Using Geostatistical Analyst for Analysis of California Air Quality

Konstantin Krivoruchko
kkrivoruchko@esri.com

For the last two centuries, air quality and air pollution have been recognized throughout the world as pressing environmental issues. From the industrial revolution in late eighteenth century in Europe through the nineteenth and twentieth centuries, the use of coal in homes and industry polluted the air. Growth continued. Smog and "pea soupers" became common in cities throughout the world. Following the famous London Smog of 1952 that lasted for five days and led to more than 5,000 deaths, pollution from industry and homes was dramatically reduced in an attempt to protect health.

Southern California experiences some of the worst air quality in the United States. Wind speeds in and around Los Angeles are low compared to other urban centers, and the region is characterized by low rates of atmospheric mixing, so pollutants become concentrated in the atmosphere. The long hours of sunshine, which attract visitors to the region's beaches, are also responsible for very high rates of photochemical reactions and the production of ozone. Federal ozone standards are commonly exceeded during spring through early autumn, whereas carbon monoxide and particulate highs are recorded in late autumn and winter.

Air quality data used in the case studies below were collected by the California Air Resources Board, Air Quality Data Branch, Statistical and Analytical Services Section, beginning in 1980. The Internet site, which provides related information, is http://www.arb.ca.gov/aqd/aqd.htm.

These case studies show spatial statistical air quality data analysis using ArcGIS Geostatistical Analyst. If you have little experience in geostatistical data analysis, we suggest reading "*Introduction to Modeling Spatial Processes Using Geostatistical Analyst*" first.

1. Analyzing Time Series Data

Since Geostatistical Analyst was designed to interpolate between variables in two dimensions, the following steps need to be used to prepare one-dimensional data for analysis:
- Use the data attribute table option "Add Field…" to create a pseudo-coordinate for the data set in the second dimension. Name the new field "Y".
- To calculate values for the new field, highlight it, right-click, and select the calculator function. In the field calculator box enter the following equation:

$$Y = 0.975 + 0.05*\text{rnd}()$$

   This generates *Y*-values equal to 1 plus or minus a small random component. This is necessary so that Geostatistical Analyst can prevent contouring in impossible situations.

- To display the data points in ArcMap, open the database layer as a DBF file. For the *x*-coordinate, select the field with daily ozone measurements, and for the *y*-coordinate, select the Y field calculated above. The *y*-values should be small, the data appearing as a line in the data view.

The data is now ready for analysis. Since Geostatistical Analyst does not provide tools to visualize the result of modeling in 1D, you will have to use other software, for example, Microsoft Excel. In the examples below, we used the Geostatistical Analyst validation option "Save Validation" to predict values of ozone for each day in 1999. The resulting DBF file was further analyzed using spreadsheet software. However, visualization of the nonspatial data using GIS is not necessary, and we advise that you export the validation or cross validation tables and cancel the analysis without visualizing the predictions.

If the input data is continuous, such as air pollutants, kriging and cokriging are powerful tools that can be used for prediction and assessing prediction uncertainties in one-dimensional data sets.

In this section, one-dimensional kriging was used to analyze daily measurements of ozone at two locations in California-San Francisco and Redlands-using the one-hour daily maximum concentration of ozone in 1999.

Our goal was to analyze dependence between observations separated by a lesser number of days, from one to seven. We predicted new values for each day using information on average dependence between following days. If the average time lag equaled one week, the resulting prediction was the average weekly ozone concentration with associated uncertainty. We used a filtered version of kriging to achieve this goal. A prediction confidence interval provided information on other likely daily ozone values.

Figure 1 shows the data in blue, and the filtered ordinary kriging interpolation, with the nugget effect treated as a measurement error (redline). Power transformation with a power value of 0.5 was used. Error bars in pink are upper and lower 95 percent confidence intervals, assuming that predictions and their standard errors are normally distributed; that is the vertical pink lines represent kriging prediction $\pm 1.96 \times$ kriging standard error values.

The green horizontal line shows the upper permissible level of 0.12 parts per million (ppm) of the annual maximum one-hour ozone concentration.
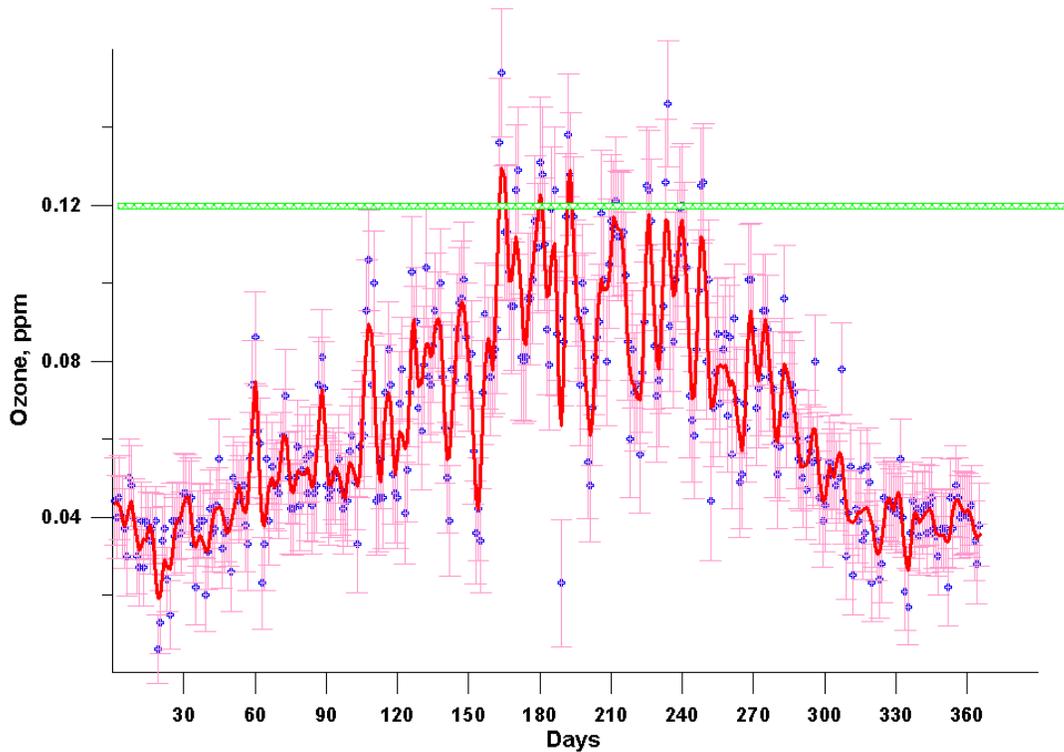
Figure 1. The Maximum Daily Concentration of Ozone per Hour in 1999 in Redlands, California. *Ordinary filtered kriging with a power transformation value of 0.5 and a semivariogram lag of one day.*

