Social Science Computer Review

http://ssc.sagepub.com

Software Review: Spatial Data Analysis of Crime: A Review of CrimeStat III

Cedrick G. Heraux

Social Science Computer Review 2007; 25; 259
DOI: 10.1177/0894439307298922

The online version of this article can be found at:
http://ssc.sagepub.com/cgi/content/abstract/25/2/259

Published by:
SAGE Publications
http://www.sagepublications.com

Social Science Computer Review
Volume 25 Number 2
Summer 2007 259–264

10.1177/0894439307298922
http://ssc.sagepub.com
hosted at
http://online.sagepub.com

Spatial Data Analysis of Crime

A Review of CrimeStat III

Cedrick G. Heraux
National Archive of Criminal Justice Data (ICPSR), University of Michigan

Title: *CrimeStat III: A Spatial Statistics Program for the Analysis of Crime Incident Locations*
Author: Ned Levine and Associates
Publisher: National Institute of Justice
Year of Publication: 2004
Version Reviewed: 3.0
Requirements: CrimeStat III is capable of running in a Windows XP Professional, Windows 2000, or Windows NT operating system. Recommended hardware includes a 1.6 GHz processor and 256 MB of RAM. Input formats are dBase (III, IV, or V), ArcView shape (.shp) files, MapInfo data (.dat) files, and numerous spreadsheet files (e.g., Excel, Lotus 1-2-3, and Microsoft Access). Output formats are tables displayed as text, and graphical files in the appropriate geographic information system (GIS) format (e.g., ArcView .shp files, and MapInfo .dat files).
Materials: CrimeStat III is distributed through the National Archive of Criminal Justice Data (ICPSR) as a zipped (.zip) file, which contains the executable file as well as the manual (which is divided into 22 separate .pdf files).

CrimeStat III is a spatial statistics program for the analysis of crime incident locations. The program has three main components: (1) the Spatial Description component allows the user to conduct analyses regarding spatial distribution of data, (2) the Spatial Modeling component allows the user to analyze the spatial behavior of data, and (3) the Crime Travel Demand Modeling component allows the user to analyze the behavior of potential serial offenders. Overall, CrimeStat III is fairly easy to use once an initial familiarity has been established with the software. This can be accomplished through the manual or the intuitive GUI. Currently, the only significant restriction to the utility of the software is its reliance on the Windows operating system. This review provides a discussion of technical utilization of the software, and a look at how this software is useful in a practical manner for the spatial data analysis of crime.

Keywords: *crime analysis; CrimeStat III; spatial statistical analysis*

Data Setup

Although the data setup component of the program does not provide analytical tools, it is perhaps the most important aspect of the software. This component is essential in allowing the user to define the files which will be used in further analyses, and thus requires the specification of numerous types of files. The user must first identify the primary file to be used as the basis for all analyses. This file must contain point data on the phenomenon of interest (e.g., assaults or home invasions) as well as X and Y (latitude and longitude) coordinates, and the user must inform the software of the appropriate variable names for these coordinates. In addition, the user may specify intensity (Z) and weight variables, as well as identify any variables related to time (e.g., days, weeks, and months) to be used in analyses. The data setup component also allows the user to specify a secondary file, which must contain the same variable names and coordinate system as the primary file. This file is used for comparison to the primary file in risk-adjusted nearest neighbor clustering (to be discussed later in this review). The reference file is used to overlay a grid file on the primary/secondary file, allowing the user to conduct analyses specific to grid-based systems. The final aspect of the data setup component is the measurement parameters, which allow the user to identify the type of distance measurement used (direct, indirect, or network) and specify the length of the street network, if necessary. The use of these parameters allows for more effective analyses of the interaction between points in the primary/secondary file.

Spatial Description

The spatial description component of the software provides the user with numerous ways to describe their data, both in terms of distribution and in the form of simple analyses. The user can obtain the mean center and standard distance for the data, as well as the standard deviational ellipse. In addition, measures of spatial autocorrelation are provided (Moran's "I" and Geary's "C"), allowing the user to determine if the data points are closer together than one would expect from a random distribution. The software also provides various types of distance analyses, which allow the user to describe various aspects of the distance between data points. These include nearest neighbor analysis and Ripley's "K" statistic, as well as the use of a distance matrix, which calculates the distance between data points in several ways (within the primary file, between the primary file and the secondary file, and between either the primary or secondary file and the grid file). The final aspect of the spatial description component is the use of hotspots analysis, which contains methods for fuzzy mode, hierarchical nearest neighbor clustering, and risk-adjusted hierarchical nearest neighbor clustering. In addition, the software allows for hotspots analysis in the form of spatial and temporal analysis of crime (STAC) and k-means clustering. The most important aspect of the spatial description component is that the results of these analyses can be saved to an ArcView- or MapInfo-compatible file, allowing the user to import the ellipses and point data back into their preferred GIS program.

