GEO 580 Lab 4
Geostatistical Analysis

In this lab, you will move beyond basic spatial analysis (querying, buffering, clipping, joining, etc.) to learn the application of some powerful *statistical* techniques to the problems of spatial patterns and distributions. You will use the ArcGIS *Geostatistical Analyst Extension* to carry out exploratory variogram analysis, and then extend this exploratory approach to spatial interpolation by way of *kriging*. You will then build a GIS model to compare the effectiveness of this *geostatistical* interpolation method to *inverse distance weighting.*

Only a limited subset of the functionality in *Geostatistical Analyst* will be covered, but you are welcome to explore this extension in more detail using the Tutorial in Chapter 2 of ESRI's *Using ArcGIS Geostatistical Analyst* manual (on the Digital Earth server). More technical information can be found in Chapters 4 and 6 of the manual. Another helpful resource is the ArcUser article, **Powerful Spatial Statistics Tools in ArcGIS 9,** available online at**:**
http://www.esri.com/news/arcuser/1104/spatial_statistics.html

Part 1 – Variogram Analysis

1 Data

The data you will use are ozone measurements for July 1996, obtained at monitoring stations in the Los Angeles basin. This data consists of two shape files: a polygon shape file, labasin.shp, with the outline of the four counties (Los Angeles, Orange, Riverside and San Bernadino), and a point shape file, oz96.shp with the ozone measurements at 32 monitoring stations.

2 Exploratory Data Analysis

Start ArcMap with a layer containing the point feature for the monitoring stations and the polygon feature with the county boundaries. Make sure the Geostatistical Analyst is active, Tools > Extensions > Geostatistical Analyst, and that its toolbar is visible, View > Toolbars. Turn on the labels for the county shapefile so they are visible on your map.

2.1 Histogram Analysis

• In the Geostatistical Analyst toolbar, go to Explore Data and select "Histogram" from the drop down menu (don't be fooled by the first graph, by default it is a histogram for the first variable, in this case STATION)
• To get a more meaningful histogram, make sure the *Layer* refers to the point layer (oz96) and change the *Attribute* to MAXDAY.
• Note how the "Statistics" check box is checked by default: this gives a summary of the data in the upper right corner of the graph; if you uncheck the box, the summary will disappear
• You can fine-tune the appearance of the histogram, by changing the number of Bars in the graph, or by choosing a data transformation to check how that affects

Thanks to Professor Luc Anselin, Spatial Analysis Laboratory, University of Illinois, Urbana-Champaign.

the distribution of the variable

- Change the number of bars to 11, and experiment with a log transformation
- The Histogram and map are *linked*. To see this, turn back to the original scale (set Transformation to None), with 8 bars in the histogram, and select (click, followed by shift-click) the two left-most bars. Note where the corresponding points are on the map.
- Using the Select Features tool from the toolbar, select the two northernmost locations in San Bernadino county (stations 5213 and 5181, make sure you have the selectable layer set to oz96).

QUESTION 1A: Where do they rate on the histogram? (Keep in mind how this may affect a semivariogram)

- Check on the distribution of MAXDAY for the four monitoring stations in Orange County.

QUESTION 1B: Where are they located on the histogram?

2.2 Assess the Normality of the Data

- In the Geostatistical Analyst toolbar, select Explore Data and choose Normal QQPlot from the drop down list. The QQPlot assesses the normality of the data, or how evenly distributed the values are respective to the mean value. A histogram of normal data will look like a bell curve, with the maximum frequency at the mean value of the distribution.
- Make sure to reset the Attribute to MAXDAY (and **not** to STATION)
- If the distribution of the data is normal, the plot will follow the diagonal line (this is only based on “visual inspection,” not on a formal test)
- Similar to the histogram, the QQ-plot has the facility to link and find the matching points on the plot for subsets of the data; click on one of the data points (perhaps one that looks like an outlier from the trend) and notice how the corresponding station is selected on the map.

2.3 Practice

Use the Histogram and Normal QQ-Plot tools to assess the distribution of the AV8TOP variable.

