CSDE GIS Workshop Series
Point Patterns and Raster Surfaces

Matt Dunbar & Chris Fowler
mddunbar@uw.edu|csfowler@uw.edu

Introduction

The goal of these exercises is to give you a chance to put the concepts we have just discussed into practice. Keep in mind we have only a limited amount of time, so our focus today will be breadth rather than depth! Don't worry, we'll offer more workshops in the future and we are always available to schedule a consultation to work on your own questions in greater detail. Please feel free to **ask questions** as we proceed if something doesn't make sense, and certainly be vocal if find you've missed a step.

Text in **bold** refers to actual commands to be performed in order to complete the tasks of this lab. Text in callout-boxes form is meant to explain or develop the actions we are taking and may be most useful if you find yourself returning to these instructions at a later date.

To begin the exercise we first need to set up our computer with the correct data files. CSSCR's computers only allow you access to read and write files from the C:\temp directory, so we will start by copying our files there.

Exercise Setup
All of the materials for the course are available on the CSDE workshop web site at:
http://csde.washington.edu/services/gis/workshops/PTSURF.shtml

- Navigate to this page and scroll down to the link for "All Workshop Materials (.zip)"
- Click through this link and Save it. A download box should appear. When the file is done downloading you can double-click on this folder icon and you will see the folder we need for our workshop.
- Open another windows explorer instance (Windows Button + E is the shortcut) and navigate to C:\temp
- If there is a folder called "PTSURF" here select and delete it.
- Drag the PTSURF folder into C:\temp.

Part 1 – *Event Data and Density Surfaces*

The purpose of this portion of the lab is to look at Point Event Data to inform ourselves of the kinds of patterns that are present in our data. We will begin with some basic data preparation and then move on to explore density surfaces and the importance of choosing appropriate boundaries and neighborhood sizes for our analysis.

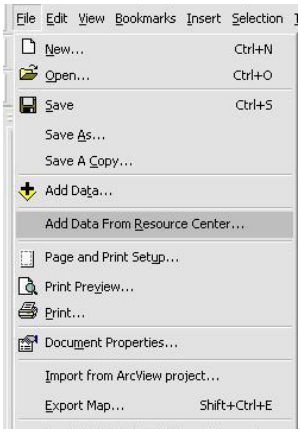
Data Prep: Sub-setting Seattle Crime Data.

1. **Open ArcMap on your computer and begin with a blank project**
2. **Add the layer Seattle_Crime.**

Layers are added in ArcMap using the "Add Data" button

3. **Add aerial photography to situate our point data in a known environment**
    a. **File -> Add Data from Resource Center**.

This will open up an internet browser and take you to an ESRI webpage. On this webpage, you can download layer files that reference web services.



    b. **Click on the "Imagery" icon to download the layer files**.
    c. **Save it to "C:\temp\PTSURF\Data\"**.
    d. **Close your internet browser**.
    e. **Add the World_Imagery.lyr file**.
4. **Subset our Data to only include Certain types of Crime Event**
    a. **Selection->Select By Attributes**
        i. **"CrimeType" = 3530 click "Apply"**
        ii. **Method = "Add to current selection"**
        iii. **"CrimeType" = 3532 click "Apply"**
        iv. **"CrimeType" = 3560 click "Apply"**
        v. **"CrimType" = 3562 click OK**
    b. **Right Click on Seattle_Crime**
        i. **Data->Export Data**
        ii. **Save file as "Pot_Coke2"**
        iii. **Add to display**
    c. **Uncheck "Seattle_Crime" to limit number of points being drawn at any one time**

At this point we have a base layer and a subset of Seattle Crime data with which we can work. There should be 634 observations in our smaller data set. Plenty to work with for the techniques we will be using.

Standard Distance

In this segment of the lab we will review some of the descriptive statistics covered in the previous workshop on Exploratory Spatial Data Analysis. This will allow us to answer a fundamental question about our data; Do criminal incidents related to the sale of drugs and the possession of drugs vary in space? Do they vary by drug?

