# An Introduction to Point Pattern Analysis using CrimeStat

Luc Anselin
Spatial Analysis Laboratory
Department of Agricultural and Consumer Economics
University of Illinois, Urbana-Champaign
http://sal.agecon.uiuc.edu/
June 24, 2003

## Introduction

This is a brief introduction to the analysis of patterns in points (as events) using Ned Levine's CrimeStat 2.0 software package. This package is freely available and can be obtained on the web from http://www.icpsr.umich.edu/NACJD/crimestat.html . The data used in this tutorial are the Pittsburgh homicide locations (various Pitt* files) and the Cardiff juvenile offender addresses (juvenile), both obtainable as shape files from the SAL sample data repository http://sal.agecon.uiuc.edu/stuff/data.html. Some familiarity is assumed with either ArcView or ArcGIS, optionally with the Spatial Analyst extension, to implement visualization of various results. CrimeStat does not have its own visualization capability, but relies on an external GIS through the export of result files.

## Getting started with CrimeStat

Start CrimeStat by double clicking its icon. Click on the welcome screen to open the main interface, shown in Figure 1.

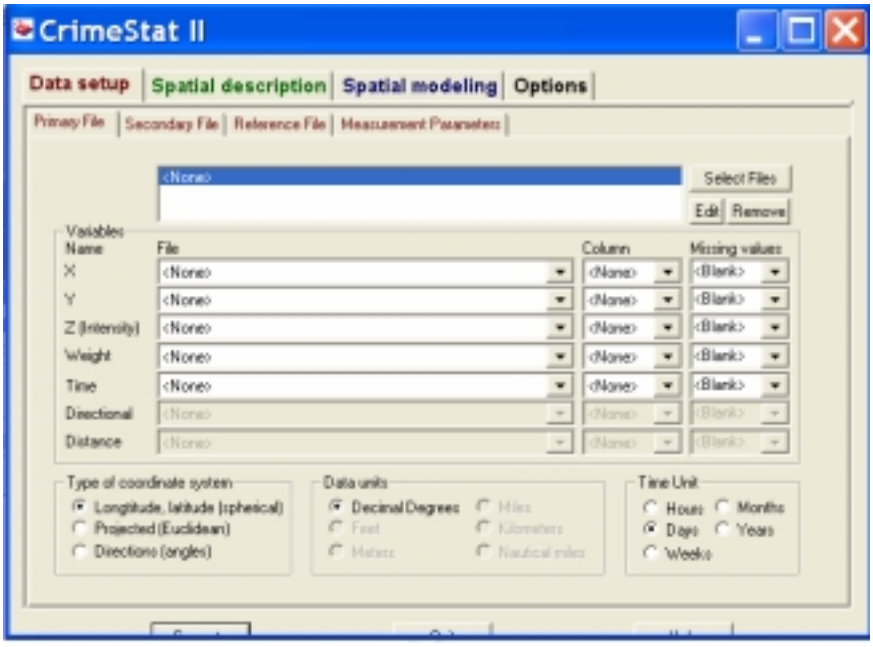


Figure 1. CrimeStat opening screen.

Note how in some systems (like the one used for this tutorial, running Windows Xp) the bottom buttons are not fully legible. They stand for, left to right, Compute, Quit and Help. Help brings up an extensive help system.

The four tabs at the top of the interface correspond to some logical steps in the way CrimeStat implements analysis. First, one needs to set up the data and possibly set some options, then choose the type of analysis (spatial description or spatial modeling). The analysis is run by clicking on the Compute button in the bottom left.

**Data setup in CrimeStat**

CrimeStat reads data from various format files, including shape files. You will be using the juvenile.shp data set from the Bailey-Gatrell text in this example. This is a very simple data file with only the X, Y coordinates of the offender addresses. You can load this into ArcView to have a quick sense of the overall pattern, as in Figure 2.[1]

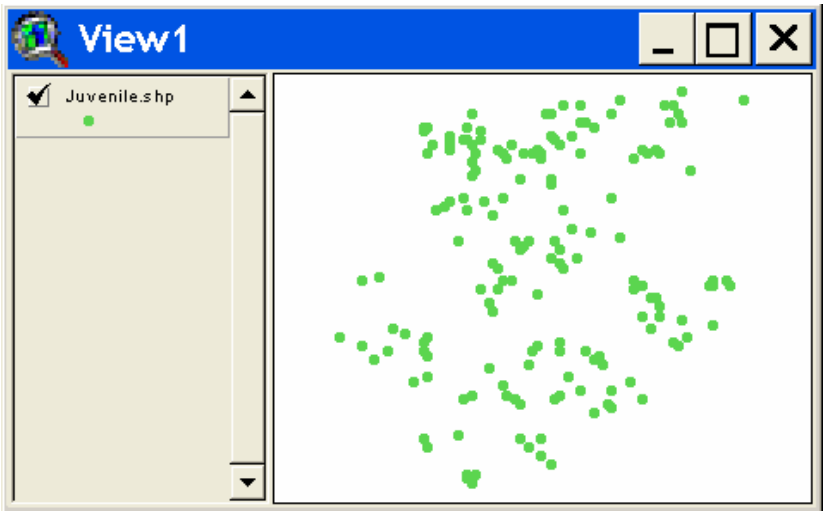


Figure 2. Juvenile point pattern in ArcView.

In CrimeStat, click on the Data setup tab to bring up the interface, shown in Figure 1. In this case, there is only one point pattern, so you only need to specify the Primary File parameters. For case-control studies, you would specify the cases as the Primary File and the controls as the Secondary File. Many routines in CrimeStat also use a Reference File, which is essentially a rectangular grid superimposed over the data points, in order to carry out density estimation or interpolation.

In the Primary File tab, specify the juvenile.shp file as the input file. Click on the Select Files button and on Browse in the File Characteristics dialog. Move around in the file system until you locate the juvenile.shp file in your working directory, as shown in Figure 3. Select this file and confirm the selection in the File Characteristics dialog (click

[1] Open ArcView, and, with the Views icon active, click on New. Then "add a theme" by clicking on the + (plus) icon and locate the juvenile.shp file. Make sure to click on the check mark to make the theme visible.

OK) and the file name will appear in all the input fields in the user interface, as shown in Figure 4.

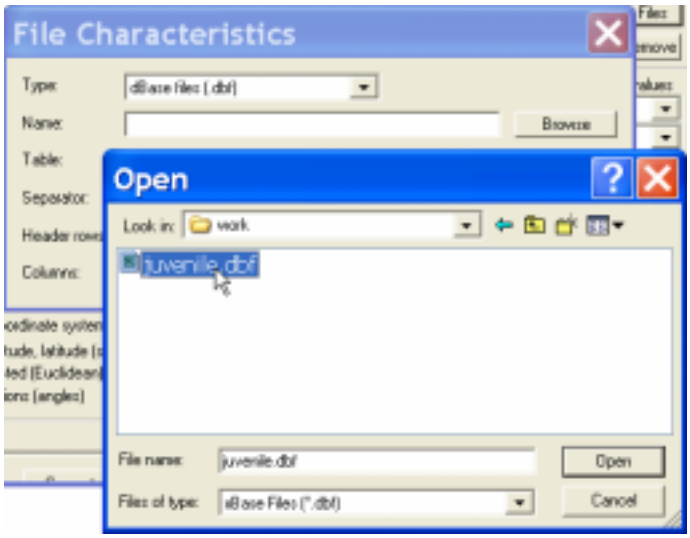


Figure 3. Select input file (Primary File)

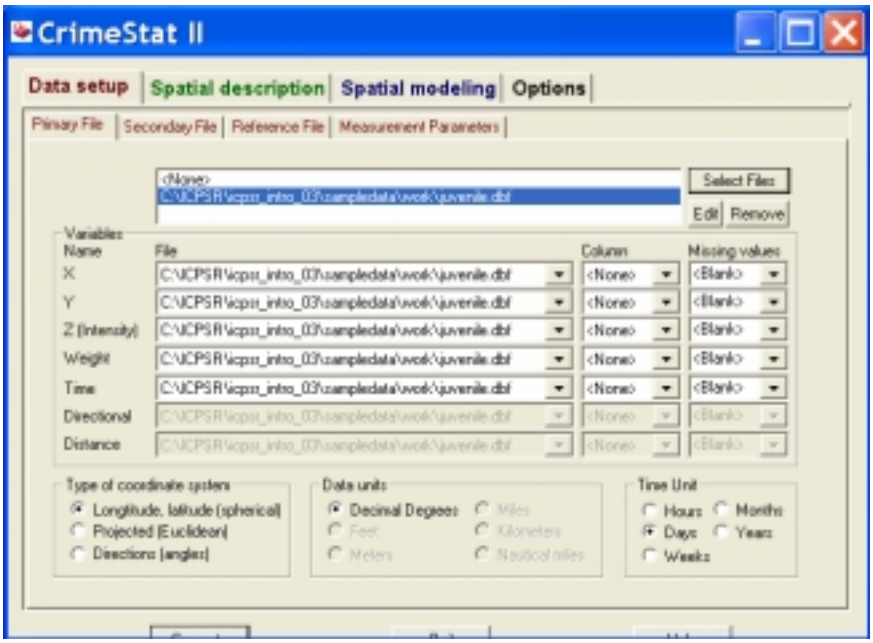


Figure 4. Primary File available for the specification of variables.

