Chapter 2:

Quickguide to *CrimeStat IV*

Ned Levine
Ned Levine & Associates
Houston, TX

Table of Contents

Table of Contents (continued)

Chapter 2:
Quickguide to *CrimeStat IV*

Introduction

The following are brief instructions for the use of *CrimeStat*® *IV* and parallels the online help menus in the program. Because there are a large number of routines in *CrimeStat*, this quickguide is very long. Detailed instructions on individual routines should be obtained from Chapters 3-32 in the documentation.

CrimeStat has five basic groupings in 27 program tabs and one option tab. Each tab lists routines, options and parameters:

Data setup

1. Primary File
2. Secondary File
3. Reference File
4. Measurement Parameters

Spatial description

5. Spatial Distribution
6. Spatial Autocorrelation
7. Distance Analysis I
8. Distance Analysis II

Hot spot analysis

9. Hot Spot Analysis I
10. Hot Spot Analysis II
11. Hot Spot Analysis of Zones

Spatial modeling I

12. Interpolation I
13. Interpolation II
14. Space-time Analysis

15. Journey-to-crime
16. Bayesian Journey-to-crime

Spatial modeling II

17. Regression I
18. Regression II
19. Discrete Choice I
20. Discrete Choice II
21. Time Series Forecasting

Crime Travel Demand

22. Project Directory
23. Trip Generation
24. Trip Distribution
25. Mode Split
26. Network Assignment
27. File Worksheet

Options

28. Saving parameters, colors and options

Throughout this chapter, figures 2.1-2.28 show the 28 tab screens with examples of data input and routine selection.

I. Data Setup

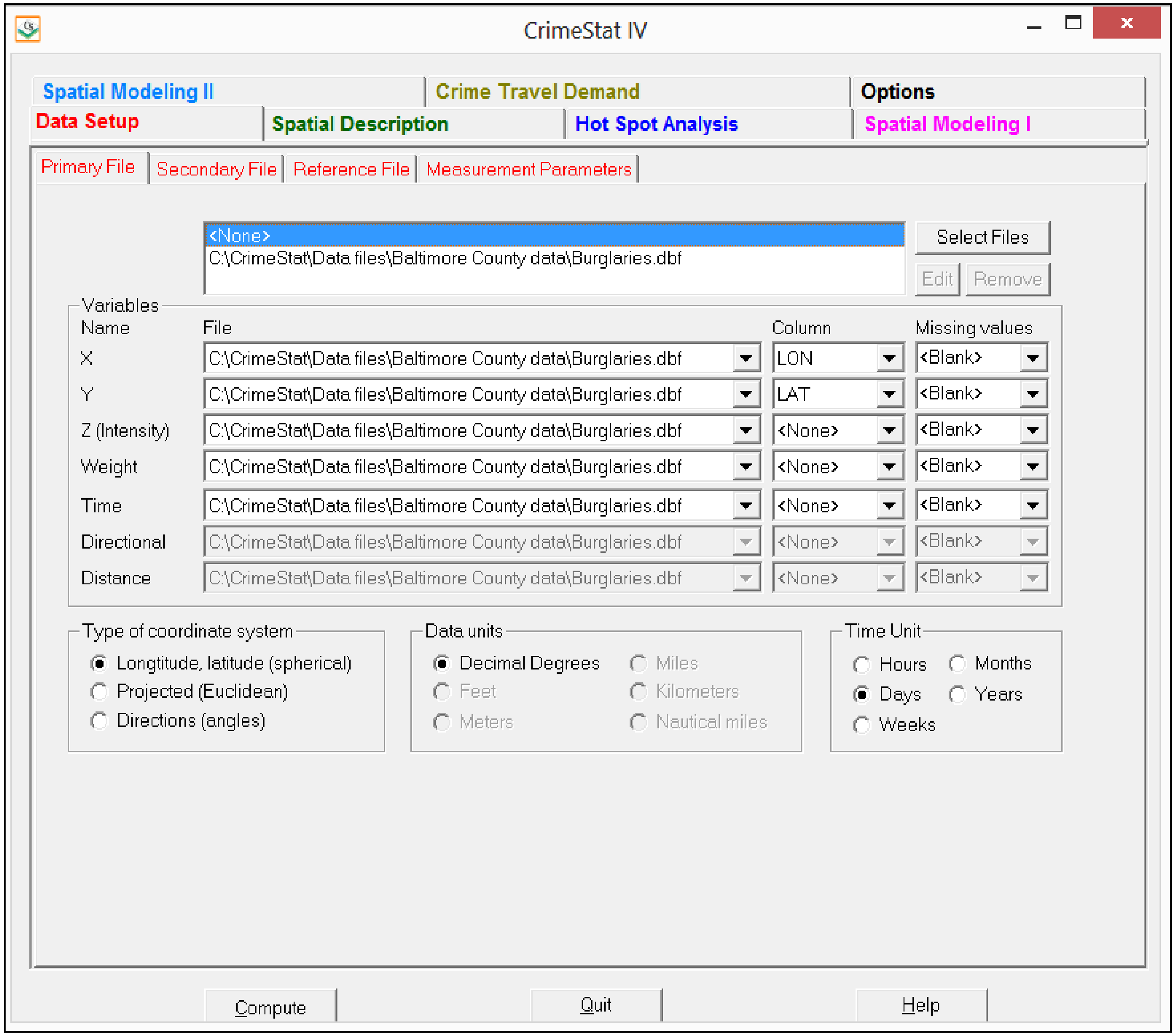
The data setup section involves defining the data set and variables for a primary file (required) and a secondary file (optional), identifying a reference grid (required for several routines), and defining measurement parameters (required for several routines).

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 have an

Figure 2.1:

Primary File Setup



<None>
C:\CrimeStat\Data files\Baltimore County data\Burglaries.dbf

Variables

Name	File	Column	Missing values
X	C:\CrimeStat\Data files\Baltimore County data\Burglaries.dbf	LON	<Blank>
Y	C:\CrimeStat\Data files\Baltimore County data\Burglaries.dbf	LAT	<Blank>
Z (Intensity)	C:\CrimeStat\Data files\Baltimore County data\Burglaries.dbf	<None>	<Blank>
Weight	C:\CrimeStat\Data files\Baltimore County data\Burglaries.dbf	<None>	<Blank>
Time	C:\CrimeStat\Data files\Baltimore County data\Burglaries.dbf	<None>	<Blank>
Directional	C:\CrimeStat\Data files\Baltimore County data\Burglaries.dbf	<None>	<Blank>
Distance	C:\CrimeStat\Data files\Baltimore County data\Burglaries.dbf	<None>	<Blank>

Type of coordinate system
- ● Longtitude, latitude (spherical)
- ○ Projected (Euclidean)
- ○ Directions (angles)

Data units
- ● Decimal Degrees
- ○ Feet
- ○ Meters
- ○ Miles
- ○ Kilometers
- ○ Nautical miles

Time Unit
- ○ Hours
- ● Days
- ○ Weeks
- ○ Months
- ○ Years

associated X and Y coordinates. Also, there can be associated weights or intensities, though these are optional. Also, there can be time references, though 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. The time references are used in the space-time analysis routines are defined in terms of hours, days, weeks, months, or years.

Select Files

Select the primary file. *CrimeStat* reads dbase 'dbf', ArcGIS point 'shp' and ASCII files. Select the type of file to be selected. Use the browse button to search for a particular file name. If the file type is ASCII, select the type of data separator (comma, semicolon, space, tab) and the number of columns. Note that there is a utility that will convert an Excel 'xls' or 'xlsx' to a 'dbf' file on the Options tab.

Variables

Define the file that contains the X and Y coordinates. If there are weights or intensities being used, define the file that contains these variables. Certain statistics (e.g., spatial autocorrelation, local Moran) require intensity values and most other statistics can use intensity values. Most other statistics can use weights. It is possible to have both an intensity variable and a weighting variable, though the user should be cautious in doing this to avoid 'double weighting'. If a time variable is used, it must be an integer or real number (e.g., 1, 36892). Do not use formatted dates (e.g., 01/01/2001, October 1, 2001). Convert these to real numbers before using the space-time analysis routines.

Columns

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. If a time variable is used, select the appropriate variable name.

Missing Values

Identify whether there are any missing values. By default, *CrimeStat* will ignore records with blank values in any of the eligible fields or records with non-numeric values (e.g., alphanumeric characters, , *). Blanks will always be excluded unless the user selects **<*none*>**. There are 8 possible options:

1. *<blank>* fields are automatically excluded. This is the default
2. *<none>* indicates that no records will be excluded. If there is a blank field, *CrimeStat* will treat it as a 0
3. *0* is excluded
4. *–1* is excluded
5. *0 and –1* indicates that both 0 and -1 will be excluded
6. *0, -1 and 9999* indicates that all three values (0, -1, 9999) will be excluded
7. *Any* other numerical value can be treated as a missing value by typing it (e.g., 99)
8. *Multiple* numerical values can be treated as missing values by typing them, separating each by commas (e.g., 0, -1, 99, 9999, -99)

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 being used, there can be an optional distance variable for the measurement. Define the file name and variable name (column) that contains the distance variable.

Time Units

Define the units for the time variable and are defined in terms of hours, days, weeks, months, or years. Time is only used for the primary file. The default value is days. Note, only integer or real numbers can be used (e.g., 1, 36892). Do not use formatted dates (e.g., 01/01/2001, October 1, 2001). Convert these to real or integer numbers before using the space-time analysis routines.

Type of Coordinate System and Data Units

Select the type of coordinate system. If the coordinates are 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 feet (e.g., State Plane), meters (e.g., UTM.), miles, kilometers, or nautical miles**.** If the coordinate system is directional, then the coordinates are angles and the data units box will be blanked out. For directions, an additional distance variable can be used. This measures the distance of the incident from an origin location; the units are undefined.

Note: if a projected coordinate system is used, but the coordinate system is defined as longitude/latitude (spherical), an error message will appear that says "Found invalid data at row 1 of the primary data set!". Change the coordinate system to Projected (Euclidean).

Secondary File

A secondary data file is optional. It is also a point file with X and Y coordinates. It 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. More than one file can be selected. Time units are not used in the secondary file.

Select Files

Select the secondary file. *CrimeStat* reads dbase ‘dbf’, ArcGIS point ‘shp’ and ASCII files. If the file type is ASCII, select the type of data separator (comma, semicolon, space, tab) and the number of columns. Note that there is a utility that will convert an Excel ‘xls’ or ‘xlsx’ to a ‘dbf’ file on the Options tab.

Variables

Define the file that contains the X and Y coordinates. If weights or intensities are being used, define the file that contains these variables. Certain statistics (e.g., spatial autocorrelation, local Moran) require intensity values and most other statistics can use intensity values. Most other statistics can use weights. It is possible to have both an intensity variable and a weighting variable, though the user should be cautious in doing this to avoid 'double weighting'. Time units are not used in the secondary file.

Columns

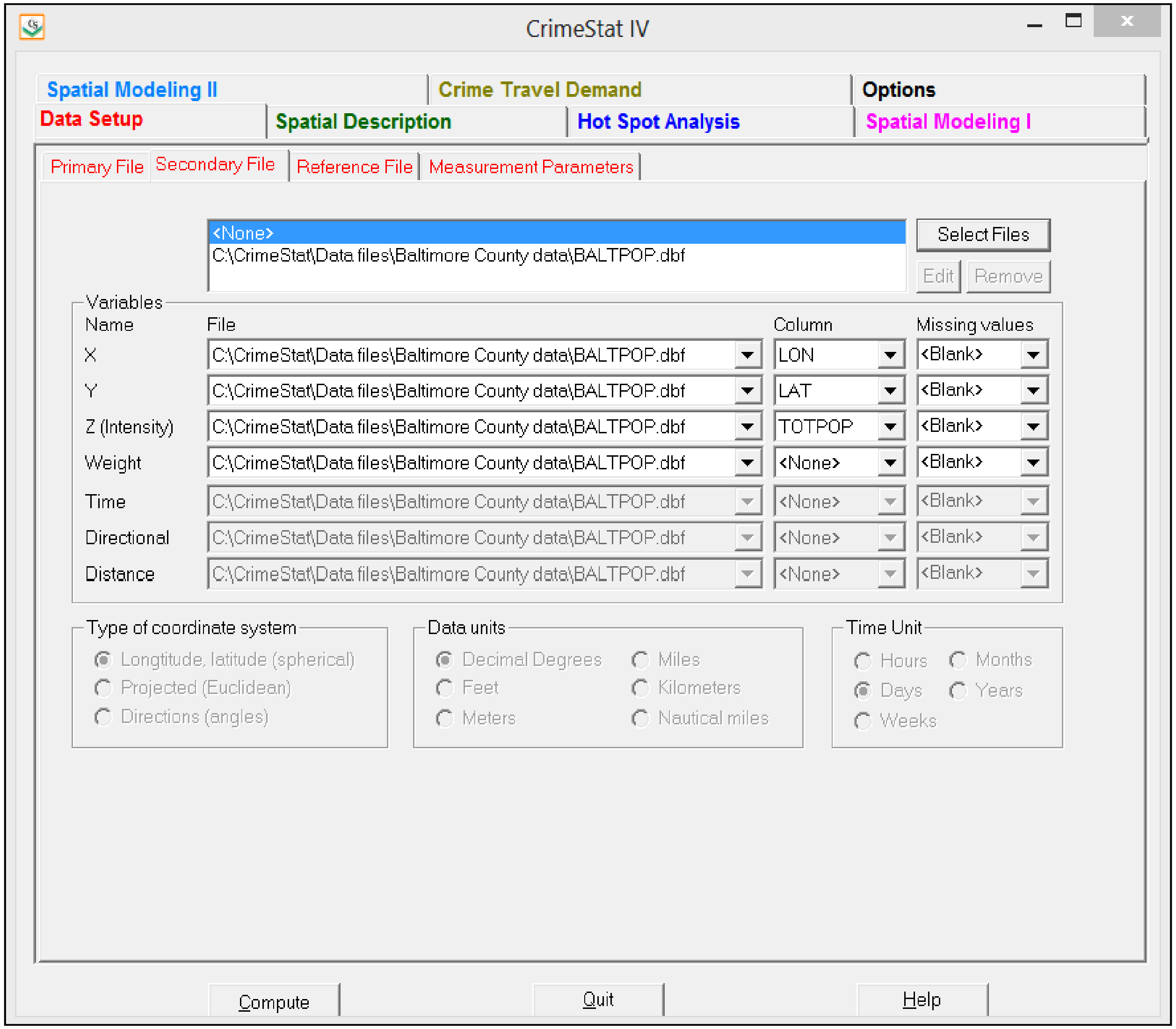
Select the variables for the X and Y coordinates respectively (e.g., Lon, Lat, Xcoord, Ycoord.) If there are weights or intensities being used, select the appropriate variable names. Time units are not used in the secondary file.

Missing Values

Identify whether there are any missing values. By default, *CrimeStat* will ignore records with blank values in any of the eligible fields or records with non-numeric values

Figure 2.2:

Secondary File Setup



<None>
C:\CrimeStat\Data files\Baltimore County data\BALTPOP.dbf

Variables

Name	File	Column	Missing values
X	C:\CrimeStat\Data files\Baltimore County data\BALTPOP.dbf	LON	<Blank>
Y	C:\CrimeStat\Data files\Baltimore County data\BALTPOP.dbf	LAT	<Blank>
Z (Intensity)	C:\CrimeStat\Data files\Baltimore County data\BALTPOP.dbf	TOTPOP	<Blank>
Weight	C:\CrimeStat\Data files\Baltimore County data\BALTPOP.dbf	<None>	<Blank>
Time	C:\CrimeStat\Data files\Baltimore County data\BALTPOP.dbf	<None>	<Blank>
Directional	C:\CrimeStat\Data files\Baltimore County data\BALTPOP.dbf	<None>	<Blank>
Distance	C:\CrimeStat\Data files\Baltimore County data\BALTPOP.dbf	<None>	<Blank>

Type of coordinate system
- ● Longtitude, latitude (spherical)
- ○ Projected (Euclidean)
- ○ Directions (angles)

Data units
- ● Decimal Degrees
- ○ Feet
- ○ Meters
- ○ Miles
- ○ Kilometers
- ○ Nautical miles

Time Unit
- ○ Hours
- ● Days
- ○ Weeks
- ○ Months
- ○ Years

(e.g., *, alphanumeric characters , *). Blanks will always be excluded unless the user selects ***<none>***. There are 8 possible options:

1. *<blank>* fields are automatically excluded. This is the default
2. *<none>* indicates that no records will be excluded. If there is a blank field, *CrimeStat* will treat it as a 0
3. *0* is excluded
4. *–1* is excluded
5. *0 and –1* indicates that both 0 and -1 will be excluded
6. *0, -1 and 9999* indicates that all three values (0, -1, 9999) will be excluded
7. *Any* other numerical value can be treated as a missing value by typing it (e.g., 99)
8. *Multiple* numerical values can be treated as missing values by typing them, separating each by commas (e.g., 0, -1, 99, 9999, -99)

Type of Coordinate System and Data Units

The secondary file must have 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. Directional coordinates (angles) are not allowed for the secondary file nor are time variables.

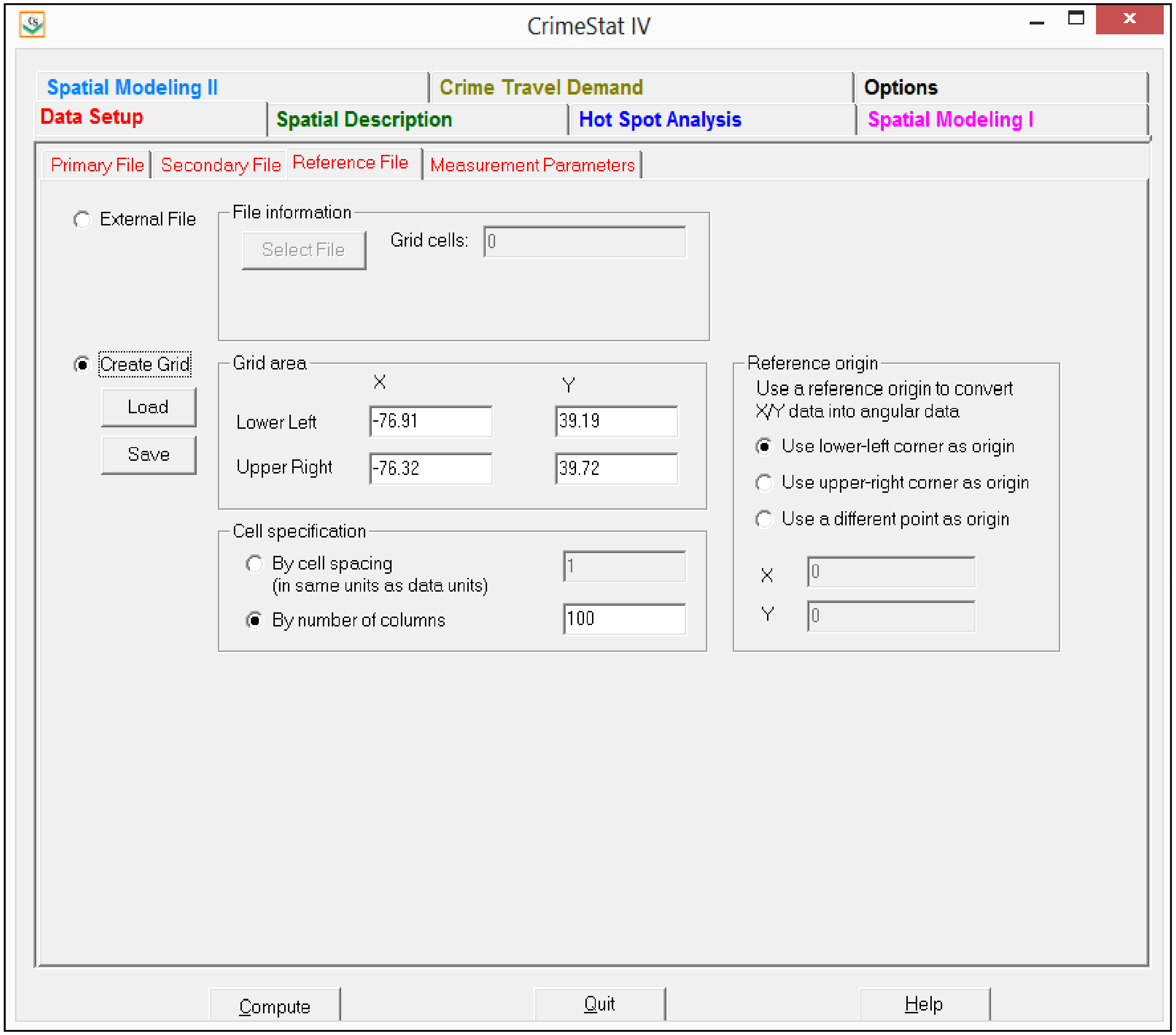
Reference File

For referencing the study area, there is a reference grid, a reference origin, and an area. The reference file is used in the risk-adjusted nearest neighbor hierarchical clustering routine, journey-to-crime estimation and in the single and dual variable kernel density estimation routines. 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. The reference origin is used in the directional mean routine. The file can be an external file that is input or can be generated by *CrimeStat*. The area is that of the study region.

Create Reference Grid

If allowing *CrimeStat* to generate a true grid, click on 'generated' and then input the lower left and upper right X and Y coordinates of a rectangle placed over the study area. Cells can be defined either by cell size, in the same coordinates and data units as the primary file, or by the number of columns in the grid (the default). In addition, a reference origin can be defined for the directional mean routine. The reference grid can be saved and re-used. Click on 'Save' and enter a file name. To use an already saved file, click on 'Load' and the file name.

Figure 2.3:
Reference File Setup



○ External File

File information

Grid cells: 0

● Create Grid

Grid area

	X	Y
Lower Left	-76.91	39.19
Upper Right	-76.32	39.72

Cell specification

○ By cell spacing (in same units as data units): 1

● By number of columns: 100

Reference origin

Use a reference origin to convert X/Y data into angular data

● Use lower-left corner as origin

○ Use upper-right corner as origin

○ Use a different point as origin

X: 0

Y: 0

The coordinates are saved in the registry, but can be re-saved in any directory. With the Load screen open, click on 'Save to file' and then enter a directory and a file name. The default file extension is 'ref.

External Reference File

If an external file that stores the coordinates of each grid cell is to be used, select the name of the reference file. *CrimeStat* reads dbase 'dbf', ArcGIS point 'shp' and ASCII files. Select 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 reference file that is read into *CrimeStat* need not be a true grid (a matrix with *k* columns and *l* rows.) However, a reference file that is read in can only be output to *Surfer for Windows* since the other output formats – *ArcGIS*, *MapInfo*, , *ArcGIS Spatial Analyst*, and ASCII grid require the reference file to be a true grid.

Reference Origin

A reference origin can be defined for the directional mean routine. The reference origin can be assigned to:

1. Use the lower-left corner defined by the minimum X and Y values. This is the default
2. Use the upper-right corner defined by the maximum X and Y values
3. Use a different origin point. With the later, the user must define the origin

Measurement Parameters

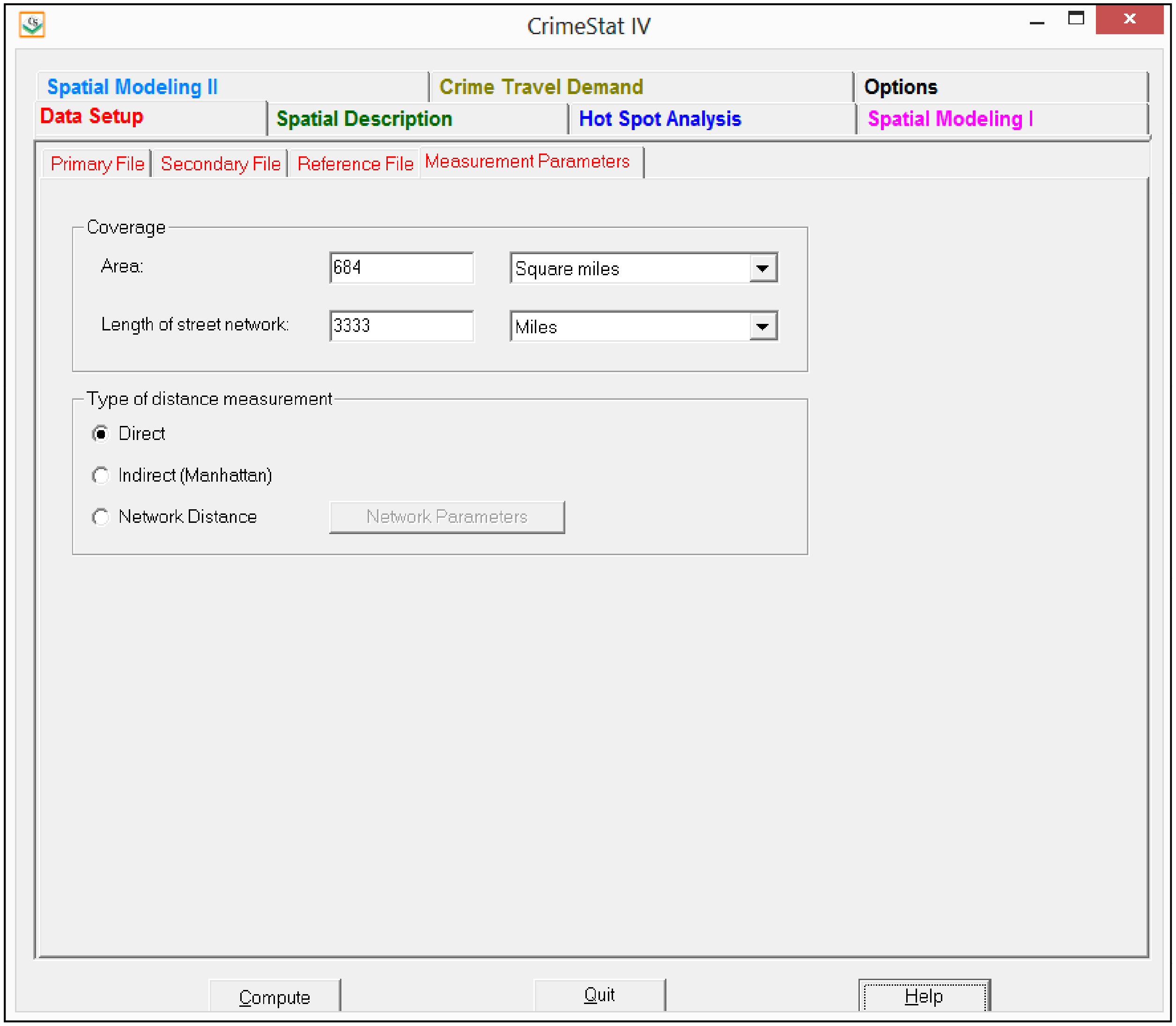
The measurement parameters page defines the measurement units of the coverage and the type of distance measurement to be used. There are three components that are defined:

Area

First, 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, risk-adjusted nearest neighbor hierarchical clustering, Stac, and K-means clustering routines.

Figure 2.4:

Measurement Parameters Setup



Coverage

Area: 684 Square miles

Length of street network: 3333 Miles

Type of distance measurement

● Direct

○ Indirect (Manhattan)

○ Network Distance

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

Second, 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, or 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

Third, define how distances are to be calculated. There are three choices:

1. Direct distance
2. Indirect (Manhattan) distance
3. Network distance

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. If the coordinates are projected, then the shortest indirect distance is the right angle of a right triangle on a two-dimensional plane

Network distance

If network distances are used, each distance is calculated as the shortest path between two points using the network. Alternatives to distance can be used including speed, travel time, or travel cost. Click on 'Network parameters' and identify a network file.

Type of network

Network files can *bi-directiona*l (e.g., a TIGER file) or *single directional* (e.g., a transportation modeling file). In a bi-directional file, travel can be in either direction. In a single directional file, travel is only in one direction. Specify the type of network to be used.

Network input file

The network file can either be a shape file (line, polyline, or polylineZ file) or another file, either a dBase IV 'dbf', ArcGIS 'shp' or ASCII file . The default is a shape file. If the file is a shape file, the routine will know the locations of the nodes. For a dBase IV or other file, the X and Y coordinate variables of the end nodes must be defined. These are called the "From" node and the "End" node. An optional weight variable is allowed for all types of file0073. The routine identifies nodes and segments and finds the shortest path. If there are one-way streets in a bi-directional file, the flag fields for the "From" and "To" nodes should be defined.

Network weight field

Normally, each segment in the network is not weighted. In this case, the routine calculates the shortest distance between two points using the distance of each segment. However, each segment can be weighted by travel time, speed or travel costs. If travel time is used for weighting the segment, the routine calculates the shortest time for any route between two points. If speed is used for weighting the segment, the routine converts this into travel time by dividing the distance by the speed. Finally, if travel cost is used for weighting the segment, the routine calculates the route with the smallest total travel cost. Specify the weighting field to be used and be sure to indicate the measurement units (distance, speed, travel time, or travel cost) at the bottom of the page. If there is no weighting field assigned, then the routine will calculate using distance.

From one-way flag and To one-way flag

One-way segments can be identified in a bi-directional file by a 'flag' field (it is not necessary in a single directional file). The 'flag' is a field for the end nodes of the segment with values of ‘0’ and ‘1’. A ‘0’ indicates that travel can pass through that node in either direction whereas a ‘1’ indicates that travel can only pass from the other node of the same segment (i.e., travel cannot occur from another segment that is connected to the node). The default assumption is for travel to be allowed through each node (i.e., there is a ‘0’ assumed for each node). For each one-way street, specify the flags for each end node. A ‘0’ allows travel from any connecting segments whereas a ‘1’ only allows travel from the other node of the same segment. Flag fields that are blank are assumed to allow travel to pass in either direction.

FromNode ID and ToNode ID

If the network is single directional, there are individual segments for each direction. Two-way streets have two segments, one for each direction. On the other hand, one-way streets have only one segment. The FromNode ID and the ToNode ID identify from which end of the segment travel should occur. If no FromNode ID and ToNode ID is defined, the routine will chose the first segment of a pair that it finds, whether travel is in the right or wrong direction. To identify correctly travel direction, define the FromNode and ToNode ID fields.

Network coordinate system

The type of coordinate system for the network is assumed to be the same as for the primary file.

Segment measurement unit

By default, the shortest path is in terms of distance. However, each segment can be weighted by travel time, travel speed, or travel cost.

1. For travel time, the units are minutes, hours, or unspecified cost units.
2. For speed, the units are miles per hour and kilometers per hour. In the case of speed as a weighting variable, it is automatically converted into travel time by dividing the distance of the segment by the speed, keeping units constant.
3. For travel cost, the units are undefined and the routine identifies routes by those with the smallest total cost.

II. Spatial Description

The spatial description section calculates spatial description, spatial autocorrelation, distance analysis, and hot spot statistics. The distance analysis and hot spot analysis statistics are on two separate tabs each.

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 six routines for describing the spatial distribution. An intensity variable and a weighting variable can be used for the first five routines, though it is not required. 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)

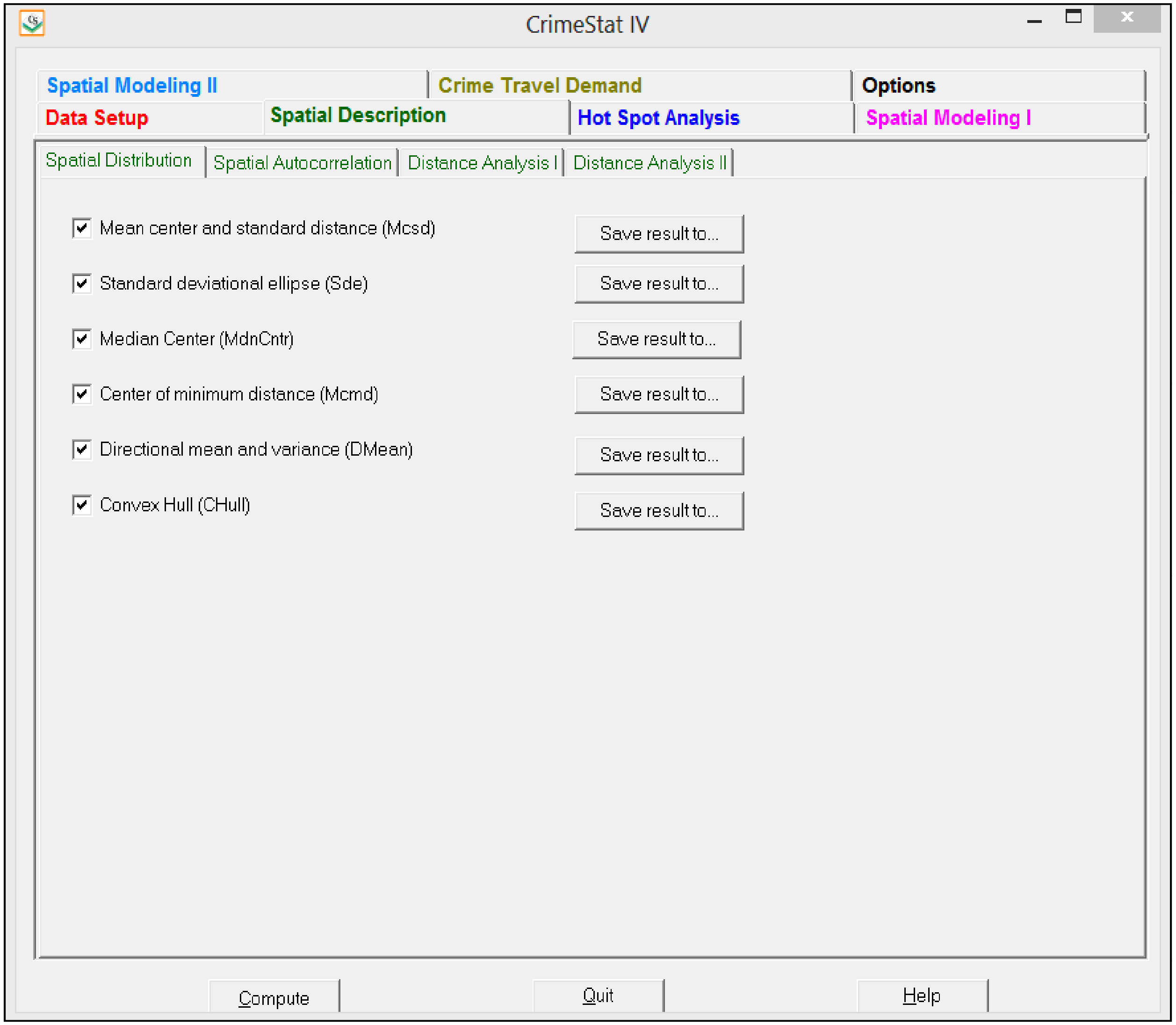
The mean center and standard distance define the arithmetic mean location and the degree of dispersion of the distribution. The Mcsd routine calculates 9 statistics:

1. The sample size
2. The minimum X and Y values
3. The maximum X and Y values
4. The X and Y coordinates of the mean center
5. The standard deviation of the X and Y coordinates
6. The X and Y coordinates of the geometric mean
7. The X and Y coordinates of the harmonic mean
8. The standard distance deviation, in meters, feet and miles. This is the standard deviation of the distance of each point from the mean center.
9. 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 geometric mean, the harmonic mean, the standard deviations of the X and Y coordinates, and the standard distance deviation can be output as graphical objects to ArcGIS ‘shp’, MapInfo ‘mif’, and *Google Earth* ‘kml’ (for spherical coordinates only) formats. A file name should be provided.

Figure 2.5:

Spatial Distribution Statistics



☑ Mean center and standard distance (Mcsd) — Save result to...

☑ Standard deviational ellipse (Sde) — Save result to...

☑ Median Center (MdnCntr) — Save result to...

☑ Center of minimum distance (Mcmd) — Save result to...

☑ Directional mean and variance (DMean) — Save result to...

☑ Convex Hull (CHull) — Save result to...

For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected. If the coordinate system is spherical (longitude/latitude), then the file can be saved as a *Google Earth* 'kml' output.

The mean center is output as a point (MC<*file name*>.) The geometric mean is output as a point (GM<*file name*>.) The harmonic mean is output as a point (HM<file name>.) The standard deviation of both the X and Y coordinates is output as a rectangle (XYD<file name>.) The standard distance deviation is output as a circle (SDD<*file name*>.)

Standard Deviational Ellipse (Sde)

The standard deviational ellipse defines both the dispersion and the direction (orientation) of that dispersion. The Sde routine calculates 9 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 X and Y axes in meters, feet and miles
5. The X and Y axes lengths in meters, feet and miles
6. The area of the ellipse defined by these axes in square meters, square feet and square miles
7. The standard deviation along the X and Y axes in meters, feet and miles for a 2X standard deviational ellipse
8. The X and Y axes lengths in meters, feet and miles for a 2X standard deviational ellipse
9. 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 ArcGIS 'shp', MapInfo 'mif', various ASCII formats, or *Google Earth* 'kml' (if the coordinates are spherical) files. A file name should be provided. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

The 1X standard deviational ellipse is output as an ellipse (SDE<*file 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<*file name*>.)

Median Center (MdnCntr)

The median center is the point at which the median of the X coordinates intersects the median of the Y coordinates. The MdnCntr routine outputs 3 statistics:

1. The sample size
2. The median value of the X coordinate
3. The median value of the Y coordinate

The tabular output can be printed and the median center can be output as a graphical object to ArcGIS 'shp', MapInfo 'mif', various ASCII formats, or *Google Earth* 'kml' (if the coordinates are spherical) files. A file name should be provided. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected. The median center is output as a point (MndCntr<file name>.)

Center of Minimum Distance (Mcmd)

The center of minimum distance defines the point at which the distance to all other points is at a minimum. Unfortunately, it is sometimes also called the 'median center', but not to be confused with median center that is the intersection of the median of X and the median of Y (see above). The Mcmd routine outputs 5 statistics:

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

The tabular output can be printed and the center of minimum distance can be output as a graphical object to ArcGIS 'shp', MapInfo 'mif', various ASCII formats, or *Google Earth* 'kml' (if the coordinates are spherical) files. A file name should be provided. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected. The center of minimum distance is output as a point (Mcmd<file name>).

Directional Mean and Variance (Dmean)

The angular mean and variance are properties of angular measurements. The angular mean is an angle defined as a bearing from true North: 0 degrees. The directional variance is a relative indicator varying from 0 (no variance) to 1 (maximal variance.) Both the angular mean and the directional variance can be calculated either through angular (directional) coordinates or through X and Y coordinates.

Output with directional coordinates

If the primary file cases are directional coordinates (bearings/angles from 0 to 360 degrees), the angular mean is calculated directly from the angles. An optional distance variable can be included. In this case, the directional mean routine will output five statistics:

1. The sample size
2. The unweighted mean angle
3. The weighted mean angle
4. The unweighted circular variance
5. The weighted circular variance

Output with X and Y coordinates

On the other hand, if the primary file incidents are defined in X and Y coordinates, the angles are defined relative to the reference origin (see Reference file) and the angular mean is converted into an equation. In this case, the directional mean routine will output nine statistics:

1. The sample size
2. The unweighted mean angle
3. The weighted mean angle
4. The unweighted circular variance
5. The weighted circular variance
6. The mean distance
7. The intersection of the mean angle and the mean distance (directional mean)
8. The X and Y coordinates for the triangulated mean
9. The X and Y coordinates for the weighted triangulated mean

The directional mean and triangulated mean can be saved as an *ArcGIS 'shp', MapInfo* 'mif', various ASCII formats, or *Google Earth* 'kml' (if the coordinates are spherical) files. A file name should be provided. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

The unweighted directional mean - the intersection of the mean angle and the mean distance is output with the prefix 'Dm' while the unweighted triangulated mean location is output with a 'Tm' prefix. The weighted triangulated mean is output with a 'TmWt' prefix. The tabular output can be printed.

Convex Hull (Chull)

The convex hull draws a polygon around the outer points of the distribution. It is useful for viewing the shape of the distribution. The routine outputs three statistics:

1. The sample size
2. The number of points in the convex hull
3. The X and Y coordinates for each of the points in the convex hull

The convex hull can be saved as an *ArcGIS 'shp', MapInfo 'mif',* , various ASCII formats, or *Google Earth* 'kml' (if the coordinates are spherical) files with a 'Chull' prefix. For *MapInfo* 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected. The convex hull is output as a graphical object with no attributes associated with it (i.e., only a polygon that defines the convex hull).

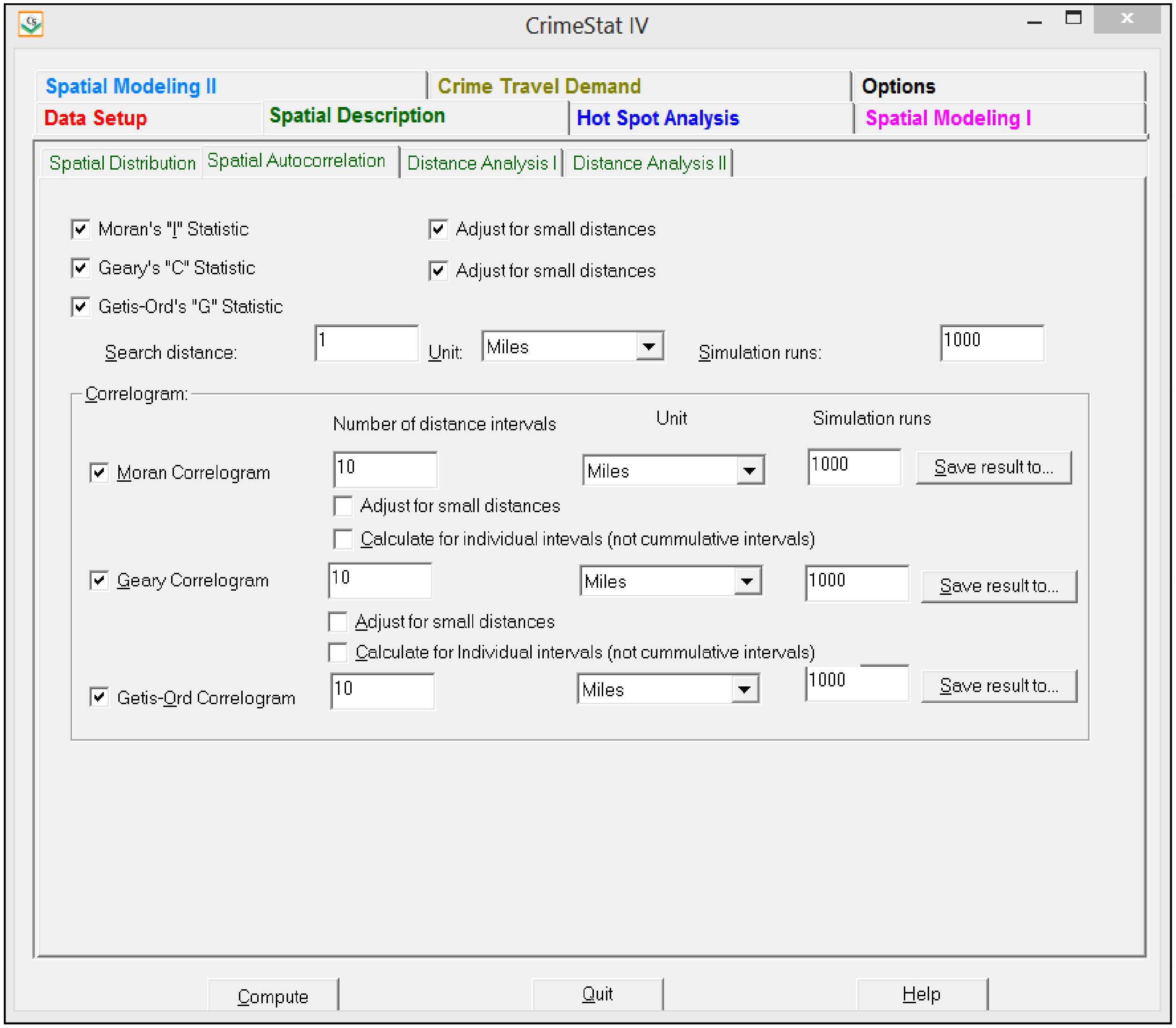
III. Spatial Autocorrelation

Spatial Autocorrelation Indices

Spatial autocorrelation indices identify whether point locations are spatially related, either clustered or dispersed. These indices would typically be applied to zonal data where an attribute value can be assigned to each zone. Six spatial autocorrelation indices are calculated. All **require** an intensity variable in the Primary File.

Figure 2.6:

Spatial Autocorrelation Statistics



☑ Moran's "I" Statistic ☑ Adjust for small distances

☑ Geary's "C" Statistic ☑ Adjust for small distances

☑ Getis-Ord's "G" Statistic

Search distance: 1 Unit: Miles Simulation runs: 1000

Correlogram:

	Number of distance intervals	Unit	Simulation runs
☑ Moran Correlogram	10	Miles	1000
	□ Adjust for small distances □ Calculate for individual intevals (not cummulative intervals)		
☑ Geary Correlogram	10	Miles	1000
	□ Adjust for small distances □ Calculate for Individual intervals (not cummulative intervals)		
☑ Getis-Ord Correlogram	10	Miles	1000

Moran's "I"(MoranI)

Moran's "I" statistic is the classic indicator of spatial autocorrelation. It is an index of co-variation between different point locations and is similar to a product moment correlation coefficient, typically varying from –1 to +1 (though these are not absolute limits). A positive value indicates that there is positive spatial autocorrelation, that in general zones are nearby other zones with similar values (either high or low) while a negative value indicates negative spatial autocorrelation, that in general zones are nearby other zones with different values (either high values next to zones with low values, or the opposite). The "I" value is calculated with the intensity variable specified on the Primary File page.

Adjust for small distances

If this box is checked, small distances are adjusted so that the maximum weighting is 1. This ensures that "I" won't become excessively large for points that are very close together. The default value is no adjustment.

Moran's "I" Output

The Moran's "I" routine calculates 6 statistics:

1. The sample size
2. Moran's "I"
3. The spatially random (expected) "I"
4. The standard error 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.

Geary's "C" (GearyC)

Geary's "C" statistic is an alternative indicator of spatial autocorrelation. It is an index of paired comparisons between different point locations and typically varies from 0 (similar values) to 2 (dissimilar values.) Theoretically, a value of +1 indicates spatial independence, that the values of one zone are unrelated to the values of nearby zones. Values less than +1 indicate positive spatial autocorrelation (zones have values similar to their neighbors) while values greater than +1 indicate negative spatial autocorrelation (zones have values different to their neighbors). The “C” value is calculated with the intensity variable specified on the Primary File page

Adjust for small distances

If this box is checked, small distances are adjusted so that the maximum weighting is 1. This ensures that “C” won't become excessively large or excessively small for points that are close together. The default value is no adjustment.

Geary “C” Output

The Geary’s “C” routine calculates 8 statistics:

1. The sample size
2. Geary's "C"'
3. Adjusted “C” (1-“C”)
4. The spatial random (expected) "C"
5. The standard error of "C"
6. A significance test of "C" under the assumption of normality (Z-test)
7. The one-tail probability level
8. The two-tail probability level

Values of “C” that are less than the expected “C” indicate clustering while values of “C” that are 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.

The adjusted “C” converts the statistic so that it varies between +1 and -1 and is similar to a Moran’s “I”. A positive value of the adjusted “C” indicates positive spatial autocorrelation while a negative value indicates negative spatial autocorrelation.

Getis-Ord General G (Getis-OrdG)

The Getis-Ord “G” statistic is an index of spatial autocorrelation for values of a variable that fall within a specified distance of each other (search distance). When compared to an expected value of G under the assumption of no spatial association, the statistic has the advantage over other global spatial autocorrelation measures (Moran, Geary) in that it can distinguish between ‘hot spots’ and ‘cold spots’. The “G” value is calculated with the intensity variable specified on the Primary File page and with respect to a specified search distance (user defined).

By itself, the G statistic is not very meaningful. The “G” value varies from 0 to 1 since it indicates the interaction of pairs of zones that are within the search distance relative to the interaction of all pairs of zones. As the search distance increases, this statistic will automatically approach 1.0. Consequently, G is compared to an expected value of G under the assumption of no significant spatial association.

Further, under the assumption that G is normally distributed, a Z-test can be constructed that tests for the significance of the actual G. A positive Z-value indicates spatial clustering of high values more than what would be expected under chance (hot spots) while a negative Z-value indicates spatial clustering of low values more than what would be expected under chance (cold spots). A “G” value around 0 typically indicates either no spatial autocorrelation at all or that the number of hot spots more or less balances the number of cold spots. The statistic requires an intensity variable in the primary file.

Search distance

The user must specify a search distance for the test and indicate the distance units (miles, nautical miles, feet, kilometers, or meters).

