The Mcsd routine calculates 11 statistics:

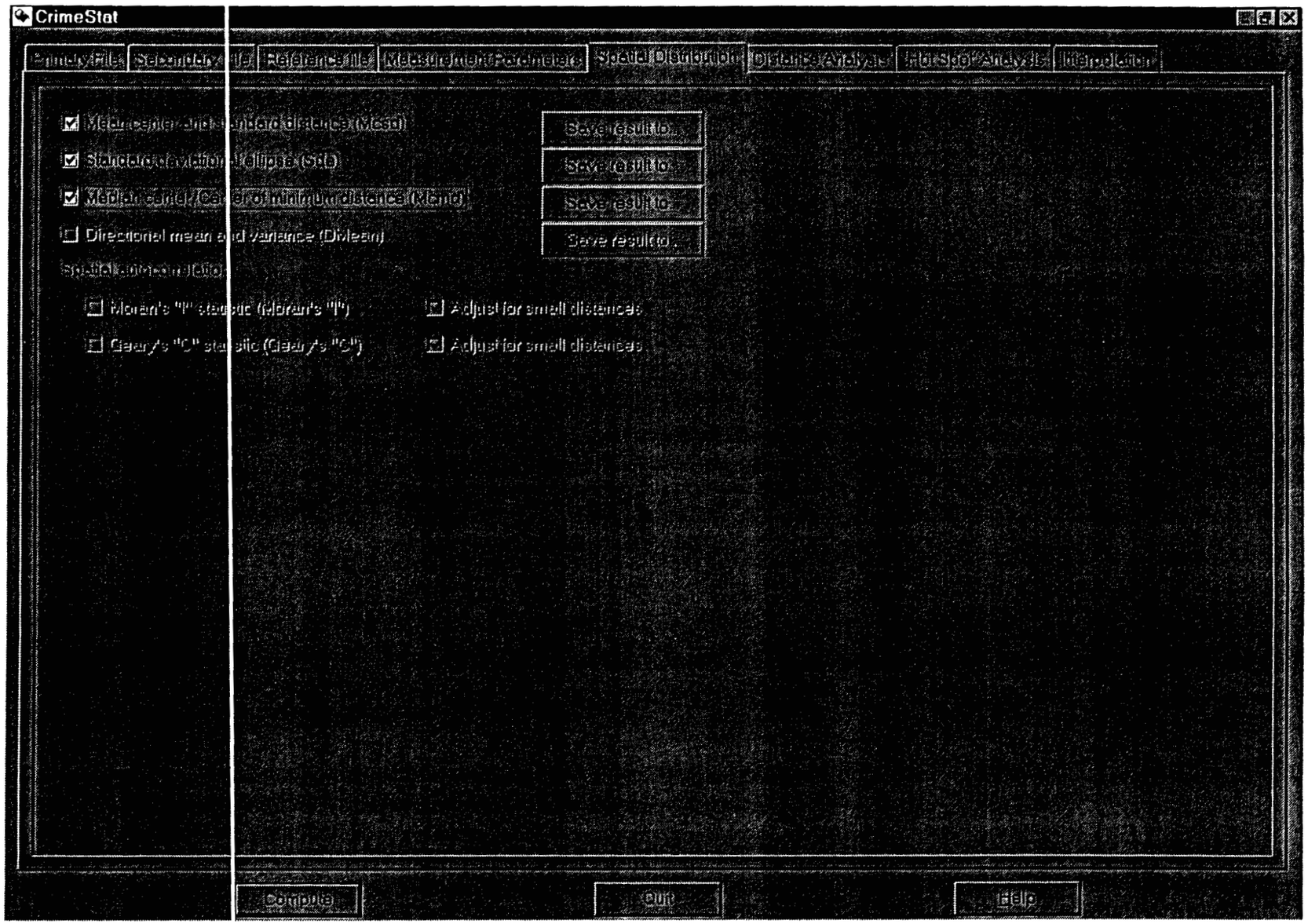
1. The sample size
2. The minimum X value
3. The minimum Y value
4. The maximum X value
5. The maximum Y value
6. The mean of the X coordinates
7. The mean of the Y coordinates
8. The standard deviation of the X coordinates
9. The standard deviation of the Y coordinates
10. The standard distance deviation, in meters, feet and miles. This is the standard deviation of the distance of each point from the mean center.
11. The circle area defined by the standard distance deviation, in square meters, square feet and square miles.