Several options are available to specify weights, time and directional effects besides the X, Y coordinates. You will not be using these additional items in the current exercise, but feel free to explore their use (CrimeStat comes with an extensive manual and several sample data sets). For now, set X to X and Y to Y. Also make sure the coordinate system is set to Projected. You can ignore the data units and time unit since there is no time in the Juvenile data set. Your Data setup interface should now look as in Figure 5.

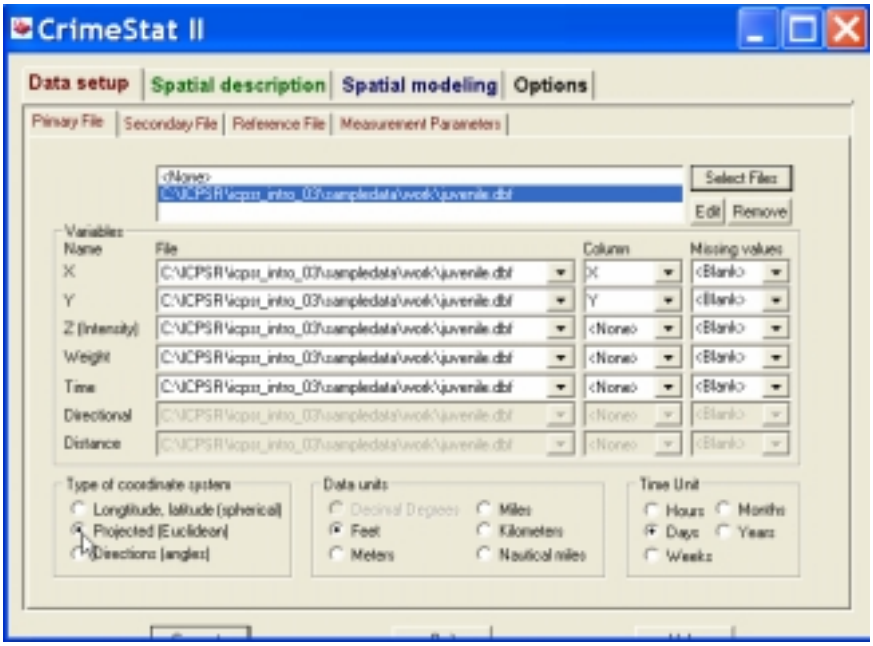


Figure 5. Variables and projection selected.

Next, specify a reference grid that will be used in the kernel density estimation routines. Click on the Reference File tab and check the Create Grid radio button. Now, specify 0, 0 as the lower left and 100, 100 as upper right as in Figure 6, and leave the default to 100 grid columns. In the Cardiff data set, the actual bounding rectangle is 2, 6 to 94, 95 (you can find out "manually" using the Identify function in ArcView).

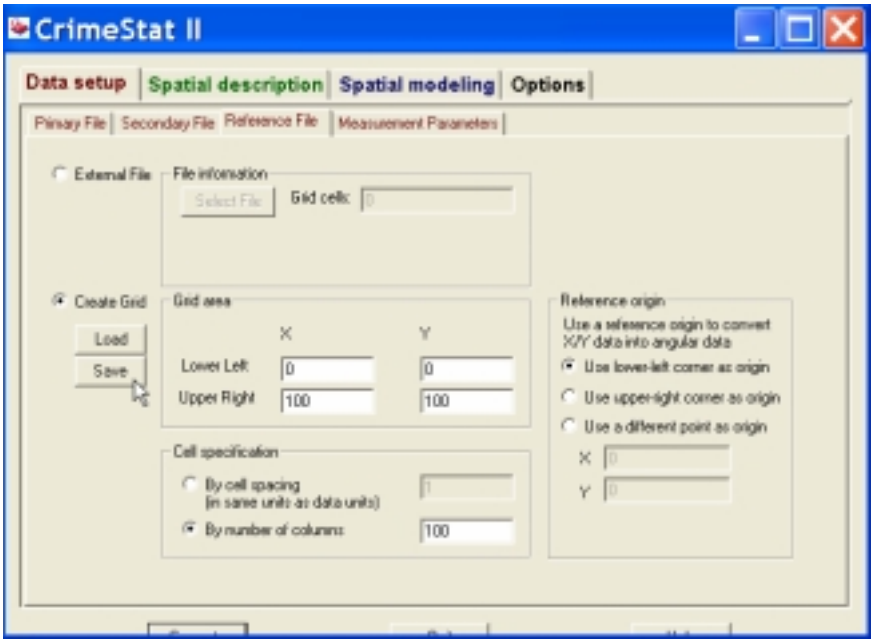


Figure 6. Specifying a Grid as a reference file.

Save the Grid specifications for later use to a file in your working directory. This is a two-step process. First, you "Save" the grid specification by clicking on the Save button in the interface (Figure 6). This brings up a dialog to name the particular grid setup, as in Figure 7. However, this does not save it to a file. To store this (and other) named grid specifications in a file, first Load them (Load button in Figure 6) and then Save to File in the following dialog, as in Figure 8.

*Practice*

Start a second instance of CrimeStat and use the Pitthom.dbf file as the primary file. Set up the coordinates and the reference grid (use the Identify button in ArcView to determine the coordinates of the lower-left and upper-right corners of the bounding rectangle).

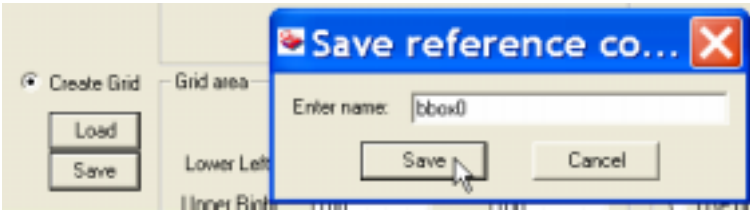


Figure 7. Save reference grid.

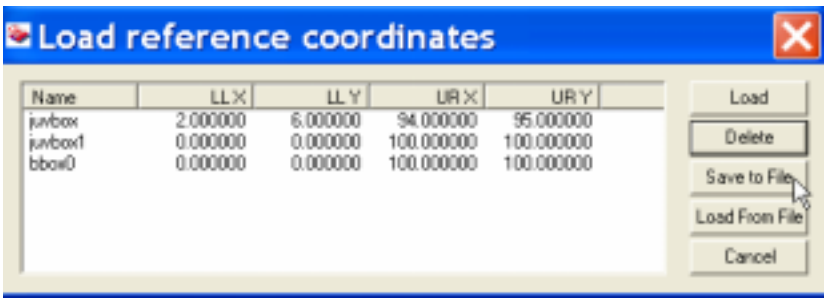


Figure 8. Load and save reference coordinates from/to a file.

**Centrography**

Basic descriptive statistics of the overall pattern of the points are the mean and median center and the standard deviational ellipse. These summaries are computed in CrimeStat as part of the Spatial Description tab under the Spatial Distribution items. The output can be saved to a shape file (as well as other formats) for overlay on the point pattern. With the juvenile point pattern as the Primary File (and with the X and Y coordinates specified), click on the Spatial Description tab to bring up the Spatial Distribution dialog. Select the check boxes next to Mean center and standard distance (Mcsd), Standard deviational ellipse (Sde) and Median Center (MdnCntr), as in Figure 9. Also specify file names for the output to be saved to a shape file (make sure to select ArcView "SHP" as the option in the Save output to list, as in Figure 9). This needs to be done for each descriptive statistic. Click on the Compute button (lower left) to start the calculations.

The results will appear in a screen, as in Figure 10. Note the tabs on the top of the screen, which let you select the output for each set of descriptive statistics (Mcsd, Sde, and MdnCntr). You can now save these results to a text file, or print them out.