Getis-Ord “G” Output

The Getis-Ord “G” routine calculates 8 statistics:

1. The sample size
2. Getis-Ord “G”
3. The spatially random (expected) "G"
4. The difference between “G” and the expected “G”
5. The standard error of "G"
6. A Z-test of "G" under the assumption of normality (Z-test)
7. The one-tail probability level
8. The two-tail probability level

Simulation of confidence intervals

Since the Getis-Ord “G” statistic may not be normally distributed, the significance test is frequently inaccurate. Instead, a permutation type Monte Carlo simulation can be run to estimate approximate confidence intervals around the “G” value. Specify the number of simulations to be run (e.g., 100, 1000, 10000). In addition to the above statistics, a simulation includes the following statistics:

9. The minimum “G” value
10. The maximum “G” value
11. The 0.5 percentile of “G”
12. The 2.5 percentile of “G”
13. The 5 percentile of “G”
14. The 10 percentile of “G”
15. The 90 percentile of “G”
16. The 95 percentile of “G”
17. The 97.5 percentile of “G”
18. The 99.5 percentile of “G”

The four pairs of percentiles (10 and 90; 5 and 95; 2.5 and 97.5; 0.5 and 99.5) create approximate 80%, 90%, 95% and 99% confidence intervals respectively. The tabular results can be printed or saved to a text file.

Moran Correlogram

The Moran Correlogram calculates the Moran’s “I” index for different distance intervals/bins (not adjusted for small distances). The “I” value typically varies between -1 and +1 though these are not absolute limits. An “I” value of 0 indicates no spatial autocorrelation. An “I” value greater than 0 indicates positive spatial autocorrelation (zones have values similar to their neighbors) while an “I” value less than 0 indicates negative spatial autocorrelation (zones have values different from their neighbors).

The Moran Correlogram calculates these “I” values as a function of distance. The user can select any number of distance intervals. The default is 10 distance intervals. The “I” value for each distance interval is calculated with the intensity variable specified on the Primary File page.

Adjust for small distances

If the item is checked, small distances are adjusted so that the maximum weighting is 1. This ensures that the “I” values for individual distances won't become excessively large or excessively small for points that are close together. The default value is no adjustment.

Calculate for individual intervals

By default, the Moran Correlogram routine calculates the “I” values for the cumulative distance from 0 to the end of the interval. If the user checks the box to ‘Calculate for individual intervals’, then the “I” values for only those pairs of points that fall within the interval are calculated. This can be useful for checking the spatial autocorrelation for a specific interval or checking whether some distances don't have sufficient numbers of points (in which case the "I" value will be unreliable).

Simulation of confidence intervals

Since the Moran "I" statistic may not be normally distributed, the significance test is frequently inaccurate. Instead, a permutation type Monte Carlo simulation can be run to estimate approximate confidence intervals around the "I" values for each distance interval. Specify the number of simulations to be run (e.g., 100, 1000, 10000).

Moran Correlogram Output

The output includes:

1. The sample size
2. The maximum distance
3. The bin (interval) number
4. The midpoint of the distance bin
5. The "I" value for the distance bin

and if a simulation is run:

6. The minimum "I" value for the distance bin
7. The maximum "I" value for the distance bin
8. The 0.5 percentile of "I" for the distance bin
9. The 2.5 percentile of "I" for the distance bin
10. The 97.5 percentile of "I" for the distance bin
11. The 99.5 percentile of "I" for the distance bin

The two pairs of percentiles (2.5 and 97.5; 0.5 and 99.5) create approximate 95% and 99% confidence interval of "I" for each distance bin. The minimum and maximum "I" values create an *envelope*. However, unless a large number of simulations are run, the actual "I" value for any bin may fall outside the envelope. The tabular results can be printed, saved to a text file or saved as a 'dbf' file (MoranCorr<file name> with the file name being provided by the user.

Graphing the "I" values by distance

A graph is produced that shows the "I" value on the Y-axis by the distance bin on the X-axis. Click on the "Graph" button. If a simulation is run, the 2.5 and 97.5 percentiles of the simulated "I" values are also shown on the graph. The graph displays the reduction in spatial autocorrelation with distance. The graph is useful for selecting the type of kernel in the Single- and Dual-kernel interpolation routines when the primary variable is weighted. For a presentation quality graph, however, the output file should be brought into Excel or another graphics program in order to display the change in "I" values and label the axes properly.

Geary Correlogram

The Geary Correlogram calculates the Geary "C" index for different distance intervals/bins (not adjusted for small distances). The "C" value typically varies between 0 and 2 though these are not absolute limits. A "C" value of 1 indicates no spatial autocorrelation. A value of "C" less than 1 indicates positive spatial autocorrelation (zones have values similar to their neighbors) while a value of "C" greater than 1 indicates negative spatial autocorrelation (zones have values different from their neighbors). The user can select any number of distance intervals. The default is 10 distance intervals. The "C" value for each distance interval is calculated with the intensity variable specified on the Primary File page.

Adjust for small distances

If the item is checked, small distances are adjusted so that the maximum weighting is 1 This ensures that the "C" values for individual distances won't become excessively large or excessively small for points that are close together. The default value is no adjustment.

Calculate for individual intervals

By default, the Geary Correlogram routine calculates the "C" values for the cumulative distance from 0 to the end of the interval. If the user checks the box to 'Calculate for individual intervals', then the "C" values for only those pairs of points that fall within the interval are calculated. This can be useful for checking whether points separated by particular distances are clustered or whether there are unreliable "C" values for particular distance intervals.

Simulation of confidence intervals

Since the Geary "C" statistic may not be normally distributed, the significance test is frequently inaccurate. Instead, a permutation type Monte Carlo simulation can be run to estimate approximate confidence intervals around the “C” values for each distance interval. Specify the number of simulations to be run (e.g., 100, 1000, 10000).

Geary Correlogram Output

The output includes:

1. The sample size
2. The maximum distance
3. The bin (interval) number
4. The midpoint of the distance bin
5. The “C” value for the distance bin
6. The Adjusted “C” value for the distance bin

and if a simulation is run:

7. The minimum “C” value for the distance bin
8. The maximum “C” value for the distance bin
9. The 0.5 percentile of “C” for the distance bin
10. The 2.5 percentile of “C” for the distance bin
11. The 97.5 percentile of “C” for the distance bin
12. The 99.5 percentile of “C” for the distance bin.

The two pairs of percentiles (2.5 and 97.5; 0.5 and 99.5) create an approximate 95% and 99% confidence interval. The minimum and maximum “C” values create an *envelope*. However, unless a large number of simulations are run, the actual “C” value for any bin may fall outside the envelope. The tabular results can be printed, saved to a text file or saved as a ‘dbf’ file (GearyCorr<file name> with the file name being provided by the user.

Graphing the “C” values by distance

A graph can be shown with the “C” value on the Y-axis by the distance bin on the X-axis. Click on the “Graph” button. If a simulation is run, the 2.5 and 97.5 percentiles of the simulated “C” values are also shown on the graph. The graph displays the reduction in spatial autocorrelation with distance. The graph is useful for selecting the type of kernel in the single- and dual-kernel interpolation routines when the primary variable is weighted. For a presentation quality graph, however, the output file should be brought into Excel or another graphics program in order to display the change in “C” values and label the axes properly.

Getis-Ord Correlogram

The Getis-Ord Correlogram calculates the Getis-Ord “G” index for different distance intervals/bins. The user can select any number of distance intervals. The default is 10 distance intervals. The statistic requires an intensity variable in the primary file.

Simulation of confidence intervals

Since the Getis-Ord “G” statistic may not be normally distributed, the significance test is frequently inaccurate. Instead, a permutation type Monte Carlo simulation can be run to estimate approximate confidence intervals around the “G” values for each distance interval. Specify the number of simulations to be run (e.g., 100, 1000, 10000).

Getis-Ord Correlogram Output

The output includes:

1. The sample size
2. The maximum distance
3. The bin (interval) number
4. The midpoint of the distance bin
5. The “G” value for the distance bin
6. The expected “G” value for the distance bin

and if a simulation is run:

7. The minimum “G” value for the distance bin
8. The maximum “G” value for the distance bin
9. The 0.5 percentile of “G” for the distance bin
10. The 2.5 percentile of “G” for the distance bin
11. The 97.5 percentile of “G” for the distance bin
12. The 99.5 percentile of “G” for the distance bin

The two pairs of percentiles (2.5 and 97.5; 0.5 and 99.5) create an approximate 95% and 99% confidence interval. The minimum and maximum “G” values create an *envelope*. However, unless a large number of simulations are run, the actual “G” value for any bin may fall outside the envelope. The tabular results can be printed, saved to a text file or saved as a ‘dbf’ file (Getis-OrdCorr<file name> with the file name being provided by the user.

Graphing the "G" values by distance

A graph can be shown that shows the "G" and Expected "G" values on the Y-axis by the distance bin on the X-axis. Click on the "Graph" button. If a simulation is run, the 2.5 and 97.5 percentiles of the simulated "G" values are also shown on the graph along with the "G"; the Expected "G" is not shown in this case. The graph displays the reduction in spatial autocorrelation with distance. Note that the "G" and expected "G" approach 1.0 as the search distance increases, that is as the pairs included within the search distance approximate the number of pairs in the entire data set. The graph is useful for selecting the type of kernel in the single- and dual-kernel interpolation routines when the primary variable is weighted. For a presentation quality graph, however, the output file should be brought into Excel or another graphics program in order to display the change in "G" values and label the axes properly.

Distance Analysis I

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. The distance routines are divided into two pages: Distance Analysis I and Distance Analysis II. On the first page, there are four routines for describing properties of the distances.

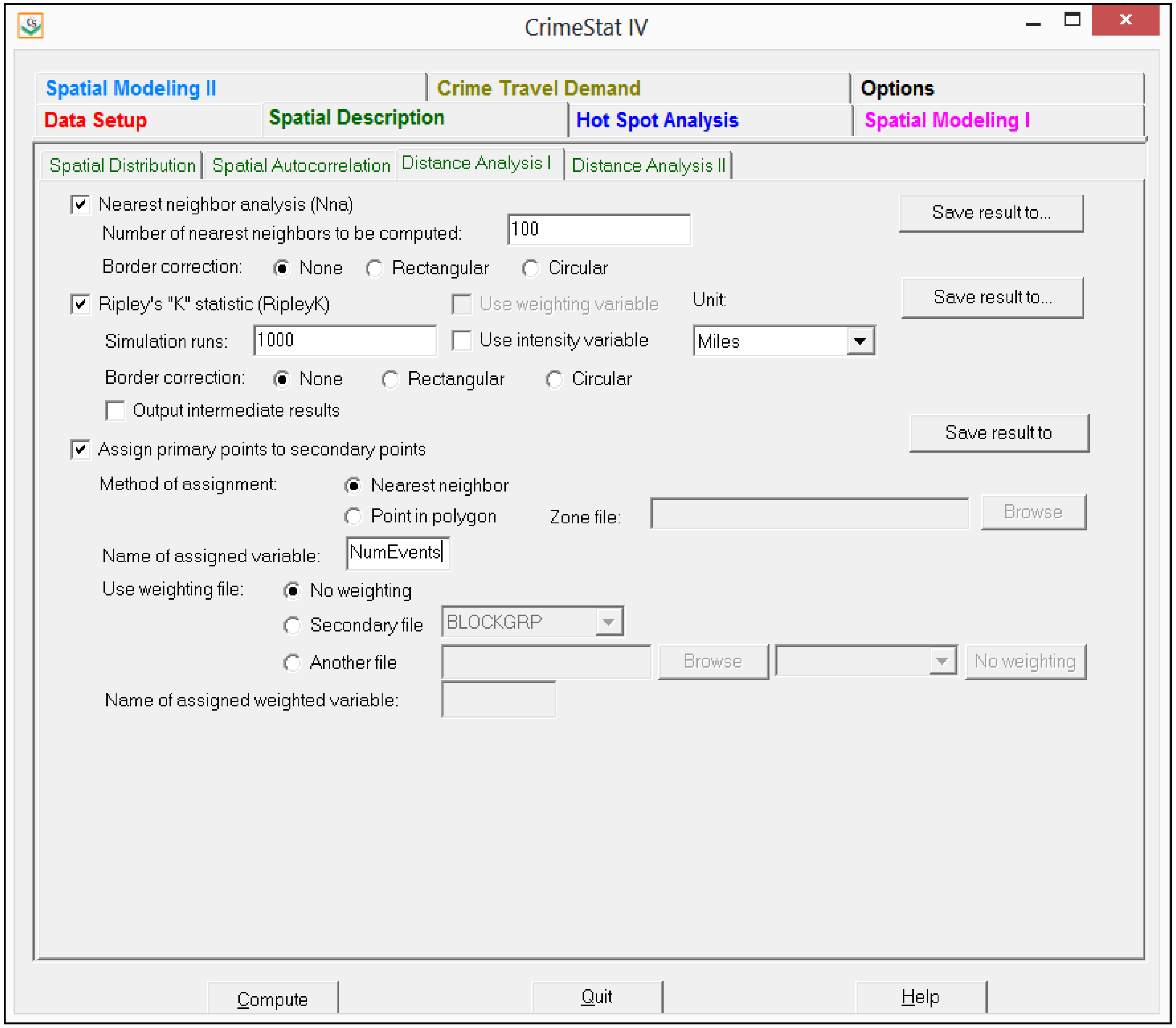
Nearest Neighbor Analysis (Nna)

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)

Figure 2.7:

Distance Analysis I Statistics



☑ Nearest neighbor analysis (Nna)

Number of nearest neighbors to be computed: 100

Border correction: ● None ○ Rectangular ○ Circular

☑ Ripley's "K" statistic (RipleyK) □ Use weighting variable Unit:

Simulation runs: 1000 □ Use intensity variable Miles

Border correction: ● None ○ Rectangular ○ Circular

□ Output intermediate results

☑ Assign primary points to secondary points

Method of assignment: ● Nearest neighbor

○ Point in polygon Zone file:

Name of assigned variable: NumEvents

Use weighting file: ● No weighting

○ Secondary file BLOCKGRP

○ Another file

Name of assigned weighted variable:

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)
11. The p-values associated with a one tail and two tail significance test

The tabular results can be printed, saved to a text file or saved as a ‘dbf’ file. For the latter, specify a file name in the “Save result to” in the dialogue box.

K-*order nearest neighbors*

The K-nearest neighbor index compares the average distance to the Kth nearest other point with a spatially random expected distance. The user can specify the number of K-nearest neighbors to be calculated, if more than one is 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. For the latter, specify a file name in the “Save result to” dialogue box.

Edge correction of nearest neighbors

The nearest neighbor analysis does not adjust for underestimation for incidents near the boundary of the study area. It is possible that there are nearest neighbors outside the boundary that are closer than the measured nearest neighbor. The nearest neighbor analysis has three edge correction options:

1. No adjustment – this is the default;
2. Adjustment that assumes the study area is a rectangle; and
3. Adjustment that assumes the study area is a circle. The rectangular and circular edge corrections adjust the nearest neighbor distances of points near the border. If a point is closer to the border (of either a rectangle or a circle) than to the measured nearest neighbor distance, then the distance to the border is taken as the adjusted nearest neighbor distance.

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 outputs 10 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
10. The p-values associated with a one tail and two tail significance test

Linear K-order nearest neighbors

NNA can calculate K-nearest linear neighbors and compare this distance the average linear distance to the Kth nearest other point with a spatially random expected distance. The user can specify 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

Edge correction of linear nearest neighbors

The nearest neighbor analysis does not adjust for underestimation for incidents near the boundary of the study area. It is possible that there are nearest neighbors outside the boundary that are closer than the measured nearest neighbor. The nearest neighbor analysis has three edge correction options: 1) No adjustment – this is the default; 2) Adjustment that assumes the study area is a rectangle; and 3) Adjustment that assumes the study area is a circle. The rectangular and circular edge corrections adjust the nearest neighbor distances of points near the border. If a point is closer to the border (of either a rectangle or a circle) than to the measured nearest neighbor distance, then the distance to the border is taken as the adjusted nearest neighbor distance.

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). 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)

The tabular results can be printed, saved to a text file, or saved as a ‘dbf’ file. For the latter, specify a file name in the “Save result to” in the dialogue box.

Simulating confidence intervals

A Monte Carlo simulation can be run to evaluate an approximate confidence interval around the L statistic. The user specifies the number of simulation runs and the L statistic is calculated for randomly assigned data. The random output is sorted and percentiles are calculated. Values of L that are greater than any particular percentile indicate more concentration while values of L less than any particular percentile indicate more dispersion. L is calculated for each of 100 distance intervals (bins.) Eight percentiles are identified for these statistics:

1. The minimum for the spatially random L value
2. The maximum for the spatially random L value
3. The 0.5 percentile for the spatially random L value
4. The 2.5 percentile for the spatially random L value

5. The 95 percentile for the spatially random L value
6. The 97.5 percentile for the spatially random L value
7. The 99 percentile for the spatially random L value
8. The 99.5 percentile for the spatially random L value

Confidence intervals can be estimated from these percentiles. The two most commonly used ones are the 95% (defined by the 2.5 and 97.5 percentiles) and the 99% (defined by the 0.5 and 99.5 percentiles). The simulated data that is used can be viewed by checking the 'Dump simulation data' box on the Options tab.

Edge correction of Ripley's K statistic

The default setting for the Ripley's "K" statistic does not adjust for underestimation for incidents near the boundary of the study area. However, it is possible that there are points outside the study area boundary that are closer than the search radius of the circle used to enumerate the "K" statistic. The Ripley's "K" statistic has three edge correction options: 1) No adjustment – this is the default; 2) Adjustment that assumes the study area is a rectangle; and 3) Adjustment that assumes the study area is a circle. The rectangular and circular edge corrections adjust the Ripley's "K" statistic for points near the border. If the distance of a point to the border (of either a rectangle or a circle) is smaller than to the radius of the circle used to enumerate the "K" statistics, then the point is weighted inversely proportional to the area of the search radius that is within the border.

Output Intermediate Results

There is a box labeled “Output intermediate results”. If checked, a separate dbf file will be output that lists the intermediate calculations. The file will be called “RipleyTempOutput.dbf”. There are five output fields:

1. The point number (POINT), starting at 0 (for the first point) and proceeding to N–1 (for the Nth point)
2. The search radius in meters (SEARCHRADI)
3. The count of the number of *other* points that are within the search radius (COUNT)
4. The weight assigned (WEIGHT)
5. The count times the weight (CTIMESW)

Assign Primary Points to Secondary Points

This routine will assign each primary point to a secondary point and then will sum by the number of primary points assigned to each secondary point. It is useful for adding up the number of primary points that are close to each secondary point. For example, in the crime travel demand module, this routine can assign incidents to zones as the module uses zonal totals. The result is a count of primary points associated with each secondary point. It is also possible to sum different variables sequentially. For example, in the crime travel demand module, both the number of crimes originating in each zone and the number of crimes occurring in each zone are needed. This can be accomplished in two runs. First, sum the incidents defined by the origin coordinates to each zone (secondary file). Second, sum the incidents defined by the destination coordinates to each zone (also secondary file). The result would be two columns, one showing the number of origins in each secondary file zone and the second showing the number of destinations in each secondary file zone.

There are two methods for assigning the primary points to the secondary.

Nearest neighbor assignment

This routine assigns each primary point to the secondary point to which it is closest. If there are two or more secondary points that are exactly equal, the assignment goes to the first one on the list.

Point-in-polygon assignment

This routine assigns each primary point to the secondary point for which it falls within its polygon (zone). A zone (polygon) shape file must be provided and the routine checks which secondary zone each primary point falls within.

Zone file for point-in-polygon assignment

If point-in-polygon assignment is used, a zonal file must be provided. This is a polygon file that defines the zones to which the primary points are assigned. The zone file should be the same as the secondary file (see Secondary file). For each point in the primary file, the routine identifies which polygon (zone) it belongs to and then sums the number of points per polygon.

Name of assigned variable

Whether nearest neighbor or point-in-polygon assignment is used, specify the name of the summed variable. The default name is FREQ.

Use weighting file

The primary file records can be weighted by another file. This would be useful for correcting the totals from the primary file. For example, if the primary file were robbery incidents from an arrest record, the sum of this variable (i.e. the total number of robberies) may produce a biased distribution over the secondary file zones because the primary file was not a random sample of all incidents (e.g., if it came from an arrest record where the distribution of robbery arrests is not the same as the distribution of all robbery incidents).

The secondary file or another file can be used to adjust the summed total. The weighting variable should have a field that identifies the ratio of the true to the measured count for each zone. A value of 1 indicates that the summed value for a zone is equal to the true value; hence no adjustment is needed. A value greater than 1 indicates that the summed value needs to be adjusted upward to equal the true value. A value less than 1 indicates that the summed value needs to be adjusted downward to equal the true value.

If another file is to be used for weighting, indicate whether it is the secondary file or, if another file, the name of the other file.

Name of assigned weighted variable

For a weighted sum, specify the name of the variable. The default will be ADJFREQ.

Save result to

For both routines, the output is a ‘dbf’ file. Define the file name. Note: be careful about using the same name as the secondary file as the saved file will have the new variable. It is best to give it a new name.

A new variable will be added to this file that gives the number of primary points in each secondary file zone and, if weighting is used, a secondary variable will be added which has the adjusted frequency.

Output intermediate results

If the label “Output intermediate results” is checked, a separate dbf file will be output that lists the intermediate calculations. The file will be called “RipleyTempOutput.dbf”. There are five output fields:

1. The point number (POINT), starting at 0 (for the first point) and proceeding to N–1 (for the Nth point)
2. The search radius in meters (SEARCHRADI)
3. The count of the number of *other* points that are within the search radius (COUNT)
4. The weight assigned (WEIGHT)
5. The count times the weight (CTIMESW)

Distance Analysis II

On the second Distance Analysis page, there are four routines that calculate distance matrices:

Distance Matrices

1. From each primary point to every other primary point
2. From each primary point to each secondary point
3. From each primary point to the centroid of each reference file grid cell. This requires a reference file to be defined or used.
4. From each secondary point to the centroid of each reference file grid cell. This requires a reference file to be defined or used

CrimeStat can calculate distances between points for a single file or distances between points for two different files. These matrices are useful for examining the frequency of different distances or for providing distances for another program. Because the output files are usually very large, only text output is allowed. This can then be read into a database or large statistical program for processing. Keep in mind that there may be storage problems for large matrices.

Within File Point-to-Point (Matrix)

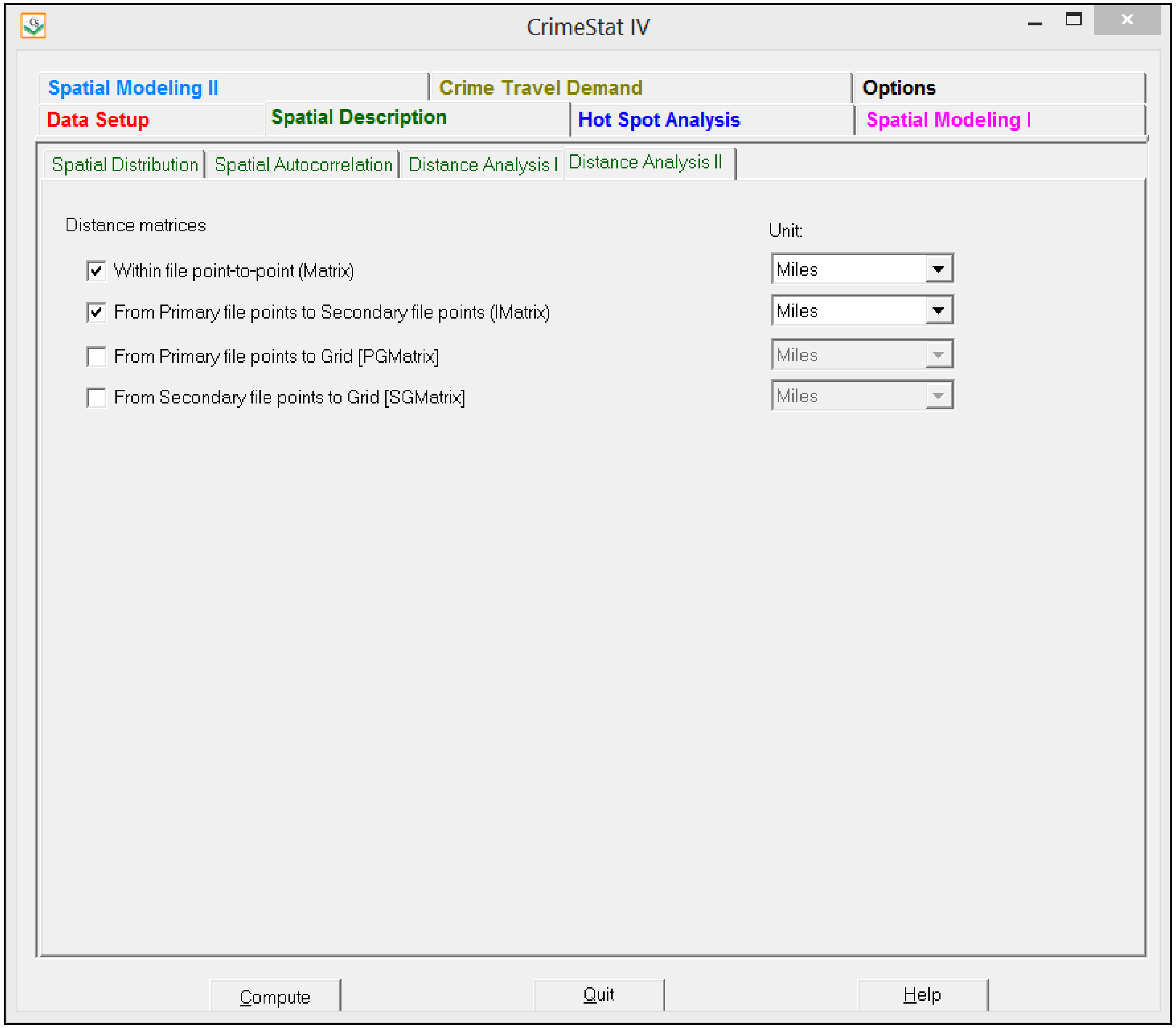
This routine outputs the distance between each point in the primary file to every other point in a specified distance unit (miles, nautical miles, feet, kilometers, or meters). The Matrix output can be saved to a text file.

From Primary File Points to 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.

Figure 2.8:

Distance Analysis II Statistics



Distance matrices

Unit:

☑ Within file point-to-point (Matrix) — Miles

☑ From Primary file points to Secondary file points (Matrix) — Miles

☐ From Primary file points to Grid [PGMatrix] — Miles

☐ From Secondary file points to Grid [SGMatrix] — Miles

From Primary File Points to Grid (PGMatrix)

This routine outputs the distance between each point in the primary file to the centroid of each cell in the reference grid. A reference has to be defined or provided on the Reference file page. Again, the distance units must be specified (miles, nautical miles, feet, kilometers, or meters). The output can be saved to a text file.

From Secondary File Points to Grid (SGMatrix)

This routine outputs the distance between each point in the secondary file to the centroid of each cell in the reference grid. A reference has to be defined or provided on the Reference file page. Again, the distance units must be specified (miles, nautical miles, feet, kilometers, or meters). The output can be saved to a text file.

III. 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 a number of different hot spot analysis routines in *CrimeStat*. They are organized on three program tabs: Hot Spot analysis I, Hot Spot analysis II, and Hot Spot Analysis of Zones.

Hot Spot Analysis I

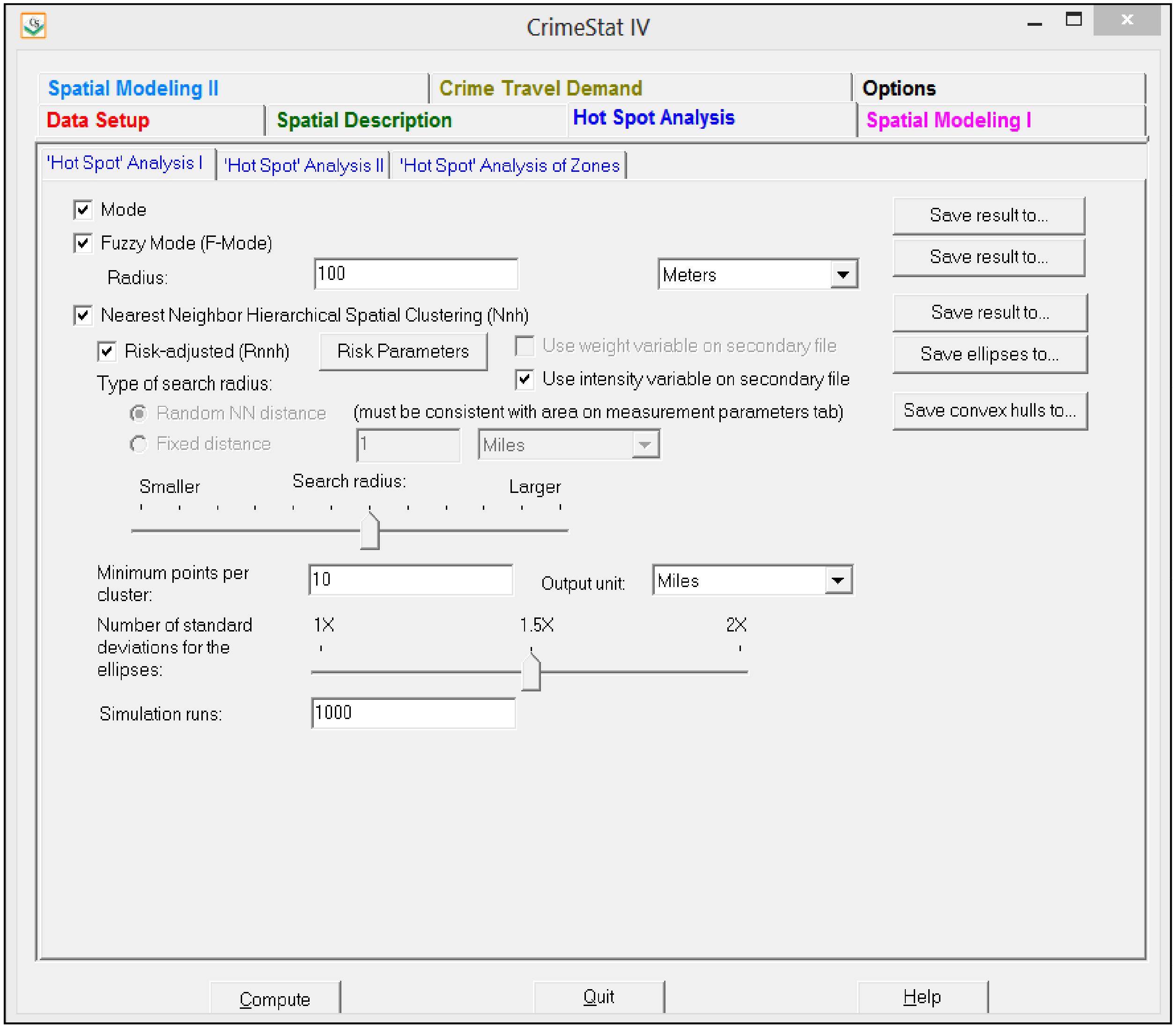
Hot spot (or cluster) analysis identifies groups of incidents that are clustered together. It is a method of second-order analysis that identifies the cluster membership of points. On the Hot Spot Analysis I page, there are four statistics that can be used to identify hot spots: 1) the mode; 2) the fuzzy mode; 3) Nearest neighbor hierarchical spatial clustering; and 4) Risk-adjusted nearest neighbor hierarchical spatial clustering.

Mode

The mode calculates the frequency of incidents for each unique location, defined by an X and Y coordinate. It will output a list of all unique locations and their X and Y coordinates and the number of incidents occurring at each, ranked in decreasing order from most frequent to least frequent. It will also list their rank order from 1 to the last unique location. The data can be output to a 'dbf' file. For the latter, specify a file name in the “Save result to” in the dialogue box.

Figure 2.9:

Hot Spot Analysis I



☑ Mode

☑ Fuzzy Mode (F-Mode)

Radius: 100 Meters

☑ Nearest Neighbor Hierarchical Spatial Clustering (Nnh)

☑ Risk-adjusted (Rnnh)

□ Use weight variable on secondary file

☑ Use intensity variable on secondary file

Type of search radius:

● Random NN distance

○ Fixed distance 1 Miles

(must be consistent with area on measurement parameters tab)

Search radius: Smaller Larger

Minimum points per cluster: 10

Output unit: Miles

Number of standard deviations for the ellipses: 1X 1.5X 2X

Simulation runs: 1000

Fuzzy Mode

The fuzzy mode calculates the frequency of incidents for each unique location within a small, user-specified distance. The user must specify the search radius and the units for the radius (miles, nautical miles, feet, kilometers, or meters). Distances should be small (e.g., less than 0.25 miles). The routine will identify each unique location, defined by its X and Y coordinates, and will calculate the number of incidents that fall within the search radius. It will output a list of all unique locations and their X and Y coordinates and the number of incidents occurring at each, ranked in decreasing order from most frequent to least frequent. It will also list their rank order from 1 to the last unique location. The data can be output to a 'dbf' file.

Nearest Neighbor Hierarchical Spatial Clustering (Nnh)

The nearest neighbor hierarchical spatial clustering routine is a constant-distance clustering routine that groups points together on the basis of spatial proximity. The user defines a threshold distance and the minimum number of points that are required for each cluster, and an output size for displaying the clusters with ellipses. The routine identifies first-order clusters, representing groups of points that are closer together than the threshold distance and in which there is at least the minimum number of points specified by the user. 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 distances between their centers are closer than the new threshold distance.

Threshold distance

The threshold distance is the search radius around a pair of points. For each pair of points, the routine determines whether they are closer together than the search radius. There are two ways to determine a threshold distance:

Random nearest neighbor distance

First, the search distance is chosen by the random nearest neighbor distance. The default value is 0.1 (i.e., fewer than 10% of the pairs could be expected to be as close or closer 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 threshold distance, the smaller the significance level that is selected and the fewer pairs will be selected. On the other hand, choosing a larger threshold distance (and, consequently, a higher significance level) will usually lead to more pairs being selected. However, the more pairs that are selected, the greater the likelihood that clusters could be chance groupings.

The slide bar is used to adjust the significance level. Move the slide bar to the left to choose a smaller threshold distance and to the right to choose a larger threshold distance.

*Fixed distanc*e

Second, a fixed distance can be selected. The default is 1 mile. In this case, the search radius uses the fixed distance and the slide bar is inoperative.

Minimum number of points

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 slide bar. These are the number of standard deviations defined by the ellipse, from one standard deviation (the default value) to three standard deviations. Typically, one standard deviation will cover about 65% of the cases whereas three standard deviations will cover more than 99% of the cases.

The tabular results can be printed, saved to a text file, or output as a 'dbf' file. The graphical results can be output as either ellipses or as convex hulls (or both) to *ArcGIS* 'shp', *MapInfo* 'mif', various ASCII formats, or *Google Earth* 'kml' (if the coordinates are spherical) files. Separate file names must be selected for the ellipse output and for the convex hull output. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Tabular output

The routine outputs six results for each cluster that is calculated:

1. The hierarchical order and the cluster number
2. The mean center of the cluster (Mean X and Mean Y)
3. The standard deviational ellipse of the cluster (the rotation and the lengths of the X and Y axes)
4. The number of points in the cluster
5. The area of the cluster

6. The density of the cluster (points divided by area)

Ellipse output

The results can be output graphically as an ellipse to *ArcGIS* 'shp', *MapInfo* 'mif', various ASCII formats, or *Google Earth* 'kml' (if the coordinates are spherical) files. A file name should be provided. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

First and higher-order ellipses will be output as separate objects. The prefix will be 'NNH1' for the first-order ellipses, 'NNH2' for the second-order ellipses, and 'NNH3' for the third-order ellipses. Higher-order ellipses will only index the number.

Output size for ellipses

The cluster output size can be adjusted by the lower slide bar. This specifies the number of ellipse standard deviations to be calculated for each cluster: one standard deviation (1X - the default value), one and a half standard deviations (1.5X), or two standard deviations (2X). The default value is one standard deviation. Typically, one standard deviation will cover more than half the cases whereas two standard deviations will cover more than 99% of the cases, though the exact percentage will depend on the distribution. Slide the bar to select the number of standard deviations for the ellipses. The output file is saved as Nnh<number><file name> with the file name being provided by the user. The number is the order of the clustering (i.e., 1, 2…).

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.

Convex hull cluster output

The clusters can also be output as convex hulls to *ArcGIS* 'shp', *MapInfo* 'mif', various ASCII formats, or *Google Earth* 'kml' (if the coordinates are spherical) files. Specify a file name. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

The name will be output with a ‘CNNH1’ prefix for the first-order clusters, a ‘CNNH2’ prefix for the second-order clusters, and a ‘CNNH3’ prefix for the third-order clusters. Higher-order clusters will index only the number.

Simulating confidence intervals

A Monte Carlo simulation can be run to estimate the approximate confidence intervals around first-order Nnh clusters; second- and higher-order clusters are not simulated since their structure depends on first-order clusters. The user specifies the number of simulation runs and the Nnh clustering is calculated for randomly assigned data. The random output is sorted and percentiles are calculated. The output includes the number of first-order clusters, the area, the number of points, and the density. Twelve percentiles are identified for these statistics:

1. The minimum for the spatially random Nnh simulations
2. The maximum for the spatially random Nnh simulations
3. The 0.5 percentile for the spatially random Nnh simulations
4. The 1 percentile for the spatially random Nnh simulations
5. The 2.5 percentile for the spatially random Nnh simulations
6. The 5 percentile for the spatially random Nnh simulations
7. The 10 percentile for the spatially random Nnh simulations
8. The 90 percentile for the spatially random Nnh simulations
9. The 95 percentile for the spatially random Nnh simulations
10. The 97.5 percentile for the spatially random Nnh simulations
11. The 99 percentile for the spatially random Nnh simulations
12. The 99.5 percentile for the spatially random Nnh simulations

Confidence intervals can be estimated from these percentiles. The two most commonly used ones are the 95% (defined by the 2.5 and 97.5 percentiles) and the 99% (defined by the 0.5 and 99.5 percentiles). The simulated data that is used can be viewed by checking the 'Dump simulation data' box on the Options tab.

Risk-Adjusted Nearest Neighbor Hierarchical Spatial Clustering (Rnnh)

The risk-adjusted nearest neighbor hierarchical spatial clustering routine groups points together on the basis of spatial proximity, but the grouping is adjusted according to the distribution of a baseline variable. The routine requires both a primary file (e.g., robberies) and a secondary file (e.g., population). For the secondary variable, if an intensity or weight variable is to be used, it should be specified.

The user selects a threshold probability for grouping a *pair* of points together by chance and the minimum number of points that are required for each cluster, and an output size for displaying the clusters with ellipses. In addition, a kernel density model for the secondary variable must be specified. The threshold distance is determined by the threshold probability and the grid cell density produced by the kernel density estimate of the secondary variable. Thus, in areas with high density of the secondary variable, the threshold distance is smaller than in areas with low density of the secondary variable.

The routine identifies first-order clusters, representing groups of points that are closer together than the threshold distance and in which there is at least the minimum number of points specified by the user. 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 the new threshold distance.

The tabular results can be printed, saved to a text file, or output as a 'dbf' file. The graphical results can be output as either ellipses or as convex hulls (or both) to *ArcGIS* 'shp', *MapInfo* 'mif', various ASCII formats, or *Google Earth* 'kml' (if the coordinates are spherical) files. Separate file names must be selected for the ellipse output and for the convex hull output. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Threshold distance

The threshold distance is the confidence interval around a random expected distance for a *pair* of points (called *credible interval*). However, unlike the Nnh routine where the threshold distance is constant throughout the study area, the threshold distance for the Rnnh routine is adjusted inversely proportional to the distribution of the secondary (baseline) variable. In areas with a high density of the secondary variable, the threshold distance will be small whereas in areas with a low density of the secondary variable, the threshold distance will be large. The default threshold probability is 0.1 (i.e., fewer than 10% of the pairs could be expected to be as close or closer 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 will be the threshold distance with, usually, fewer pairs being selected. On the other hand, choosing a higher significance level, the larger the threshold distance and, usually, the more pairs will be selected. However, the higher the significance level chosen, the greater the likelihood that clusters could be chance groupings.

Move the slide bar to the left to choose a smaller threshold distance and to the right to choose a larger threshold distance.

Risk parameters

A density estimate of the secondary variable must be calculated to adjust the threshold distance of the primary variable. This is done through kernel density estimation. The risk parameters tab defines this model. The secondary variable is automatically assumed to be the 'at risk' (baseline) variable. The user specifies a method of interpolation (normal, uniform, quartic, triangular, and negative exponential kernels) and the choice of bandwidth (fixed interval or adaptive interval). If an adaptive interval is used, the minimum sample size for the band width (search radius) must be specified. If a fixed interval is used, the size of the interval (radius) must be specified along with the measurement units (miles, nautical miles, feet, kilometers, or meters). Finally, the units of the output density must be specified (squared miles, squared nautical miles, squared feet, squared kilometers, squared meters).

The routine overlays a 50 x 50 grid on the study area and calculates a kernel density estimate of the secondary variable. The density is then re-scaled to equal the sample size of the primary variable. For each grid cell, a cell-specific threshold distance is calculated for grouping a pair of points together by chance. The threshold probability selected by the user is applied to this cell-specific threshold distance to produce a threshold distance that corresponds to the cell-specific confidence interval. Pairs of points that are closer than the cell-specific threshold distance are selected for first-order clustering.

Use of intensity variable

If an intensity variable has been used in the secondary file, the intensity box should be checked.

Minimum number of points

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 slide bar. These are the number of standard deviations defined by the ellipse, from one standard deviation (the default value) to three standard deviations. Typically, one standard deviation will cover about 65% of the cases whereas three standard deviations will cover more than 99% of the cases.

Tabular output

The routine outputs six results for each cluster that is calculated:

1. The hierarchical order and the cluster number
2. The mean center of the cluster (Mean X and Mean Y)
3. The standard deviational ellipse of the cluster (the rotation and the lengths of the X and Y axes)
4. The number of points in the cluster
5. The area of the cluster
6. The density of the cluster (points divided by area)

Ellipse output

The results can be output graphically as an ellipse to *ArcGIS* 'shp', *MapInfo* 'mif', various ASCII formats, or *Google Earth* 'kml' (if the coordinates are spherical) files. A file name should be provided. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

First- and higher-order ellipses will be output as separate objects. The prefix will be 'RNNH1' for the first-order ellipses, 'RNNH2' for the second-order ellipses, and 'RNNH3' for the third-order ellipses. Higher-order ellipses will only index the number.

Output size for ellipses

The cluster output size can be adjusted by the lower slide bar. This specifies the number of ellipse standard deviations to be calculated for each cluster: one standard deviation (1X - the default value), one and a half standard deviations (1.5X), or two standard deviations (2X). The default value is one standard deviation. Typically, one standard deviation will cover more than half the cases whereas two standard deviations will cover more than 99% of the cases, though the exact percentage will depend on the distribution. Slide the bar to select the number of standard deviations for the ellipses. The output file is saved as Rnnh<number><file name> with the file name being provided by the user. The number is the order of the clustering (i.e., 1, 2…).

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.

Convex hull cluster output

The clusters can also be output as convex hulls to *ArcGIS* 'shp', *MapInfo* 'mif', various ASCII formats, or *Google Earth* 'kml' (if the coordinates are spherical) files. Specify a file name. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

First- and higher-order clusters will be output as separate objects. The clusters will have a 'CRNNH1' prefix for the first-order clusters, a 'CRNNH2' prefix for the second-order clusters, and a 'CRNNH3' prefix for the third-order clusters. Higher-order clusters will index only the number.

Simulating confidence intervals

A Monte Carlo simulation can be run to estimate the approximate confidence intervals around first-order Rnnh clusters; second- and higher-order clusters are not simulated since their structure depends on first-order clusters. The user specifies the number of simulation runs and the Rnnh clustering is calculated for randomly assigned data. The random output is sorted and percentiles are calculated. The output includes the number of first-order clusters, the area, the number of points, and the density.

Twelve percentiles are identified for these statistics:

1. The minimum for the spatially random Rnnh simulations
2. The maximum for the spatially random Rnnh simulations
3. The 0.5 percentile for the spatially random Rnnh simulations
4. The 1 percentile for the spatially random Rnnh simulations
5. The 2.5 percentile for the spatially random Rnnh simulations
6. The 5 percentile for the spatially random Rnnh simulations
7. The 10 percentile for the spatially random Rnnh simulations
8. The 90 percentile for the spatially random Rnnh simulations
9. The 95 percentile for the spatially random Rnnh simulations
10. The 97.5 percentile for the spatially random Rnnh simulations
11. The 99 percentile for the spatially random Rnnh simulations

12. The 99.5 percentile for the spatially random Rnnh simulations

Confidence intervals can be estimated from these percentiles. The two most commonly used ones are the 95% (defined by the 2.5 and 97.5 percentiles) and the 99% (defined by the 0.5 and 99.5 percentiles). The simulated data that is used can be viewed by checking the 'Dump simulation data' box on the Options tab.

Hot Spot Analysis II

On the Hot Spot Analysis II page, there are two statistics that can be used to identify hot spots: 1) STAC; and 2) K-means clustering.

Spatial and Temporal Analysis of Crime (STAC)

The Spatial and Temporal Analysis of Crime (STAC) routine is a variable-distance clustering routine. It initially groups points together on the basis of a constant search radius, but then combines clusters that overlap. On the STAC Parameters tab, define a search radius, the minimum number of points that are required for each cluster, and an output size for displaying the clusters with ellipses.

The tabular results can be printed, saved to a text file, or output as a ‘dbf’ file. The graphical results can be output as either ellipses or as convex hulls (or both) to *ArcGIS* ‘shp’, *MapInfo* ‘mif’, various ASCII formats, or *Google Earth* ‘kml’ (if the coordinates are spherical) files. Separate file names must be selected for the ellipse output and for the convex hull output. For MapInfo ‘mif’ format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

STAC parameters

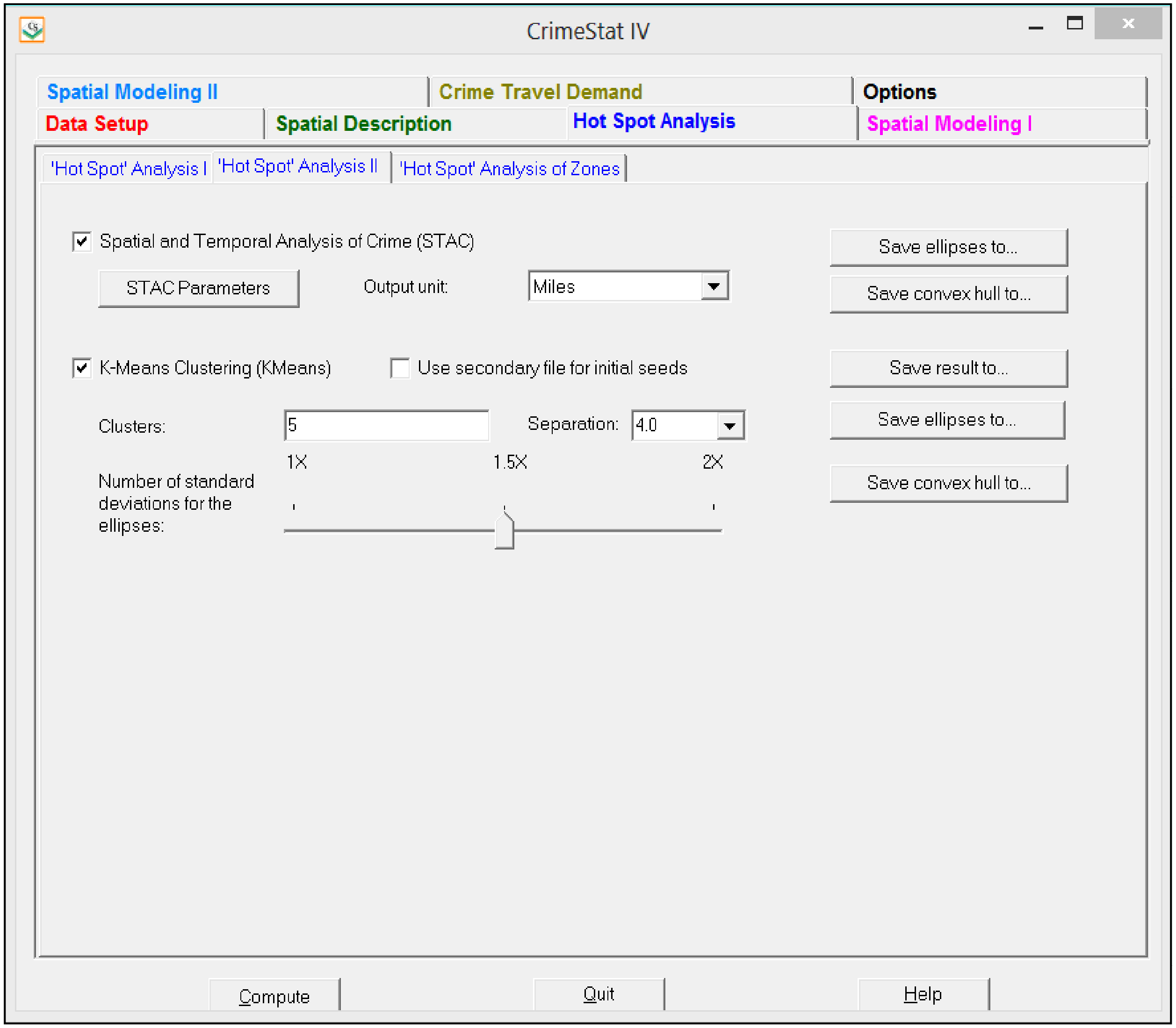
The STAC parameters tab allows the selection of a search radius, the minimum number of points, the scan type, the boundary definition, the number of simulation runs, and the output size of the STAC ellipses.