QUESTION 2: Is the AV8TOP variable more normally distributed than the MAXDAY? How do the Normal QQ-Plots compare?

3 Modeling Spatial Trends

One of the fundamental assumptions underlying variogram modeling is the absence of a spatial trend (no spatial autocorrelation). In other words, the mean value of the process should be constant throughout space. The Geostatistical Analyst contains functionality to assess the presence of a spatial trend by visual inspection. Note that in this exploratory phase, you cannot save a fitted trend surface (any trend surface needed in the variogram model is fit within the "Geostatistical Wizard").

3.1 Trend Analysis

Select Trend Analysis from the Explore Data drop down list to assess spatial trends in the MAXDAY variable.

- Make sure to set the Attribute correctly
- A three-dimensional graph will appear in the window, with the point locations on the horizontal plane, the values floating in the vertical dimension, spikes connecting the point locations to their values and a projection on the plane to the right and to the back (you may want to expand the dialog box to enlarge the graph)
- Check out the various options by unchecking and checking Sticks (takes out the spikes), Projected Data (takes out the projected data points on the side and back panel), Input Data Points (takes out the floating points in 3-D space) and Trend on Projections (takes out the projected trends in the Y and X plane)
- Note the trend in the pollution measure away from the coast and from North to South
- You can rotate the view, but it takes some practice to realize what is being rotated. It can be either the points (Locations) or the graph itself; the latter gives you a better idea of how the projected lines reflect a trend in the data
- To rotate the graph, select Graph from the drop down list and click on and drag the arrows directly to the right of that item. Isn't that neat?
- As with the Histogram and QQPlot, you can select items on the map and see their "spikes" highlighted in the graph (note that the points are not highlighted in the projected trend surface)

QUESTION 3: Describe the spatial trend in the AV8TOP variable compared to the MAXDAY variable.

4 Exploring the Semivariogram

You can continue exploring the data by taking a closer look at the variogram structure. This is implemented in the Semivariogram/Covariance Cloud item in the Explore Data menu. All the analyses will be for the MAXDAY variable. Make sure to set Attribute to that variable, since the default is to use STATION, which is rather meaningless.

4.1 Creating a Semivariogram

- After invoking Semivariogram/Covariance Cloud, keep the Tab on *Semivariogram*
- The scatterplot-like graph is the *semivariogram cloud plot*, it shows for each distance *bin* (horizontal axis) the squared difference between each *pair* of observations on the vertical axis. This is how different in value each pair of stations is. We are asking, "are two stations at x distance (lag) apart quite different in their daily ozone maximum, or are they similar?"
- Change the Lag Size to 0.2 and the Number of Lags to 20. You will see that the cloud plot does not contain any values beyond a distance of about 2.4 units on the x-axis. This means that the maximum distance between any two stations is 2.4 units.
- A good rule of thumb is to constrain the variogram to 1/2 the maximum distance (2.4 units) and to make sure there are at least 30 pairs are in each bin
- So, change the Number of Lags to 10 and the Lag Size to 0.12 and note how the cloud plot is truncated
- Experiment with some other combinations of Lag Size and Number of Lags combinations to get a good sense for what is going on

4.2 Exploratory Variography – Linking and Selection

- Use the semivariogram settings with Number of Lags to 10 and the Lag Size to 0.12 and select the three points that look like outliers in the upper left corner of the semivariogram (NOTE: it is easiest to click on the graph and drag the box over the three points, so all three are selected at once)
- locate the corresponding *pairs* in the map and note how all three have a common *origin* (station 5213) (remember what you saw in the histogram? remember those extreme values?)
- In variogram terms, this means that the squared difference of the MAXDAY variable observed for pairs this distance apart is much *higher* than for the others, suggesting that *station 5213 may be a true outlier*, or, that it may represent a *pocket of local nonstationarity*
- Select the two top most points in the semi-variogram and assess their relative location; how would you interpret this finding? (remember what high semivariance means)
- Check the semivariogram for other possible outliers

