Overview

In this lab you will think critically about the functionality of spatial interpolation, improve your kriging skills, and learn how to use several geostatistical tools in ArcMap. The following exercises guide you through a rainfall mapping procedure for Guatemala, a country dependent on rainfed agriculture, where one-out-of-three people is malnourished, and child stunting is very prevalent. Rainfall mapping is one method of identifying food insecurity at the sub national level, a critically important tool for distribution of food aid. Given station precipitation data for the Guatemala agricultural season during El Nino years, we identify regions of Guatemala where drought has caused crop failures. This gives us a good idea the locations we should be concerned about as we approach the next El Nino year. We examine two types of kriging. First, we use kriging to interpolate the ‘raw’ station rainfall data throughout the country. Second, we interpolate the station data using a high resolution background mean field. As we move through the exercises we find that the incorporation of gridded climatological data, and the use of station anomalies instead of measurements, improves results. Finally, we discuss what the rainfall maps might mean for maize production in typical El Nino years.

Materials

<table>
<thead>
<tr>
<th>Data name</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guatemala.shp</td>
<td>Shapefile</td>
<td>Guatemala boundary</td>
</tr>
<tr>
<td>Guatemala_elnino.shp</td>
<td>Shapefile</td>
<td>Guatemala met stations. Average total precipitation (mm) of the major agricultural season (June, July, August) in El Nino years</td>
</tr>
<tr>
<td>fclim_jja</td>
<td>Raster</td>
<td>Gridded climatological mean total precipitation (mm) for June, July, August</td>
</tr>
</tbody>
</table>

Prerequisites

1. Create a folder and download to it the contents of the table above. Open ArcMap. Before you begin, you must enable the Geostatistical Analyst and Spatial Analyst extensions. In ArcMap, click Tools, click Extensions, then check Geostatistical Analyst and Spatial Analyst. Click Close. The click View, point to Toolbars, then check Geostatistical Analyst and Spatial Analyst.
2. Add the two shapefiles to your map via File…Add Data and navigating to the folder you created. Throughout this lab feel free to explore ways of changing the map display in ArcMap. You can move layers into and out of sight and change their visual properties using the Layers menu. Save your map File…Save, naming it “GuatemalaAid_ElNino.mxd.”

**Explore station data**

**3. Data details.** Notice that the points, the met stations, extend outside Guatemala’s national boundary. These stations are necessary to create a kriged surface that is accurate close to the country edges. Below “Layers,” you’ll see the list of map contents. Right-click upon the station data layer and choose Open Attribute Table to view the station data. Each row is a station. The column (field) “Mean_jja” is the average rainfall total for the June, July, August season during El Nino years (1972, 1982, 1987, 1991, 1997, and 2002). This is the field we will be working with. Close table.

**4. Visualize data.** The interpolation methods used to generate a surface give the best results if the data is normally distributed (a bell-shaped curve). If your data is skewed (lopsided), you might choose to transform the data to make it normal. Thus, it is important to understand the distribution of your data before creating a surface.

Using three useful visualization techniques, let’s explore the station data.

4.A. Examining the data distribution

To start, click the Geostatistical Analyst drop-down arrow, point to Explore Data, then click Histogram. Click the Layer drop-down arrow and choose “Guatemala_elnino.” Click the Attribute drop-down arrow and choose “Mean_jja.” You may want to resize the Histogram dialog box so you can also see the map. The distribution of the Mean_jja attribute is depicted by a histogram with the range of values separated into 10 classes. The frequency of data within each class is represented by the height of each bar.

Experiment with the number of classes. Note that by clicking on a bar the stations grouped into that range of values are highlighted upon the map.

Generally, the important features of the distribution are its central value, spread, and symmetry. As a quick check, if the mean and the median are approximately the same value, you have one piece of evidence that the data may be normally distributed.

4.A.1. *What are the units of the x-axis?*

4.A.2. *Using the histogram shape and a statistical measure of the precipitation data to justify, is the station data normally distributed?*

Next, visualize the data with the Normal QQplot. Take care to change the attribute to Mean_jja.

4.A.3. *Briefly explain what the Normal QQplot shows, in general terms and in relation to the station data.*
4.B. Examining spatial trends

If a trend exists in your data, it is the nonrandom (deterministic) component of a surface that can be represented by a mathematical formula. For instance, a gently sloping hillside can be represented by a plane. A valley would be represented by a more complex formula (a second-order polynomial) that creates a U shape. This formula may produce the representation of the surface you want. However, many times the formula is too smooth to accurately depict the surface because no hillside is a perfect plane nor is a valley a perfect U shape. If the trend surface does not adequately portray the surface well enough for your particular need, you may want to remove it and continue with your analysis, modeling the residuals, which is what remains after the trend is removed. When modeling the residuals, you will be analyzing the short-range variation in the surface. This is the part that isn't captured by the perfect plane or the perfect U shape.