5. **Spatial Statistics Tools->Measuring Geographic Distributions->Directional Distribution**
    a. **Case Field = CrimeDescr**
    b. **Double-click on Pot_Coke_DirectionalDistrib layer**
        i. **Symbology->Categories->Unique values**
        ii. **Value Field = CRIMEDESCR**
        iii. **Add All Values**
        iv. **Double click on colored box next to each value to modify display settings**
            1. **Fill Color = No Color**
            2. **Outline Width =2**
            3. **Outline Color = Pink-Possession Cocaine, Light Blue- Possession Marijuana, Red- Sell Cocaine, Blue – Sell Marijuana**

These circles represent the minimum elliptical area necessary to contain 66% of the crime events associated with each type of crime. We can see real variation in the distributions of these crime types, and given an actual research question related to differences in these crime types we would probably want to create each of the density surfaces below for each of the four types of data. These ellipses are crude tools for describing the locations of our points, but the fact that they show differences suggests that we may learn something from parsing this data further.

6. **At this point it may be helpful to turn off our Imagery and Seattle_Crime layers so we can focus on the patterns in our data without distraction and long load times. We can come back to them if need be.**

Point Density Surface

In point density we first overlay a grid on our study area. Next, for each cell in our grid we count the number of events occurring within a given distance from the center of our cell. This value is then divided by the area of the shape we used to calculate number of events giving us a local value for the points per unit squared. Point Density surface gives us a way to see changes in density in our data. As we will see, its values are open to a wide range of interpretations, and it is best understood as an exploratory technique or as a measure to be added into some other analysis.

7. **Point Density-Quadrat Counts**

For our first density surface we will replicate the quadrat counts from the lecture. This is something of an abuse of the point density function, but it gives a sense of how point density works and will provide a good lead into the other point density examples.

    a. **Spatial Analyst Tools->Density->Point Density**
    b. **Change the name of the output raster to "…\Quadrat"**
    c. **Output cell size = 5280 (our cells will be 1 square mile)**
    d. **Neighborhood = Rectangle**
    e. **Units = Map**

f. **Height and Width = 5280**
g. **Run**
h. **Right-click on Quadrat layer->Properties->Symbology**
    i. **Alter Color Ramp to something more varied**
    ii. **Classify->Method=Geometric Interval**

By setting our cell size and Neighborhood to the same size and by setting that size to be 1 square mile (the units in which the result is reported) we have replicated the simple quadrat counts from the lecture. The results can be understood as "X pot and cocaine related crimes per square mile." This value is reported for each 1 mile x 1 mile square within the study area. Some of the results may not appear correct based on visual inspection, but this is because points may be stacked on top of one another for various reasons and so what looks like 3 distinct points may, in fact, represent 5 crimes.

8. Point Density with circular neighborhood

a. **Spatial Analyst Tools->Density->Point Density**
b. **Change the name of the output raster to "…\Pt_Den"**
c. **Output cell size = 1320 (our cells will be 1/4 mile square)**
d. **Neighborhood = Circle**
e. **Units = Map**
f. **Radius = 2979 (gives a circle with area equivalent to 1 square mile)**
g. **Run**
h. **Right-click on Pt_Den layer->Properties->Symbology**
    i. **Alter Color Ramp to something more varied**
    ii. **Classify->Method=Geometric Interval**

When we move to a circular neighborhood and change cell size to be smaller than the neighborhood we begin to take information from outside the cell and use it to generate the value within the cell. We chose our radius to get a circle with the same area (in square feet) as the square in the previous exercise (e.g. from Area =πr^2 we get r = $\sqrt{\frac{(5280x5280)}{\pi}} = 2979ft.$) This allows us to read our results as comparable to the "crimes per square mile" as above.

9. Point density –using a smaller neighborhood

a. **At the bottom of the Tools Menu select the "Results" tab**
b. **Expand "Current Session" and Double click on the top instance of Point Density**

The results tab allows us to easily access a record of all the functions we have used in our session so far. When we save our project as a .mxd file this record will be saved until the next time we open the project

c. **Adjust the Raster name to Pt_Den2**
d. **Radius = 2000**

When we make the radius of our circle smaller we reduce the extent of the high density areas and get a tighter resolution with respect to our centers of criminal activity. Note also that the highest value (408 crimes per square mile) is double the highest value (216) obtained using the larger circle radius.