The tabular output can be printed and the mean center (mean X, mean Y), the standard deviations of the X and Y coordinates, and the standard distance deviation can be output as graphical objects to ArcView '.shp', MapInfo® '.mif' and Atlas*GIS™ '.bna' formats. A root name should be provided. The mean center is output as a point (MC<root name>). The standard deviations of the X and Y coordinates are output as a rectangle (XYD<root name>). The standard distance deviation is output as a circle (SDD<root name>).

Standard Deviational Ellipse (Sde)

The standard deviational ellipse defines both the dispersion and the direction (orientation) of that dispersion. The Sde routine calculates 13 statistics:

Figure 5: Spatial Distribution Screen



1. The sample size
2. The clockwise angle of Y-axis rotation in degrees
3. The ratio of the long to the short axis after rotation
4. The standard deviation along the new Y axis in meters, feet and miles
5. The standard deviation along the new X axis in meters, feet and miles
6. The Y axis length in meters, feet and miles
7. The X axis length in meters, feet and miles
8. The area of the ellipse defined by these axes in square meters, square feet and square miles
9. The standard deviation along the Y axis in meters, feet and miles for a 2X standard deviational ellipse
10. The standard deviation along the X axis in meters, feet and miles for a 2X standard deviational ellipse
11. The Y axis length in meters, feet and miles for a 2X standard deviational ellipse
12. The X axis length in meters, feet and miles for a 2X standard deviational ellipse
13. The area of the 2X ellipse defined by these axes in square meters, square feet and square miles.

The tabular output can be printed and the 1X and 2X standard deviational ellipses can be output as graphical objects to ArcView '.shp', MapInfo '.mif' and Atlas*GIS '.bna' formats. A root name should be provided. The 1X standard deviational ellipse is output as an ellipse (SDE<root name>). The 2X standard deviational ellipse is output as an ellipse with axes that are twice as large as the 1X standard deviational ellipse (2SDE<root name>).

Median Center/Center of Minimum Distance (Mcmd)

The center of minimum distance (or median center) defines the point at which the distance to all other points is at a minimum. The Mcmd routine outputs 7 statistics:

1. The sample size
2. The mean of the Y coordinates
3. The mean of the X coordinates
4. The number of iterations required to identify a median center
5. The degree of error (tolerance) for stopping the iterations
6. The Y coordinate which defines the center of minimum distance (median center)
7. The X coordinate which defines the center of minimum distance (median center).

The tabular output can be printed and the median center can be output as a graphical object to ArcView '.shp', MapInfo '.mif or Atlas*GIS '.bna' files. A root name should be provided. The median center is output as a point (MDN<root name>).

Directional Mean and Variance (DMean)

The directional mean and variance are calculated if the input variable is a collection of angular measures ($0^0 - 360^0$) and the coordinate system is defined as directions on the primary file screens. The directional mean is an angle while the directional variance is a relative indicator varying from 0 (no variance) to 1 (maximal variance). The tabular output can be printed. If an additional distance variable is input, then the mean distance of the measurements is also output.

Spatial Autocorrelation Indices

Spatial autocorrelation indices identify whether point locations are spatially related, either clustered or dispersed. Two spatial autocorrelation indices are calculated. Both require an intensity variable in the primary file.

Moran's "I" (MoranI)

Moran's "I" statistic is the classic indicator of spatial autocorrelation. It is an index of covariation between different point locations and is similar to a product moment correlation coefficient, varying from –1 to +1. The MoranI routine calculates 6 statistics:

1. The sample size
2. Moran's "I"
3. The spatially random (expected) "I"
4. The standard deviation of "I"
5. A significance test of "I" under the assumption of normality (Z-test)
6. A significance test of "I" under the assumption of randomization (Z-test)

Values of *I* greater than the expected *I* indicate clustering while values of *I* less than the expected *I* indicate dispersion. The significance test indicates whether these differences are greater than what would be expected by chance. The tabular output can be printed.

Adjust for small distances

If checked, small distances are adjusted so that the maximum distance weighting is 1 (see documentation for details). This ensures that *I* will not become excessively large for points that are close together. This is the default setting.

Geary's "C" (GearyC)

Geary's "C" statistic is an alternative indicator of spatial autocorrelation. It is an index of paired comparison between different point locations and varies from 0 (similar values) to 2 (dissimilar values). The GearyC routine calculates 5 statistics:

1. The sample size
2. Geary's "C""
3. The spatial random (expected) "C"
4. The standard deviation of "C"
5. A significance test of "I" under the assumption of normality (Z-test)

Values of *C* less than the expected *C* indicate clustering while values of *C* greater than the expected *C* indicate dispersion. The significance test indicates whether these differences are greater than what would be expected by chance. The tabular output can be printed.

