CHAPTER 1

STATISTICAL APPROACH TO GIS DATA ANALYSIS

This chapter begins with a discussion of the difference between deterministic and statistical spatial data analysis. Definitions are given, and an example of modeling radiocesium soil contamination data with measurement errors is presented.

Spatial statistical models are classified into geostatistical, regional, and point pattern models. They correspond to the data classified by their *locations* into continuous, regional, and discrete types. There are four types of data *values* that can be associated with any spatial statistical model: continuous, ordinal, categorical, and binary. Continuous means that data values come from the real number line. Ordinal means that they are in discrete categories, and the order of those categories is meaningful; for example, small, medium, and large. Categorical means that data values are in discrete categories, and order is not meaningful; for example, red, yellow, and green. Binary data take either the value 0 or 1. In this book, we usually use the word “continuous” when describing data locations, not data values.

The differences in modeling assumptions for each data type are briefly discussed, and an example is given of data (lightning strike locations) that can be modeled using any of the three types of spatial statistical models.

Spatial data dependency is a common feature of any spatial statistical model. The importance of modeling the extent of spatial dependency is illustrated using simulated data.

Spatial data dependency is based on the distance between spatial objects, but distance is not necessarily related to the length of a line connecting any two objects. Several examples illustrate the problem of selecting the appropriate measure of distance between spatial objects.

The chapter ends with an example of regional data visualization problems using family size data in counties of the United States.

An addendum to the chapter illustrates how spatial dependence is estimated and used in geostatistics for prediction to the unsampled locations.

Many readers begin a book at chapter 1, skipping the preface. We suggest reading the preface before this chapter because it provides additional motivation for studying and using spatial statistics.

SPATIAL ANALYSIS AND SPATIAL DATA ANALYSIS

The use of GIS functionality to understand and interpret spatial information is called *spatial analysis*. GIS tools are powerful, and the results they return are often so visually impressive that it is easy to lose sight of something that everyone knows at some level, namely, that most datasets have errors both in attributes and in locations, and that processing data—with overlay, buffering, and the like—propagates errors. Much of the current progress in GIS consists of providing open and consistent database support and making it easier to construct maps. If there is a need to test hypotheses and to make predictions in fields such as agriculture, epidemiology, or hydrology, it is also necessary to be informed about the uncertainty of these tests and predictions in order to fully understand the decisions that can be made using the results of the analysis.

Spatial data analysis uses statistical theory and software to analyze data with location coordinates. Nonstatistical models (often called deterministic) postulate data relationships by the imposition of *a priori* models, whereas statistical models estimate the model parameters from the data at hand. For example, inverse distance weighted (IDW) interpolation is a deterministic model because the dependence between pairs of values is defined simply by the distance between data locations raised to a predefined power value. This model has been applied to fields as diverse as air pollution, meteorology, and the study of plant nutrient distributions in soils. The main shortcomings of deterministic models are the arbitrariness of the postulated spatial data similarity and the absence of information on their prediction uncertainties.

Kriging is a statistical model because it infers the dependence between pairs of points from examination of all similarly distant pairs in the data. This allows calculation of averages and their associated properties along with an estimate of the uncertainty of the prediction. The parameters of the kriging model are different for any two different datasets.

A GIS provides (1) spatial data management, (2) mapping, (3) spatial data analysis, and (4) decision-making assistance. A GIS has outstanding tools for the first two and parts of the latter two. ArcGIS Geostatistical Analyst is a spatial data analysis program, an extension to Esri ArcGIS software developed to integrate GIS data management and mapping capabilities with statistical modeling of spatially continuous data such as temperature and elevation—data for which Tobler's first law of geography applies: "Everything is related to everything else, but near things are more related than distant things." Incorporating spatial statistical modeling into the GIS

environment is natural because GIS tools describe things over distance, while statistics can explain how these same things change and interrelate. With the addition of statistical methods, GIS is enhanced with key tools for decision support and evaluating the uncertainty associated with those decisions.

Statistics is not the only tool for decision making, but much of statistics is concerned with quantifying uncertainty. Making decisions requires managing risk, and this, in turn, requires understanding uncertainty. Integration of 1 through 4 above provides scientists and managers with an inclusive environment in which data are managed, models are constructed, and prediction uncertainty is made explicit.

In Belarus, after the Chernobyl accident in 1986, decisions had to be made quickly about whom to evacuate. The maximum permissible exposure was considered one millisievert per year. The amount of exposure is not important for our purposes here, only that there is a measurable threshold. The decision to evacuate had to proceed in the face of two types of uncertainty: uncertainty about the quality of the data and uncertainty about the model of the various processes through which the data was run to get results.

In the years since the accident, territory zoning has been based on the available experimental data:

- Direct measurements of the dose accumulated in a person, or
- If the individual dose is unavailable, dose estimation from radioactive food contamination, or
- If irradiation from eating contaminated food cannot be calculated, the level of radioactive contamination of soil with cesium-137, strontium-90, and actinoids

Decision making based on individual doses is the easiest because the data uncertainty is small.

Several elements go into the calculation of dose: contamination of food and characteristics of a person's daily diet, age, and body weight. Each of these is necessarily an average. Not every person weighs the same, nor does everyone eat the same amount of the same food. Contamination varies from place to place and from foodstuff to foodstuff. Calculations (see chapter 3) show that uncertainty of the dose based on the data collected in one of the villages is 47 percent, uncertainty of intake (average daily diet plus average contamination) is 42 percent, and uncertainty of weight is 21 percent. Such a great amount of uncertainty makes identifying which areas to evacuate a difficult task. It is no small effort (or expense) to evacuate people, and there is significant risk in leaving people in dangerous areas.

In practice, threshold values of soil contamination were used in zoning instead of the dose measuring or calculation, since soil samples are much easier to collect than doses are to measure or calculate. The adopted thresholds are the calculated effective values of soil contamination that are believed to correspond to some critical values of the individual dose. Although measurements of the radionuclide soil contamination are considered accurate, the uncertainty of such territory zoning for the purpose of the dose control is large. The major source of uncertainty is concealed in the coefficient of correlation between the dose and the soil contamination. Another component of uncertainty is that of the measurements of soil contamination themselves.

Filtered prediction is one of many methods of measuring data quality and accounting for uncertainties. This method improves decisions in one area by using information from other areas. The map in figure 1.1 shows the locations of settlements close to Chernobyl, along with numbers representing each settlement's level of radiocesium soil contamination in curies per square kilometer (Ci/km^2). The upper permissible limit is 15. The error of radioactive soil contamination measurements is about 20 percent, and two circled settlements are close to the upper permissible level limit. When information from other settlements is used to predict the level of contamination at the two circled settlements, the one measured at 14.56 Ci/km^2 is predicted to be 17.05 Ci/km^2, and the one marked 15.45 is predicted to be 16.88. According to the kriging model with a measurement error

component, their true contamination values are in the intervals 14.2–19.9 and 13.8–20.0. Thus, both settlements are rather unsafe, and people living in and around them should probably be evacuated.