Search radius

The search radius is the distance within the STAC routine searches. The default is 0.5 miles. A 20 x 20 grid is overlaid on the study area. At each intersection of a row and a

Figure 2.10:

Hot Spot Analysis II



☑ Spatial and Temporal Analysis of Crime (STAC)

Output unit: Miles

☑ K-Means Clustering (KMeans)

□ Use secondary file for initial seeds

Clusters: 5

Separation: 4.0

Number of standard deviations for the ellipses: 1X 1.5X 2X column, the routine counts all points that are closer than the search radius. Overlapping circles are combined to form variable-size clusters. The smaller the search radius that is selected, the fewer points will be selected. On the other hand, choosing a larger search area, the more points will be selected. However, the larger the search area, the greater the likelihood that clusters could be chance groupings. On the STAC Parameters tab, type the search radius into the box and specify the measurement units (miles, nautical miles, feet, kilometers, or meters).

Scan type

The scan type is the type of grid overlaid on the study area. There are two choices: rectangular (default) and triangular.

Boundary

The study area boundaries can be defined from the data set or the reference grid.

Minimum number of points

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 5 points. 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. On the STAC Parameters tab, type the minimum number of points each cluster is required to have.

Tabular output

The routine outputs eight results for each cluster that is calculated:

1. The hierarchical order and the cluster number
2. The mean center of the cluster (Mean X and Mean Y)
3. The standard deviational ellipse of the cluster (the rotation and the lengths of the X and Y axes)
4. The number of points in the cluster
5. The area of the cluster
6. The density of the cluster (cluster points divided by area)
7. The number of points in the ellipse
8. The density of the ellipse (ellipse points divided by area)

Ellipse output

The results can be output graphically as an ellipse to *ArcGIS* 'shp', *MapInfo 'mif', various ASCII formats*, or *Google Earth* 'kml' (if the coordinates are spherical) files. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected. The ellipses will be output as combined objects. The prefix will be 'ST'.

Output size for ellipses

The cluster output size can be adjusted by the lower slide bar This specifies the number of ellipse standard deviations to be calculated for each cluster: one standard deviation (1X - the default value), one and a half standard deviations (1.5X), or two standard deviations (2X). The default value is one standard deviation. Typically, one standard deviation will cover more than half the cases whereas two standard deviations will cover more than 99% of the cases, though the exact percentage will depend on the distribution. The output file is saved as St<file name> with the file name being provided by the user. On the STAC Parameters tab, slide the bar to select the number of standard deviations for the ellipses.

Convex hull cluster output

The clusters can also be output as convex hulls to *ArcGIS* 'shp', *MapInfo* 'mif', various ASCII formats, or *Google Earth* 'kml' (if the coordinates are spherical) files. Specify a file name. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected. The name will be output with a 'CST' prefix.

Simulating confidence intervals

A Monte Carlo simulation can be run to estimate the approximate confidence intervals around the STAC clusters. The user specifies the number of simulation runs and the STAC clustering is calculated for randomly assigned data. The random output is sorted and percentiles are calculated. The output includes the number of clusters, the area, the number of points, and the density. Fifteen percentiles are identified for these statistics:

1. The minimum for the spatially random STAC simulations
2. The minimum for the spatially random STAC simulations

3. The minimum for the spatially random STAC simulations
4. The minimum for the spatially random STAC simulations
5. The maximum for the spatially random STAC simulations
6. The 0.5 percentile for the spatially random STAC simulations
7. The 1 percentile for the spatially random STAC simulations
8. The 2.5 percentile for the spatially random STAC simulations
9. The 5 percentile for the spatially random STAC simulations
10. The 10 percentile for the spatially random STAC simulations
11. The 90 percentile for the spatially random STAC simulation
12. The 95 percentile for the spatially random STAC simulations
13. The 97.5 percentile for the spatially random STAC simulations
14. The 99 percentile for the spatially random STAC simulations
15. The 99.5 percentile for the spatially random STAC simulations

Confidence intervals can be estimated from these percentiles. The two most commonly used ones are the 95% (defined by the 2.5 and 97.5 percentiles) and the 99% (defined by the 0.5 and 99.5 percentiles). The simulated data that is used can be viewed by checking the 'Dump simulation data' box on the Options tab.

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 routine finds K seed locations in which points are assigned to the nearest cluster. The default K is 5. If K is small, the clusters will typically cover larger areas.

The tabular results can be printed, saved to a text file, or output as a ‘dbf’ file. The graphical results can be output as either ellipses or as convex hulls (or both) to *ArcGIS* ‘shp’, *MapInfo ‘mif’*, various ASCII formats, or *Google Earth* ‘kml’ (if the coordinates are spherical) files. Separate file names must be selected for the ellipse output and for the convex hull output. For MapInfo ‘mif’ format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Initial cluster locations

The routine starts with an initial guess (seed) for the K locations and then conducts local optimization. The user can modify the location of the initial clusters in two ways, which are not mutually exclusive:

Separation

1. The separation between the initial clusters can be increased or decreased. There is a separation scale with pre-defined values from 1 to 10; the default is 4. The user can type in any number, however (e.g., 15). Increasing the number increases the separation between the initial cluster locations while decreasing the number decreases the separation.

Initial seed locations

2. The user can define the initial seed locations and the number of clusters, K, with a secondary file. The routine takes K from the number of points in the secondary file and takes the X/Y coordinates of the points as the initial seed locations.

Tabular output

The routine outputs seven characteristics for each cluster that is calculated:

1. The cluster ID
2. The center of minimum distance of the cluster (Mean X and Mean Y)
3. The standard deviational ellipse of the cluster (the rotation and the lengths of the X and Y axes)
4. The area of the cluster
5. The sum of squares in distances between the center of minimum distance of the cluster and each point that is part of the cluster
6. The mean squared error of the distances between the center of minimum distance of the cluster and each point that is part of the cluster
7. The number of points in the cluster

Ellipse output

The results can be output graphically as an ellipse to *ArcGIS* ‘shp’, *MapInfo* ‘mif’, various ASCII formats, or *Google Earth* ‘kml’ (if the coordinates are spherical) files. Specify a file name. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected. The ellipses will be output as separate objects with a 'KM' prefix.

Output size for ellipses

For both methods, the cluster output size can be adjusted with the lower slide bar. This specifies the number of ellipse standard deviations to be calculated for each cluster: one standard deviation (1X - the default value), one and a half standard deviations (1.5X), or two standard deviations (2X). The default value is one standard deviation. Typically, one standard deviation will cover more than half the cases whereas two standard deviations will cover more than 99% of the cases, though the exact percentage will depend on the distribution. Slide the bar to select the number of standard deviations for the ellipses. The output file is saved as Km<file name> with the file name being provided by the user.

Convex hull cluster output

The clusters can also be output as convex hulls to *ArcGIS* 'shp', *MapInfo* 'mif', various ASCII formats, or *Google Earth* 'kml' (if the coordinates are spherical) files. Specify a file name. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected. The convex hulls will be output as separate objects with a 'CKM' prefix.

Hot Spot Analysis of Zones

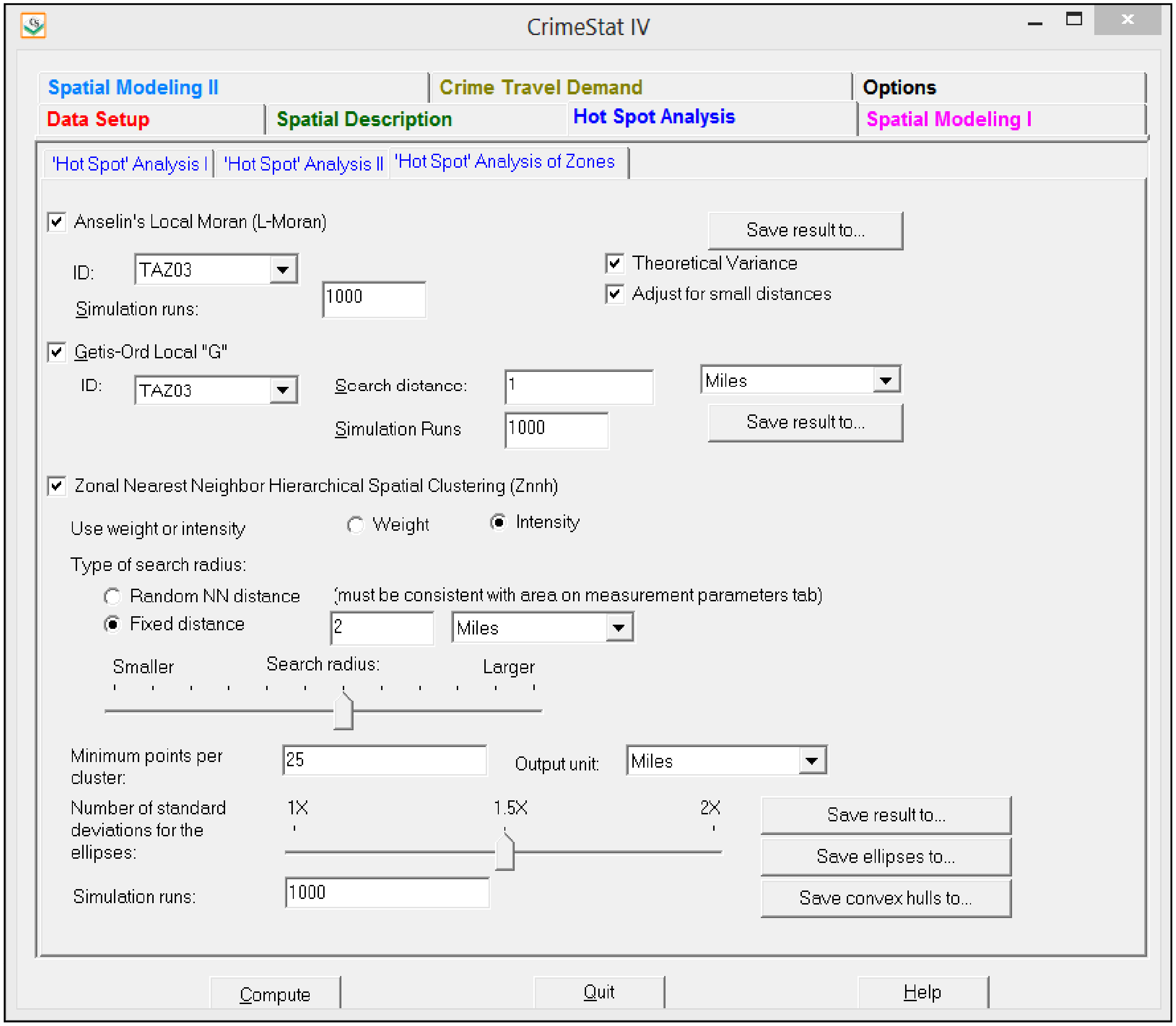
The Hot Spot Analysis of Zones section includes clustering statistics for zonal data. These include 1) Anselin's local Moran; 2) the Getis-Ord local "G", and 3) the zonal nearest neighbor hierarchical clustering algorithm.

Anselin's Local Moran (L-Moran)

Anselin's 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 zone. The index points to

Figure 2.11:

Hot Spot Analysis of Zones



☑ Anselin's Local Moran (L-Moran)

ID: TAZ03

Simulation runs: 1000

☑ Theoretical Variance

☑ Adjust for small distances

☑ Getis-Ord Local "G"

ID: TAZ03

Search distance: 1 Miles

Simulation Runs 1000

☑ Zonal Nearest Neighbor Hierarchical Spatial Clustering (Znnh)

Use weight or intensity ○ Weight ● Intensity

Type of search radius:

○ Random NN distance (must be consistent with area on measurement parameters tab)

● Fixed distance 2 Miles

Smaller Search radius: Larger

Minimum points per cluster: 25

Output unit: Miles

Number of standard deviations for the ellipses: 1X 1.5X 2X

Simulation runs: 1000

clustering or dispersion relative to the local neighborhood. Zones with high "I" values have an intensity value that is higher than their neighbors while zones with low "I" values have intensity values lower than their neighbors. The output can be printed or output as a 'dbf' file.

ID field

The user should indicate a field for the ID of each point (or zone). This ID will be saved with the output and can then be linked with the input file (Primary File) for mapping.

Theoretical variance

If checked, the routine will calculate the theoretical variance of the "I" value for each zone (see documentation for details).

Adjust for small distances

If checked, small distances are adjusted so that the maximum weighting is no higher than 1. This ensures that the local "I" won't become excessively large for points that are grouped together. The default setting is no adjustment.

Simulation of confidence intervals

A Monte Carlo simulation can be run to estimate approximate confidence intervals around the "I" value for each zone. Note, a simulation may take time to run especially if the data set is large or if a large number of simulation runs are requested. Specify the number of simulations to be run (e.g., 100, 1000, 10000).

Output

The output is for each zone and includes:

1. The sample size
2. The ID for the zone
3. The X coordinate for the zone
4. The Y coordinate for the zone
5. The "I" for the zone
6. The expected "I" for the zone

and if the theoretical variance is checked:

7. The theoretical variance of the “I” for the zone
8. A Z-test of the “I” under the assumption of normality

and if a simulation is run:

9. The 0.5 percentile of “I” for the zone
10. The 2.5 percentile of “I” for the zone
11. The 97.5 percentile of “I” for the zone
12. The 99.5 percentile of “I” for the zone

The two pairs of percentiles (2.5 and 97.5; 0.5 and 99.5) create approximate 95% and 9% confidence interval of “I” for each zone. The tabular results can be printed, saved to a text file or saved as a ‘dbf’ file (LMoranCorr<file name> with the file name being provided by the user.

The ‘dbf’ output file can then be linked to the input ‘dbf’ file by using the ID field as a matching variable. This would be done if the user wants to map the “I” variable, the Z-test, or those zones for which the “I” value is either higher than the 97.5 or 99.5 percentiles or lower than the 2.5 or 0.5 percentiles of the simulation results.

Getis-Ord Local “G” (L-Getis-Ord)

The Getis-Ord “G” statistic is an index of spatial autocorrelation for values of a variable that fall within a specified distance of each other. When compared to an expected value of G under the assumption of no spatial association, it has the advantage over other global spatial autocorrelation measures (Moran, Geary) in that it can distinguish between hot spots and cold spots. The “G” value is calculated with the intensity variable specified on the Primary File page and with respect to a specified search distance (defined by the user).

The Getis-Ord Local “G” statistic applies the Getis-Ord "G" statistic to individual zones to assess whether particular zones are spatially related to the nearby ones (‘neighbors’). Unlike the global Getis-Ord “G”, the Getis-Ord Local “G” is applied to each individual zone.

By itself, the G statistic for an individual zone is not very meaningful. The “G” value varies from 0 to 1 since it indicates the interaction of pairs of zones that are within the search distance relative to the interaction of all pairs of zones. As the search distance increases, this statistic will automatically approach 1.0. Consequently, G is compared to an expected value of G under the assumption of no significant spatial association.

Further, under the assumption that G is normally distributed, a Z-test can be constructed that tests for the significance of the actual G. A positive Z-value indicates spatial clustering of high values more than what would be expected under chance (hot spots) while a negative Z-value indicates spatial clustering of low values more than what would be expected under chance (cold spots). A "G" value around 0 indicates no spatial autocorrelation.

ID field

The user should indicate a field for the ID of each point (or zone). This ID will be saved with the output and can then be linked with the input file (Primary File) for mapping.

Search distance

The user must specify a search distance for the test and indicate the distance units (miles, nautical miles, feet, kilometers, or meters).

Simulation of confidence intervals

Since the Getis-Ord "G" statistic may not be normally distributed, the significance test is frequently inaccurate. Instead, a permutation type Monte Carlo simulation can be run to estimate approximate confidence intervals around the "G" value. Specify the number of simulations to be run (e.g., 100, 1000, 10000).

Output

The output is for each zone and includes:

1. The sample size
2. The ID for the zone
3. The X coordinate for the zone
4. The Y coordinate for the zone
5. The "G" for the zone
6. The expected "G" for the zone
7. The standard deviation of "G" for the zone
8. A Z-test of "G" under the assumption of normality for the zone

and if a simulation is run:

9. The 0.5 percentile of "G" for the zone

10. The 2.5 percentile of “G” for the zone
11. The 97.5 percentile of “G” for the zone
12. The 99.5 percentile of “G” for the zone

The two pairs of percentiles (2.5 and 97.5; 0.5 and 99.5) create approximate 95% and 99% confidence interval of “G” for each zone. The tabular results can be printed, saved to a text file or saved as a ‘dbf’ file (LGetis-OrdCorr<file name> with the file name being provided by the user.

The ‘dbf’ output file can then be linked to the input ‘dbf’ file by using the ID field as a matching variable. This would be done if the user wants to map the “G” variable, the expected “G” or those zones for which the “G” value is either higher than the 97.5 or 99.5 percentiles or lower than the 2.5 or 0.5 percentiles of the simulation results.

Zonal Nearest Neighbor Hierarchical Clustering (Znnh)

The zonal nearest neighbor hierarchical spatial clustering routine applies the nearest neighbor hierarchical clustering algorithm. The point-based Nnh is a constant-distance clustering routine that groups points together on the basis of spatial proximity. A threshold distance is defined and the minimum number of points that are required for each cluster specified. The output can be displayed with ellipses or convex hulls.

On the other hand, in the zonal Nnh (Znnh), the algorithm is adjusted to allow *weighting* of each zone, usually applied to a single point within the zone (e.g., a centroid). Thus, if the ‘point’ is a centroid of a zone, then the weighting is an attribute assigned to that centroid (e.g., population, employment, median household income). Clusters are groups of adjacent zones that have much higher weights than non-clustered zones.

The routine requires a primary file (e.g., robberies) that is weighted with the weight or intensity variable (see Primary File). On the Znnh routine, the user defines a weighting variable, a threshold distance and the minimum number of zones that are required for each cluster, and an output size for displaying the clusters with ellipses or convex hulls.

The routine identifies first-order clusters that represent groups of zones that are closer together than the threshold distance, that have the highest weights, and in which there is at least the minimum number of zones specified by the user (the minimum is 3 zones). Clustering is hierarchical in that the first-order clusters are treated as separate zones to be clustered into second-order clusters, and the second-order clusters are treated as separate zones to be clustered into third-order clusters, and so on. Higher-order clusters will be identified only if the distances between their centers are closer than the new threshold distance.

Weighting variable

Each zone must be weighted by a variable. This can be either the intensity variable or the weight variable defined on the Primary File page (but not both). The user specifies whether the intensity or the variable variable is to be used. The default is Intensity.

Threshold distance

The threshold distance is the search radius around a zone centroid. For each pair of zones, the routine determines whether they are closer together than the search radius. There are two ways to determine a threshold distance:

Random nearest neighbor distance

First, the search distance is chosen by the random nearest neighbor distance. The default value is 0.1 (i.e., fewer than 10% of the pairs could be expected to be as close or closer by chance.) Pairs of zones that are closer together than the threshold distance are grouped together whereas pairs of zones that are greater than the threshold distance are ignored. The smaller the threshold distance and the smaller the significance level that is selected, then the fewer numbers of paired zones will be selected. On the other hand, choosing a larger threshold distance (and, consequently, a higher significance level) will usually lead to more pairs being selected. However, the more pairs that are selected, the greater the likelihood that clusters could be chance groupings.

The slide bar is used to adjust the significance level. Move the slide bar to the left to choose a smaller threshold distance and to the right to choose a larger threshold distance.

Fixed distance

Second, a fixed distance can be selected. The default is 1 mile. In this case, the search radius uses the fixed distance and the slide bar is inoperative.

Minimum number of zones

The minimum number of zones required for each cluster allows the user to specify a minimum number of zones for each cluster. The default is 10 zones and the minimum is 3. Third, the output size for the clusters can be adjusted by the second slide bar. These are the number of standard deviations defined by the ellipse, from one standard deviation (the default value) to three standard deviations. Typically, one standard deviation will cover about 65% of the cases whereas three standard deviations will cover more than 99% of the cases.

The tabular results can be printed, saved to a text file, or output as a 'dbf' file. The graphical results can be output as either ellipses or as convex hulls (or both) to *ArcGIS* 'shp', *MapInfo* 'mif', various ASCII formats, or Google Earth 'kml' (if the coordinate system is spherical) files. Separate file names must be selected for the ellipse output and for the convex hull output. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Simulating confidence intervals

A Monte Carlo simulation can be run to estimate the approximate confidence intervals around first-order Nnh clusters; second- and higher-order clusters are not simulated since their structure depends on first-order clusters. The user specifies the number of simulation runs and the Nnh clustering is calculated for randomly assigned data. The random output is sorted and percentiles are calculated. The output includes the number of first-order clusters, the area, the number of zones, and the density.

Confidence intervals can be estimated from these percentiles. The two most commonly used ones are the 95% (defined by the 2.5 and 97.5 percentiles) and the 99% (defined by the 0.5 and 99.5 percentiles). The simulated data that is used can be viewed by checking the 'Dump simulation data' box on the Options tab.

Type of graphical output

The type of graphical output is specified, either standard deviational ellipses or convex hulls around the zones identified in each cluster. If the output is to be ellipses, then the output size for the clusters can be adjusted by the second slide bar. These are the number of standard deviations defined by the ellipse, from one standard deviation (the default value) to three standard deviations. Typically, one standard deviation will cover about 50-60% of the zones (and a higher percentage of the total of the weighting variable) whereas three standard deviations will cover more than 99% of the zones. On the other hand, if the output is to be convex hulls, the routine outputs a convex hull for each identified cluster.

Ellipse output

The results can be output graphically as an ellipse to *ArcGIS* 'shp', *MapInfo* 'mif', various ASCII formats, or Google Earth 'kml' (if the coordinate system is spherical) files. A file name should be provided. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

First and higher-order ellipses will be output as separate objects. The prefix will be 'NNH1' for the first-order ellipses, 'NNH2' for the second-order ellipses, and 'NNH3' for the third-order ellipses. Higher-order ellipses will only index the number.

Output size for ellipses

The cluster output size can be adjusted by the lower slide bar. This specifies the number of ellipse standard deviations to be calculated for each cluster: one standard deviation (1X - the default value), one and a half standard deviations (1.5X), or two standard deviations (2X). The default value is one standard deviation. Typically, one standard deviation will cover more than half the zones in a cluster whereas two standard deviations will cover more than 99% of the zones in a cluster, though the exact percentage will depend on the distribution. Slide the bar to select the number of standard deviations for the ellipses. The output file is saved as Znnh<number><file name> with the file name being provided by the user. The number is the order of the clustering (i.e., 1, 2…).

Restrictions on the number of clusters can be placed by defining a minimum number of zones that are required. The default is 10 and the minimum is 3. If there are too few zones allowed, then there will be many very small clusters. By increasing the number of required zones, the number of clusters will be reduced.

Convex hull cluster output

The clusters can also be output as convex hulls to *ArcGIS* 'shp', *MapInfo* 'mif', various ASCII formats, or Google Earth 'kml' (if the coordinate system is spherical) files. Specify a file name. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

The name will be output with a 'CNNH1' prefix for the first-order clusters, a 'CNNH2' prefix for the second-order clusters, and a 'CNNH3' prefix for the third-order clusters. Higher-order clusters will index only the number.

Note that ellipses may extend beyond the zones that are clustered together and may also leave out zones that are part of the cluster. Ellipses are abstractions and, while good for visualization, are not precise. Convex hulls are more precise since they define only those zones that are part of the cluster.

Tabular output

The routine outputs six results for each cluster that is calculated:

1. The hierarchical order and the cluster number
2. The mean center of the cluster (Mean X and Mean Y)
3. The standard deviational ellipse of the cluster (the rotation and the lengths of the X and Y axes)
4. The number of zones in the cluster
5. The area of the cluster
6. The density of the cluster (the total weight of the zones divided by area)

and if a simulation is run:

7. The minimum for the spatially random Znnh simulations:
8. The maximum for the spatially random Znnh simulations
9. The 0.5 percentile for the spatially random Znnh simulations
10. The 1 percentile for the spatially random Znnh simulations
11. The 2.5 percentile for the spatially random Znnh simulations
12. The 5 percentile for the spatially random Znnh simulations
13. The 10 percentile for the spatially random Znnh simulations
14. The 90 percentile for the spatially random Znnh simulations
15. The 95 percentile for the spatially random Znnh simulations
16. The 97.5 percentile for the spatially random Znnh simulations
17. The 99 percentile for the spatially random Znnh simulations
18. The 99.5 percentile for the spatially random Znnh simulations

IV. Spatial Modeling I

The first spatial modeling section conducts kernel density estimation, Head Bang statistics, space-time analysis, journey-to-crime calibration and estimation, and Bayesian journey-to-crime diagnostics and estimation. The spatial modeling section is made up of five distinct tabs: Interpolation I, Interpolation II, Space-time analysis, Journey-to-crime estimation, and Bayesian Journey-to-crime estimation.

Interpolation I

The interpolation I 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 that compares 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 are 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.

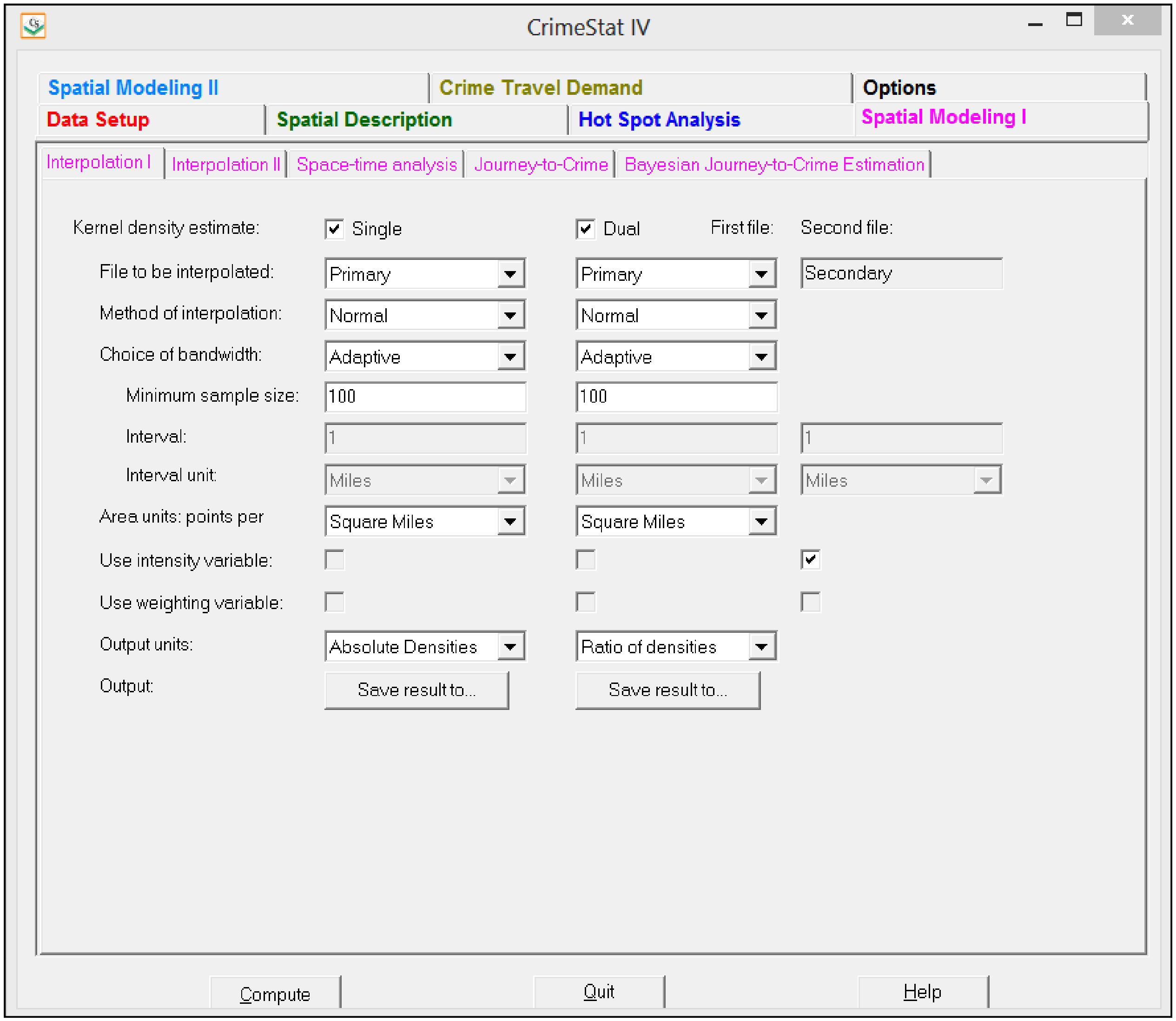
Method of interpolation

There are five types of kernel distributions that can be used to estimate point density:

1. The **normal** kernel overlays a three-dimensional normal distribution over each point that then extends over the area defined by the reference file. This is the default kernel function.

Figure 2.12:

Interpolation I Statistics



Kernel density estimate:	☑ Single	☑ Dual First file:	Second file:
File to be interpolated:	Primary	Primary	Secondary
Method of interpolation:	Normal	Normal	
Choice of bandwidth:	Adaptive	Adaptive	
Minimum sample size:	100	100	
Interval:	1	1	1
Interval unit:	Miles	Miles	Miles
Area units: points per	Square Miles	Square Miles	
Use intensity variable:	□	□	☑
Use weighting variable:	□	□	□
Output units:	Absolute Densities	Ratio of densities	
Output:	Save result to...	Save result to...	

2. The **uniform** kernel overlays a uniform function over each point that only extends for a limited distance.

3. The **quartic** kernel overlays a quartic function over each point that only extends for a limited distance.

4. The **triangular** kernel overlays a three-dimensional triangle over each point that only extends for a limited distance.

5. The **negative exponential** kernel overlays a three dimensional negative exponential function over each point that only extends for a limited distance

The methods produce similar results though 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 uniform, quartic, triangular and negative exponential kernels, bandwidth is the radius of a circle defined by the surface. For all types, larger bandwidth will produce smoother density estimates and 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, or meters).

Output (areal) units

Specify the areal 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

The density estimate for each cell can be calculated in one of three ways:

1. **Absolute densities**. This is the number of points per grid cell and is scaled so that the sum of all grid cells equals the sample size. This is the default.

2. **Relative densities**. For each grid cell, this is the absolute density divided by the grid cell area and is expressed in the output units (e.g., points per square mile)

3. **Probabilities**. This is the proportion of all incidents that occur in the grid cell. The sum of all grid cells equals a probability of 1

Select whether absolute densities, relative densities, or probabilities are to be output for each cell. The default is absolute densities.

Output

The results can be output as a *Surfer for Windows* file (for both an external or generated reference file) or as an *ArcGIS* ‘shp’, *MapInfo* ‘mif’, *ArcGIS Spatial Analyst* 'asc', or ASCII grid 'grd' file (only if the reference file is generated by *CrimeStat*). For MapInfo ‘mif’ format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

The output file is saved as K<file name> with the file name being provided by the user.

Dual Kernel Density Estimate (DualKernel)

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 five types of kernel distributions that can be used to estimate point density:

1. The **normal** kernel overlays a three-dimensional normal distribution over each point that then extends over the area defined by the reference file. This is the default kernel function.

2. The **uniform** kernel overlays a uniform function over each point that only extends for a limited distance.

3. The **quartic** kernel overlays a quartic function over each point that only extends for a limited distance.

4. The **triangular** kernel overlays a three-dimensional triangle over each point that only extends for a limited distance.

5. The **negative exponential** kernel overlays a three dimensional negative exponential function over each point that only extends for a limited distance

The methods produce similar results though 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 uniform, quartic, triangular and negative exponential kernels, bandwidth is the radius of a circle defined by the surface. For all types, larger bandwidth will produce smoother density estimates and 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, or 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, or meters). The default is one mile for both the first and second files.

Output (areal) units

Specify the areal 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

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

The density estimate for each cell can be calculated in one of six ways:

1. Ratio of densities - this is the ratio of the density for the first file divided by the density of the second file
2. Log ratio of densities - this is the natural logarithm of the ratio of the density for the first file divided by the density of the second file.
3. Absolute difference in densities - this is the difference between the absolute density of the first file and the absolute density of the second file. It is the *net* difference. The densities of each file are scaled so that the sum of the grid cells equals the sample size.
4. Relative difference in densities - this is the difference between the relative density of the first file and the relative density of the second file. It is the *relative* difference. The cell densities of each file are divided by the grid cell area to produce a measure of relative density in the specified output units (e.g., points per square mile). The relative density of the second file is then subtracted from the relative density of the first file.
5. Absolute sum of densities - this is the sum of the absolute density of the first file and the absolute density of the second file. It is the *net* sum. The densities of each file are scaled so that the sum of the grid cells equals the sample size.
6. Relative sum of densities - this is the sum of the relative density of the first file and the relative density of the second file. It is the *relative* sumThe cell densities of each file are divided by the grid cell area to produce a measure of relative density in the specified output units (e.g., points per square mile). The relative density of the second file is then added to the relative density of the first file.

Select whether the ratio of densities, the log ratio of densities, the absolute difference in densities, the relative difference in densities, the absolute sum of densities, or the relative sum of densities are to be output for each cell. The default is the ratio of densities.

Output

The results can be output as a *Surfer for Windows* file (for both an external or generated reference file) or as an *ArcGIS* ‘shp’, *MapInfo* ‘mif’, *ArcGIS Spatial Analyst* 'asc', or ASCII grid 'grd' file (only if the reference file is generated by *CrimeStat*). For MapInfo ‘mif’ format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

The output file is saved as *DK<file name>* with the file name being provided by the user.

Interpolation II

The interpolation II tab allows the implementation of the Head Bang statistic for zonal data and its interpolation to a grid.

Head Bang

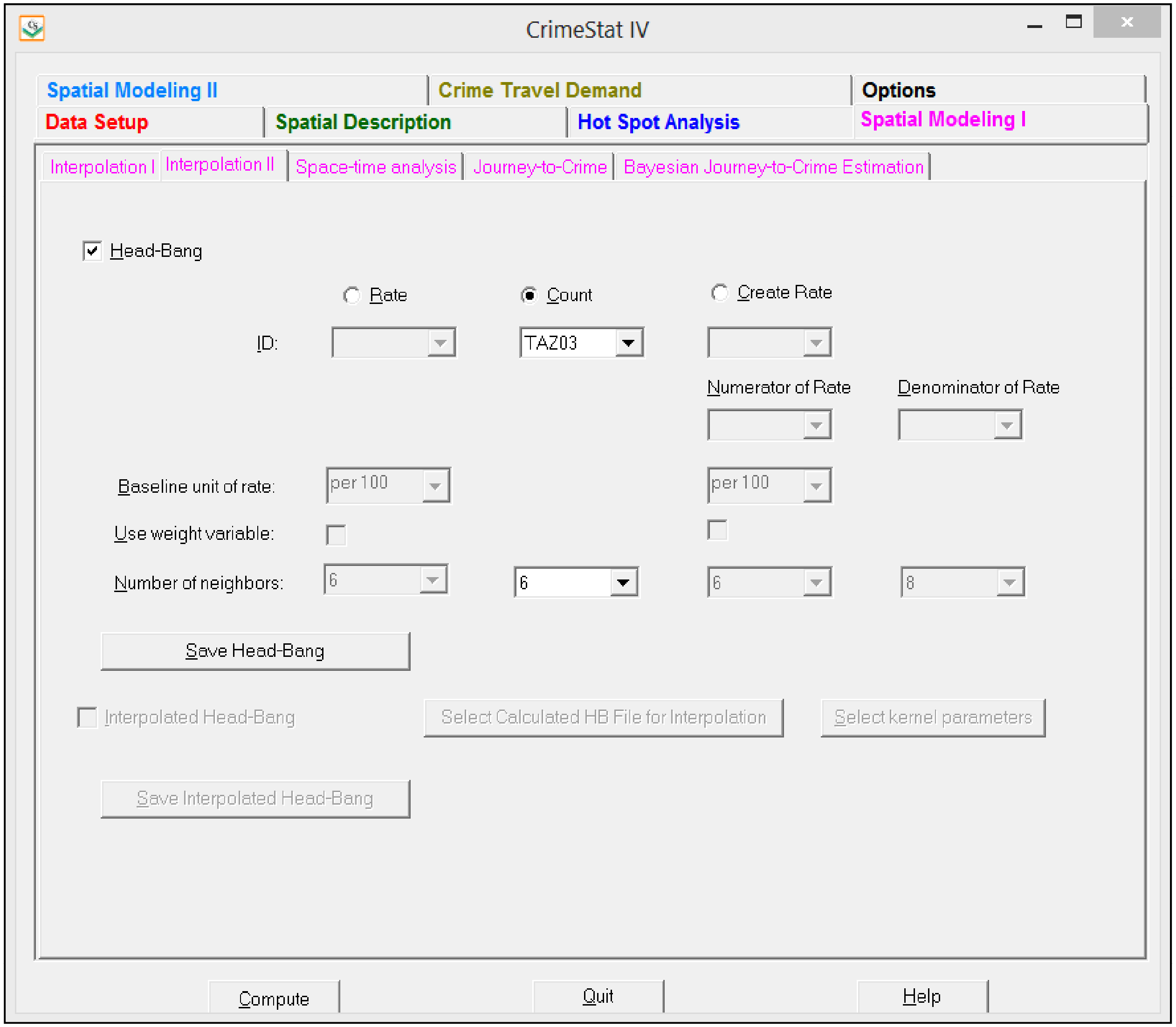
The Head Bang statistic is a weighted two-dimensional smoothing algorithm that is applied to zonal data. It is useful for eliminating extreme values in a distribution and adjusting the values of zones to be similar to their neighbors. The statistic requires an intensity variable in the primary file. The value of the intensity variable for each zone is compared to its neighbors with the number of neighbors defined by the user. The intensity values of the neighbors are rank-ordered and then divided into two equal-sized groups, high and low. The median of the high group of neighbors and the median of the low group of neighbors are calculated. The intensity value of the zone is then compared to these two medians. If it falls between the two medians, then the zone keeps its intensity value. If its value is higher than the high median, then the zone takes the high median as its value unless it has a weighting which is greater than its neighbors. If its value is lower than the low median, then the zone takes the low median as its value unless it has a weighting which is greater than its neighbors.

Type of Variable to be Smoothed

The user must specify whether the variable to be smoothed is a rate variable, a volume variable, or two variables that are to be combined into a rate.

Figure 2.13:

Interpolation II Statistics



☑ Head-Bang

○ Rate ● Count ○ Create Rate

ID: TAZ03

Numerator of Rate

Denominator of Rate

Baseline unit of rate: per 100 per 100

Use weight variable: □ □

Number of neighbors: 6 6 6 8

□ Interpolated Head-Bang

Select Calculated HB File for Interpolation
Select kernel parameters
Save Interpolated Head-Bang
Compute
Quit
Help

Rate variable

If the variable to be smoothed is a rate variable, the variable that is smoothed must be defined in the Z(Intensity) field on the Primary File. Also, a weight variable should be chosen and should be defined in the Weight field on the Primary File.

ID field

The ID field that identifies zones must be defined.

Baseline unit for rate

The rate is an index of one variable relative to another variable, the baseline. Specify the unit that the rate is expressed by powers of 10. The range is from 1 (absolute rate) to 1 per 1,000,000. The default is 1 per 100 (or percentages).

Use weight variable

The rate can (and probably should) be weighted by an additional weight variable specified on the Primary File page. Check the 'Use weight variable' box to weight the rate. Otherwise, the weight is 1. A typical weight variable would be the population size of the zone.

Number of neighbors

The user must also specify the number of neighbors to be used for the comparison. The number of neighbors can run from 4 through 40. The default is 6. If the number of neighbors selected is even, the routine divides the data set into two equal-sized groups. If the number of neighbors selected is odd, then the middle zone is used in calculating both the low median and the high median.

Volume variable

If the variable to be smoothed is a volume variable, the variable that is smoothed must be defined in the Z(Intensity) field on the Primary File.

ID field

The ID field that identifies zones must be defined.

Number of neighbors

The user must also specify the number of neighbors to be used for the comparison. The number of neighbors can run from 4 through 40. The default is 6. If the number of neighbors selected is even, the routine divides the data set into two equal-sized groups. If the number of neighbors selected is odd, then the middle zone is used in calculating both the low median and the high median.

Create rate

Unlike the rate and volume calculations, the user must specify which two variables (fields) must be related to create a rate. One of these is to be defined in the *numerator of the rate* box and one in the *denominator of the rate* box. For example, if the data include number of robberies as one field in the data set and population as another field, then the number of robberies would identified as the numerator of the rate while population would be identified as the denominator of the rate. Both variables should be volumes.

Also, a weight variable should be chosen and should be defined in the Weight field on the Primary File. The weight is applied to the created rate after it is calculated. A typical weight variable would be the population size of the zone.

ID field

The ID field that identifies zones must be defined.

Baseline unit for rate

The rate is an index of one variable relative to another variable, the baseline. The result of the division of the numerator by the denominator will then be multiplied by the base unit of the baseline. Specify the unit that the rate is expressed by powers of 10. The range is from 1 (absolute rate) to 1,000,000 (resulting in an index of 1:1,000,000). The default is 100 (resulting in an index of 1:100, or percentages).

Use weight variable

The rate can (and probably should) be weighted by an additional weight variable specified on the Primary File page. Check the 'Use weight variable' box to weight the rate. Otherwise, the weight is 1. A typical weight variable would be the population size of the zone.

Number of neighbors

The user must also specify the number of neighbors to be used for the comparison. The number of neighbors can run from 4 through 40. The default is 6. If the number of neighbors selected is even, the routine divides the data set into two equal-sized groups. If the number of neighbors selected is odd, then the middle zone is used in calculating both the low median and the high median.

Output for each zone

The output is for each zone and includes:

1. The ID field
2. The X coordinate
3. The Y coordinate
4. The smoothed intensity variable (Z_MEDIAN)
5. The weight of the zone (WEIGHT). The default is 1.0.

Select output file

The tabular results can be printed, saved to a text file or saved as a 'dbf' file. For saving to a 'dbf' file, specify a file name in the "Save result to" in the dialogue box.

1. If the routine is run on a volume, then the file is saved as VolHB<file name> with the file name being provided by the user.

2. If the routine is run on a rate, then the file is saved as RateHB<file name> with the file name being provided by the user.

3. If the routine is run with a rate being created from two variables in the file, then the file is saved as CRateHB< file name> with the file name being provided by the user.

The 'dbf' file can then be linked to the input 'dbf' file by using the ID field as a matching variable. This would be done if the user wants to map the smoothed variable.

Interpolated Head Bang (IHB)

The Head Bang calculations can be interpolated to a grid. If the user checks this box, then the routine will also interpolate the calculations to a grid using kernel density estimation. An output file from the Head Bang routine is required. Also, a reference file is required to be defined on the Reference File page.

Essentially, the routine takes a Head Bang output and interpolates it to a grid using a kernel density function. The same results can be obtained by inputting the Head Bang output on the Primary File page and using the single kernel density routine on the Interpolations I page. However, there is no intensity variable in the Interpolated Head Bang because the intensity has already been incorporated in the Head Bang output. Also, there is no weighting of the Head Bang estimate.

The user must then define the parameters of the interpolation.

Method of interpolation

There are five types of kernel distributions that can be used to interpolate the Head Bang to the grid:

1. The **normal** kernel overlays a three-dimensional normal distribution over each point that then extends over the area defined by the reference file. This is the default kernel function.

2. The **uniform** kernel overlays a uniform function over each point that only extends for a limited distance.

3. The **quartic** kernel overlays a quartic function over each point that only extends for a limited distance.

4. The **triangular** kernel overlays a three-dimensional triangle over each point that only extends for a limited distance

5. The **negative exponential** kernel overlays a three dimensional negative exponential function over each point that only extends for a limited distance.

The different kernel functions produce similar results though 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 uniform, quartic, triangular and negative exponential kernels, bandwidth is the radius of a circle defined by the surface. For all types, larger bandwidth will produce smoother density estimates and 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, or meters).

Output (areal) units

Specify the areal 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.

Calculate densities or probabilities

The density estimate for each cell can be calculated in one of three ways:

1. **Absolute densities**. This is the number of points per grid cell and is scaled so that the sum of all grid cells equals the sample size. This is the default.

2. **Relative densities**. For each grid cell, this is the absolute density divided by the grid cell area and is expressed in the output units (e.g., points per square mile)
3. **Probabilities**. This is the proportion of all incidents that occur in the grid cell. The sum of all grid cells equals a probability of 1

Select whether absolute densities, relative densities, or probabilities are to be output for each cell. The default is absolute densities.

Output

The results can be output as an *ArcGIS* 'shp', *MapInfo* 'mif', *ArcGIS Spatial Analyst* 'asc', *Surfer for Windows* file (for both an external or generated reference file), or as or ASCII grid 'grd' file (only if the reference file is generated by *CrimeStat*). For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

The output file is saved as IHB<file name> with the file name being provided by the user.

Space-time Analysis

The space-time analysis tab allows the analysis of the interaction between space and time. There are four routines. First, there is the Knox index that shows the simple binomial relationship between events occurring in space and in time. Second, there is the Mantel index that shows the correlation between closeness in space and closeness in time. Third, there is a spatial-temporal moving average that calculates a mean center for a temporal span. Fourth, there is a Correlated Walk Analysis that diagnoses the spatial and temporal sequencing of incidents committed by a serial offender.

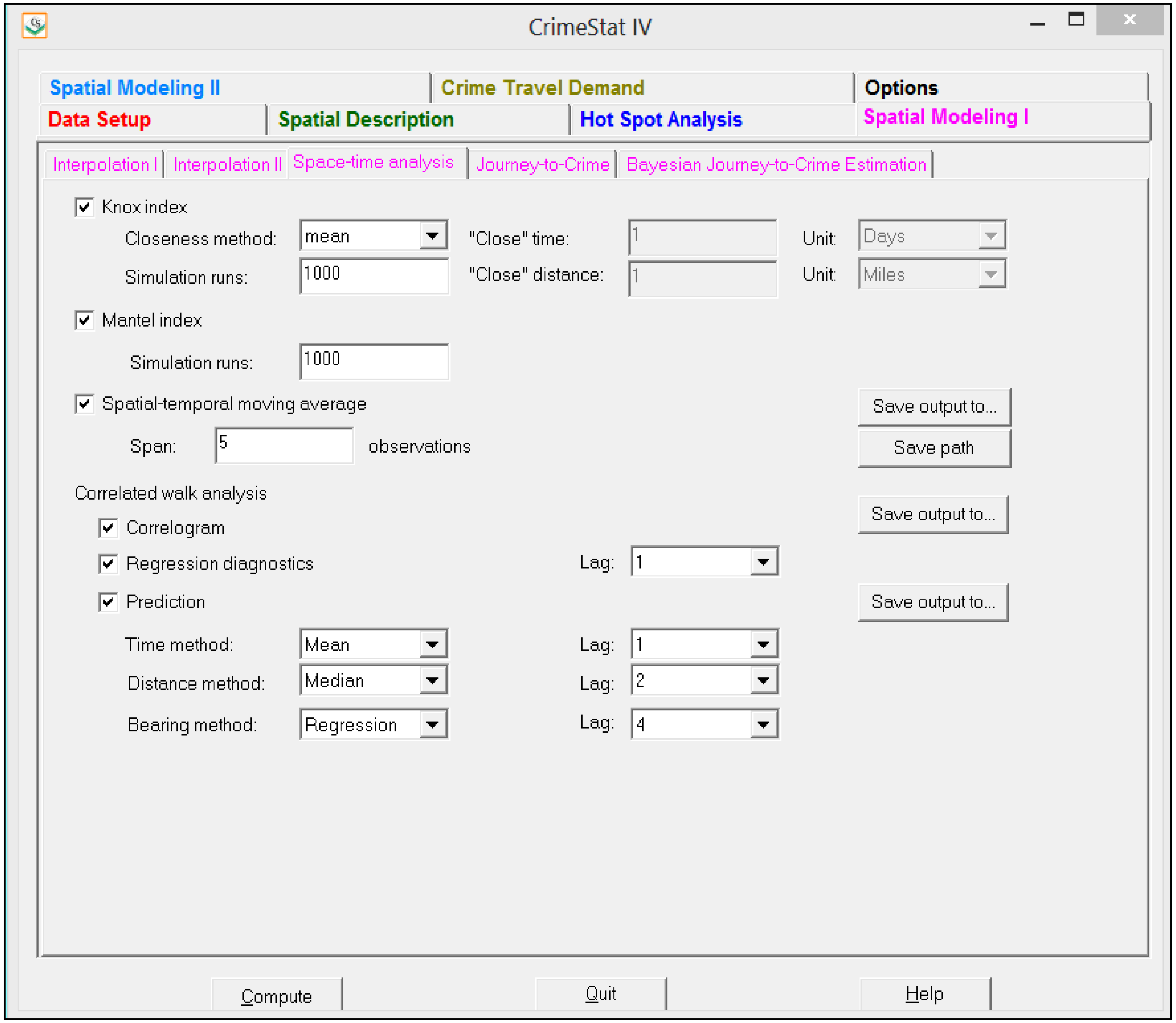
For each of these routines, time **must** be defined by an integer or real variable, and **not** by a formatted date. For example, 3 days, 2.1 weeks, 4.3 months, or the number of days from January 1, 1900 (e.g., 37174) are all eligible time values. 'November 1, 2001', '07/30/01' or '19th October, 2001' are not eligible values. Convert all formatted dates into a real number. Time units must be consistent across all observations (i.e., all values are hours or days or weeks or months or years, but not two or more these units). If these conditions are violated, *CrimeStat* will calculate results, but they won't be correct.