Adjust for small distances

If checked, small distances are adjusted so that the maximum distance weighting is 1 (see documentation for details). This ensures that *C* will not become excessively large or excessively small for points that are close together. This is the default setting.

Distance Analysis

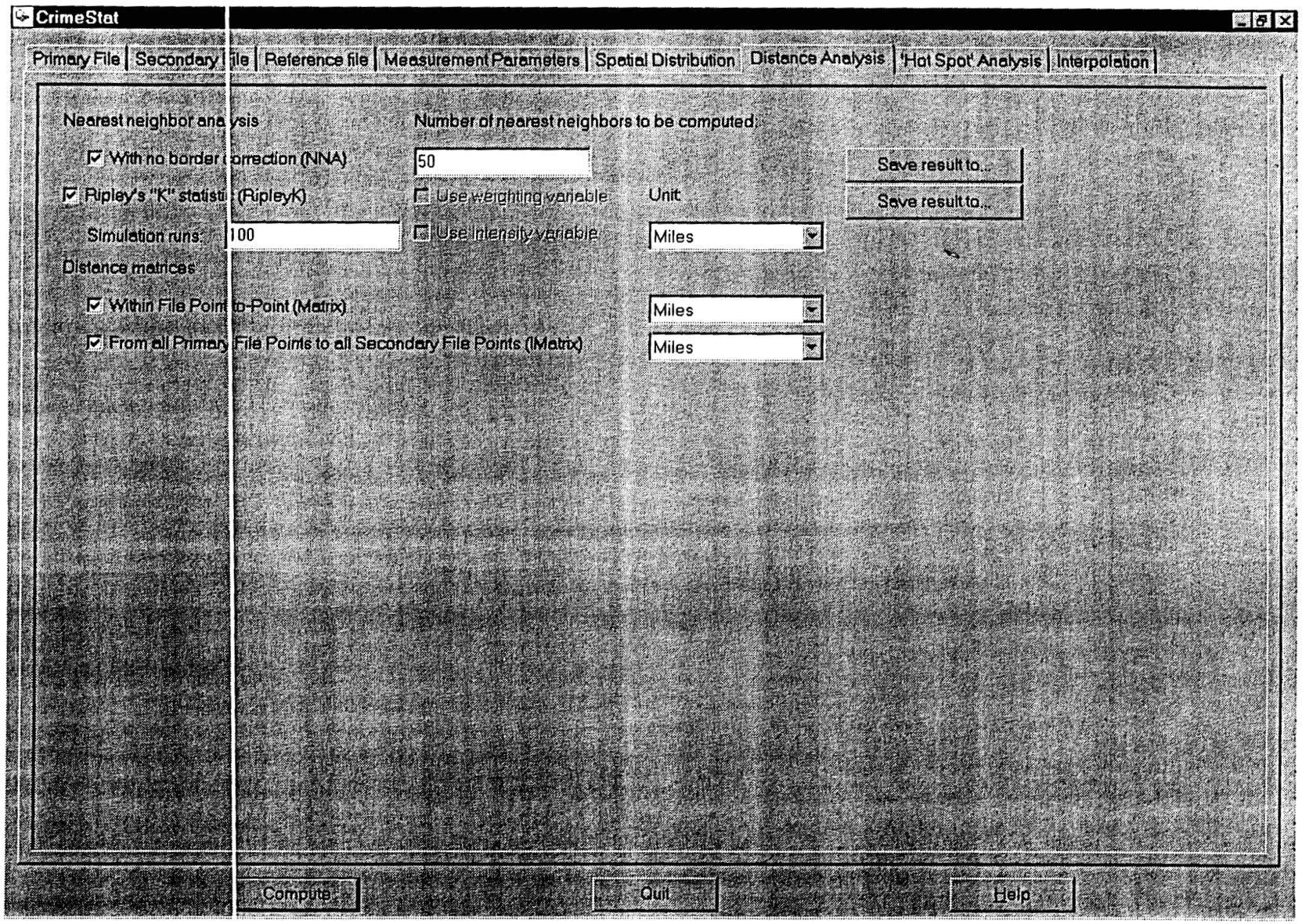
Distance analysis provides statistics about the distances between point locations. It is useful for identifying the degree of clustering of points. It is sometimes called second-order analysis. There are three routines for describing properties of the distances and there are two routines that output distance matrices.