Because of uncertainties associated with observations as well as great data variability, it is risky to predict a day's ozone concentration using the measurements from a few days before. However, the weather is stable in Southern California, and usually it is not difficult to predict tomorrow's temperature based on today's weather conditions.

Figure 2 shows semivariogram models for transformed data averaging using a one-day lag. The semivariogram model is a J-Bessel of an order of 1.065 with a partial sill equaling 0.00257, a range of 5.62 days, and a nugget of 0.00203.
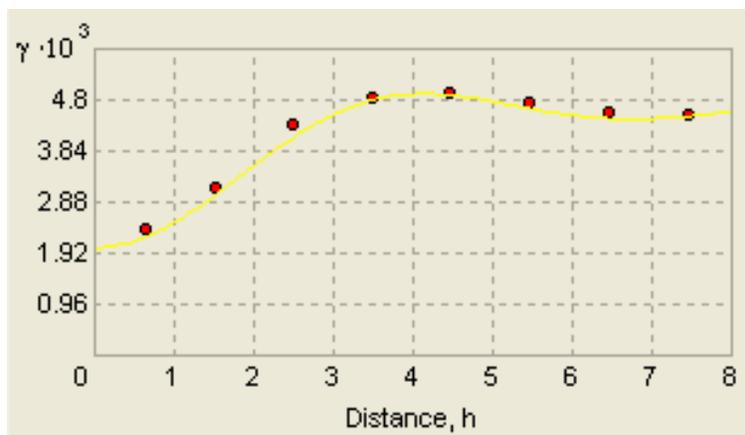
Figure 2. Semivariogram Models Using a One-Day Lag.

3

A great variation in values separated by just one day is evidence of measurement errors, not just measurement device inaccuracy but also positional errors (data is not separated by exactly one day) and time integration errors (the measurements lasted approximately one hour).

Figure 3 shows a one-dimensional analysis of the same data in figure 1 using a lag of seven days. Averaging using a greater time interval makes prediction smoother.
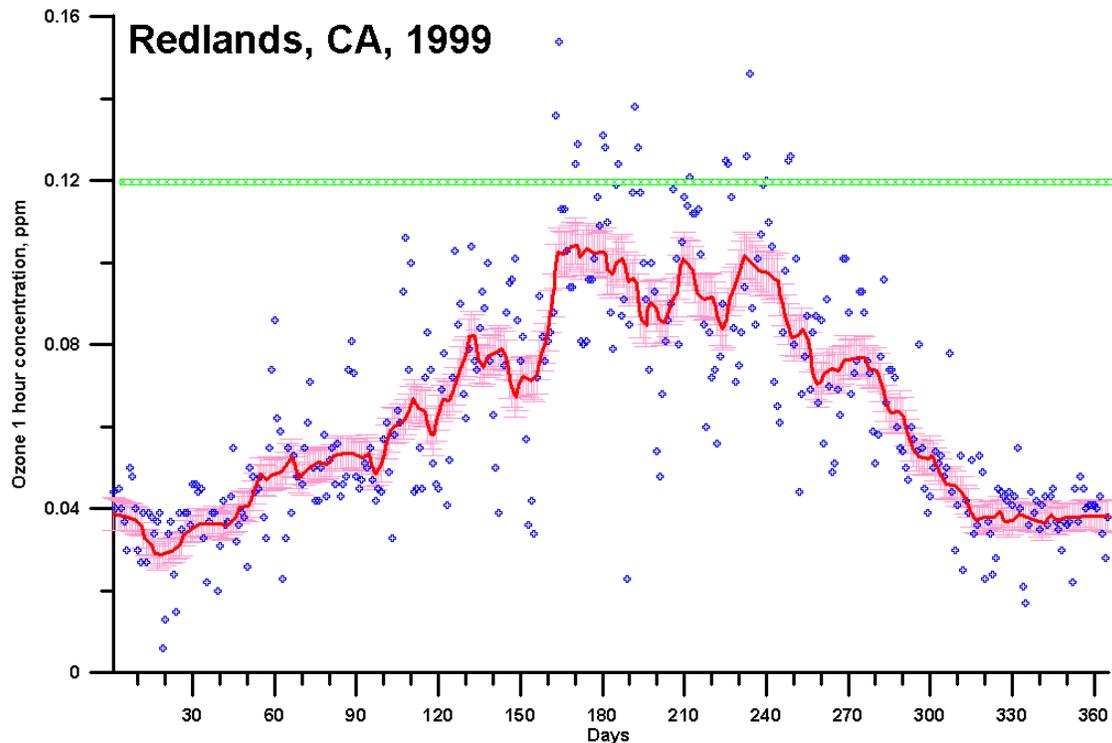
Figure 3. The Same as Figure 1 With Lag of Seven Days.