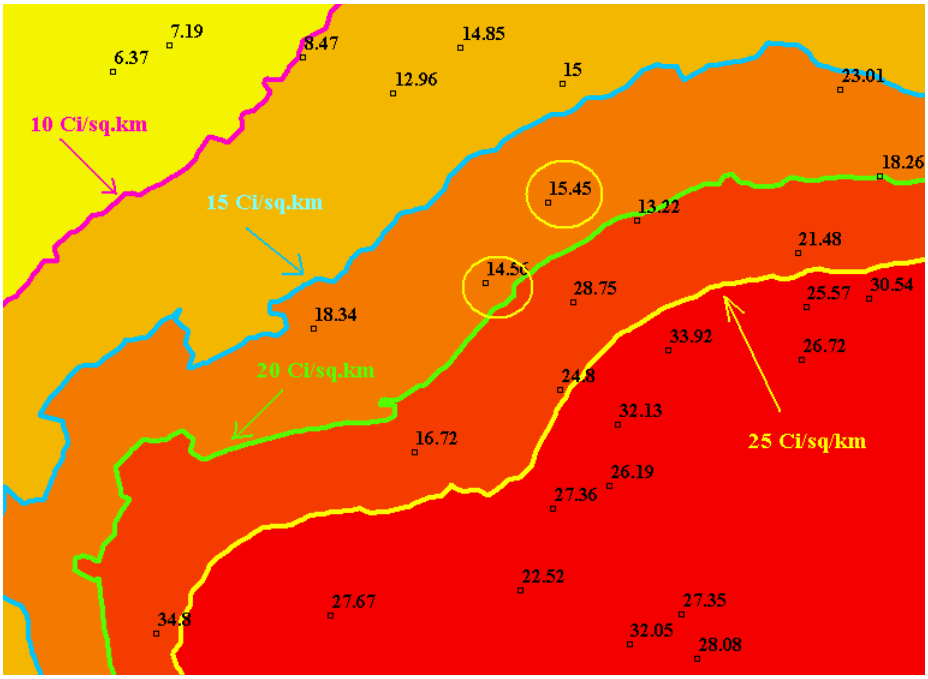


Figure 1.1 Courtesy of International Sakharov Environmental University, Minsk, Belarus; Institution of Radiation Safety, Belrad, Republic of Belarus.

In the Chernobyl example, it is evident that the data were uncertain, both in the accuracy of the measurements themselves and in the prediction of the values at the unsampled locations. Similarly, much spatial data contains errors in both attribute values and locations because extremely precise measurements are very costly or impractical to obtain. Although this should not limit the use of the data for decision making, neither should it give the decision maker the false impression that maps based on such data are exact.

In summary, a model is an approximation of reality, both incomplete and imperfect. The data from which models are made is a sampling of reality, also incomplete and imperfect. To use models intelligently, their limitations must be recognized. Where is the error? How much of it is there? How much is too much? Only with answers to these questions can uncertain data be used and models incorporating that uncertainty be created and understood to make sensible and reliable decisions.

The statistical methods covered in this book provide objective tools for measuring data quality, accounting for the inevitable errors and uncertainties involved in mapping, assessing spatial patterns, and even deciding whether what seems to be a pattern is in fact a pattern. Humans are good at seeing patterns, even where there are not any. Just think how easy it is for people to see faces everywhere—in rocks, in clouds, in the moon.

TYPES OF SPATIAL DATA AND RELATED STATISTICAL MODELS

There are three broad groups of spatial data classified according to their locations: discrete, continuous, and regional. All are illustrated in figure 1.2.

Discrete data are isolated on the earth's surface, for example, tree locations, which cannot be found at every point. Continuous data exist at every point (e.g., elevation), but they are measured in a finite number of locations. Regional data are taken collectively or in summary form; for example, the incidence of a particular cancer in the United States.

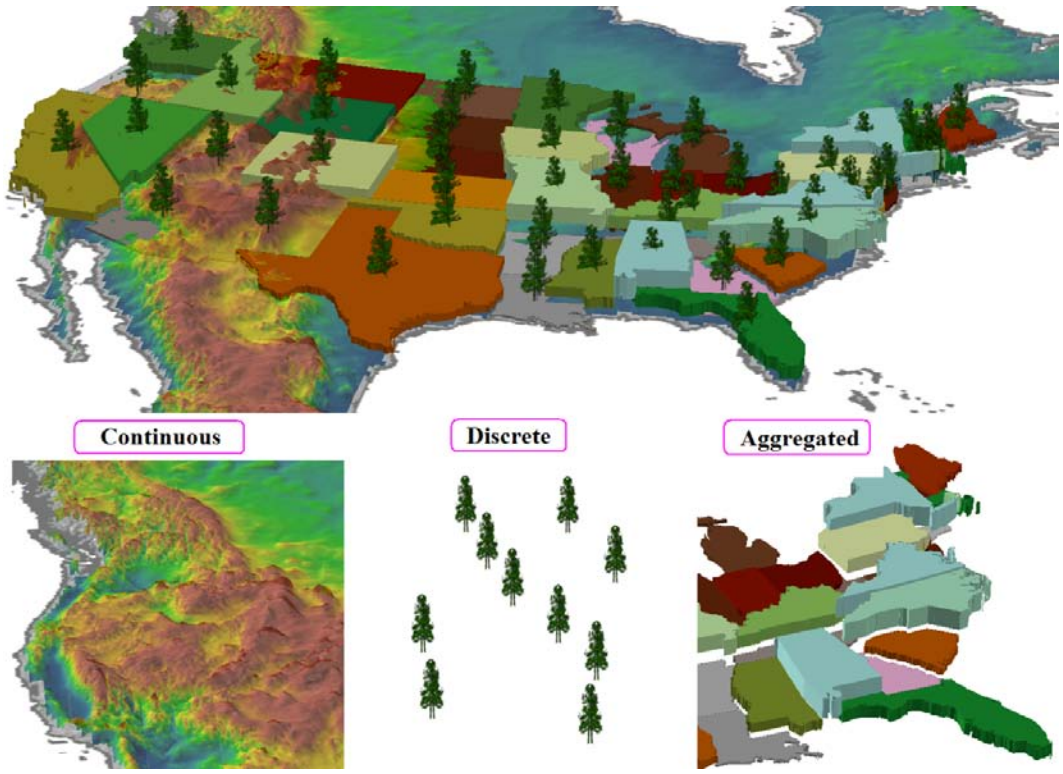


Figure 1.2

Statistical models for discrete, continuous, and regional data are different. In statistical literature, modeling of discrete point data is called point pattern analysis, modeling of spatially continuous data is called geostatistics, and modeling of regional data is called lattice analysis. Applying tools developed for continuous data to data that are discrete or regional will produce anomalous results. This may not be immediately obvious, for nothing in those results will necessarily announce that there are anomalies.

