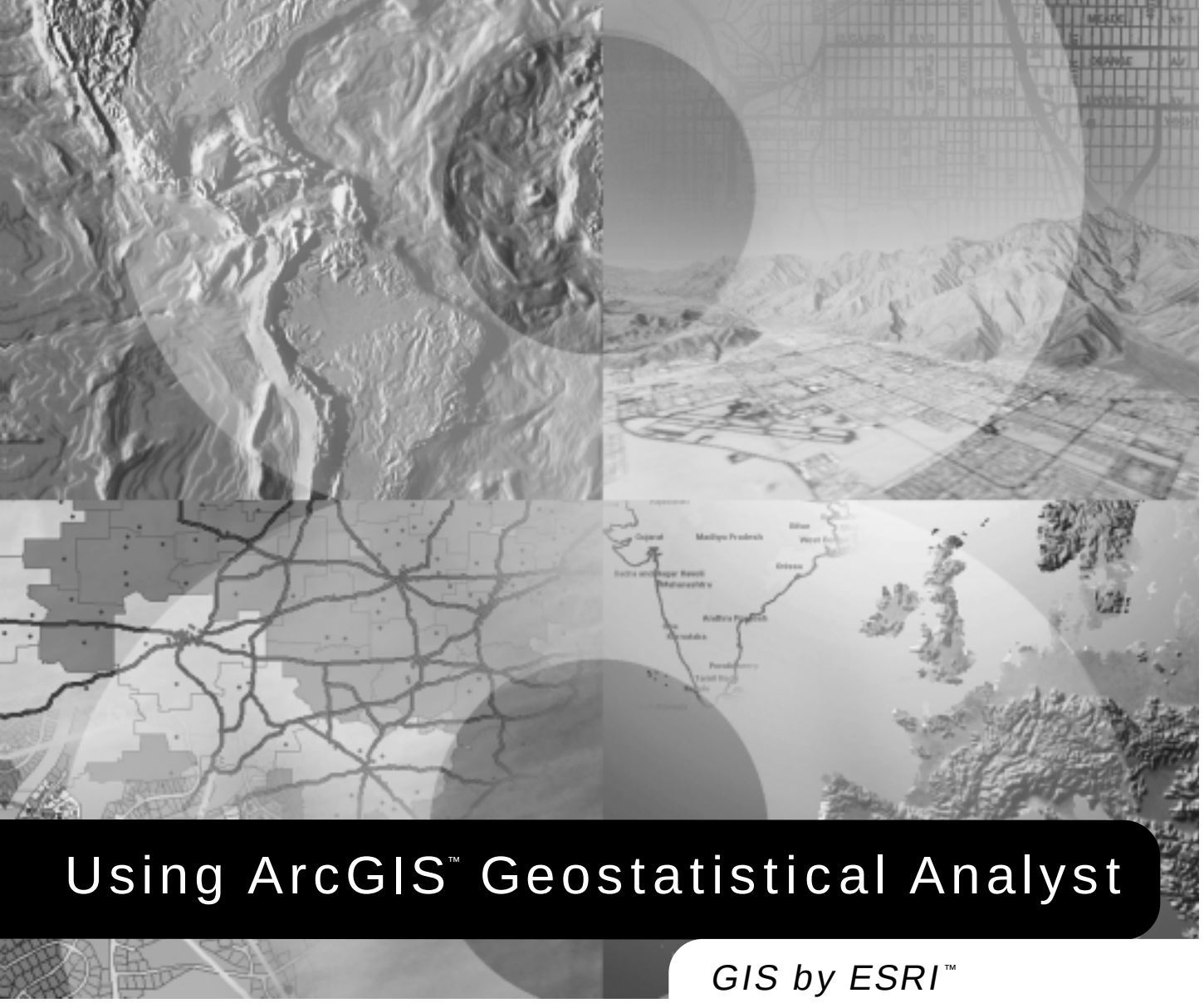


Using ArcGIS™ Geostatistical Analyst

GIS by ESRI™

DATA CREDITS

Carpathian Mountains data supplied by USDA Forest Service, Riverside, California, and is used here with permission.

Radioceasium data supplied by International Sakharov Environmental University, Minsk, Belarus, and is used here with permission. Copyright © 1996.

Air quality data for California supplied by California Environmental Protection Agency, Air Resource Board, and is used here with permission. Copyright © 1997.

Radioceasium contamination in forest berries data supplied by the Institute of Radiation Safety "BELRAD", Minsk, Belarus, and is used here with permission. Copyright © 1996.

CONTRIBUTING WRITERS

Kevin Johnston, Jay M. Ver Hoef, Konstantin Krivoruchko, and Neil Lucas

DATA DISCLAIMER

THE DATA VENDOR(S) INCLUDED IN THIS WORK IS AN INDEPENDENT COMPANY AND, AS SUCH, ESRI MAKES NO GUARANTEES AS TO THE QUALITY, COMPLETENESS, AND/OR ACCURACY OF THE DATA. EVERY EFFORT HAS BEEN MADE TO ENSURE THE ACCURACY OF THE DATA INCLUDED IN THIS WORK, BUT THE INFORMATION IS DYNAMIC IN NATURE AND IS SUBJECT TO CHANGE WITHOUT NOTICE. ESRI AND THE DATA VENDOR(S) ARE NOT INVITING RELIANCE ON THE DATA, AND ONE SHOULD ALWAYS VERIFY ACTUAL DATA AND INFORMATION. ESRI DISCLAIMS ALL OTHER WARRANTIES OR REPRESENTATIONS, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. ESRI AND THE DATA VENDOR(S) SHALL ASSUME NO LIABILITY FOR INDIRECT, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES, EVEN IF ADVISED OF THE POSSIBILITY THEREOF.

U. S. GOVERNMENT RESTRICTED/LIMITED RIGHTS

Any software, documentation, and/or data delivered hereunder is subject to the terms of the License Agreement. In no event shall the U.S. Government acquire greater than RESTRICTED/LIMITED RIGHTS. At a minimum, use, duplication, or disclosure by the U.S. Government is subject to restrictions as set forth in FAR §52.227-14 Alternates I, II, and III (JUN 1987); FAR §52.227-19 (JUN 1987) and/or FAR §12.211/12.212 (Commercial Technical Data/Computer Software); and DFARS §252.227-7015 (NOV 1995) (Technical Data) and/or DFARS §227.7202 (Computer Software), as applicable. Contractor/Manufacturer is ESRI, 380 New York Street, Redlands, CA 92373-8100, USA.

Quick-start tutorial

2

IN THIS CHAPTER

- **Exercise 1: Creating a surface using default parameters**
- **Exercise 2: Exploring your data**
- **Exercise 3: Mapping ozone concentration**
- **Exercise 4: Comparing models**
- **Exercise 5: Mapping the probability of ozone exceeding a critical threshold**
- **Exercise 6: Producing the final map**

With the Geostatistical Analyst, you can easily create a continuous surface, or map, from measured sample points stored in a point-feature layer, raster layer, or by using polygon centroids. The sample points may be measurements such as elevation, depth to the water table, or levels of pollution, as is the case in this tutorial. When used in conjunction with ArcMap, the Geostatistical Analyst provides a comprehensive set of tools for creating surfaces that can be used to visualize, analyze, and understand spatial phenomena.

Tutorial scenario

The U.S. Environmental Protection Agency is responsible for monitoring atmospheric ozone concentration in California. Ozone concentration is measured at monitoring stations throughout the State. The locations of the stations are shown here. The concentration levels of ozone are known for all of the stations, but we are also interested in knowing the level for every location in California. However, due to cost and practicality, monitoring stations cannot be everywhere. The Geostatistical Analyst provides tools that make the best predictions possible by examining the relationships between all of the sample points and producing a continuous surface of ozone concentration, standard errors (uncertainty) of predictions, and probabilities that critical values are exceeded.

Introduction to the tutorial

The data you'll need for this tutorial is included on the Geostatistical Analyst installation disk. The datasets were provided courtesy of the California Air Resources Board.

The datasets are:

Dataset	Description
ca_outline	Outline map of California
ca_ozone_pts	Ozone point samples (ppm)
ca_cities	Location of major Californian cities
ca_hillshade	A hillshade map of California

The Ozone dataset (ca_ozone_pts) represents the 1996 maximum eight-hour average concentration of ozone in parts per million (ppm). (The measurements were taken daily and grouped into eight-hour blocks.) The original data has been modified for the purposes of the tutorial and should not be taken to be accurate data.

From the ozone point samples (measurements), you will produce two continuous surfaces (maps), predicting the values of ozone concentration for every location in the State of California based on the sample points that you have. The first map that you create will simply use all default options to show you how easy it is to create a surface from your sample points. The second map that you produce will allow you to incorporate more of the spatial relationships that are discovered among the points. When creating this second map, you will use the ESDA tools to examine your data. You will also be introduced to some of the geostatistical options that you can use to create a surface such as removing trends and modeling spatial autocorrelation. By using the ESDA tools and working with the geostatistical parameters, you will be able to create a more accurate surface.

Many times it is not the actual values of some caustic health risk that is of concern, but rather if it is above some toxic level. If this is the case immediate action must be taken. The third surface you create will assess the probability that a critical ozone threshold value has been exceeded.

For this tutorial, the critical threshold will be if the maximum average of ozone goes above 0.12 ppm in any eight-hour period during the year; then the location should be closely monitored. You will use the Geostatistical Analyst to predict the probability of values complying with this standard.

This tutorial is divided into individual tasks that are designed to let you explore the capabilities of the Geostatistical Analyst at your own pace. To get additional help, explore the ArcMap online Help system or see *Using ArcMap*.

- Exercise 1 takes you through accessing the Geostatistical Analyst and through the process of creating a surface of ozone concentration to show you how easy it is to create a surface using the default parameters.
- Exercise 2 guides you through the process of exploring your data before you create the surface in order to spot outliers in the data and to recognize trends.
- Exercise 3 creates the second surface that considers more of the spatial relationships discovered in Exercise 2 and improves on the surface you created in Exercise 1. This exercise also introduces you to some of the basic concepts of geostatistics.
- Exercise 4 shows you how to compare the results of the two surfaces that you created in Exercises 1 and 3 in order to decide which provides the better predictions of the unknown values.
- Exercise 5 takes you through the process of mapping the probability that ozone exceeds a critical threshold, thus creating the third surface.
- Exercise 6 shows you how to present the surfaces you created in Exercises 3 and 5 for final display, using ArcMap functionality.

You will need a few hours of focused time to complete the tutorial. However, you can also perform the exercises one at a time if you wish, saving your results after each exercise.

Exercise 1: Creating a surface using default parameters

Before you begin you must first start ArcMap and enable Geostatistical Analyst.

Starting ArcMap and enable Geostatistical Analyst

Click the Start button on the Windows taskbar, point to Programs, point to ArcGIS, and click ArcMap. In ArcMap, click Tools, click Extensions, and check Geostatistical Analyst. Click Close.

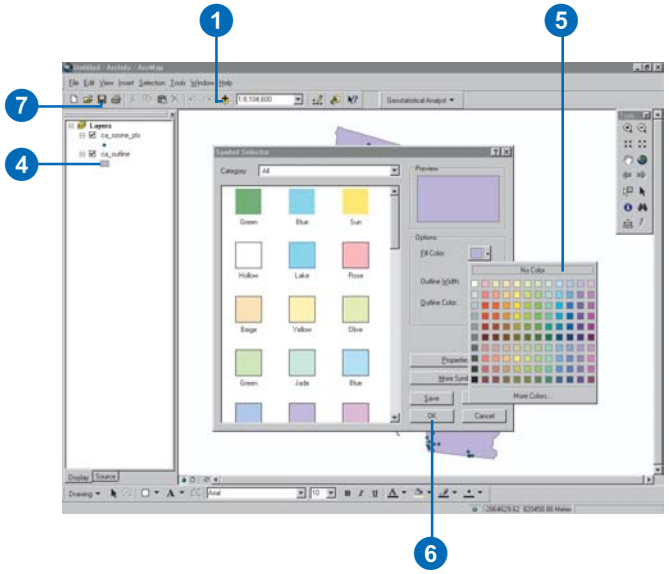
Adding the Geostatistical Analyst toolbar to ArcMap

Click View, point to Toolbars, and click Geostatistical Analyst.

Adding data layers to ArcMap

Once the data has been added, you can use ArcMap to display the data and, if necessary, to change the properties of each layer (symbology, and so on).

1. Click the Add Data button on the Standard toolbar.
2. Navigate to the folder where you installed the tutorial data (the default installation path is C:\ArcGIS\ArcTutor\Geostatistics), hold down the Ctrl key, then click and highlight the ca_ozone_pts and ca_outline datasets.
3. Click Add.
4. Click the ca_outline layer legend in the table of contents to open the Symbol Selector dialog box.
5. Click the Fill Color dropdown arrow and click No Color.
6. Click OK on the Symbol Selector dialog box.



The ca_outline layer is now displayed transparently with just the outline visible. This allows you to see the layers that you will create in this tutorial underneath this layer.

Saving your map

It is recommended that you save your map after each exercise.

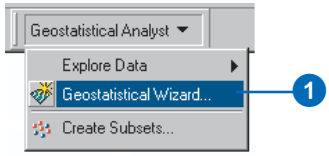
7. Click the Save button on the Standard toolbar.

   You will need to provide a name for the map because this is the first time you have saved it (we suggest Ozone Prediction Map.mxd). To save in the future, click Save.

Creating a surface using the defaults

Next you will create (interpolate) a surface of ozone concentration using the default settings of the Geostatistical Analyst. You will use the ozone point dataset (ca_ozone_pts) as the input dataset and interpolate the ozone values at the locations where values are not known using ordinary kriging. You will click Next in many of the dialog boxes, thus accepting the defaults. Do not worry about the details of the dialog boxes in this exercise. Each dialog box will be revisited in later exercises. The intent of this exercise is to create a surface using the default options.