4.3 Directional Effects

You can check the semi-variogram for directional effects using the Semivariogram Surface. The surface is shown in the lower left corner of the window and contains the average squared difference for all pairs that match a given distance/direction combination. These values are contained in square grid cells. *Asymmetries in the surface suggest directional effects* (i.e. values are more different or more similar in a particular direction, like north, or south)

- Activate the directional analysis by checking the Show Search Direction check box
- Note the angle and parallel lines on the surface plot; these can be manipulated with the cursor to select a directional range (Hey, check this out. If you point in the direction of the BLUE colors, where values are pretty similar from the center out, the variogram values are low, right, because the difference between them is low. Now, point in the direction where the colors are RED, where values are pretty different compared to the blue center. See the variogram values jump up? There is more difference in value between any two points going in this direction…this is the value/take home message of a semivariogram)
- Note how the number of points in the semivariogram is a lot less than in the original plot (for all the data pairs): only those pairs matching the directional criterion are kept
- Select some of the points in the semivariogram and check on the map how the lines connecting the pairs are all in the same general angle
- Move the direction selection tool, reselect points in the semivariogram (with high value) and check how the pairs correspond to a different angle
- Experiment with some different angles.

QUESTION 4: In what cardinal direction is the directional effect evident?

4.4 Practice

Carry out the exploratory variography for the AV8TOP variable and compare the outliers and other distinct patterns (directional effect) to those found for MAXDAY.

Part 2 – Kriging

1 Basics of Spatial Prediction

Kriging, or spatial interpolation, is based on the spatial information in the semivariogram and is implemented in a *Geostatistical Analyst.* This guides you step-by-step through the process. You can leave all options to their default settings and the Geostatistical Analyst will produce a predicted surface map. However, you should exercise some judgment in the choice of variogram function, data transformations, directional effects, etc.

Start ArcMap with a layer containing the point feature for the monitoring stations and the polygon feature with the county boundaries (just like in part one). Make sure the Geostatistical Analyst is active: go to Tools > Extensions > Geostatistical Analyst, and see that its toolbar is visible, under View > Toolbars.

In the Geostatistical Analyst toolbar, select Geostatistical Wizard and make sure the Input Data (Dataset 1) is set to the point layer (oz96) and the Attribute to

MAXDAY. Set the Method as *Kriging*. Start the Wizard by clicking on Next.

1.1 Geostatistical Wizard

• **Step 1** in the interactive process is the choice of Method; for now, leave the default to Ordinary Kriging, Prediction Map, without transformations or trend removal (Ordinary Kriging will take out an estimate for the mean value), click on Next to move on

• **Step 2** is Semivariogram modeling, based on the same two graphs as in the exploratory part, a semivariogram cloud plot and a surface.
• Note how a variogram function has been fit through the cloud plot (the blue line) with as default the Spherical model
• Note the values for the parameters (Major Range, Partial Sill, and Nugget) and the choices for the Lag Size and Number of Lags (later, you will change these), move on by clicking on Next

• **Step 3** specifies the Searching Neighborhood; ignore this for now and click on Next

• **Step 4** provides some indication of model fit by systematically dropping an observation from the data set, and refitting the model, or *cross-validation*
• Make a note of some of the measures of fit (mean, root-mean-square, which is comparable to a goodness-of-fit) for future comparison
• click Finish and a summary of the model parameters will be presented in a window
Complete the process by clicking on OK in the last dialog to obtain a predicted surface in ArcMap, covering a bounding rectangle around the sample points. Rearrange the layers so that you can see the original monitor locations on top of the predicted surface.

1.2 Practice

Go through the steps in the Geostatistical Wizard to create a predicted surface for AV8TOP, using the default settings.

2 Customizing Spatial Prediction

There are several ways in which you can customize the presentation of the predicted surface and the application of various geostatistical methods to obtain the interpolation. Refer back to ESRI’s *Using ArcGIS Geostatistical Analyst* manual for specifics.