For example, locations of lightning strikes on one particular day are displayed in figure 1.3. The red circles reflect a radius of uncertainty in the location of the strikes. In addition to the approximate data location, there is information on the polarity and strength of the lightning strikes. (Positive polarity is more likely to ignite a wildfire.)

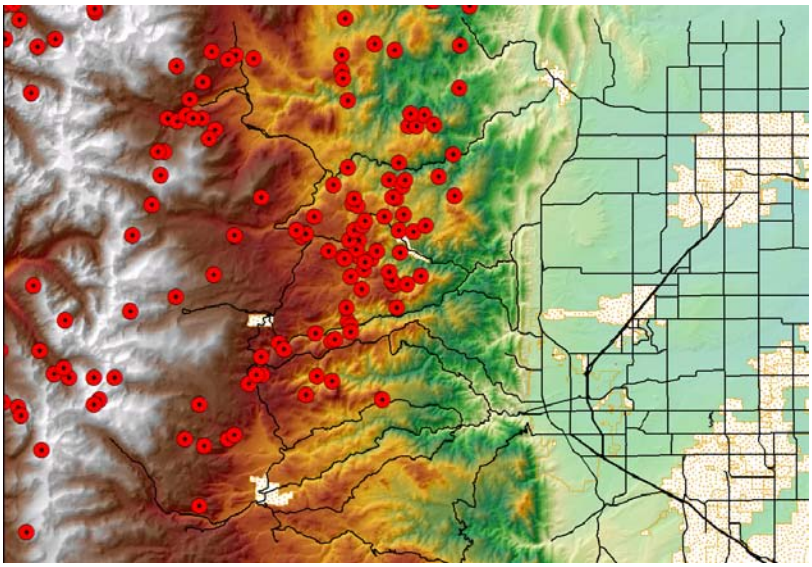


Figure 1.3

A continuous map of positive-polarity lightning strikes using a statistical interpolation method (kriging) can be created and used to indicate the risk of a wildfire. However, this approach does not account for the spatial distribution of the lightning strikes. There may be far more strikes with negative polarity, with increased fire risk simply because of the number of strikes. More important, prediction of the polarity and strength between observation locations is questionable because lightning strikes did not happen between actual strike locations.

To account for errors in identifying the locations of lightning strikes, polygons can be defined that partition the data according to soil and forest types, then the number of lightning strikes can be counted and polarity distribution estimated in the polygons. Regional data analysis on such aggregated data would be challenging because this requires sophisticated methods for the creation of polygons and their neighborhood selection.

In spatial statistics, the lightning strike data would be considered as the realization of a marked point pattern process, which is a combination of two processes. The first process locates the lightning strikes (a discrete point process), while a second process controls the strength and polarity associated with strikes at the recorded locations (a continuous process).

Statistical models for discrete, continuous, and regional data can be employed and visually appealing maps created with the lightning strike data, which are discrete by nature. The results of modeling will be different but equally appealing, although most likely erroneous, if the geostatistical method is used.

The analysis of the spatial point pattern usually starts with finding the most similar pattern among those that can be generated with known statistical features. A typical goal of statistical analysis of discrete points is estimation of the points' intensity at each location in the study area. The intensity surface can be a function of covariates. For example, the density of plants can be a function of the soil properties.

A model of point pattern intensity can be developed using spatial trend, dependence on spatial covariates, and interaction between points of the pattern. Models called marked point processes allow dependence between point locations and their associated values. For example, tree location and diameter at breast height could interact because trees compete for light and nutrients.

Geostatistical models (kriging) predict values where no measurements have been taken assuming that data have similar statistical properties in any part of the study area. This assumption is called stationarity. More precisely, the mean and variance of a variable at one location should be equal to the variable's mean and variance at other locations, and the correlation between data at any two locations should depend only on the direction and magnitude of the vector that separates them and not on their exact locations. Stationarity is a property of the kriging model; it is not necessarily a feature of the phenomena under study.

In kriging, data are assumed to be stationary even if there is uncertainty as to what "stationary" means. If conventional kriging is used on nonstationary data, such as disease rates (the incidence of disease observed in a geographical area per 1,000 or 100,000 persons) in which the data variability (variance) of the rate depends on the population, which varies with geographical region, conventional kriging will not return results that can be used safely to make decisions because the fundamental model assumption has been violated, and the accuracy of the analysis cannot be trusted.

Statistical models always use some assumptions, and if these assumptions are not fulfilled, decision making based on the results of the modeling may be incorrect. Once it is known what assumptions are made by the model, the appropriate data can be matched to it or altered to fit the assumptions. This can be done, for example, by removing large-scale data variation and transforming data to approximate a theoretical data distribution such as Gaussian.

For instance, the popular indexes of spatial data association, Moran's *I* and Geary's *c,* are widely used for analyzing rates. However, these indices were derived under the assumption that data mean and variance are constants, a condition that hardly applies for rates. Therefore, other measures that more explicitly take into account the aggregated nature of the data are to be preferred.

In contrast to continuous data, regional data are usually observed at every location (that is at every polygon) and there is usually no need in interpolation. Researchers working with regional data are typically interested in a way of validly presenting a map of counts or rates. One problem with this is that regions with smaller populations may have an abnormally high rate simply because there are fewer people in the region. Other typical goals of regional data analysis are finding significant clusters, homogeneous groups of regions that vary in similar manner, and variables that vary together with the variables of interest (covariates).

Statistical models for continuous, regional, and discrete data are discussed in chapters 8–13.

SPATIAL DATA DEPENDENCY

The applicability of spatial statistics demands that the data be spatially dependent, at least to a certain extent. If data are spatially independent, there is no way to predict a data value at an unsampled location with reasonable accuracy. If there is data dependency, data contain information about the relationship between values at nearby locations. As the dependence on data grows stronger, the prediction uncertainty lessens, and the least amount of data will be required for reasonable prediction. In figure 1.4, it would not be difficult to predict the value of one or several blue points if they were removed, but prediction of a removed green point is more difficult. The best that could be done would be to predict the removed green point within the area bounded by the red lines.