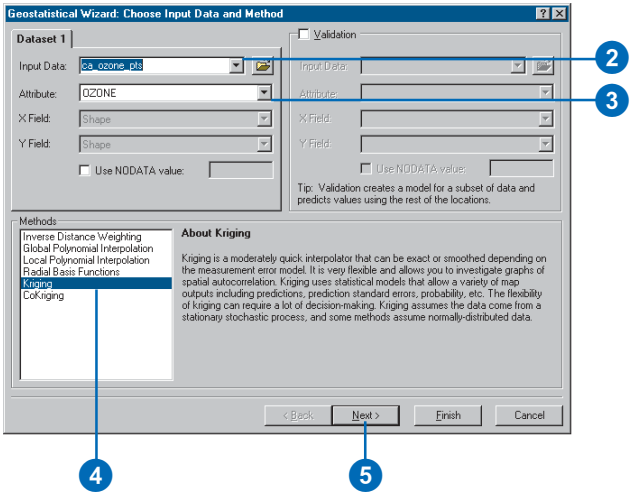
1. Click the Geostatistical Analyst toolbar, then click Geostatistical Wizard.



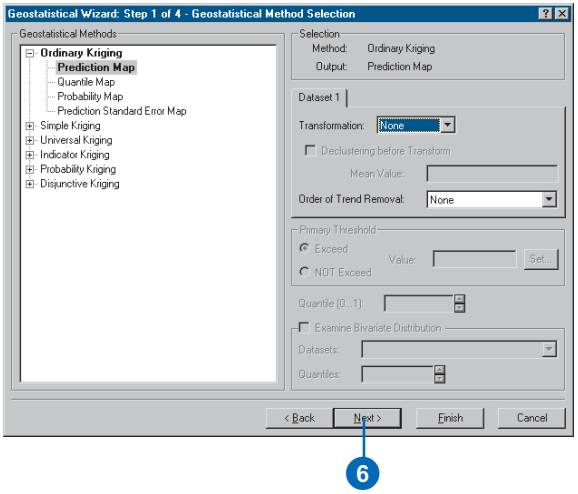
2. Click the Input Data dropdown arrow and click ca_ozone_pts.
3. Click the Attribute dropdown arrow and click the OZONE attribute.
4. Click Kriging in the Methods dialog box.
5. Click Next.

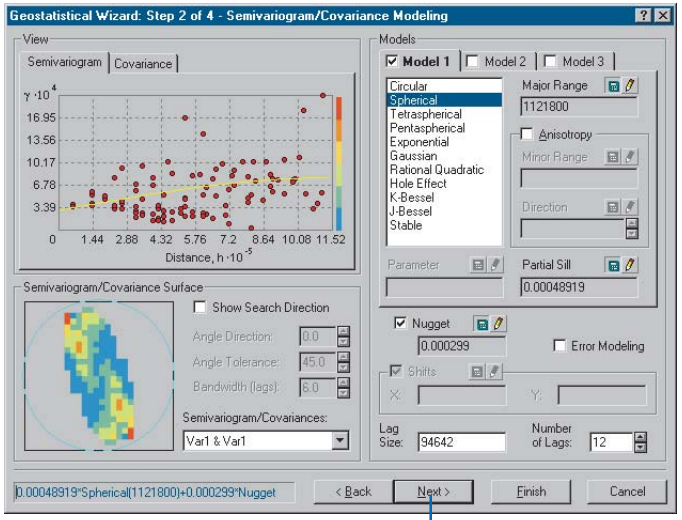
   By default, Ordinary Kriging and Prediction Map will be selected in the Geostatistical Method Selection dialog box.

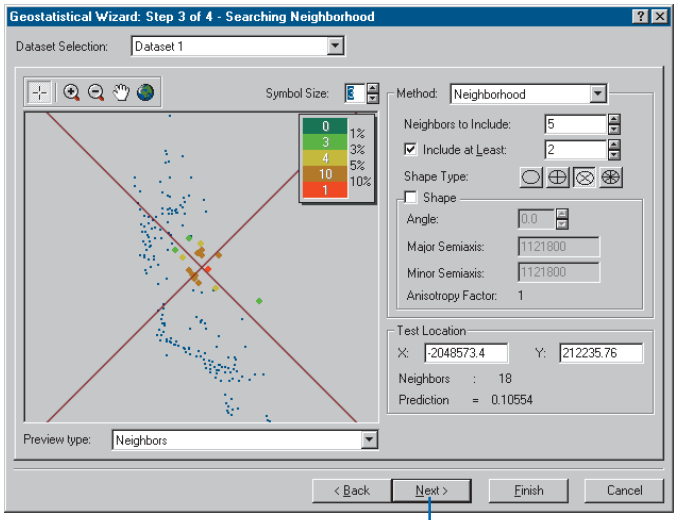
Note that having selected the method to map the ozone surface, you could click Finish here to create a surface using the default parameters. However, steps 6 to 10 will expose you to many of the different dialog boxes.



6. Click Next on the Geostatistical Method Selection dialog box.





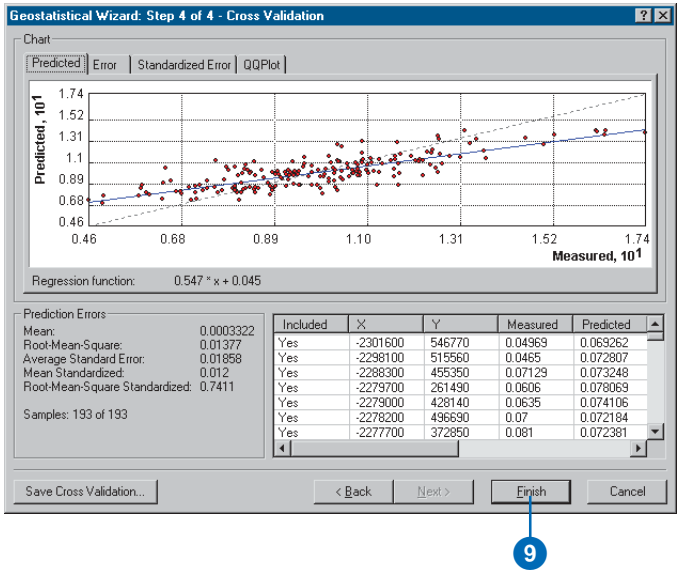


The Semivariogram/Covariance Modeling dialog box allows you to examine spatial relationships between measured points. You assume things that are close are more alike. The semivariogram allows you to explore this assumption. The process of fitting a semivariogram model while capturing the spatial relationships is known as variography.

7. Click Next.

The crosshairs show a location that has no measured value. To predict a value at the crosshairs you can use the values at the measured locations. You know that the values of the closer measured locations are more like the value of the unmeasured location that you are trying to predict. The red points in the above image are going to be weighted (or influence the unknown value) more than the green points since they are closer to the location you are predicting. Using the surrounding points, with the model fitted in the Semivariogram Modeling dialog box, you can predict a more accurate value for the unmeasured location.

8. Click Next.



The Cross Validation dialog box gives you some idea of "how well" the model predicts the values at the unknown locations. You will learn how to use the graph and understand the statistics in Exercise 4.

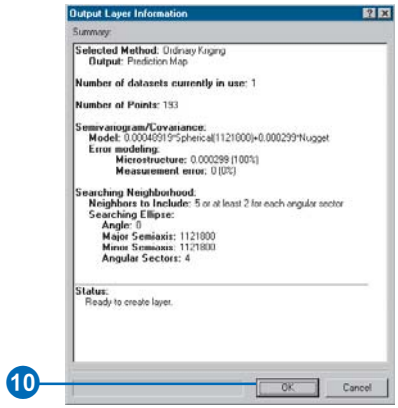
9. Click Finish.

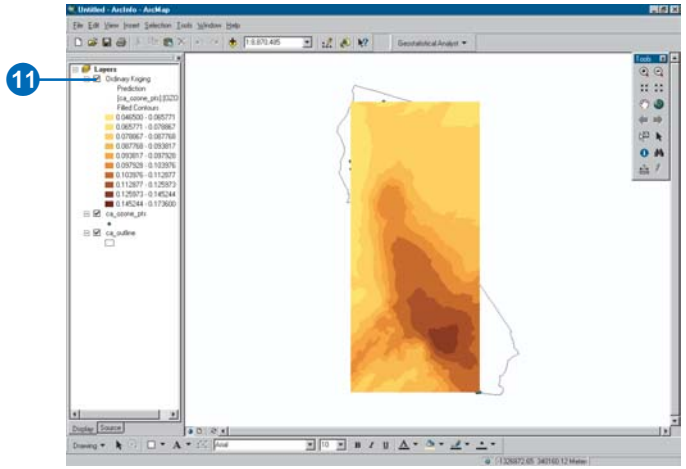
   The Output Layer Information dialog box summarizes information on the method (and its associated parameters) that will be used to create the output surface.

10. Click OK.

    The predicted ozone map will appear as the top layer in the table of contents.

11. Click the layer in the table of contents to highlight it, then click again, and change the layer name to "Default".





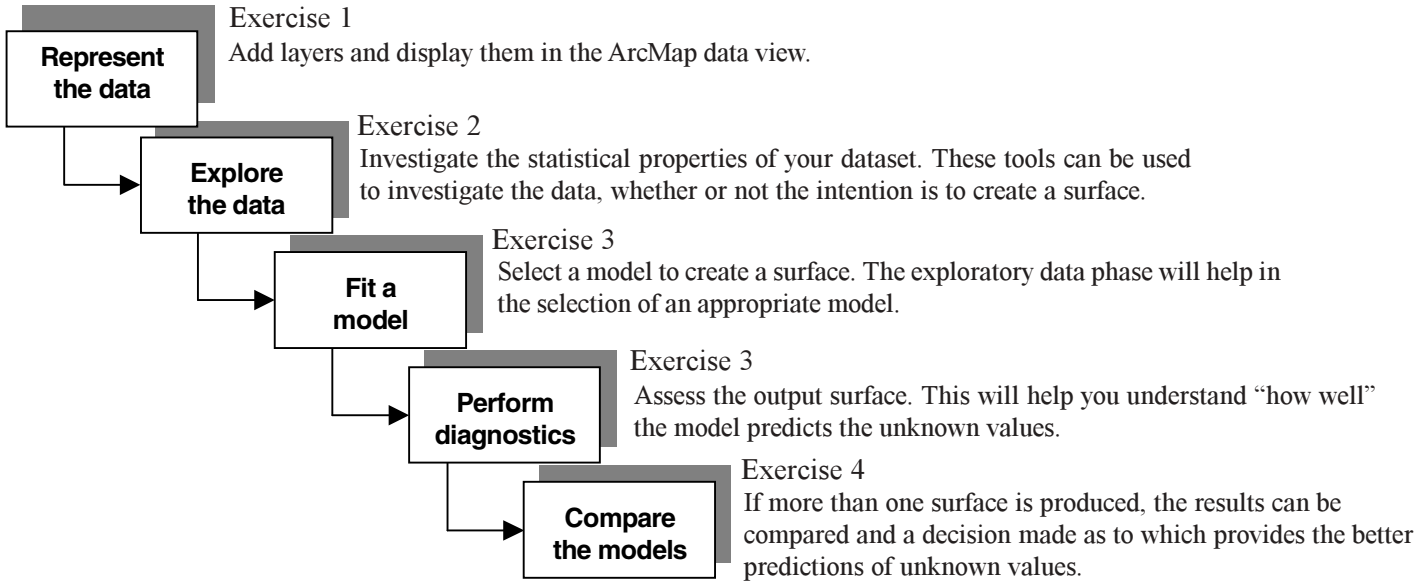
   This name change will help you distinguish this layer from the one you will create in Exercise 4.

12. Click save on the ArcMap Standard toolbar.

Notice that the interpolation continues into the ocean. You will learn in Exercise 6 how to restrict the prediction surface to stay within California.

Surface-fitting methodology

You have now created a map of ozone concentration and completed Exercise 1 of the tutorial. While it is a simple task to create a map (surface) using the Geostatistical Analyst, it is important to follow a structured process as shown below:



You will follow this structured process in the following exercises of the tutorial. In addition, in Exercise 5, you will create a surface of those locations that exceed a specified threshold and, in Exercise 6, you will create a final presentation layout of the results of the analysis performed in the tutorial.

Note that you have already performed the first step of this process, representing the data, in Exercise 1. In Exercise 2, you will explore the data.

Exercise 2: Exploring your data

In this exercise you will explore your data. As the structured process on the previous page suggests, to make better decisions when creating a surface you should first explore your dataset to gain a better understanding of it. When exploring your data you should look for obvious errors in the input sample data that may drastically affect the output prediction surface, examine how the data is distributed, look for global trends, etc.

The Geostatistical Analyst provides many data-exploration tools.

In this tutorial you will explore your data in three ways:

- Examine the distribution of your data.
- Identify the trends in your data, if any.
- Understand the spatial autocorrelation and directional influences.

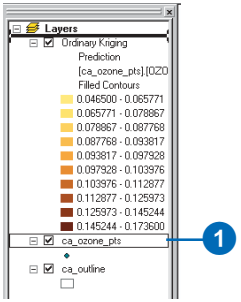
If you closed the map after Exercise 1, click the File menu and click Open. In the dialog box, click the Look in box dropdown arrow and navigate to the folder where you saved the map document (Ozone Prediction Map.mxd). Click Open.