2.1 Fine Tuning Presentation

There are other ways in which you can customize the presentation of the predicted surface. For example, you can change the extent of the rectangle containing the predicted surface:
• In the Ordinary Kriging layer for MAXDAY (i.e., the one containing the predicted surface),

right click and select Properties; select the Extent tab
- Check out the options in the "*Set the extent to"* drop down list. For example, choose the rectangular extent of labasin (the LA county boundary layer); click on OK to see the effect
- Note how bad the predictions get as you move out of the core window used for estimation!

You can also *clip* the rectangle with the predicted surface such that it matches the irregular outline of another shape. For example, you may want to clip it to match the boundaries of one of the LA Basin counties. First you need to create a separate layer (shapefile) that contains only the selected county:
- Start by selecting Los Angeles county, using the Select Features item from the toolbar and click anywhere in the county to select it
- Right click on the labasin layer item in the legend and select Data > Export Data from the menu
- Leave the option to Export selected features and enter an appropriate file name for the shape file to be created (e.g., LAcounty); click on OK to save the new shape file
- Reply "Yes" when asked to add the exported data to the map as a layer

To limit the interpolated surface map to the extent of Los Angeles county, you need to change the Properties of the Layers:
- Right click on "Layers" (in bold at the top), select Properties and click on the Data Frame tab
- In the Data Frame Properties dialog, focus on the Clip to Shape item and check the Enable box; specify the LAcounty layer as the shape to use as a clip
- Click OK; now the predicted map coincides with the outline of the single county

The interpolated surface can be portrayed in a number of different ways. This is set in the Symbology Property of the layer. The default is Filled Contours, but you can also choose Hillshade, Contours, or Grid:
- Right click on the Ordinary Kriging layer for MAXDAY and select Properties
- Select the Symbology tab
- Uncheck Filled Contours and check Contours instead; click OK to see the effect
- Experiment with the other types of symbology for the predicted surface

The predicted surface can be turned into a permanent feature class or shape file by means of the Data Export function. This will allow you to then use the predicted surface in other analyses. For example, an exported grid surface (or raster) can be used to compare the predicted values to those obtained with another technique using the ArcGIS Spatial Analyst Extension raster calculator:

- Right click on the Ordinary Kriging layer; in the Data item, select Export to Raster
- Choose a file name for the output file (try LAkrig) and click OK; add the raster to

the layers if you wish
- Alternatively, you could use the Data > Export to Vector function to create a shapefile with the interpolated contours

2.2 Fine Tuning Analysis

So far, the predicted surface was based on using the defaults in the Wizard. However, this is by no means necessarily the best way to proceed. It is useful to reanalyze the spatial correlation structure in the MAXDAY variable using different variogram models and/or by including trend removal. Also, the fitted model is sensitive to the choice of the *number of bins* and their *range.*

Create some new predicted surfaces by using different settings:

- Start the Geostatistical Wizard, and make sure the Input Data are set to the point layer (oz96), and the Attribute to MAXDAY
- Keep the Methods setting to Kriging and click on Next
- In the following dialog, set the Order of Trend Removal to *second* and click Next to continue
- The window will show the predicted second order trend surface; click Next again
- In the semivariogram dialog, set the Lag Size to 0.1 and the Number of Lags to 12, and observe the change in the parameter estimates (you should get a mean of 0.1048, rms of 3.068, and SE of 2.353)
- Now, run the analysis for an *exponential* variogram

2.3 Practice

Experiment with model selection and changing lag size for the AV8TOP variable. Present the results in different ways, as filled contours, contour lines, and raster surfaces. Try using various clips to portray the predicted value surface for administrative units.

3 Assessing and Comparing Model Fit

In order to assess how good the predicted surface is, its fit needs to be quantified. There are basically two methods to accomplish this. In one, an observation point is left out and then predicted using the surface fit for the remaining observations. This is done for each observation in turn. The overall fit can then be summarized as a *mean squared error* or other indicator. This method is referred to as *cross-validation*. The other approach extends this idea to a specific subset of the data. For example, several observation points are left out of the analysis (the model fitting) and the estimated surface is used to predict their values. Those predictions are then compared to the actual values.