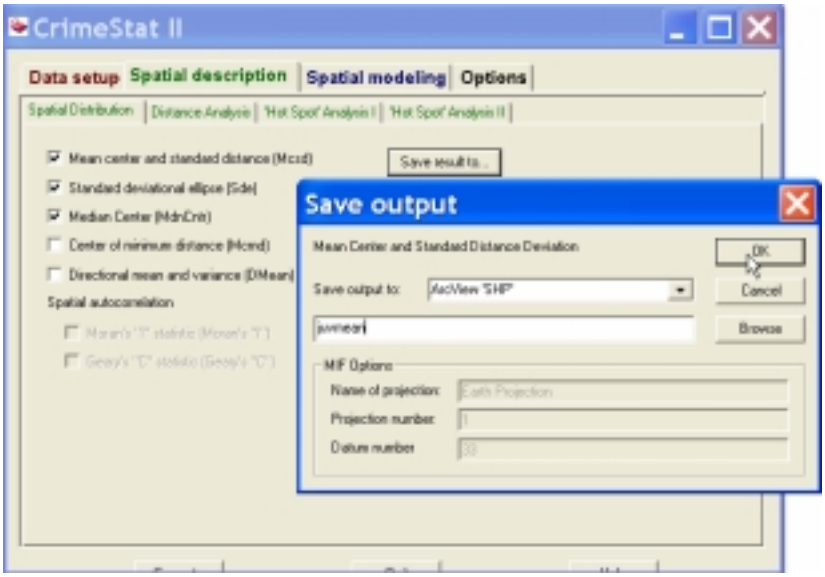


Figure 9. Centrography settings.

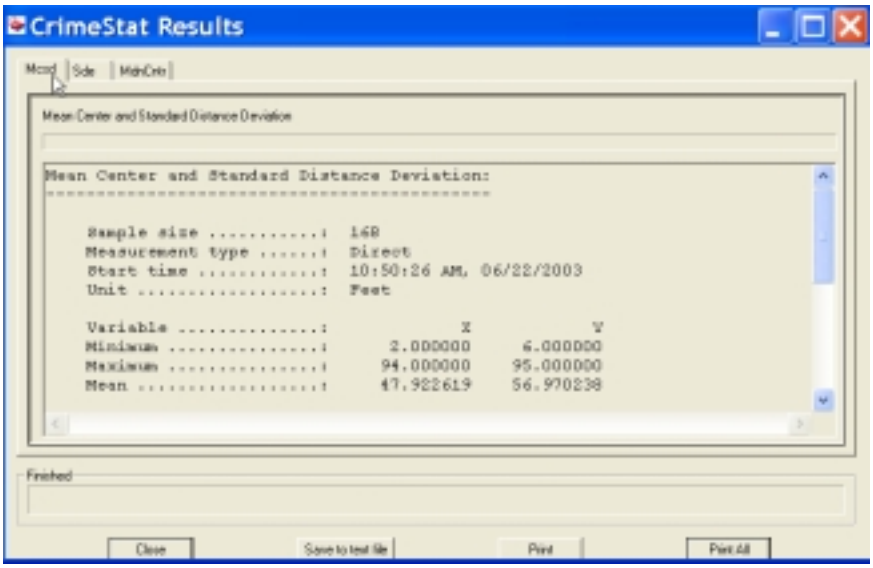


Figure 10. Mean center result screen.

A more visual representation of the centrographic characteristics of the point pattern is obtained by overlaying the computed shape files on the original pattern. When the “Save Result to” option is used, a number of shape files are created in the current working directory. These have the same name as specified in the file dialog, with a prefix indicating the type of results contained in them. The mean center is in the MC*file*.shp shape file, median center in MdnCntr*file*.shp, standard deviational ellipse in

*2SDE*file.shp, etc. (see the CrimeStat manual for a full list of options). For example, in Figure 11, the mean center (green cross), median center (red dot), standard deviational rectangle (black rectangle), standard distance deviation (blue circle) and standard deviational ellipse (green ellipse) are illustrated for the juvenile point pattern. The matching shape file names are given in the legend panel. In this particular application, the mean and median centers are practically the same and there is only a slight indication of a directional effect (the circle and ellipse are very close, a strong directional effect would be shown when the ellipse would be very elongated along one axis).

*Practice*

Carry out an centrographic analysis of the point pattern for the Pittsburgh homicides (use Pitthom.shp as the Primary File). Overlay the results on the map of points in ArcView (or ArcGIS). For a challenge, compare and visualize the summary statistics between the homicides in 93 and 94. To accomplish this, you will need to build a query in ArcView/ArcGIS to select those observations for which the Event_yr is 93 (or 94), followed by a Theme > Convert to Shape file command to create a separate shape file for the selected observations. Then specify the new shape file as the Primary File in the CrimeStat analysis.

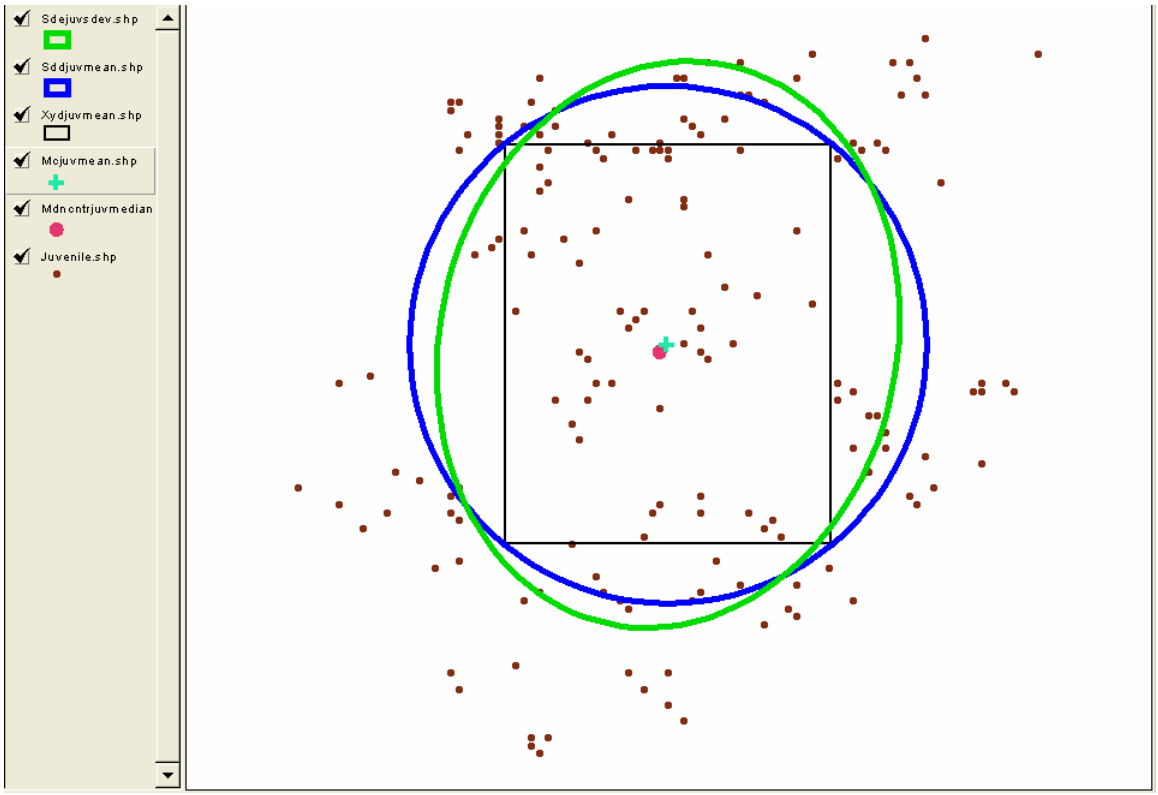


Figure 11. Centrography of Cardiff juvenile offender addresses.

**Kernel density estimation**

Kernel density estimation is implemented under the Interpolation functionality of the Spatial modeling tab in CrimeStat. Note that, strictly speaking, kernel density estimation is not an interpolation technique, but more precisely the estimation of a probability surface. CrimeStat includes five kernel estimators: the normal, uniform, spherical, conical and negative exponential (see the CrimeStat manual, pp. 301-309, for mathematical details). The main difference between these is that the normal includes all points in the pattern, whereas the others have a distance cut-off beyond which no points are included in the kernel estimation. The kernel is estimated for each point on a grid that is overlaid on the point pattern (the Reference File, Grid).

Each of the kernels allows you to specify a kernel bandwidth as a fixed interval or to choose an adaptive interval (the default). It is important to understand that the default for an adaptive interval is to include 100 points in the kernel estimation for each location on the grid. For small data sets, this will tend to lead to a very smooth (and largely uninformative) surface. You can also set the bandwidth to a fixed distance range (interval). This requires that the point coordinates are in meaningful units to yield distance measures in common units, such as feet, meters or miles. This is not the case in the juvenile example. Some experimentation with the bandwidth specification is typically necessary.

The kernel values are computed as totals, or absolute density (points per grid area, rescaled such that the sum over all grids adds up to the observed total), as points per areal units, or relative density (i.e., points per square mile), or as probabilities. Output can be saved in a number of formats, such as a shape file of grid polygons.

To compute some different kernel surfaces for the Cardiff juvenile data, make sure to use the juvenile.shp file as the Primary File and set the Grid option in the Reference file to 100 grids starting at 0, 0 (lower left) up to 100, 100 (upper right). In the Spatial Modeling tab, select Interpolation. Keep the method as normal, but specify 25 as the minimum sample size in the Adaptive bandwidth specification. For now, keep the calculation option to the default of Absolute Densities. Your setup should look as in Figure 12. You can practice later with changing these settings.