Nearest Neighbor Analysis (Nna)

The nearest neighbor index provides an approximation about whether points are more clustered or dispersed than would be expected on the basis of chance. It compares the average distance of the nearest other point (nearest neighbor) with a spatially random expected distance by dividing the empirical average nearest neighbor distance by the expected random distance (the nearest neighbor index). The nearest neighbor routine requires that the geographical area be entered on the Measurement Parameters page and that direct distances be used. The *Nna* routine calculates 10 statistics:

1. The sample size
2. The mean nearest neighbor distance in meters, feet and miles
3. The standard deviation of the nearest neighbor distance in meters, feet and miles
4. The minimum distance in meters, feet and miles
5. The maximum distance in meters, feet and miles
6. The mean random distance (for both the maximum bounding rectangle and the user input area, if provided
7. The mean dispersed distance in meters, feet and miles (for both the maximum bounding rectangle and the user input area, if provided)
8. The nearest neighbor index (for both the maximum bounding rectangle and the user input area, if provided)

Figure 6: Distance Analysis Screen



9. The standard error of the nearest neighbor index (for both the maximum bounding rectangle and the user input area, if provided)
10. A significance test of the nearest neighbor index (Z-test)

The tabular results can be printed, saved to a text file, or saved as a '.dbf' file.

Number of nearest neighbors

The K-nearest neighbor index compares the average distance to the K^{th} nearest other point with a spatially random expected distance. The user can indicate the number of K-nearest neighbors to be calculated, if more than one are to be calculated. *CrimeStat* will calculate 3 statistics for each order specified:

1. The mean nearest neighbor distance in meters for the order
2. The expected nearest neighbor distance in meters for the order
3. The nearest neighbor index for the order

The *Nna* routine will use the user-defined area unless none is provided in which case it will use the maximum bounding rectangle. The tabular results can be printed, saved to a text file or output as a '.dbf' file.

Linear Nearest Neighbor Analysis

The linear nearest neighbor index provides an approximation about whether points are more clustered or dispersed along road segments than would be expected on the basis of chance. It is used with indirect (Manhattan) distances and requires the input of the total length of a road network on the measurement parameters page (see Measurement Parameters). That is, if indirect distances are checked on the measurement parameters page, then the linear nearest neighbor will be calculated. The linear nearest neighbor index is the ratio of the empirical average linear nearest neighbor distance to the expected linear random distance. The *Nna* routine calculates 9 statistics for the linear nearest neighbor index:

1. The sample size
2. The mean linear nearest neighbor distance in meters, feet and miles
3. The minimum distance between points along a grid network
4. The maximum distance between points along a grid network
5. The mean random linear distance
6. The linear nearest neighbor index
7. The standard deviation of the linear nearest neighbor distance in meters. feet and miles
8. The standard error of the linear nearest neighbor index
9. A t-test of the difference between the empirical and expected linear nearest neighbor distance

Number of linear nearest neighbors

Nna can calculate K-nearest linear neighbors and compare this distance the average linear distance to the K^{th} nearest other point with a spatially random expected distance. The user can indicate the number of K-nearest linear neighbors to be calculated, if more than one are to be calculated. *CrimeStat* will calculate 3 statistics for each order specified:

1. The mean linear nearest neighbor distance in meters for the order
2. The expected linear nearest neighbor distance in meters for the order
3. The linear nearest neighbor index for the order

Ripley's "K" Statistic (RipleyK)

Ripley's "K" statistic compares the number of points within any distance to an expected number for a spatially random distribution. The empirical count is transformed into a square root function, called L (see documentation for more details). Values of L that are greater than the upper limit of the simulations indicate concentration while values of L less than the lower limit of the simulations indicate dispersion. L is calculated for each of 100 distance intervals (bins). The RipleyK routine calculates 6 statistics:

1. The sample size
2. The maximum distance in meters, feet and miles
3. 100 distance bins
4. The distance for each bin
5. The transformed statistic, L(t), for each distance bin
6. The expected random L under complete spatial randomness, L(csr)

In addition, *CrimeStat* can estimate the sampling distribution by running spatially random simulations over the study area. If one or more spatially random simulations are specified, there are 6 additional statistics:

7. The minimum L value for the spatially random simulations
8. The maximum L value for the spatially random simulations
9. The 2.5 percentile L value for the spatially random simulations
10. The 97.5 percentile L value for the spatially random simulations
11. The 0.5 percentile L value for the spatially random simulations
12. The 99.5 percentile L value for the spatially random simulations

The tabular results can be printed, saved to a text file, or saved as a '.dbf' file.

Distance Matrices

CrimeStat can calculate the distances between points for a single file or the distances between points for two different files. These matrices can be useful for examining the frequency of different distances or for providing distances for another program.

Within File Point-to-Point (Matrix)

This routine outputs the distance between each point in the primary file to every other point in a specified distance unit (miles, nautical miles, feet, kilometers, or meters). The Matrix output can be saved to a text file.