Kernel Density

Point Density generated a surface that conveyed information centered on a given cell—events per square mile. We could vary the cell size and vary the neighborhood considered, but we were always drawing some boundary and counting points within it. This is good because we are generating "real" information; the values report characteristics of the data at a given location. Kernel density changes the game around, and instead of drawing a circle (or square) around our cell, we weight each event so that its "impact" decreases with distance. Next, for each grid cell, we sum all of the impact values for points with a positive impact overlaying the center of that cell. The total is reported as the kernel density for that cell and generates a surface. This method of calculation has the advantage of smoothing the harsh transitions of the point density method since events are not in/out of an area, but contribute to a lesser extent as they move away from the center of any given grid cell. However, in the process of introducing this smoothing function we are making claims about the impact of each event that are significant. In this case, the presence of a crime is assumed to impact an area surrounding the actual crime, and so our surface generates something akin to a *perceived criminality* index. Note that this is actually somewhat reasonable under the circumstances since the events we are using have been shifted in space somewhat to marginally protect the identity of individuals—the point is our best guess at crime location, but we know it to be at least partially false.

1. **Kernel Density**

    a. **Select the "Favorites" Tab to go back to the list of Tools**
    b. **Spatial Analyst Tools->Density->Kernel Density**
    c. **Input feature = Pot_Coke**
    d. **Output Raster = "…\Kernel"**
    e. **Output cell size = 1320 (our cells will be 1/4 mile square)**
    f. **Search Radius = 2000**
    g. **Run**
    h. **Right-click on Pt_Den layer->Properties->Symbology**
        i. **Alter Color Ramp to something more varied**
        ii. **Classify->Method=Geometric Interval**

Part 2 –*Point Pattern Analysis*

In the first part of the lab we used density surfaces to explore patterns of intensity in our data. We could visualize where events were more tightly clustered and where they were more dispersed. This is a good step for understanding our data. It is also a good step if what we want to do is generate a spatial variable that summarizes event data for inclusion in some larger analytic project. Missing from this process is any effort to quantify the extent of clustering or dispersion in our data. The second portion of this lab will explore two techniques for doing just this: Nearest Neighbor Analysis and Ripley's K

Nearest Neighbor Analysis

This technique calculates the average distance from each point to its closest neighbor and compares this distance to the expected average distance if the points were generated randomly. It is possible to estimate not only the average distance, but also its standard error, so we can then generate an associated Z-score and p value to give us some sense of the probability that the observed distribution of distances could have occurred at random.

1. **Spatial Statistics Tools>Analyzing Patterns>Average Nearest Neighbor**
    a. **Input Feature Class = Pot_Coke**
    b. **Display Output Graphically**

Our analysis indicates the near impossibility that the observed density of observations could have occurred by chance. Before we assign too much value to this result however, it is worth noting that the nearest neighbor statistic uses a minimum bounding rectangle to supply the Area value used to generate the expected distances. Given the irregular shape of the City of Seattle and the presence of a lot of water, our Area value is clearly inflated. Let's go again with a more accurate Area estimation

2. **Switch to the Results Tab and Open up the Average Nearest Neighbor function**
    a. **Change Area to** 2,340,211,390 (no commas)

The area input above is based on a summary of the areas of all block groups in Seattle with water features removed (see Intro to GIS II lab for details on how to access and clip this data). We then add an additional field to the attribute table (Type = Double) and Calculate Geometry to obtain the area for each block group. Finally we can right click on our new Area field and generate statistics for the field. Having done all this our Nearest Neighbor Distance is substantially closer to our expected distance, but still at the very clustered scale of things. Even though we have a better measure now, however, we need to be careful. For privacy reasons our data has all been moved to the nearest street intersection, so disparate points are recorded as piled up at these locations. In practice, our data is almost certainly clustered—we have done nothing to control for population density at the very least, since population is clustered, it would make sense that crimes involving persons are also clustered. All this is just a way of exhorting you to be careful about employing this particular measure.

Ripley's K function

Ripley's K is an attempt to estimate the scale at which clustering or dispersion is most concentrated in our data. It represents kind of a cross between point density surface (we start with a point and calculate how many other events are within a given radius of our point) and average nearest neighbor (we take the aforementioned values and average them across all points in our data, comparing the result to that expected under the assumption of complete spatial randomness). It is an excellent technique to use for empirically estimating the extent at which clustering is meaningful in your data, but it suffers from the same weaknesses with respect to borders, irregular study areas, sensitivity to radii used, etc… of these techniques as well.