Knox Index

The Knox index is an index showing the relationship between 'closeness in time' and 'closeness in distance'. Pairs of events are compared in distance and in time and are represented

Figure 2.14:

Space-Time Analysis



☑ Knox index

Closeness method: mean | "Close" time: 1 | Unit: Days

Simulation runs: 1000 | "Close" distance: 1 | Unit: Miles

☑ Mantel index

Simulation runs: 1000

☑ Spatial-temporal moving average

Span: 5 observations

Correlated walk analysis

☑ Correlogram

☑ Regression diagnostics | Lag: 1

☑ Prediction

Time method: Mean | Lag: 1

Distance method: Median | Lag: 2

Bearing method: Regression | Lag: 4

as a 2 x 2 table. If there is a relationship, it would normally be positive, that is events that are close together in space (i.e., in distance) are also occurring in a short time span. There are three methods for defining closeness in time or in distance:

1. Mean. That is, events that are closer together than the mean time interval or are closer together than the mean distance are defined as 'Close' whereas events that are father together than mean time interval or are farther together than the mean distance are defined as 'Not close'. This is the default.

2. Median. That is, events that are closer together than the median time interval or are closer together than the median distance are defined as 'Close' whereas events that are father together than median time interval or are farther together than the median distance are defined as 'Not close'.

3. User defined. The user can specify any value for distinguishing 'Close' and 'Not close' for either time or distance.

The output includes a 2 x 2 table of the distribution of pairs categorized as 'Close' or 'Not close' in time and in distance. Note, that since pairs of events are being compared, there are N*(N-1)/2 pairs in a data set where N is the number of events. The output also includes a table of the expected of the distribution of pairs on the assumption that events in time are space are independent of each other. Finally, the output includes a chi-square test of the differences between the observed and expected distributions. Note, that since pairs are being compared, independence of observations is not true and a usual p-value associated with the chi-square test cannot be properly calculated.

Simulating confidence intervals

A Monte Carlo simulation can be run to estimate the approximate Type I error probability levels for the Knox index. The user specifies the number of simulation runs. Data are randomly assigned and the chi-square value for the Knox index is calculated for each run. The random output is sorted and percentiles are calculated. Twelve percentiles are identified for this index:

1. The minimum for the spatially random Knox chi-square
2. The maximum for the spatially random Knox chi-square
3. The 0.5 percentile for the spatially random Knox chi-square
4. The 1 percentiles for the spatially random Knox chi-square
5. The 2.5 percentile for the spatially random Knox chi-square
6. The 5 percentile for the spatially random Knox chi-square

7. The 10 percentile for the spatially random Knox chi-square
8. The 90 percentile for the spatially random Knox chi-square
9. The 95 percentile for the spatially random Knox chi-square
10. The 97.5 percentile for the spatially random Knox chi-square
11. The 99 percentile for the spatially random Knox chi-square
12. The 99.5 percentile for the spatially random Knox chi-square

Confidence intervals can be estimated from these percentiles. The two most commonly used ones are the 95% (defined by the 2.5 and 97.5 percentiles) and the 99% (defined by the 0.5 and 99.5 percentiles). The simulated data that is used can be viewed by checking the 'Dump simulation data' box on the Options tab.

Mantel Index

The Mantel index is the correlation between closeness in time and closeness in distance across pairs. Each pair of events is compared for the time interval and the distance between them. If there is a positive relationship between closeness in time and closeness in space (distance), then there should be a sizeable positive correlation between the two measures. Note, that since pairs of events are being compared, there are N*(N-1)/2 pairs in the data set where N is the number of events.

Simulating confidence intervals

A Monte Carlo simulation can be run to estimate the approximate confidence intervals around the Mantel correlation. The user specifies the number of simulation runs and the Mantel index is calculated for randomly assigned data. The random output is sorted and percentiles are calculated. Twelve percentiles are identified for this index:

1. The minimum for the spatially random Mantel index
2. The maximum for the spatially random Mantel index
3. The 0.5 percentile for the spatially random Mantel index
4. The 1 percentiles for the spatially random Mantel index
5. The 2.5 percentile for the spatially random Mantel index
6. The 5 percentile for the spatially random Mantel index
7. The 10 percentile for the spatially random Mantel index
8. The 90 percentile for the spatially random Mantel index
9. The 95 percentile for the spatially random Mantel index
10. The 97.5 percentile for the spatially random Mantel index
11. The 99 percentile for the spatially random Mantel index

12. The 99.5 percentile for the spatially random Mantel index

Confidence intervals can be estimated from these percentiles. The two most commonly used ones are the 95% (defined by the 2.5 and 97.5 percentiles) and the 99% (defined by the 0.5 and 99.5 percentiles). The simulated data that is used can be viewed by checking the 'Dump simulation data' box on the Options tab.

Spatial-Temporal Moving Average

This routine calculates the mean center as it changes over the sequence of the events. The routine sorts the incidents in the order in which they occur. The user defines a *span* of sequential incidents; the default is five observations. The routine places a window covering the span over the incidents and calculates the mean center (the mean X coordinate and the mean Y coordinate). It then moves the window one observation. Approximations are made at the beginning and end observations for the sequence. The result is a set of mean centers ordered from the first through last observations. This statistic is useful for identifying whether the central location for a set of incidents (perhaps committed by a serial offender) has moved over time.

There are four outputs for this routine:

1. The sample size
2. The number of observations making up the span
3. The span number
4. The X and Y coordinates for each span window.

The tabular results are output as a dBase ‘dbf’ file. 0020A line showing the sequential output cal also be output as an *ArcGIS* ‘shp’, *MapInfo* ‘mif’ or ‘bna’ ASCII formats. For MapInfo ‘mif’ format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected. The object will be output with a “STMA” prefix.

Correlated Walk Analysis

The Correlated Walk Analysis (CWA) analyzes the sequential movements of a serial offender and makes predictions about the time and location of the next event. Sequential movements are analyzed in terms of three parameters: Time difference between events (e.g., the number of days between two consecutive events), Distance between events – the distance between two consecutive events, and Bearing (direction) between events – the angular direction between two consecutive events in degrees (from 0 to 360).

There are three CWA routines for analyzing sequential events:

1. Correlogram (CWA-C)
2. Regression diagnostics (CWA-D)
3. Prediction (CWA-P)

Correlated Walk Analysis Correlogram (CWA-C)

The correlogram presents the lagged correlations between events for time difference, distance, and bearing (direction). The lags are the sequential comparisons. A lag of 0 is the sequence compared with itself; by definition, the correlation is 1.0. A lag of 1 is the sequence compared with the previous sequence. A lag of 2 is the sequence compared with two previous sequences. A lag of 3 is the sequence compared with three previous sequences, and so forth. In total, comparisons are made up to seven previous sequences (a lag of 7).

Typically, for time difference, distance and location separately, the lag with the highest correlation is the strongest. However, with each consecutive lag, the sample size decreases by one and a high correlation associated with a high lag comparison can be unreliable if the sample size is small. Consequently, the adjusted correlogram discounts the correlations by the number of lags.

The CWA correlogram is output as a dBase 'dbf' file.

Correlated Walk Analysis Regression Diagnostics (CWA-D)

The regression diagnostics presents the regression statistics for different lag models. The lag must be specified; the default is a lag of 1 (the sequential events compared with the previous events). Three regression models are run for time difference, direction, and bearing. The output includes statistics for:

1. The sample size
2. The distance and time units
3. The lag of the model (from 1 to 7)
4. The multiple R (correlation) between the lags
5. The squared multiple R (i.e., R-squared)
6. The standard error of estimate for the regression

7. The coefficient, standard error, t-value, and probability value (two-tail) for the constant.
8. The coefficient, standard error, t-value, and probability value (two-tail) for the coefficient.
9. The analysis of variance for the regression model, including the sum-of-squares and the mean-square error for the regression model and the residual (error), the F-test of the regression mean-square error divided by the residual mean-square error, and the probability level for the F-test.

In general, the model with the lowest standard error of estimate (and, consequently, highest multiple R) is best. However, with a small sample size, the model can be unreliable. Further, with each consecutive lag, the sample size decreases by one and a high multiple R associated with a high lag comparison can be unreliable if the sample size is small.

Correlated Walk Analysis Prediction (CWA-P)

The prediction routine allows the prediction of a next event, in time, distance, and direction. For each parameter – time difference, distance, and bearing, there are three models that can be used:

1. The mean difference (i.e., use mean time difference, mean distance, mean bearing)

2. The median difference (i.e., use median time difference, median distance, median bearing)

3. The regression model (i.e., use the estimated regression coefficient and intercept)

For each of these, a different lag comparison can be used, from 1 to 7. The lag defines the sequence from which the prediction is made. Thus, for a lag of 1, the interval from the next-to-last to the last event is used as a reference (i.e., between events N-1 and N); for a lag of 2, the interval from the third-to-last to the next-to-last event is used as a reference (i.e.,between events N-2 and N-1); and so forth. The particular model selected is then added to the reference sequence.

Example 1: with a lag of 1 and the use of the mean difference, the mean time difference is added to the time of the last event, the mean distance is added to the location of the last event, and the mean bearing is added to the location of the last event.

Example 2: with a lag of 2 and the use of the regression model, the predicted time difference is added to the time of the next-to-last event; the predicted distance is added to the location of the next-to-last event and the prediction bearing is added to the location of the last event. Note: if the regression model is used, the lag for distance and bearing must be the same.

Example 3: with a lag of 1 for time, a lag of 2 for distance and the use of the mean distance, and a lag of 3 for bearing and the use of the median bearing, the predicted time difference is added to the last event, the mean distance is added to the location of the next-to-last event, and the median bearing is added to the location of the third-from-last event.

Tabular output

The tabular output includes:

1. The method used for time, distance, and bearing
2. The lag used for time, distance, and bearing
3. The predicted time difference
4. The predicted distance
5. The predicted bearing
6. The final predicted time
7. The X-coordinate of the final predicted location
8. The Y-coordinate of the final predicted location

Graphical output

If the user specifies an output file name, there are five graphical objects that are output as an *ArcGIS* 'shp', *MapInfo* 'mif' or various ASCII formats:

1. The sequence of incidents from the first to the last. This object has a prefix of 'Events' before the file name provided by the user.
2. The predicted location of the next event. This is the event after the last in the input sequence. This object has a prefix of 'Preddest' before the file name.
3. The predicted path between the last event in the sequence and the expected next event. This object has a prefix of 'Pw' before the file name.
4. The center of minimum distance for the sequence of events. This is the single best measure of the likely origin location of the offender

5. The expected path between the center of minimum distance and the predicted location of the next event. This is a guess about the likely origin and likely destination for a next event by the offender.

For MapInfo ‘mif’ format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Journey to Crime Estimation (Jtc)

The journey to crime (Jtc) routine estimates the likelihood that a serial offender lives at any location within the study area. Both a primary file and a reference file are required. The locations of the serial crimes are defined in the primary file while all locations within the study area are identified in the reference file. The Jtc routine can use two different travel distance functions: 1) An already-calibrated distance function; and 2) A mathematical formula. Either direct or indirect (Manhattan) distances can be used though the default is direct (see Measurement parameters.)

Calibrate Journey to Crime Function

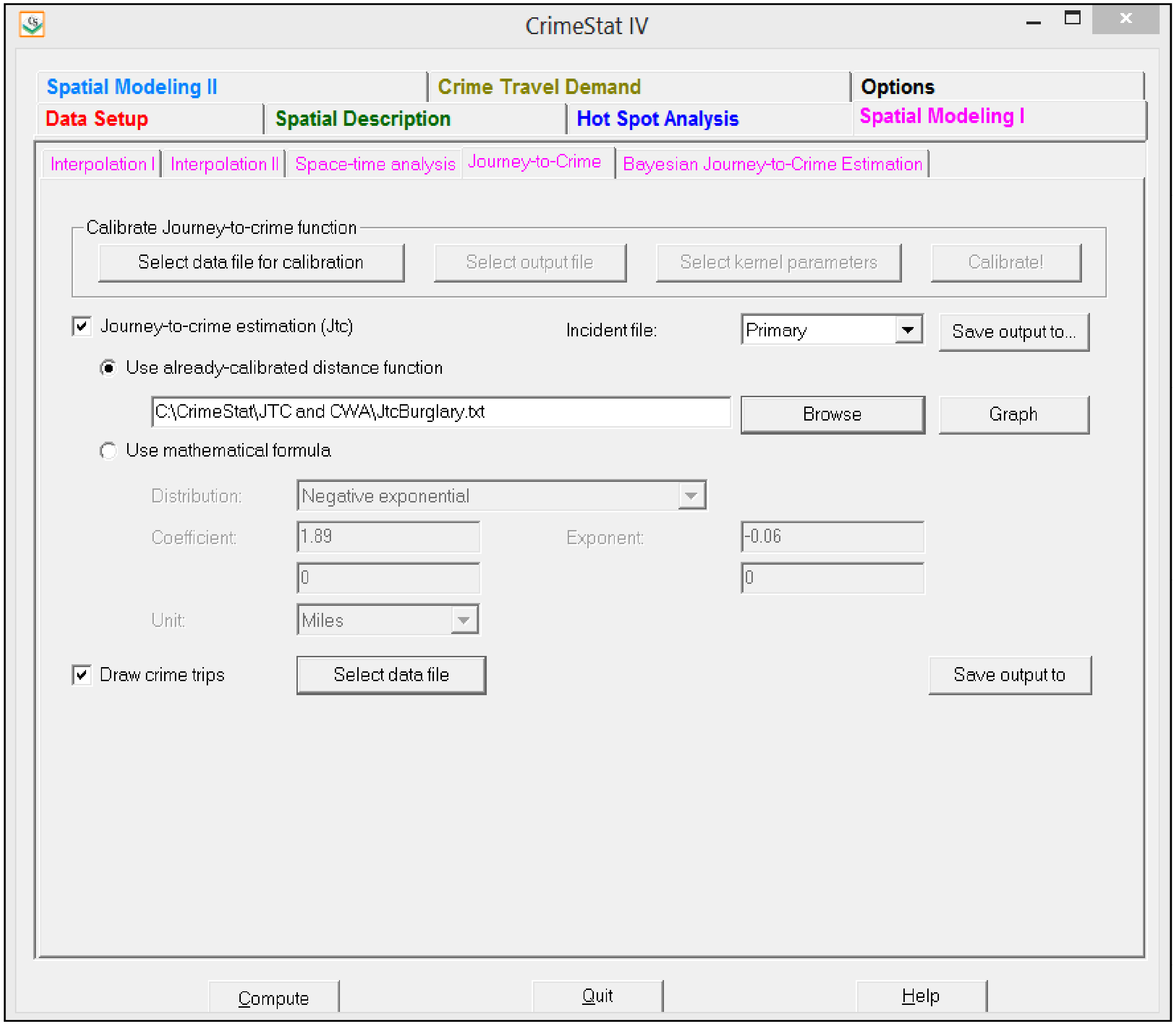
This routine calibrates a journey to crime distance function for use in the estimation routine. A file is input which has a set of incidents (records) that includes both the X and Y coordinates for the location of the offender's residence (origin) and the X and Y coordinates for the location of the incident that the offender committed (destination.) The routine estimates a travel distance function (trip lengths) using a one-dimensional kernel density method. For each record, the distance between the origin location and the destination location is calculated and is represented on a distance scale. The maximum distance is calculated and divided into a number of intervals; the default is 100 equal sized intervals, but the user can modify this. For each distance (point) calculated, a one-dimensional kernel is overlaid. For each distance interval, the values of all kernels are summed to produce a smooth function of journey to crime distance. The results are saved to a file that can be used in the journey to crime estimation routine.

Select data file for calibration

Select the file that has the X and Y coordinates for the origin and destination locations. *CrimeStat* reads dbase ‘dbf’, ArcGIS point ‘shp’ and ASCII files. Select the tab and specify 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 2.15:

Journey-to-crime Analysis



☑ Journey-to-crime estimation (Jtc)

Incident file: Primary

● Use already-calibrated distance function

C:\CrimeStat\JTC and CWA\JtcBurglary.txt

○ Use mathematical formula

Distribution: Negative exponential

Coefficient: 1.89

Exponent: -0.06

0

0

Unit: Miles

☑ Draw crime trips

Variables

Define the file which contains the X and Y coordinates for both the origin (residence) and destination (crime) locations

Columns

Select the variables for the X and Y coordinates respectively for *both* the origin and destination locations (e.g., Lon, Lat, HomeX, HomeY, IncidentX, IncidentY.) Both locations must be defined for the routine to work.

Missing values

Identify whether there are any missing values for these four fields (X and Y coordinates for both origin and destination locations). By default, *CrimeStat* will ignore records with blank values in any of the eligible fields or records with non-numeric values (e.g., alphanumeric characters, , *). Blanks will always be excluded unless the user selects **<*none*>**. There are 8 possible options:

1. <*blank*> fields are automatically excluded. This is the default
2. <*none*> indicates that no records will be excluded. If there is a blank field, *CrimeStat* will treat it as a 0
3. *0* is excluded
4. *–1* is excluded
5. *0 and –1* indicates that both 0 and -1 will be excluded
6. *0, -1 and 9999* indicates that all three values (0, -1, 9999) will be excluded
7. *Any* other numerical value can be treated as a missing value by typing it (e.g., 99)
8. *Multiple* numerical values can be treated as missing values by typing them, separating each by commas (e.g., 0, -1, 99, 9999, -99)

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.) Directional coordinates are not allowed for this routine.

Select kernel parameters

There are five parameters that must be defined.

Method of interpolation

There are five types of kernel distributions that can be used to estimate the distance decay density of the trip lengths:

1. The **normal** kernel overlays a three-dimensional normal distribution over each point that then extends over the area defined by the reference file. This is the default kernel function.

2. The **uniform** kernel overlays a uniform function over each point that only extends for a limited distance.

3. The **quartic** kernel overlays a quartic function over each point that only extends for a limited distance.

4. The **triangular** kernel overlays a three-dimensional triangle over each point that only extends for a limited distance.

5. The **negative exponential** kernel overlays a three dimensional negative exponential function over each point that only extends for a limited distance

The methods produce similar results though 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 uniform, quartic, triangular and negative exponential kernels, bandwidth is the radius of a circle defined by the surface. For all types, larger bandwidth will produce smoother density estimates and both adaptive and fixed bandwidth intervals can be selected.

Fixed bandwidth

A fixed bandwidth distance is a fixed interval for each point. The user must define the interval, the interval size, and the distance units by which it is calculated (miles, nautical miles, feet, kilometers, or meters). The default bandwidth setting is fixed with intervals of 0.25 miles each. The interval size can be changed.

Adaptive bandwidth

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

Specify interpolation bins

The interpolation bins are defined in one of two ways:

1. By the number of bins. The maximum distance calculated is divided by the number of specified bins.This is the default with 100 bins. The user can change the number of bins

2. By the distance between bins. The user can specify a bin width in miles, nautical miles, feet, kilometers, and meters

Output (areal) units

Specify the areal density units as points per mile, nautical mile, foot, kilometer, or meter. The default is points per mile.

Calculate densities or probabilities

The density estimate for each cell can be calculated in one of three ways:

1. **Absolute densities**. This is the number of points per grid cell and is scaled so that the sum of all grid cells equals the sample size. This is the default.

2. **Relative densities**. For each grid cell, this is the absolute density divided by the grid cell area and is expressed in the output units (e.g., points per square mile)

3. **Probabilities**. This is the proportion of all incidents that occur in the grid cell. The sum of all grid cells equals a probability of 1.

Select whether absolute densities, relative densities, or probabilities are to be output for each cell. The default is absolute densities.

Select output file

The output *must* be saved to a file. *CrimeStat* can save the calibration output to either a dbase 'dbf' or ASCII text 'txt' file.

Calibrate!

Click on 'Calibrate!' to run the routine. The output is saved to the specified file upon clicking on 'Close'. The output file is saved as JtcCalib<file name> with the file name being provided by the user.

Graph of journey to crime travel function

Click on 'View graph' to see the journey crime travel distance function (journey to crime likelihood by distance.) The screen view can be printed by clicking on 'Print'. For a better quality graph, however, the output should be imported into a graphics package.

Journey to Crime Estimation (Jtc)

The journey to crime (Jtc) routine estimates the likelihood that a serial offender lives at any location within the study area. Both a primary file and a reference file are required. The locations of the serial crimes are defined in the primary file while all locations within the study area are identified in the reference file. The Jtc routine can use two different travel distance functions: 1) An already-calibrated distance function; and 2) A mathematical formula.

Use an already-calibrated distance function

If a travel distance function has already been calibrated (see 'Calibrate journey to crime function'), the file can be directly input into the Jtc routine.

Input

The user selects the name of the already-calibrated travel distance function. *CrimeStat* reads dbase 'dbf'', ArcGIS 'shp' and ASCII text files.

Output

The Jtc routine calculates a relative likelihood estimate for each cell of the reference file. Higher values indicate higher relative likelihoods. The results can be output as a *Surfer for Windows* file (for both an external or generated reference file) or as an *ArcGIS* 'shp', *MapInfo* 'mif', *ArcGIS Spatial Analyst* 'asc', or ASCII grid 'grd' file (only if the reference file is generated by *CrimeStat).* For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

The output file is saved as Jtc<file name> with the file name being provided by the user.

Use a mathematical formula

A mathematical formula can be used instead of a calibrated distance function. To do this, it is necessary to specify the type of distribution. There are five mathematical models that can be selected:

1. Negative exponential
2. Normal
3. Lognormal
4. Linear
5. Truncated negative exponential

The normal is the default. For each mathematical model, two or three different parameters must be defined:

1. For the negative exponential, the coefficient and exponent
2. For the normal distribution, the mean distance, standard deviation and coefficient
3. For the lognormal distribution, the mean distance, standard deviation and coefficient
4. For the linear distribution, an intercept and slope
5. For the truncated negative exponential, a peak distance, peak likelihood, intercept, and exponent

Output

The Jtc estimation routine calculates a relative likelihood estimate for each cell of the reference file. Higher values indicate higher relative likelihoods. The results can be output as a *Surfer for Windows* file (for both an external or generated reference file) or as an *ArcGIS* 'shp', *MapInfo* 'mif', *ArcGIS Spatial Analyst* 'asc', or ASCII grid 'grd' file (only if the reference file is generated by *CrimeStat*). For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

The output file is saved as Jtc<file name> with the file name being provided by the user.

Draw Crime Trips

This routine is a utility for both the Journey to crime routine and the Trip Distribution routine (in the Crime Travel Demand module). If given a file with origins and destinations, the routine will draw a line between the origin and destination for each record. It is useful for examining the actual trip links made by an offender.

Select data file

Select the file that has the X and Y coordinates for the origin and destination locations. *CrimeStat* reads dbase 'dbf', ArcGIS point 'shp' and ASCII files. Select the tab and specify 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 for both the origin (residence) and destination (crime) locations

Columns

Select the variables for the X and Y coordinates respectively for *both* the origin and destination locations (e.g., Lon, Lat, HomeX, HomeY, IncidentX, IncidentY.) Both locations must be defined for the routine to work.

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.) Directional coordinates are not allowed for this routine.

Save output to

The graphical results can be output as lines in *ArcGIS* 'shp', *MapInfo* 'mif' or various ASCII formats. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Bayesian Journey to Crime Estimation (Jtc)

The Bayesian Journey-to-crime module (BJtc) is a tool for estimating the likely residence location of a serial offender. It is an extension of the Journey-to-crime routine (Jtc) which uses a travel distance function to make guesses about the likely residence location. The Bayesian Journey to crime routine estimates the likelihood that a serial offender lives at any location within the study area using two pieces of information: 1) the distribution of incidents committed by the offender; and 2) the distribution of origins by other offenders who committed crimes in the same location as the offender, based on an origin-destination matrix.

A travel distance function is applied to the distribution of incidents to produce one estimate of the likely origin of the offender while an origin-destination matrix is used to produce another estimate of the likely origin of the offender based on the origins of other offenders who committed crimes in the same locations. Both estimates can be combined in several ways to produce a joint estimate of the likely origin of the offender.

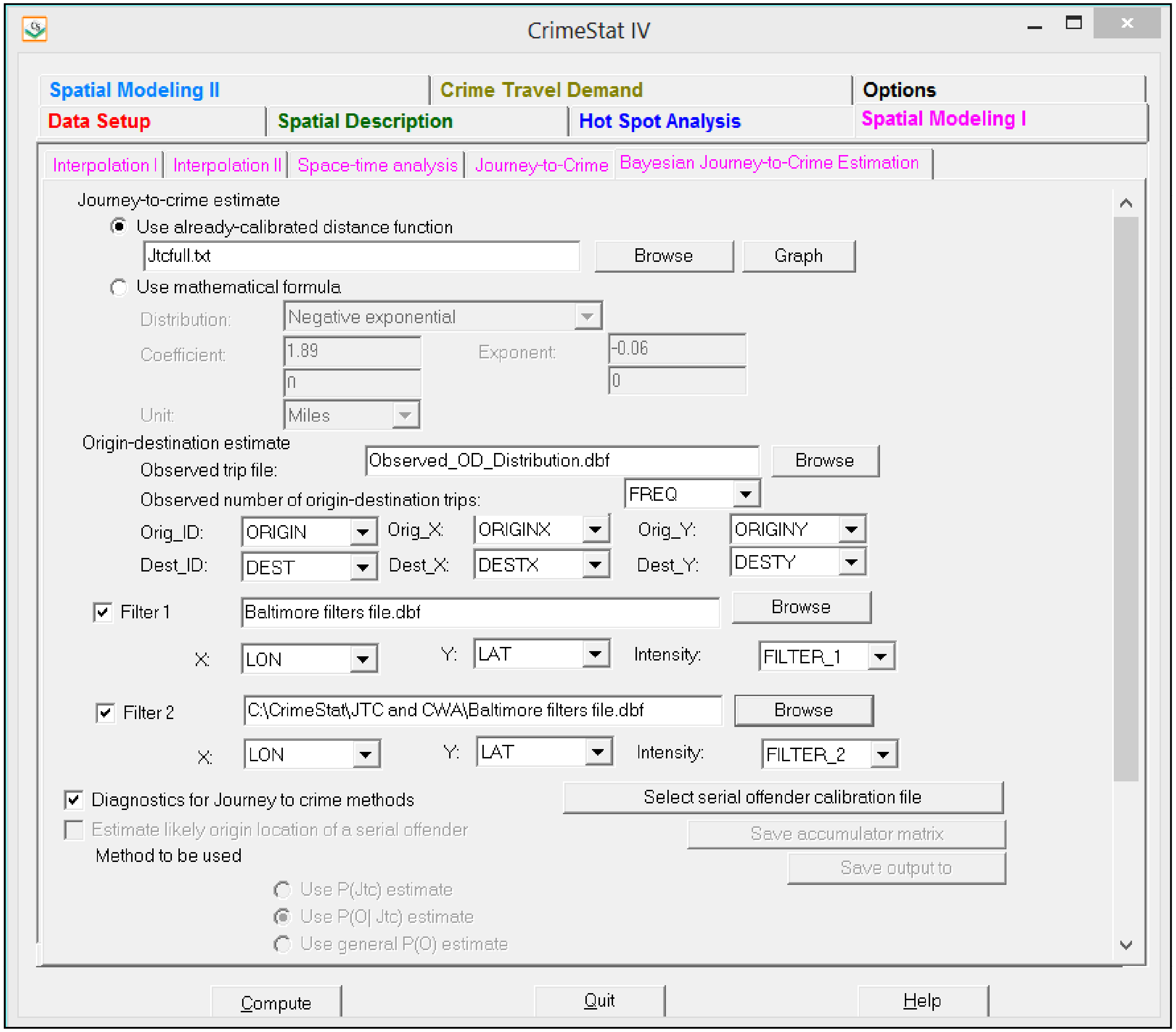
There are two routines in the Bayesian Journey to Crime module:

1. Diagnostics for comparing different journey-to-crime methods; and

2. A routine for estimating the likely origin of a serial offender using a selected journey-to-crime method.

Figure 2.16:

Bayesian Journey-to-crime Analysis



CrimeStat IV

Journey-to-crime estimate

● Use already-calibrated distance function

Jtcfull.txt

○ Use mathematical formula

Distribution: Negative exponential

Coefficient: 1.89

n

Exponent: -0.06

0

Unit: Miles

Origin-destination estimate

Observed trip file: Observed_OD_Distribution.dbf

Observed number of origin-destination trips: FREQ

Orig_ID: ORIGIN Orig_X: ORIGINX Orig_Y: ORIGINY

Dest_ID: DEST Dest_X: DESTX Dest_Y: DESTY

☑ Filter 1 Baltimore filters file.dbf

X: LON Y: LAT Intensity: FILTER_1

☑ Filter 2 C:\CrimeStat\JTC and CWA\Baltimore filters file.dbf

X: LON Y: LAT Intensity: FILTER_2

☑ Diagnostics for Journey to crime methods

☐ Estimate likely origin location of a serial offender

Method to be used

○ Use P(Jtc) estimate

● Use P(O| Jtc) estimate

○ Use general P(O) estimate

The routines are applications of Bayes Theorem to Journey to Crime estimation.

Bayes Theorem

Bayes Theorem is defined as:

$$P(B|A) = \frac{P(B)*P(A|B)}{P(A)} \qquad (2.1)$$

where P(B|A) is the probability of event B given event A (the conditional probability of B given A), P(B) is the simple probability of event B, P(A|B) is the probability of event A given event B (the conditional probability of A given B), and P(A) is the probability of event A.

Bayesian Inference

In the statistical interpretation of Bayes Theorem, the probabilities are estimates of a random variable. Let θ be a parameter of interest and let X be some data. Thus, Bayes Theorem can be expressed as:

$$P(\theta|X) = \frac{P(\theta)*P(X|\theta)}{P(X)} \qquad (2.2)$$

where P(θ|X) is the posterior probability of θ given the data, *X*, and P(θ) is the probability that θ has a certain distribution and is often called the *prior probability*. P(X|θ) is the probability that the data would be obtained given that θ is true and is often called the *likelihood function* (i.e., it is the likelihood that the data will be obtained given the distribution of θ). Finally, P(X) is the marginal probability of the data, the probability of obtaining the data under all possible scenarios; essentially, it is the data.

The equation can be rephrased in logical terms:

$$\text{The posterior probability that } \theta \text{ is true given the data, X} = \frac{\text{Likelihood of obtaining the data given } \theta \text{ is true} \; * \; \text{Prior probability of } \theta}{\text{Marginal probability of X}} \qquad (2.3)$$

In other words, this formulation allows an estimate of the probability of a particular parameter, θ, to be updated given new information. Since θ is the prior probability of an event, given some new data, X, Bayes Theorem can be used to update the estimate of θ. The prior probability of θ can come from prior studies, an assumption of no difference between any of the conditions affecting θ, or an assumed mathematical distribution. The likelihood function can also come from empirical studies or an assumed mathematical function. Irrespective of how these are interpreted, the result is an estimate of the parameter, θ, given the evidence, X. This is called the *posterior probability* (or posterior distribution).

Application of Bayesian inference to Journey to Crime Estimation

Applying Bayesian inference to journey to crime estimation, there are three different estimates of where an offender lives:

1. An estimate of the residence location of a single offender based on the location of the incidents that this person committed and an assumed travel distance function, P(Jtc).

2. An estimate of the residence location of a single offender based on a general distribution of all offenders, irrespective of any particular destinations for incidents, P(O). Essentially, this is the distribution of origins irrespective of the destinations.

3. An estimate of the residence location of a single offender based on the distribution of offenders given the distribution of incidents committed by the single offender, P(O|Jtc).

The Bayesian formula can now be approximated by:

$$P(Jtc|O) \approx \frac{P(O|Jtc)*P(Jtc)}{P(O)} \qquad (2.4)$$

where P(Jtc|O) is the probability that a particular serial offender lives at any one location given both an estimate of where the offender lives given a travel distance function and an estimate of where an offender lives given the distribution of origins by other offenders who committed crimes in the same locations. The numerator expresses this relationship and is called the *Bayesian product term*. Since obtaining the probability of the data under all scenarios is virtually impossible to estimate, the equation is an approximation, relating this product term to the distribution of all offenders, P(O). This is called *Bayesian risk*.

The Bayesian Journey to Crime Estimation Module

The Bayesian Journey-to-crime estimation module is made up of two routines, one for diagnosing which Journey-to-crime method is best and one for applying that method to a particular serial offender.

Data Preparation for Bayesian Journey to Crime Estimation

There are three sets of data that are required and one optional data set. The three required ones are:

1. The incidents committed by a single offender for which an estimate will be made of where that individual lives

2. A journey-to-crime function that estimates the likelihood of an offender committing crimes at a certain distance (or travel time if a network is used)

3. An origin-destination matrix

The fourth, optional data set is a diagnostics file of multiple known serial offenders for which both their residence and crime locations are known.

Both a primary file and a reference file are also required. For the Bayesian Jtc Diagnostics routine, any point file can be used as the primary file. For the Bayesian Jtc Estimation routine, the primary file should be the locations of the crimes committed by the single serial offender for whom the estimate is being obtained. The reference file also needs to be defined and should include all locations where crimes have been committed (see Reference File).

Serial offender data

For each serial offender for whom an estimate will be made of where that person lives, the data set should include the location of the incidents committed by the offender. The data are set up as a series of records in which each record represents a single event. On each data set, there are X and Y coordinates identifying the location of the incidents this person has committed.

Journey-to-crime travel function

The Journey-to-crime function is an estimate of the likelihood of an offender traveling a certain distance. Typically, it represents a frequency distribution of distances traveled, though it could be a frequency distribution of travel times if a network was used to calibrate the function with the Journey to crime estimation routine (see Journey to crime estimation). It can come from an a priori assumption about travel distances, prior research, or a calibration data set of offenders who have already been caught. The "Calibrate journey-to-crime function" routine (on the Journey-to-crime page under Spatial modeling) can be used to estimate this function.

The BJtc routine can use two different travel distance functions: 1) An already-calibrated distance function; and 2) A mathematical formula. Either direct or indirect (Manhattan) distances can be used though the default is direct (see Measurement parameters.)

Origin-destination matrix

The origin-destination matrix relates the number of offenders who commit crimes in one of N zones who live (originate) in one of M zones. It can be created from the "Calculate observed origin-destination trips" routine (on the 'Describe origin-destination trips' page under the Trip distribution module of the Crime Travel Demand model).

Diagnostics file for Bayesian Jtc routine

The aim of the diagnostics file is to provide information to the analyst about which of several parameters (to be described below) are best at guessing where an offender lives. The assumption is that if a particular parameter was best with the K offenders in a diagnostics file in which the residence location was known, then the same parameter will also be best for a serial offender for whom the residence location is not known.

How many serial offenders are needed to make up a diagnostics file? There is no simple answer to this. Clearly, the more, the better since the aim is to identify which parameter is most sensitive with a certain level of precision and accuracy. Certainly, a minimum of 10 would be necessary. But, more would certainly be more accurate. Further, the offender records used in the diagnostics file should be similar in other dimensions to the offender that is being tracked. However, this may be impractical.

Once the data sets have been collected, they need to be placed in an *appended* file, with one serial offender on top of another. Each record has to represent a single incident. Further, the records have to be arranged sequentially with all the records for a single offender being grouped together. The routine automatically sorts the data by the offender ID. But, to be sure that the result is consistent, the data should be prepared in this way.

Regarding the fields in each record, at the minimum there is a need for an ID field, and the X and Y coordinates of both the crime location and the residence location. The ID field is any string variable.

Diagnostics for Journey to Crime Methods

The following applies to the "diagnostics" routine only.

Data Input

The user inputs the five required data sets and two optional data sets.

1. Any primary file with an X and Y location. A suggestion is to use one of the files for the serial offender, but this is not essential
2. A grid that will be overlaid on the study area. Use the Reference File under Data setup to define the X and Y coordinates of the lower-left and upper-right corners of the grid as well as the number of columns
3. A journey-to-crime function that estimates the likelihood of an offender committing crimes at a certain distance (or travel time if a network is used)
4. An origin-destination matrix
5. The diagnostics file of known serial offenders in which both their residence and crime locations are known
6. (Optional) A data set that includes a filter variable (see below)
7. (Optional) A data set that includes a second filter variable (see below)

Methods Tested

The "diagnostics" routine compares seven methods for estimating the likely location of a serial offender:

1. The Journey-to-crime distance method, P(Jtc).
2. The general crime distribution based on the origin-destination matrix, P(O). Essentially, this is the distribution of origins irrespective of the destinations.
3. The distribution of origins based only on the incidents committed by the serial offender, P(O|Jtc).
4. The product of the Journey-to-crime estimate (1 above) and the distribution of origins based only on the incidents committed by the serial offender (3 above), P(Jtc)*P(O|Jtc). This is the numerator of the Bayesian function, the product of the prior probability times the likelihood estimate.

5. The simple average of the Journey-to-crime estimate (1 above) and the distribution of origins based only on the distribution of incidents committed by the serial offender (3 above), P(Jtc) + P(O|Jtc). This is an alternative to the product term (4 above).
6. The Bayesian risk estimate as indicated in the discussion above (method 4 above divided by method 2 above), P(Bayes risk).
7. The center of minimum distance, Cmd. Previous research has indicated that the center of minimum of distance produces the least error in minimizing the distance between where the method predicts the most likely location for the offender and where the offender actually lives.

Interpolated Grid

With each serial offender, in turn, and with each method, the routine overlays a grid over the study area. The grid is defined by the Reference File parameters (see Data setup). The routine then interpolates each input data set into a probability estimate for each grid cell with the sum of the cells equaling 1.0 (within three decimal places). The manner in which the interpolation is done varies by the method:

1. For the Journey-to-crime method, P(Jtc), the routine interpolates the selected distance function to each grid cell to produce a density estimate. The density estimates are converted to probabilities so that the sum of the grid cells equals 1.0.

2. For the general crime distribution method, P(O), the routine sums up the incidents by each origin zone from the origin-destination matrix and interpolates that using the normal distribution method of the single kernel density routine (see Kernel Density Interpolation). The density estimates are converted to probabilities so that the sum of the grid cells equals 1.0.

3. For the distribution of origins based only on the incidents committed by the serial offender, from the origin-destination matrix the routine identifies the zone in which the incidents occur and reads only those origins associated with those destination zones. Multiple incidents committed in the same origin zone are counted multiple times. The routine then uses the single kernel density routine to interpolate the distribution to the grid (see Kernel Density Interpolation). The density estimates are converted to probabilities so that the sum of the grid cells equals 1.0.

4. For the product of the Journey-to-crime estimate and the distribution of origins based only on the incidents committed by the serial offender, the routine multiples the

probability estimate obtained in 1 above by the probability estimate obtained in 3 above. The product density estimates are converted to probabilities so that the sum of the grid cells equals 1.0.

5. For the simple average of the Journey-to-crime estimate and the distribution of origins based only on the incidents committed by the serial offender, the routine adds the probability estimate obtained in 1 above to the probability estimate obtained in 3 above and divides by two. The average density estimates are converted to probabilities so that the sum of the grid cells equals 1.0.

6. For the Bayesian risk estimate, the routine takes the product estimate (4 above) and divides it by the general crime distribution estimate (2 above). The resulting density estimates are converted to probabilities so that the sum of the grid cells equals 1.0.

7. Finally, for the center of minimum distance estimate, the routine calculates the center of minimum distance for each serial offender in the "diagnostics" file and calculates the distance between this statistic and the location where the offender is actually residing. This is used only for the distance error comparisons.

Note in all of the probability estimates (excluding 7), the cells are converted to probabilities prior to any multiplication or division. The results are then re-scaled so that the resulting grid is a probability (i.e., all cells sum to 1.0).

Additional Filtering

A filter is a probability matrix that is applied to the estimate but is not conditioned on the existing variables in the model. For example, an opportunity matrix that was independent of the distribution of offences by a single serial offender or the origins of other offenders who committed crimes in the same locations could be applied as an alternative (equation 14.14):

$$P(Jtc|O) \propto P(Jtc) * P(O|Jtc) * P(A) \tag{2.5}$$

In this case, P(A) is an independent matrix. Another filter that could be applied is residential land use. The vast majority of offenders are going to live in residential areas. Thus, a residential land use filter estimates the probability of a residential land use for every cell, P(A), could be applied to screen out cells that are not residential, such as

$$P(Jtc|O) \propto P(Jtc) * P(O|Jtc) * P(A) \tag{2.6}$$

In this way, additional information can be integrated into the methodology to improve the accuracy and precision of the estimates. Clearly, having additional variables be conditioned upon existing variables in the model would be ideal since that would fit the true Bayesian approach. But, even if independent filters were brought in, the model could be improved.

Defining up to two filters

The Bayesian Journey-to-crime routine allows the addition of up to two filters, called **F1** and **F2**. If one filter variable is defined as a data set, then F1 will be applied to the probability components. If two filter variables are defined as data sets, then both F1 and F2 will be applied simultaneously to the probability components.

Output of Routine

For each offender in the “diagnostics” file, the routine calculates three different statistics for each of the methods:

1. The estimated **probability** in the cell where the offender actually lives. It does this by, first, identifying the grid cell in which the offender lives (i.e., the grid cell where the offender’s residence X and Y coordinate is found) and, second, by noting the probability associated with that grid cell. The higher the probability, the better the estimate.

2. The **percentile** of all grid cells in the entire grid that have to be searched to find the cell where the offender lives based on the probability estimate from 1, ranked from those with the highest probability to the lowest. Obviously, this percentile will vary by how large a reference grid is used (e.g., with a very large reference grid, the percentile where the offender actually lives will be small whereas with a small reference grid, the percentile will be larger). But, since the purpose is to compare methods, the actual percentage should be treated as a relative index. The result is sorted from low to high so that the smaller the percentile, the better. For example, a percentile of 1% indicates that the probability estimate for the cell where the offender lives is within the top 1% of all grid cells. Conversely, a percentile of 30% indicates that the probability estimate for the cell where the offender lives in within the top 30% of all grid cell.

3. The **distance** between the cell with the highest probability and the cell where the offender lives. The smaller the distance between the cell with the highest probability and the cell where the offender lives, the better.

Output matrices

The "diagnostics" routine outputs two separate matrices. The probability estimates (numbers 1 and 2 above) are presented in a separate matrix from the distance estimates (number 3 above). The user can save the total output as a text file or can copy and paste each of the two output matrices into a spreadsheet separately. We recommend the copying-and-pasting method into a spreadsheet as it will be difficult to line up differing column widths for the two matrices and summary tables at the bottom of each.

Summary statistics

The "diagnostics" routine will also provide summary information at the bottom of each matrix. For the probability matrix, these include:

1. The mean (probability or percentile)
2. The median (probability or percentile)
3. The standard deviation (probability or percentile)
4. The number of times the Jtc estimate produces the highest probability
5. The number of times the O|Jtc estimate produces the highest probability
6. The number of times the O estimate produces the highest probability
7. The number of times the "product" term estimate produces the highest probability
8. The number of times the Bayesian risk estimate produces the highest probability
9. If filter variable F1 has been defined:
   A. The number of times the Jtc*F1 estimate produces the highest probability
   B. The number of times the O|Jtc*F1 estimate produces the highest probability
   C. The number of times the "Product"*F1 estimate produces the highest probability
   D. The number of times the Bayesian risk*F1 estimate produces the highest probability
10. If both filter variable F1 and filter variable F2 have been defined:
    A. The number of times the Jtc*F1*F2 estimate produces the highest probability
    B. The number of times the O|Jtc*F1*F2 estimate produces the highest probability
    C. The number of times the "Product"*F1*F2 estimate produces the highest probability
    D. The number of times the Bayesian risk*F1*F2 estimate produces the highest probability

For the distance matrix, these include:

1. The mean distance
2. The median distance
3. The standard deviation distance
4. The number of times the Jtc estimate produces the closest distance
5. The number of times the O|Jtc estimate produces the closest distance
6. The number of times the O estimate produces the closest distance
7. The number of times the "product" term estimate produces the closest distance
8. The number of times the Bayesian risk estimate produces the closest distance
9. The number of times the center of minimum distance (CMD) produces the closest distance
10. *If* filter variable F1 has been defined:
    A. The number of times the Jtc*F1 estimate produces the closest distance
    B. The number of times the O|Jtc*F1 estimate produces the closest distance
    C. The number of times the "Product"*F1 estimate produces the closest distance
    D. The number of times the Bayesian risk*F1 estimate produces the closest distance
11. *If* both filter variable F1 and filter variable F2 have been defined:
    A. The number of times the Jtc*F1*F2 estimate produces the closest distance
    B. The number of times the O|Jtc*F1*F2 estimate produces the closest distance
    C. The number of times the "Product"*F1*F2 estimate produces the closest distance
    D. The number of times the Bayesian risk*F1*F2 estimate produces the closest distance

These statistics, especially the summary statistics, should indicate which of the methods produces the best accuracy, defined in terms of highest probability (for the probability matrix) and closest distance (for the distance matrix), and efficiency, defined in terms of the smallest search area to locate the serial offender.

Estimate Likely Origin Location of a Serial Offender

The following applies to the Bayesian Jtc "Estimate likely origin of a serial offender" routine. Once the "diagnostic" routine has been run and a preferred method selected, the next routine allows the application of that method to a single serial offender.

Data Input

The user inputs the three required data sets and a reference file grid. The two filter variables can also be applied, but are optional

1. The incidents committed by a single offender that we're interested in catching. This must be the primary file.
2. A journey-to-crime function that estimates the likelihood of an offender committing crimes at a certain distance (or travel time if a network is used).
3. An origin-destination matrix.
4. The reference file also needs to be defined and should include all locations where crimes have been committed (see Reference File).
5. (Optional) A data set that includes a filter variable (see above).
6. (Optional) A data set that includes a second filter variable (see above).

Methods Tested

The Bayesian Jtc "Estimate" routine interpolates the incidents committed by the serial offender to a grid, allowing the user to estimate where the offender is liable to live. There are 13 different methods for estimating the likely location of a serial offender that can be used, depending on whether filter variables are used or not. However, the user has to choose only one of these:

1. The Journey-to-crime distance method, P(Jtc).

2. The general crime distribution based on the origin-destination matrix, P(O). Essentially, this is the distribution of origins irrespective of the destinations.

3. The distribution of origins based only on the incidents committed by the serial offender, P(O|Jtc).
4. The product of the Journey-to-crime estimate (1 above) and the distribution of origins based only on the incidents committed by the serial offender (3 above), P(Jtc)*P(O|Jtc). This is the numerator of the Bayesian function discussed above, the product of the prior probability times the likelihood estimate.

5. The weighted average of the Journey-to-crime estimate (1 above) and the distribution of origins based only on the distribution of incidents committed by the serial offender (3 above), P(Jtc) + P(O|Jtc). This is an alternative to the product term (4 above). The user must select weights for each of the two estimates such

that the sum of the weights equal 1.0. The default weights are 0.5 for each estimate.

6. The Bayesian risk estimate as indicated above (method 4 above divided by method 2 above), P(Bayes risk).
7. *If* one filtering variable, F1, has been used:

   A. P(Jtc)*F1
   B. P(O|Jtc)*F1
   C. "Product"*F1
   D. Bayesian risk*F1

8. *If* two filtering variables have been used:

   A. P(Jtc)*F1*F2
   B. P(O|Jtc)*F1*F2
   C. "Product"*F1*F2
   D. Bayesian risk*F1*F2

Interpolated Grid

For the estimation method that is selected, the routine overlays a grid on the study area. The grid is defined by the reference file parameters (see Reference File). The routine then interpolates the input data set (the primary file) into a probability estimate for each grid cell with the sum of the cells equaling 1.0 (within three decimal places). The manner in which the interpolation is done varies by the method chosen:

1. For the Journey to crime method, P(Jtc), the routine interpolates the selected distance function to each grid cell to produce a density estimate. The density estimates are converted to probabilities so that the sum of the grid cells equals 1.0;
2. For the general crime distribution method, P(O), the routine sums up the incidents by each origin zone and interpolates that using the normal distribution method of the single kernel density routine (see Kernel Density Interpolation). The density estimates are converted to probabilities so that the sum of the grid cells equals 1.0.