Spatial Modeling

As with any social scientific endeavor, although data description is useful, data modeling allows the social scientist to better understand that data. The spatial modeling component of the software includes three different types of analysis for this purpose. In interpolation, the user can specify either a single-variable kernel density estimate, which produces a surface or contour estimate of point density (e.g., police use of force), or a dual-variable kernel density estimate, which compares the point density (e.g., police use of force) to the density of a baseline measure (e.g., total number of arrests). In a further application of modeling clustering, the space-time analysis routine allows the user to analyze the relationship between time and space for the existing data points (through the Knox Index or the Mantel Index), or to predict the behavior of a serial offender (through the correlated walk analysis). In addition, the software has recently added a spatial-temporal moving average, which allows the user to compensate for small sample size. The final aspect of the spatial modeling component is the journey to crime analysis, which allows the user to estimate the expected location of a serial offender (e.g., a rapist) given the distribution of data points (e.g., a series of rapes) and a model of traveling distance (e.g., using a street or transportation network). As with the spatial description component, the various results (e.g., density patterns, graphs for the moving average, and crime trips) can be output in formats compatible with numerous GIS programs.

Crime Travel Demand Modeling

The final component of the software to be discussed is that of crime travel demand modeling, which was recently added to CrimeStat III. This analysis is conducted by defining a crime trip as any link between an offender's origin (e.g., residence, place of employment, or site of immediately preceding crime) and a crime location. Dividing the data points into zones, the routine counts the number of crimes which either originate or terminate (or both) in each zone. This is accomplished by calibrating the model, and separately predicting the number of crimes originating in each zone (origins) and terminating in each zone (destinations) using variants of a Poisson regression model. Once these models have been estimated, the software can run a balancing routine to ensure that there are an equal number of origins and destinations. The crime travel demand component then analyzes the distribution of crime trips by using a gravity model, which predicts movement between two places based on the relative importance of the origin and destination locations, and the functional distance between those locations. This allows for calculation of the observed distribution from the point data as well, and the software then provides a separate routine for comparing the predicted and observed trip distributions. In consideration of the availability of different networks to the offender, the software also provides the user with the means to predict numerous trip models based on a variety of methods of travel (on foot, on a bicycle, in a car, on a bus, and on a train). The final aspect of the crime travel demand modeling component is that of network assignment, by which users can analyze the most likely route for each trip. The software allows the user to determine these routes using a "shortest path" algorithm (much as would one using a website such as Mapquest), taking into account factors such as one-way streets (for cars and buses) or transit station/stop locations (for buses and trains). As with all previous components, the crime travel demand modeling component can save out the results (e.g., predicted and observed points, trip routes, and completed networks) in a format compatible with numerous GIS programs.

Practical Utilization of CrimeStat III

Although CrimeStat has been used throughout the past several years as a means of spatial data analysis for a variety of crime types, it is most often used to examine serial crimes such as murder, rape, and burglary. In keeping with this tradition, this review will examine the practical utilization of CrimeStat III in the analysis of serial crimes. As an example, consider a law enforcement database of numerous burglaries over the past 3 months within a certain jurisdiction. Once this database is properly mapped in a GIS package such as ArcMap or MapInfo, the corresponding shapefile database (i.e., the database with all burglaries mapped as point data) is imported into CrimeStat III using the Data Setup tab. After providing the software with the variable names for the *X* and *Y* coordinates (often labeled either *X* and *Y*, or *Latitude* and *Longitude*, in the database), the user defines the type of coordinate system being used and indicates any time units needed for space-time analyses (noting that formatted dates such as 01/01/2007 are not accepted). In addition, the Measurement Parameters subtab should be used to identify the area being studied, particularly with regard to the street network, in order to perform more complex analyses. Using the line or polyline street file from a GIS program allows the user to indicate that the network to be used is a multidirectional network consisting of street segments upon which the offender can travel.