Examining the distribution of your data

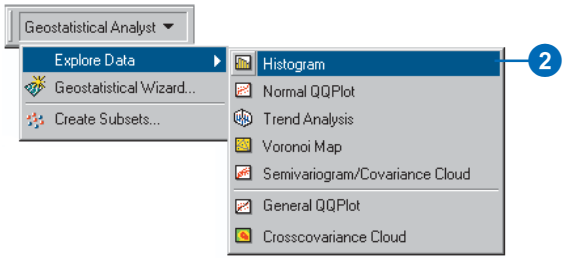
Histogram

The interpolation methods that are 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 may choose to transform the data to make it normal. Thus, it is important to understand the distribution of your data before creating a surface. The Histogram tool plots frequency histograms for the attributes in the dataset, enabling you to examine the univariate (one-variable) distribution of the dataset for each attribute. Next, you will explore the distribution of ozone for the ca_ozone_pts layer.

1. Click ca_ozone_pts, move it to the top of the table of contents, then place ca_outline underneath ca_ozone_pts.

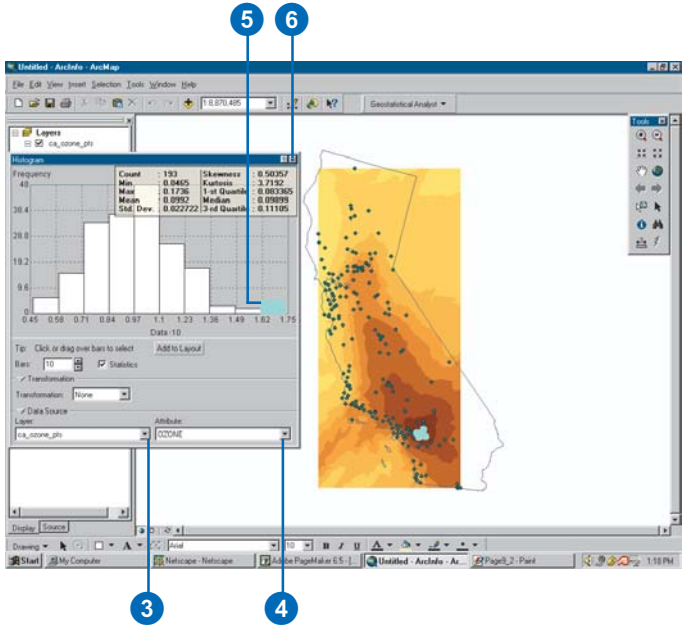


2. Click the Geostatistical Analyst toolbar, point to Explore Data, and click Histogram.



You may wish to resize the Histogram dialog box so you can also see the map, as the following diagram shows.

3. Click the Layer dropdown arrow and click ca_ozone_pts.
4. Click the Attribute dropdown arrow and click OZONE.



The distribution of the ozone attribute is depicted by a histogram with the range of values separated into 10 classes. The relative proportion (density) of data within each class is represented by the height of each bar.

Generally, the important features of the distribution are its central value, its spread, and its 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.

The histogram shown above indicates that the data is unimodal (one hump) and fairly symmetric. It appears to be close to a normal distribution. The right tail of the distribution indicates the presence of a relatively small number of sample points with large ozone concentration values.

5. Click the histogram bar with ozone values ranging from 0.162 to 0.175 ppm.

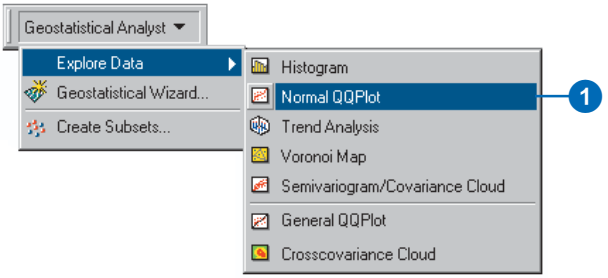
   The sample points within this range are highlighted on the map. Note that these sample points are located within the Los Angeles region.

6. Click to close the dialog box.

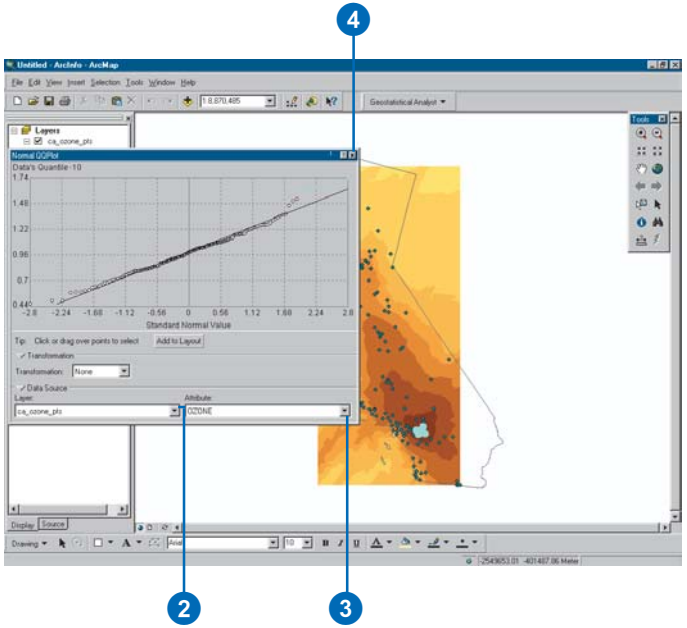
Normal QQPlot

The QQPlot is where you compare the distribution of the data to a standard normal distribution, providing yet another measure of the normality of the data. The closer the points are to creating a straight line, the closer the distribution is to being normally distributed.

1. Click the Geostatistical Analyst toolbar, point to Explore Data, and click Normal QQPlot.



2. Click the Layer dropdown arrow and click ca_ozone_pts.
3. Click the Attribute dropdown arrow and click OZONE.



A General QQPlot is a graph on which the quantiles from two distributions are plotted versus each other. For two identical distributions, the QQPlot will be a straight line. Therefore, it is possible to check the normality of the ozone data by plotting the quantiles of that data versus the quantiles of a standard normal distribution. From the Normal QQPlot above you can see that the plot is very close to a straight line. The main departure from this line occurs at high values of ozone concentration (which were highlighted in the histogram plot so they are highlighted here also).

If the data did not exhibit a normal distribution in either the Histogram or the Normal QQPlot, it may be necessary to transform the data to make it comform to a normal distribution before using certain kriging interpolation techniques.

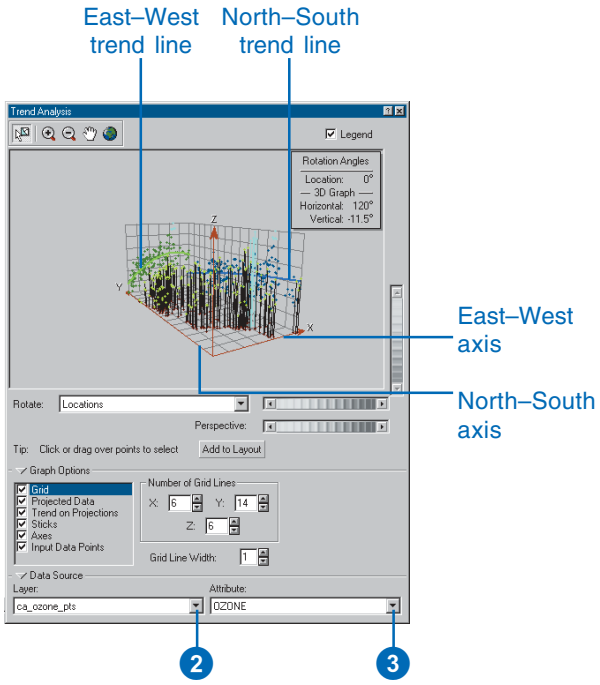
4. Click to exit the dialog box.

Identifying global trends in your data

If a trend exists in your data, it is the nonrandom (deterministic) component of a surface that can be represented by some 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 desire. 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 your surface 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".

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

1. Click the Geostatistical Analyst toolbar, point to Explore Data, and click Trend Analysis.
2. Click the Layer dropdown arrow and click ca_ozone_pts.



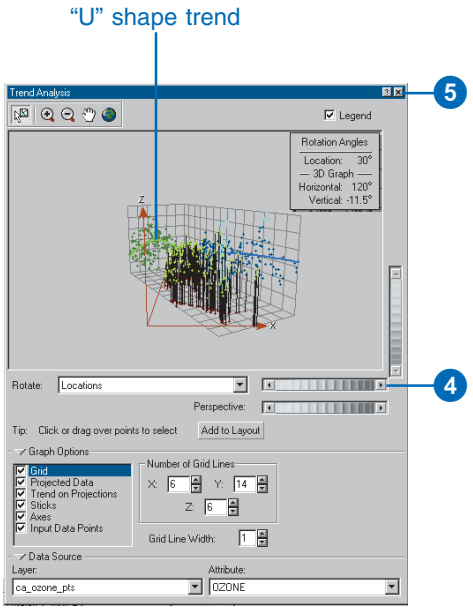
3. Click the Attribute dropdown arrow and click OZONE.

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. However, if you look at the light green line in the image above, you can see it starts out with low values and increases as it moves east until it levels out. This demonstrates that the data seems to exhibit a strong trend in the east–west direction and a weaker one in the north–south direction.

4. Click the Rotate Projection scroll bar and scroll left until the rotation angle is 30°.

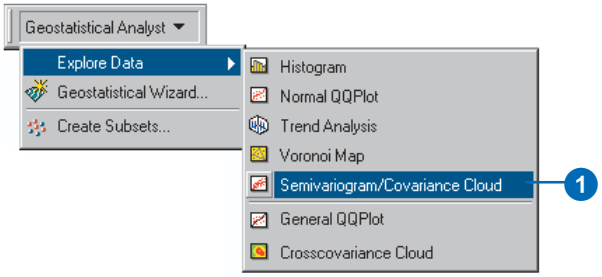
This rotation enables you to see the shape of the east–west trend more clearly. You can see that the projection actually exhibits an upside-down "U" shape. Because the trend is "U" shaped, a second-order polynomial is a good choice to use for the global trend. Even though the trend is being exhibited on the east–west projection plane, because we rotated the points 30°, the actual trend is northeast to southwest. The trend seen is possibly caused by the fact that the pollution is low at the coast, but moving inland there are large human populations that taper off again at the mountains. You will remove these trends in Exercise 4.

5. Click to exit the dialog box.

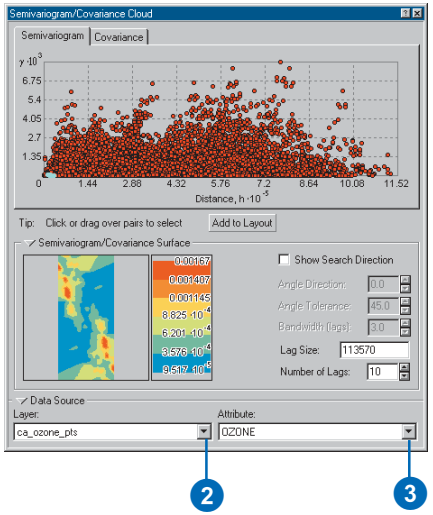


Understanding spatial autocorrelation and directional influences

1. Click the Geostatistical Analyst toolbar, point to Explore Data, and click Semivariogram/Covariance Cloud.



2. Click the Layer dropdown arrow and click ca_ozone_pts.
3. Click the Attribute dropdown arrow and click OZONE.

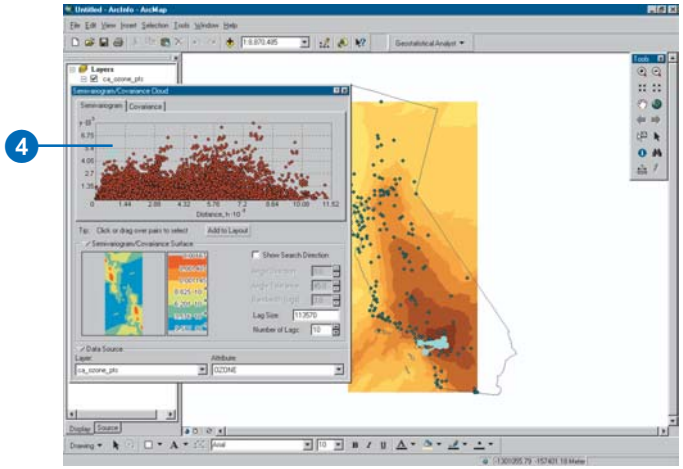


The Semivariogram/Covariance Cloud allows you to examine the spatial autocorrelation between the measured sample points. In spatial autocorrelation, it is assumed that things that are close to one another are more alike. The Semivariogram/Covariance Cloud lets you examine this relationship. To do so, 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.

Each red dot in the Semivariogram/Covariance Cloud represents a pair of locations. Since closer locations should be more alike, in the semivariogram the close locations (far left on the x-axis) should have small semivariogram values (low on the y-axis). As the distance between the pairs of locations increases (move right on the x-axis), the semivariogram values should also increase (move up on the y-axis). 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 should investigate these pairs of locations to see if there is the possibility that the data is inaccurate.

4. Click and drag the Selection pointer over these points to highlight them. (Use the following diagram as a guide. It is not important to highlight the exact points the diagram displays.)



The pairs of sample locations that are selected in the semivariogram are highlighted on the map, and lines link the locations, indicating the pairing.