3. For the distribution of origins based only on the incident committed by the serial offender, the routine identifies the zone in which the incident occurs and reads only those origins associated with those destination zones in the origin-destination matrix. Multiple incidents committed in the same origin zone are counted multiple

times. The routine then uses the single kernel density routine to interpolate the distribution to the grid (see Kernel Density Interpolation). The density estimates are converted to probabilities so that the sum of the grid cells equals 1.0.

4. For the product of the Journey-to-crime estimate and the distribution of origins based only on the incidents committed by the serial offender, the routine multiples the probability estimate obtained in 1 above by the probability estimate obtained in 3 above. The product density estimates are converted to probabilities so that the sum of the grid cells equals 1.0.

5. For the Bayesian risk estimate, the routine takes the product estimate (4 above) and divides it by the general crime distribution estimate (2 above). The resulting densities are converted to probabilities so that the sum of the grid cells equals 1.0.

6. If one or two filter variables are used, each filter variable is interpolated to the reference grid and then converted into probabilities. The filter probability grid is then multiplied by the P(Jtc), P(O|Jtc), "Product" or Bayesian risk grids to produce a filtered grid.

Note that in all estimates, the cells are converted to probabilities prior to any multiplication or division. The results are then re-scaled so that the resulting grid is a probability (i.e., all cells sum to 1.0).

Output of Routine

Once the method has been selected, the routine interpolates the data to the grid cell and outputs it as a 'shp', 'mif/mid', or Ascii file for display in a GIS program. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

The tabular output shows the probability values for each cell in the matrix and also indicates which grid cell has the highest probability estimate.

Accumulator Matrix

There is also an intermediate output, called the *accumulator matrix*, which the user can save. This lists the number of origins identified in each origin zone for the specific pattern of incidents committed by the offender, prior to the interpolation to grid cells. That is, in reading the origin-destination file, the routine first identifies which zone each incident committed by the offender falls within. Second, it reads the origin-destination matrix and identifies which origin zones are associated with incidents committed in the particular destination zones. Finally, it sums up the number of origins by zone ID associated with the incident distribution of the offender. This can be useful for examining the distribution of origins by zones prior to interpolating these to the grid.

V. Spatial Modeling II

The second spatial modeling section conducts regression modeling of a dependent variable, either binomial, unconstrained, or a count variable. It also includes a module for modeling discrete (nominal) choices. There are five sets of routines in the section: 1) Regression I for modeling multivariate predictors of a continuous or binary variable; 2) Regression II for making predictions on a new data set based on a regression model; 3) Discrete choice I for modeling discrete decisions; 4) Discrete choice II for making predictions on a new data set based on a discrete choice model; and 5) Temporal modeling for predicting the expected number of counts of an incident variable by zones and for detecting when the actual number exceeds a threshold prediction.

Regression Modeling I

The aim of a regression model is to estimate a functional relationship between a dependent variable and one or more independent variables. In the current version, 18 possible regression models are available with several options for each of these:

MLE Normal (OLS)
MCMC Normal
MCMC Normal-CAR
MCMC Normal-SAR
MLE Poisson
MLE Poisson with linear dispersion correction (NB1)
MLE Poisson-Gamma (NB2)
MCMC Poisson-Gamma (NB2)
MCMC Poisson-Gamma-CAR
MCMC Poisson-Gamma-SAR
MCMC Poisson-Lognormal
MCMC Poisson-Lognormal-CAR

MCMC Poisson-Lognormal-SAR
MLE Binomial Logit
MLE Binomial Probit
MCMC Binomial Logit
MCMC Binomial Logit-CAR
MCMC Binomial Logit-SAR

In addition, each of the 12 MCMC models can be run with an exposure (offset) variable used to define the population 'at risk' allowing a total of 30 possible regression models to be run.

There are two pages in the module. The Regression I page allows the testing of a model while the Regression II page allows a prediction to be made based on an already-estimated model. Also, since the Regression I module and Trip Generation module in the Crime Travel Demand Model duplicate regression functions, only one of these can be run at a time.

Input Data set

The data set for the regression module can be the Primary file or another file. If it is the Primary file, then it must be a spatial data file with and X and Y coordinates defined on each record. If it is another file, click on 'Other' and then identify the file. Only 'dbf' or 'txt' files are allowed.

Dependent Variable

To start loading the module, click on the 'Calibrate model' tab. A list of variables from the Primary File is displayed. There is a box for defining the dependent variable. The user must choose one dependent variable.

Independent Variables

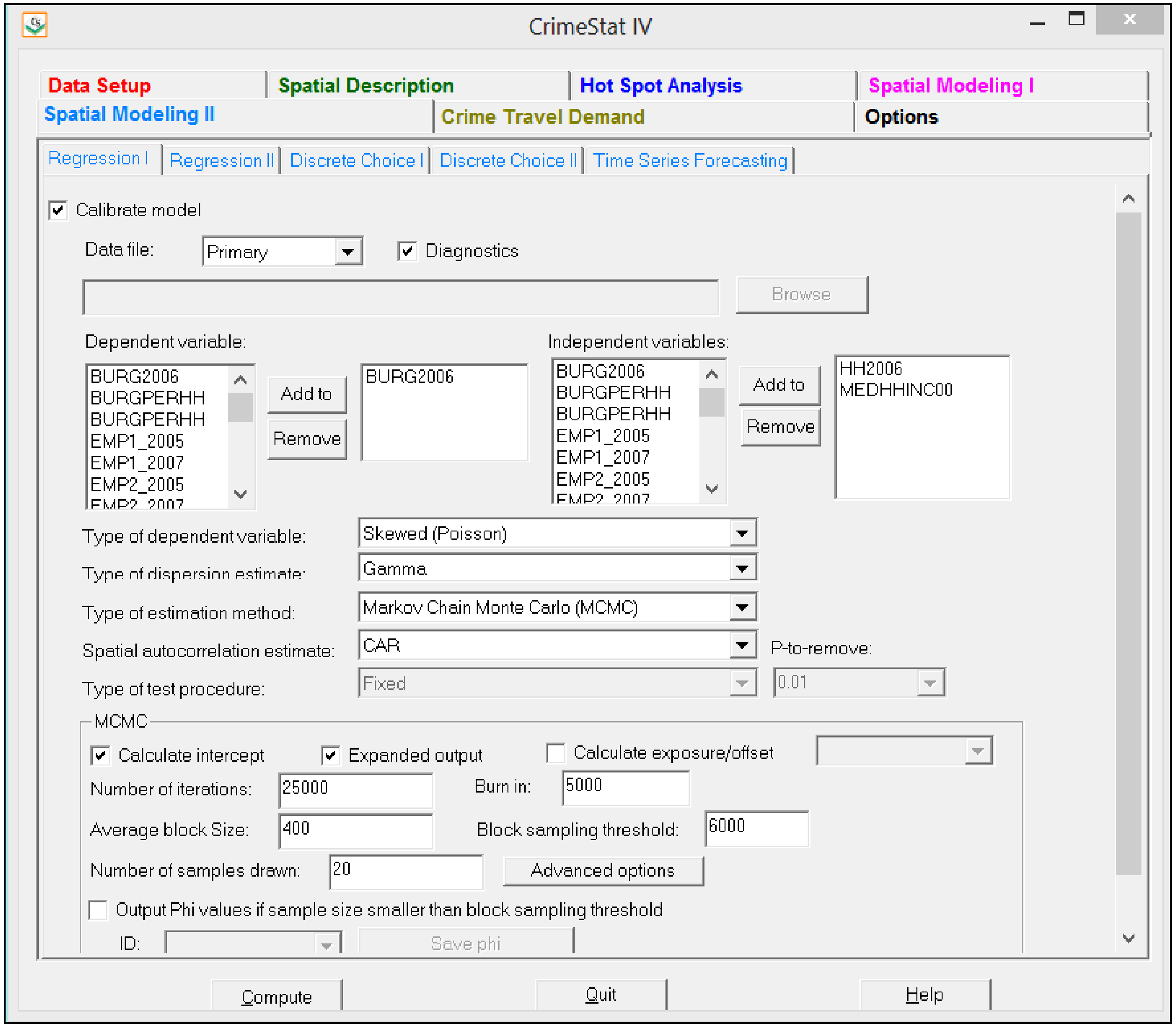
There is a box for defining the independent variables. The user must choose one or more independent variables. There is no limit to the number. The variables are output in the same order as specified in the dialogue so a user should consider how these are to be displayed.

Model decisions

There are five decisions that must be made for each regression model.

Figure 2.17:

Regression Modeling I



☑ Calibrate model

Data file: Primary ☑ Diagnostics

Dependent variable: BURG2006, BURGPERHH, BURGPERHH, EMP1_2005, EMP1_2007, EMP2_2005, EMP2_2007 — selected: BURG2006

Independent variables: BURG2006, BURGPERHH, BURGPERHH, EMP1_2005, EMP1_2007, EMP2_2005, EMP2_2007 — selected: HH2006, MEDHHINC00

Type of dependent variable: Skewed (Poisson)

Type of dispersion estimate: Gamma

Type of estimation method: Markov Chain Monte Carlo (MCMC)

Spatial autocorrelation estimate: CAR

P-to-remove: 0.01

Type of test procedure: Fixed

MCMC

☑ Calibrate intercept ☑ Expanded output □ Calculate exposure/offset

Number of iterations: 25000 Burn in: 5000

Average block Size: 400 Block sampling threshold: 6000

Number of samples drawn: 20

□ Output Phi values if sample size smaller than block sampling threshold

ID:

Type of Dependent Variable

The first model decision is the type of dependent variable: Skewed (Poisson), Normal/OLS, Binomial Probit, or Binomial Logit/Logistic. The default is a Poisson.

Type of Dispersion Estimate

The second model decision is the type of dispersion estimate to be used. The choices are Gamma, Poisson, Poisson with linear correction, and Lognormal. The default is Gamma. For the MLE and MCMC Normal (OLS) models, the dispersion is automatically normal. For the binomial logit or binomial probit, the dispersion is automatically binomial.

Type of Estimation Method

The third model decision is the type of estimation method to be used: Maximum Likelihood (MLE) or Markov Chain Monte Carlo (MCMC). The default is MLE.

Spatial Autocorrelation Regression Model

If the user accepts an MCMC algorithm, then a fourth decision is whether to run a spatial autocorrelation estimate along with it (a Conditional Autoregressive function – CAR, or a Simultaneous Autoregressive function - SAR). The MCMC Normal, MCMC Poisson-Gamma, MCMC Poisson-Lognormal, and MCMC Logit functions can be run with a spatial autocorrelation parameter. Note that the CAR model runs quite quickly whereas the SAR model runs very slowly. Unless the data set is small or a SAR model is absolutely essential, we recommend using a CAR function for the spatial regression models.

Type of Test Procedure

The fifth, and last model decision, is whether to run a fixed model or a backward elimination *stepwise* procedure (only with the MLE models). A fixed model includes all selected independent variables in the regression whereas a backward elimination model starts with all selected variables in the model but proceeds to drop variables that fail the P-to-remove test, one at a time. Any variable that has a significance level in excess of the P-to-remove value is dropped from the equation.

Specify whether a fixed model (all selected independent variables are used in the regression) or a backward elimination stepwise model is used. The default is a fixed model. If a backward elimination stepwise model is selected, choose the P-to-remove value (default is .01).

MCMC Model Choices

If the user chooses the MCMC algorithm, then eight *additional* decisions have to be made.

Number of Iterations

The first MCMC decision is the number of iterations. The default is 25,000. The number should be sufficient to produce reliable estimates of the parameters. Check the MC Error/Standard deviation ratio and the G-R statistic after the run to be sure most parameters are below 1.05 and 1.20 respectively. If not, increase the number of iterations and 'burn in' iterations.

'Burn in' iterations

The second MCMC decision is the number of initial iterations that will be dropped from the final distribution (the 'burn in' period). The default is 5,000. The number of 'burn in' iterations should be sufficient for the algorithm to reach an equilibrium state and produce reliable estimates of the parameters. Check the MC Error/Standard deviation ratio and the G-R statistic after the run to be sure most parameters are below 1.05 and 1.20 respectively. If not, increase the number of iterations and 'burn in' iterations.

Block Sampling Threshold

The third MCMC decision is whether to run all the records through the MCMC algorithm or whether to draw block samples. The algorithm will be run on all records unless the number of records exceeds number specified in the block sampling threshold. The default threshold is 6000 records. To run all the records through the MCMC algorithm, change this value to be greater than the number of records in the database. Note that calculation time will increase substantially if all records in a large database are run through the algorithm.

Average Block Size

The fourth MCMC decision is the number of records to be drawn for each block sample if the total number of records is greater than the block sampling threshold. The default is 400 records per block sample. Note that this is an average. Actual samples will vary in size. The output will display the expected sample size and the average sample size that was drawn.

Number of Samples Drawn

The fifth MCMC decision is the number of samples to be drawn if the total number of records is greater than the block sampling threshold. The default is 25 block samples. Typically, 20-30 block samples will achieve stable model results.

Calculate Intercept

The sixth MCMC decision is whether to run a model with or without an intercept (constant). The default is with an intercept estimated. To run the model without the intercept, uncheck the ‘Calculate intercept’ box.

Calculate Exposure/Offset

The seventh MCMC decision is whether to run a risk model. If the model is a risk or rate model, then an exposure (offset) variable needs to be defined. The exposure (offset) choice is available for the MCMC Poisson-Gamma, MCMC Poisson-Lognormal, and MCMC Binomial Logit models plus their spatial autocorrelation options. It is not available for the MCMC Normal or MCMC Normal-CAR/SAR models. Check the ‘Calculate exposure/offset’ box and identify the variable that will be used as the exposure variable. The coefficient for this variable will automatically be 1.0.

Advanced Options

The eighth, and last, MCMC decision is the prior values used for the different parameters being estimated. The MCMC algorithm requires an initial estimate for each parameter. There is a dialogue of advanced options for the MCMC algorithm by which they can be changed.

Initial Parameters Values

For the beta coefficients (including the intercept), the default values are 0. These are displayed as a blank screen for the Beta box. However, other prior estimates of the beta coefficients can be substituted for the assumed 0 coefficients. To do this, all independent variable coefficients plus the intercept (if used) must be listed in the order in which they appear in the model and must be separated by commas. Do not include the beta coefficients for the spatial autocorrelation term (if used) or the error (Taupsi) term.

Taupsi (error term)

The output of the MCMC always includes an error term, called *Taupsi* (τ_ψ). This is an exponent of the error term, $e^{\tau\psi}$, which together is called the *dispersion parameter*. The default value for Taupsi is 1.0. The user can substitute an alternative value.

Rho and Tauphi

The spatial autocorrelation component is made up of three separate sub-components, called Rho, Tauphi, and Alpha and are additive. Rho is roughly a global component that applies to the entire data set. Tauphi is roughly a neighborhood component that applies to a sub-set of the data. Alpha is essentially a localized effect. The default initial values for Rho and Tauphi are 0.5 and 1 respectively. The user can substitute alternative values for these parameters.

Alpha

Alpha is the exponent for the distance decay function in the spatial model. Essentially, the distance decay function defines the weight to be applied to the values of nearby records. The weight can be defined by one of three mathematical functions. First, the weight can be defined by a negative exponential function.

Second, the weight can be defined by a restricted negative exponential with the negative exponential operating up to the specified search distance, whereupon the weight becomes 0 for greater distances.

Third, the weight can be defined as a uniform value for all other observations within a specified search distance. This is a *contiguity* (or adjacency) measure. Essentially, all other observations have an equal weight within the search distance and 0 if they have a greater distance.

For the negative exponential and restricted negative exponential functions, substitute the selected value for alpha in the alpha box and for the restricted negative exponential and uniform functions, specify the search distance and distance units. The default is a negative exponential with an alpha of -1.0 in miles.

Value for 0 distance between records

The advanced options dialogue has a parameter for the minimum distance to be assumed between different records. If two records have the same X and Y coordinates (which could happen if the records are individual events, for example), then the distance between these records will be 0. This could cause unusual calculations in estimating spatial effects. Instead, it is more reliable to assume a slight difference in distance between all records. The default is 0.005 miles but the user can modify this (including substituting 0 for the minimal distance).

Output

The output depends on whether an MLE or an MCMC model has been run.

Maximum Likelihood (MLE) Model Output

The MLE routines (Normal/OLS, Poisson, Poisson with linear correction, MLE Poisson-Gamma, Binomial Probit, and Binomial Logit/Logistic) produce a standard output that includes summary statistics and estimates for the individual coefficients.

MLE Summary Statistics

The summary statistics include:

Information about the model

1. The data file
2. The dependent variable
3. The number of records
4. The residual degrees of freedom (N – number of parameters estimated)
5. The type of regression model (Normal/OLS, Poisson, Poisson with linear correction, Poisson-Gamma, Binomial Logit, Binomial Probit)
6. The method of estimation (MLE)

Likelihood statistics

7. Log-likelihood estimate, which is a negative number. For a set number of independent variables, the more negative the log-likelihood the better.
8. Log-likelihood per case. This divides the log-likelihood by the sample size (N). This indicates the average contribution to the log-likelihood of each observation. The more negative, the better.
9. Akaike Information Criterion (AIC) adjusts the log-likelihood for the degrees of freedom. The smaller the AIC, the better.

10. AIC per case. This divides the AIC statistic by the sample size (N). This indicates the average contribution to the AIC of each observation. The smaller, the better.
11. Bayesian Information Criterion (BIC), sometimes known as the Schwartz Criterion (SC), adjusts the log-likelihood for the degrees of freedom. The smaller the BIC, the better.
12. BIC per case. This divides the BIC/SC statistic by the sample size (N). This indicates the average contribution to the BIC/SC of each observation. The smaller, the better.
13. Deviance compares the log-likelihood of the model to the log-likelihood of a model that fits the data perfectly. A smaller deviance is better.
14. The probability value of the deviance based on a Chi-square test with N-K-1 degrees of freedom where K is the number of independent variables.
15. Pearson Chi-square is a test of how closely the predicted model fits the data. A smaller Chi-square is better since it indicates the model fits the data better.
16. The probability value of the Pearson Chi-square based on a Chi-square test with N-K-1 degrees of freedom where K is the number of independent variables.

Model error estimates

17. Mean Absolute Deviation (MAD). For a set number of independent variables, a smaller MAD is better.
18. Quartiles for the Mean Absolute Deviation. For any one quartile, smaller is better.
19. Mean Squared Predictive Error (MSPE). For a set number of independent variables, a smaller MSPE is better.
20. Quartiles for the Mean Squared Predictive Error. For any one quartile, smaller is better.
21. Squared multiple R (for Normal/OLS models only). This is the percentage of the dependent variable accounted for by the independent variables.
22. Adjusted squared multiple R (for Normal/OLS models only). This is the squared multiple R adjusted for degrees of freedom.

Dispersion tests

23. Adjusted deviance. This is a measure of the difference between the observed and predicted values (the residual error) adjusted for degrees of freedom. The smaller the adjusted deviance, the better. A value greater than 1 indicates over-dispersion.

24. Probability of adjusted deviance. This is the probability associated with the adjusted deviance test with 1 degree of freedom.

25. Adjusted Pearson Chi-square. This is the Pearson Chi-square adjusted for degrees of freedom. The smaller the Pearson Chi-square, the better. A value greater than 1 indicates over-dispersion.

26. Probability of Adjusted Pearson Chi-square. This is the probability associated with the Pearson Chi-square test with 1 degree of freedom.

27. Dispersion multiplier. This is the ratio of the expected variance to the expected mean. For a set number of independent variables, the smaller the dispersion multiplier, the better. For example, in a pure Poisson distribution, the dispersion should be 1.0. In practice, a ratio greater than 10 indicates that there is too much variation that is unaccounted for in the model. Either add more variables or change the functional form of the model.

28. Z-test for dispersion multiplier (Poisson models only). This is a test for whether the dispersion parameter is significantly greater than that assumed by the Poisson model. It is a test of over-dispersion.

29. P-value for Z-test of dispersion parameter (Poisson models only). This is the one-tail probability level associated with the Z-test.

30. Inverse dispersion multiplier. For a set number of independent variables, a larger inverse dispersion multiplier is better. A ratio close to 1.0 is considered good.

MLE Individual Coefficient Statistics

For the individual coefficients, the following are output:

31. The coefficient. This is the estimated value of the coefficient from the maximum likelihood estimate.

32. Standard Error. This is the estimated standard error from the maximum likelihood estimate.

33. Pseudo-tolerance. This is the tolerance value based on a linear prediction of the variable by the other independent variables. See equation Up. 2.18.

34. Z-value. This is asymptotic Z-test that is defined based on the coefficient and standard error. It is defined as Coefficient/Standard Error.

35. p-value. This is the two-tail probability level associated with the Z-test.

Markov Chain Monte Carlo (MCMC) Model Output

The MCMC routines (Poisson-Gamma, Poisson-Gamma-CAR/SAR, Poisson-Lognormal, Poisson-Lognormal-CAR/SAR, Binomial Logit, Binomial Logit-CAR/SAR) produce a standard output and an optional expanded output**.** The standard output includes summary statistics and estimates for the individual coefficients.

MCMC Summary Statistics

The summary statistics include:

Information about the model

1. The dependent variable
2. The number of records
3. The sample number. This is only output when the block sampling method is used.
4. The number of cases for the sample. This is only output when the block sampling method is used.
5. Date and time for sample. This is only output when the block sampling method is used
6. The residual degrees of freedom (N – number of parameters estimated)
7. The type of regression model (Normal, Normal-CAR/SAR, Poisson-Gamma, Poisson-Gamma-CAR/SAR, Poisson-Lognormal, Poisson-Lognormal-CAR/SAR, Binomial Logit, Binomial Logit-CAR/SAR)
8. The method of estimation
9. The number of iterations
10. The 'burn in' period
11. The block size is the expected number of records selected for each block sample. The actual number may vary.
12. The number of samples drawn. This is output when the block sampling method used.
13. The average block size. This is output when the block sampling method used.
14. The type of distance decay function used. This is output for models that use CAR or SAR spatial autocorrelation functions.
15. Condition number for the distance matrix. If the condition number is large, then the model may not have properly converged. This is output for the Poisson-Gamma-CAR model only.
16. Condition number for the inverse distance matrix. If the condition number is large, then the model may not have properly converged. This is output for the Poisson-Gamma-CAR/SAR or Poisson-Lognormal-CAR/SAR models only.

Likelihood statistics

17. Log-likelihood estimate, which is a negative number. For a set number of independent variables, the smaller the log-likelihood (i.e., the most negative) the better.
18. Log-likelihood per case. This divides the log-likelihood by the sample size (N). This indicates the average contribution to the log-likelihood of each observation. The more negative, the better.
19. Deviance Information Criterion (DIC) for Poisson models only. This adjusts the log-likelihood for the effective degrees of freedom. The smaller the DIC, the better.
20. Akaike Information Criterion (AIC) adjusts the log-likelihood for the degrees of freedom. The smaller the AIC, the better.
21. AIC per case. This divides the AIC statistic by the sample size (N). This indicates the average contribution to the AIC of each observation. The smaller, the better.
22. Bayesian Information Criterion (BIC), sometimes known as the Schwartz Criterion (SC), adjusts the log-likelihood for the degrees of freedom. The smaller the BIC, the better.
23. BIC per case. This divides the BIC/SC statistic by the sample size (N). This indicates the average contribution to the BIC/SC of each observation. The smaller, the better.
24. Deviance compares the log-likelihood of the model to the log-likelihood of a model that fits the data perfectly. A smaller deviance is better.
25. The probability value of the deviance based on a Chi-square test with N-K-1 degrees of freedom where *K* is the number of independent variables.
26. Pearson Chi-square is a test of how closely the predicted model fits the data. A smaller Chi-square is better since it indicates the model fits the data well
27. The probability value of the Pearson Chi-square based on a Chi-square test with N-K-1 degrees of freedom where *K* is the number of independent variables.

Model error estimates

28. Mean Absolute Deviation (MAD). For a set number of independent variables, a smaller MAD is better.
29. Quartiles for the Mean Absolute Deviation. For any one quartile, smaller is better.
30. Mean Squared Predictive Error (MSPE). For a set number of independent variables, a smaller MSPE is better.

31. Quartiles for the Mean Squared Predictive Error. For any one quartile, smaller is better.

Dispersion tests

32. Adjusted deviance. This is a measure of the difference between the observed and predicted values (the residual error) adjusted for degrees of freedom. The smaller the adjusted deviance, the better. A value greater than 1 indicates over-dispersion.
33. The probability value of the adjusted deviance based on a Chi-square test with 1 degree of freedom.
34. Adjusted Pearson Chi-square. This is the Pearson Chi-square adjusted for degrees of freedom. The smaller the Pearson Chi-square, the better. A value greater than 1 indicates over-dispersion.
35. The probability value of the adjusted Pearson Chi-square based on a Chi-square test with 1 degree of freedom.
36. Dispersion multiplier. This is the ratio of the expected variance to the expected mean. For a set number of independent variables, the smaller the dispersion multiplier, the better. In a pure Poisson distribution, the dispersion should be 1.0. In practice, a ratio greater than 10 indicates that there is too much variation that is unaccounted for in the model. Either add more variables or change the functional form of the model.
37. Inverse dispersion multiplier. For a set number of independent variables, a larger inverse dispersion multiplier is better. A ratio close to 1.0 is considered good.

MCMC Individual Coefficients Statistics

For the individual coefficients, the following are output:

38. The mean coefficient. This is the mean parameter value for the *N-k* iterations where *k* is the 'burn in' samples that are discarded. With the MCMC block sampling method, this is the mean of the mean coefficients for all block samples.
39. The standard deviation of the coefficient. This is an estimate of the standard error of the parameter for the *N-k* iterations where *k* is the 'burn in' samples that are discarded. With the MCMC block sampling method, this is the mean of the standard deviations for all block samples.
40. t-value. This is the t-value based on the mean coefficient and the standard deviation. It is defined by Mean/Std.
41. p-value. This is the two-tail probability level associated with the t-test.

42. Adjusted standard deviation (Adj. Std). The block sampling method will produce substantial variation in the mean standard deviation, which is used to estimate the standard error. Consequently, the standard error will be too large. An approximation is made by multiplying the estimated standard deviation by

$$\sqrt{\frac{\bar{n}}{N}}$$

where $\bar{n}$ is the average sample size of the block samples and *N* is the number of records. If no block samples are taken, then this statistic is not calculated.

43. Adjusted t-value. This is the t-value based on the mean coefficient and the adjusted standard deviation. It is defined by Mean/Adj_Std. If no block samples are taken, then this statistic is not calculated.

44. Adjusted p-value. This is the two-tail probability level associated with the adjusted t-value. If no block samples are taken, then this statistic is not calculated.

45. MC error is a Monte Carlo simulation error. It is a comparison of the means of *m* individual chains relative to the mean of the entire chain. By itself, it has little meaning.

46. MC error/Std is the MC error divided by the standard deviation. If this ratio is less than .05, then it is a good indicator that the posterior distribution has converged.

47. G-R stat is the Gelman-Rubin statistic which compares the variance of *m* individual chains relative to the variance of the entire chain. If the G-R statistic is under 1.2, then the posterior distribution is commonly considered to have converged.

48. Spatial autocorrelation term (Phi) for CAR/SAR models only. This is the estimate of the fixed effect spatial autocorrelation effect. It is made up of three components: a global component (Rho); a local component (Tauphi); and a local neighborhood component (Alpha, which is defined by the user).

49. The log of the error in the model (Taupsi). This is an estimate of the unexplained variance remaining. Taupsi is the exponent of the dispersion multiplier, $e^{\tau\psi}$. For any fixed number of independent variables, the smaller the Taupsi, the better.

Expanded Output (MCMC only)

If the expanded output box is selected, additional information on the percentiles from the MCMC sample are displayed. If the block sampling method is used, the percentiles are the means of all block samples. The percentiles are:

50. 2.5th percentile

51. 5th percentile
52. 10th percentile
53. 25th percentile
54. 50th percentile (median)
55. 75TH percentile
56. 90th percentile
57. 95th percentile
58. 97.5th percentile

The percentiles can be used to construct confidence intervals around the mean estimates or to provide a non-parametric estimate of significance as an alternative to the estimated t-value in the standard output. For example, the 2.5th and 97.5th percentiles provide approximate 95 percent confidence intervals around the mean coefficient while the 0.5th and 99.5th percentiles provide approximate 99 percent confidence intervals.

The percentiles will be output for all estimated parameters including the intercept, each individual predictor variable, the spatial effects variable (Phi), the estimated components of the spatial effects (Rho and Tauphi), and the overall error term (Taupsi).

Output Phi Values (CAR/SAR models only)

For CAR or SAR models only, the individual Phi values can be output. This will occur if the sample size is smaller than the block sampling threshold. Check the ‘Output Phi value if sample size smaller than block sampling threshold’ box. An ID variable must be identified and a DBF output file defined.

Multicollinearity Among the Independent Variables in the Regression Model

A major consideration in any regression model is that the independent variables are statistically independent. Non-independence is called *Multicollinearity*. Non-independence means that there is overlap in prediction among two or more independent variables. This can lead to uncertainty in interpreting coefficients as well as an unstable model that may not hold in the future. Generally, it is a good idea to reduce Multicollinearity as much as possible.

A tolerance test is given for each coefficient. This is defined as 1 – the R-square of the independent variable predicted by the remaining independent variables in the equation using an Ordinary Least Squares model. It is an indicator of how much the other independent variables in a model account for the variance of any particular independent variable. Since the method uses the Ordinary Least Squares methods, it is an approximate (pseudo) test for the functions estimated by maximum likelihood. A message is displayed that indicates probable or possible Multicollinearity. If there is substantial Multicollinearity (indicated by low tolerance values), it is a good idea is to drop one of the multicolinear independent variables and re-run the model. However, each of the coefficients should be inspected carefully before accepting a final model.

Graph of Residual Errors

While the output page is open, clicking on the graph button will display a graph of the residual errors (on the Y axis) against the predicted values (on the X axis). Only residual errors that vary between -200 and +200 are shown to allow most of the errors to be displayed.

Save Output

The predicted values and the residual errors can be output to a DBF file with a RegOut<*root name*> with the root name being provided by the user. The output includes all the variables in the input data set plus two new ones: 1) the predicted values of the dependent variable for each observation (with the field name PREDICTED); and 2) the residual error values, representing the difference between the actual /observed values for each observation and the predicted values (with the field name RESIDUAL). The file can be imported into a spreadsheet or graphics program and the errors plotted against the predicted dependent variable.

Save Estimated Coefficients

The individual coefficients can be output to a DBF file with a RegCoeff <*root name*> with the root name being provided by the user. This file can be used in the ‘Make Prediction’ routine under Regression II.

Diagnostic Tests

The regression module has a set of diagnostic tests for evaluating the characteristics of the data and the most appropriate model to use. There is a diagnostics box on the Regression I page.

Diagnostics are provided on:

1. The minimum and maximum values for the dependent and independent variables
2. Skewness in the dependent variable
3. Spatial autocorrelation in the dependent variable
4. Estimated values for the distance decay parameter – alpha, for use in the CAR or SAR models

5. Multicolinarity among the independent variables

Minimum and Maximum Values for the Variables

First, the minimum and maximum values of both the dependent and independent variables are listed. A user should look for ineligible values (e.g., -1) as well as variables that have a very high range. The MLE routines are sensitive to variables with very large ranges.

Skewness Tests

Skewness in the dependent variable can distort a linear model by allowing high values to be underestimated while allowing low values to be overestimated and a Poisson-type model is preferred over the linear for highly skewed variables.

The diagnostics utility tests for skewness using two different measures: 1) the "***g"*** statistic, and 2) the ratio of the simple variance to the simple mean. Either significant "g" scores or variance-to-mean ratios greater than about 2:1 should make the user cautious about using a linear model. If either measure indicates skewness, *CrimeStat* prints out a message indicating the dependent variable appears to be skewed and that a Poisson or Poisson-Gamma model should be used.

Testing for Spatial Autocorrelation in the Dependent Variable

The third type of test in the diagnostics utilities is the Moran's "I" coefficient for spatial autocorrelation. If the "I" is significant, *CrimeStat* outputs a message indicating that there is definite spatial autocorrelation in the dependent variable and that it needs to be accounted for, either by a proxy variable or by estimating a CAR or SAR model.

Estimating the Value of Alpha for CAR or SAR Models

The fourth type of diagnostic test is an estimate of a plausible value for the distance decay function, α, in CAR or SAR models. Three values of alpha are given in different distance units, one associated with a weight of 0.9 (a very steep distance decay), one associated with a weight of 0.75 (a moderate distance decay), and one associated with a weight of 0.5 (a shallow distance decay). Users should run the Moran Correlogram and examine the graph of the drop off in spatial autocorrelation to assess what type of decay function most likely exists. The user should choose an alpha value that best represents the distance decay and should define the distance units for it.

Multicollinearity Test

The fifth type of diagnostic test is for Multicollinearity among the independent predictors. The pseudo-tolerance test is presented for each independent variable. This is defined as 1-R^2 for the other independent variables in the equation. Each independent variable should have a high tolerance (0.90 or higher). *CrimeStat* prints out an error message if tolerance is not high.

Regression Modeling II

The Regression II module allows the user to apply a model to another data set and make a prediction. The ‘Make prediction’ routine allows the application of coefficients to a data set. There are two types of models that are fitted – linear and Poisson. For both types of model, the coefficients file must include information on the intercept and each of the coefficients. The user reads in the saved coefficient file and matches the variables to those in the new data set based on the order of the coefficients file.

If the model had estimated a general spatial effect from a CAR or SAR model, then the general Phi will have been saved with the coefficient files. If the model had estimated specific spatial effects from a CAR or SAR model, then the specific Phi values will have been saved in a separate Phi coefficients file. In the latter case, the user must read in the Phi coefficients file along with the general coefficient file.

Data File

The data set for the regression module can be the Primary file or another file. If it is the Primary file, then it must have X and Y coordinates defined on each record. If it is another file, click on ‘Other’ and then identify the file. Only ‘dbf’ or ‘txt’ files are allowed.

Saved Coefficients File

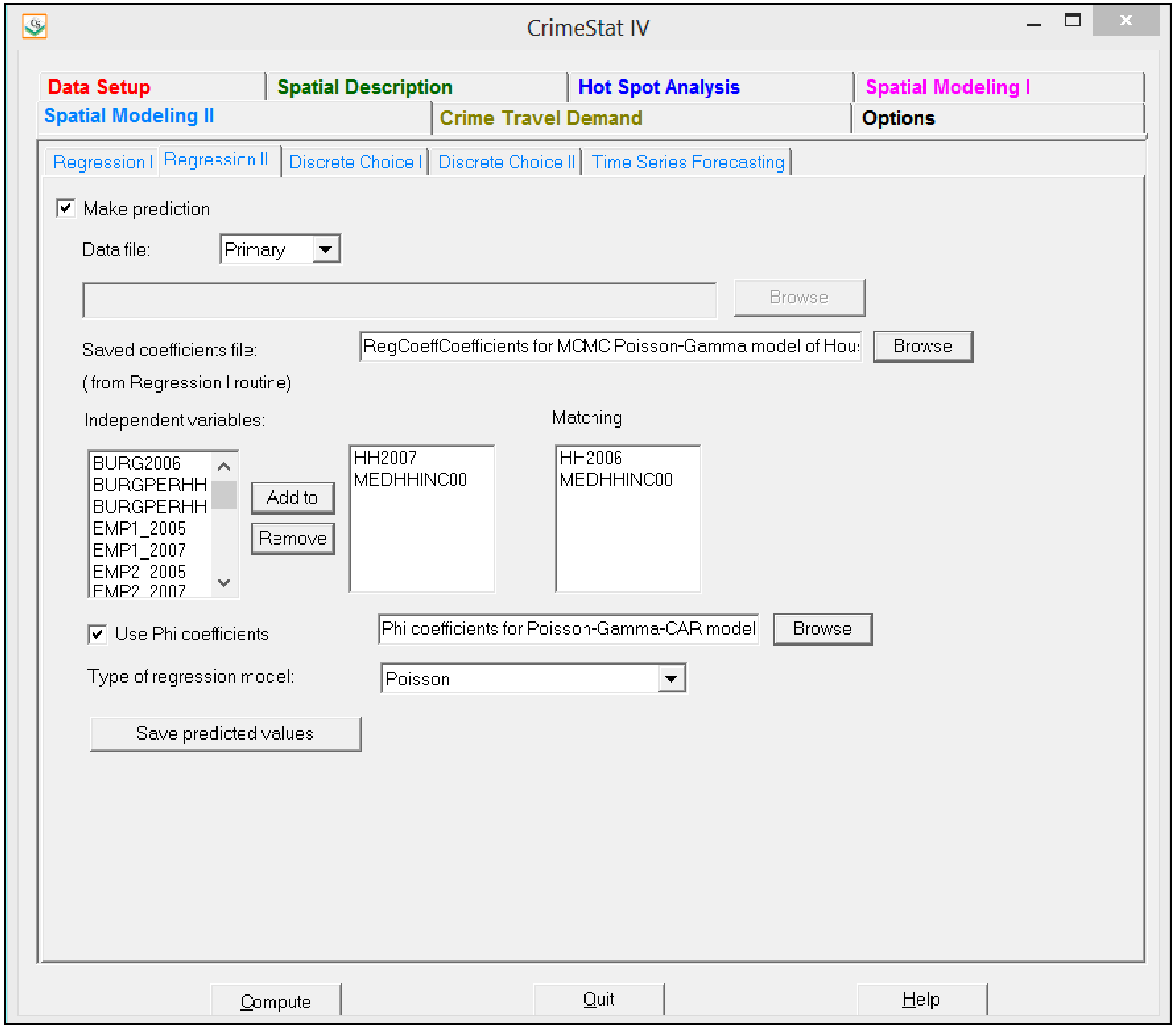
In order to make a prediction, a model must have already been calibrated and the coefficients saved in a coefficients file. Point to the directory where the coefficients file has been saved and identify it.

Matching Independent Variables

The independent variables that were used in the calibrated coefficients file will be listing in the matching column. Select corresponding variables from the input data file. The items should be listed in the same order as in the matching column.

Figure 2.18:

Regression Modeling II



☑ Make prediction

Data file: Primary

Saved coefficients file: RegCoeffCoefficients for MCMC Poisson-Gamma model of Hou:

(from Regression I routine)

Independent variables:

BURG2006
BURGPERHH
BURGPERHH
EMP1_2005
EMP1_2007
EMP2_2005
EMP2_2007

	Matching
HH2007	HH2006
MEDHHINC00	MEDHHINC00

☑ Use Phi coefficients: Phi coefficients for Poisson-Gamma-CAR model

Type of regression model: Poisson

Use Phi coefficients

If the saved coefficients file was from a model that was a spatial regression, the saved Phi coefficients can be also applied to the new data set. The number of Phi coefficients must match the number of records in the input data file, however. For example, this would be appropriate when a model is calibrated on zones which do not change over time. Therefore, the Phi coefficients estimated for the zones in one time period could be applied to the same zones to make a prediction for a later time period.

Point to the directory where the Phi coefficients have been saved and identify the file.

Output

The screen output provides predictions of the value of the dependent variable in the same order as in the input data set.

Save Predicted Values

The predicted values and the residual errors can be output to a DBF file with a RegMakePred<*root name*> with the root name being provided by the user. A column called PREDICTED will be added that contains the predicted value of the dependent variable.

Discrete Choice Modeling I

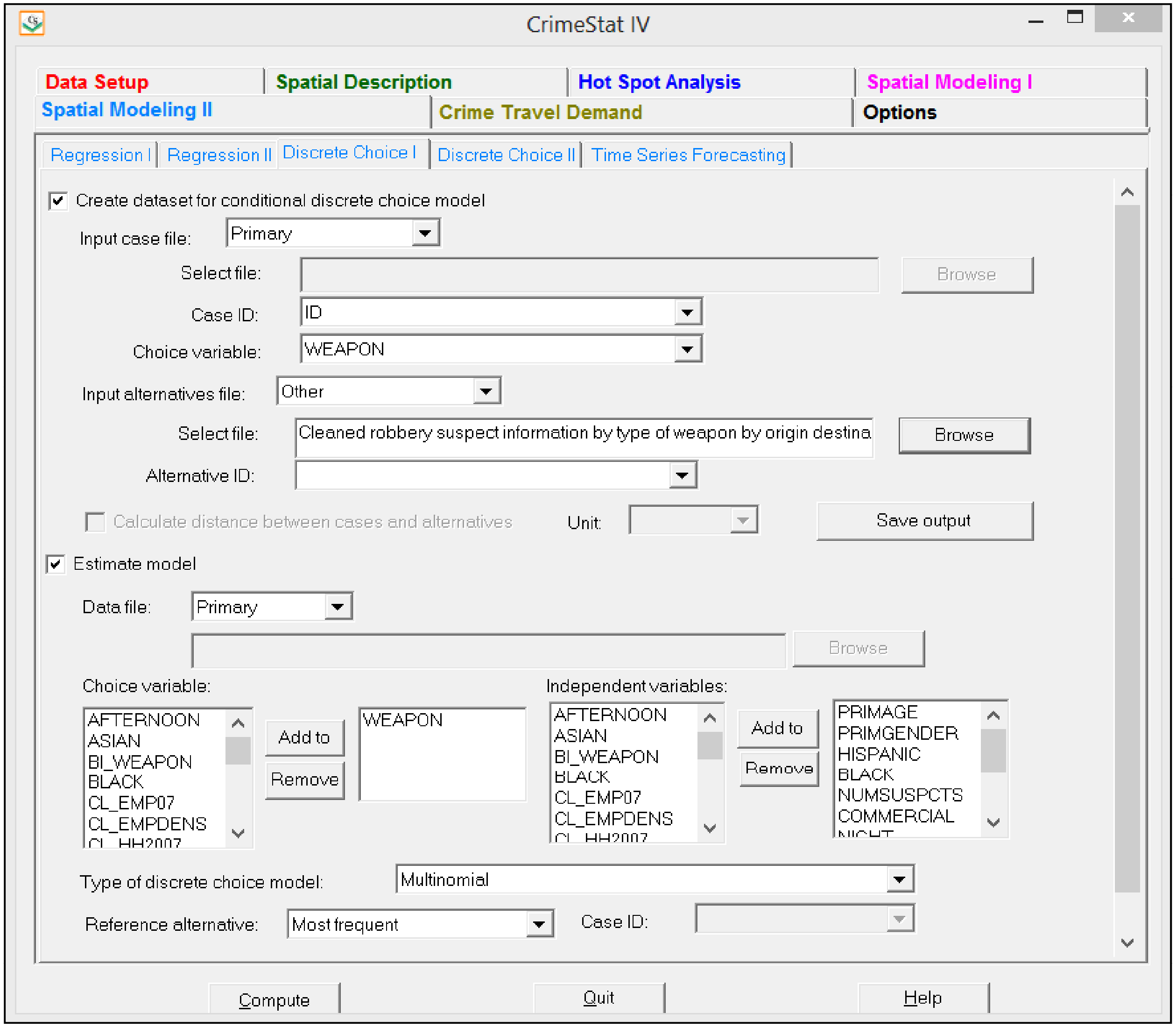
The aim of the discrete choice I module is to estimate a functional relationship between a discrete (nominal) dependent variable and one or more independent variables. It is a statistical method that is derived from utility theory, i.e. random utility maximization (RUM) theory. A 'decision maker' (e.g., an offender committing a crime) is faced with a set of alternatives, labeled 1 through *J*, from which s/he has to select exactly one.

The probability that an alternative will be chosen is a function of its observed and unobserved utility to the decision maker. The observed utility is a function of known variables and can be expressed as a linear combination of the independent variables. The unobserved utility is the random error component of the model. The estimated probability is the exponentiated observed utility of a specific alternative, *J*, divided by the sum of the exponentiated observed utilities of all available alternatives.

There are two general forms of the discrete choice model, multinomial logit and conditional logit. The *multinomial logit* model estimates the probability that a specific

Figure 2.19:

Discrete Choice Modeling I



☑ Create dataset for conditional discrete choice model

Input case file: Primary

Select file:

Case ID: ID

Choice variable: WEAPON

Input alternatives file: Other

Select file: Cleaned robbery suspect information by type of weapon by origin destina

Alternative ID:

□ Calculate distance between cases and alternatives Unit:

☑ Estimate model

Data file: Primary

Choice variable:

AFTERNOON
ASIAN
BI_WEAPON
BLACK
CL_EMP07
CL_EMPDENS
CL_HH2007

WEAPON

Independent variables:

AFTERNOON
ASIAN
BI_WEAPON
BLACK
CL_EMP07
CL_EMPDENS
CL_HH2007

PRIMAGE
PRIMGENDER
HISPANIC
BLACK
NUMSUSPCTS
COMMERCIAL
NIGHT

Type of discrete choice model: Multinomial

Reference alternative: Most frequent Case ID:

alternative, 1 to *J*, as a function of characteristics of the decision makers, either personal characteristics (e.g., age, gender, ethnicity) or environmental characteristics (e.g., the median household income of the block in which the decision maker lives). The probability that any one alternative is chosen is estimated as a function of these characteristics. Per variable (characteristic), there is one parameter estimated for every alternative, one of which is the reference alternative in which the coefficients are automatically set to 0.

The multinomial logit model is most appropriate when the outcome of the choice is expected to depend mostly on characteristics of the decision maker (and not on observed characteristics of the alternatives) and when there are only a limited number of alternatives available (e.g., 5 weapon choices). The *conditional logit* model is a more general model and estimates the probability of a set of alternatives, 1 to *J*, as a function of characteristics of the alternatives themselves, possibly in interaction with characteristics of the decision maker. The conditional logit model is most appropriate when the outcome of the choice is expected to depend mostly on the characteristics of the alternatives, and can handle a large number of alternatives. However, the analysis file becomes very large. There is a single parameter estimated for every characteristic of the alternative.

Although the multinomial and the conditional logit are based on a single underlying statistical model, their estimation requires different data structures. In the multinomial logit model, the data contain a single record for every decision maker, and a single dependent (nominal) variable that indicates which alternative (*1..J*) was chosen. Thus, if there are *N* decision makers, there are *N* records and at least one varible indicates which alternative was chosen. The file structure is thus similar to that used in the regression module.

In the conditional logit model, for each decision maker there is a record for every choice that this decision maker is faced. Thus, if there are *N* decision makers and *J* alternatives available to every decision maker, then the data set has *N*J* records, one for every alternative faced by the decision maker. In this case, the alternative that was selected has to be indicated by a dichotmous (dummy) variable (1 for chosen and 0 for not chosen).

Create Data set for Conditional Logit Model

This routine is optional. It simplifies the task of creating a database for use in the conditional logit model. It matches a *case* database with a alternatives data base, producing the cross join of both databases. The ***case*** database is the database for the multinomial logit model. It will thus have the individual records of the decision makers – offenders, individuals, organizations. It will include at least one variable indicating the alternative that the decision maker selected (e.g., type of crime committed, the type of weapon used, the location where the crime was committed) as well as characteristics of the individuals or characteristics associated with the individuals (e.g., age, gender, ethnicity, median household income of the zone where the decision maker lives time of event, day of week of event).

The ***alternatives*** database, on the other hand, lists the individual alternatives that were available (e.g., all the locations where a crime could be committed, all the different types of weapons that were used by different offenders) as well as attributes associated with the alternatives themselves (e.g., median household income or number of employees working at the locations, or characteristics associated with each type of weapon).

The joined has one record per alternative for each case. Thus, if there are *N* individuals faced with *J* choices, then the matching routine will create *N**J* records. It should be noted that the matching assigns every characteristic associated with a choice to every case associated with a decision maker. A field, called CHOSEN, is automatically added to every record. This field has the value 1 for alternatives that were chosen and 0 for alternatives that were not chosen. The Chosen field should thus sum to N (i.e., only one record per decision maker should have a selected alternative). Also, as an option, and only if both the individuals and the alternatives have geographic coordinates, a second field called DISTANCE will be added that calculates the distance from each case record to each alternative record. The user must specify which distance units are to be used (miles, kilometers, meters, feet, or nautical miles).

For example, if both the case database and the alternatives database contain X and Y coordinates, then it is possible to calculate the distance between every decision maker and every choice. In most situations, locations at shorter distances are more likely to be chosen.

The routine cannot calculate other interactions associated with a specific alternative and particular decision maker, and such interactions must be added to the data outside CrimeStat. Interactions between variables in the data can be calculated. For example, to test whether increasing distance makes alternatives less attractive for juvenile offenders but not for adult offenders, an interaction DISTANCE x AGE can be calculated. Other interactions require additional information, for example if location choice is what is modeled, one may want to add a variable indicating, for each alternative location, how many prior offences the offender has committed before in that alternative location. In these cases the external file is constructed by the user, and the step “Create data set for conditional discrete choice model” is skipped.