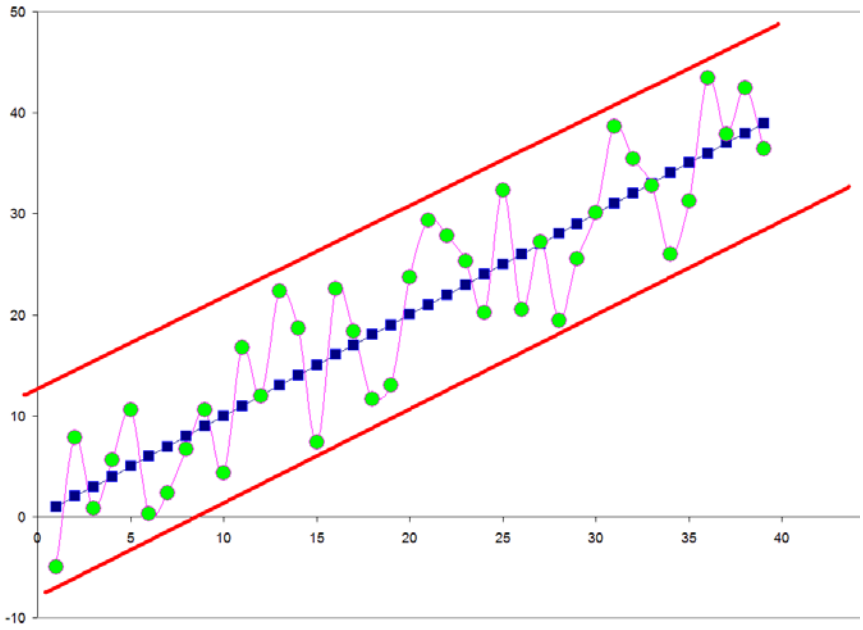


Figure 1.4

The importance of spatial data dependency in the case of spatial interpolation can be illustrated by using deterministic and statistical models implemented in Geostatistical Analyst.

Fifty data points are displayed in figure 1.5 (left). Using this data, a surface was created using an inverse distance weighted interpolation (IDW) with the Geostatistical Analyst's default parameters. The predictions are plotted in figure 1.5 (right).

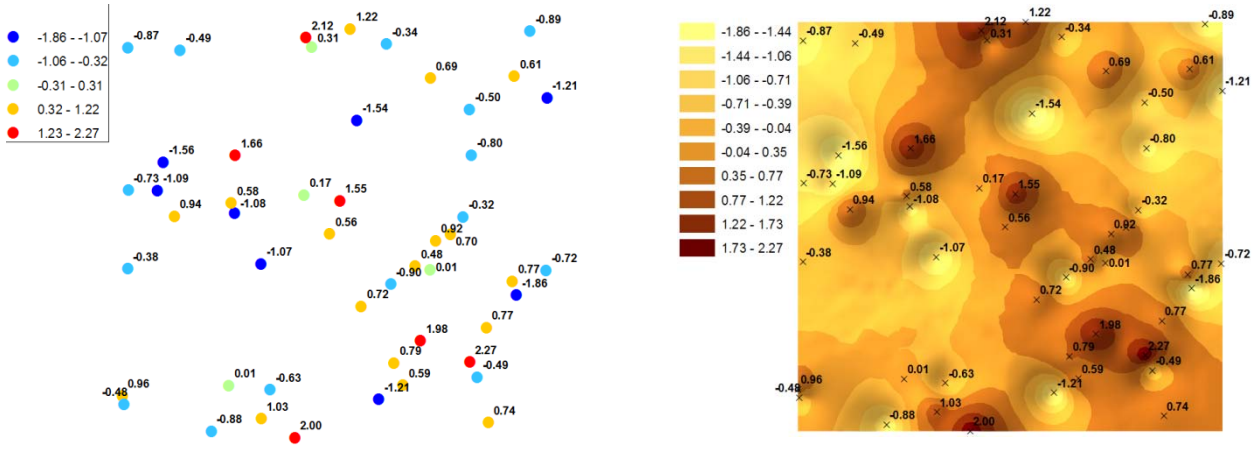


Figure 1.5

Using the same data, another surface was created using a radial basis function with a thin-plate spline kernel (figure 1.6 at left). The surfaces differ, although not greatly. Each surface is the product of different assumptions about the relation of data points to each other. These assumptions are not connected in any fundamental way to characteristics abstracted from the dataset but are predefined. Thus, the models are deterministic.

However, the default statistical model, ordinary kriging (figure 1.6 at right), applied to the same data produces a very different surface—a noisy, flat surface—unlike the chain of valleys and hills produced by the deterministic inverse distance weighted interpolation and the radial basis function models. Is this surface better in some way?

The relationship between points on the surface produced by ordinary kriging is defined by characteristics drawn from the data. All prediction surfaces are not absolutely accurate, but error in the surface produced by ordinary kriging can be evaluated.

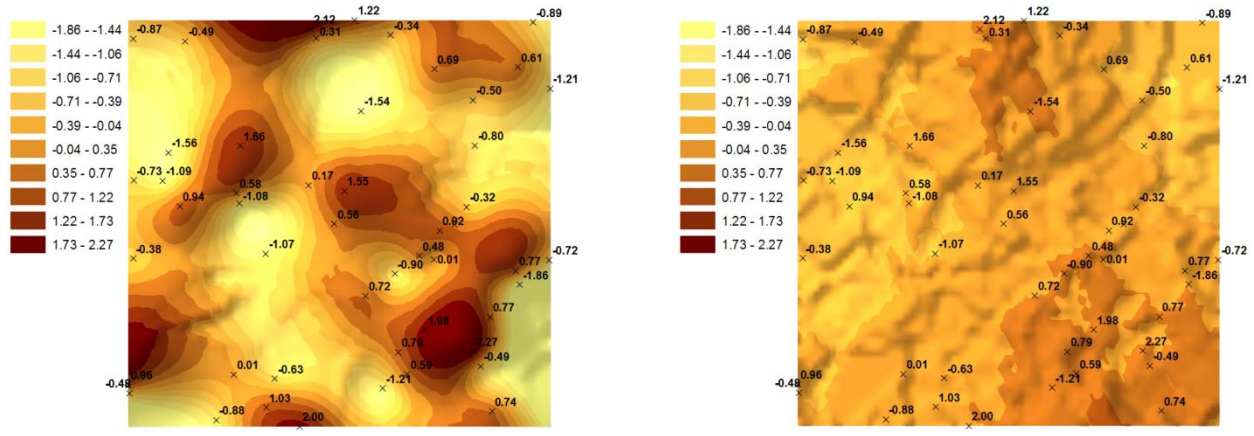


Figure 1.6

The cross-validation diagnostics in figure 1.7 compare results for ordinary kriging and IDW. Ordinary kriging predictions are preferred over predictions from the IDW model because the prediction mean and root-mean-square errors are smaller for kriging than for IDW, as is the spread of points around the fitted line.