Because the mountains surrounding it are natural barriers to smog propagation from Los Angeles, as shown in Figure 4a, Redlands is one of the most contaminated cities in California, with ozone concentrations above standard in summer. However, there are places where the situation is different. For example, in San Francisco, the maximum ozone concentration in 1999, shown in Figure 4b, was registered on September 30, while the maximum ozone concentration in Redlands in 1999 was registered three and one-half months before, on June 13.

(a)	(b)
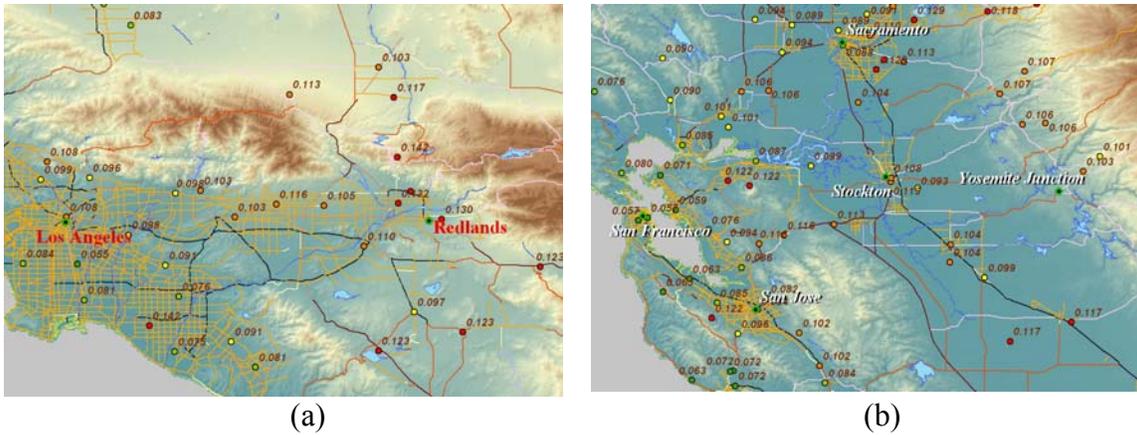
Figure 4. Areas Around Redlands (a) and San Francisco (b). Maximum one hour concentrations of ozone at the monitoring stations in 1999 are displayed. Major roads mark the source of air pollution, and hills and mountains are natural barriers to smog propagation.

Predictions of one-hour ozone concentration in San Francisco are presented in figure 5. The redline shows the filtered ordinary kriging prediction; pink and blue lines are upper and lower 95 percent confidence intervals, assuming that the predictions and their standard errors are distributed normally. The semivariogram model used for predictions is in the upper left corner of Figure 5. A lag interval of six days was chosen for averaging.
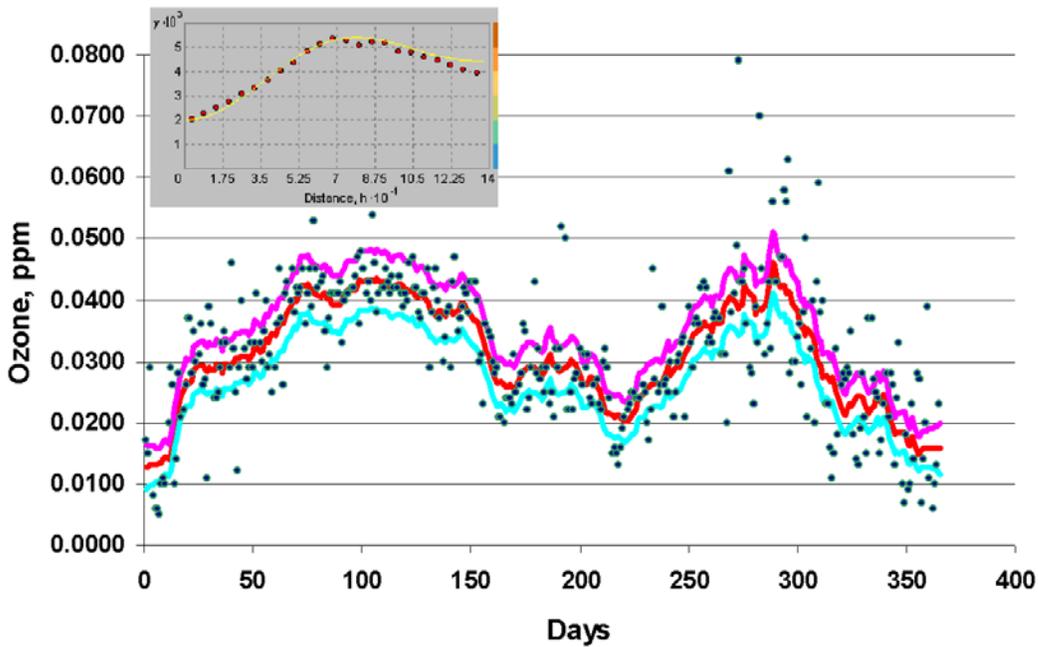
Figure 5. Daily Maximum One-Hour Concentration of Ozone in 1999 in San Francisco, California. In the upper left corner is a semivariogram for the lag of six days.