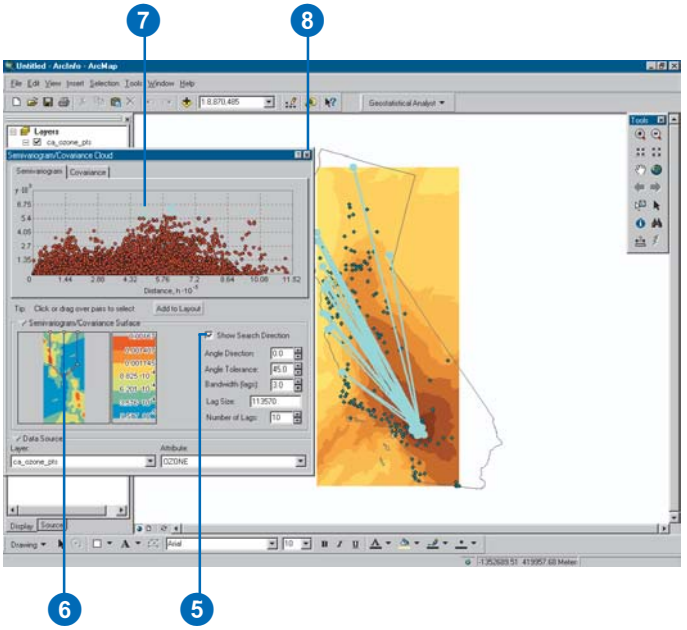
There are many reasons why the data values differ more among sample locations between the Los Angeles area and other areas. One possibility is that there are more cars in the Los Angeles area than in other areas, which will invariably produce more pollution, contributing to a higher ozone buildup in the Los Angeles area.

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, the Geostatistical Analyst provides tools to account for it in the surface-creation process. To explore for a directional influence in the semivariogram cloud, you use the Search Direction tools.

5. Check Show Search Direction.
6. Click and move the directional pointer to any angle.

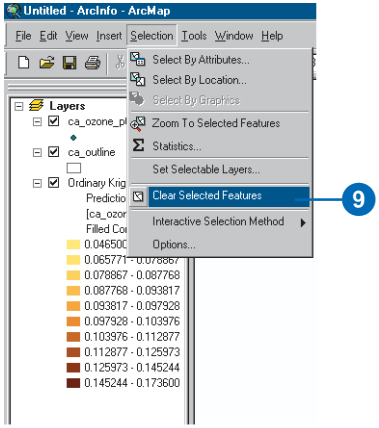
The direction the pointer is facing determines which pairs of data locations are plotted on the semivariogram. For example, if the pointer is facing an east–west direction, only the pairs of data locations that are east or west of one another will be plotted on the semivariogram. This enables you to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

7. Click and drag the Selection tool along the values with the highest semivariogram values to highlight them on the plot and in the map. (Use the following diagram as a guide. It is not important to highlight the exact points in the diagram or to use the same search direction.)



You will notice that the majority of the linked locations (representing pairs of points on the map), regardless of distance, correspond to one of the sample points from the Los Angeles region. Taking more pairs of points, at any distance, into consideration, shows that it is not just pairs of points from the Los Angeles region out to the coast that have high semivariogram values. Many of the pairs of data locations from the Los Angeles region to other inland areas also have high semivariogram values. This is because the values of ozone in the Los Angeles area are so much higher than anywhere else in California.

8. Click to exit the dialog box.
9. Click Selection and click Clear Selected Features to clear the highlighted points on the map.



In this exercise we learned:

1. The ozone data is close to a normal distribution. They are unimodal and fairly symmetrical around the mean/median line as seen in the histogram.
2. The Normal QQPlot reaffirmed that the data is normally distributed since the points in the plot created a fairly straight line, and transformation is not necessary.
3. Using the Trend Analysis tool you saw that the data exhibited a trend and, once refined, identified that the trend would be best fit by a second-order polynomial in the southeast to northwest direction (330 degrees).
4. From the Semiovariogram/Covariance Cloud we found that the high values of ozone concentration in Los Angeles create high semivariance values with locations nearby as well as far away.
5. The semivariogram surface indicates there is a spatial autocorrelation in the data.

Knowing that there are no outlier (or erroneous) sample points in the dataset and that the distribution is close to normal, you can proceed with confidence to the surface interpolation. Also, you will be able to create a more accurate surface because you know that there is a trend in the data that you can adjust for in the interpolation.

Exercise 3: Mapping ozone concentration

In Exercise 1, you used the default parameters to map ozone concentration. However, you did not take into account the statistical properties of the sample data. For example, from exploring the data in Exercise 2, it appeared that the data exhibited a trend. This can be incorporated into the interpolation process.

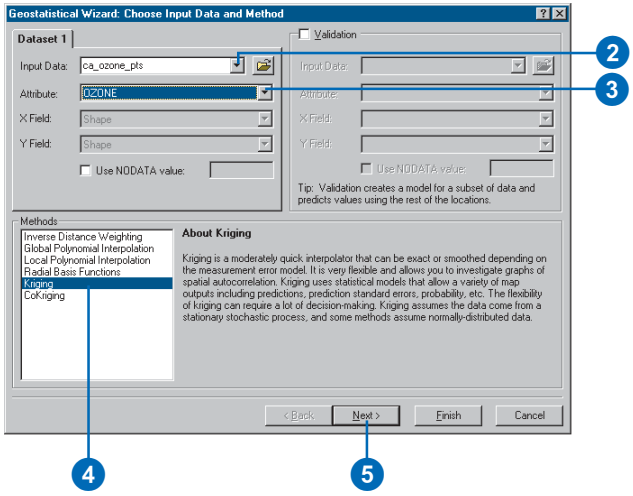
In this exercise you will:

- Improve on the map of ozone concentration created in Exercise 1.
- Be introduced to some basic geostatistical concepts.

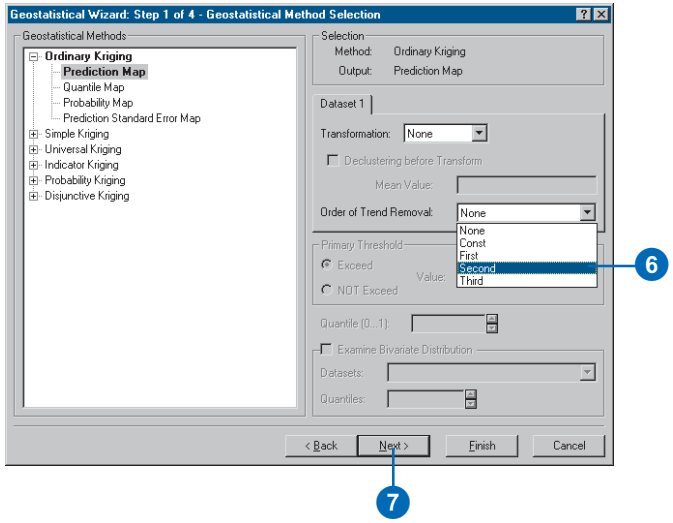
Again you will use the ordinary kriging interpolation method and will incorporate the trend in your model to create better predictions.

1. Click the Geostatistical Analyst toolbar and click Geostatistical Wizard.
2. Click the Input Data dropdown and click ca_ozone_pts.
3. Click the Attribute dropdown arrow and click the OZONE attribute.
4. Click Kriging in the Methods box.
5. Click Next.

   By default, Ordinary Kriging and Prediction are selected.



From the exploration of your data in Exercise 2, you discovered that there was a global trend in your data. After refinement with the Trend Analysis tool, you discovered that a second-order polynomial seemed reasonable and the trend was from the southeast to the northwest. This trend can be represented by a mathematical formula and removed from the data. Once the trend is removed, the statistical analysis will be performed on the residuals or the short-range variation component of the surface. The trend will automatically be added back before the final surface is created so that the predictions will produce meaningful results. By removing the trend, the analysis that is to follow will not be influenced by the trend, and once it is added back a more accurate surface will be produced.

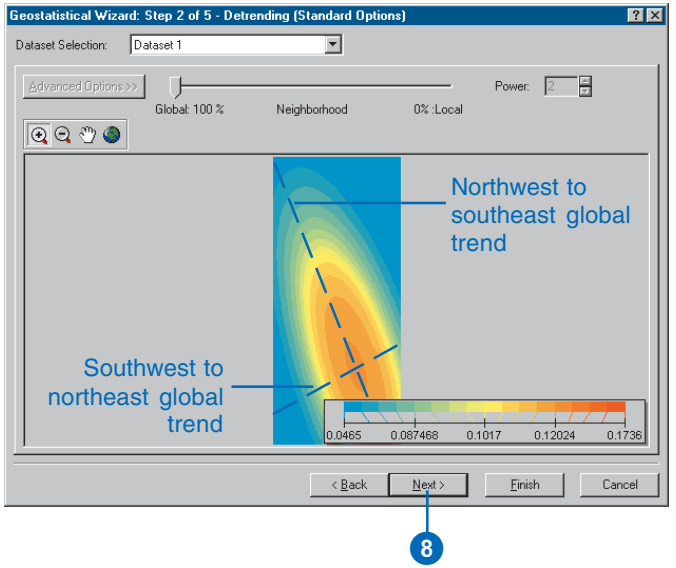


6. On the Geostatistical Method Selection dialog box, click the Order of Trend Removal dropdown arrow and click Second.

   A second-order polynomial will be fitted because a U-shaped curve was detected in the southwest to northeast direction in the Trend Analysis dialog box in Exercise 2.

7. Click Next on the Geostatistical Method Selection dialog box.

   By default, the Geostatistical Analyst maps the global trend in the dataset. The surface indicates the most rapid change in the southwest to northeast direction and a more gradual change in the northwest–southeast direction (causing the ellipse shape).



Trends should only be removed if there is justification for doing so. The southwest to northeast trend in air quality can be attributed to an ozone buildup between the mountains and the coast. The elevation and the prevailing wind direction are contributing factors to the relatively low values in the mountains and at the coast. The high concentration of humans also leads to high levels of pollution between the mountains and coast. The northwest to southeast trend varies much more slowly due to the higher populations around Los Angeles and extending to lesser numbers in San Francisco. Hence we can justifiably remove these trends.

8. Click Next on the Detrending dialog box.

Semivariogram/Covariance modeling

In the Semivariogram/Covariance Cloud in Exercise 2, you explored the overall spatial autocorrelation of the measured points. To do so, you examined the semivariogram, which showed the difference-squared of the values between each pair of points at different distances. The goal of Semivariance/Covariance modeling is to determine the best fit for a model that will pass through the points in the semivariogram (the yellow line in the diagram).

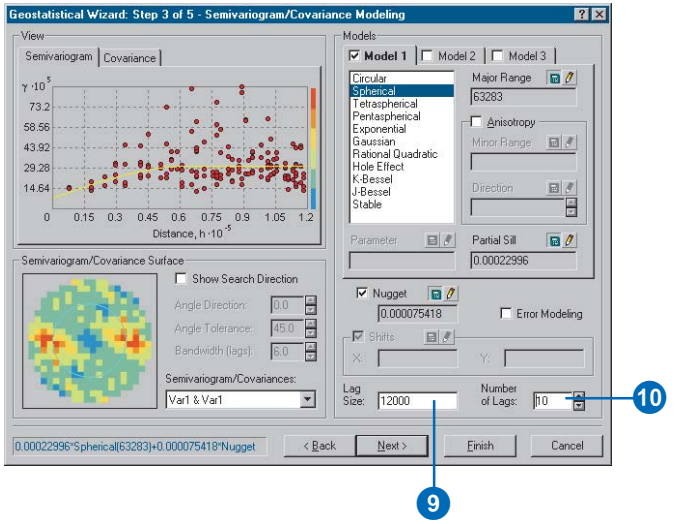
The semivariogram is a function that relates semivariance (or dissimilarity) of data points to the distance that separates them. Its graphical representation can be used to provide a picture of the spatial correlation of data points with their neighbors.

The Semivariogram/Covariance Modeling dialog box allows you to model the spatial relationship in the dataset. By default, optimal parameters for a spherical semivariogram model are calculated. The Geostatistical Analyst first determines a good lag size for grouping semivariogram values. The lag size is the size of a distance class into which pairs of locations are grouped in order to reduce the large number of possible combinations. This is called binning. As a result of the binning, notice that there are fewer points in this semivariogram than the one in Exercise 2. A good lag distance can also help reveal spatial correlations. The dialog box displays the semivariogram values as a surface and as a scatterplot related to distance. Then it fits a spherical semivariogram model (best fit for all directions) and its associated parameter values, which are typically called the nugget, range, and partial sill.

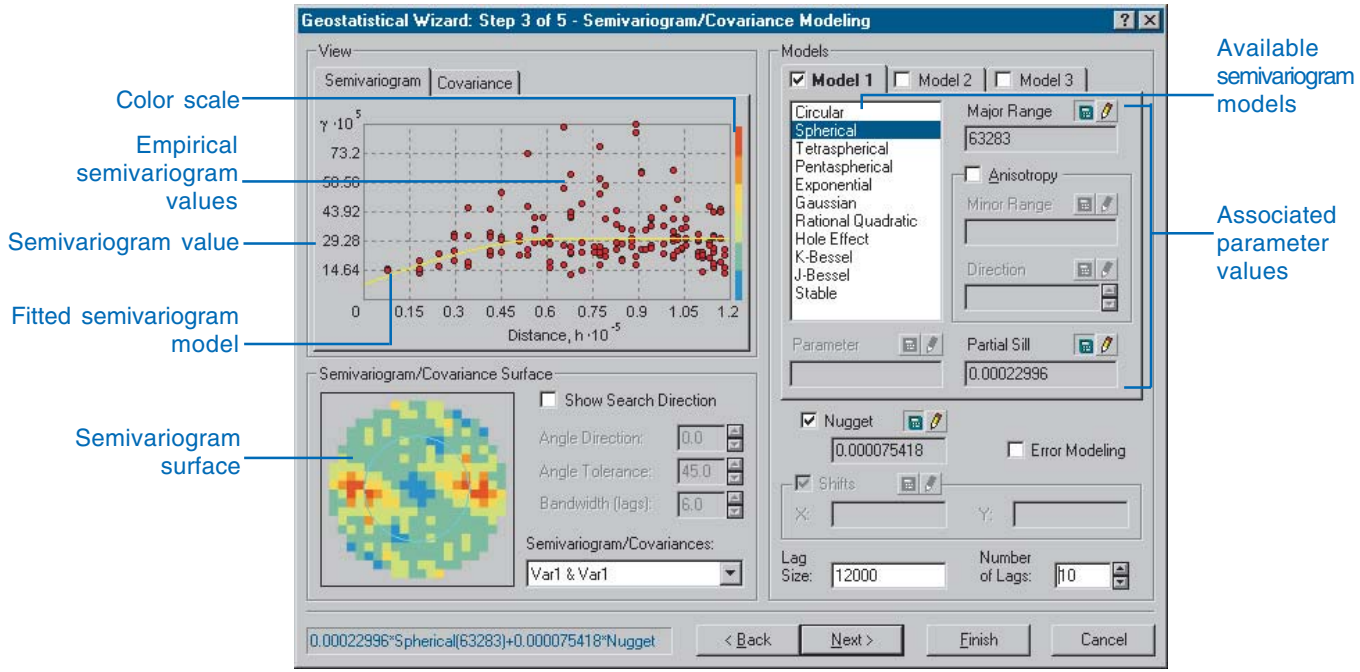
Try to fit the semivariogram at small lags (distances). It is possible to use different bin sizes and refit the default spherical model by changing the lag size and number of lags.