From All Primary File Points to All Secondary File Points (IMatrix)

This routine outputs the distance between each point in the primary file to each point in the secondary file in a specified distance unit (miles, nautical miles, feet, kilometers, or meters). The IMatrix output can be saved to a text file.

'Hot Spot' Analysis

'Hot spot' (or cluster) analysis identifies groups of incidents that are clustered together. It is a method of second-order analysis that identifies the cluster membership of points. There are three statistics that can be used to identify 'hot spots': 1) Nearest neighbor hierarchical spatial clustering; 2) K-means clustering; and 3) Local Moran Statistics.

Nearest Neighbor Hierarchical Spatial Clustering (Nnh)

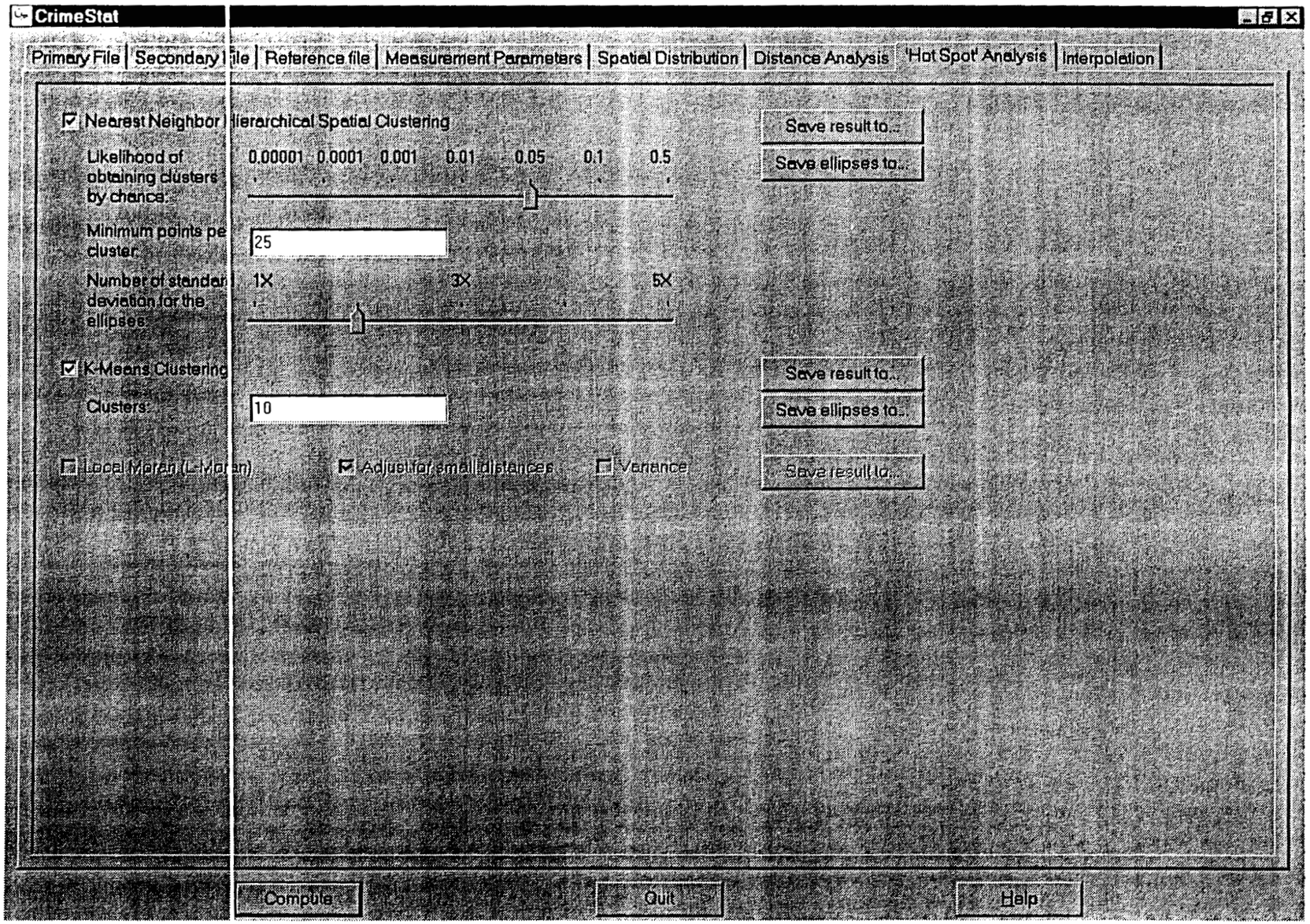
The nearest neighbor hierarchical spatial clustering routine groups points together on the basis of spatial proximity. The user defines a significance level associated with a threshold, a minimum number of points that are required for each cluster, and an output size for displaying the clusters with ellipses.