According to the EPA, the expected annual number of maximum hourly average ozone values exceeding 0.12 ppm must be equal to or less than one. The implementation rule

allows the state no more than three daily, maximum hourly average measurements in excess of 0.12 ppm during three years at each approved monitoring site.

One correct alternative to this rule is the probability for each day of exceeding the threshold. Figure 6 presents an ordinary transGaussian kriging estimation of the probability that the 0.12 ppm level was exceeded in Redlands in 1987 and in 1999, shown as red crests. We used a power transformation with the power value equal to 0.5. Semivariograms were estimated using averaging by a time lag of one day.

(a)   (b)
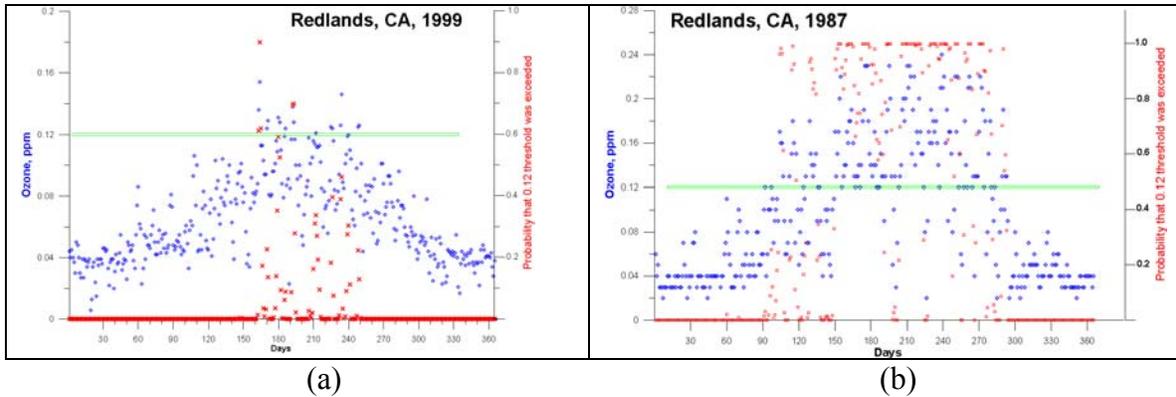
Figure 6. Daily Maximum One-Hour Concentration of Ozone (blue, left *y*-axis) and the Probability (red, right *y*-axis) That 0.12 was Exceeded in 1999 (a) and 1987 (b).

On only one day in 1999 was there a probability greater than 0.9 that the ozone level would exceed 0.12. The situation was much worse in the summer of 1987, when the ozone level was above the threshold approximately every other day.

The goal of ecological analysis is to estimate the risk of health hazard to the people who live in the contaminated territory. Dose, the parameter used to establish safety rules and regulations, is either proportional to the area under the prediction lines in figures 1, 3, and 5 above,

$$Dose = function~(time) \times Area$$

or it is a function of that area,

$$Dose = function~(time) \times function~(Area)$$

Rather than use a single measurement, which can contain errors, it is better to integrate exposure. The proportional coefficient *function (time)* can, for instance, be made equal to the proportion of the days the population at risk spent out doors.

Indoor exposure can be even higher than outdoor exposure. If indoor exposure is not negligible, it should be estimated separately and added to the outdoor exposure.

2. Mapping SO$_2$ (*focusing on weak spatial autocorrelation*)

Sulphur dioxide (SO$_2$) is a colorless, nonflammable gas. The sources of sulphur dioxide include fossil fuel combustion, smelting, manufacture of sulphuric acid, conversion of wood pulp to paper, incineration of refuse, and production of elemental sulphur. Coal

burning is the single, largest man-made source of sulphur dioxide, accounting for about 50 percent of annual global emissions, with oil burning accounting for a further 25-30 percent. A natural source of sulphur dioxide is volcanoes. SO_2 has a penetrating odor that irritates the eyes and air passages. Even moderate concentrations of sulphur dioxide may result in health problems for asthmatics.

Ambient air pollution is often compared with disease rates, as in a series of articles published on the correlation between such air pollutants as SO_2, O_3, NO_2, PM_{10}, and $PM_{2.5}$ and respiratory symptoms, cancer rates, and some other chronic diseases in nonsmokers such as Seventh-Day Adventists. Results of the studies suggest a relationship between long-term exposure to air pollution and the development of specific chronic diseases including bronchitis, asthma, and cancer. In these case studies, pollution measured at the monitoring stations was interpolated by an inverse distance weighted (IDW) algorithm to Zip Code centroids, and predictions were used to assess the long-term effects of air pollution on various chronic diseases.

We used 90 measurements of the high 24-hour concentration of SO_2 for November 1980, the month when the highest concentration of the chemical was recorded, to produce maps using IDW and kriging interpolators. Figure 7a presents the result of IDW interpolation using a power value of 6.7, derived using the Geostatistical Analyst "Optimize Power Value" option that finds the minimum root-mean-square error of several cross validation exercises using different power values. This map has better cross-validation statistics than those created using the default power value of 2.

The California air quality standard for SO_2 is 0.04 ppm. Brown and red indicate areas with predictions above that standard. Very large, warm-colored areas around a few measurements with the highest values of ozone cause suspicion. Large, blue-colored areas considered safe.