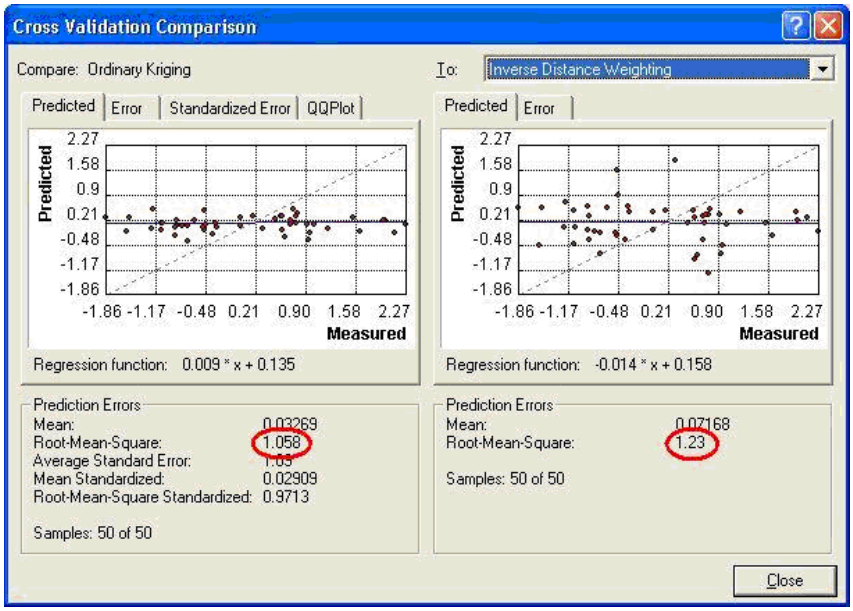


Figure 1.7

The smooth-looking surfaces created using deterministic interpolation models are less accurate than the rough surface produced by ordinary kriging, because the data used in this example are spatially independent. Spatially independent data can be produced by a random number generator (as these were), and the points are unrelated to each other in the sense that near datum is not more alike than data separated by greater distances. Because the data do not contain information about similarity with their neighbors, interpolation is not appropriate. Although none of the models can predict values removed from the dataset, kriging produces a surface close to the average value of the data. This is the best prediction surface when data are spatially independent.

Real data are spatially dependent more often than not. If the elevation is 500 feet in one place and 600 feet in another, there must be at least one location somewhere in between with an elevation of 550 feet. Spatial data dependency is not simply present or absent; it occurs in varying degrees. Spatial statistics provide methods to define the extent of spatial data dependency for use in modeling.

DISTANCE BETWEEN GEOGRAPHIC OBJECTS

Distance between geographic objects is an important attribute of models in spatial statistics. Specification of distances between points will influence the conclusions drawn from data. The terrain that forms the foundation of the spatial effects under investigation is not a flat surface. Spatial objects are frequently separated by barriers that may be natural, political, or conceptual as shown in figure 1.8. Therefore, distance in spatial statistics is a more complicated concept than simply the length of a line connecting two objects.

Figure 1.8

Straight-line distance may not be the best measure of the proximity between geographic objects. In figure 1.9, lines connect locations into pairs that are all about an equal distance apart. Add the geographic layers in figure 1.9 (right), and it is easy to see that many pairs cross bodies of water. Therefore, if the points represent animals that do not swim, then the lines that cross water do not represent true shortest distances between those points for those animals.

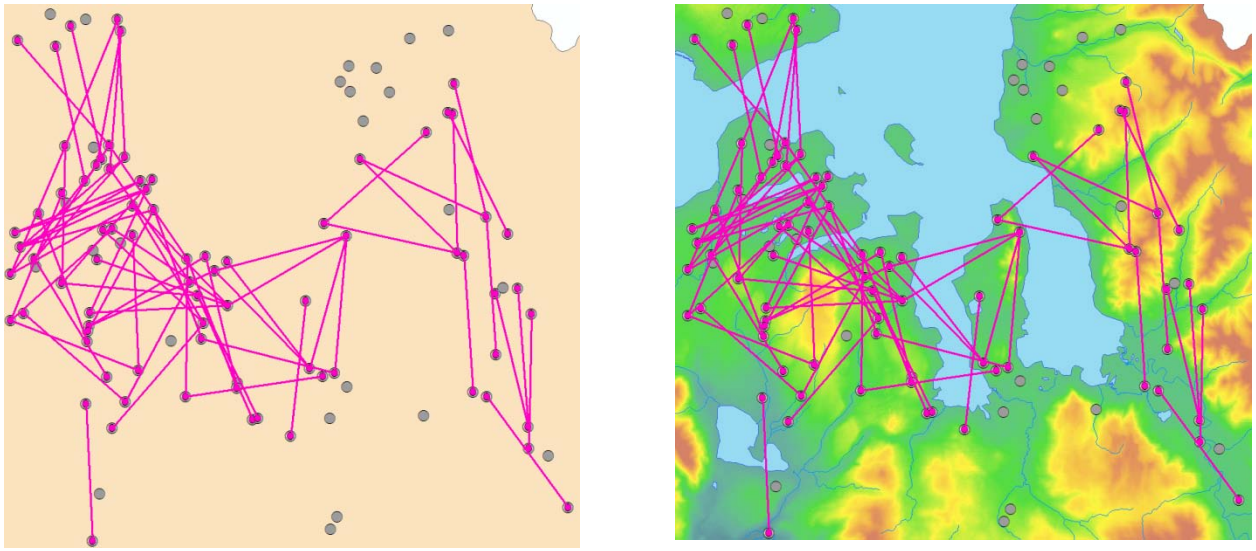


Figure 1.9

Figure 1.10 at left shows the straight-line (Euclidean) distance to the closest measurement point for each cell of the 285-by-253 grid with a cell size of 180 meters. This map of distances can be compared to a distance map in figure 1.10 at right, which is the least accumulative cost distance over a cost surface calculated as a sum of three grids:

1. *Cost value equals 1 + elevation/(lowest elevation value).*
2. *Cost value equals 5 if cell intersects stream and 0 otherwise.*
3. *Cost value equals 20 if cell intersects lake and 0 otherwise.*

The points' density will be very different if different distance metrics are used (see example in chapter 13).