To visualize the estimated surface in ArcView, make sure to set the “save results to” option to the shape file option, as in Figure 13. This will write a shape file to the working directory that has the interpolated value as the Z variable for each square grid. In Figure 13, the file name juvgrid0 was specified. Click on Compute to start the process.

The results appear in a CrimeStat results window, as shown in Figure 14. You must click on the Kernel Density tab to see the results (the default is to show the results for the first analysis tab). You can scroll through the results window or save the results to a text file. You must use the slider bar to the left of the results and the Go button to see more than the first 45 interpolated values. In addition, a shape file Kjuvgrid0.shp has been added to your working directory.

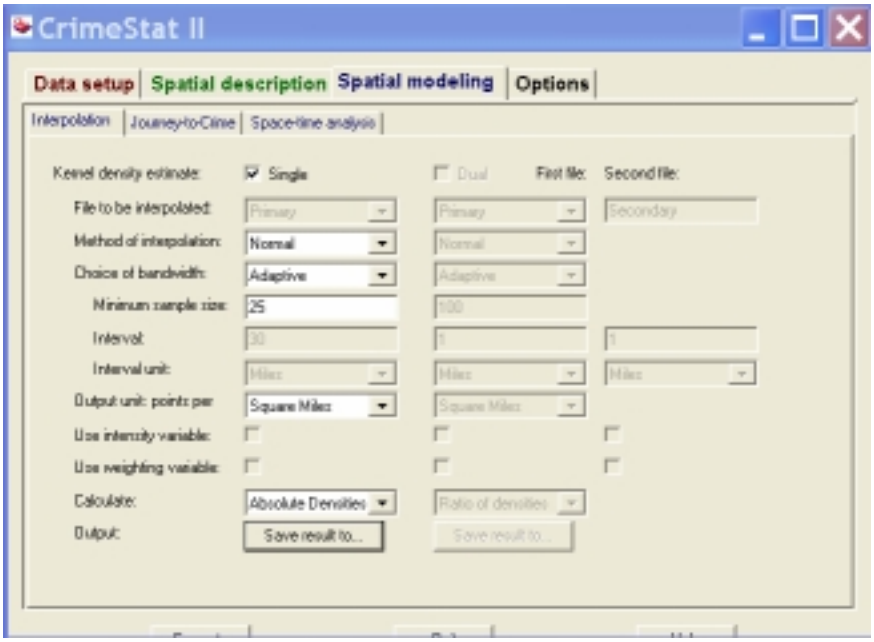


Figure 12. Normal kernel density estimation setup.

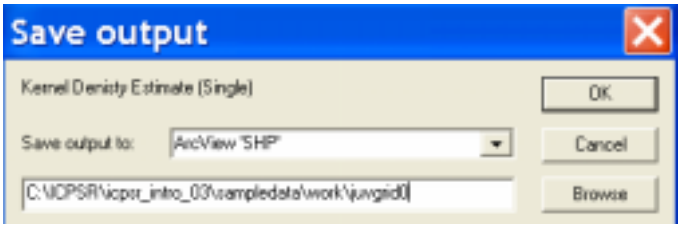


Figure 13. Kernel output file specification.

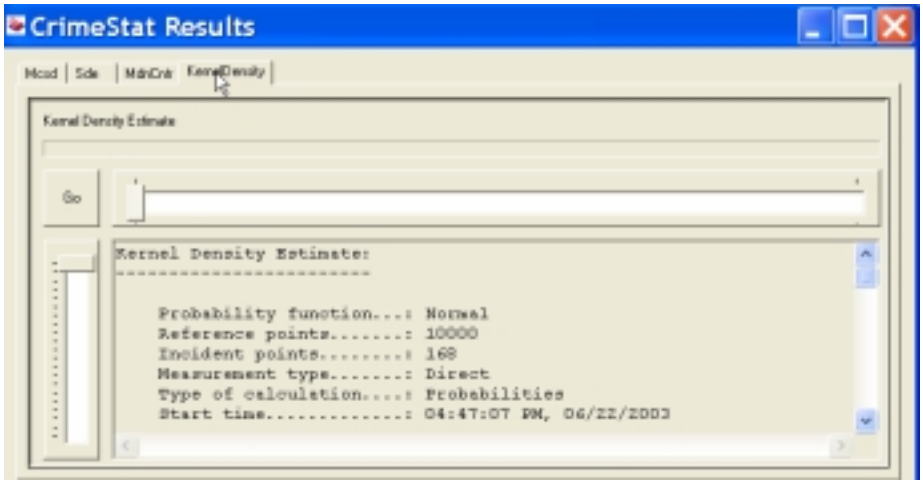


Figure 14. Kernel calculation results.

To visualize the estimated surface, open up ArcView and add the Kjuvgrid0 theme. You must use the legend editor to turn this into a meaningful grid map. In ArcView (use similar commands in other GIS software) select Graduated Color as the legend type and use Z as the Calculation Field. For now, you can keep the default Natural Breaks classification (you can experiment with different types of classifications later). The resulting grid map will be as in Figure 15, with the original point pattern superimposed.

If you are familiar with ESRI's Spatial Analyst extension, you can convert the polygon grid shape file to Spatial Analysts's "grid" format and then use the Surface analysis functionality to superimpose contour lines on the density surface. For example, Figure 16 is the result of such an operation, with the extent of the grid and contour themes set to the extent of juvenile.shp, and with 0.005 as the contour step.[2] Note how the normal kernel density tends to smooth the surface and remove a lot of the underlying detail in the original pattern. Experiment with changing some of the parameters, such as the kernel bandwidth.

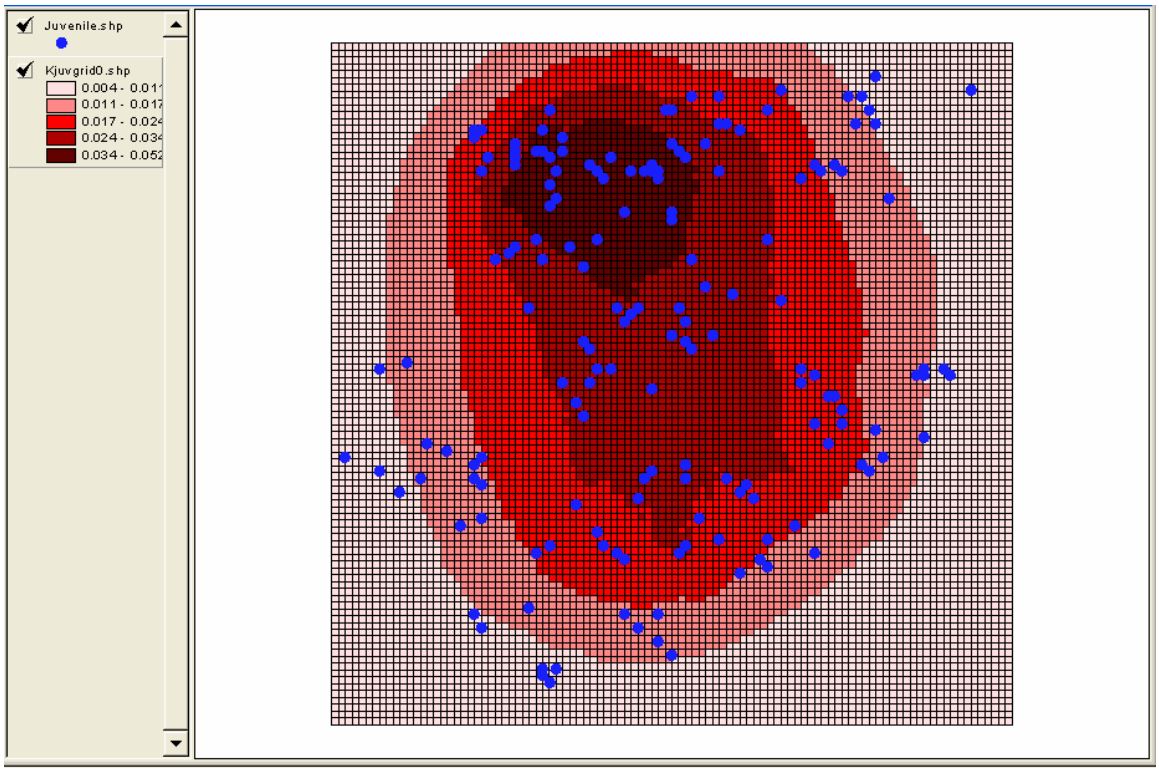


Figure 15. Normal kernel density.