Figure 7b presents a simple kriging interpolation with normal score data transformation. Experimental covariance and a J-Bessel model are presented in the upper right corner.
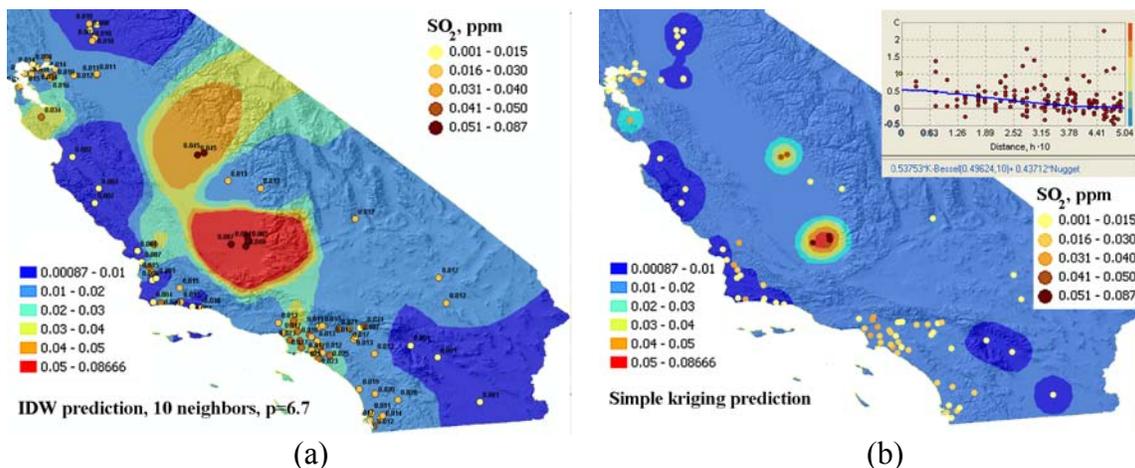
Figure 7. Prediction of Maximum 24-hour Concentration of SO_2 in November 1980 Using IDW (a) and simple Kriging interpolators (b).

Spatial correlation in Figure 7b is weak and smooth, with a large nugget effect and a range of data correlation. Although this map looks more logical and has more accurate cross validation statistics than the map in Figure 7a, we do not think that it can help in epidemiological studies. This is because the weaker the spatial correlation, the more data is needed for meaningful interpolation. The number of SO_2 measurements is insufficient for ecological analysis.

The main difference between interpolation using deterministic and statistical methods is that kriging prediction is accompanied by estimated prediction standard error. Such a map is presented in figure 8a. Areas with few prediction errors, shown in blue, occupy a relatively small part of the territory under study.

The number of SO_2 measurements decreases each year. Figure 8b presents 62 available measurements of high 24-hour concentration of SO_2 in May 1999. The covariance model in the upper right corner of Figure 8b shows very weak spatial correlation. As a result, it is impossible to produce an interpolation map with reasonably few prediction errors using this data.

(a)   (b)
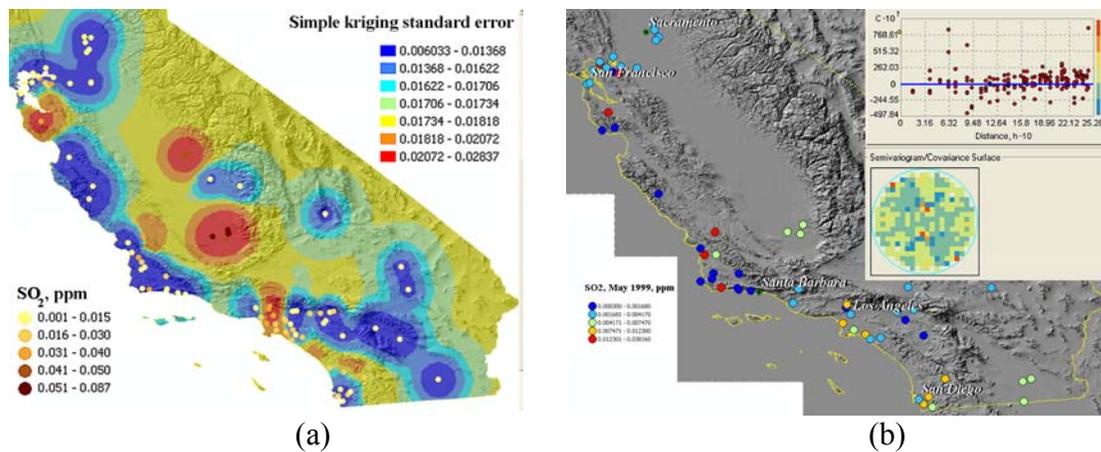
Figure 8 (a)Simple Kriging Prediction Standard Errors of Maximum 24-Hour Concentration of SO_2 in November 1980;(b) Observations of the 24-Hour Maximum Concentration of SO_2 in May 1999. Covariance in the upper right corner indicates that there is no spatial autocorrelation.

Figure 9a presents an ordinary kriging prediction of the maximum 24-hour annual value of SO_2 in 1999 to the centers of Zip Codes in Southern California. Measurement values are printed in black. Figure 9b shows the prediction and 90 percent confidence interval of SO_2 in a randomly selected 10 percent of Zip polygons, based on the assumption that predictions and their standard errors are distributed normally.