9. Type a new Lag Size value of 12000.
10. Click in the input box and type 10 for the Number of Lags.

Reducing the lag size means that you are effectively zooming in to model the details of the local variation between neighboring sample points. You will notice that with a smaller lag size, the fitted semivariogram (the yellow line) rises sharply and then levels off. The range is the distance where it levels off. This flattening out of the semivariogram indicates that there is little autocorrelation beyond the range.



By removing the trend, the semivariogram will model the spatial autocorrelation among data points without having to consider the trend in the data. The trend will be automatically added back to the calculations before the final surface is produced.



The color scale, which represents the calculated semivariogram value, provides a direct link between the empirical semivariogram values on the graph and those on the semivariogram surface. The value of each "cell" in the semivariogram surface is color coded, with lower values blue and green and higher values orange and red. The average value for each cell of the semivariogram surface is plotted on the semivariogram graph. The x-axis on the semivariogram graph is the distance from the center of the cell to the center of the semivariogram surface. The semivariogram values represent dissimilarity. For our example, the semivariogram starts low at small distances (things close together are more similar) and increases as distance increases (things get more dissimilar farther apart). Notice from the semivariogram surface that dissimilarity increases more rapidly in the southwest to northeast direction than in the southeast to northwest direction. Earlier, you removed a coarse-scale trend. Now it appears that there are directional components to the autocorrelation at finer scales, so we will model that next.

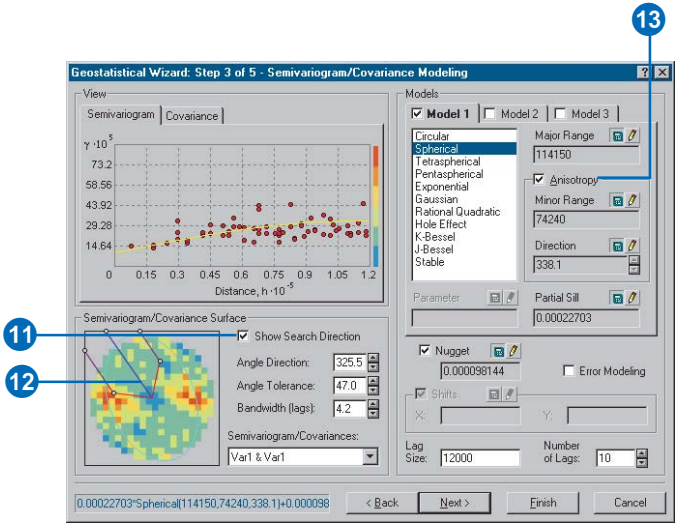
Directional semivariograms

A directional influence will affect the points of the semivariogram and the model that will be fit. In certain directions closer things may be more alike than in other directions. Directional influences are called anisotropy, and the Geostatistical Analyst can account for them. Anisotropy can be caused by wind, runoff, a geological structure, or a wide variety of other processes. The directional influence can be statistically quantified and accounted for when making your map.

You can explore the dissimilarity in data points for a certain direction with the Search Direction tool. This allows you to examine directional influences on the semivariogram chart. It does not affect the output surface. The following steps show you how to achieve this.

11. Check Show Search Direction. Note the reduction in the number of semivariogram values. Only those points in the direction of the search are displayed.
12. Click and hold the cursor on the center line in the Search Direction. Move the direction of the search tool. As you change the direction of the search, note how the semivariogram changes. Only the semivariogram surface values within the direction of the search are plotted on the semivariogram chart above.

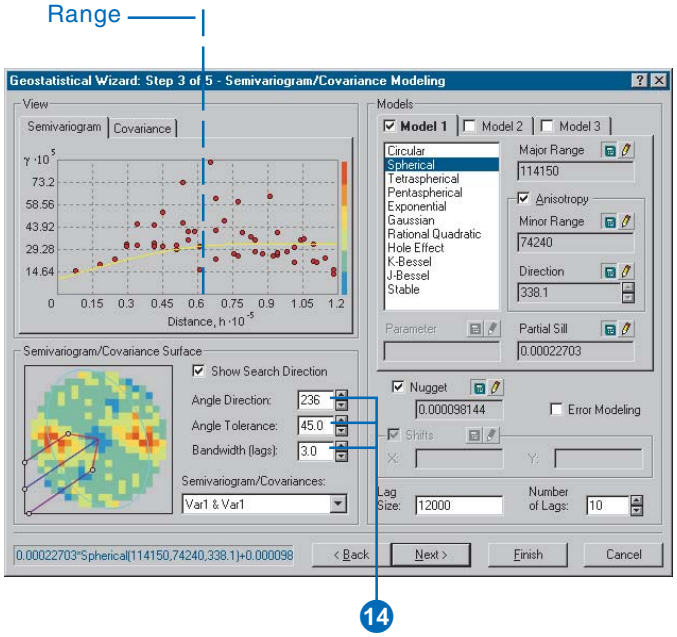
To actually account for the directional influences on the semivariogram model for the surface calculations, you must calculate the anisotropical semivariogram or covariance model.



13. Check Anisotropy.

The blue ellipse on the semivariogram surface indicates the range of the semivariogram in different directions. In this case the major axis lies approximately in the NNW–SSE direction.

Anisotropy will now be incorporated into the model to adjust for the directional influence of autocorrelation in the output surface.



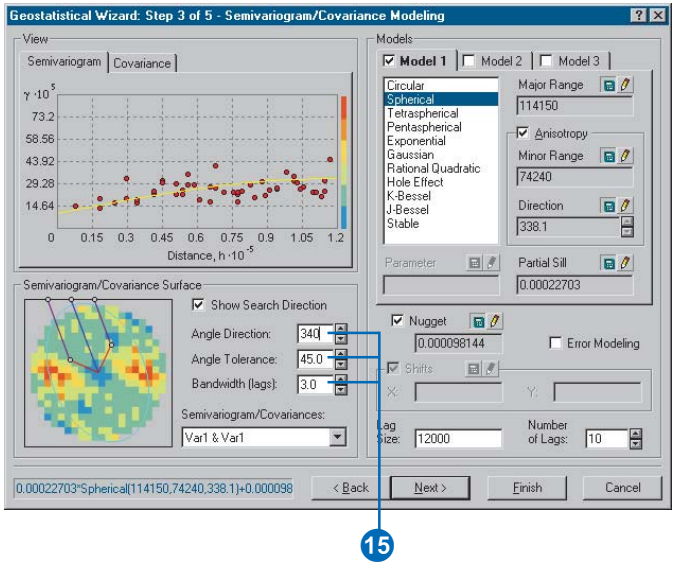
14. Type the following parameters for the Search direction to make the directional pointer coincide with the minor axis of the anisotropical ellipse:

    Angle Direction: 236.0

    Angle Tolerance: 45.0

    Bandwidth (lags): 3.0

Note that the shape of the semivariogram curve increases more rapidly to its sill value. The x- and y-coordinates are in meters, so the range in this direction is approximately 74 km.



15. Type the following parameters for the Search direction to make the directional pointer coincide with the major axis of the anisotropical ellipse:

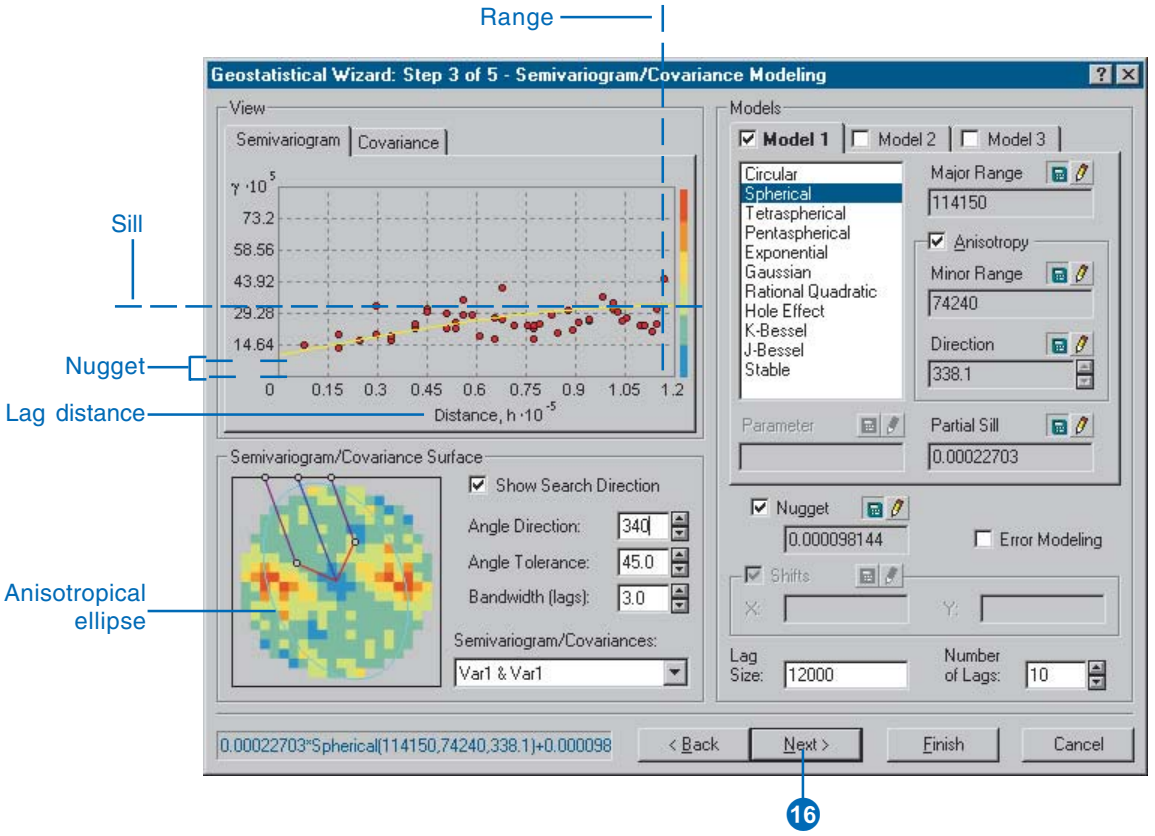
    Angle Direction: 340.0

    Angle Tolerance: 45.0

    Bandwidth (lags): 3.0

The semivariogram model increases more gradually, then flattens out. The range in this direction is 114 km.

The plateau that the semivariogram models reach in both steps 14 and 15 is the same and is known as the sill. The range is the distance at which the semivariogram model reaches its limiting value (the sill). Beyond the range, the dissimilarity between points becomes constant with increased lag distance. The lag is defined by the distance



between pairs of points. Points separated by a lag distance greater than the range are spatially uncorrelated. The nugget represents measurement error and/or microscale variation (variation at spatial scales too fine to detect). It is possible to estimate the measurement error if you have multiple observations per location, or you can decompose the nugget into measurement error and microscale variation by checking the Nugget Error Modeling check box.

16. Click Next.

Now you have a fitted model to describe the spatial autocorrelation, taking into account detrending and directional influences in the data. This information, along with the configuration and measurements of locations around the prediction location, is used to make a prediction. But how should man-measured locations be used for the calculations?