[2] The steps involved in obtaining this result are as follows, illustrated for ArcView (they are slightly different in ArcGIS). First, make sure the Spatial Analyst extension has been activated. With the grid polygon theme active, select Theme > Convert to Grid, set the extent to that of juvenile.shp, the number of rows, columns to 100, and Z as the field for the grid values. Add the new grid theme to a View (make sure to set the input file type to grid instead of Feature) and then select Surface > Create Contours. Set the interval to 0.005 and you should see the same result as in Figure 16. Make sure to use the legend editor to turn the contour theme into a "graduated color" with "contour" as the variable.

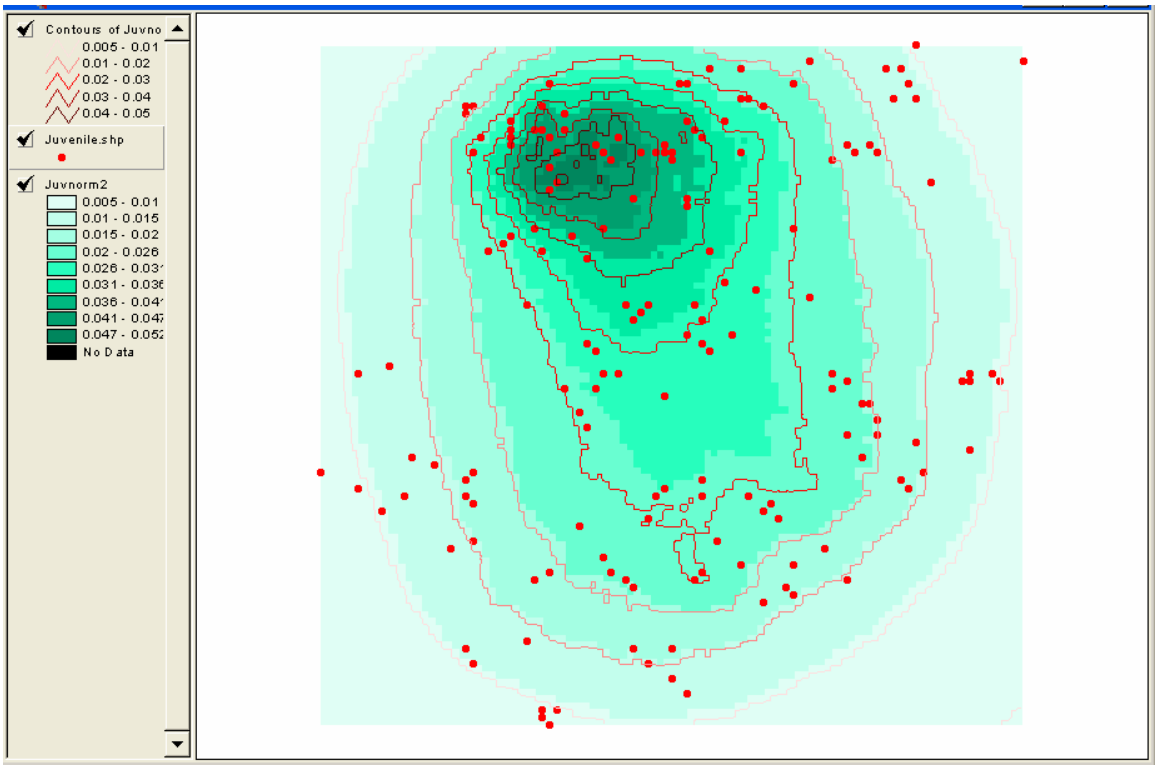


Figure 16. Normal density kernel with contour lines.

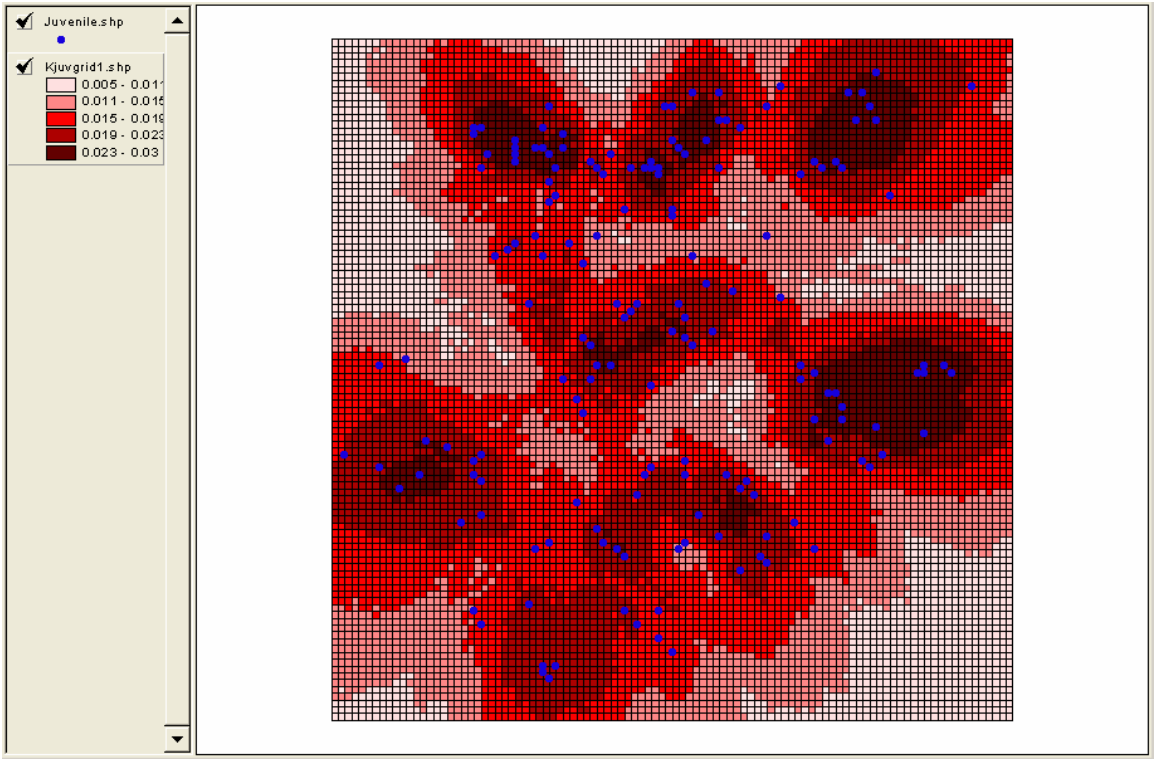


Figure 17. Triangular kernel density.

The kernel density results are very sensitive to the choice of kernel model and settings. To illustrate this, consider a Triangular kernel with the same settings as for the Normal (25 points cutoff for the Adaptive bandwidth). The result is as in Figure 17, which is much spikier and focuses on several "hot spots" rather than the central tendency reflected in Figures 15-16. Experiment with different settings, for example, changing the bandwidth to 50. The kernel densities differ with respect to how steep the cutoff is and how smooth the resulting surface will be. Some trial and error will start to show some persistent patterns in the data, suggesting potential clusters or hot spots.

*Practice*

Use the Pittburgh homicide file to construct a normal and triangular (or other type) kernel density surface. If you are comfortable with the Spatial Analyst, display the surfaces as isoline maps (contour maps). If you have created separate shape files for 93 and 94, you can compare the "clusters" and "hot spots" suggested by the kernel densities between the two years. You can also create different shape files using the other variables to distinguish between point patterns, such as gang-related vs. non gang-related, or using guns vs. not using guns.

**Nearest Neighbor distance statistic**

In order to more formally assess the extent to which a point pattern shows clustering or dispersion, two main classes of techniques can be applied. The first uses the magnitude and/or distribution of inter-point distances, or the distances between the points and reference locations as an indicator (distance based tests). The second set of methods uses the number of points within a given area as the basis for test statistics (quadrat counts). The simplest of the distance based statistics uses the distribution of the distance to the nearest neighbor as a measure. If this distance tends to be smaller than what it would be under complete spatial randomness, this suggests clustering. If, on the other hand, it tends to be larger, then dispersion is the suggested alternative.

A Nearest Neighbor statistic is implemented in CrimeStat in the Distance Analysis tab of Spatial Description. Make sure you have the juvenile.shp file as the primary file with the proper coordinates and projection set. You do not need the Reference File for these calculations. In the Distance Analysis dialog, check the box next to Nearest Neighbor Analysis (Nna) and leave the defaults to their settings, as in Figure 18.

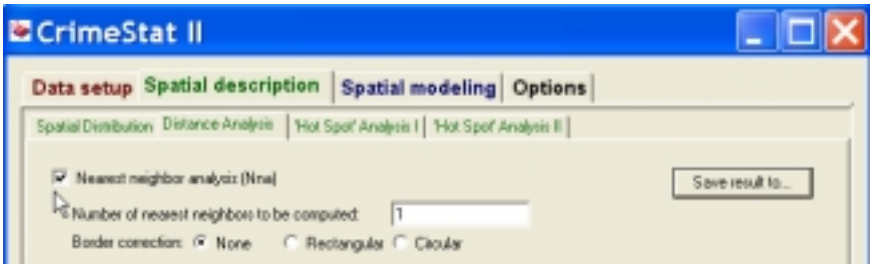


Figure 18. Nearest Neighbor analysis setup.