Input Case File

The case data set for the Discrete Choice I module can be the Primary file or another file. If it is the Primary file, then it must have X and Y coordinates defined on each record. If it is another file, then there are no coordinates defined. Click on ‘Other’ and then identify the file. Only ‘dbf’ or ‘txt’ files are allowed. To avoid confusion, the user must verify that no variable/field in the input case file has the same name as any variable in the Input Alternatives File (see below).

Case ID

Select the Case ID. The Input Case File must have a Case ID, a variable that uniquely identifies cases in the Input Case File.

Choice variable

Select the Choice Variable. The Input Case File must contain a variable (field) that identifies alternative chosen by the decision maker. For example, if the choice is about the type of weapon used, then the Choice Variable indicates whether it was a gun, a knife, strong arm, and so forth. Or, if the choice is the census tract in which a crime was perpetrated, then the Choice Variable identifies the census tract where the incident occurred.

Input Alternatives File

The alternatives data set for the Discrete Choice I module can be the Primary file or another file. If it is the Primary file, then it is a spatial file and must have X and Y coordinates on each record. If it is another file, then there are no coordinates defined. Click on ‘Other’ and then identify the file. Only ‘dbf’ or ‘txt’ files are allowed. To avoid confusion, the user must verify that no variable in the input alternative file has the same name as any variable in the Input Case File.

Alternatives ID

Select the Alternatives ID. The Alternatives File must have an Alternative ID, a variable that uniquely identifies records in file. The coding of the Alternative ID variable must exactly match the coding of the Choice Variable in the Input Case File. Be careful about ID names. If the ID names are the same, the name will appear twice in the file with the first use representing the case file and the second use representing the alternatives file. The names reflect the link between each case ID and each alternatives ID. It will be better to use different names to confusion.

Calculate Distance between Cases to Alternatives

There is an optional box that allows the routine to calculate the distance from each case record to each alternative record. If checked, the routine will calculate the distance. This only applies if both the case file and the alternatives file are either the Primary file or Secondary. The user must specify the distance units to be used in the calculation (in miles, kilometers, feet, meters, or nautical miles). The box is checked by default. The saved filed will have a new field called DIST. For example, if the X/Y coordinates for an offender's home address are coded in the Input Case File while the coordinates for census tract are recorded in the Input Alternatives File, then the distances from the offender's home to each alternative census tract will be calculated.

Save Output

The matched Input Case and Input Alternatives file is saved as a new file in 'dbf' format, that can subsequently be used to estimate a conditional (but not multinomial) logit model, as described below under 'Estimating a conditional logit model'. The user should define the name of the file and point to the directory where it is saved. The output includes all fields from the case file and all fields from the choice file, and optionally a field DIST containing calculated distances. There will be *J* records for each of the *N* cases. There will be an automatically added field called CHOSEN that takes the value '1' for the choice that was selected and '0' for choices that were not selected.

Note that because the joined data base can be very large, before you start creating a data set for conditional discrete choice model, be careful to include in the alternatives and choice files only variables that you are likely to use in your analysis, and to format them to be as small as possible.

Estimate Model

The Estimate Model routine will estimate a discrete choice model, either the multinomial logit or the conditional logit.

Estimating a Multinomial Logit Model

The *multinomial logit* model is used when there is one record per decision maker with a choice having been made by the decision maker. The model estimates the effect of each independent variable on the probability of each distinct alternative. The data are structured so that there is one record per decision maker with the choice variable indicating which alternative was chosen. The data set is similar to that of the regression model in that there is one record per decision maker.

The model then estimates the effects of the independent variables on the probability of each alternative. By definition, one of the alternatives (by default the most frequently chosen alternative, otherwise to be chosen by the user) is the reference alternative to which the other alternatives are compared.

The multinomial logit model is always estimated with a constant. This type of model is appropriate when values of the predictor variables only vary across cases (decision-makers), not across alternatives.

Estimating a Conditional Logit Model

The *conditional logit* model, on the other hand, is used when the values of the predictor variables vary across alternatives. In that case, there is one record per alternative per decision maker. That is, the decision maker is faced with *J* alternatives but chooses only one. The database must indicate which of the *J* alternatives was selected and the model estimates the effect of each independent variable on choosing an alternative. There is a record for every alternative faced by the decision maker. The parameter estimates indicate the effects of the independent variables on the likelihood that the alternative is selected.

Typically, if there are *N* decision makers and *J* alternatives, then there will be normally *N x J* records. It is possible for a particular decision maker to have fewer than *J* alternatives. The model will still work.

Data File

The data set for the model can be either the Primary file or another file (the Secondary file is not available). If the Primary file is used, the coordinate system and distance units are the same as were defined on the Primary file page.

Select file for other discrete choice file

If the discrete choice file is another file than the Primary file, the user must browse and identify the file.

Choice Variable

A list of variables from the discrete choice file is displayed. There is a box for defining the choice variable. The user must select one choice variable. . For the conditional logit model, on the other hand, the variable contains a set of 1's (for selected alternatives) or 0's (for alternatives that were not selected). If the data set was constructed with the *CrimeStat* 'Create data set for conditional discrete choice model' routine, then the field CHOSEN should be used.

Note that the field that is added for the choice variable (whether CHOSEN or another variable) is inspected for unique values. If the data set is large, it may take awhile to filter through those values. Eventually, though, the variable will be added to the choice variable dialogue.

Independent Variables

There is a box for defining the independent variables. The user must choose one or more independent variables. There is no limit to the number. The variables are output in the same order as specified in the dialogue so a user should consider how these are to be displayed. The order in which the variables are entered does not affect the estimated parameters.

Type of Discrete Choice Model

The type of discrete choice model to be estimated must be specified. The choices are *Multinomial* (logit) or *Conditional* (logit). The default model is the Conditional logit. NOTE: the file used for a Multinomial Logit model is different than the file used for a Conditional Logit model. With the file used in the Multinomial Logit model, there is one record per case with the choice specified on the record. With the file used in the Conditional Logit model, there is one record per alternative with *J* records per case (where *J* is the number of alternatives). Be sure to use the correct file type. The routine ***assumes*** that the data are ***consistent*** with the type of model chosen. For a multinomial logit model, the routine will treat each record as a separate decision maker and will estimate a model for each choice less the reference choice. For a conditional logit model, the routine will treat each record as one of *J* choices (where *J* is defined by the user – see below) and will estimate a single model for the decision.

The user needs to be very careful that the correct data set is used with the appropriate model because the routine can estimate its equations with either of these data sets. That is, if the data set is appropriate for the multinomial logit model but the user specifies a conditional logit model, the routine will estimate a single equation treating multiples of *J* records as a single decision maker. Similarly, if the data set is appropriate for a conditional logit model but the user specifies a multinomial logit model, the routine will treat each record as if it were a separate decision maker and will estimate one equation for each choice that it finds in the choice variable. The results in both these cases will be meaningless since the there is a mismatch between the data set and the type of model selected. In short, the user should be aware of this.

Reference alternative *(multinomial logit model only)*

For the multinomial logit model, the user should specify which choice is to be used as the reference. The constant and the coefficients for the reference choice will automatically be 0. The user should specify a particular choice from the list of available alternatives or select the most frequently used alternative as the reference choice. Keep in mind that the coefficients will change depending on which alternative is selected as the reference choice since a comparison is always relative. This will affect the interpretation of the coefficients though not the estimated probabilities.

For the conditional logit model, however, there is no reference choice. Therefore, this field will be blanked out when the type of discrete choice model is conditional.

Case ID *(conditional logit model only)*

When a conditional logit model is estimated, each case contributes multiple records to the data file (as many as there are alternatives). In order for *CrimeStat* to know which records belong to the same case (decision maker), the user must specify a Case ID variable, i.e. a variable that uniquely identifies cases (decision makers). If the data set was created with the *CrimeStat* 'Create Data set for Conditional Logit Model' routine, the variable is the Case ID variable specified in that routine.

Output for the Discrete Choice Model

The output includes both summary statistics and individual variable coefficients estimates. The output will vary between the multinomial logit and conditional logit models.

Discrete Choice Model Summary Statistics

The summary statistics include:

Information about the model

1. Date and time
2. The data file
3. The dependent (choice) variable
4. The number of records

5. The degrees of freedom
6. The type of choice model (multinomial discrete or conditional discrete)
7. Number of alternatives
8. The method of estimation (MLE – maximum likelihood estimation, only in this version).

Discrete choice model likelihood statistics

9. Log-likelihood estimate, which is a negative number. For a set number of independent variables, the smaller the log-likelihood (i.e., the most negative) the better.
10. Log-likelihood per case. Smaller (more negative) values are better. This is useful when comparing a similar model but with different numbers of records.
11. Akaike Information Criterion (AIC) adjusts the log-likelihood for the degrees of freedom. The smaller the AIC, the better.
12. AIC per case. Smaller values are better.
13. Bayesian Information Criterion (BIC), sometimes known as the Schwartz Criterion (SC), adjusts the log-likelihood for the degrees of freedom. The smaller the BIC, the better.
14. BIC per case. Smaller values are better.
15. Mean Absolute Deviation (MAD). For a set number of independent variables, a smaller MAD is better.
16. Mean Squared Predictive Error (MSPE). For a set number of independent variables, a smaller MSPE is better.

Discrete Choice Individual Coefficients Statistics

There is a different coefficient output for the multinomial logit model than for the conditional logit model. The multinomial logit model will output constants ***and*** individual coefficients for each of *J*-1 alternatives (where *J* is the total number of alternatives). The constant and coefficients for the reference alternative are automatically defined as zero (0). For example, if there are four alternatives, then three sets of equations will be output, one for each of the *J-1* (4-1=3) alternatives.

The coefficients are always relative to the reference alternative. Therefore, a positive coefficient indicates that the independent variable contributes more for that alternative than for the reference alternative while a negative coefficient indicates that the independent variable contributes less for that choice than for the reference choice. The significance test of the coefficient indicates whether the difference is statistically significant or not compared to the reference alternative. Note that the multinomial logit model *always* has a constant.

On the other hand, the conditional logit model will output a single set of individual coefficients with ***no*** constant. There is no reference choice and the coefficients are relative to not choosing a particular alternative (i.e., having a value of 0 for CHOSEN).

For the individual coefficients, the following are output for each independent variable:

1. The coefficient.
2. The standard error of the coefficient.
3. t-value.
4. p-value. This is the two-tail probability level associated with the t-test.
5. Odds ratio. This is the exponentiation of the coefficient (i.e., e^{β}). It indicates the relative odds of that variable affecting the choice relative to the reference choice (multinomial logit model) or relative to 0 (conditional logit model).

Average predicted probability

For the conditional logit model only, an additional table is output that indicates the average predicted probability of the model for those cases that were selected (i.e., in which CHOSEN=1), for those cases that were not selected (i.e., in which CHOSEN=0), and for all cases. The number of records associated with each category and the standard deviation are given.

Multicollinearity Among Independent Variables in the Discrete Choice Model

A major consideration in any regression model (including discrete choice) is that the independent variables are statistically independent. Non-independence is called *Multicollinearity* and means that there is overlap in prediction among two or more independent variables. This can lead to uncertainty in interpreting coefficients as well as to an unstable model that may not hold in the future. Generally, it is a good idea to reduce Multicollinearity as much as possible.

A tolerance test is given for each coefficient. This is defined as 1 – the R-square of the independent variable predicted by the remaining independent variables in the equation using an Ordinary Least Squares model. It is an indicator of how much the remaining variables in a model account for the variance of any particular independent variable. Since the method uses the Ordinary Least Squares (OLS) methods, it is an approximate (pseudo) test for the discrete choice routines. OLS assumes normality and constant residual errors. However, many independent variables are not normally distributed (e.g., income, distance traveled, number of persons living in poverty).

Consequently, the use of OLS to test for Multicollinearity is exact only when the independent variable being examined for tolerance is normally distributed; otherwise, it is an approximate test. Nevertheless, it is useful indicator of multicollinearity. If the tolerance is low, that definitely indicates that there is multicollinearity. On the other hand, a high tolerance level does not necessarily indicate that there is little multicollinearity. From the test, a guidance message is displayed that indicates probable or possible Multicollinearity. If there is substantial Multicollinearity (indicated by low tolerance values), it is a good idea is to drop one of the multicolinear independent variables and re-run the model.

Save Output

The output from the discrete choice model can be saved.

Saved Multinomial Logit Output

For the multinomial logit model, the output is a ‘dbf’ file that includes all the input variables along with the estimated probability for each choice and the residual error for each choice (the observed choice, 1 or 0, minus the predicted probability). The probability and residual error is presented for each of the *J* alternatives. These are labeled with a ‘P_ ‘ for probability and ‘R_’ for residual error. The different alternatives are indicated by a subscript from 0 (for the reference choice) through *J-1* (for the other alternatives) in the same order in which they are listed in Reference Choice dialogue (excluding the reference choice itself). For example, P_Choice0 is the estimated probability for choice 0 (the reference choice) while R_Choice3 is the estimated residual error for choice 3 (the third one listed in the list under Reference Choice excluding the reference choice itself).

Saved Conditional Logit Output

For the conditional logit model, the output is a ‘dbf’ file and includes all the input variables along with the estimated probability and the residual error for the case. For each case ID, there will be only one record that was chosen. Further, since the conditional logit model produces only one equation, there is only one probability and one residual error. The probability is labeled PREDPROB and the residual error is labeled RESID. The residual error can be used to compare different models. The MAD and MSPE statistics (discussed above) summarize the residual errors. But, a user might want to plot the residuals against one of the independent

Save Estimated Coefficients

The coefficients from either the multinomial logit or the conditional logit models can be saved for use with other data sets. Specify a directory where the coefficients file is to be saved and provide a root name. The saved coefficients file for the multinomial logit model will have a DCCoeffMNL prefix while the saved coefficients file for the conditional logit model will have a DCCoeffCNL prefix before the user defined root name.

Discrete Choice Modeling II

The Discrete Choice II module allows the user to apply the estimated coefficients from a discrete choice model to another data set (or a subset of the same data set) and calculate predicted probabilities, for either the multinomial logit or the conditional logit model. The 'Make prediction' routine allows the application of coefficients to a data set. The saved coefficients are applied to similar independent variables and to corresponding values of the choice variable to produce an estimated probability of an alternative.

Make Prediction

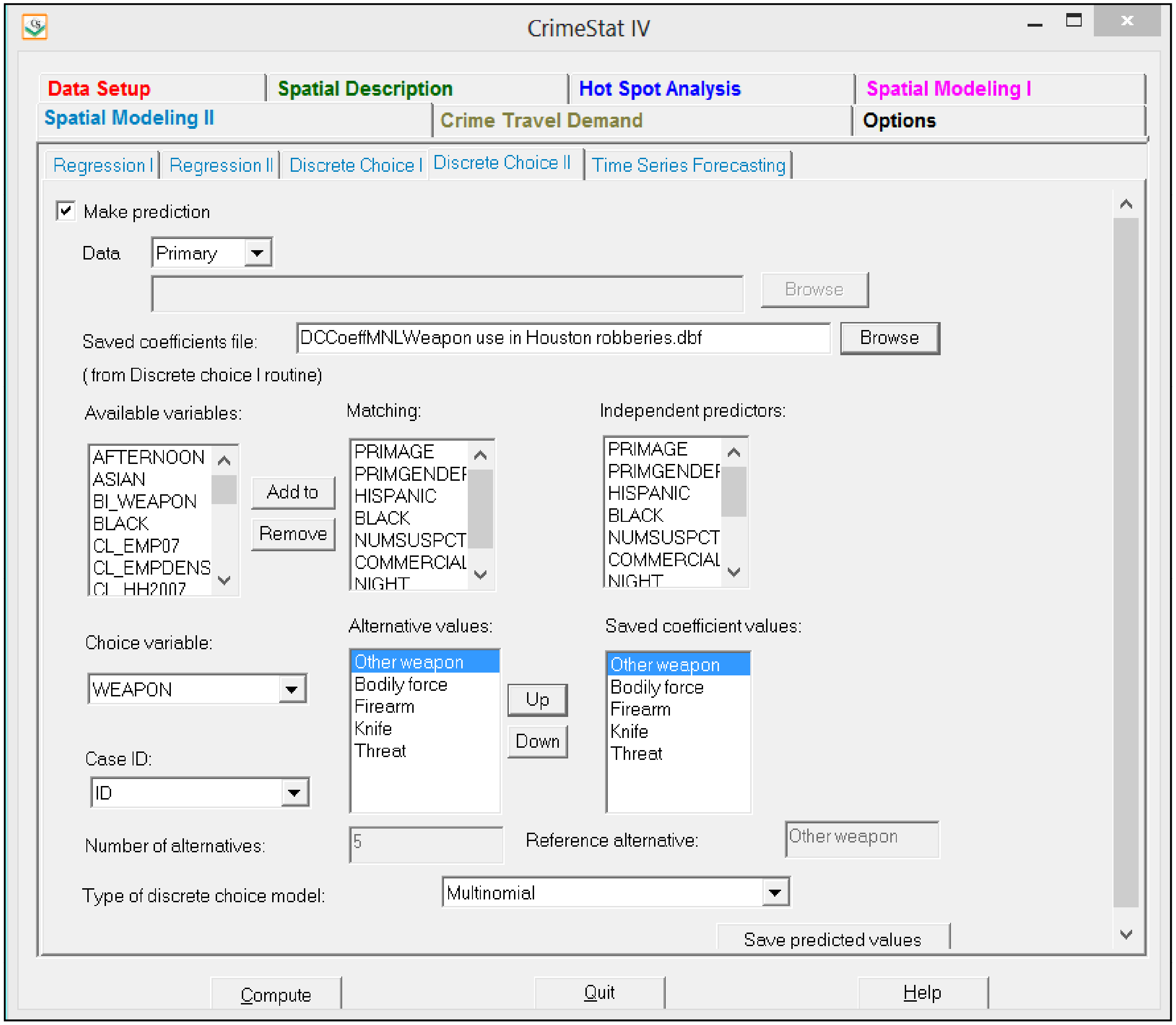
There are two types of models that can be fitted – multinomial logit or conditional logit. For both types of model, the coefficients file must include information on each of the coefficients. In addition, the coefficients model for the multinomial must include the value of the constant. The user reads in the saved coefficient file and matches the variables to those in the new data set based on the order of the coefficients file.

Discrete Choice Data File

The new data set can be either the Primary file or another file. If another file is being used, point to the directory where it is stored and identify it. The structure of the file for which a prediction is made must be the same as that from which the model was initially calibrated. That is, for a multinomial logit prediction, there must be a file with one record per decision maker and which includes and ID and each of the independent variables used in the prediction. For a

Figure 2.20:

Discrete Choice Modeling II



☑ Make prediction

Data: Primary

Saved coefficients file: DCCoeffMNLWeapon use in Houston robberies.dbf

(from Discrete choice I routine)

Available variables:
AFTERNOON
ASIAN
BI_WEAPON
BLACK
CL_EMP07
CL_EMPDENS
CL_HH2007

Matching:
PRIMAGE
PRIMGENDEF
HISPANIC
BLACK
NUMSUSPCT
COMMERCIAL
NIGHT

Independent predictors:
PRIMAGE
PRIMGENDEF
HISPANIC
BLACK
NUMSUSPCT
COMMERCIAL
NIGHT

Choice variable: WEAPON

Case ID: ID

Alternative values:
Other weapon
Bodily force
Firearm
Knife
Threat

Saved coefficient values:
Other weapon
Bodily force
Firearm
Knife
Threat

Number of alternatives: 5

Reference alternative: Other weapon

Type of discrete choice model: Multinomial conditional logit prediction, there must be a joined file with a record for every combination of case and alternative.

Discrete Choice Saved Coefficients File

In order to make a prediction, a model must have already been estimated and the coefficients saved in a coefficients file. Point to the directory where the coefficients file has been saved and identify it.

Available Variables

The box labeled 'Available variables' will list all the fields on the input data set.

Independent Predictors

The independent variables that were used in the estimated coefficients file will be listed in the right column. They will be in the same order as was estimated in the calibration file.

Matching variables

Select corresponding variables from the input data file for the middle column. The items should be listed in the same order as in the 'independent predictors' column. They should be similar variables in content but need not have the names as in the original data file.

Alternative Values (multinomial logit only)

The values of the choice variables from the input file will be displayed in the middle column. The order should match the values in the adjacent saved coefficients file column. The 'Up' and 'Down' buttons can be used to re-order the values to be sure they are matched exactly.

Saved coefficient values *(multinomial logit model only)*

The values of the saved coefficients file will be displayed in the right column. Additional values can be added with the "Add to" button and existing values can be removed with the "Remove" button. It is essential that the values in the middle column match ***exactly*** their corresponding values in the right column.

Reference alternative *(multinomial logit only)*

The reference alternative value is displayed. If it is not correct, type in the correct value to be used or, better yet, re-estimate the original model. This field will be blanked out for the conditional logit model since it is not appropriate.

Discrete Choice Prediction Output

The screen output provides predictions of the value of the dependent variable in the same order as in the input data set. For the multinomial logit model, the predictions are labeled as CHOICE0 (for the reference choice), CHOICE1, CHOICE2, and so forth, in the same order as in the input data set. For each alternative, these predictions represent the probability that this alternative is chosen, given the values of the predictor variables.

For the conditional logit model, the prediction is applied to each available alternative. The screen output presents the predictions in matrix format with the case ID listed on the vertical axis and the choices listed on the horizontal axis (labeled CHOICE0, CHOICE1, CHOICE2, and so forth, in the same order as in the input data set).

Save Predicted Values for Discrete Choice Prediction

The predicted values and the residual errors can be output to a DBF file with a DCMakePredMNL<*root name*> for the multinomial logit and DCMakePredCNL<*root name*> for the conditional logit with the root name being provided by the user. The output files differ between the multinomial and conditional logit models.

Multinomial Logit Prediction Output

For the multinomial logit prediction, there is probability produced for each of the *J* alternatives. The probabilities are labeled P_CHOICE0 (for the reference choice), P_CHOICE1, P_CHOICE2, and so forth in the same order as in the Choice Values dialogue (with the exception of the reference choice which is always defined as P_CHOICE0). The probabilities will sum to 1.0 for all alternatives (within rounding-off error).

Conditional Logit Prediction Output

For the conditional logit prediction, there is a single probability output which is applied to the particular record. Since the data for the conditional logit model has a single record for each choice faced by the decision maker, the probability applies to that choice. The probabilities will sum to 1.0 for all alternatives (within rounding-off error). The column is labeled PREDPROB.

Time Series Forecasting

The Time Series Forecasting module is designed for the forecasting of crime or other counts by specific geographical areas (districts) and the detection of unusual levels of activity above-and-beyond the forecast. The methods are useful for tactical deployment of police resources but can be used by other fields where the monitoring of events by time is a regular part of their procedures. The module has a single interface page. It requires a user to specify an input file – either the Primary File or another file, identify variables in the file used for forecasting, select a seasonality adjustment, specify an exponential smoothing model, turn on the Trigg Tracking Signal, define Trigg parameter values, and save the output.

Input File

This is the file with the data for the Time Series Forecasting module. The data set for the regression module can be the Primary file or another file. If it is the Primary file, then it must have X and Y coordinates defined on each record. If it is another file, click on 'Other' and then identify the file. Only 'dbf' or 'txt' files are allowed.

Each record represents a unique combination of an area unit and a season number. A minimum of three years worth of data is required. For example, if there are 20 districts and monthly counts of the number of events over three years, then there will be 720 records (20 districts x 12 months x 3 years).

Areal Unit

The areal unit is the name or identifier for the district of the incident being forecasted. The name can be alphanumeric or numeric.

Year

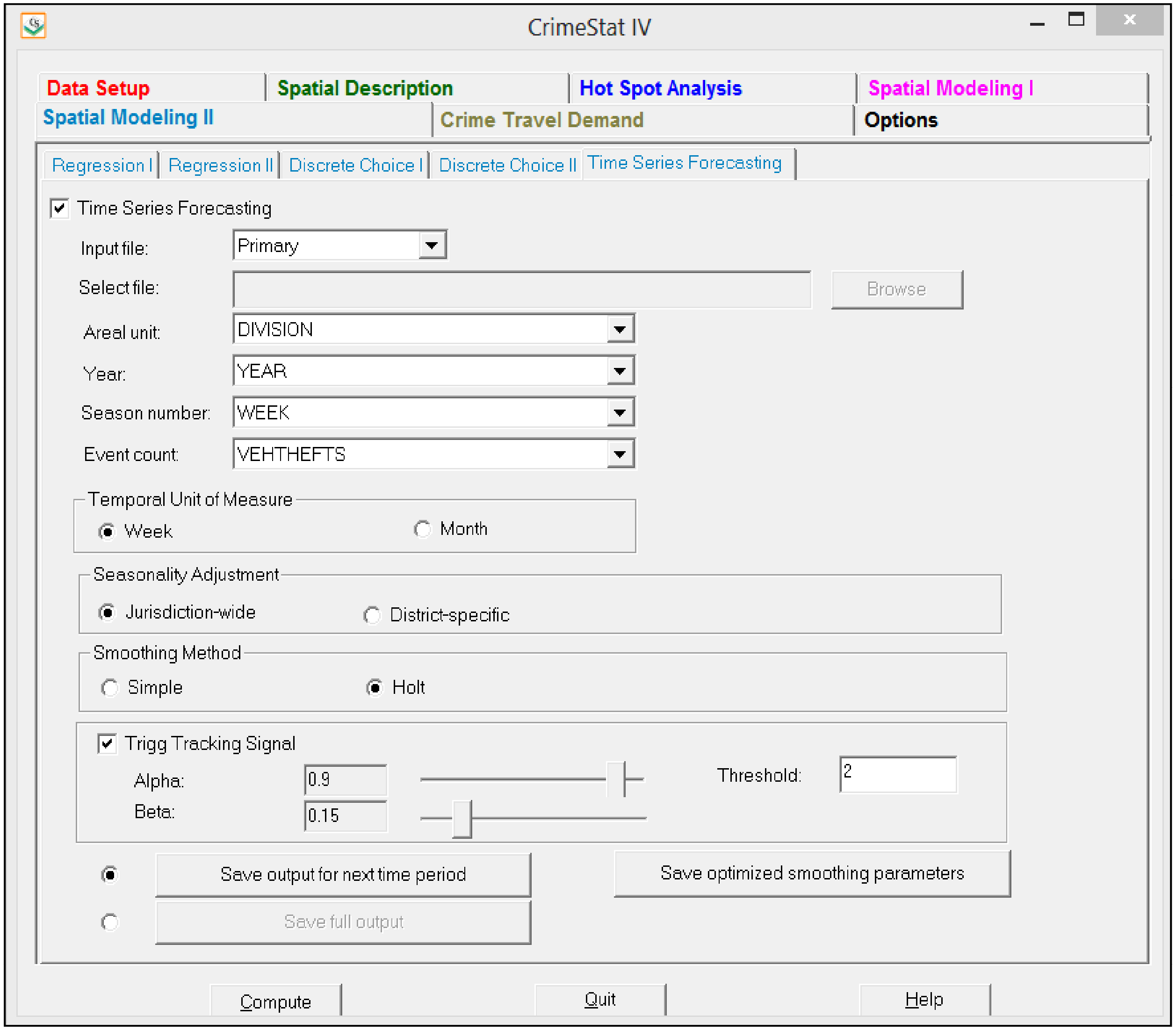
The year is the calendar year such as 2012 of each data record. This must be recorded. As mentioned above, there must be at least three years of data. This is a numeric variable.

Season Number

This is the season number. A season is the unique temporal identifier. With this module, only months or weeks are allowed. Thus, the season number is 1 through 12 for months and 1 through 52 for weeks. Note that there cannot be partial weeks. Since a year has 365 or

Figure 2.21:

Time Series Forecasting



☑ Time Series Forecasting

Input file: Primary

Select file:

Areal unit: DIVISION

Year: YEAR

Season number: WEEK

Event count: VEHTHEFTS

Temporal Unit of Measure: ● Week ○ Month

Seasonality Adjustment: ● Jurisdiction-wide ○ District-specific

Smoothing Method: ○ Simple ● Holt

☑ Trigg Tracking Signal

Alpha: 0.9

Beta: 0.15

Threshold: 2

● Save output for next time period

○ Save full output

366 days, there are 1 or 2 extra days left over. These must be assigned to either the first week of the next year or the last week of the current year.

Event Count

This is the count of the number of events for a given areal unit, year, and time period.

Temporal Unit of Measure

This field defines the type of season used, either week or month.

Seasonality Adjustment

The seasonality adjustment is the adjustment made for each time observation for seasonal patterns such as when, for example, crime is low in February and high in July relative to the time series trend line. The routine uses either the data from the entire jurisdiction (e.g., the entire city) - jurisdiction-wide, and applies this to each district or it uses individual data from each district so that each gets its own unique seasonal pattern - district-specific.

Smoothing Method

The smoothing method provides a more reliable estimate of the expected number of events based on past trends. The routine provides two alternative models, simple smoothing or Holt exponential smoothing. Simple smoothing assumes that there is no trend and that future values will follow past values. Holt smoothing adds a trend line into the expected number of future events. The models have smoothing parameters which CrimeStat automatically chooses by minimizing one-step-ahead forecast errors.

Trigg Tracking Signal

The Trigg Tracking Signal provides a test statistic for unusual activity in the number of events. If the absolute value of the signal exceeds a pre-specified threshold value, then there is a "signal trip" meaning that it is likely that there is an unusual change in events. The signal has three parameters with default values provided, alpha, beta and the threshold value.

Alpha and beta are parameters that vary between 0 and 1. An alpha of 0.9 makes the tracking signal very reactive to current data on the anticipation of changes in a time series pattern. A value of beta of 0.15 smoothes the measure of spread used to standardize the Trigg signal and retains some history. Cohen, Garman, and Gorr (2009) found that these are the best performing parameter values. However, the user can experiment.

Alpha

Alpha is a smoothing parameter that varies between 0 and 1. An alpha of 0.9 (the default value) makes the tracking signal very reactive to current data on the anticipation of changes in a time series pattern. Note that "Alpha" is the same parameter as used in simple exponential smoothing for forecasting, but here is used to smooth the Trigg tracking signal instead of crime counts. Decreasing the parameter alpha below 0.9 will reduce the importance of more recent events.

Beta

Beta is a smoothing parameter that varies between 0 and 1. A value of beta of 0.15 (the default value) smooths the measure of spread used to standardize the Trigg signal and retains a good amount of history while allowing estimates to drift and follow changing spread in the data. Increasing beta above 0.15 will smooth the data more and will reduce the Trigg more towards the mean.

Threshold

The threshold is the value of the Trigg Tracking Signal that indicates whether the expected number of events will be greater than what is normally expected ("business-as-usual"). The default threshold of 1.5 is somewhat liberal in the sense that it will signal more periods of unusual activity. However, most police organizations would rather respond to more expected events even if the increased activity does not materialize (i.e., are false positives) than not respond and have events blow up. To use more conservative values, try 1.75 or 2.0 to get fewer signal trips.

Output

There are three types of output – full, one-step ahead, and the optimized smoothing parameters. The first two outputs produce the following calculated values:

1. DE_SEASON is the number of events per period (EVENTCOUNT) divided by the seasonal factor for the current observation's season (December) and, thus, is a de-seasonalized count of events. To calculate the seasonal factor for each record divide EVENTCOUNT by DE_SEASON.

2. SMTH_LEVEL is the smoothed estimate for the current observation (e.g., December 2012).

3. When using the Holt smoothing method, there is one additional estimated parameter. SMTH_SLOPE is the change in estimated crime for each step ahead. If, for example, you need the forecasts for February 2013 and your current time period is December 2012, you add two times SMTH_SLOPE to SMTH_LEVEL because February 2013 is two steps ahead of December 2012.

4. SQ_ERROR is the squared forecast error of the current observation from the forecast made for it from the previous period (e.g., November 2012 if the current period is December 2012).

5. TRIGG is the value of the Trigg Tracking Signal for the current observation.

6. SIGNALTRIP indicates whether the Trigg level was higher than the threshold. If it was, this field will have a **1** to indicate that the Trigg value was greater than or equal to the threshold selected and the detected change is an increase, a **-1** if the Trigg value is greater than or equal to the threshold but the detected change was a decrease, and a **0** otherwise.

7. FORECAST is the one-step-ahead forecast, for the next observation in time. For example, if the the current period is December 2012, then one-step ahead forecast is for January 2013. For a January 2013 forecast and simple exponential smoothing it is SMTH_LEVEL for December 2012 multiplied by the seasonal factor for January 2013. For January 2013 and Holt smoothing it is the sum of SMTH_LEVEL and SMTH_SLOPE times the seasonal factor for January 2013.

Save Full Output

The full output includes all input fields plus the calculated values. If the user clicks the Save full output button and then clicks the Save full output button, a save output window opens. Select dBase ‘DBF’ for the Save output to field, browse to the folder of your choice, and type a file name. Both the input data and the one-step ahead forecast are output to the screen and to a ‘dbf’ file. The file will be saved with a “TS_F” prefix before the defined file name.

Save Output for Next Time Period

The next time period output includes only the calculated fields for both the screen and saved file. The word "next" refers to the forecast made for the next time period, while the Trigg tracking signal evaluates the current period. Again, in the dialog for saving the output file, type the .dbf extension in the chosen file name. The file is saved as a 'dbf' file with a "TS_C" (for 'current') prefix.

Save Optimized Smoothing Parameters

The third type of output shows the results of the optimization process for exponential smoothing. This provides information on the parameters used to optimize the smoothing for each district. Define the file name and it will be saved as an ASCII text file with a 'txt' extension. The output fields are:

1. Optimum Alpha is the smoothing parameter value for *level* of a time series that minimizes the one-step-ahead forecast sum of squared errors.

2. Optimum Gamma is the smoothing parameter value for *time trend slope* of a time series that minimizes the one-step-ahead forecast sum of squared errors.

3. SSE is the resulting optimal sum of squared errors for the time series.

It is valuable to review the optimal parameters to see which areas have stable versus dynamic time series. Note that for the Trigg calculation, we want a large alpha to detect large changes in the number of recent events. That is why the default value of alpha is 0.9. However, for forecasting, we want a low alpha in order to smooth the data to produce a stable forecast.

VI. Crime Travel Demand Modeling

The crime travel demand module is a sequential model of crime travel by zone over a metropolitan area. Crime incidents are allocated to zones, both by the location where the crime occurred (destinations) and the location where the offender started (origins). A crime trip is defined as a crime event that originates at one location and ends at another location; the two locations can be the same. For each zone, the number of crimes originating in the zone and the number of crimes ending (occurring) in that zone are enumerated. Thus, the model is for count (or volumes), not rates. Other zonal data must be obtained to be used as predictor variables of the origin and destination counts.

The model is made up four sequential steps, each of which can involve smaller steps:

1. Trip generation – separate models are developed for predicting the number of crimes originating or ending in each zone. There are, therefore, two models. One is a model of the predicted number of crime trips that originate in each zone while the other is a model of the predicted number of crime trips that end in each zone. The number of origin zones can be greater than the number of destination zones.

2. Trip distribution – A model is developed for the number of crimes originating in each zone that go to each destination zone. The result is a prediction of the number of crimes originating in each zone that end in each zone (trip links).

3. Mode split – A model is developed that splits the number of predicted trips from each origin zone to each destination zone by travel mode (e.g., walking, bicycle, driving, bus, train). Thus, each zone-to-zone trip link is separated into different travel modes.

4. Network assignment – A model is developed for the route taken for each crime trip link (whether for all modes or by separate modes). Thus, the shortest path through a network is determined. Different travel modes will have different routes since bus and train, in particular, must use a separate network.

Crime Travel Demand Data Preparation

In order to run the crime travel demand module, particular data must be obtained and prepared. These involve:

1. A zonal framework that will be used for the modeling. In general, it is best to select the smallest zone size for which data can be obtained (e.g., block groups, census tracts, traffic analysis zones). However, it is often difficult to obtain data for the smallest units (e.g., blocks, grid cells). The larger the zone size, the more there will be intra-zonal trips and the greater the error in the model. Thus, the user must balance the need for small zones with the availability of data. Since crimes can occur outside a study area, the number of origin zones can be (and probably should be) greater than the number of destination zones. However, each destination zone should be included within the origin zone collection. Typically, there will be separate data sets for the origin zones and for the destination zones.

2. Data on crime origins and crime destinations are obtained (usually from arrest records) and are allocated to zones. The incidents are then summed by zone to produce a count. The "Assign primary points to secondary points" routine (under Distance analysis) can be used for this purpose. Thus, each origin zone has a count of the number of crimes originating in that zone and each destination zone has separate counts of the number of crimes originating in that zone and the number of crimes occurring (ending) in that zone. Crimes can be sub-divided into types (e.g., robbery, burglary, vehicle theft).

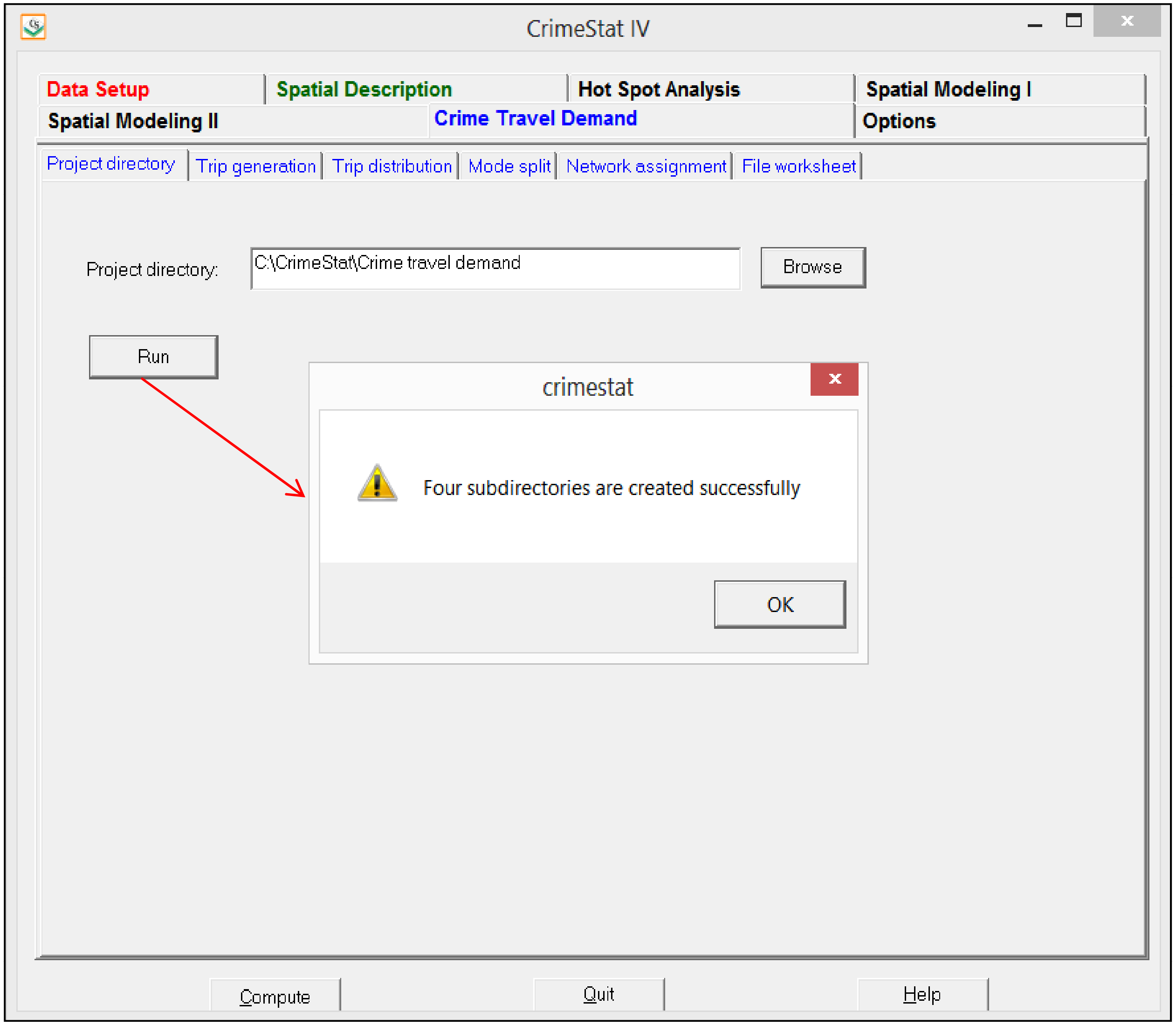
3. Additional data for the zones are obtained. These would include population (or households), sub-populations (e.g., age groups, race/ethnic groups), income levels, poverty levels, employment (retail and non-retail), land use, particular types of land use (e.g., drug locations, markets, parking lots), policing variables (e.g., personnel deployment, beat frequency), intervention variables (e.g., drug treatment centers), and other variables. It's important that all variables included must cover all zones for either the origin data set or the destination data set. For example, if poverty is used a variable in the origin model, then all origin zones must have an enumeration of poverty. Similarly, if retail employment is used as a variable in the destination model, then all destination zones must have an enumeration of retail employment.

4. Data on dummy variables and special generators are also obtained. Dummy variables would be a proxy for a condition that does or does not exist. Zones that have the condition are assigned a '1' whereas zones that do not have the condition are assigned a '0'. For example, if a freeway cross a zone, then a freeway dummy variable would assign '1' to that zone (and all others that the freeway crossed) whereas all other zones received a '0' for this variable. A special generator is a land use that attracts trips (e.g., a stadium, a railroad station). All zones that have the special generator are assigned a value whereas all other zones receive a '0'; the value can either be a dummy variable (i.e., a '1') or the actual count if that can be obtained (e.g., the number of patrons at a football stadium event).

Project Directory

The Crime Travel Demand module is a complex model that involves many different files. Because of this, we recommend that the separate steps in the model be stored in separate directories under a main project directory. While the user can save any file to any directory

Figure 2.22:
Crime Travel Demand Project Directory



Project directory: C:\CrimeStat\Crime travel demand

Four subdirectories are created successfully within the module, keeping the inputs and output files in separate directories can make it easier to identify files as well as examine files that have already been used at some later time.

Project Directory Utility

The project directory utility allows the creation of a master directory for a project and four separate sub-directories under the master directory that correspond to the four modeling stages.

The user puts in the name of a project in the dialogue box and points it to a particular drive and directory location (depending on the number of drives available to the user). For example, a project directory might be called "Robberies 2003" or "Bank robberies 2005". The utility then creates this directory if it does not already exist and creates four sub-directories underneath the project directory:

1. Trip generation
2. Trip distribution
3. Mode split
4. Network assignment

The user can then save the different output files into the appropriate directories. Further, for each sequential step in the crime travel demand model, the user can easily find the output file from the previous step which would become input file for the next step.

Trip Generation

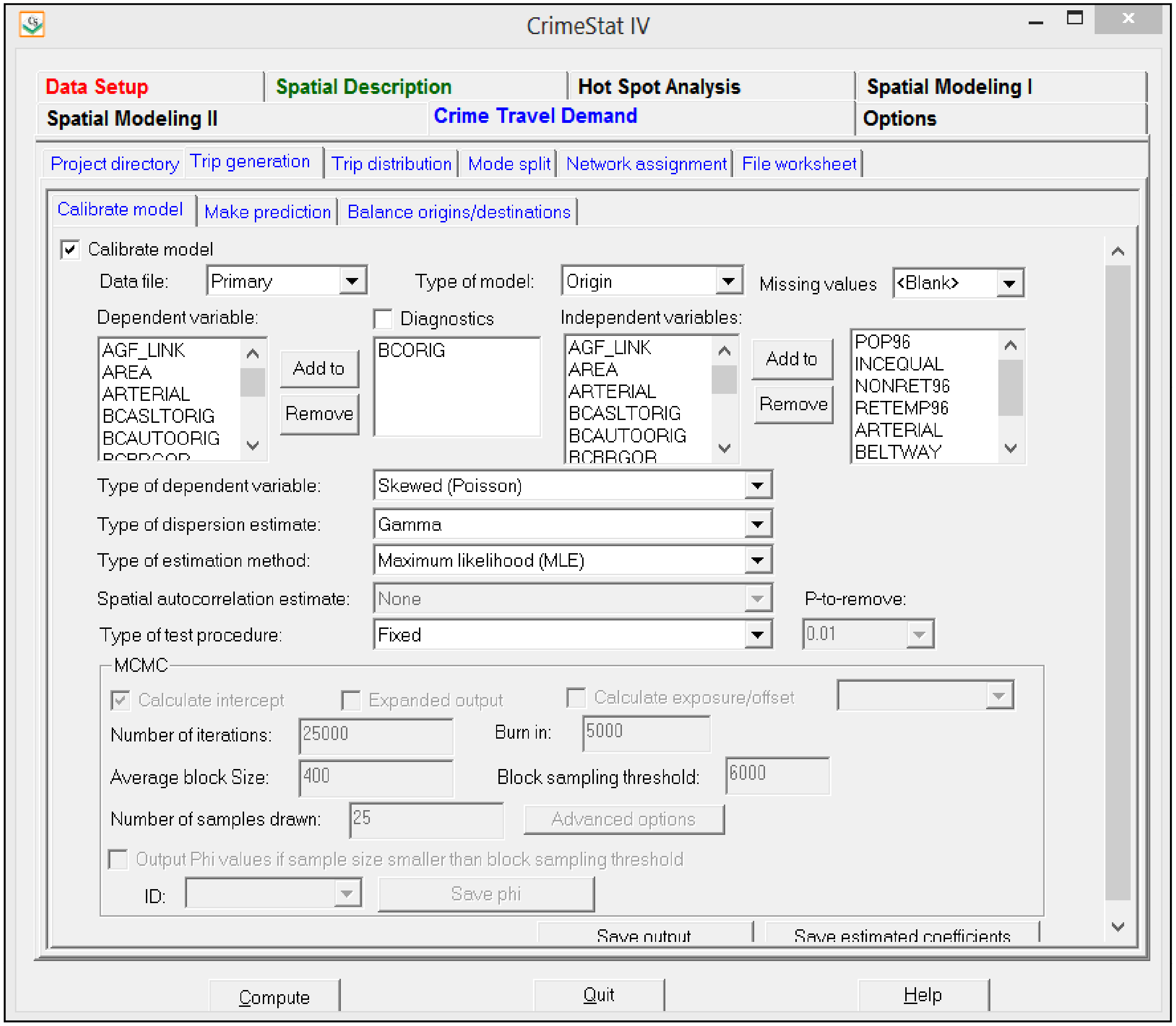
Trip generation involves the development of separate models for predicting the number of crimes originating in each zone and the number of crimes occurring (ending) in each zone. There are three steps to the trip generation:

1. **Calibrate model.** A step that calibrates the model against known data using regression techniques. The result is a prediction of the number of trips either originating in a zone (the origin model) or the number of trips ending in a zone (the destination model).

2. **Make prediction**. A step that applies the calibrated model to a data set and also allows the addition of trips from outside the study area (external trips).

Figure 2.23:

Trip Generation Modeling



☑ Calibrate model

Data file: Primary

Type of model: Origin

Missing values: <Blank>

Dependent variable: AGF_LINK, AREA, ARTERIAL, BCASLTORIG, BCAUTOORIG, BCBRGOR

□ Diagnostics

BCORIG

Independent variables: AGF_LINK, AREA, ARTERIAL, BCASLTORIG, BCAUTOORIG, BCBRGOR

POP96, INCEQUAL, NONRET96, RETEMP96, ARTERIAL, BELTWAY

Type of dependent variable: Skewed (Poisson)

Type of dispersion estimate: Gamma

Type of estimation method: Maximum likelihood (MLE)

Spatial autocorrelation estimate: None

P-to-remove: 0.01

Type of test procedure: Fixed

MCMC

☑ Calculate intercept

□ Expanded output

□ Calculate exposure/offset

Number of iterations: 25000

Burn in: 5000

Average block Size: 400

Block sampling threshold: 6000

Number of samples drawn: 25

□ Output Phi values if sample size smaller than block sampling threshold

ID:

1. **Balance predicted origins & destinations**. A step that ensures that the number of predicted origins equals the number of predicted destinations. Since a trip involves an origin and a destination, it is essential that the number of origins equal the number of destinations.

Calibrate Trip Generation Model

This step involves calibrating a regression model against the zonal data. Two separate models are developed, one for trip origins and one for trip destinations. The dependent variable is the number of crimes originating in a zone (for the trip origin model) or the number of crimes ending in a zone (for the trip destination model). The independent variables are zonal variables that may predict the number of origins or destinations.