Searching neighborhood

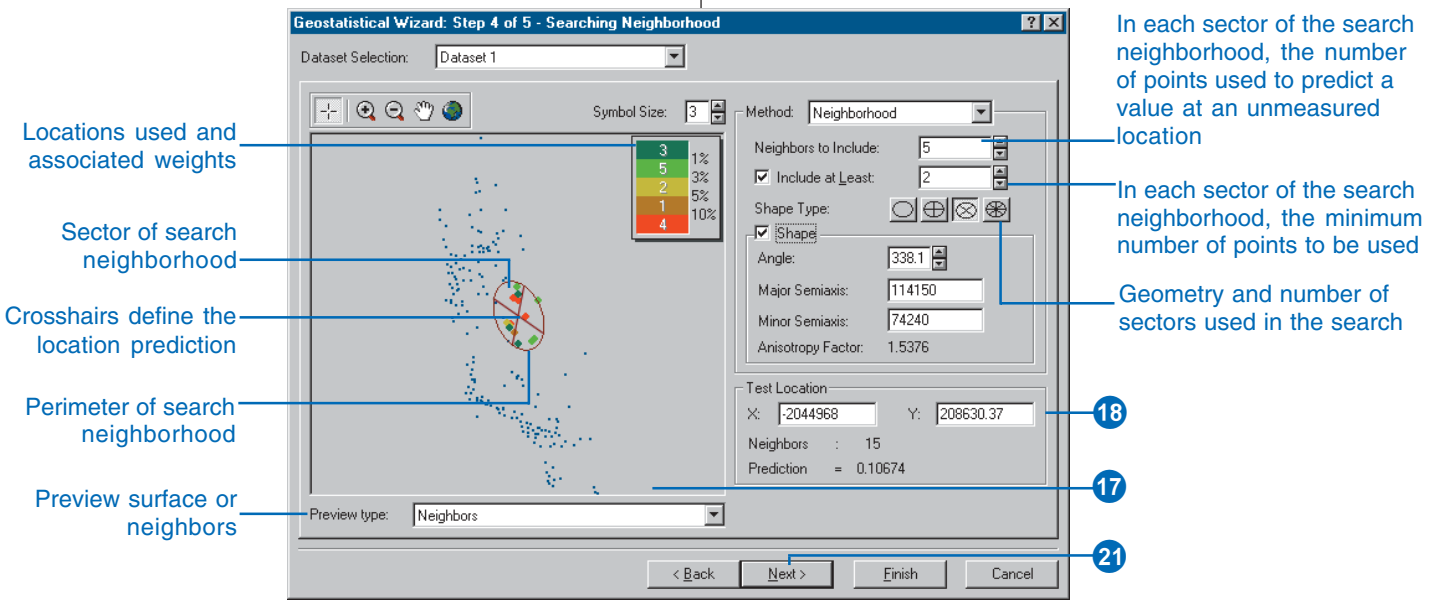
It is common practice to limit the data used by defining a circle (or ellipse) to enclose the points that are used to predict values at unmeasured locations.

Additionally, to avoid bias in a particular direction, the circle (or ellipse) can be divided into sectors from which an equal number of points are selected. By using the Searching Neighborhood dialog box, you can specify the number of points (a maximum of 200), the radius (or major/minor axis), and the number of sectors of the circle (or ellipse) to be used for prediction.

The points highlighted in the data view window give an indication of the weights that will be associated with each location in the prediction of unknown values. In this example, four locations (red) have weights of more than 10 percent. The larger the weight, the more impact that location will have on the prediction of unknown values.

17. Click inside the graph view to select a prediction location (where the crosshairs meet). Note the change in the selection of data location (together with their associated weights) that will be used for calculating the value at the prediction location.
18. For the purpose of this tutorial, type the following coordinates in the Test Location input boxes:

    X = -2044968 and Y = 208630.37.
19. Check the Shape check box and type 90 in the Angle input box. Notice how the shape changes. However, to account for the directional influences, change the angle back to 338.1.



20. Uncheck the Shape check box—the Geostatistical Analyst will use the default values (calculated in the Semivariogram/Covariance dialog earlier).

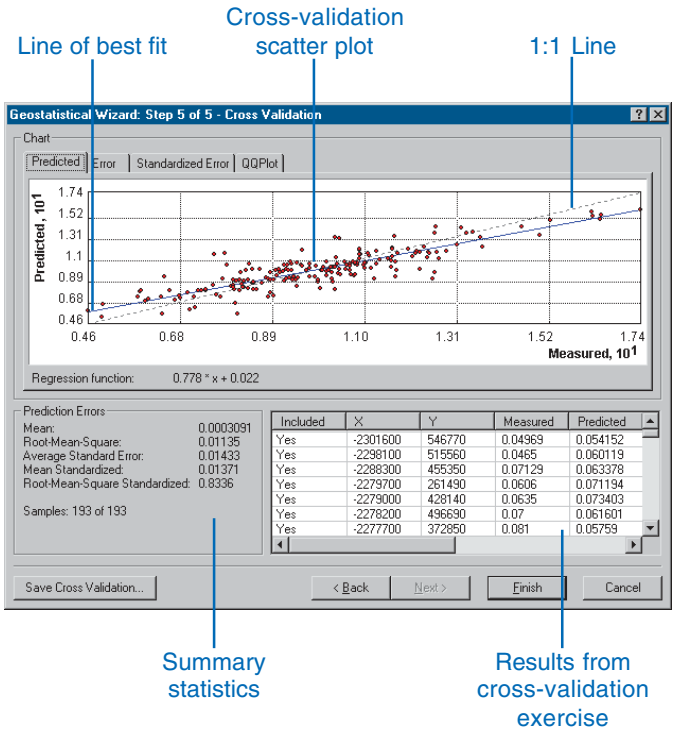
21. Click Next on the Searching Neighborhood dialog box.

Before you actually create the surface, you next use the Cross-validation dialog to perform diagnostics on the parameters to determine "how good" your model will be.

Cross-validation

Cross-validation gives you an idea of "how well" the model predicts the unknown values.

For all points, cross-validation sequentially omits a point, predicts its value using the rest of the data, and then 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. For a model that provides accurate predictions, the mean error should be close to 0, the root-mean-square error and average standard error should be as small as possible (this is useful when comparing models), and the root-mean-square stardardized error should be close to 1.

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, the root-mean-square standardized prediction error should be close to 1 if the standard errors are accurate, and the root-mean-square prediction error should be small if the predictions are close to the measured values.

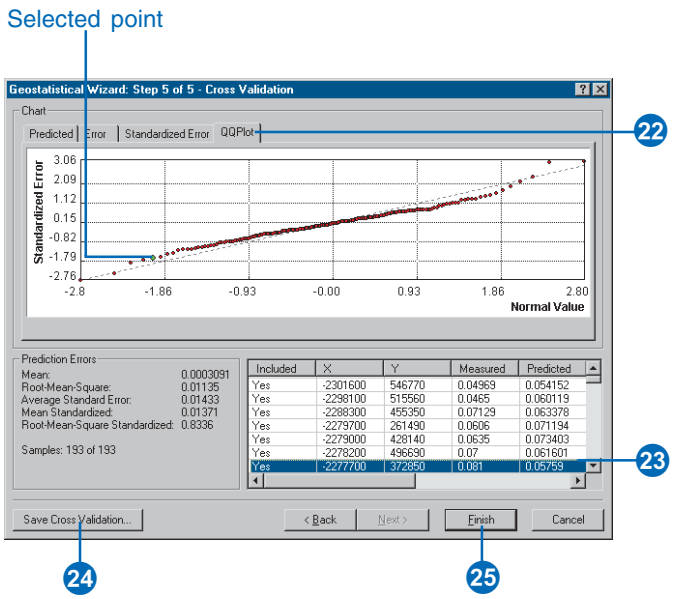
The Cross Validation dialog box also allows you to display scatterplots that show the Error, Standardized Error, and QQPlot for each data point.

22. Click the QQPlot tab to display the QQPlot.

From the QQPlot you can see that some values fall slightly above the line and some slightly below the line, but most points fall very close to the straight dashed line, indicating that prediction errors are close to being normally distributed.

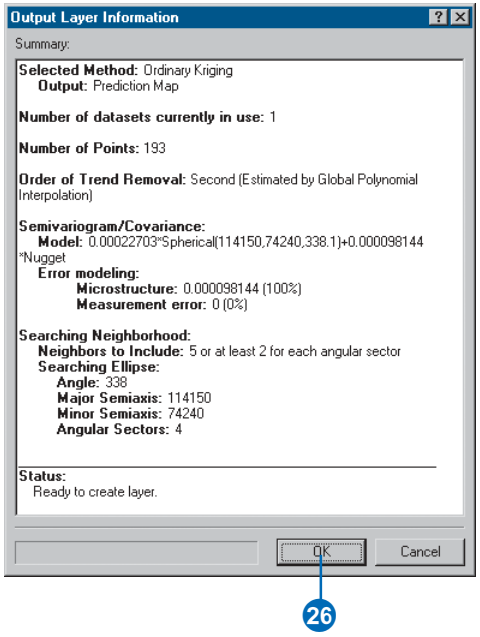
23. To highlight the location for a particular point, click on the row that relates to the point of interest in the table. The selected point is highlighted in green on the scattergram.

24. Optionally, click Save Cross Validation to save the table for further analysis of the results.



25. Click Finish.

The Output Layer Information dialog box provides a summary of the model that will be used to create a surface.

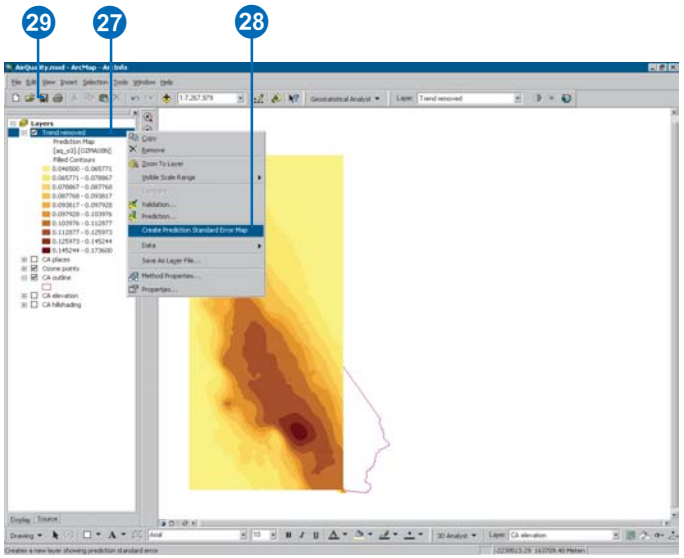


26. Click OK.

The predicted ozone map will appear as the top layer in ArcMap.

By default, the layer assumes the name of the kriging method used to produce the surface (e.g., Ordinary Kriging).

27. Click the layer name to highlight it, then click it again and change it to "Trend removed".



You can also create a Prediction Standard Error surface to examine the quality of the predictions.

28. Right-click on the "Trend removed" layer that you created and click on Create Prediction Standard Error Map.

29. Click Save on the Standard toolbar.

The Prediction Standard Errors quantify the uncertainty for each location in the surface that you created. A simple rule of thumb is that 95 percent of the time, the true value of the surface will be within the interval formed by the predicted value ± 2 times the prediction standard error if data are normally distributed. Notice in the Prediction Standard Error surface that locations near sample points generally have lower error.

The surface you created in Exercise 1 simply used the defaults of the Geostatistical Analyst, with no consideration of trends in the surface, of using smaller lag sizes, or of using an anisotropic semivariogram model. The prediction surface you created in this exercise took into consideration the global trends in the data, adjusted the lag size, and adjusted for the local directional influence (anisotropy) in the semivariogram.

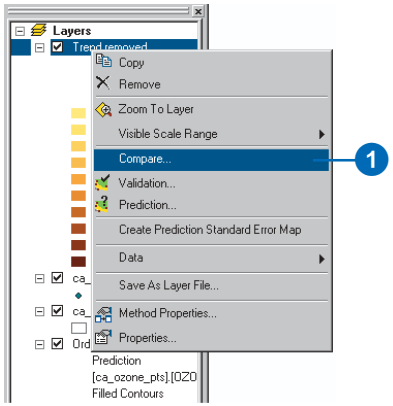
In Exercise 4, you will compare the two models to see which one provides a better prediction of unknown values.

Note: Once again, you see that the interpolation continues into the ocean. You will learn in Exercise 6 how to restrict the prediction surface to stay within California.

Exercise 4: Comparing models

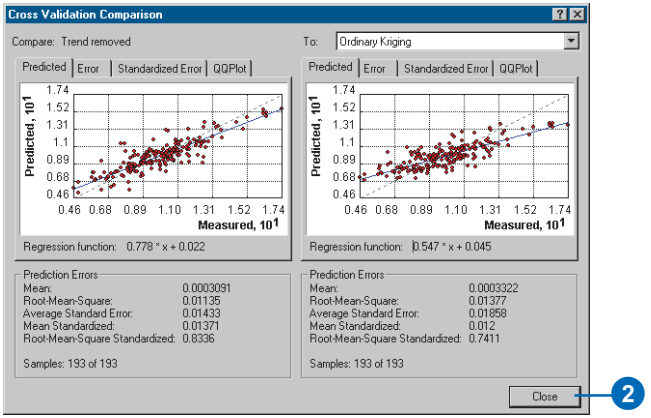
Using the Geostatistical Analyst, you can compare the results of two mapped surfaces. This allows you to make an informed decision as to which provides more accurate predictions of ozone concentration based on cross-validation statistics.

1. Right-click the "Trend removed" layer, point to Compare.... You will be comparing the "Trend removed" layer with the Default layer you created in Exercise 2.

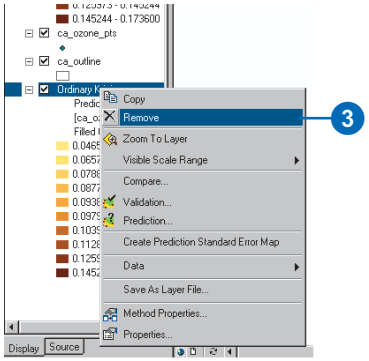


Because the root-mean-square prediction error is smaller for the Trend removed layer, the root-mean-square standardized prediction error is closer to one for the Trend removed layer, and the mean prediction error is also closer to zero for the Trend removed layer, you can state with some evidence that the Trend removed model is better and more valid. Thus, you can remove the default layer since you no longer need it.

2. Click Close on the Cross Validation Comparison dialog box.



3. Right-click the Default layer and click Remove.



4. Click the Trend removed layer and move it to the bottom of the table of contents so that you can see the sample points and outline of California.
5. Click Save on the Standard toolbar.