Click on the Compute button to carry out the analysis. The result window will contain the summary statistics. You can save these to a text file, the contents of which are as in Figure 19. Note that it does not make much sense to save the results to a dbf file, since only the summary distance statistic will be saved, not the full distribution of nearest neighbor distances. The results yield a Nearest Neighbor Index of 0.7003, which is obtained by taking the ratio of the observed mean nearest neighbor distance to the mean random distance. The value less than 1 suggests clustering. A test statistic can be constructed by taking the difference between the observed and random mean nearest neighbor distance and standardizing by the standard error. The resulting Z-value of -7.43 is well above the usual critical values, suggesting "significant" clustering. However, these tests have to be interpreted with some caution. Also, there are many nearest neighbor based statistics, and they don't necessarily lead to the same conclusion.

You can assess the sensitivity of the results to a number of settings, such as the use of border corrections.[3]

*Practice*

Compute the nearest neighbor index to assess the extent of clustering of the Pittsburgh homicide point pattern. Compare the results for the two periods combined and for each time period separately. Also compare the findings for different types of crimes (guns or not, gangs or not). Try different border corrections to assess the sensitivity of the results.

```
Nearest neighbor analysis:
--------------------------

    Sample size........: 168
    Measurement type...: Direct
    Start time.........: 06:18:04 PM, 06/22/2003

    Mean Nearest Neighbor Distance ..:  2.44 ft
    Standard Dev of Nearest
    Neighbor Distance ...............:  1.84 ft
    Minimum Distance ................:  0.00 ft
    Maximum Distance ................:  106.89 ft

    Based on Bounding Rectangle:
    Area ............................:  8188.00 sq ft
    Mean Random Distance ............:  3.49 ft
    Mean Dispersed Distance .........:  7.50 ft
    Nearest Neighbor Index ..........:  0.7003
    Standard Error ..................:  0.14 ft
    Test Statistic (Z) ..............:  -7.4315
    p-value (one tail) ..............:  0.0001
    p-value (two tail) ..............:  0.0001

             Mean Nearest           Expected Nearest       Nearest
  Order      Neighbor Distance (m)  Neighbor Distance (m)  Neighbor Index
  *****      **********************  **********************  **************
    1                 2.4445                 3.4906              0.70030
```

Figure 19. Results of nearest neighbor analysis.

[3] Note that in order to use the Manhattan distance (linear nearest neighbor index) feature, you must specify the total length of the street network, which is not available for the Juvenile or Pittsburgh data sets.

**Ripley's K function**

The nearest neighbor distance statistics are described as "first order" statistics, since they only consider the distance to the nearest point. Second order distance statistics consider the complete distribution of all distances in the point pattern. Ripley's K function is an example of such a second order statistic, and is essentially a test on the cumulative distribution function of the full set of inter-point distances. This distribution can be compared to a reference distribution under complete spatial randomness. A higher proportion of shorter distances than random would suggest clustering, whereas a higher proportion of longer distances suggests dispersion.

CrimeStat implements Ripley's K function under the Distance Analysis of the Spatial Description tab. The program does not report the actual K function results, but instead the L function, which is simply a rescaled K function such that the reference for complete spatial randomness is linear and horizontal (at zero).

Make sure the juvenile.shp primary file is set, with the proper coordinates and projection. Check the box next to Ripley's K statistic (and uncheck the box next to Nna), set the number of simulation runs to 1000 and specify a dBase file for the output, as illustrated in Figure 20. The cumulative distribution, organized in 100 distance bins will then be written to a dBase file. Note that the program will add the prefix Ripley to whatever filename you specify, so in the example of Figure 20, the dBase file will be called Ripleyjuvripley.dbf (and not juvripley.dbf as you might expect).

Click on Compute to start the calculation and simulation runs. This may take a while, depending on how many simulation runs you specified. When the program is finished, click on the Ripley tab in the results window to see the output, as illustrated in Figure 21.

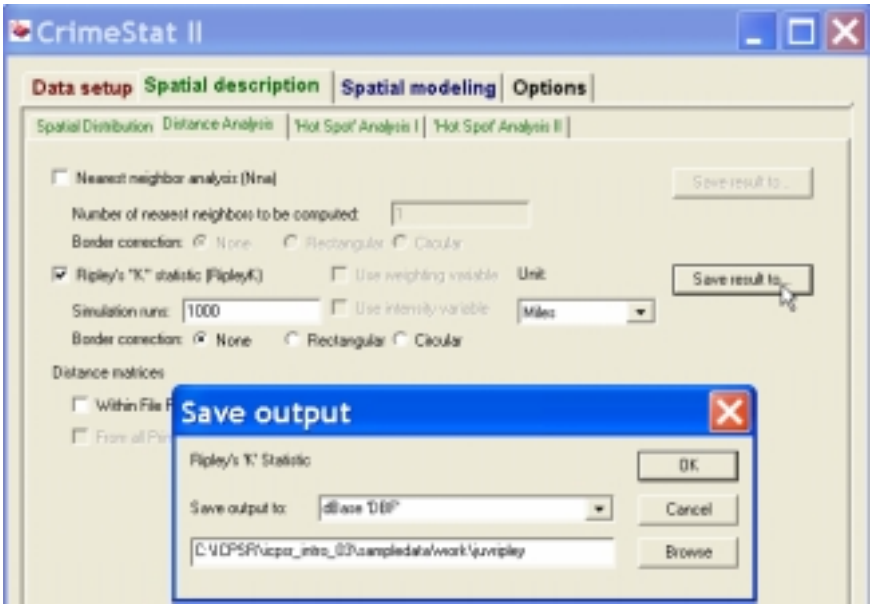


Figure 20. Ripley's K setup.

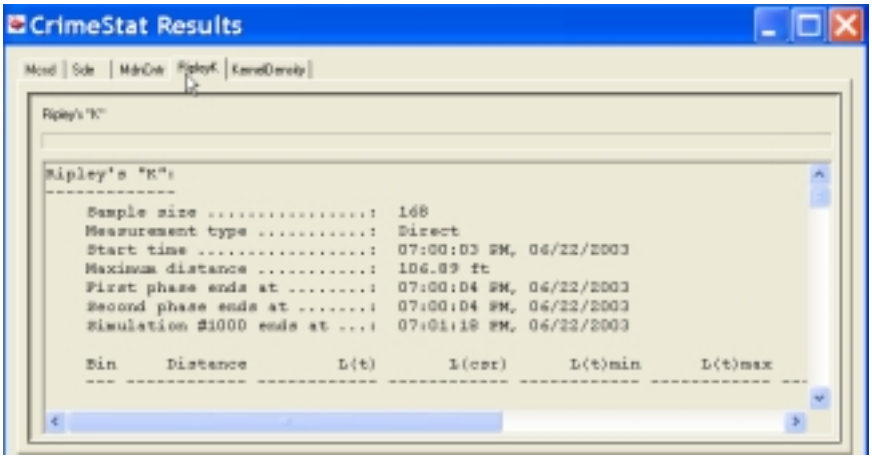


Figure 21. Ripley's K results window.

The results window is not that easy to interpret, since it is basically a list of values of the L function and selected quantiles under complete spatial randomness for 100 distance bins. To better visualize this, load the output dbf file (Ripleyjuvripley.dbf) into a spreadsheet or graphing package and turn it into a graph, as in Figure 22. This was accomplished by using an Excel scatter graph with the bin distances as the X-axis and on the vertical axis the L values, L(csr), a horizontal line at zero, L(t)max and L(t)min, the max and min of the randomization envelope. Note how the dark blue line is outside the randomization envelope for shorter distances, suggesting clustering.