(a)   (b)
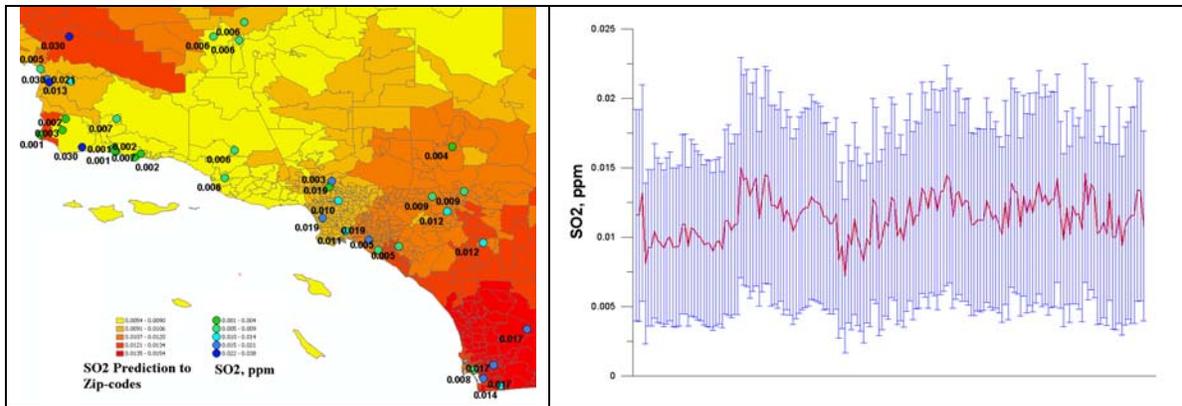
Figure 9 (a) Ordinary Kriging Prediction of SO_2 Maximum Annual Value in 1999 to the centers of Zip Codes; (b) Prediction and Its 90 Percent Confidence Interval of SO_2 in a Randomly Selected 10 Percent of Zip Polygons.

Prediction errors are very large. Reliable interpolation is not possible using available sulfur dioxide data nor is searching for a correlation between exposure from SO_2 and people's health.

3. Probability Mapping

Because predictions are not true values, the uncertainty associated with predictions should be provided. Figures 8a and 10b are examples of prediction standard error maps, quantified by the minimized prediction root-mean-squared error that makes kriging optimum.

In addition to the combination of prediction and prediction standard error maps, two other possibilities of presenting this information on just one map are provided by Geostatistical Analyst; namely, probability and quantile mapping. Probability maps show the degree to which the interpolated values exceed a specified variable's threshold. Quantile maps are particular probability maps in which the thresholds are the quantiles of the predicted distribution.

Figure 10a shows the probability that in 1999 the maximum 24-hour $PM_{2.5}$ value would exceed the California Ambient Air Quality Standard of 65 μg/m^3. We used disjunctive kriging with normal score transformation to create the map. Prediction standard errors are displayed in Figure 10b, together with populated places shown as circles. The redline is the standard error of prediction of 0.3. Blue areas showing prediction errors are large, therefore, decisions about $PM_{2.5}$ should be made there with great care.

(a)   (b)
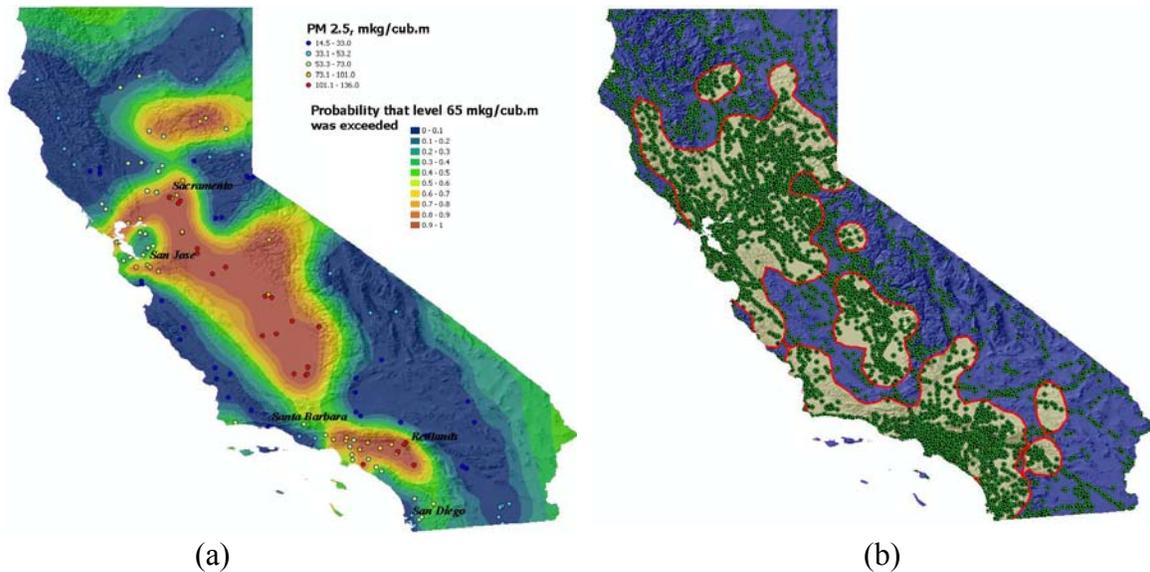
Figure 10 Probability That the $PM_{2.5}$ Value Exceeded 65 μg/m^3 in 1999 (a); Prediction Standard Errors and Populated Places (b). Disjunctive kriging with normal score transformation was used.

Figure 11 shows PM_{10} quantile maps with underestimated (Figure 11a) and overestimated (Figure 11c) predicted mean values. Figure 11b shows neither overestimated nor underestimated prediction of PM_{10}. Often it is better to overestimate contamination than miss areas with high contamination, and a quantile map is the right tool for that.