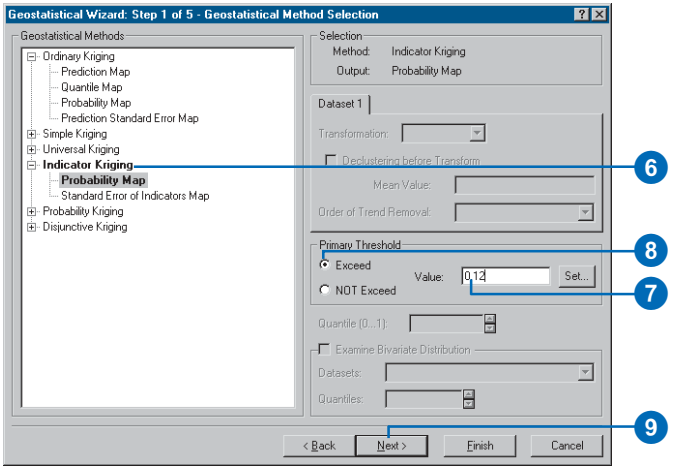
You have now identified the best prediction surface, but there may be other types of surfaces that you might wish to create.

Exercise 5: Mapping the probability of ozone exceeding a critical threshold

In Exercises 1 and 3 you used ordinary kriging to map ozone concentration in California using different parameters. In the decision making process, care must be taken in using a map of predicted ozone for identifying unsafe areas because it is necessary to understand the uncertainty of the predictions. For example, suppose the critical threshold ozone value is 0.12 ppm for an eight-hour period, and you would like to decide if any locations exceed this value. To aid the decision making process, you can use the Geostatistical Analyst to map the probability that ozone values exceed the threshold.

While the Geostatistical Analyst provides a number of methods that can perform this task, for this exercise you will use the indicator kriging technique. This technique does not require the dataset to conform to a particular distribution. The data values are transformed to a series of 0s and 1s according to whether the values of the data are below or above a threshold. If a threshold above 0.12 ppm is used, any value below this threshold will be assigned a value of 0, whereas the values above the threshold will be assigned a value of 1. Indicator kriging then uses a semivariogram model that is calculated from the transformed 0–1 dataset.

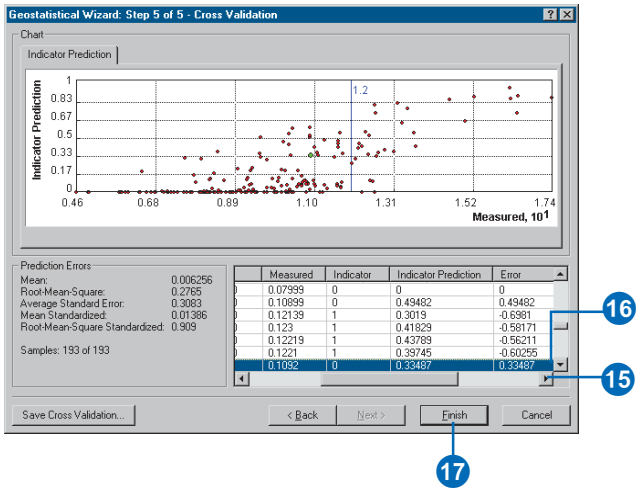
1. Click the Geostatistical Analyst toolbar and click Geostatistical Wizard.
2. Click the Layer dropdown arrow and click ca_ozone_pts.
3. Click the Attribute dropdown arrow and click the OZONE attribute.
4. Click Kriging in the method box.
5. Click Next on the Choose Input Data and Method dialog box.
6. Click Indicator Kriging; notice that Probability Map is selected.
7. Set the Primary Threshold Value to 0.12.
8. Click the Exceed radial button to select it.
9. Click Next on the Geostatistical Method Selection dialog box.



10. Click Next on the Additional Cutoffs Selection dialog box.
11. Click Anisotropy to account for the directional nature of the data.
12. Type 25000 for the lag size and 10 for the number of lags.

13. Click Next on the Semivariogram/Covariance Modeling dialog box.
14. Click Next on the Searching Neighborhood dialog box.

    The blue line represents the threshold value (0.12 ppm). Points to the left have an indicator-transform value of 0, whereas points to the right have an indicator-transform value of 1.
15. Click and scroll right until the Measured, Indicator, and Indicator Prediction columns are displayed.
16. Click and highlight a row in the table with an indicator value of 0. That point will be highlighted in green on the scattergraph, to the left of the blue threshold line.



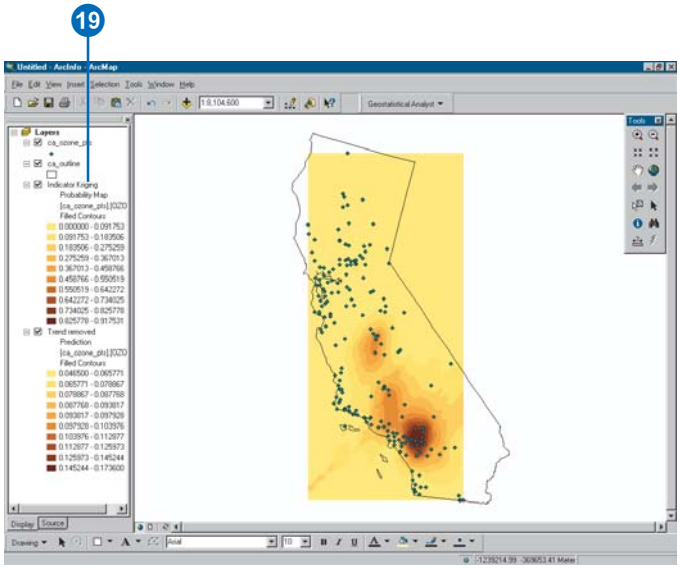
The measured and indicator columns display the actual and transformed values for each sample location. The indicator prediction values can be interpreted as the probability of exeeding the threshold. The indicator prediction values are calculated using the semivariogram modeled from the binary (0,1) data, created as indicator transformations of your original data. Cross-validation sequentially omits a point and then calculates indicator prediction values for each.

For example, the highest measured value is 0.1736. If this location had not actually been measured, a prediction of about an 85 percent chance that it was above the threshold based on the indicator kriging model would have been made.

17. Click Finish on the Cross Validation dialog box.
18. Click OK on the Output Layer Information dialog box.

    The probability map will appear as the top layer in the ArcMap data view.

    The map displays the indicator prediction values, interpreted as the probability that the threshold value of 0.12 ppm was exceeded on one or more days in the year 1996.



It is clear from the map that near Los Angeles the probability of values exceeding our threshold (staying, on average, below 0.12 ppm for every eight-hour period during the year) is likely.

19. Click and hold the Indicator Kriging layer. Drag the layer and reposition between the ca_outline and trend removed layers.

Click Save on the Standard toolbar to save your map. Exercise 6 will show you how you can use the functionality within ArcMap to produce a cartographically pleasing map of the prediction surface that you created in Exercise 3 and the probability surface that you created in this exercise.

Exercise 6: Producing the final map

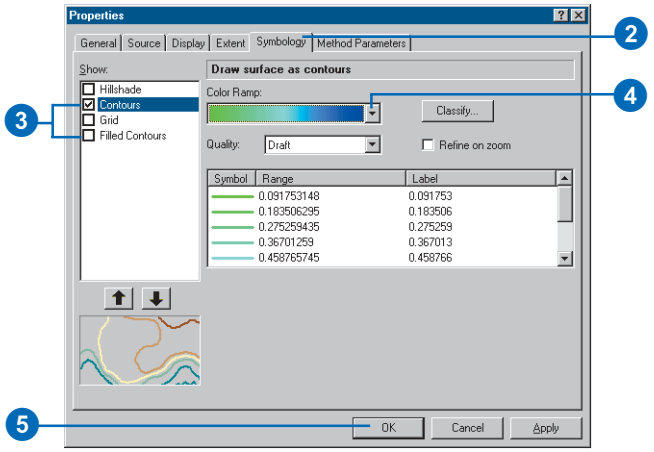
You will now produce a final map for presentation. You will use ArcMap to produce a final output map in which the prediction and the probability surfaces will be displayed.

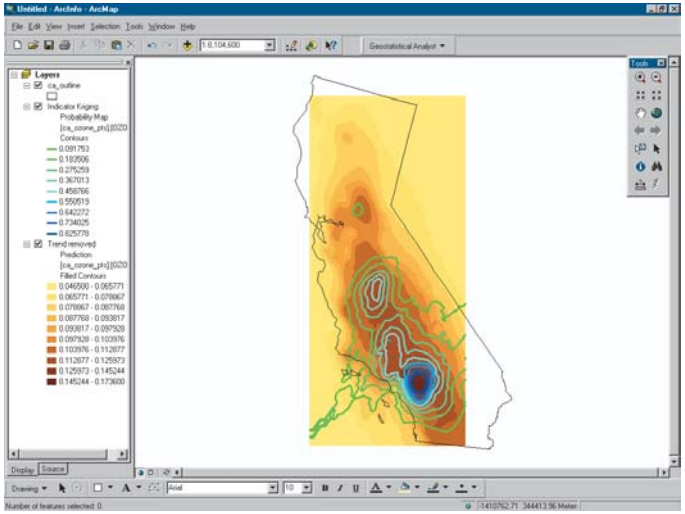
Displaying both surfaces

You can change the display of the probability map so you will be able to see both the prediction and the probability maps at the same time. The probability levels will be displayed as a contour map.

1. Right-click the Indicator Kriging layer. Click Properties.
2. Click the Symbology tab.
3. Uncheck the Filled Contours check box, then check the Contours check box.
4. Click the Color Ramp dropdown arrow and choose an alternative color ramp.
5. Click OK.

   You now see both the probability map (the contours) and the prediction map as the diagram to the right shows.





Extrapolating ozone values

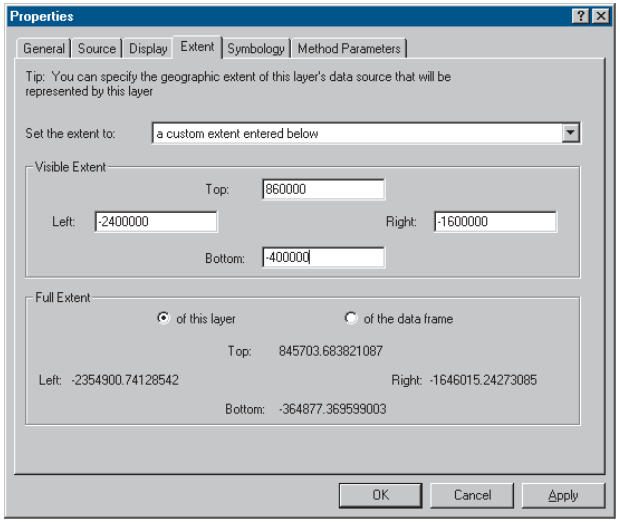
By default, the Geostatistical Analyst interpolates the value of the selected variable at any location that lies within the area defined by the north–south and east–west limits of the sample point data. However, the map of predicted ozone does not cover the geographical extent of California (the ca_outline layer). To overcome this problem you will extrapolate values (predict values outside the default bounding box) for both surfaces.

1. Right-click the Indicator Kriging layer in the table of contents and click Properties. Click the Extent tab. In Set the extent to: select a custom extent entered below and type the following values for the Visible Extent, then click OK:

   Left: -2400000 Right: -1600000

   Top: 860000 Bottom: -400000

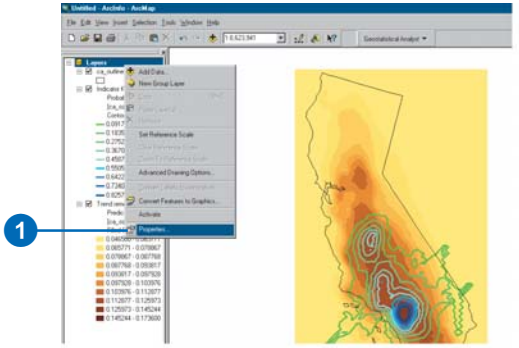
   Repeat this step for the Trend removed layer.



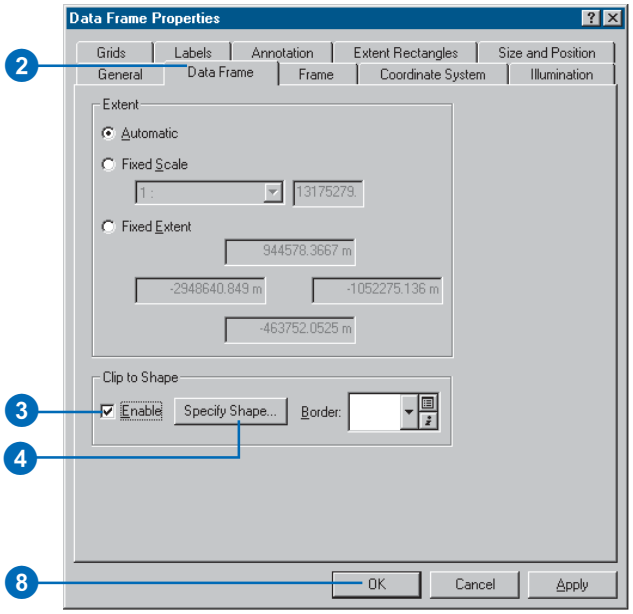
Clipping the layers to the California State outline

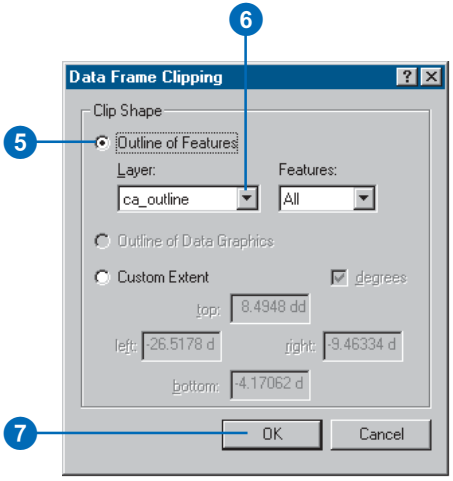
You will now clip the layers to the ca_outline layer as you are only interested in mapping the ozone levels within the State of California and this will produce a more appealing map.