First, the threshold distance is the lower limit of the confidence interval around a random expected distance. The default value is 0.1 (i.e., fewer than 10% of distances could be expected to be as small or smaller by chance). Pairs of points that are closer together than the threshold distance are grouped together, whereas pairs of points that are greater than the threshold distance are ignored. The smaller the significance level that is selected, the smaller the threshold distance. Move the slider bar to the desired likelihood level.

Second, the minimum number of points required for each cluster allows the user to specify a minimum number of points for each cluster. The default is 10 points. Third, the output size for the clusters can be adjusted by the second slider bar. These are the number of standard deviations defined by the ellipse, from one standard deviation (the default value) to five standard deviations. Typically, one standard deviation will cover about 65% of the cases whereas five standard deviations will cover more than 99% of the cases.

Clustering is hierarchical in that the first-order clusters are treated as separate points to be clustered into second-order clusters, and the second-order clusters are treated as separate points to be clustered into third-order clusters, and so on. Higher-order clusters will be identified only if the distance between their centers are closer than a new

Figure 7: 'Hot Spot' Analysis Screen



threshold distance. The results can be printed, saved to a text file, output as a '.dbf' file, or output as ellipses to *ArcView* '.shp', *MapInfo* '.mif' or *Atlas*GIS* '.bna' files

Significance level

The threshold distance is adjusted by a significance level on the upper slider bar that indicates the Type I error for a one-tailed lower range of a confidence interval around an expected spatially random distance. Distances smaller than this threshold are candidates for clustering. The range of values vary from 0.5 likelihood for a Type I error (i.e., a distance that is equal to an expected spatially random distance) to 0.0001 likelihood for a Type I error (i.e., a distance that is very unlikely to come from a spatially random process). The higher the p-level chosen, the larger the area the clusters will cover with larger ellipses. The smaller the likelihood, then clusters will cover smaller areas with smaller ellipses. However, the higher the p-level chosen, the greater the likelihood that clusters could be chance groupings. Slide the bar to choose a significance level.

Minimum points per cluster

Restrictions on the number of clusters can be placed by defining a minimum number of points that are required. The default is 10. If there are too few points allowed, then there will be many very small clusters. By increasing the number of required points, the number of clusters will be reduced.

Output size for ellipses

The output size for the clusters can be adjusted by the lower slider bar. These are the number of standard deviations defined by the ellipse, from one standard deviation (1X - the default value) to five standard deviations (5X). The default value is one standard deviation. Typically, one standard deviation will cover about 60% of the cases whereas five standard deviations will cover more than 99% of the cases, although the exact percentage will depend on the distribution. Slide the bar to select the number of standard deviations for the ellipses.

K-means Clustering (KMeans)

The K-means clustering routine is a procedure for partitioning all the points into K groups in which K is a number assigned by the user. The default K is 5. The routine finds K seed locations in which the distance between points within a cluster are small but the distances between seed locations are large. If K is small, the clusters will typically cover larger areas. Conversely, if K is large, the clusters will typically cover smaller areas. The results can be printed, saved to a text file, output as a '.dbf' file, or output as ellipses to *ArcView* '.shp', *MapInfo* '.mif' or *Atlas*GIS* '.bna' files

Local Moran Statistics (L-Moran)

The local Moran statistic applies the Moran's "I" statistic to individual points (or zones) to assess whether particular points/zones are spatially related to the nearby points (or zones). The statistic requires an intensity variable in the primary file. Unlike the global Moran's "I" statistic, the local Moran is applied to each individual point/zone. The index points to clustering or dispersion relative to the local neighborhood. Points (or zones) with high "I" values have an intensity value that is higher than their neighbors while points with low "I" values have intensity values lower than their neighbors. The output can be printed or output as a '.dbf' file.

Adjust for small distances

If checked, small distances are adjusted so that the maximum distance weighting is no higher than 1 (see documentation for details). This ensures that the local "I" will not become excessively large for points that are grouped together. This is the default setting.

Interpolation

The interpolation tab allows estimates of point density using the kernel density smoothing method. There are two types of kernel density smoothing: one applied to a single distribution of points and the other applied to two different distributions. Each type has variations on the method that can be selected. Both types require a reference file that is overlaid on the study area (see Reference File). The kernels are placed over each point and the distance between each reference cell and each point is evaluated by the kernel function. The individual kernel estimates for each cell are summed to produce an overall estimate of density for that cell. The intensity and weighting variables can be used in the kernel estimate. The densities can be converted into probabilities.