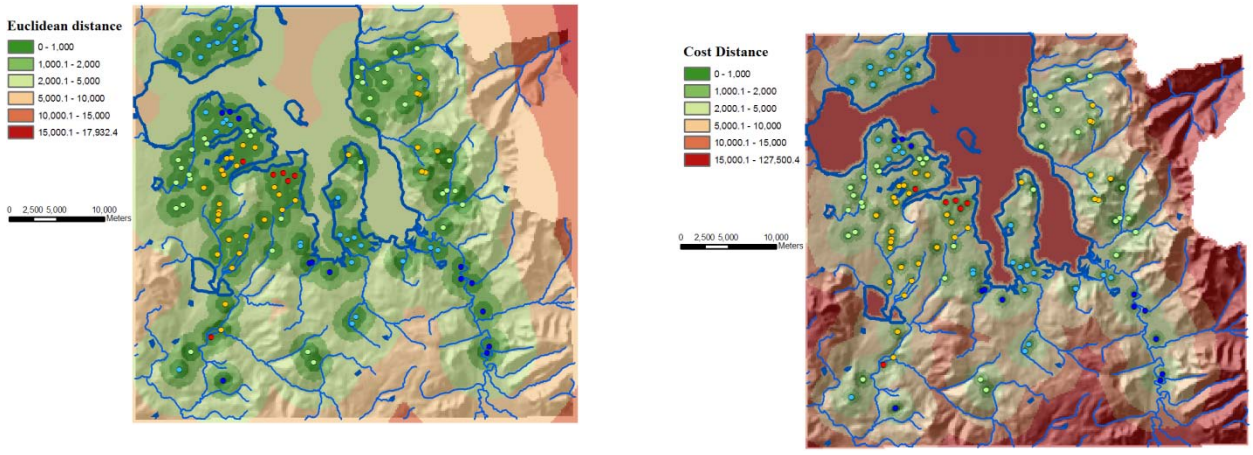
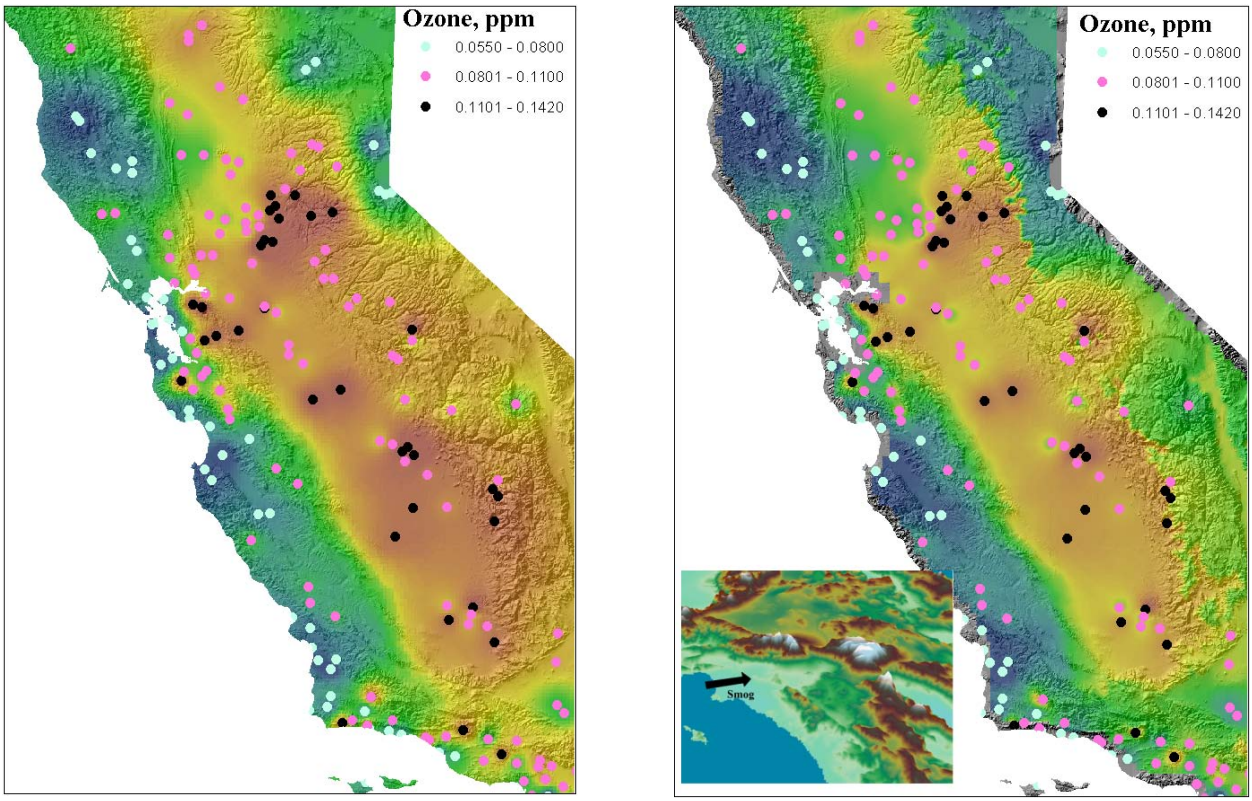


Figure 1.10

Figure 1.11 shows different interpolations of ozone concentration. Figure 1.11 at left, which shows the predicted concentration of ozone in California, neglects to account for the effects of mountains. But since mountains are as effective a barrier to industrial ozone as water is to nonswimmers, taking the effect of the mountains into

consideration gives a more realistic picture of ozone distribution, as shown in figure 1.11 at right, which depicts it traveling up gorges, for instance.



From California Ambient Air Quality Data CD, 1980-2000, December 2000.
Courtesy of The California Air Resources Board, Planning and Technical Support Division, Air Quality Data Branch.

Figure 1.11

The distance between spatial objects with nonzero areas cannot be defined uniquely. Figure 1.12 shows zip

polygons in the Bay Area of Northern California. There are infinitely many ways to specify distances between polygons, which, for instance, can be based on the travel time between randomly selected points inside polygons.