3. **Spatial Statistics Tools >Analyzing Patterns>Multi-Distance Spatial Cluster Analysis**
    a. **Number of Distance Bands = 10**
    b. **Compute Confidence Intervals = 9 Permutations**

c. **Display Results Graphically**
d. **Boundary Correction Method = Ripley's Edge Correction Formula**

The output from this analysis indicates strong patterns of clustering at all distances. We could refine this by altering the study area to exclude water and even non street locations if we were really trying to be fancy. As it is, this statistic tells us that the greatest difference between point counts and expected point counts occurs at a distance of ~9000 feet, a bit under 2 miles. This is probably driven by the dominant role of the Central Business District in our analysis and its water and hill constrained geography leading to sharp gradients on its borders. Given more time we might take this newly recognized distance and reinsert it in some of our kernel density or point density measures from the previous lab to see if these empirically generated cluster ranges told us something more compelling about where possession and sales of cocaine and marijuana are occurring in Seattle.

Part 3 –I*nterpolation*

Inverse Distance Weighted Interpolation

1. **Start ArcMap, and add the following files to your blank map from "C:\temp\PTSURF\Data\":**
   a. **Rain_Stations (shapefile)**
   b. **Kenya_outside_boundry (shapefile)**
2. **What is the coordinate system of these files? In particular, note the units of measurement.**
3. **Spatial Analyst Tools->Interpolation->IDW**
   a. **Input point features = Rain_stations**
   b. **Z value field = RAIN_MM_**

Limiting the points used for interpolation

The characteristics of the interpolated surface can also be controlled by limiting the input points used in the calculation of each output cell value. You can specify the number of points to use directly, or specify a fixed radius within which points will be included in the interpolation.

Variable search radius

With a variable search radius, the number of points used in calculating the value of the interpolated cell is specified, which makes the radius distance vary for each interpolated cell, depending on how far it has to search around each interpolated cell to reach the specified number of input points. Thus, some neighborhoods will be small and others will be large, depending on the density of the measured points near the interpolated cell. You can also specify a maximum distance (in map units) that the search radius cannot exceed. If the radius for a particular neighborhood reaches the maximum distance before obtaining the specified number of points, the prediction for that location will be performed on the number of measured points within the maximum distance. Generally, you will use smaller neighborhoods or a minimum number of points when the phenomenon has a great amount of variation.

Fixed search radius

A fixed search radius requires a neighborhood distance and a minimum number of points. The distance dictates the radius of the circle of the neighborhood (in map units). The distance of the radius is constant, so for each interpolated cell, the radius of the circle used to find input points is the same. The minimum number of points indicates the minimum number of measured points to use within the

neighborhood. All the measured points that fall within the radius will be used in the calculation of each interpolated cell. When there are fewer measured points in the neighborhood than the specified minimum, the search radius will increase until it can encompass the minimum number of points. The specified fixed search radius will be used for each interpolated cell (cell center) in the study area; thus, if your measured points are not spread out equally (which they rarely are), there are likely to be different numbers of measured points used in the different neighborhoods for the various predictions.

c. Try with both Search radius options: Variable (12 points) and Fixed (25,000 m)

Controlling the influence with the Power parameter

IDW relies mainly on the inverse of the distance raised to a mathematical power. The Power parameter lets you control the significance of known points on the interpolated values based on their distance from the output point. It is a positive, real number, and its default value is 2.

By defining a higher power value, more emphasis can be put on the nearest points. Thus, nearby data will have the most influence, and the surface will have more detail (be less smooth). As the power increases, the interpolated values begin to approach the value of the nearest sample point. Specifying a lower value for power will give more influence to surrounding points that are farther away, resulting in a smoother surface.

Since the IDW formula is not linked to any real physical process, there is no way to determine that a particular power value is too large. As a general guideline, a power of 30 would be considered extremely large and thus of questionable use. Also keep in mind that if the distances or the power value are large, the results may be incorrect.

An optimal value for the power can be considered to be where the minimum mean absolute error is at its lowest. ArcGIS Geostatistical Analyst provides a way to investigate this.

d. Try adjusting the Power parameter from its default to a value of 1.

4. **Spatial Analyst Tools -> Extraction -> Extract by Mask**
   a. **Input raster = Choose from IDW raster**
   b. **Input raster or feature mask data = kenya_outside_boundary**
   c. **Output raster = IDW_Mask.img**

Trend Surface Analysis

1. **Spatial Analyst Tools->Interpolation->Trend**
   a. **Input point features = Rain_stations**
   b. **Z value field = RAIN_MM_**
   c. **Try with 1st, 2nd, and 3rd order polynomials**