The Trend Analysis tool enables you to identify the presence or absence of trends in the input dataset.

To do so, click the Geostatistical Analyst drop-down arrow, point to Explore Data, then click Trend Analysis. Take care to change the attribute to Mean_jja.

Each vertical stick in the trend analysis plot represents the location and value (height) of each data point. The points are projected onto the perpendicular planes, an east-west and a north-south plane. A best-fit line (a polynomial) is drawn through the projected points, which model trends in specific directions. If the line were flat, this would indicate that there would be no trend.

Explore the 3D plot by rotating the graph and/or the locations.

4.B.1. Describe the spatial patterns of the station precipitation data (trends in the east-west and north-south directions). What could be a driver of the east-west trend?

4.C. Examining the semivariogram

Now we’ll examine the semivariogram cloud for evidence of autocorrelation and directional trends.

The semivariogram/covariance cloud allows you to examine the spatial autocorrelation between the measured sample points. In spatial autocorrelation, it is assumed that things that are close to one another are more alike. The semivariogram/covariance cloud lets you examine this relationship. To do so, a semivariogram value, which is the difference squared between the values of each pair of locations, is plotted on the y-axis relative to the distance separating each pair on the x-axis.

To view the semivariogram cloud, click the Geostatistical Analyst drop-down arrow, point to Explore Data, then click Semivariogram/Covariance Cloud. Take care to change the attribute to Mean_jja.

Each red dot in the semivariogram/covariance cloud represents a pair of locations. Since locations that are close to each other should be more alike, in the semivariogram the close locations (far left on the x-axis) should have small semivariogram values. As the distance between the pairs of locations increases the semivariogram values should also increase. However, a certain distance is reached where the cloud flattens out, indicating that the relationship between the pairs of locations beyond this distance is no longer correlated.
Looking at the semivariogram, if it appears that some data locations that are close together (near zero on the x-axis) have a higher semivariogram value (high on the y-axis) than you would expect, you would investigate these pairs of locations to see if there is a possibility that the data is inaccurate.

Try dragging the Selection pointer over points in the semivariogram. This highlights pairs of sample locations on the map, and shows the lines linking the locations.

**4.C.1 What are the units of the x-axis?**

Also note that you can change the lag size. The lag size is the size of a distance class into which pairs of locations are grouped to reduce the large number of possible combinations. This is binning.

Change it to 0.3 and, using the mouse, select all the plotted points. Look at the map. Repeat these steps with a lag size of 0.1, then 0.05.

**4.C.2. How does changing the lag size affect the construction of the semivariogram cloud? Hint: Pay attention to the x-axis values.**

For your information:

Besides global trends that were discussed in the previous section, there may also be directional influences affecting the data. The reasons for these directional influences may not be known, but they can be statistically quantified. These directional influences will affect the accuracy of the surface you create in the next exercise. However, once you know if one exists, Geostatistical Analyst provides tools to account for it in the surface-creation process. To explore for a directional influence in the semivariogram cloud, you can use the Search Direction tools.

Check Show search direction, then click and drag the directional pointer to any angle.

The direction the pointer is facing determines which pairs of data locations are plotted on the semivariogram. For example, if the pointer is facing an east-west direction, only the pairs of data locations that are east or west of one another will be plotted on the semivariogram. This enables you to eliminate pairs you are not interested in and to explore the directional influences on the data.

**5. Calculate average station distance.** Close the semivariogram tool. In the main menu area of ArcMap, click Window, then ArcToolbox. In the ArcToolbox window, navigate to Spatial Statistics Tools, then to Analyzing Patterns, then to Average Nearest Neighbor to calculate average station distance. Use Euclidean Distance and do NOT display output graphically. When the tool has finished scroll up in its box to see the results.

**5.1 What is the observed mean distance between stations?**

**Kriging with Geostatistical Analyst**

6. Krige the station data using default tool parameters.
Under Geostatistical Analyst, click Geostatistical Wizard. Select kriging as the method of spatial interpolation and take care to choose the appropriate layer and attribute as the input data. Click Next.

To get acquainted with the tool we'll use the default parameters, so leave the parameters selected as Ordinary Kriging and Prediction Map. Do not transform the data or remove any trends. Click Next.

Now we need to decide which model we want to use to predict values at unsampled points. A model that fits the empirical semivariogram will result in a better spatial interpolation. Experiment with different model choices.