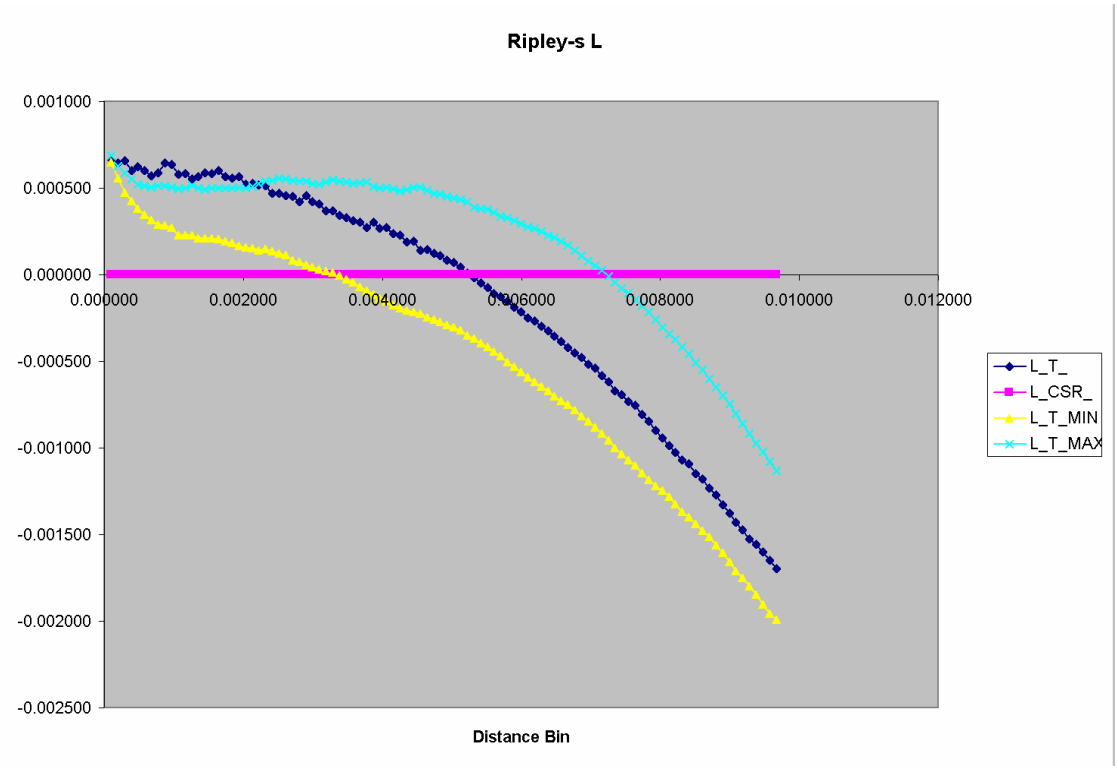


Figure 22. Ripley's L function.

*Practice*

As before, use the Pittsburgh homicide data and Ripley's K function to assess overall clustering of homicides (overall, by year and/or by type). Visualize the computed distributions in a spreadsheet or graphing package. Experiment with border adjustments to assess the sensitivity of the results.

**Hot Spot detection (STAC)**

Quadrat methods assess the presence of clusters by comparing the number of events (points) within a given region to the number expected under complete spatial randomness. The STAC (Spatial and Temporal Analysis of Crime) method is a form of quadrat method. More precisely, it is a combination of a scan statistic (counting the number of events within a circle) and a hierarchical clustering technique (points that are present in more than one identified "clustered" circles result in all the points in the two circles to be combined). The results are visualized as a standard deviational ellipse computed for the points identified to be a "cluster" or "hot spot." The significance of the identified cluster can be assessed by means of a Monte Carlo randomization method.

STAC is implemented in CrimeStat under the Hot Spot Analysis II tab of the Spatial Description tab. Make sure the juvenile.shp file is set as the Primary File with the proper coordinates and projection specified, and set the Reference File as the 100 x 100 grid with origin at 0, 0, as before. Also set the Data Units to Kilometers. Check the box next to STAC on the interface, and set the Output Units to Kilometers, as in Figure 23. Click on the "save ellipses to" button to specify the output file for the standard deviational ellipses as a shape file and enter the file name in the text box, as in Figure 24. Finally, you need to set the parameters for the STAC algorithm (make sure you have specified the Grid option in the Reference File tab or STAC won't work). As in any clustering operation, the results of STAC are quite sensitive to these parameters. The most important ones are the search radius (STAC uses a circle with a fixed radius in the scan operation) and the minimum number of points to consider a cluster. Both of these are context specific and may require some trial and error. For example, setting the search radius too large or too small may not yield any clusters. In Figure 25, the settings are 10 for the search radius and 5 for the minimum number of points. Also specify 1000 for the number of randomizations (this is not required for the STAC algorithm to work). Click on Compute to start the analysis. This yields 3 clusters, as shown in the results window in Figure 26.

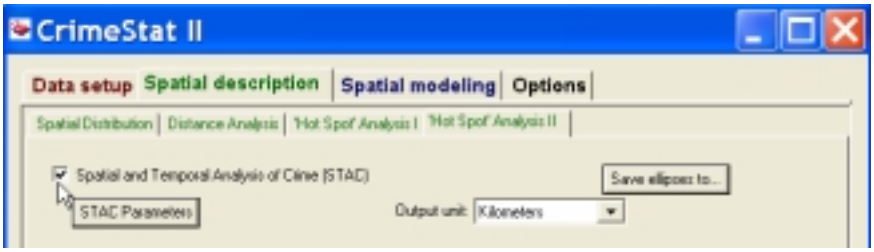


Figure 23. STAC setup interface.

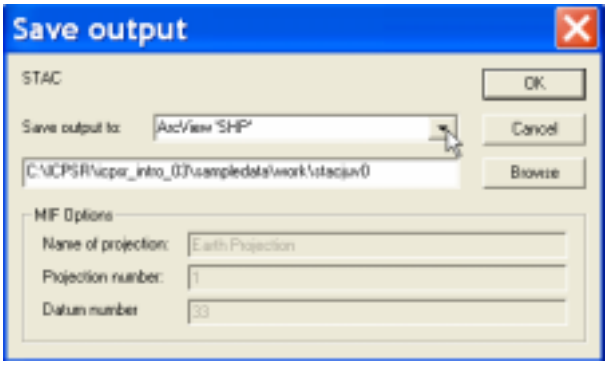


Figure 24. STAC output file specification.

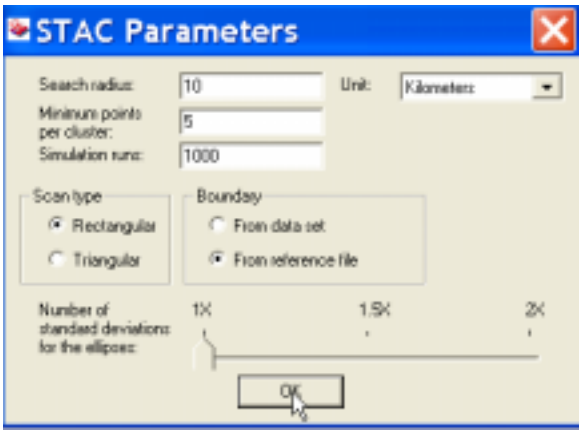


Figure 25. STAC parameters.

```
Spatial and Temporal Analysis of Crime:
---------------------------------------

    Sample size ...........:  168
    Measurement type ......:  Direct
    Scan type..... ......:  Rectangular
    Input units .... ......:  Kilometers
    Output units ... ......:  Kilometers, Square Kilometers, Points per Square Kilometers
    Standard Deviations ...:  1.0
    Start time ............:  11:19:44 AM, 06/23/2003
    Search radius..........:  10.000000
    Boundary...............:  0.00000,0.00000 to 100.00000,100.00000
    Points inside boundary.:  168
    Simulation runs .......:  1000

    Cluster          Mean X          Mean Y     Rotation       X-Axis       Y-Axis            Area Points Cluster Density
    ------- --------------- --------------- ------------ ------------ ------------ --------------- ------ ---------------
          1        43.85938        61.28125      8.67519     22.30442     35.22888      2468.53727    128         0.051853
          2        77.04762        47.00000     24.11872     10.73316      8.14698       274.70964     21         0.076444
          3        28.42857        10.85714     42.92881      9.78010      3.65763       112.38099      7         0.062288

    Distribution of the number of clusters found in simulation (percentile):
    Percentile Clusters            Area       Points          Density
    ---------- -------- --------------- ------------ ----------------
           min        1        17.44290            5         0.017810
           0.5        1        30.28371            5         0.020087
           1.0        1        34.63255            5         0.022049
           2.5        2        43.87791            6         0.023968
           5.0        2        55.88136            6         0.026582
          10.0        2        68.65395            6         0.028804
          90.0        6      2821.55054          107         0.106745
          95.0        7      3086.45174          115         0.129728
          97.5        7      3361.59426          120         0.159534
          99.0        9      3557.21758          123         0.188806
          99.5        9      3933.35003          124         0.205723
           max       11      4216.99908          129         0.343979
```

Figure 26. STAC results

For the search radius of 10, the results are not that useful. Three clusters are identified, and their mean center, area, number of points and density are listed in the results page (Figure 26). When superimposing the ellipse shape file on the point pattern, it is obvious that the first (largest) cluster is not a useful "hot spot" in that it contains 128 out of the 168 points in the pattern, as shown in Figure 27. Resetting the search radius to 5 yields 10 clusters, shown in Figure 28. You can further experiment with setting a different search radius, changing the minimum number of points for a cluster, etc.

Another interesting comparison is to overlay the STAC ellipses on the kernel density grid, to get further insight into the overall patterns in the points. As shown in Figure 29, there is some correspondence between some of the clusters and the higher elevation densities, but not total. In part this is due to the different densities in the clusters (not all of them are high density since they may have resulted from collapsing several initial clusters).