(a)   (b)   (c)
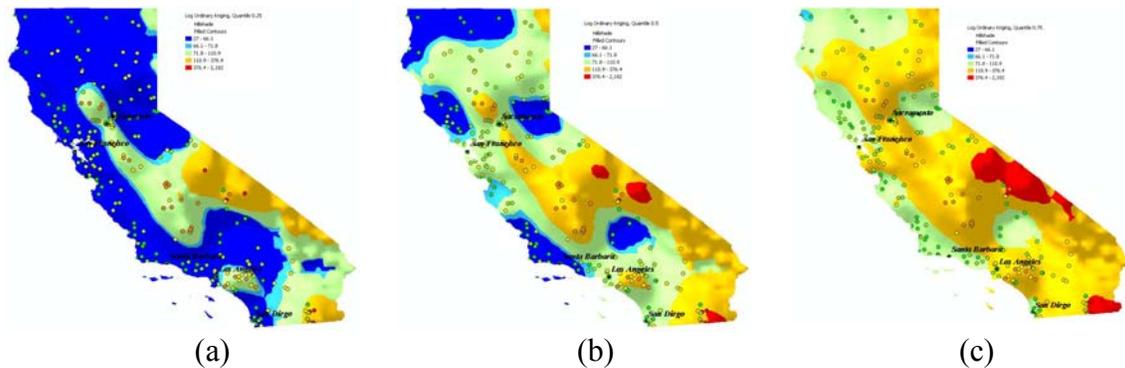
Figure 11. PM_{10} 0.25 (first quartile) (a), 0.5 (median) (b), and 0.75 (third quartile) Quantile (c) Maps, Created Using Lognormal ordinary Kriging.

4. Using Additional Variables to Improve Predictions

For data interpolation of air pollution in California, we have a limited number of measurements of several air pollution variables and detailed information on geographic variables such as elevation, distance to the ocean, and distance to the road (this variable is useful because cars are major sources of air contamination in most parts of California). Among the possible approaches to interpolate multivariate data is cokriging, which combines spatial data on several variables to make a single map of one of the variables. It

is appealing to use information from other variables to help make predictions, but it comes at a price. Cokriging requires more estimations than kriging including estimating the autocorrelation for each variable as well as all the cross correlations.

Cross correlation between variables can be very informative. Figures 12a and 12b show directional semivariogram clouds and models of ozone and NO_2, and figure 12c displays the directional cross covariance cloud and cross covariance model between maximum annual one-hour concentrations of ozone and NO_2 in 1996. The cross correlation between NO_2 and ozone is asymmetric; the highest cross covariance occurs for ozone and NO_2 at different locations. The highest correlation occurs when taking NO_2 values that are shifted to the west of the ozone values. The shift finds the distance and direction where the calculated cross covariance is at its maximum value. This adds two parameters to the cross covariance model to describe the shift in the x- and y-coordinates.

(a)   (b)   (c)
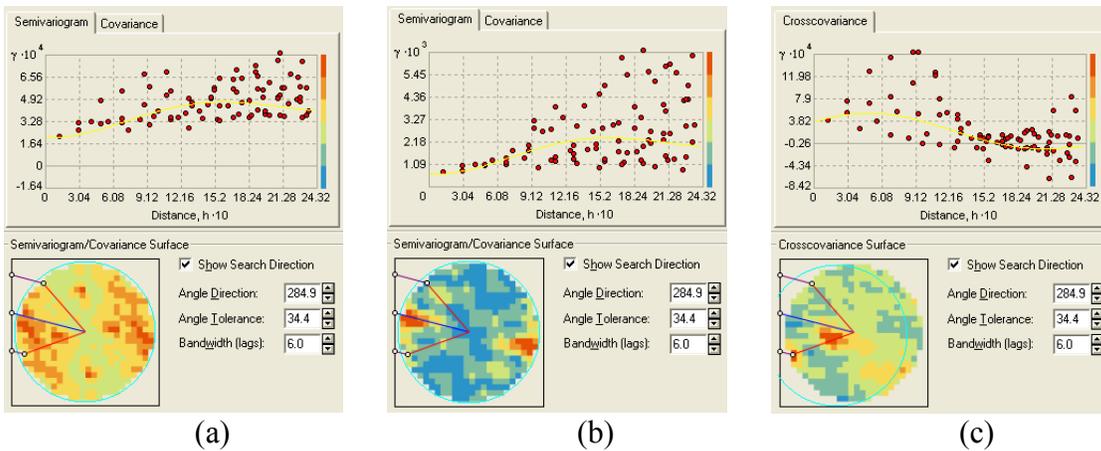
Figure 12 Directional Semivariogram of Ozone (a). Directional Semivariogram of Nitrogen Dioxide (b). Directional Cross Covariance of Ozone and Nitrogen Dioxide (c)

Ozone and nitrogen dioxide maximum values shift because ozone is produced as a result of chemical reactions between pollutants. That means ozone reaches maximum concentration several hours after NO_2. During this time, pollutants shift to the east because, typically, the wind direction in summer is from the ocean.