Single Kernel Density Estimate (KernelDensity)

The single kernel density routine estimates the density of points for a single distribution by overlaying a symmetrical surface over each point, evaluating the distance from the point to each reference cell by the kernel function, and summing the evaluations at each reference cell.

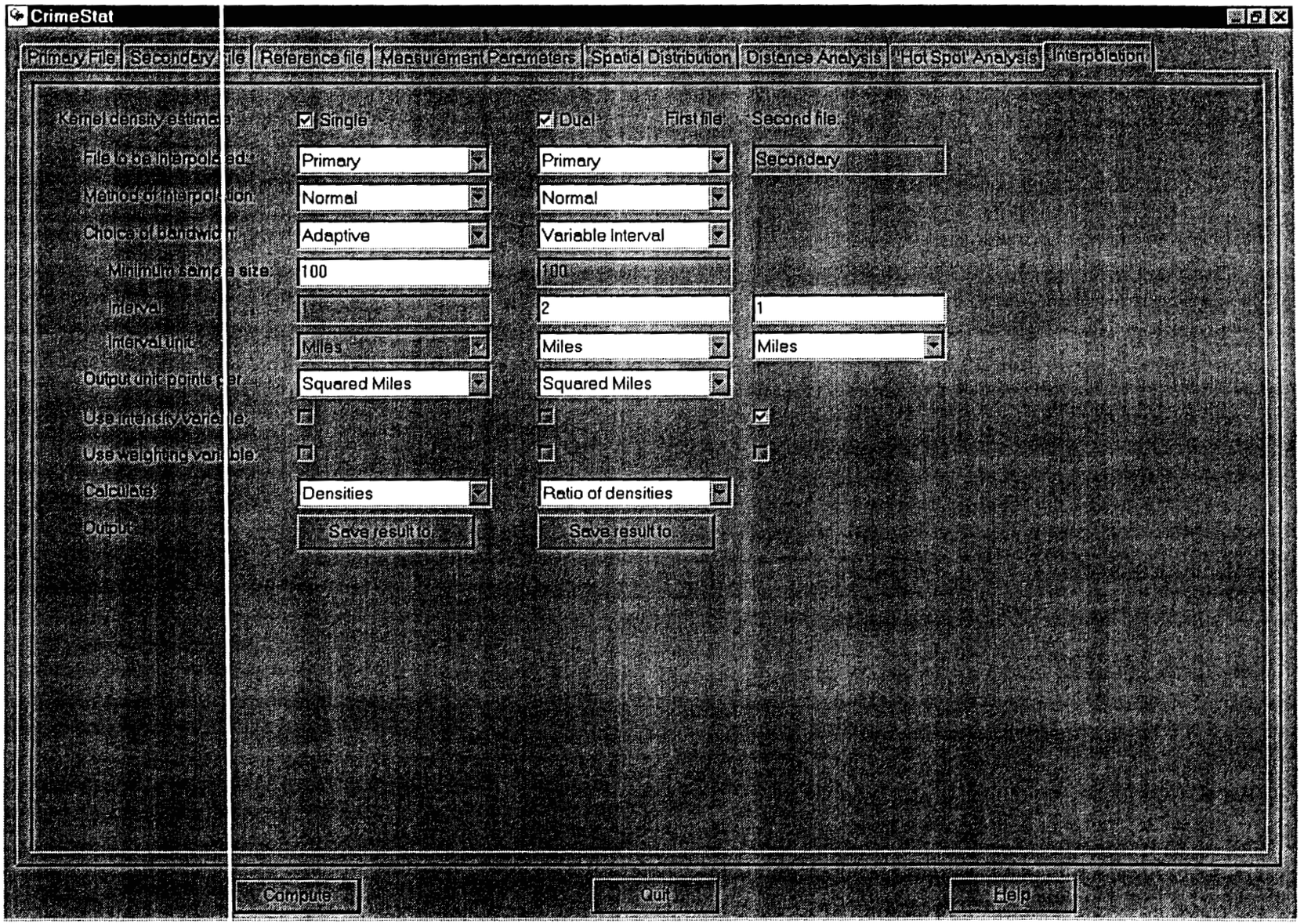
File to be interpolated

The estimate can be applied to either the primary file (see Primary file) or a secondary file (see Secondary File). Select which file is to be interpolated. The default is the Primary.

Method of interpolation

There are two types of kernel distributions that can be used to estimate the density points. The normal distribution overlays a normal distribution over each point, which then

Figure 8: Interpolation Screen



extends over the area defined by the reference file. This is the default kernel function. The quartic kernel overlays a surface that only extends for a limited distance from each point. The two methods produce similar results although the normal is generally smoother for any given bandwidth.