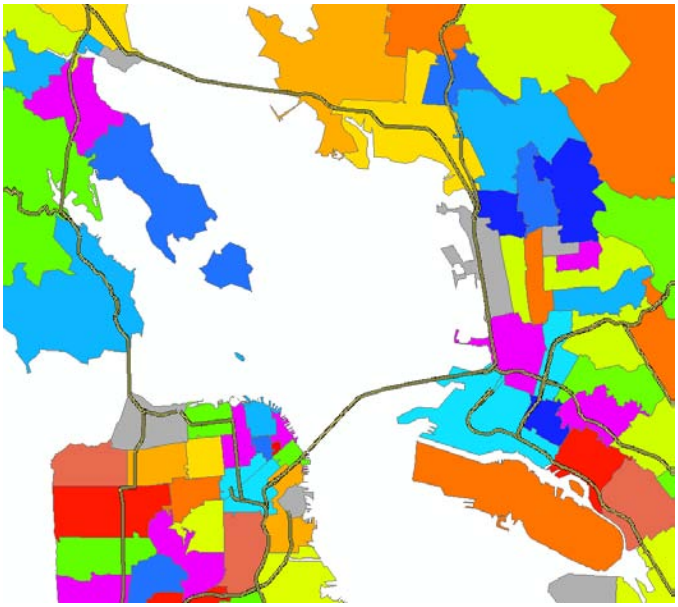


Figure 1.12

AN EXAMPLE OF STATISTICAL SMOOTHING OF REGIONAL DATA

The ease of spatial visualization provided by GIS entices users to rely on visual comparison between maps for reaching conclusions. For example, an impression that frequently arises in disease mapping is the apparent clustering of disease cases on a map. Apparent clustering tempts the researcher to look for an environmental cause for the disease outbreak, while this clustering may simply be caused by mapping the raw number of disease cases instead of rates (the number of cases per capita).

It is also easy to produce misleading impressions from data points just by classifying them for mapping and manipulating the color palette of the map. Figure 1.13 shows the average family size in United States counties using two widely used classification schemes: natural breaks (top) and quantiles (bottom). The bottom map suggests that most large families prefer to live near the borders, while the top map suggests a more homogeneous distribution.

The population in the bottom map appears to be higher, too. Which map is correct? Which map more closely corresponds to reality, and how can that be determined? Answering these questions requires understanding the data.

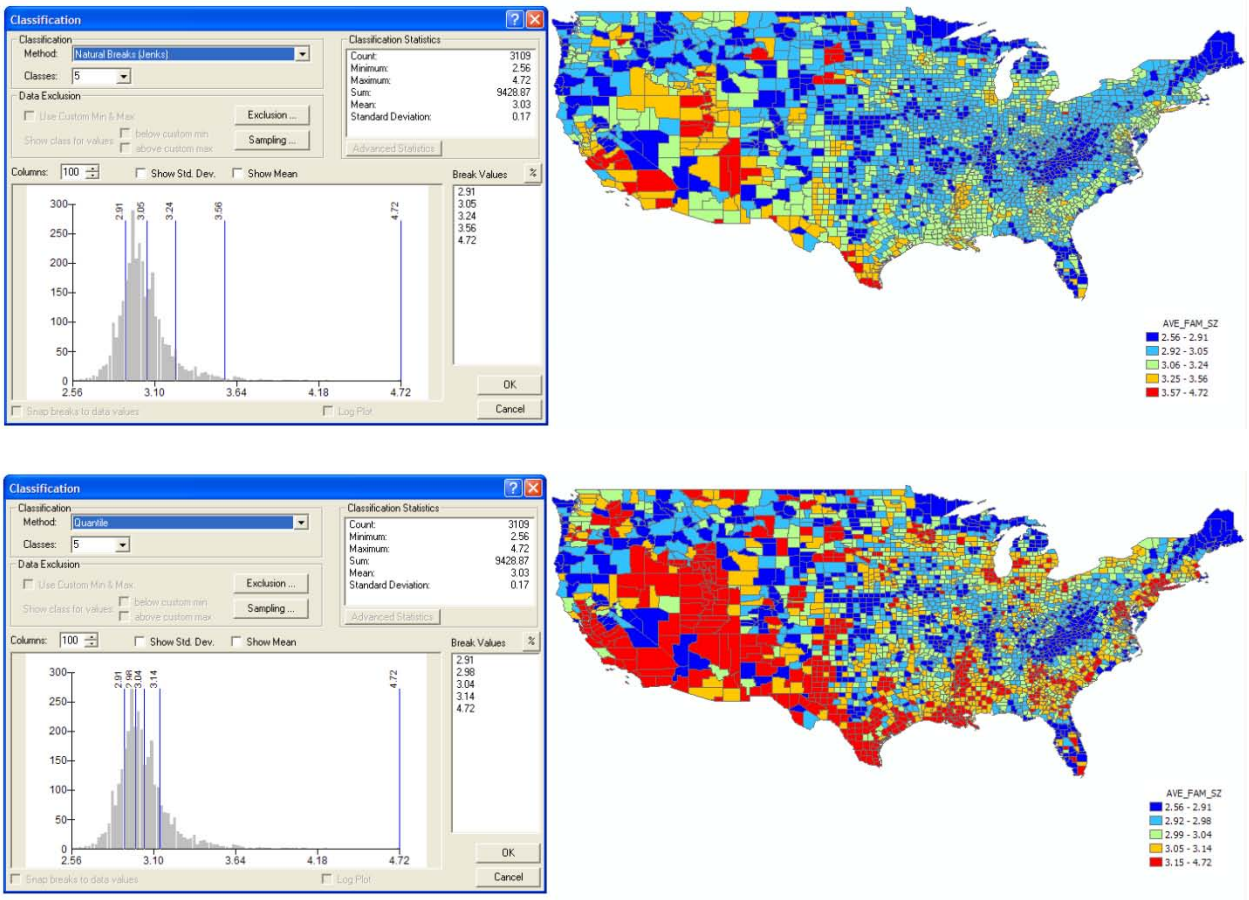


Figure 1.13

Data in their aggregated form exist not by their nature, but by the way they are assembled into regional subdivisions. Figure 1.14 at left shows hypothetical locations of families of different sizes, stars, and the administrative borders. The distribution of family sizes is affected by the administrative divisions within which the average family size is quantified. But, the idea of family size distribution has a continuous component as shown in figure 1.14 at right, because family sizes are likely to grade into each other, however steeply.