Figure 13a presents the result of ozone prediction using a cokriging model, with ozone as the primary variable and a grid of distances from major California roads as the secondary variable. Major roads are displayed at the top layer of the map. Cross validation statistics show that using distance from a road as a secondary variable improves the prediction of ozone pollution. Figure 13b shows the cross covariance cloud and the exponential model used to create the map in Figure 13a. The largest correlation occurs at the nonzero distance between the monitoring stations and the data on the grid. Cross correlation is anisotropical and shifted, so it takes some time to find the optimal cross covariance model in this situation.
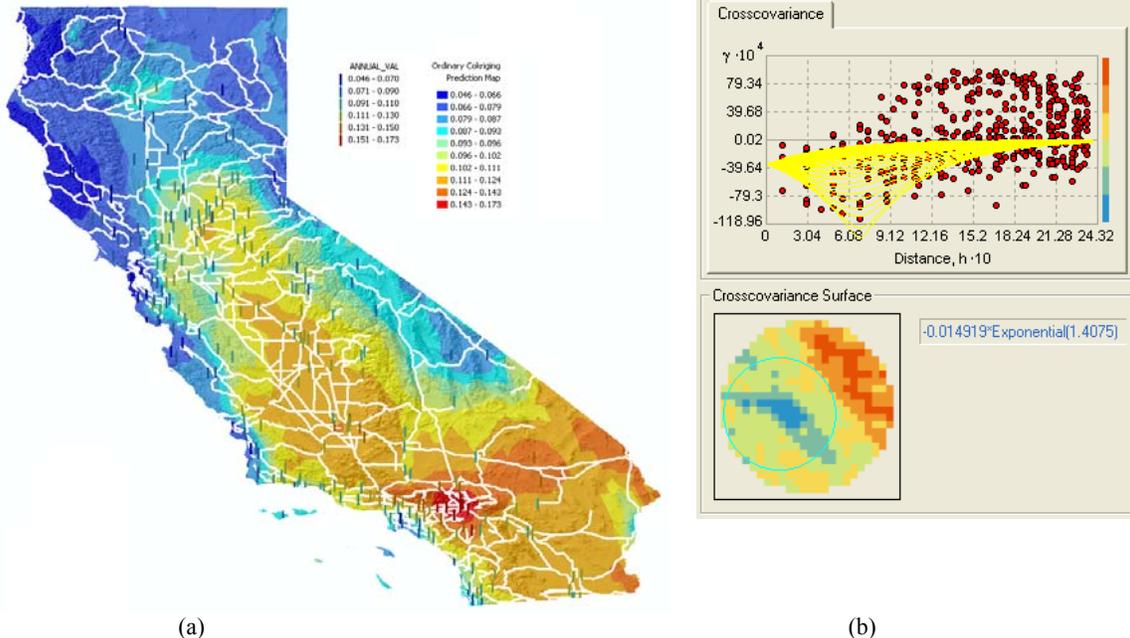
(a)   (b)

Figure 13 Ordinary Cokriging Prediction of the Maximum One-hour Annual Value of Ozone in California in 1999 Using Distance to the Road from the Monitoring Stations as the Secondary Variable (a). Cross covariance Between Ozone and Distance from the Major Roads (c)

5. Space-Time Data Analysis Using 2D Geostatistics

The current version of Geostatistical Analyst cannot work with 3D or space-time data. However, with a little creativity, some space-time data can be analyzed and visualized. Two examples of such analysis are presented in Figure 14.

Figure 14a shows the result of predictions using the estimated coefficients of a linear regression trend in maximum one-hour July values of ozone from 1980 to 1999 in 105 monitoring stations. Ozone tends to decrease in the southern and middle parts of California and increase in the north. Fortunately, the level of ozone there is much lower (see Figure 13a).

Maximum ozone concentration each July from 1980 to 1999 was used to create prediction maps. The result of the modeling was exported to grids and displayed using animation in ArcGIS 3D Analyst (currently at \\terrytate\ProductVideos\Geostatistical Analyst\CA ozone concentration from 1980 to 1999). The tendency of decreasing ozone concentration is clear here as well.

Figure 14b shows a surface prepared using kriging based on monthly data of PM_{10} in the city of Riverside, using years as the *x*-axis, months as *y*-axis, and maximum annual 24 hours as the data value. The result of predictions was visualized using 3D Analyst. We can see again that pollutant concentration in one of the more contaminated cities of Southern California has tended to decrease over time.
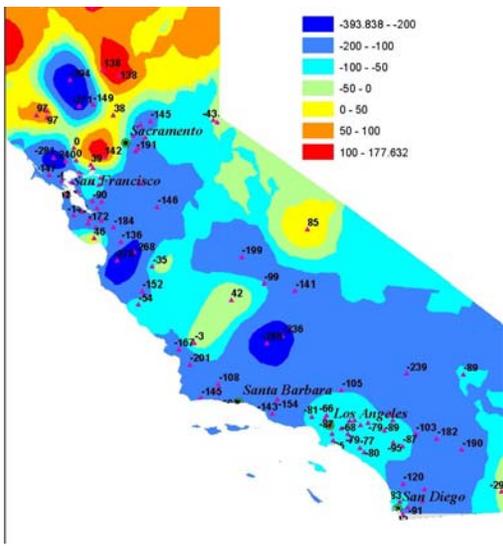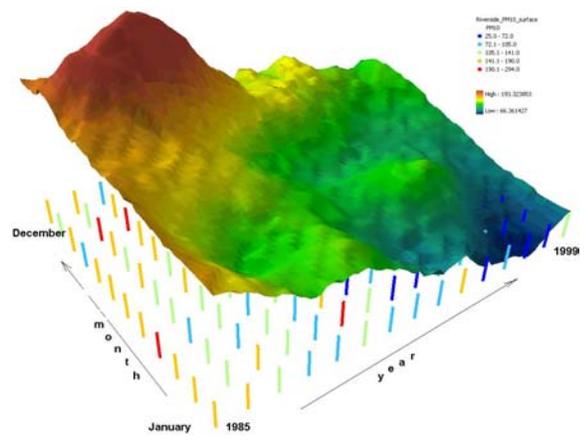
(a)   (b)

Figure 14 Interpolation of Ozone Trends in 1980-1999 (a). Negative values in cool colors represent areas where contamination decreases and positive values in warm colors where contamination increases (a). Ordinary kriging surface based on monthly measurements of PM_{10} in the city of Riverside (b)