*Practice*

Use the Pittsburgh homicide data (pitthom.shp) to carry out a hot spot analysis using STAC. Experiment with different search radii. Start with 500, using miles as the distance unit and 50194,87016 – 110183,127712 as the bounding box. Increase the radius and assess the effect. As before, you can also carry out analyses for the individual years and/or crime types. Compare the STAC ellipses to one of the kernel density estimates and assess the degree of similarity in the suggestion of clusters and hot spots.

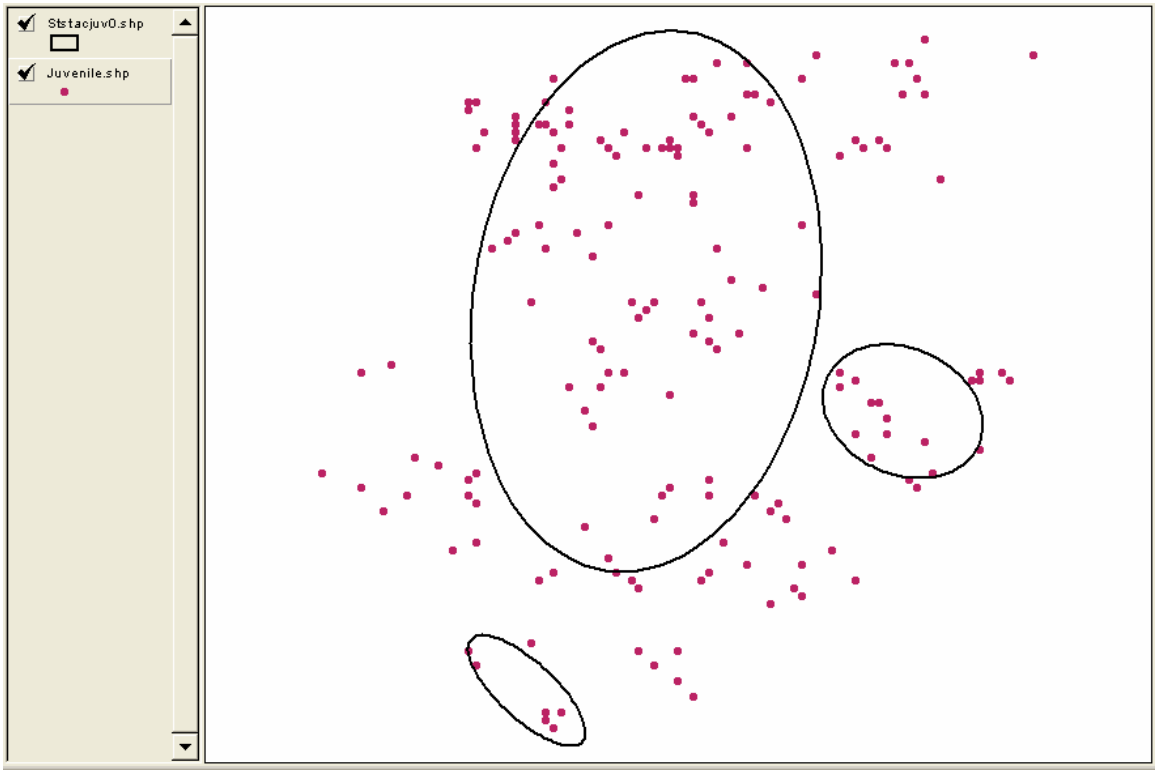


Figure 27. STAC ellipses on point pattern.

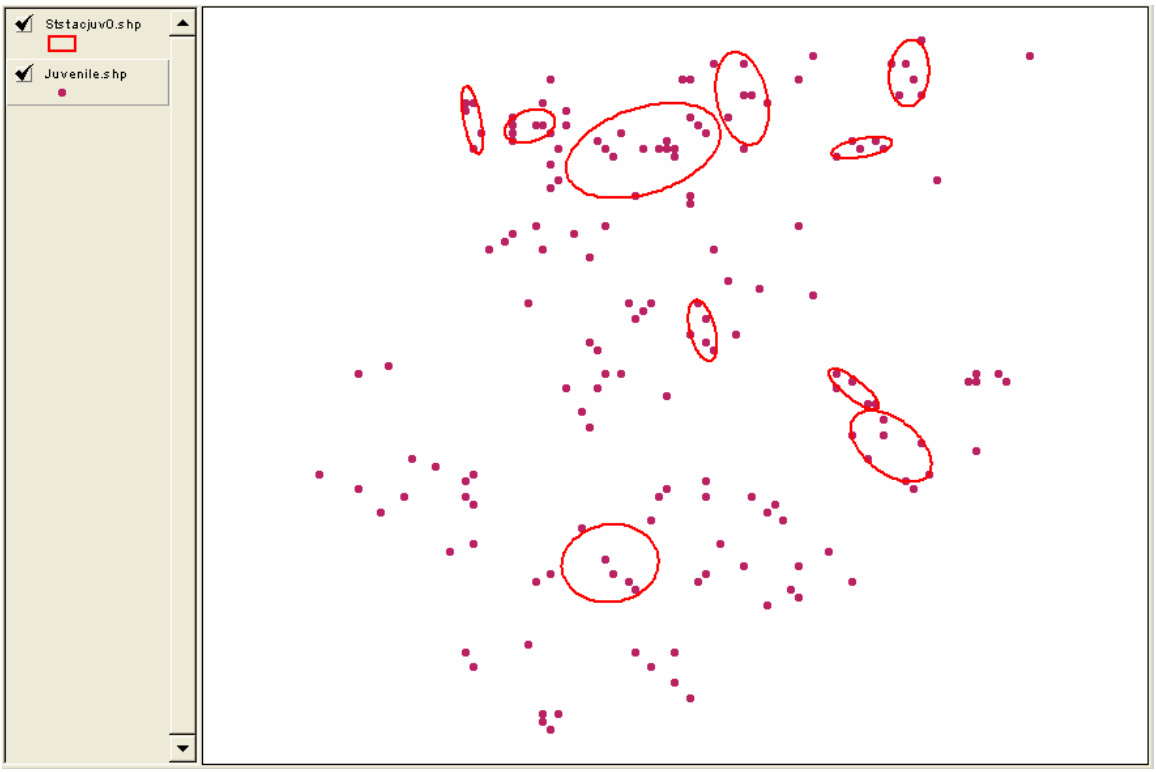


Figure 28. Refined STAC ellipses.

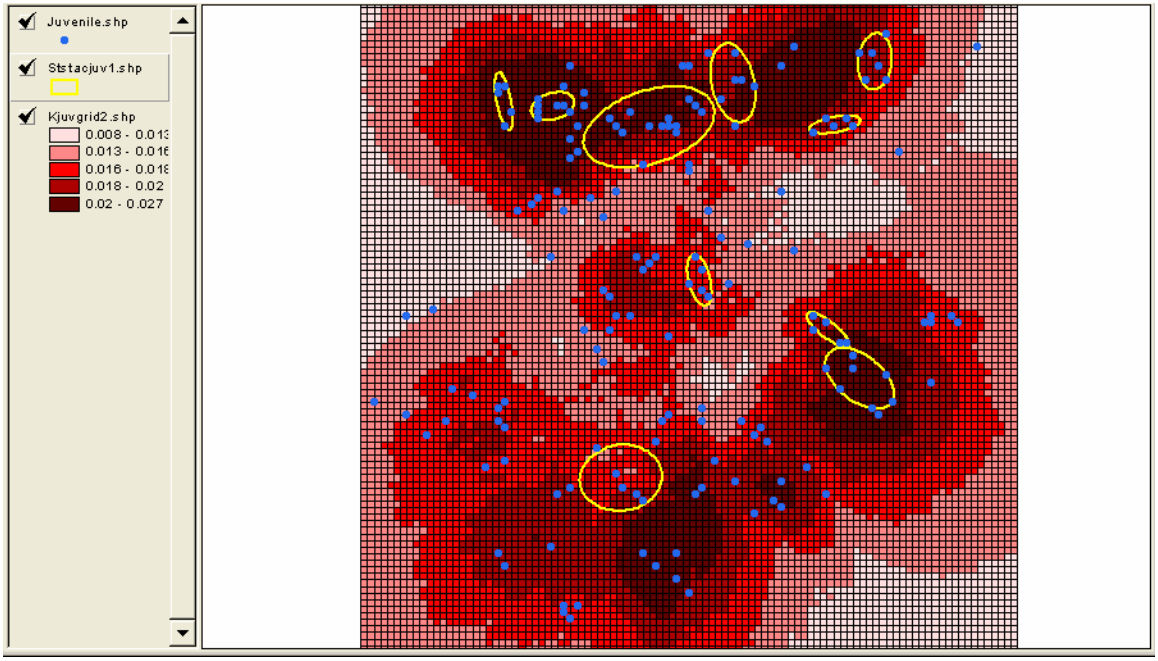


Figure 29. STAC ellipses and triangular kernel density.