6.1. What are two things about the semivariogram that change when you change the model?

For this analysis we will use the exponential model, the average station distance we calculated in 5.1 as our lag size, and set the number of lags to 10.

6.2. When we use these parameters and change the lag size from the default of ~0.086 to the average station distance, what happens to the nugget? What assumption does this imply?

Click Next. Now we can visualize how predictions will be made at unsampled locations using the Search Neighborhood Dialog Box. The blue points are the meteorological stations. Experiment by moving the crosshairs, paying attention to how the weights given to neighboring stations change. Expand the Show Weights tab.

6.3. Manually type in this coordinate: x= -90, y=15 to view how a prediction at that location is made. How many neighbors are considered? How many of them are given a weight greater than 10%? Which station (number) is assigned the largest weight, and what is this weight?

6.4. When you move the cross hairs to a location with no stations within the search radius, how does the distribution of station weights (with distance) change?

Click Next. Now we can use cross-validation to get an idea how well the model is predicting unknown values. For all points, cross-validation sequentially omits a point, predicts its value using the rest of the data, and compares the measured and predicted values. The calculated statistics serve as diagnostics that indicate whether the model is reasonable for map production.

In addition to visualizing the scatter of points around this 1:1 line, a number of statistical measures can be used to assess the model's performance. The objective of cross-validation is to help you make an informed decision about which model provides the most accurate predictions.

In addition to visualizing the scatter of points around this 1:1 line, a number of statistical measures can be used to assess the model's performance. The objective of cross-validation is to help you make an informed decision about which model provides the most accurate predictions.

Here the term prediction error is used for the difference between the prediction and the actual measured value. For a model that provides accurate predictions, the mean prediction error should be close to 0 if the predictions are unbiased and the root-mean-square prediction error should be small. The average standard error should also be as small as possible (this is useful when comparing models) and the root-mean-square standardized prediction error should be close to 1 if the standard errors are accurate.
The Cross Validation dialog box also allows you to display scatterplots that show the error, standardized error, and QQ plot for each data point.

6.5. What is the mean prediction error of this model (include units in your answer)? What is the largest prediction error of all the stations? Is it an over or underestimation?

Click Finish, then OK to create the kriged surface. Notice that it now appears in the Layers menu and has been automatically named “Ordinary Kriging...Prediction Map.” Rename it by right-clicking, Properties...General. Rename it “Kriging with Measurements.”

**Kriging with station anomalies**

7. Using anomalies. Recall that spatial interpolation works best when the data is normally distributed. Because the station data is not normally distributed, in order to make a better rainfall surface we need to transform it.

We can do so by converting the (El Nino) station rainfall measurements to anomalies, or residuals. An anomaly is the difference from the mean value. In this case, anomalies will tell us what locations tend to be wetter or drier than normal during El Nino, and by how much (in mm). As opposed to one-time rainfall measurements, a dataset of anomalies has the advantage of central tendency. Anomalies are much more likely to be normally distributed, hence are better for kriging. Once we have made (via spatial interpolation) a surface of rainfall anomalies we can add it to a surface of climatological means to create a better rainfall estimation surface for El Nino years.

To calculate El Nino rainfall anomalies we need to know the climatological mean at each station. For this we will use the gridded climatological dataset “fclim_jja.” In this case, as in most others, the climatological mean data is composed with a higher station density than a single year of measurements and has been quality checked to a much greater extent.

Add this raster layer to the map.

To give the station dataset information about the climatological mean season (Mean_jja) rainfall totals we need to extract the raster information to the station location. In ArcToolbox...Spatial Analyst Tools...Extraction... select Extract Values to Points. Choose from in the appropriate dropdown boxes the station and raster data. Name your output shapefile “elnino_with_fclim.shp.” When the operation is complete this shapefile will be added to the Layers menu.

Open the attribute table of “elnino_with_fclim.shp.” The climatological mean for each station is in the field RASTERVALU.

Now compute the El Nino anomaly for each station. To do so, we first need to create a field in the attribute table for the anomaly values to be held. At the bottom of the attribute table, select Options...Add Field.
Here we need to tell ArcMap characteristics about the data (that we are about to create) so that it knows how it should be stored and how to create the new field. While not of major importance in this case, data type and storage decisions are vital when dealing with large spatial datasets. To minimize data processing time and memory needed to hold it, we want to store data values in a way that minimizes storage needs while still maintaining necessary data characteristics.

Data saved as the type “short integer” only takes up 2 bytes and can hold whole integer values between -32,768 and 32,768. Because our anomaly values will be much smaller than these bounds, short integer is a good choice. If the data were to have a decimal place, we would use the data type “float” (which takes up more storage space) or scale the data by a constant so that the decimal place is no longer needed.