In the current version, 13 possible regression models are available with several options for each of these:

MLE Normal (OLS)
MCMC Normal
MCMC Normal-CAR
MCMC Normal-SAR
MLE Poisson
MLE Poisson with linear dispersion correction (NB1)
MLE Poisson-Gamma (NB2)
MCMC Poisson-Gamma (NB2)
MCMC Poisson-Gamma-CAR
MCMC Poisson-Gamma-SAR
MCMC Poisson-Lognormal
MCMC Poisson-Lognormal-CAR
MCMC Poisson-Lognormal-SAR

Since the Regression I module and Trip Generation module duplicate most of the regression functions, only one of these can be run at a time.

Input Data set

The data set for the trip generation must be the Primary File data set. The coordinate system and distance units are also the same.

Dependent Variable

To start loading the module, click on the ‘Calibrate model’ tab. A list of variables from the Primary File is displayed. There is a box for defining the dependent variable. The user must choose one dependent variable.

Independent Variables

There is a box for defining the independent variables. The user must choose one or more independent variables. There is no limit to the number. The variables are output in the same order as specified in the dialogue so a user should consider how these are to be displayed.

Model decisions

There are five decisions that must be made for each regression model.

Type of Dependent Variable

The first model decision is the type of dependent variable The first model decision is the type of dependent variable: Skewed (Poisson) or Normal (OLS). The default is a Poisson.

Type of Dispersion Estimate

The second model decision is the type of dispersion estimate to be used. The choices are Gamma, Poisson, Poisson with linear correction, Normal (automatically defined for the Normal model), or lognormal. The default is Gamma.

Type of Estimation Method

The third model decision is the type of estimation method to be used: Maximum Likelihood (MLE) or Markov Chain Monte Carlo (MCMC). The default is MLE.

Spatial Autocorrelation Regression Model

If the user accepts an MCMC algorithm, then a fourth decision is whether to run a spatial autocorrelation estimate along with it. This can only be run if the dependent variable is Poisson and MCMC has been chosen as the type of estimation method. The spatial autocorrelation choices are Conditional Autoregressive (CAR) or Simultaneous Autoregression (SAR).

Type of Test Procedure

The fifth, and last model decision, is whether to run a fixed model or a backward elimination *stepwise* procedure (only with an MLE model). A fixed model includes all selected independent variables in the regression whereas a backward elimination model starts with all selected variables in the model but proceeds to drop variables that fail the P-to-remove test, one at a time. Any variable that has a significance level in excess of the P-to-remove value is dropped from the equation.

Specify whether a fixed model (all selected independent variables are used in the regression) or a backward elimination stepwise model is used. The default is a fixed model. If a backward elimination stepwise model is selected, choose the P-to-remove value (default is .01).

MCMC Model Choices

If the user chooses the MCMC algorithm, then eight additional decisions have to be made.

Number of Iterations

The first MCMC decision is the number of iterations to be run. The default is 25,000. The number should be sufficient to produce reliable estimates of the parameters. Check the MC Error/Standard deviation ratio and the G-R statistic after the run to be sure most parameters are below 1.05 and 1.20 respectively. If not, increase the number of iterations and 'burn in' iterations.

'Burn in' iterations

The second MCMC decision is the number of initial iterations that will be dropped from the final distribution (the 'burn in' period). The default is 5,000. The number of 'burn in' iterations should be sufficient for the algorithm to reach an equilibrium state and produce reliable estimates of the parameters. Check the MC Error/Standard deviation ratio and the G-R statistic after the run to be sure most parameters are below 1.05 and 1.20 respectively. If not, increase the number of iterations and 'burn in' iterations.

Block Sampling Threshold

The third MCMC decision is whether to run all the records through the MCMC algorithm or whether to draw block samples. The algorithm will be run on all records unless the number of records exceeds the block sampling threshold. The default threshold is 6000 records. To run all the records through the MCMC algorithm, change this value to be greater than the number of records in the database. Note that calculation time will increase substantially if all records in a large database are run through the algorithm.

Average Block Size

The fourth MCMC decision is the number of records to be drawn for each block sample if the total number of records is greater than the block sampling threshold. The default is 400 records per block sample. Note that this is an average. Actual samples will vary in size. The output will display the expected sample size and the average sample size that was drawn.

Number of Samples Drawn

The fifth MCMC decision is the number of samples to be drawn if the total number of records is greater than the block sampling threshold. The default is 25 block samples. Typically, 20-30 block samples will achieve stable model results.

Calculate Intercept

The sixth MCMC decision is whether to run a model with or without an intercept (constant). The default is with an intercept estimated. To run the model without the intercept, uncheck the 'Calculate intercept' box.

Calculate Exposure/Offset

The seventh MCMC decision is whether to run a risk model. If the model is a risk or rate model, then an exposure (offset) variable needs to be defined. Check the 'Calculate exposure/offset' box and identify the variable that will be used as the exposure variable. The coefficient for this variable will automatically be 1.0.

Advanced Options

The eighth MCMC decision is the prior values used for the different parameters being estimated. The MCMC algorithm requires an initial estimate for each parameter. There is a dialogue of advanced options for the MCMC algorithm by which they can be changed.

Initial Parameters Values

For the beta coefficients (including the intercept), the default values are 0. These are displayed as a blank screen for the Beta box. However, other prior estimates of the beta coefficients can be substituted for the assumed 0 coefficients. To do this, all independent variable coefficients plus the intercept (if used) must be listed in the order in which they appear in the model and must be separated by commas. Do not include the beta coefficients for the spatial autocorrelation term (if used) or the error (Taupsi) term.

Taupsi (error term)

The output of the MCMC always includes an error term, called *Taupsi* (τ_ψ). This is an exponent of the error term, $e^{\tau\psi}$, which together is called the *dispersion parameter*. The default value for Taupsi is 1.0. The user can substitute an alternative value.

Rho and Tauphi

The spatial autocorrelation component is made up of three separate sub-components, called Rho, Tauphi, and Alpha and are additive. Rho is roughly a global component that applies to the entire data set. Tauphi is roughly a neighborhood component that applies to a sub-set of the data. Alpha is essentially a localized effect. The default initial values for Rho and Tauphi are 0.5 and 1 respectively. The user can substitute alternative values for these parameters.

Alpha

Alpha is the exponent for the distance decay function in the spatial model. Essentially, the distance decay function defines the weight to be applied to the values of nearby records. The weight can be defined by one of three mathematical functions. First, the weight can be defined by a negative exponential function.

Second, the weight can be defined by a restricted negative exponential with the negative exponential operating up to the specified search distance, whereupon the weight becomes 0 for greater distances.

Third, the weight can be defined as a uniform value for all other observations within a specified search distance. This is a *contiguity* (or adjacency) measure. Essentially, all other observations have an equal weight within the search distance and 0 if they are greater than the search distance.

For the negative exponential and restricted negative exponential functions, substitute the selected value for alpha in the alpha box and for the restricted negative exponential and uniform functions, specify the search distance and distance units. The default is a negative exponential with an alpha of -1.0 in miles.

Value for 0 distance between records

The advanced options dialogue has a parameter for the minimum distance to be assumed between different records. If two records have the same X and Y coordinates (which could happen if the records are individual events, for example), then the distance between these records will be 0. This could cause unusual calculations in estimating spatial effects. Instead, it is more reliable to assume a slight difference in distance between all records. The default is 0.005 miles but the user can modify this (including substituting 0 for the minimal distance).

Output

The output depends on whether an MLE or an MCMC model has been run.

Maximum Likelihood (MLE) Model Output

The MLE routines (Normal/OLS, Poisson, Poisson with linear correction, MLE Poisson-Gamma, Binomial Probit, MLE Binomial Logit) produce a standard output that includes summary statistics and estimates for the individual coefficients.

MLE Summary Statistics

The summary statistics include:

Information about the model

1. The dependent variable
2. The number of records
3. The degrees of freedom (N – number of parameters estimated)
4. The type of regression model (Normal/OLS, Poisson, Poisson with linear correction, Poisson-Gamma, Binomial Probit, Binomial Logit)
5. The method of estimation (MLE)

Likelihood statistics

6. Log-likelihood estimate, which is a negative number. For a set number of independent variables, the more negative the log-likelihood the better.
7. Akaike Information Criterion (AIC) adjusts the log-likelihood for the degrees of freedom. The smaller the AIC, the better.

8. Bayesian Information Criterion (BIC), sometimes known as the Schwartz Criterion (SC), adjusts the log-likelihood for the degrees of freedom. The smaller the BIC, the better.
9. Deviance compares the log-likelihood of the model to the log-likelihood of a model that fits the data perfectly. A smaller deviance is better.
10. The probability value of the deviance based on a Chi-square with k-1 degrees of freedom.
11. Pearson Chi-square is a test of how closely the predicted model fits the data. A smaller Chi-square is better since it indicates the model fits the data well.

Model error estimates

12. Mean Absolute Deviation (MAD). For a set number of independent variables, a smaller MAD is better.
13. Quartiles for the Mean Absolute Deviation. For any one quartile, smaller is better.
14. Mean Squared Predictive Error (MSPE). For a set number of independent variables, a smaller MSPE is better.
15. Quartiles for the Mean Squared Predictive Error. For any one quartile, smaller is better.
16. Squared multiple R (for linear model only). This is the percentage of the dependent variable accounted for by the independent variables.
17. Adjusted squared multiple R (for linear model only). This is the squared multiple R adjusted for degrees of freedom.

Over-dispersion tests

18. Adjusted deviance. This is a measure of the difference between the observed and predicted values (the residual error) adjusted for degrees of freedom. The smaller the adjusted deviance, the better. A value greater than 1 indicates over-dispersion.
19. Adjusted Pearson Chi-square. This is the Pearson Chi-square adjusted for degrees of freedom. The smaller the Pearson Chi-square, the better. A value greater than 1 indicates over-dispersion.
20. Dispersion multiplier. This is the ratio of the expected variance to the expected mean. For a set number of independent variables, the smaller the dispersion multiplier, the better. For example, in a pure Poisson distribution, the dispersion should be 1.0. In practice, a ratio greater than 10 indicates that there is too much

variation that is unaccounted for in the model. Either add more variables or change the functional form of the model

21. Inverse dispersion multiplier. For a set number of independent variables, a larger inverse dispersion multiplier is better. A ratio close to 1.0 is considered good.

MLE Individual Coefficient Statistics

For the individual coefficients, the following are output:

22. The coefficient. This is the estimated value of the coefficient from the maximum likelihood estimate.
23. Standard Error. This is the estimated standard error from the maximum likelihood estimate.
24. Pseudo-tolerance. This is the tolerance value based on a linear prediction of the variable by the other independent variables. See equation Up. 2.18.
25. Z-value. This is asymptotic Z-test that is defined based on the coefficient and standard error. It is defined as Coefficient/Standard Error.
26. p-value. This is the two-tail probability level associated with the Z-test.

Markov Chain Monte Carlo (MCMC) Model Output

The MCMC routines (Poisson-Gamma, Poisson-Gamma-CAR/SAR, Poisson-Lognormal, Poisson-Lognormal-CAR/SAR, Binomial Logit, Binomial Logit-CAR/SAR) produce a standard output and an optional expanded output**.** The standard output includes summary statistics and estimates for the individual coefficients.

MCMC Summary Statistics

The summary statistics include:

Information about the model

1. The dependent variable
2. The number of records
3. The sample number. This is only output when the block sampling method is used.
4. The number of cases for the sample. This is only output when the block sampling method is used.

5. Date and time for sample. This is only output when the block sampling method is used
6. The degrees of freedom (N – number of parameters estimated)
7. The type of regression model (Poisson-Gamma, Poisson-Gamma-CAR/SAR, Poisson-Lognormal, Poisson-Lognormal-CAR/SAR, Binomial Logit, Binomial Logit-CAR/SAR)
8. The method of estimation
9. The number of iterations
10. The 'burn in' period
11. The distance decay function used. This is output for CAR/SAR models only.
12. The block size is the expected number of records selected for each block sample. The actual number may vary.
13. The number of samples drawn, output when the block sampling method used.
14. The average block size. This is output when the block sampling method used.
15. The type of distance decay function. This is output for CAR/SAR models only.
16. Condition number for the distance matrix. If the condition number is large, then the model may not have properly converged. This is output for CAR/SAR models only.
17. Condition number for the inverse distance matrix. If the condition number is large, then the model may not have properly converged. This is output for CAR/SAR models only.

Likelihood statistics

18. Log-likelihood estimate, which is a negative number. For a set number of independent variables, the smaller the log-likelihood (i.e., the most negative) the better.
19. Deviance Information Criterion (DIC) adjusts the log-likelihood for the effective degrees of freedom. The smaller the DIC, the better.
20. Akaike Information Criterion (AIC) adjusts the log-likelihood for the degrees of freedom. The smaller the AIC, the better.
21. Bayesian Information Criterion (BIC), sometimes known as the Schwartz Criterion (SC), adjusts the log-likelihood for the degrees of freedom. The smaller the BIC, the better.
22. Deviance compares the log-likelihood of the model to the log-likelihood of a model that fits the data perfectly. A smaller deviance is better.
23. The probability value of the deviance based on a Chi-square with k-1 degrees of freedom.

24. Pearson Chi-square is a test of how closely the predicted model fits the data. A smaller Chi-square is better since it indicates the model fits the data well.

Model error estimates

25. Mean Absolute Deviation (MAD). For a set number of independent variables, a smaller MAD is better.
26. Quartiles for the Mean Absolute Deviation. For any one quartile, smaller is better.
27. Mean Squared Predictive Error (MSPE). For a set number of independent variables, a smaller MSPE is better.
28. Quartiles for the Mean Squared Predictive Error. For any one quartile, smaller is better.

Over-dispersion tests

29. Adjusted deviance. This is a measure of the difference between the observed and predicted values (the residual error) adjusted for degrees of freedom. The smaller the adjusted deviance, the better. A value greater than 1 indicates over-dispersion.
30. Adjusted Pearson Chi-square. This is the Pearson Chi-square adjusted for degrees of freedom. The smaller the Pearson Chi-square, the better. A value greater than 1 indicates over-dispersion.
31. Dispersion multiplier. This is the ratio of the expected variance to the expected mean. For a set number of independent variables, the smaller the dispersion multiplier, the better. In a pure Poisson distribution, the dispersion should be 1.0. In practice, a ratio greater than 10 indicates that there is too much variation that is unaccounted for in the model. Either add more variables or change the functional form of the model.
32. Inverse dispersion multiplier. For a set number of independent variables, a larger inverse dispersion multiplier is better. A ratio close to 1.0 is considered good.

MCMC Individual Coefficients Statistics

For the individual coefficients, the following are output:

33. The mean coefficient. This is the mean parameter value for the *N*-*k* iterations where *k* is the ‘burn in’ samples that are discarded. With the MCMC block sampling method, this is the mean of the mean coefficients for all block samples.
34. The standard deviation of the coefficient. This is an estimate of the standard error of the parameter for the *N*-*k* iterations where *k* is the ‘burn in’ samples that are

discarded. With the MCMC block sampling method, this is the mean of the standard deviations for all block samples.

35. t-value. This is the t-value based on the mean coefficient and the standard deviation. It is defined by Mean/Std.

36. p-value. This is the two-tail probability level associated with the t-test.

37. Adjusted standard deviation (Adj. Std). The block sampling method will produce substantial variation in the mean standard deviation, which is used to estimate the standard error. Consequently, the standard error will be too large. An approximation is made by multiplying the estimated standard deviation by $\sqrt{\frac{\bar{n}}{N}}$ where $\bar{n}$ is the average sample size of the block samples and *N* is the number of records. If no block samples are taken, then this statistic is not calculated.

38. Adjusted t-value. This is the t-value based on the mean coefficient and the adjusted standard deviation. It is defined by Mean/Adj_Std. If no block samples are taken, then this statistic is not calculated.

39. Adjusted p-value. This is the two-tail probability level associated with the adjusted t-value. If no block samples are taken, then this statistic is not calculated.

40. MC error is a Monte Carlo simulation error. It is a comparison of the means of *m* individual chains relative to the mean of the entire chain. By itself, it has little meaning.

41. MC error/Std is the MC error divided by the standard deviation. If this ratio is less than .05, then it is a good indicator that the posterior distribution has converged.

42. G-R stat is the Gelman-Rubin statistic which compares the variance of *m* individual chains relative to the variance of the entire chain. If the G-R statistic is under 1.2, then the posterior distribution is commonly considered to have converged.

43. Spatial autocorrelation term (Phi) for Poisson-Gamma-CAR models only. This is the estimate of the fixed effect spatial autocorrelation effect. It is made up of three components: a global component (Rho); a local component (Tauphi); and a local neighborhood component (Alpha, which is defined by the user).

Expanded Output (MCMC only)

If the expanded output box is selected, additional information on the percentiles from the MCMC sample are displayed. If the block sampling method is used, the percentiles are the means of all block samples. The percentiles are:

44. 2.5th percentile
45. 5th percentile
46. 10th percentile
47. 25th percentile
48. 50th percentile (median)
49. 75TH percentile
50. 90th percentile
51. 95th percentile
52. 97.5th percentile

The percentiles can be used to construct confidence intervals around the mean estimates or to provide a non-parametric estimate of significance as an alternative to the estimated t-value in the standard output. For example, the 2.5th and 97.5th percentiles provide approximate 95 percent confidence intervals around the mean coefficient while the 0.5th and 99.5th percentiles provide approximate 99 percent confidence intervals.

The percentiles will be output for all estimated parameters including the intercept, each individual predictor variable, the spatial effects variable (Phi), the estimated components of the spatial effects (Rho and Tauphi), and the overall error term (Taupsi).

Output Phi Values (CAR/SAR models only)

For CAR or SAR models only, the individual Phi values can be output. This will occur if the sample size is smaller than the block sampling threshold. Check the 'Output Phi value if sample size smaller than block sampling threshold' box. An ID variable must be identified and a DBF output file defined.

Multicollinearity Among the Independent Variables

A major consideration in any regression model is that the independent variables are statistically independent. Non-independence is called *Multicollinearity*. Non-independence means that there is overlap in prediction among two or more independent variables. This can lead to uncertainty in interpreting coefficients as well as an unstable model that may not hold in the future. Generally, it is a good idea to reduce Multicollinearity as much as possible. A tolerance test is given for each coefficient. This is defined as 1 – the R-square of the independent variable predicted by the remaining independent variables in the equation using an Ordinary Least Squares model. It is an indicator of how much the other independent variables in the equation account for the variance of any particular independent variable. Since the method uses the Ordinary Least Squares methods, it is an approximate (pseudo) test for the Poisson regression routines. A message is displayed that indicates probable or possible Multicollinearity. A good idea is to drop one of the multicolinear independent variables and re-run the model. However, each of the coefficients should be inspected carefully before accepting a final model.

Graph of Residual Errors

While the output page is open, clicking on the graph button will display a graph of the residual errors (on the Y axis) against the predicted values (on the X axis). Only residual errors that vary between -200 and +200 are shown to allow most of the errors to be displayed.

Save Output

The predicted values and the residual errors can be output to a DBF file with a TripGenOut<*root name*> with the root name being provided by the user. The output includes all the variables in the input data set plus two new ones: 1) the predicted values of the dependent variable for each observation (with the field name PREDICTED); and 2) the residual error values, representing the difference between the actual /observed values for each observation and the predicted values (with the field name RESIDUAL). The file can be imported into a spreadsheet or graphics program and the errors plotted against the predicted dependent variable.

Save Estimated Coefficients

The individual coefficients can be output to a DBF file with a TripGenCoeff<*root name*> with the root name being provided by the user. This file can be used in the 'Make Prediction' routine of the Trip Generation module.

Diagnostic Tests

The regression module has a set of diagnostic tests for evaluating the characteristics of the data and the most appropriate model to use. There is a diagnostics box on the 'Calibrate model' page.

Diagnostics are provided on:

1. The minimum and maximum values for the dependent and independent variables
2. Skewness in the dependent variable
3. Spatial autocorrelation in the dependent variable

4. Estimated values for the distance decay parameter – alpha, for use in CAR/SAR models
5. Multicolinarity among the independent variables

Minimum and Maximum Values for the Variables

First, the minimum and maximum values of both the dependent and independent variables are listed. A user should look for ineligible values (e.g., -1) as well as variables that have a very high range. The MLE routines are sensitive to variables with very large ranges.

Skewness Tests

Skewness in the dependent variable can distort a linear model by allowing high values to be underestimated while allowing low values to be overestimated and a Poisson-type model is preferred over the linear for highly skewed variables.

The diagnostics utility tests for skewness using two different measures: 1) the "***g"*** statistic, and 2) the ratio of the simple variance to the simple mean. Either significant "g" scores or variance-to-mean ratios greater than about 2:1 should make the user cautious about using a linear model. If either measure indicates skewness, *CrimeStat* prints out a message indicating the dependent variable appears to be skewed and that a Poisson-based model should be used.

Testing for Spatial Autocorrelation in the Dependent Variable

The third type of test in the diagnostics utilities is the Moran's "I" coefficient for spatial autocorrelation. If the "I" is significant, *CrimeStat* outputs a message indicating that there is definite spatial autocorrelation in the dependent variable and that it needs to be accounted for, either by a proxy variable or by estimating a CAR or SAR model.

Estimating the Value of Alpha for CAR or SAR Models

The fourth type of diagnostic test is an estimate of a plausible value for the distance decay function, α, in CAR or SAR models. Three values of alpha are given in different distance units, one associated with a weight of 0.9 (a very steep distance decay), one associated with a weight of 0.75 (a moderate distance decay), and one associated with a weight of 0.5 (a shallow distance decay). Users should run the Moran Correlogram and examine the graph of the drop off in spatial autocorrelation to assess what type of decay function most likely exists. The user should choose an alpha value that best represents the distance decay and should define the distance units for it.

Multicollinearity Test

The fifth type of diagnostic test is for Multicollinearity among the independent predictors. The tolerance test is presented for each independent variable. This is defined as 1-R^2 for the other independent variables in the equation. Each independent variable should have a high tolerance (0.90 or higher). *CrimeStat* prints out an error message if tolerance is not high.

Make Trip Generation Prediction

This routine applies an already-calibrated regression model to a data set. This would be useful for several reasons: 1) if external trips are to be added to the model (which is normally preferred); 2) if the model is applied to another data set; and 3) if variations on the coefficients are being tested with the same data set. The model will need to be calibrated first (see Calibrate trip generation model) and the coefficients saved as a parameters file. The coefficient parameter file is then re-loaded and applied to the data.

Data Input File

The data file is input as either the primary or secondary file. Specify whether the data file is the primary or secondary file.

Type of Model

Specify whether the model is for origins or destinations. This will be printed out on the output header.

Trip Generation Parameters (coefficients) File

This is the saved coefficient parameter file. It is DBF with a TRIPGENCOEFF prefix. Load the file by clicking on the Browse button and finding the file. Once loaded, the variable names of the saved coefficients are displayed in the “Matching parameters” box.

Independent Variables

Select independent variables from the list of variables in the data file. Up to 15 variables can be selected.

Matching Parameters

The selected independent variables need to be matched to the saved variables in the trip generation parameters file **in the same order**. Add the appropriate variables one by one in the order in which they are listed in the matching parameters box. It is essential that the order by the same otherwise the coefficients will be applied to the wrong variables.

If the model had estimated a general spatial effect from a CAR or SAR model, then the general Phi will have been saved with the coefficient files. If the model had estimated specific spatial effects from a CAR or SAR model, then the specific Phi values will have been saved in a separate Phi coefficients file. In the latter case, the user must read in the Phi coefficients file along with the general coefficient file.

Missing Values

Specify any missing value codes for the variables. Blank records will automatically be considered as missing. If any of the selected dependent or independent variables have missing values, those records will be excluded from the analysis.

Add External Trips

External trips are trips that start outside the modeled study area. Because they are crimes that originate outside the study area, they were not included in the zones used for the origin model. Therefore, they have to be independently estimated and added to the origin zone total to make the number of origins equal to the number of destinations. Click on the "Add external trips" button to enable this feature.

Number of external trips

Add the number of external trips to the box. This number will be added as an extra origin zone (the External zone).

Origin ID

Specify the origin ID variable in the data file. The external trips will be added as an extra origin zone, called the "External" zone. Note: all destination ID's should be in the origin zone file and must have the same names. This is necessary for subsequent modeling stages.

Type of Regression Model

Specify the type of regression model to be used. The default is a Poisson regression and the other alternative is a Linear (Ordinary Least Squares) regression.

Save Predicted Values

The output is saved as a 'dbf' file under a different file name with a TripGenMakePred<*root name*> with the root name being provided by the user. The output includes all the variables in the input data set plus the predicted values of the dependent variable for each observation (with the name PREDICTED. In addition, *if* external trips were added, then there is a new record with the name EXTERNAL listed in the Origin ID column. This record lists the added trips in the PREDICTED column and zeros (0) for all other numeric fields.

Output

The tabular output includes summary information about file and lists the predicted values for each input zone.

Balance Origins and Destinations

Since, by definition, a 'trip' has an origin and a destination, the number of predicted origins must equal the number of predicted destinations. Because of slight differences in the data sets of the origin model and the destination model, it is possible that the total number of predicted origins (including any external trips – see Make trip generation prediction) may not equal the total number of predicted destinations. This step, therefore, is essential guarantee that this condition will be true. The routine adjusts either the number of predicted origins or the number of predicted destinations so that the condition holds. The trip distribution routines will not work unless the number of predicted origins equals the number of predicted destinations (within a very small rounding-off error).

Predicted origin file

Specify the name of the predicted origin file by clicking on the Browse button and locating the file.

Origin variable

Specify the name of the variable for the predicted origins (e.g., PREDICTED).

Predicted destination file

Specify the name of the predicted destination file by clicking on the Browse button and locating the file.

Destination variable

Specify the name of the variable for the predicted origins (e.g., PREDICTED).

Balancing method

Specify whether origins or destinations are to be held constant. The default is 'Hold destinations constant'.

Save predicted origin/destination file

The output is saved as a 'dbf' file under a different file name. The output includes all the variables in the input data set plus the adjusted values of the predicted values of the dependent variable for each observation. If destinations are held constant, the adjusted variable name for the predicted trips is ADJORIGIN. If origins are held constant, the adjusted variable name for the predicted trips is ADJDEST.

Output

The tabular output includes file summary information plus information about the number of origins and destinations before and after balancing. In addition, the predicted values of the dependent variable are displayed.

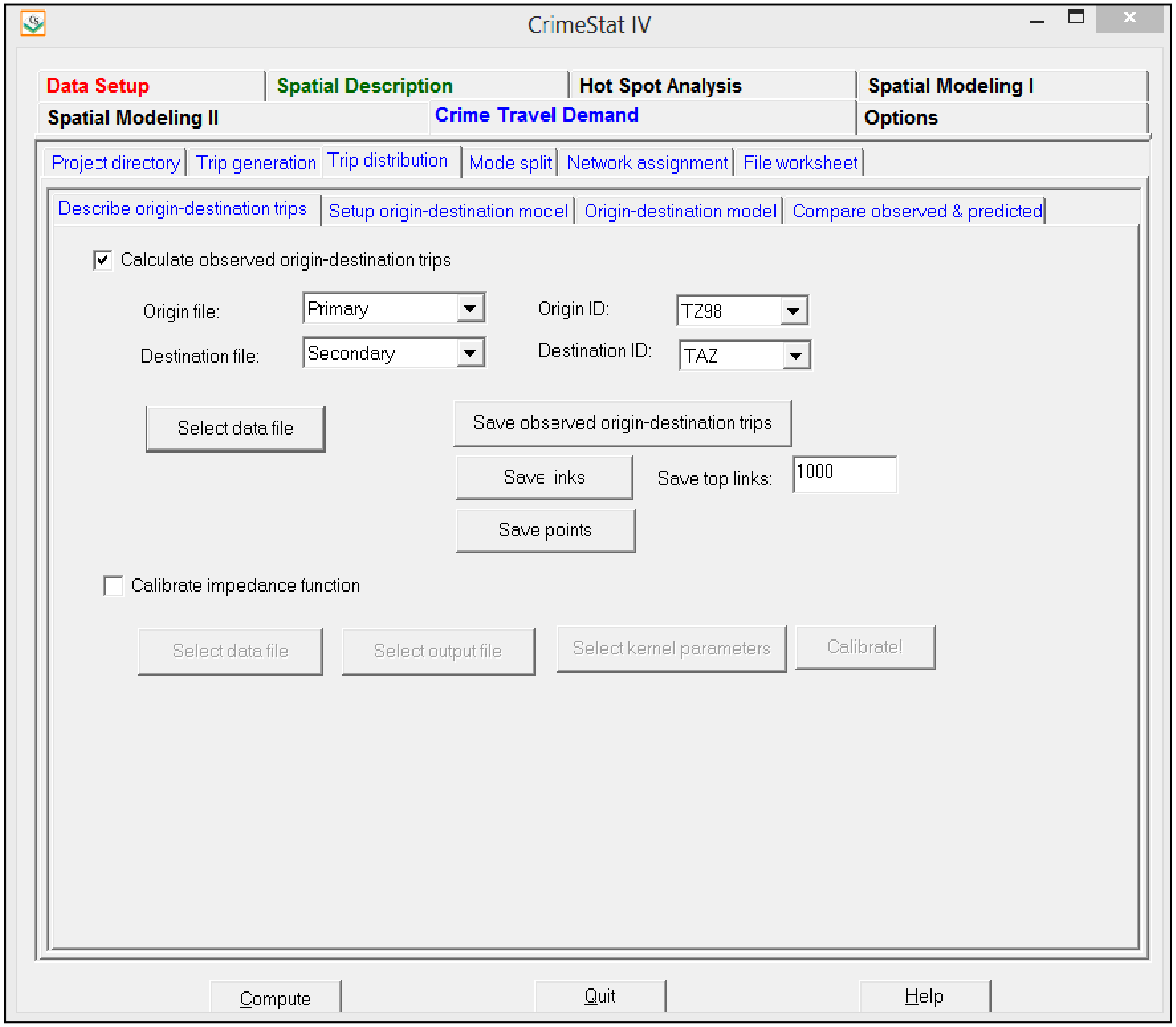
Trip Distribution

Trip distribution involves the estimation of the number of trips that travel from each origin zone (including the 'external' zone) to each destination zone. The estimation is based on a gravity-type model. The determining variables are the number of predicted origins, the number of predicted destinations, the impedance (or cost) of travel between the origin zone, coefficients for the origins and destinations, and exponents of the origins and destinations.

The user inputs the number of predicted origins and predicted destinations and specifies an impedance model (which can be mathematical or calibrated from an existing data set). In addition,

Figure 2.24:

Trip Distribution Modeling



☑ Calculate observed origin-destination trips

Origin file: Primary

Destination file: Secondary

Origin ID: TZ98

Destination ID: TAZ

Save top links: 1000

□ Calibrate impedance function

the user specifies exponents for the origin and destination values. The model iteratively estimates the coefficients. In addition, the routine can calculate the actual (observed) trip distribution with an existing data set that lists individual origin and destination locations. Finally, a comparison between the observed distribution and that predicted by the model can be made.

Describe Origin-Destination trips

An empirical description of the actual trip distribution matrix can be made if there is a data set that includes individual origin and destination locations. The user defines the origin location and the destination location for each record and a set of zones from which to compare the individual origins and destinations. The routine matches up each origin location with the nearest zone, each destination location with the nearest zone, and calculates the number of trips from each origin zone to each destination zone. This is an *observed* distribution of trips by zone.

Calculate Observed Origin-Destination trips

Check if an empirical origin-destination trip distribution is to be calculated.

Origin file

The origin file is a list of origin zones with a single point representing the zone (e.g., the centroid). There can be more origin zones than destination zones, but **all** destination zones must be included among the origin zone list. The origin file must be input as either the primary or secondary file. Specify whether the data file is the primary or secondary file.

Origin ID

Specify the origin ID variable in the data file (e.g., CensusTract, Block, TAZ). **Note**: all destination ID’s should be in the origin zone file and must have the same names.

Destination file

The destination file is a list of destination zones with a single point representing the zone (e.g., the centroid). It must be input as either the primary or secondary file. Specify whether the data file is the primary or secondary file.

Destination ID

Specify the destination ID variable in the data file (e.g., CensusTract, Block, TAZ). Note: the ID's used for the destination file zones must be the same as in the origin file.

Select data file

The data set must have individual origin and destination locations. Each record must have the X/Y coordinates of an origin location and the X/Y coordinates of a destination location. For example, an arrest file might list individual incidents with each incident having a crime location (the destination) and a residence or arrest location (the origin). Select the file that has the X and Y coordinates for the origin and destination locations. *CrimeStat* reads dbase 'dbf', ArcGIS 'shp' and ASCII text files. Select the tab and specify 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 for both the origin (residence) and destination (crime) locations.

Columns

Select the variables for the X and Y coordinates respectively for *both* the origin and destination locations (e.g., Lon, Lat, HomeX, HomeY, IncidentX, IncidentY.) Both locations must be defined for the routine to work.

Missing values

Identify whether there are missing values for these four fields (X and Y coordinates for both origin and destination locations). By default, *CrimeStat* will ignore records with blank values in any eligible field or records with non-numeric values (e.g., alphanumeric characters, , *). Blanks will always be excluded unless the user selects **<*none*>**. There are 8 possible options:

1. <*blank*> fields are automatically excluded. This is the default
2. <*none*> indicates that no records will be excluded. If there is a blank field, *CrimeStat* will treat it as a 0
3. *0* is excluded

4. *–1* is excluded
5. *0 and –1* indicates that both 0 and -1 will be excluded
6. *0, -1 and 9999* indicates that all three values (0, -1, 9999) will be excluded
7. *Any* other numerical value can be treated as a missing value by typing it (e.g., 99)
8. *Multiple* numerical values can be treated as missing values by typing them, separating each by commas (e.g., 0, -1, 99, 9999, -99).

Type of coordinate system and data units

The coordinate system and data units are listed for information. 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.)

Table output

The entire origin-destination matrix is output as a table to the screen including summary file information and:

1. The origin zone (ORIGIN)
2. The destination zone (DEST)
3. The number of observed trips (FREQ)

Save observed origin-destination trips

If specified, the full origin-destination output is saved as a 'dbf' file named by the user.

File output

The file output includes:

1. The origin zone (ORIGIN)
2. The destination zone (DEST)
3. The X coordinate for the origin zone (ORIGINX)
4. The Y coordinate for the origin zone (ORIGINY)
5. The X coordinate for the destination zone (DESTX)
6. The Y coordinate for the destination zone (DESTY)
7. The number of trips (FREQ)

Note: each record is a unique origin-destination combination and there are M x N records where M is the number of origin zones (including the external zone) and N is the number of destination zones.

Save links

The top observed origin-destination trip links can be saved as separate **line** objects for use in a GIS. Specify the output file format (*ArcGIS* 'shp', *MapInfo* 'mif' or ASCII) and the file name. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Save top links

Because the output file is very large (number of origin zones x number of destination zones), the user can select a sub-set of zone combinations with the most observed trips. Indicating the top K links will narrow the number down to the most important ones. The default is the top 100 origin-destination combinations. Each output object is a line from the origin zone to the destination zone with an ODT prefix. The prefix is placed before the output file name. The line graphical output for each object includes:

1. An ID number from 1 to K, where K is the number of links output (ID)
2. The feature prefix (ODT)
3. The origin zone (ORIGIN)
4. The destination zone (DEST)
5. The X coordinate for the origin zone (ORIGINX)
6. The Y coordinate for the origin zone (ORIGINY)
7. The X coordinate for the destination zone (DESTX)
8. The Y coordinate for the destination zone (DESTY)
9. The number of observed trips for that combination (FREQ)
10. The distance between the origin zone and the destination zone.

Save points

Intra-zonal trips (trips in which the origin and destination are the same zone) can be output as separate **point** objects as an *ArcGIS* 'shp', *MapInfo* 'mif' or various ASCII formats. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Again, the top *K* points are output (default=100). Each output object is a point representing an intra-zonal trip with an ODTPOINTS prefix. The prefix is placed before the output file name. The point graphical output for each object includes:

1. An ID number from 1 to K, where K is the number of links output (ID)
2. The feature prefix (POINTSODT)
3. The origin zone (ORIGIN)
4. The destination zone (DEST)
5. The X coordinate for the origin zone (ORIGINX)
6. The Y coordinate for the origin zone (ORIGINY)
7. The X coordinate for the destination zone (DESTX)
8. The Y coordinate for the destination zone (DESTY)
9. The number of observed trips for that combination (FREQ)

Calibrate Impedance Function

This function allows the calibration of an approximate travel impedance function based on actual trip distributions. It is used to describe the travel distance of an actual sample (the calibration sample). A file is input which has a set of incidents (records) that includes both the X and Y coordinates for the location of the offender's residence (origin) and the X and Y coordinates for the location of the incident that the offender committed (destination.) The routine estimates a travel distance function using a one-dimensional kernel density method. For each record, the distance between the origin location and the destination location is calculated and is represented on a distance scale. The maximum distance is calculated and divided into a number of intervals; the default is 100 equal sized intervals, but the user can modify this. For each distance (point) calculated, a one-dimensional kernel is overlaid. For each distance interval, the values of all kernels are summed to produce a smooth function of travel impedance. The results are saved to a file that can be used origin-destination model. Note, however, that this is an empirical distribution and represents the combination of origins, destinations, and costs. It is not necessarily a good description of the impedance (cost) function by itself. Many of the mathematical functions produce a better fit than the empirical impedance function.

Select data file for calibration

Select the file that has the X and Y coordinates for the origin and destination locations. *CrimeStat* reads dbase 'dbf', ArcGIS ‘shp’ and ASCII files. Select the tab and select 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 for both the origin (residence) and destination (crime) locations

Columns

Select the variables for the X and Y coordinates respectively for *both* the origin and destination locations (e.g., Lon, Lat, HomeX, HomeY, IncidentX, IncidentY.) Both locations must be defined for the routine to work.

Missing values

Identify whether there are any missing values for these four fields (X and Y coordinates for both origin and destination locations). By default, *CrimeStat* will ignore records with blank values in any of the eligible fields or records with non-numeric values (e.g., alphanumeric characters, , *). Blanks will always be excluded unless the user selects **<*none*>**. There are 8 possible options:

1. <*blank*> fields are automatically excluded. This is the default
2. <*none*> indicates that no records will be excluded. If there is a blank field, *CrimeStat* will treat it as a 0
3. *0* is excluded
4. *–1* is excluded
5. *0 and –1* indicates that both 0 and -1 will be excluded
6. *0, -1 and 9999* indicates that all three values (0, -1, 9999) will be excluded
7. *Any* other numerical value can be treated as a missing value by typing it (e.g., 99)
8. *Multiple* numerical values can be treated as missing values by typing them, separating each by commas (e.g., 0, -1, 99, 9999, -99)

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.) Directional coordinates are not allowed for this routine.

Select kernel **parameters**

There are five parameters that must be defined.

Method of **interpolation**

There are five types of kernel distributions that can be used to estimate point density:

1. The **normal** kernel overlays a three-dimensional normal distribution over each point that then extends over the area defined by the reference file. This is the default kernel function.

2. The **uniform** kernel overlays a uniform function over each point that only extends for a limited distance.

3. The **quartic** kernel overlays a quartic function over each point that only extends for a limited distance.

4. The **triangular** kernel overlays a three-dimensional triangle over each point that only extends for a limited distance.
5. The **negative exponential** kernel overlays a three dimensional negative exponential function over each point that only extends for a limited distance

The methods produce similar results though 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 uniform, quartic, triangular and negative exponential kernels, bandwidth is the radius of a circle defined by the surface. For all types, larger bandwidth will produce smoother density estimates and both adaptive and fixed bandwidth intervals can be selected.

Fixed bandwidth

A fixed bandwidth distance is a fixed interval for each point. The user must define the interval, the interval size, and the distance units by which it is calculated (miles, nautical miles, feet, kilometers, or meters). The default bandwidth setting is fixed with intervals of 0.25 miles each. The interval size can be changed.

Adaptive bandwidth

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

Specify interpolation bins

The interpolation bins are defined in one of two ways:

1. By the number of bins. The maximum distance calculated is divided by the number of specified bins. The default is 100 bins. The user can change the number of bins.

2. By the distance between bins. The user can specify a bin width in miles, nautical miles, feet, kilometers, and meters.

Output (areal) units

Specify the areal density units as points per mile, nautical mile, foot, kilometer, or meter. The default is points per mile.

Calculate densities or probabilities

The density estimate for each cell can be calculated in one of three ways:

1. **Absolute densities.** This is the number of points per grid cell and is scaled so that the sum of all grid cells equals the sample size.
2. **Relative densities**. For each grid cell, this is the absolute density divided by the grid cell area and is expressed in the areal output units (e.g., points per square mile)

3. **Probabilities**. This is the proportion of all incidents that occur in the grid cell. The sum of all grid cells equals a probability of 1. Unlike the Jtc calibration routine, this is the default. In most cases, a user would want a proportional (probability) distribution as the relative differences in impedance for different costs are what is of interest.

Select whether absolute densities, relative densities, or probabilities are to be output for each cell. The default is probabilities.

Select output file

The output *must* be saved to a file. *CrimeStat* can save the calibration output to either a dbase 'dbf' or ASCII text 'txt' file.

Calibrate!

Click on 'Calibrate!' to run the routine. The output is saved to the specified file upon clicking on 'Close'.

Graphing the travel impedance function

Click on 'View graph' to see the travel impedance function. The screen view can be printed by clicking on 'Print'. For a better quality graph, however, the output should be imported into a graphics package.

Setup Origin-Destination Model

The page is for the setup of the origin-destination model. All the relevant files, models and exponents are input on the page.

Predicted origin file

The predicted origin file is a file that lists the origin zones with a single point representing the zone (e.g., the centroid) and also includes the predicted number of crimes by origin zone. The file must be input as either the primary or secondary file. Specify whether the data file is the primary or secondary file.

Origin variable

Specify the name of the variable for the predicted origins (e.g., PREDICTED, ADJORIGINS).

Origin ID

Specify the origin ID variable in the data file (e.g., CensusTract, Block, TAZ). Note: all destination IDs should be in the origin zone file and must have the same names.

Predicted destination file

The predicted destination file is a list of destination zones with a single point representing the zone (e.g., the centroid) and also includes the predicted number of crimes by destination zone. It must be input as either the primary or secondary file. Specify whether the data file is the primary or secondary file.

Destination variable

Specify the name of the variable for the predicted destination (e.g., PREDICTED, ADJDEST).

Destination ID

Specify the destination ID variable in the data file (e.g., CensusTract, Block, TAZ). Note: the ID's used for the destination file zones must be the same as in the origin file.

Exponents

The exponents are power terms for the predicted origins and destinations and indicate the relative strength of those variables. For example, compared to an exponent of 1.0 (the default), an exponent greater than 1.0 will strengthen that variable (origins or destinations) while an exponent less than 1.0 will weaken that variable. They can be considered 'fine tuning' adjustments.

Origins

Specify the exponent for the predicted origins. The default is 1.0.

Destinations

Specify the exponent for the predicted origins. The default is 1.0.

Impedance function

The trip distribution routine can use two different travel distance functions: 1) An already-calibrated distance function; and 2) A mathematical formula. The default is a mathematical formula.

Use an already-calibrated distance function

If a travel distance function has already been calibrated (see 'Calibrate impedance function' under trip distribution), the file can be directly input into the routine. The user selects the name of the already-calibrated travel distance function. *CrimeStat* reads dbase 'dbf'', ArcGIS 'shp' and ASCII files.

Use a mathematical formula

A mathematical formula can be used instead of a calibrated distance function. To do this, it is necessary to specify the type of distribution. There are five mathematical models that can be selected:

1. Negative exponential
2. Normal
3. Lognormal
4. Linear
5. Truncated negative exponential

The lognormal is the default. For each mathematical model, two or three different parameters must be defined:

1. For the negative exponential, the coefficient and exponent
2. For the normal distribution, the mean distance, standard deviation and coefficient
3. For lognormal distribution, the mean distance, standard deviation and coefficient
4. For the linear distribution, an intercept and slope
5. For the truncated negative exponential, a peak distance, peak likelihood, intercept, and exponent.

Measurement unit

The routine can calculate impedance in four ways, by:

1. Distance (miles, nautical miles, feet, kilometers, or meters)
2. Travel time (minutes, hours)
3. Speed (miles per hour, kilometers per hour)
4. General travel costs (unspecified units).

These must be setup under Network distance on the Measurement Parameters page. Specify the appropriate units. In the Network Parameters dialogue, specify the measurement units. The default is distance in miles.

Assumed impedance for external zones

For trips originating outside the study area (external trips), specify the amount and the units that will be assumed for these trips. The default is 25 miles.

Assumed impedance for intra-zonal trips

For trips originating and ending in the same zone (intra-zonal trips), specify the amount and the units that will be assumed for these trips. The default is 0.25 miles.

Minimum number of trips per cell

The parameter allows a minimum number of predicted trips for each origin-destination combination (cell). It will return a zero (0) if the predicted number is less than the minimum. This can be adjusted to avoid many cells with very small numbers of predicted trips. Care must be taken, though, as this can alter the overall distribution. The default minimum is 0.05 trips per cell.

Model constraints

In calibrating a model, the routine must constrain either the origins or the destinations (single constraint) or constrain both the origins and the destinations (double constraint). In the latter case, it is an iterative solution. The default is to constrain destinations as it is assumed that the destinations totals (the number of crimes occurring in each zone) are probably more correct than the number of crimes originating in each zone. . Specify the type of constraint for the model.

Constrain origins

If constrain origins is selected, the total number of trips from each origin zone will be held constant.

Constrain destinations

If constrain destinations is selected, the total number of trips from each destination zone will be held constant.

Constrain both origins and destinations

If constrain both origins and destinations is selected, the routine iteratively works out a balance between the number of origins and the number of destinations.

Origin-Destination Model

The trip distribution (origin-destination) model is implemented in two steps. First, the coefficients are calculated according to the exponents and impedance functions specified on the setup page. Second, the coefficients and exponents are applied to the predicted origins and destinations resulting in a predicted trip distribution. Because these two steps are iterative, they cannot be run simultaneously.

Calibrate origin-destination model

Check the 'Calibrate origin-destination model' box to run the calibration model.

Save modeled coefficients (parameters)

The modeled coefficients are saved as a 'dbf' file. Specify a file name.

Apply predicted origin-destination model

Check the 'Apply predicted origin-destination model' box to run the trip distribution prediction.

Modeled coefficients file

Load the modeled coefficients file saved in the 'Calibrate origin-destination model' stage.

Assumed coordinates for external zone

In order to model trips from the 'external' zone (trips from outside the study area), specify coordinates for this zone. These coordinates will be used in drawing lines from the predicted origins to the predicted destinations. There are four choices:

1. Mean center (the mean X and mean Y of all origin file points are taken). This is the default.
2. Lower-left corner (the minimum X and minimum Y values of all origin file points are taken).
3. Upper-right corner (the maximum X and maximum Y values of all origin file points are taken).
4. Use coordinates (user-defined coordinates). Indicate the X and Y coordinates that are to be used.

Table output

The table output includes summary file information and (with default names):

1. The origin zone (ORIGIN)
2. The destination zone (DEST)
3. The number of predicted trips (PREDTRIPS)

Save predicted origin-destination trips

Define the output file. The output is saved as a 'dbf' file with the file name specified by the user.

File output

The file output includes (with default names):

1. The origin zone (ORIGIN)
2. The destination zone (DEST)
3. The X coordinate for the origin zone (ORIGINX)
4. The Y coordinate for the origin zone (ORIGINY)
5. The X coordinate for the destination zone (DESTX)
6. The Y coordinate for the destination zone (DESTY)
7. The number of predicted trips (PREDTRIPS)

Note: each record is a unique origin-destination combination and there are M x N records where M is the number of origin zones (including the external zone) and N is the number of destination zones.

Save links

The top predicted origin-destination trip links can be saved as separate **line** objects for use in a GIS. Specify the output file format (*ArcGIS* 'shp', *MapInfo* 'mif' or various ASCII formats) and the file name. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Save top links

Because the output file is very large (number of origin zones x number of destination zones), the user can select a sub-set of zone combinations with the most predicted trips. Indicating the top K links will narrow the number down to the most important ones. The default is the top 100 origin-destination combinations. Each output object is a line from the origin zone to the destination zone with an ODT prefix. The prefix is placed before the output file name. The graphical output includes (with default names):

1. An ID number from 1 to K, where K is the number of links output (ID)
2. The feature prefix (ODT)
3. The origin zone (ORIGIN)
4. The destination zone (DEST)
5. The X coordinate for the origin zone (ORIGINX)
6. The Y coordinate for the origin zone (ORIGINY)
7. The X coordinate for the destination zone (DESTX)
8. The Y coordinate for the destination zone (DESTY)
9. The number of predicted trips for that combination (PREDTRIPS)
10. The distance between the origin zone and the destination zone.