Choice of bandwidth

The kernels are applied over a limited search distance, called 'bandwidth'. For the normal kernel, bandwidth is the standard deviation of the normal distribution. For the quartic kernel, bandwidth is the radius of a circle defined by the surface. For both types, a larger bandwidth will produce smoother density estimates. For each, both adaptive and fixed bandwidth intervals can be selected.

Adaptive bandwidth

An adaptive bandwidth distance is identified by the minimum number of other points found within a circle drawn around a single point. A circle is placed around each point, in turn, and the radius is increased until the minimum sample size is reached. Thus, each point has a different bandwidth interval. This is the default bandwidth setting. The user can modify the minimum sample size. The default is 100 points.

Fixed bandwidth

A fixed bandwidth distance is a fixed interval for each point. The user must define the interval and the distance units by which it is calculated (miles, nautical miles, feet, kilometers, meters).

Output units

Specify the density units as points per square mile, per squared nautical miles, per square feet, per square kilometers, or per square meters. The default is points per square mile.

Use intensity variable

If an intensity variable is being interpolated, then this box should be checked.

Use weighting variable

If a weighting variable is being used in the interpolation, then this box should be checked.

Calculate densities or probabilities

Select whether densities (points per square unit of area) or probabilities (the proportion of all points) are to be output for each cell. The default is densities.

Output

The results can be output as a *Surfer for Windows* file (for both an external or generated reference file) or as an *ArcView* '.shp', *MapInfo* '.mif', *Atlas*GIS* '.bna', or *ArcView Spatial Analyst*© file (only if the reference file is generated by *CrimeStat*).

Dual Kernel Density Estimate (DuelKernel)

The dual kernel density routine compares two different distributions involving the primary and secondary files. A 'first' file and 'second' file need to be defined. The comparison allows the ratio of the first file divided by the second file, the logarithm of the ratio of the first file divided by the second file, the difference between the first file and second file (i.e., first file – second file), or the sum of the first file and the second file.

File to be interpolated

Identify which file is to be the 'first file' (primary or secondary) and which is to be the 'second file (primary or secondary). The default is Primary for the first file and Secondary for the second file.

Method of interpolation

There are two types of kernel distributions that can be used to estimate the density points. The normal distribution overlays a normal distribution over each point, which then extends over the area defined by the reference file. This is the default kernel function. The quartic kernel overlays a surface that extends only for a limited distance from each point. The two methods produce similar results although the normal is generally smoother for any given bandwidth.

Choice of bandwidth

The kernels are applied to a limited search distance, called 'bandwidth'. For the normal kernel, bandwidth is the standard deviation of the normal distribution. For the quartic kernel, bandwidth is the radius of a circle defined by the surface. For both types, larger bandwidth will produce smoother density estimates. For each, adaptive, fixed and variable bandwidth intervals can be selected.

Adaptive bandwidth

An adaptive bandwidth distance is identified by the minimum number of other points found within a circle drawn around a single point. A circle is placed around each point, in turn, and the radius is increased until the minimum sample size is reached. Thus, each point has a different bandwidth interval. This is the default bandwidth setting. The user can modify the minimum sample size. The default is 100 points.

Fixed bandwidth

A fixed bandwidth distance is a fixed interval for each point. The user must define the interval and the distance units by which it is calculated (miles, nautical miles, feet, kilometers, meters). The default is one mile.

Variable bandwidth

A variable bandwidth allows separate fixed intervals for both the first and second files. For each, the user must define the interval and the distance units by which it is calculated (miles, nautical miles, feet, kilometers, meters). The default is one mile for both the first and second files.

Output units

Specify the density units as points per square mile, per square nautical miles, per square feet, per square kilometers, or per square meters. The default is points per square mile.

Use intensity variable

For the first and second files separately, check the appropriate box if an intensity variable is being interpolated.

Use weighting variable

For the first and second files separately, check the appropriate box if a weighting variable is being used in the interpolation.

Calculate densities or probabilities

Select whether densities (points per square unit of area) or probabilities (the proportion of all points) are to be output for each cell. The default is densities.

Output

The results can be output as a *Surfer for Windows* file (for both an external or generated reference file) or as an *ArcView* '.shp', *MapInfo* '.mif', *Atlas*GIS* '.bna', or *ArcViewSpatial Analyst* only if the reference file is generated by *CrimeStat*).

Dynamic Data Exchange (DDE) Support

CrimeStat supports Dynamic Data Exchange (DDE). See Appendix A in the documentation or the online help screens for more information.