Precision is the maximal number of digits you will need to store in each row of the field. Choose an appropriate value for precision, name the field “Anom” and click ok.

7.1 What value did you choose for precision, and how did you come to this conclusion?

To calculate anomalies and put them in this new field, right-click on “Anom” and choose Field Calculator. Say yes. To input the calculation for “Anom,” double-click on “Mean_jja” then single-click on the subtract button, then double-click on “RASTERVALU,” then click ok.

Visualize the distribution of the El Nino anomalies using the same methods as before.

7.2 Are anomalies more normally distributed than the measurements? What is the mean for the anomalies? Why is it not zero?

Now let’s krig with the anomalies. Using the same procedure as section 6, we’ll create a kriged surface with the anomalies.

First, use the same lag size as before, a lag number of 10, and the exponential model.

7.3 Explain the structure of the semivariogram.

Now, use a lag size of 0.31 and change the number of lags to 9.

7.4 What has changed?

Click Finish. Rename your surface “Kriging with Anomalies.”

8. Comparison of outputs. In the Layers menu right-click upon the “Kriging with Anomalies” layer, then select Compare. Compare to the “Kriging with Measurements” layer with cross validation.

8.1 Using the root-mean-square standardized error and one of the available plots to justify, which kriged surface is better? Why can’t we directly compare the mean error?
Create a precipitation surface for typical El Nino years

9. Combine the anomaly and climatological surfaces. Now we will add the kriged anomaly surface to the climatological mean precipitation surface.

To do so, both surfaces must be of the same data type. So first, right click on the “Kriging with Anomalies” layer and select Export to Raster. Save the output to your folder and name it “elnino_anom.” Click OK, then yes to add this layer to your map.

Under the Spatial Analyst Tool (on the ArcMap Toolbar), select Raster Calculator. By double clicking, build the equation “elnino_anom” + “fclim_jja” then click Evaluate. The result will appear in the Layer menu with the name “Calculation.” This is your final precipitation surface for typical El Nino years. Rename this layer “elnino_precip.”

What does it mean for maize production?

10. The agricultural context. Maize, the dominant crop of rainfed agricultural systems in Guatemala, needs at minimum 500 mm of total seasonal rainfall to produce satisfactory yields. To identify which regions of Guatemala typically do not receive satisfactory rainfall for maize production during El Nino, we will apply a threshold to the final precipitation surface.

Again under Spatial Analyst, select Raster Calculator. This time, build the equation “elnino_precip” < 500 and click Evaluate. The result will appear in the Layers menu as “Calculation.”

Pixels with the value of 1 identify the areas which historically have tended to receive less than 500 mm of seasonal rainfall during El Nino.

10.1. Assuming that maize is planted is ubiquitously in Guatemala, according to this threshold analysis what region of the country has historically experienced drought-inducing crop failure during El Nino?

Major problems can arise if food aid is sent to a region that does not need it. To get an idea about accuracy of our regional drought identification, we need to consider uncertainty of our interpolated precipitation surface.

10.2. What are some sources of uncertainty in this analysis?

One way of identifying uncertainty is to display a map of prediction standard errors. To do this, in the Layer menu, right-click upon the “Kriging with Anomalies” layer and select Create Prediction Standard Error Map. The result will appear as “Kriging with Anomalies_2.”

10.3. What is the main driver of spatial variation in standard error predictions? In which parts of the country can we be the most and least certain about our rainfall estimates?
We can use the standard error predictions to create confidence bounds around our rainfall estimates. This is especially important for the region we have identified as prone to serious drought. Assuming that the standard errors are approximately normally distributed, we can say with p% confidence that a location typically receives “x” amount of rainfall +/- (z-score_{p%} * standard error).

For example, assuming that the standard errors are approximately normally distributed, we can say with 95% confidence that a location receives between x - (1.96 * standard error) and x + (1.96 * standard error) of rainfall during typical El Nino years.

Using Raster Calculator, create a gridded surface for the lower confidence bound of rainfall estimates at the 80% probability level. (The z-score at the 80% probability level is 1.282). Make one for the upper confidence bound as well. Rename these “lower bound estimate” and “upper bound estimate,” respectively.

Using the same thresholding technique as before, identify the areas that we are 80% confident receive less than the required amount of rainfall to produce satisfactory yields during El Nino years.

10.4. Did you use the estimated rainfall surface for the upper or lower confidence bound? Why?

10.5. How has your identified area of drought concern changed?

10.6. Imagine we are approaching the Guatemala cropping season during an El Nino year. Explain to a policymaker now the likely post-season food aid needs of the country. Justify your claims with what you have learned throughout this analysis, including the level of certainty you have and why.