1. Right-click Layers and click Properties.

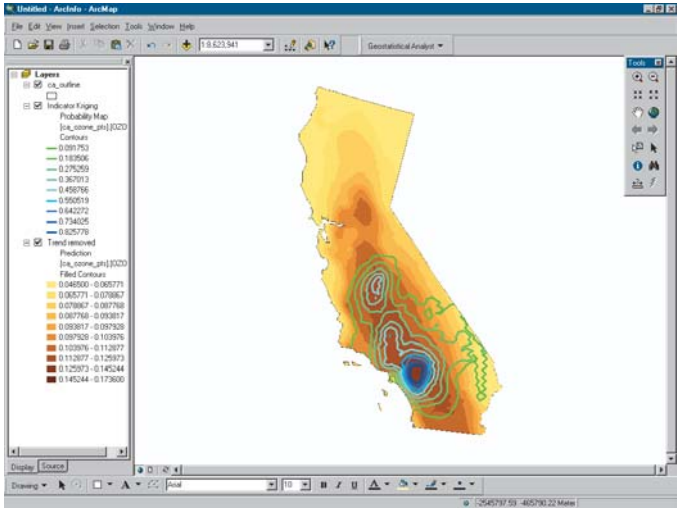


2. Click the Data Frame tab.
3. Check the Enable Clip to Shape check box.
4. Click Specify Shape.
5. Click Outline of Features.
6. Click the Layer dropdown and click ca_outline.
7. Click OK.
8. Click OK to close the Data Frame Properties dialog box.





The clipped map should look like the following diagram.



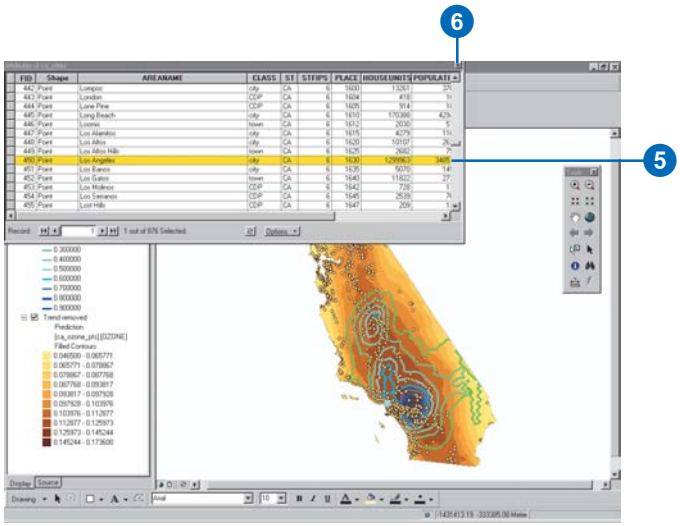
Locating the City of Los Angeles

1. Click the Add Data button on the Standard toolbar.
2. Navigate to the folder where you installed the tutorial data (the default installation path is C:\ArcGIS\ArcTutor\Geostatistics), then click ca_cities.
3. Click Add.

   A map of the location of cities in California will be displayed.

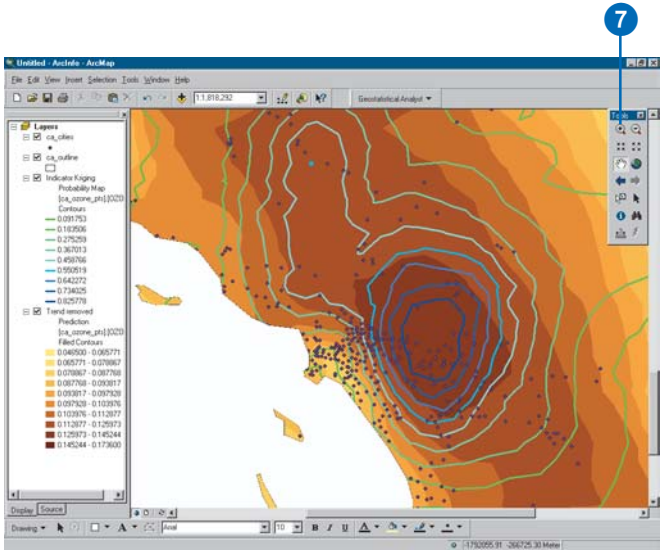
4. Right-click the ca_cities layer and click Open Attribute Table.
5. Scroll through the table and find the AreaName called Los Angeles. Click this row.

   The City of Los Angeles is highlighted on the map.
6. Click to close the attribute table.



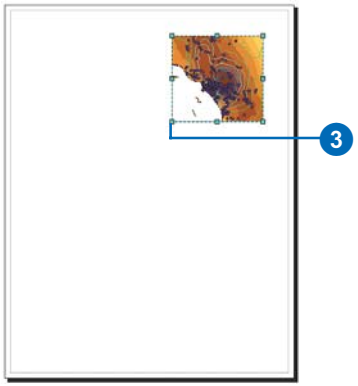
7. Click the Zoom In tool on the Tools toolbar and zoom in on the City of Los Angeles.

Notice that the area with the highest ozone concentration is actually located just to the east of Los Angeles.



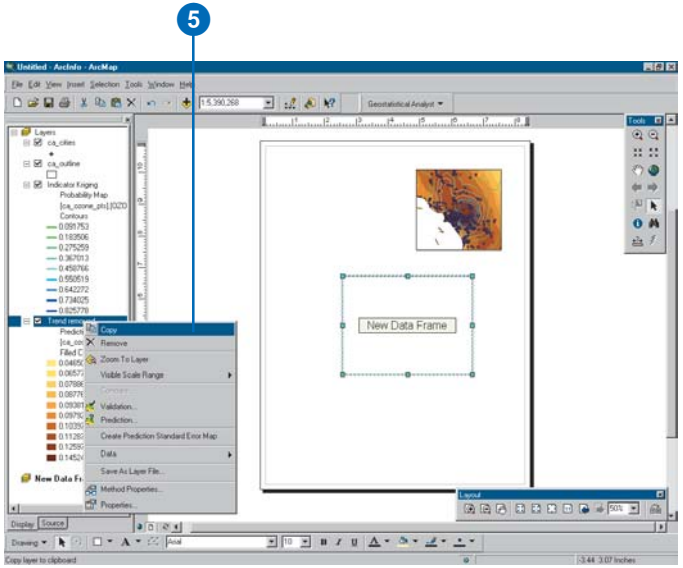
Create a layout

1. Click View on the Main menu and click Layout View.
2. Click the map to highlight it.
3. Click and drag the bottom-left corner of the Data Frame to resize the map.

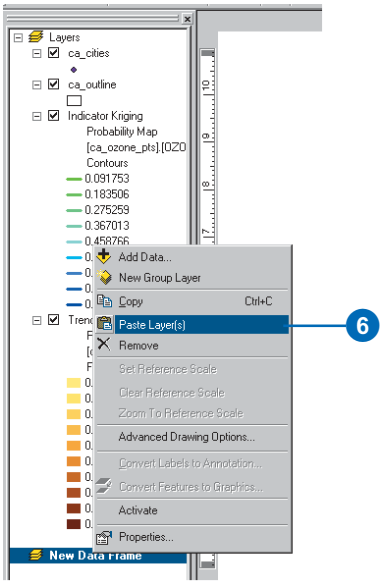


4. Click Insert on the Main menu and click Data Frame.

   A new data frame is inserted on the map. You can now copy all the layers in the first data frame into this one in order to display a map of ozone values for the whole of California alongside the ozone map, which zooms in on the Los Angeles area.

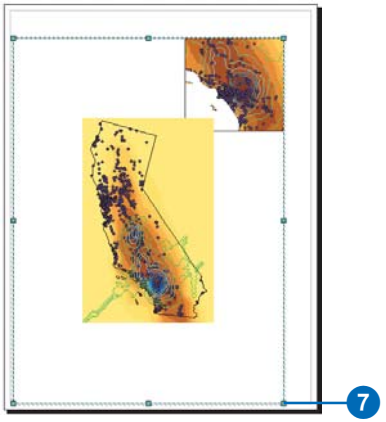


5. Right-click the Trend removed layer and click Copy.
6. Right-click the New Data Frame in the table of contents and click Paste Layer(s).



Follow steps 5 and 6 for all the other layers.

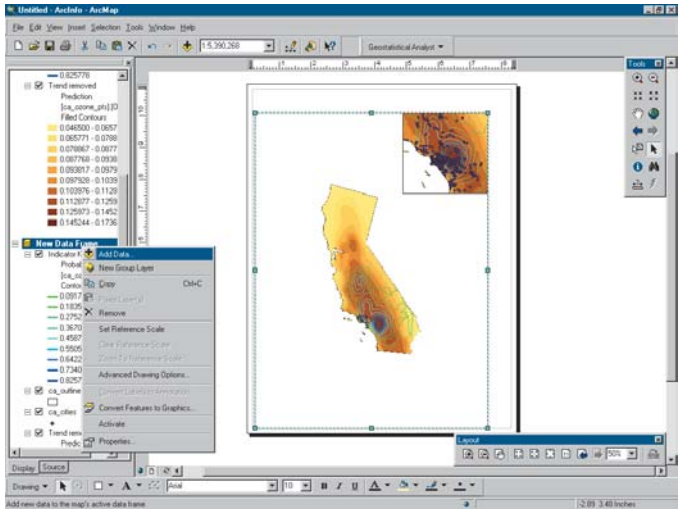
7. Click and drag the New Data Frame to fit the whole page.



8. Click the Full Extent button on the Tools toolbar to view the full extent of the map in the New Data Frame.
9. Right-click the New Data Frame and click Properties.
10. Click the Data Frame tab and, as you did for the first Data Frame, check Enable Clip to Shape and click the Specify Shape button. Choose ca_outline as the layer to clip to, then click OK.

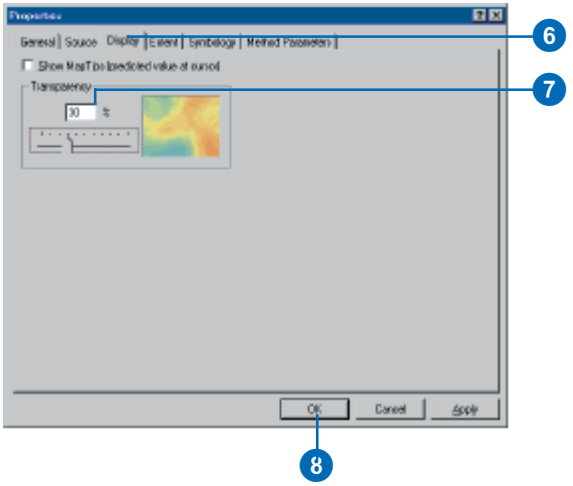
Adding a hillshade and transparency

1. Right-click the New Data Frame and click Add Data.



2. Navigate to the folder where you installed the tutorial data (the default installation path is C:\ArcGIS\ArcTutor\Geostatistics), then click ca_hillshade.
3. Click Add.

   A hillshade map of California will be displayed.
4. Click ca_hillshade and move it to the bottom of the table of contents.
5. Right-click the Trend removed layer in the New Data Frame table of contents and click Properties.



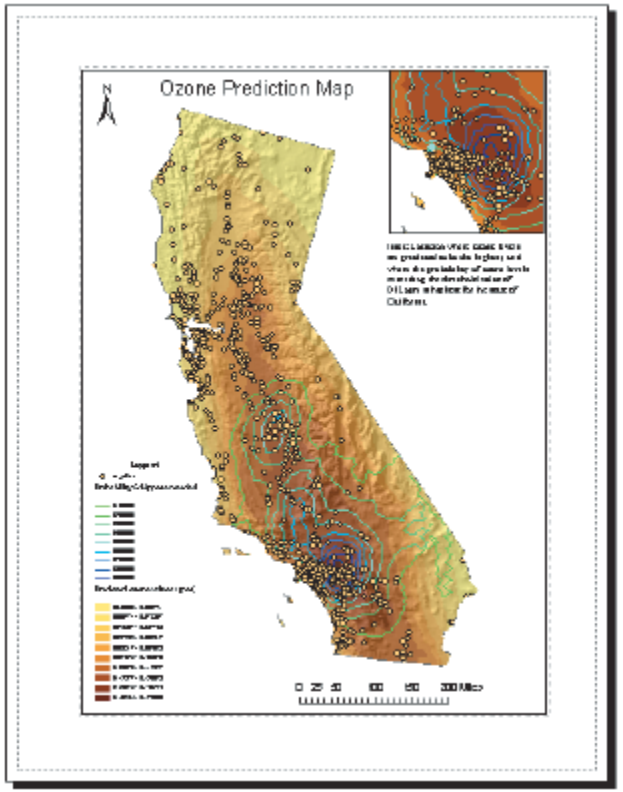
6. Click the Display tab.
7. Type 30 for the percentage of transparency.
8. Click OK.

   The hillshade should now partially display underneath the Trend removed layer.

Adding map elements

1. Click Insert on the Main menu and click Legend.
2. Move the legend to the bottom-left corner of the layout.
3. Optionally, click Insert and add a North arrow, a Scale bar, and Text.

The following diagram shows the type of finished map you could produce using the functionality of ArcMap. Refer to *Using ArcMap* if necessary to learn about inserting elements into the layout.



The map shows that the area east of Los Angeles has the highest predicted levels of ozone and the highest probability of exceeding the critical average threshold (0.12 ppm) on at least one eight-hour period during 1996. Since this is the case in the analysis (but remember the original data has been altered), you may wish to focus on these areas and analyze time series measurements of ozone to accurately identify the areas at potential risk.