3.1 Cross-Validation

The Geostatistical Wizard generates the cross-validation statistics:

- make sure you have at least two predicted surfaces for MAXDAY in your table of contents, such as the ordinary kriging one obtained previously (no trend removal, spherical model), and ordinary kriging using an exponential model with second order trend removed

- Right click on one of the predicted surfaces in the table of contents and select Compare...
- A dialog will open with the cross-validation statistics for the current surface on the left hand side, below Compare
- On the right you can select the surface to compare it to in the To: drop down list (select your other predicted surface)
- When comparing two predicted surfaces in this way, focus on the mean prediction error (should be close to 0), compare the root mean square (smaller is better) and check the root-mean-square standardized (closest to 1 is best)

QUESTION 5: Describe which two surfaces you created, and compare them by describing the mean prediction error, root mean squares, and root-mean-square standardized.

To illustrate the second method, first create two new shapefiles from the ozone station point layer by using the selection tool and Data > Export Data. One file should contain only the four monitors for Orange County, the other should contain all monitors *except* those in Orange County. (you should un-do the enabled clip option from before if you haven't already…click on layers, properties, data frame and un-click it to get the whole extent back). (remember, to select more than one station, hold down the shift key and click on more than one!)

Now proceed with creating an interpolated surface based on the incomplete data set (all points *except* those in Orange County). (Just do the ordinary kriging with no trend removed, spherical model, and make sure you're attribute is set to MAXDAY and the layer to the incomplete data one described above)

To compare the results:

- In the layer for the predicted surface, right-click and select Validation
- In the dialog, select the layer with the points for Orange county as the Input Data
- Make sure the attribute is set to MAXDAY and specify a shape file to save the output (you can name it "NoOC_validation", if you fancy)
- The new layer will be created with the original attributes for the point layer (the orange county only info) as well as new variables, including Measured, Predicted and Error (you can see this in the attribute table)
- Use the Statistics on the Error field to summarize the prediction error (right click on Error and select Statistics, **not** Summarize)

QUESTION 6: What is the mean and range (min to max) of the error using this validation method?

Part 3 – Spatial Interpolation Comparison

1 Data

Use the sample data provided in the file oz9799.shp in the lab folder. This is a point shape file that contains two sets of monthly ozone measures (maximum and average over the highest 8 hours) for monitoring stations in the Los Angeles basin during a three-year period, from 1997-1999. So, you'll see columns entitled, M971-M9912 (max reading, Jan 1997- Dec 1999) and A971-A9912 (average reading, Jan 1997-Dec 1999).

You are allowed (encouraged) to work in teams of up to three members. All members of the team will receive the score for this portion of the lab. There are plenty of data for **each group to choose a different month**. Make sure to pick a month/variable for which there are no missing values (negative values). This means that June-Nov of 1997 is out…choose another month. Try something in 1998 why don't ya, there as don't seem to be any negatives there. You can choose to analyze the max value, or the average for that particular month.

1. Explore the variogram cloud plot, identify (suggestive) spatial outliers.

2. Choose two different variogram models (i.e. spherical, exponential, Gaussian) to compare them.

3. Construct a predicted surface using both models.

4. Assess the fit by cross-validation, and choose the best model. In your report, list the prediction errors and make a case for why you choose which model. (note: it is also helpful to label oz9799 with the max or average value of your month, so you can see how the prediction surfaces are interpolating the real data points)
5. Now cross-validate your selected model using a subset of the data. Compare the predicted surface to one with 3 points left out (see page 8 of this lab, you just did this!) Make sure to report the mean and range of the error.
6. Perform an inverse-distance-weighted derived surface on your month. Report the mean prediction error, and the root-mean-square error.

Prepare a 2 to 3-page, 1.5 spaced report that compares the effectiveness of inverse distance weighting to geostatistical methods in creating a believable spatially-interpolated surface from sample point data. Keep the discussion brief and to the point, but please insert any relevant maps and figures (no page length limitation for maps or figures). Make your decision based on comparison of error, and think about how valid the spatially interpolated surface is when cross validated on itself. Also, visually

compare them…what do you see? How does the surface match with the real values? Which method best captures the outliers?