Once the appropriate files have been identified, the user should turn to the Spatial Description tab in order to perform simple analyses which will help to describe the data. The subtab for Spatial Distribution allows the user to obtain first-order spatial statistics such as the mean center and standard distance. Using various *X* and *Y* coordinates, the software is able to calculate the arithmetic mean location (i.e., the "average" incident location) and the degree of dispersion of all burglary incidents. The standard deviational ellipse allows the user to examine the directionality of that dispersion, whereas the median center provides the user with point data for the middle of the distribution. With respect to spatial autocorrelation, the software is able to provide a Moran's "I" statistic, which is an indicator of the spatial autocorrelation (i.e., covariance) among burglary incidents, or a Geary's "C" statistic, which is an index of paired comparisons between burglary incidents assigning a 0 for completely similar values and a 2 for completely dissimilar values. The two subtabs for Distance Analysis provide second-order statistics regarding the degree of clustering of data points. The user can select a nearest neighbor analysis to determine if the clustering of points is more or less than would be expected by chance. In the context of data based on a network, the user can select a linear nearest neighbor analysis to determine if data points are more or less clustered along particular road segments than would be expected by chance. From discussions with users of CrimeStat III, it has become apparent that many choose to focus on the two Hot Spots Analysis subtabs when examining point data in a potential crime series. Building on the routines in the Distance Analysis subtab, the nearest neighbor hierarchical spatial clustering routine allows the user to identify clusters of points on the basis of spatial proximity through the output of an ellipse able to be imported into standard GIS software. The user can also adjust for the distribution of a baseline variable, such as population or number of occupied houses, using the risk-adjusted nearest neighbor hierarchical spatial clustering routine. Further refinements can be obtained by using the spatial and temporal analysis of crime routine, which uses a variable-distance clustering to combine clusters that overlap one another. As an example, consider a jurisdiction interested in determining if a series of burglary incidents over the past 3 months has been the work of only one offender. Whereas the spatial distribution statistics will identify the general pattern of these incidents, the second-order distance analysis statistics will identify if these burglaries are more clustered along particular city streets than would be expected by chance. Using Hot Spots Analysis, the investigator could then determine the extent of the clustering and identify particular areas of concern for law enforcement based on the ellipses output by the software.

Although the Spatial Description routines allow the user to examine how the point data are distributed, including clustering, the Spatial Modeling tab allows for more specific analyses related to building a theoretical model. In keeping with our example of burglaries over the past 3 months, here the user would select a dual-variable kernel density estimation routine (under the Interpolation subtab) which would compare the density of incidents (i.e., burglaries) to the density of the baseline variable (i.e., occupied houses). A variety of variables could be used as the baseline in this situation (e.g., population and total number of crimes reported) in order to allow the user to explore several explanatory models for the series of burglaries. In addition, under the Space-Time Analysis subtab, the user can conduct a correlated walk analysis, which analyzes the distribution of incidents over space and time in order to determine a sequence of events and make predictions regarding the location and time of the next event. Thus, using data on the location and time of prior burglaries, the jurisdiction can make predictions regarding the details of the next burglary. Under the Journey to Crime Analysis subtab, the software estimates a travel model of all crime incidents and then provides an estimate of the offender's central location given the spatial-temporal distribution of incidents and available methods of travel. Continuing our example, law enforcement could then obtain an estimate of the area where the serial burglar is likely to live based on all of this information.

The Crime Travel Demand tab is a new addition to the CrimeStat software and focuses on sequential modeling of crime travel by zones over a particular network or jurisdiction. Each crime trip is defined as belonging to a specific zone based on where the crime occurred (destinations) and where the offender started the journey (origins). In the Trip Generation subtab, separate models are produced to predict the number of origins and destinations in each zone, with a balancing procedure ensuring that the former are equal to the latter. In the Trip Distribution subtab, crime trips are distributed using a gravity model, which predicts the relative strength of the bond, and movement, between two points based on population size and distance. The Mode Split subtab allows the user to examine a variety of possible travel modes for the offender between origins and destinations, in order to predict the most likely mode of travel. The Network Assignment subtab builds on this work by assigning predicted trips from each zone to the most likely trip route based on travel mode using the shortest possible path. Thus, continuing with our burglary example, the Crime Travel Demand tab could be used to (a) examine each trip from the offender's predicted location to a crime incident, (b) make predictions regarding crime trips based on available travel methods, and (c) make predictions regarding the offender's most likely route(s) to crime incidents.

Overall, CrimeStat III provides numerous statistical analyses which are beneficial in the spatial analysis of crime. The software allows for both simple analyses regarding the distribution of crime incidents and more complex analyses dealing with model testing.

Although the manual is somewhat lengthy, the software itself is easy to use, and users will gain familiarity with repeated usage. This particular version of the software introduces a new module (Crime Travel Demand) which serves to significantly increase its utility and allows users to experience a full range of applications to spatial data analysis of crime.

Cedrick Heraux recently received his PhD in criminal justice from Michigan State University. His research focuses on police behavior and on the use of geographic information systems for spatial statistical analysis. He can be reached at cheraux@isr.umich.edu.