Save points

Intra-zonal trips (trips in which the origin and destination are the same zone) can be output as separate **point** objects as an *ArcGIS* 'shp', *MapInfo* 'mif' or various ASCII formats. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Again, the top K points are output (default=100). Each output object is a point representing an intra-zonal trip with an ODTPOINTS prefix. The prefix is placed before the output file name. The graphical output for each includes (with default names):

1. An ID number from 1 to K, where K is the number of links output (ID)
2. The feature prefix (POINTSODT)
3. The origin zone (ORIGIN)
4. The destination zone (DEST)
5. The X coordinate for the origin zone (ORIGINX)
6. The Y coordinate for the origin zone (ORIGINY)
7. The X coordinate for the destination zone (DESTX)
8. The Y coordinate for the destination zone (DESTY)
9. The number of predicted trips for that combination (PREDTRIPS)

Compare Observed & Predicted Origin-Destination Trip Lengths

The predicted trip distribution model can be compared with the observed (actual) trip distribution. Since there are many cells for this comparison (M origins x N destinations), a comparison is usually conducted for the trip length distributions. Each origin-destination link (whether the observed distribution or that predicted by the model) is converted into a trip length. The maximum distance between an origin and a destination is then divided into K bins (intervals), where K can be defined by the user; the default is 25. The two distributions are compared with two statistics: 1) the coincidence ratio (essentially a positive correlation index that varies between 0 and 1 with 0 representing little coincidence and 1 representing perfect coincidence) and 2) the Komolgorov-Smirnov two-sample test (a test of the difference between the cumulative proportions of the observed and predicted distributions). There is also a graph that compares the two distributions.

Observed trip file

Select the observed trip distribution file by clicking on the Browse button.

Observed number of origin-destination trips

Specify the variable for the observed number of trips. The default name is FREQ.

Orig_ID

Specify the ID name for the origin zone. The default name is ORIGIN. Note: the origin ID's should be the same as in the predicted file in order to compare the top links.

Orig_X

Specify the name for the X coordinate of the origin zone. The default name is ORIGINX.

Orig_Y

Specify the name for the Y coordinate of the origin zone. The default name is ORIGINY.

Dest_ID

Specify the ID name for the destination zone. The default name is DEST. Note: the destination ID's should be the same as in the predicted file in order to compare the top links.

Dest_X

Specify the name for the X coordinate of the destination zone. The default name is DESTX.

Dest_Y

Specify the name for the Y coordinate of the destination zone. The default name is DESTY.

Predicted trip file

Select the predicted trip distribution file by clicking on the Browse button and finding the file.

Predicted number of origin-destination trips

Specify the variable for the predicted number of trips. The default name is PREDTRIPS

Orig_ID

Specify the ID name for the origin zone. The default name is ORIGIN. Note: the origin ID's should be the same as in the observed file in order to compare the top links.

Orig_X

Specify the name for the X coordinate of the origin zone. The default name is ORIGINX.

Orig_Y

Specify the name for the Y coordinate of the origin zone. The default name is ORIGINY.

Dest_ID

Specify the ID name for the destination zone. The default name is DEST. Note: the destination ID's should be the same as in the observed file in order to compare the top links.

Dest_X

Specify the name for the X coordinate of the destination zone. The default name is DESTX.

Dest_Y

Specify the name for the Y coordinate of the destination zone. The default name is DESTY.

Select bins

Specify how the bins (intervals) will be defined. There are two choices. One is to select a fixed number of bins. The other is to select a constant interval.

Fixed number

This sets a fixed number of bins. An interval is defined by the maximum distance between zone divided by the number of bins. The default number of bins is 25. Specify the number of bins.

Constant interval

This defines an interval of a specific size. If selected, the units must also be chosen. The default is 0.25 miles. Other distance units are nautical miles, feet, kilometers, and meters. Specify the interval size.

Save comparison

The output is saved as a 'dbf' file with the file name specified by the user.

Table output

The table output includes summary information and:

1. The number of trips in the observed origin-destination file
2. The number of trips in the predicted origin-destination file
3. The number of intra-zonal trips in the observed origin-destination file
4. The number of intra-zonal trips in the predicted origin-destination file
5. The number of inter-zonal trips in the observed origin-destination file
6. The number of inter-zonal trips in the predicted origin-destination file
7. The average observed trip length
8. The average predicted trip length
9. The median observed trip length
10. The median predicted trip length
11. The Coincidence Ratio (an indicator of congruence varying from 0 to 1)
12. The D value for the Komolgorov-Smirnov two-sample test
13. The critical D value for the Komolgorov-Smirnov two-sample test
14. The p-value associated with the D value of Komolgorov-Smirnov two-sample test relative to the critical D value.

and for each bin:

15. The bin number
16. The bin distance
17. The observed proportion
18. The predicted proportion

File output

The saved file includes (with default names):

1. The bin number (BIN)
2. The bin distance (BINDIST)
3. The observed proportion (OBSERVPROP)
4. The predicted proportion (PREDPROP)

Graph of observed and predicted trip lengths

While the output page is open, clicking on the graph button will display a graph of the observed and predicted trip length proportions on the Y-axis by the trip length distance on the X-axis.

Compare Top Links

As an alternative to a comparison of trip lengths for the observed and predicted distributions, the top links can be compared with a pseudo-Chi square test. Since the top links have the most trips, the Chi square distribution can be used for comparison. However, because the rest of the distribution is not being used, significance tests are invalid.

The statistic compares the number of trips for the top links in the observed distribution with the number of trips for the same links in the predicted model. The routine calulates a Chi square value.

The statistic is useful for comparing different models. The *lower* the Chi square value, the better the fit between the predicted model and the observed for the top links. The aim is to find the model that gives the lowest possible Chi square value.

Note: in order to use this routine, the origin and destination ID's ***must*** be the same for both the observed and predicted trip files.

Click the box and specify the number of links to be compared. The default value is 100. The output includes:

1. The number of links that are compared

and for each trip pair in order of the number of trips:

2. The zone ID of the origin zone (FromZone)
3. The zone ID of the destination zone (ToZone)
4. The observed (actual) number of trips
5. The predicted number of trips.

At the bottom of the page is a Chi-square test of the difference between the observed and predicted number of trips for the top links. Since not all trips have been included in this distribution, no significance test is conducted. The aim should be to find the model with the lowest Chi-square value.

Optimizing the Fit Between the Observed and Predicted Links

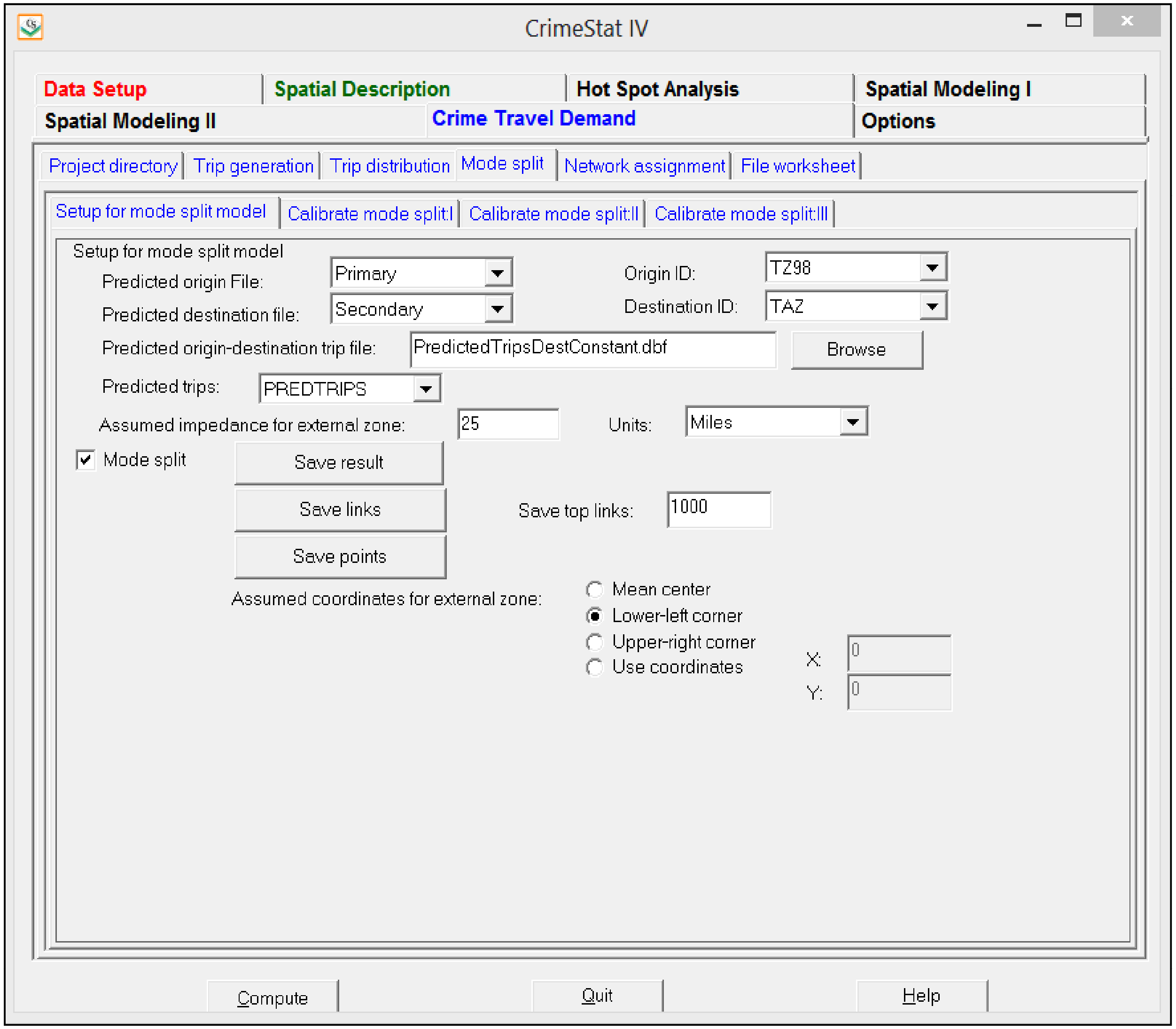
Ideally, the best model would fulfill three comparison tests. First, the number of intra-zonal tests (and, by implication, the number of inter-zonal trips) in the predicted trip distribution would be identical to the number of intra-zonal trips in the observed distribution. Second, the overall model would have a high coincidence ratio and a non-significant Komolgorov-Smirnov test for the trip length comparison. Third, the Chi square value for the top links would be the lowest possible. In practice, an optimal model may have to balance these three criteria, producing a good match in the number of intra-zonal trips, a reasonably low Chi square value for the top links, and a reasonably high coincidence ratio for the trip length comparison. There may not be a single, optimal model.

Mode Split

Mode split involves separating the predicted trips by link (i.e., the trips from any one origin zone, A, to any one destination zone, B) into distinct travel modes (e.g., walk, bicycle, drive, bus, train). The basis of the separation is an aggregate relative impedance function. This is, essentially, the ‘cost’ of traveling by any one mode relative to all modes, whether cost is defined in terms of distance, travel time, or generalized costs. The model can be determined by either an empirically-derived impedance function or a mathematical function. The empirically-derived impedance function would come from a calibration data set whereas the mathematical function is selected on the basis of either previous experience or other studies. The separate impedance functions can be constrained to a network in order to prevent trips from being allocated that are nearly impossible (e.g., train trips where there are no train lines and bus trips where there are no bus routes).

Figure 2.25:

Mode Split Modeling



Setup for mode split model

Predicted origin File: Primary

Predicted destination file: Secondary

Origin ID: TZ98

Destination ID: TAZ

Predicted origin-destination trip file: PredictedTripsDestConstant.dbf

Predicted trips: PREDTRIPS

Assumed impedance for external zone: 25

Units: Miles

☑ Mode split

Save top links: 1000

Assumed coordinates for external zone:

○ Mean center
● Lower-left corner
○ Upper-right corner
○ Use coordinates

X: 0

Y: 0

The steps of the routine are as follows. First, the user inputs a file of predicted trips (i.e., the number of predicted trips from every origin zone to every destination zone). Second, the user defines which travel modes are to be modeled. Up to five separate modes are allowed.

Third, the user sets up an impedance model for **each** travel mode. Any of the impedance models can be constrained to a particular network (e.g., bus mode constrained to a bus network; train mode constrained to a train network). This would normally be desired even for modes where travel in any direction is possible (e.g., walk, bicycle, drive modes). Fourth, and finally, after all impedance models have been defined, the routine is run and splits the predicted trips into the defined modes on the basis of the relative impedance of each mode to all impedances.

Setup for Mode Split Model

This page defines the predicted trip file and the output file. It also allows a definition of where external trips are assumed to come from.

Predicted origin file

The predicted origin file is a file that lists the origin zones with a single point representing the zone (e.g., the centroid) and also includes the predicted number of crimes by origin zone. The file must be input as either the primary or secondary file. Specify whether the data file is the primary or secondary file.

Origin variable

Specify the name of the variable for the predicted origins (e.g., PREDICTED, ADJORIGINS).

Origin ID

Specify the origin ID variable in the data file (e.g., CensusTract, Block, TAZ). Note: all destination ID's should be in the origin zone file and must have the same names.

Predicted destination file

The predicted destination file is a list of destination zones with a single point representing the zone (e.g., the centroid) and also includes the predicted number of crimes by destination zone. It must be input as either the primary or secondary file. Specify whether the data file is the primary or secondary file.

Destination variable

Specify the name of the variable for the predicted destination (e.g., PREDICTED, ADJDEST).

Destination ID

Specify the destination ID variable in the data file (e.g., CensusTract, Block, TAZ). Note: the ID's used for the destination file zones must be the same as in the origin file.

Predicted origin-destination trip file

The predicted origin-destination trip file lists the predicted number of trips from every origin zone to every destination zone. On the mode split setup page, select the predicted trip file (i.e., the predicted origin-destination trip file by clicking on the 'Browse' button.

Predicted trips

Specify the variable for the predicted number of trips. The default name is PREDTRIPS

Assumed impedance for external zone

In order to model trips from the 'external zone' (trips from outside the study area), specify an impedance to be assumed. The default is 25 miles.

Assumed coordinates for external zone

In order to model trips from the 'external' zone (trips from outside the study area), specify coordinates for this zone. These coordinates will be used in drawing lines from the predicted origins to the predicted destinations. There are four choices:

1. Mean center (the mean X and mean Y of all origin file points are taken). This is the default.
2. Lower-left corner (the minimum X and minimum Y values of all origin file points are taken).
3. Upper-right corner (the maximum X and maximum Y values of all origin file points are taken).
4. Use coordinates (user-defined coordinates). Indicate the X and Y coordinates that are to be used.

Run Mode Split

Check the "Mode split" box to enable the routine. It will run when the "Compute" button is clicked.

Mode Split Output

There are three types of output for the mode split routine.

1. The zone-to-zone trip file for **each** mode separately can be output as a dbf file.
2. The most frequent inter-zonal (i.e., trips between different zones) trips for **each** mode separately can be output as polylines.
3. The most frequent intra-zonal (i.e., trips within the same zone) trips for **each** mode separately can be output as points.

Output file name (save result)

Define the output file name by clicking on 'Save result'. The output will be saved as a 'dbf' file with the file name specified by the user. For **each** mode, the prefix 'TMode' will be prefaced before the file. For example, if the name provided by the user is "robberies.dbf" and if there are three travel modes modeled, then there will be three output files (TMode1robberies.dbf; TMode2robberies.dbf; TMode3robberies.dbf).

File output

The file output includes:

1. The origin zone (ORIGIN)
2. The destination zone (DEST)
3. The X coordinate for the origin zone (ORIGINX)
4. The Y coordinate for the origin zone (ORIGINY)
5. The X coordinate for the destination zone (DESTX)
6. The Y coordinate for the destination zone (DESTY)
7. The number of predicted trips (PREDTRIPS)

Note: each record is a unique origin-destination combination and there are M x N records where M is the number of origin zones (including the external zone) and N is the number of destination zones.

Save links

The top predicted origin-destination trip links can be saved as separate **line** objects for use in a GIS. Specify the output file format (*ArcGIS* 'shp', *MapInfo* 'mif' or various ASCII formats) and the file name. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

For **each** mode, the prefix 'TripMode' will be prefaced before the file. For example, if the name provided by the user is "robberies" and if there are three travel modes modeled, then there will be three graphical output files (TripMode1robberies.shp/mif; TripMode2robberies.shp/mif; TripMode3robberies.shp/mif).

Save top links

Because the output file is very large (number of origin zones x number of destination zones), the user can select a sub-set of zone combinations with the most predicted trips. Indicating the top K links will narrow the number down to the most important ones. The default is the top 100 origin-destination combinations. Each output object is a line from the origin zone to the destination zone with a TripMode prefix where '' is the mode number. The prefix is placed before the output file name. The graphical output includes:

1. An ID number from 1 to K, where K is the number of links output (ID)
2. The feature prefix (ODT)
3. The origin zone (ORIGIN)
4. The destination zone (DEST)
5. The X coordinate for the origin zone (ORIGINX)
6. The Y coordinate for the origin zone (ORIGINY)
7. The X coordinate for the destination zone (DESTX)
8. The Y coordinate for the destination zone (DESTY)
9. The number of predicted trips for that combination (PREDTRIPS)
10. The distance between the origin zone and the destination zone.

Save points

Intra-zonal trips (trips in which the origin and destination are the same zone) can be output as separate **point** objects as an *ArcGIS* 'shp', *MapInfo* 'mif' or various ASCII formats. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Again, the top K points are output (default=100). Each output object is a point representing an intra-zonal trip with a TripModePoints prefix where ‘’ is the mode number. The prefix is placed before the output file name. For example, if the name provided by the user is “robberies” and if there are three travel modes modeled, then there will be three graphical output files (TripModePoints1robberies.shp/mif; TripModePoints2robberies.shp/mif; TripModePoints3robberies.shp/mif).

The graphical output for each includes:

1. An ID number from 1 to K, where K is the number of links output (ID)
2. The feature prefix (POINTSODT)
3. The origin zone (ORIGIN)
4. The destination zone (DEST)
5. The X coordinate for the origin zone (ORIGINX)
6. The Y coordinate for the origin zone (ORIGINY)
7. The X coordinate for the destination zone (DESTX)
8. The Y coordinate for the destination zone (DESTY)
9. The number of predicted trips for that combination (PREDTRIPS)

Calibrate Mode Split: I-III

For each mode (up to five), the impedance parameters have to be set. There are three pages for this:

1. “Calibrate mode split: I” covers modes 1 and 2.
2. “Calibrate mode split: II’ covers modes 3 and 4.
3. “Calibrate mode split: III” covers mode 5.

For each mode, the user should indicate whether the mode is to be used, the name to be used for the mode, whether a default impedance will be calculated directly or if it should be constrained to a network, and the specific impedance model used. If any mode is not used, then it will not be part of the calculations. Use only those modes that are relevant, but, also, be sure not to leave out any important ones.

The following instructions apply to each of the five modes.

Mode

Check the box if the mode is to be used.

Label

Put in a label for the mode. Default names are provided (walk, bicycle, drive, bus, train), but the user is not required to use those.

Impedance constraint

The impedance will be calculated either directly or is constrained to a network. The default impedance is defined with the type of distance measurement specified on the Measurement Parameters page (under Data setup). On the other hand, if the impedance is to be constrained to a network, then the network has to be defined.

Default

The default impedance is that specified on the Measurement parameters page. If direct distance is the default distance (on the measurement parameters page), then all impedances are calculated as a direct distance. If indirect distance is the default, then all impedances are calculated as indirect (Manhattan) distance. If network distance is the default, then all impedances are calculated using the specified network and its parameters; travel impedance will automatically be constrained to the network under this condition.

Constrain to network

An impedance calculation should be constrained to a network where there are limited choices. For example, a bus trip requires a bus route; if a particular zone is not near an existing bus route, then a direct distance calculation will be misleading since it will probably underestimate true distance. Similarly, for a train trip, there needs to be an existing train route. Even for walking, bicycling and driving trips, an existing network might produce a more realistic travel impedance than simply assuming a direct travel path. If the impedance calculation is to be constrained to a network, then the network must be defined.

Check the 'Constrain to network' box and click on the 'Parameters' button. The network file can be either a shape **line** or **polyline** file (the default) or another file, either dBase IV 'dbf' or ASCII. If the file is a shape file, the routine will know the locations of the nodes. All the user needs to do is identify a weighting variable, if used.

For a dBase IV or other file, the X and Y coordinate variables of the end nodes must be defined. These are called the "From" node and the "End" node, though there is no particular order. An optional weight variable is allowed for both a shape or dbf file. The routine identifies nodes and segments and finds the shortest path. By default, the shortest path is in terms of distance though each segment can be weighted by travel time, travel speed, or generalized cost; in the latter case, the units are minutes, hours, or unspecified cost units.

Note: using network distance for distance calculations can be a very slow process (i.e., taking many hours or even up to several days for calculating a large matrix).

Minimum absolute impedance

If the mode is constrained to a network, an additional constraint is needed to ensure realistic allocations of trips. This is the minimum absolute impedance between zones. The default is 2 miles. For any zone pair (an origin zone and a destination zone) that is closer together (in distance, time interval, or cost) than the minimum specified, no trips will be allocated to that mode. This constraint is to prevent unrealistic trips being assigned to intra-zonal trips or trips between nearby zones. *CrimeStat* uses three impedances for a constrained network: 1) the impedance from the origin zone to the nearest node on the network (e.g., nearest rail station); b) the impedance along the network to the node nearest to the destination; and c:) the impedance from that node to the destination zone. Since most impedance functions for a mode constrained to a network will have the highest likelihood some distance from the origin, it's possible that the mode would be assigned to, essentially, very short trips (e.g., the distance from an origin zone to a rail network and then back again might be modeled as a high likelihood of a train trip even though such a trip is very unlikely).

For each mode that is constrained to a network, specify the minimum absolute impedance. The units will be the same as that specified by the measurement units. The default is 2 miles. If the units are distance, then trips will only be allocated to those zone pairs that are equal to or greater in distance than the minimum specified. If the units are travel time or speed, then trips will only be allocated to those zone pairs that are farther apart than the distance that would be traveled in that time at 30 miles per hour. If the units are cost, then the routine calculates the average cost per mile along the network and only allocates trips to those zone pairs that are farther apart than the distance that would be traveled at that average cost.

Impedance function

The model split routine can use two different travel distance functions: 1) An already-calibrated distance function; and 2) A mathematical formula. The default is a mathematical formula.

Use an already-calibrated distance function

If a travel distance function for the specific mode has already been calibrated (see 'Calibrate impedance function' under trip distribution), the file can be directly input into the routine. That routine can be used to calibrate a function if there are data on origins and destinations for individual travel modes.

The user selects the name of the already-calibrated travel distance function. *CrimeStat* reads dbase 'dbf', ArcGIS ‘shp’, and ASCII files.

Use a mathematical formula

A mathematical formula can be used instead of a calibrated distance function. To do this, it is necessary to specify the type of distribution. There are five mathematical models that can be selected:

1. Negative exponential – the default
2. Normal distribution
3. Lognormal distribution
4. Linear distribution
5. Truncated negative exponential

For each mathematical model, two or three different parameters must be defined:

1. For the negative exponential, the coefficient and exponent. This is the default and default values are provided.
2. For the normal distribution, the mean distance, standard deviation and coefficient.
3. For lognormal distribution, the mean distance, standard deviation and coefficient.
4. For the linear distribution, an intercept and slope.
5. For the truncated negative exponential, a peak distance, peak likelihood, intercept, and exponent.

Segment measurement unit

The routine can calculate impedance in four ways, by:

1. Distance (miles, nautical miles, feet, kilometers, or meters)
2. Travel time (minutes, hours)
3. Speed (miles per hour, kilometers per hour)
4. General travel costs (unspecified units).

Specify the appropriate units. The default is distance in miles.

Network Assignment

Network assignment involves assigning predicted trips (either all trips or by separate travel modes) to a particular route on a network. That is, for every origin-destination trip link, a particular route is found along a network (roadway, transit). The routine does this using a shortest path algorithm. The user must provide the network with its parameters. The routine allows the definition of one-way streets in order to produce a more realistic representation. In the current version, the assignment routine works on one predicted trip file at a time.

Predicted Origin-Destination file

The predicted origin-destination trip file is a file that lists the predicted number of trips from every origin zone to every destination zone. Select the predicted trip file (i.e., the predicted origin-destination trip file) by clicking on the ‘Browse’ button.

Origin ID

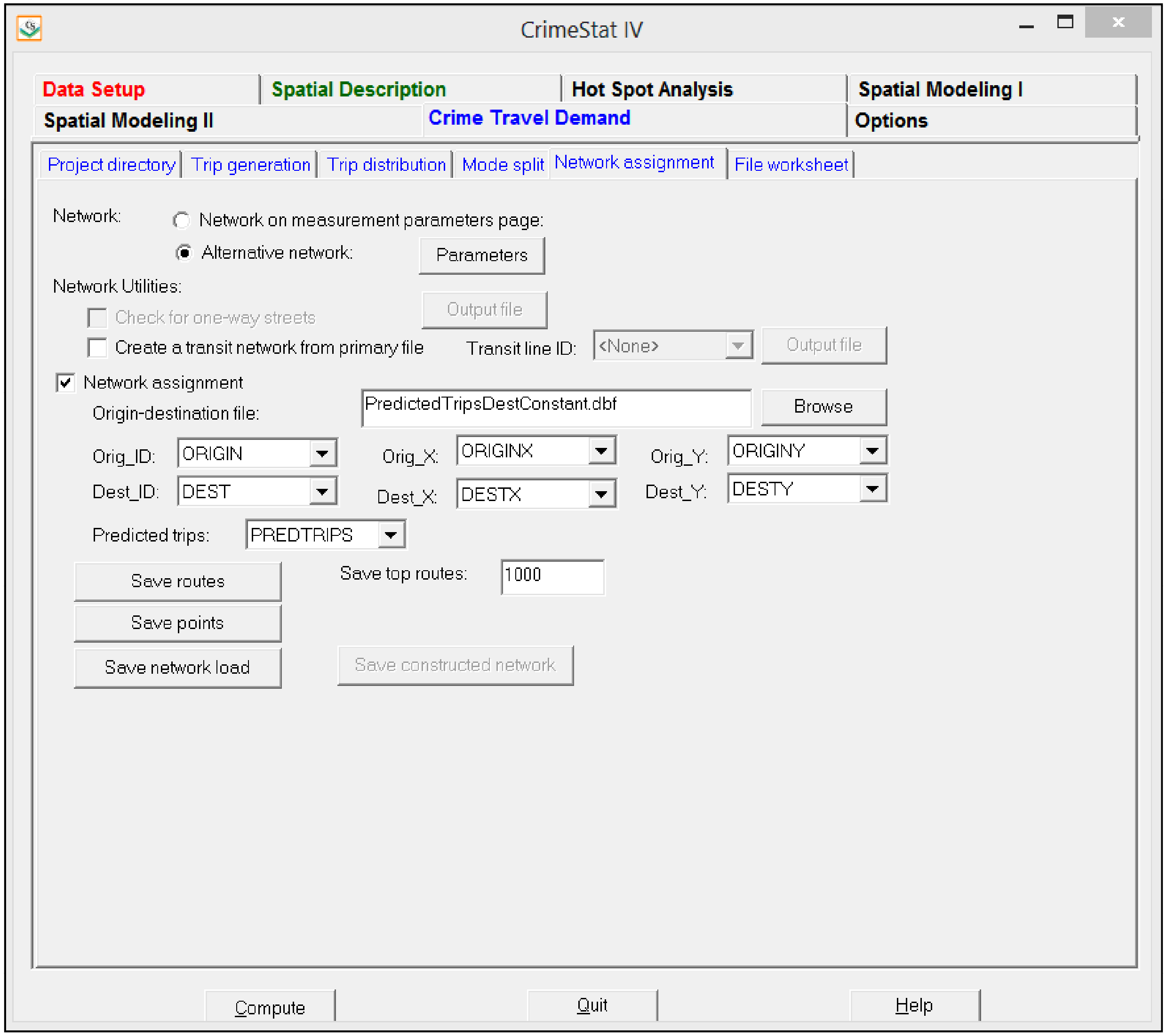
Specify the origin zone ID variable in the data file. The default name is ORIGIN.

Origin_X

Specify the name of the variable for the X coordinate of the origin zone. The default name is ORIGINX.

Figure 2.26:

Network Assignment Modeling



Network: ○ Network on measurement parameters page:
● Alternative network:

Network Utilities:

□ Check for one-way streets

□ Create a transit network from primary file Transit line ID: <None>

☑ Network assignment

Origin-destination file: PredictedTripsDestConstant.dbf

Orig_ID: ORIGIN Orig_X: ORIGINX Orig_Y: ORIGINY

Dest_ID: DEST Dest_X: DESTX Dest_Y: DESTY

Predicted trips: PREDTRIPS

Save top routes: 1000

Origin_Y

Specify the name of the variable for the Y coordinate of the origin zone. The default name is ORIGINY.

Destination ID

Specify the destination zone ID variable in the data file. The default name is DEST.

Destination_X

Specify the name of the variable for the X coordinate of the destination zone. The default name is DESTX.

Destination_Y

Specify the name of the variable for the Y coordinate of the destination zone. The default name is DESTY.

Predicted trips

Specify the variable for the predicted number of trips. The default name is PREDTRIPS

Network Used

The network assignment routine requires a network from which the shortest path from every origin zone to every destination zone can be computed. To run this routine, check the 'Network assignment' box at the top of the page.

The user must specify the network that is to be used. There are two choices.

1. If a network was defined on the Measurement parameters page (Data setup), that network can be used to calculate the shortest path.
2. Whether a network has been defined on the Measurement parameters page or not, an alternative network can be selected. This will take priority if a network has been defined on both pages.

Network on m*easurement parameters* page

Check the 'Network on Measurement parameters page' box to use that network. All the parameters will have been defined for that setup (see Measurement parameters page).

Alternative network

If an alternative network is to be used, it must be defined. Check the 'Alternative network' box and click on the 'Parameters' button.

Note: if a network is also used on the Measurement Parameters page, then it must be defined there as well. CrimeStat will check whether that file exists; if it does not, the routine will stop and an error message will be issued. Therefore, if an alternative network is used, the user should probably change the distance measurement on the Measurement Parameters page to direct or indirect distance.

Type of network

Network files can *bi-directiona*l (e.g., a TIGER file) or *single directional* (e.g., a transportation modeling file). In a bi-directional file, travel can be in either direction. In a single directional file, travel is only in one direction. Specify the type of network to be used.

Network input file

The network file can either be a shape file (line, polyline, or polylineZ file) or another file, either dBase IV 'dbf' or ASCII. The default is a shape file. If the file is a shape file, the routine will know the locations of the nodes. For a dBase IV or other file, the X and Y coordinate variables of the end nodes must be defined. These are called the "From" node and the "End" node. An optional weight variable is allowed for both a shape or dbf file. The routine identifies nodes and segments and finds the shortest path. If there are one-way streets in a bi-directional file, the flag fields for the "From" and "To" nodes should be defined.

Network weight field

Normally, each segment in the network is not weighted. In this case, the routine calculates the shortest distance between two points using the distance of each segment. However, each segment can be weighted by travel time, speed or travel costs. If travel time is used for weighting the segment, the routine calculates the shortest time for any route between two points. If speed is used for weighting the segment, the routine converts this into travel time by dividing the distance by the speed. Finally, if travel cost is used for weighting the segment, the routine calculates the route with the smallest total travel cost. Specify the weighting field to be used and be sure to indicate the measurement units (distance, speed, travel time, or travel cost) at the bottom of the page. If there is no weighting field assigned, then the routine will calculate the path using distance.

From one-way flag and To one-way flag

One-way segments can be identified in a bi-directional file by a ‘flag’ field (it is not necessary in a single directional file). The ‘flag’ is a field for the end nodes of the segment with values of ‘0’ and ‘1’. A ‘0’ indicates that travel can pass through that node in either direction whereas a ‘1’ indicates that travel can only pass from the other node of the same segment (i.e., travel cannot occur from another segment that is connected to the node). The default assumption is for travel to be allowed through each node (i.e., there is a ‘0’ assumed for each node). There is a ‘From one-way flag’ field and a ‘To one-way flag’ field. For each one-way street, specify the flags for each end node. A ‘0’ allows travel from any connecting segments whereas a ‘1’ only allows travel from the other node of the same segment. Flag fields that are blank are assumed to allow travel to pass in either direction.

FromNode ID and ToNode ID

If the network is single directional, there are individual segments for each direction. Typically, two-way streets have two segments, one for each direction. On the other hand, one-way streets have only one segment. The FromNode ID and the ToNode ID identify from which end of the segment travel should occur. If no FromNode ID and ToNode ID is defined, the routine will chose the first segment of a pair that it finds, whether travel is in the right or wrong direction. To identify correctly travel direction, define the FromNode and ToNode ID fields.

Network coordinate system

The type of coordinate system for the network file is the same as for the primary file.

Segment measurement unit

By default, the shortest path is in terms of distance. However, each segment can be weighted by travel time, travel speed, or travel cost.

1. For travel time, the units are minutes, hours, or unspecified cost units.

2. For speed, the units are miles per hour and kilometers per hour. In the case of speed as a weighting variable, it is automatically converted into travel time by dividing the distance of the segment by the speed, keeping units constant.
3. For travel cost, the units need to be defined in terms of cost per unit distance (e.g., per mile, per kilometer). The routine will then identify routes by those with the smallest total cost.

Network Utilities

There are two network utilities that can be used.

Check for one-way streets

First, there is a routine that will identify one-way streets *if* the network is single directional. In a single directional file, one-way streets do not have a reciprocal pair (i.e., a segment traveling in the opposite direction). This is indicated by a reciprocal pair of ID's for the "From" and "To" nodes. If checked, the routine identifies those segments that do not have reciprocal node ID's. The network is saved with a new field called "**Oneway**". One-way segments are assigned a value of '1' value and two-way segments are assigned a value of '0'. The output is saved as an *ArcGIS* 'shp', *MapInfo* 'mif' or various ASCII formats. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Create a transit network from primary file

Second, there is a routine that will create a network from the primary file. This is useful for creating a transit network from a collection of bus stops (bus network) or rail stations (rail network). If checked, the routine will read the primary file and will draw lines from one point to another *in the order* in which the points appear in the primary file. Note, it is essential to order the points in the same order in which the network should be drawn (otherwise, an illogical network will be obtained). It is easy to do this in a spreadsheet program.

Transit Line ID

The routine can handle multiple lines, for example different rail lines or bus routes (e.g., Line A, Line B, Route 1, Route 2). In the primary file, the points must be grouped by lines, however, and must be classified by an ID field. Within each group, the points must be arranged in order of occurrence; the routine will draw a lines from one point to another in that order. In the Transit Line ID field, indicate which variable is the classification variable.

The output is saved as an *ArcGIS* 'shp', *MapInfo* 'mif' or various ASCII formats. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Network Output

There are three types of output for the network assignment routine. First, the most frequent inter-zonal (i.e., trips between different zones) routes can be output as polylines. Second, the most frequent intra-zonal (i.e., trips within the same zone) routines can be output as points. Third, the entire network can be output in terms of the total number of trips that occur on each segment (network load).

Save routes

The shortest routes can be saved as separate **polyline** objects for use in a GIS. Specify the output file format (*ArcGIS* 'shp', *MapInfo* 'mif' or various ASCII formats) and the file name. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Save top routes

Because the output file is very large (number of origin zones x number of destination zones), the user can select a zone-to-zone route with the most predicted trips. The default is the top 100 origin-destination combinations. Each output object is a line from the origin zone to the destination zone with a Route prefix. The prefix is placed before the output file name. The graphical output includes:

1. An ID number from 1 to K, where K is the number of links output (ID)
2. The feature prefix (ROUTE)
3. The origin zone (ORIGIN)
4. The destination zone (DEST)
5. The X coordinate for the origin zone (ORIGINX)

6. The Y coordinate for the origin zone (ORIGINY)
7. The X coordinate for the destination zone (DESTX)
8. The Y coordinate for the destination zone (DESTY)
9. The number of trips on that particular route (FREQ)
10. The distance between the origin zone and the destination zone (DIST).

Save points

Intra-zonal trips (trips in which the origin and destination are the same zone) can be output as separate **point** objects as an *ArcGIS* 'shp', *MapInfo* 'mif' or various ASCII formats. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Again, the top K points are output (default=100). Each output object is a point representing an intra-zonal trip with a RoutePoints. The prefix is placed before the output file name.

The graphical output for each includes:

1. An ID number from 1 to K, where K is the number of links output (ID)
2. The feature prefix (ROUTEPoints)
3. The origin zone (ORIGIN)
4. The destination zone (DEST)
5. The X coordinate for the origin zone (ORIGINX)
6. The Y coordinate for the origin zone (ORIGINY)
7. The X coordinate for the destination zone (DESTX)
8. The Y coordinate for the destination zone (DESTY)
9. The number of trips on that particular route (FREQ)
10. The distance between the origin zone and the destination zone (DIST).

Save network load

It is also possible to save the total network *load* as an *ArcGIS* 'shp', *MapInfo* 'mif' or ASCII file. This is the total number of trips on each segment of the network. The routine takes every origin zone to destination zone combination and sums the number of trips that occur on each segment of the network. For MapInfo 'mif' format, the user has to define up to nine parameters including the name of the projection and the projection number. If the MapInfo system file MAPINFOW.PRJ is placed in the same directory as CrimeStat, then a list of common projections with their appropriate parameters is available to be selected.

Click on the “Save output network” box and specify a file name for the output.

Crime Travel Demand Case Studies

Chapters 31 and 32 present two case studies on crime travel demand, one by Richard Block and one by Dan Helms.

File Worksheet

The file worksheet allows the saving of names for the files in the crime travel demand module. Because there are a large number of files used (many used in multiple routines), saving the names will make it easier to keep track of the files. The file worksheet is not required for use in the crime travel demand module. But we do recommend using it remember the names of files in a particular travel demand model. There are five worksheets for keeping track of the different routines.

File Worksheet 1

This worksheet keeps track of the files used in the trip generation step. These include:

Trip generation

Calibrate model
Make prediction
Balance origins with destinations

File Worksheet 2

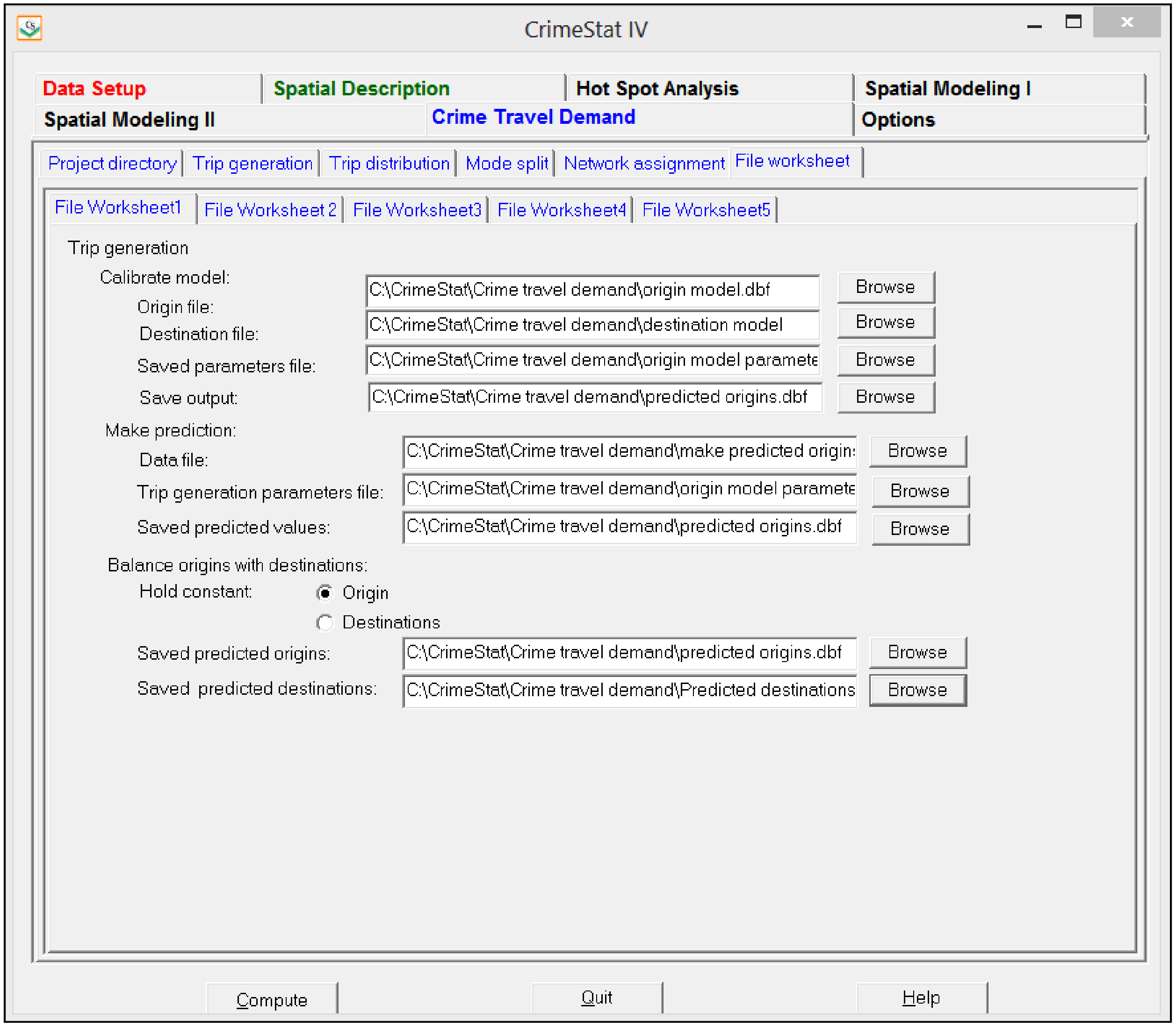
This worksheet keeps track of some used in the trip distribution step, in particular the observed trip distribution and trip distribution model setup. These include:

Trip distribution

Describe origin-destination trips
Setup origin-destination model

Figure 2.27:

Crime Travel Demand File Worksheet



Trip generation

Calibrate model:

Origin file: C:\CrimeStat\Crime travel demand\origin model.dbf

Destination file: C:\CrimeStat\Crime travel demand\destination model

Saved parameters file: C:\CrimeStat\Crime travel demand\origin model paramete

Save output: C:\CrimeStat\Crime travel demand\predicted origins.dbf

Make prediction:

Data file: C:\CrimeStat\Crime travel demand\make predicted origin:

Trip generation parameters file: C:\CrimeStat\Crime travel demand\origin model paramete

Saved predicted values: C:\CrimeStat\Crime travel demand\predicted origins.dbf

Balance origins with destinations:

Hold constant: ● Origin ○ Destinations

Saved predicted origins: C:\CrimeStat\Crime travel demand\predicted origins.dbf

Saved predicted destinations: C:\CrimeStat\Crime travel demand\Predicted destinations

File Worksheet 3

This worksheet also keeps track files used in the trip distribution step, in particular the trip distribution model and the comparison between the observed and predicted trip length distributions. These include:

Origin-destination model
Compare observed and predicted origin-destination trip lengths

File Worksheet 4

This worksheet keeps track of the files used in the mode split step, including the mode split setup and modes 1-3. These include:

Mode split

Setup for mode split
Modes modeled
Modes 1-3

File Worksheet 5

This worksheet keeps track of the remaining files used in the mode split step (modes 4-5) as well as network assignment routine. These include:

Mode split (continued)

Modes modeled
Modes 4-5

Network assignment

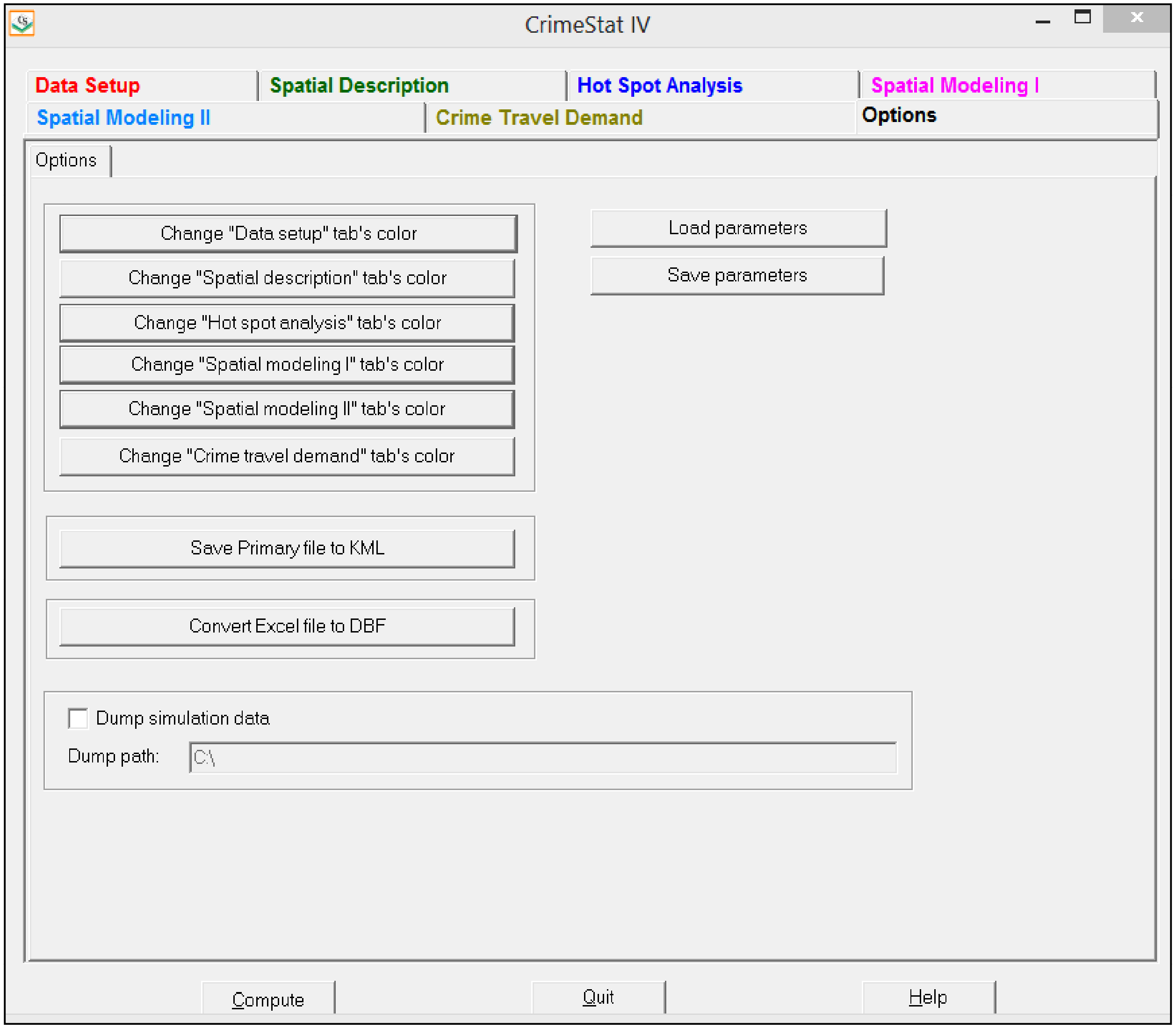
VII. Options

The Options page includes six features that can improve the usability of CrimeStat.

1. Colors for tabs. The user can select one of tens of thousands of colors for each of the major tabs. The Options tab remains black.

Figure 2.28:

CrimeStat Options


CrimeStat IV
Data Setup
Spatial Description
Hot Spot Analysis
Spatial Modeling I
Spatial Modeling II
Crime Travel Demand
Options
Options
Change "Data setup" tab's color
Change "Spatial description" tab's color
Change "Hot spot analysis" tab's color
Change "Spatial modeling I" tab's color
Change "Spatial modeling II" tab's color
Change "Crime travel demand" tab's color
Load parameters
Save parameters
Save Primary file to KML
Convert Excel file to DBF
Compute
Quit
Help

□ Dump simulation data

Dump path: C:\

2. There is a utility for saving the Primary file to *Google Earth* ‘kml’ files ***if*** the coordinate system is spherical (longitude and latitude). *Google Earth* only accepts universal, spherical coordinates so that this option is not available if the data are projected. Many of the routines in *CrimeStat* can save objects as ‘kml’ if the coordinate system is spherical. This utility allows the primary file to be also converted for display in *Google Earth*.

3. There is a utility for converting Excel ‘xls’ and ‘xlsx’ files to ‘dbf’. Excel is a very common format for data storage. However, *CrimeStat* was designed around ‘dbf’ files. The utility allows Excel spreadsheets to be quickly converted to ‘dbf’. Note that only single sheet (page) Excel files can be converted (not multi-sheet files).

   A. Click on the utility and then find the file to be converted.

   B. Define the output name

   C. Click ‘O.K.’ and the file will be converted to a ‘dbf’ file.

4. The user can specify a directory for dumping simulation files (the default is none).

5. The user can *save CrimeStat* parameters in a parameter ‘param’ file. Only top level parameters can be saved, however. The parameters selected on dialogues that open (e.g., Advanced options) cannot be saved.

6. The user can *load* a *CrimeStat* parameter file. Again, Only top level parameters can be loaded.