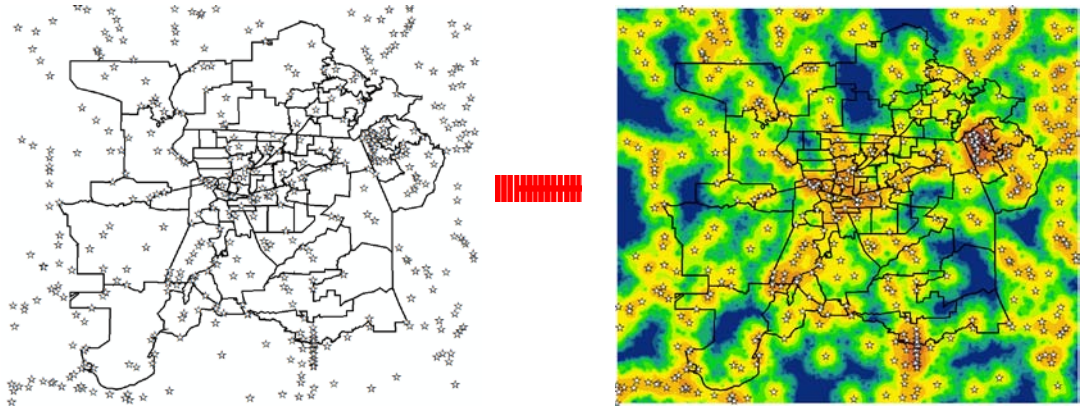


Figure 1.14

Averages for administrative subdivisions are calculated simply by counting the number of events, in this case the number of people (stars in figure 1.15). This answers the question of which choropleth map of average family size

is better: the map that is closer to the population density map (figure 1.14 at right) because a spatial point pattern intensity surface should be related to counts aggregated over regions.

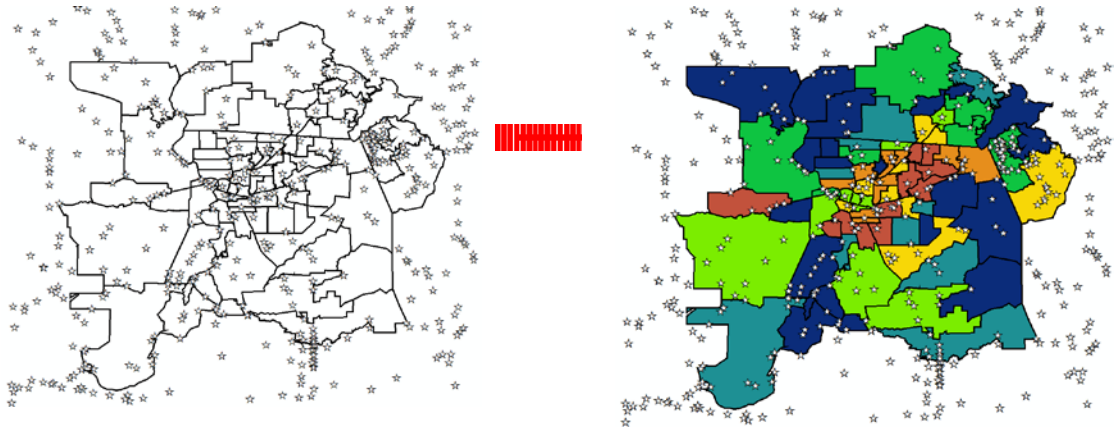


Figure 1.15

The intensity map, in other words, the map of the population distribution, is usually unknown. Smoothing techniques try to approximate it by averaging values in each county with those of the neighboring counties (figure 1.16).

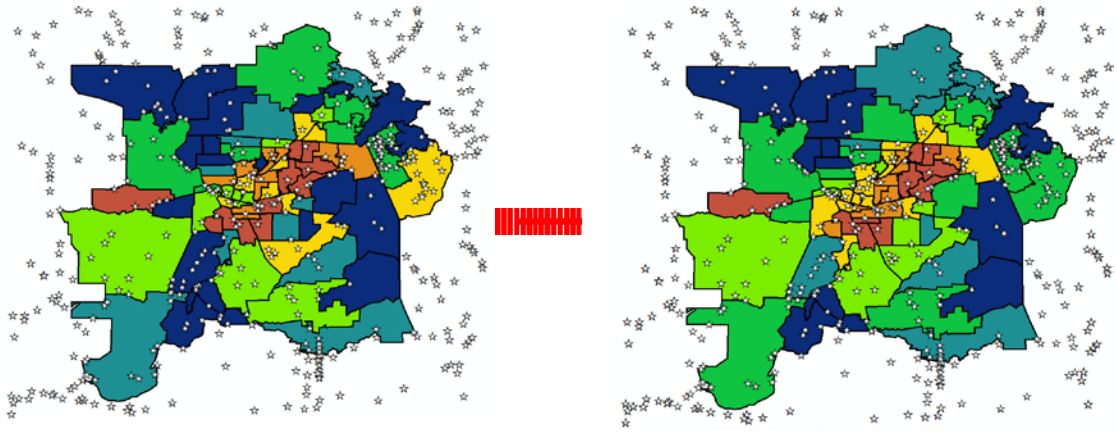


Figure 1.16

In this example, data smoothing makes sense because the average family size in a given county does not completely characterize the distribution of the variable within that county. For example, figure 1.17 shows the distribution of family size in block groups in the part of the city of Redlands, California. Family size variation is large, and values between the first and third data quartiles (3.17 and 3.91 people per family, respectively) have approximately the same frequency of appearance as the mean family size value, 3.54. In other words, the number of families with three and four members is approximately the same. The uncertainty of the family size is described by standard deviation. If standard deviation of the family size for each administrative unit is available, a pair of maps of the average family size and its standard deviation is much more informative than the mean family size map alone.

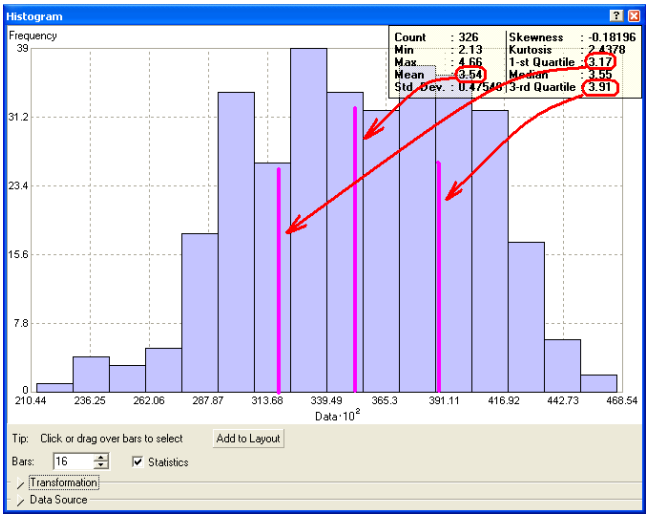


Figure 1.17

Figure 1.18 shows a smoothed map of average family size. A new value is predicted in each county, using values in the neighboring counties and the distance between county centroids (note that more accurate smoothing can be done using areal interpolation model, see chapter 12). After smoothing, it is easier to see which areas have a large family size and which have a small family size.

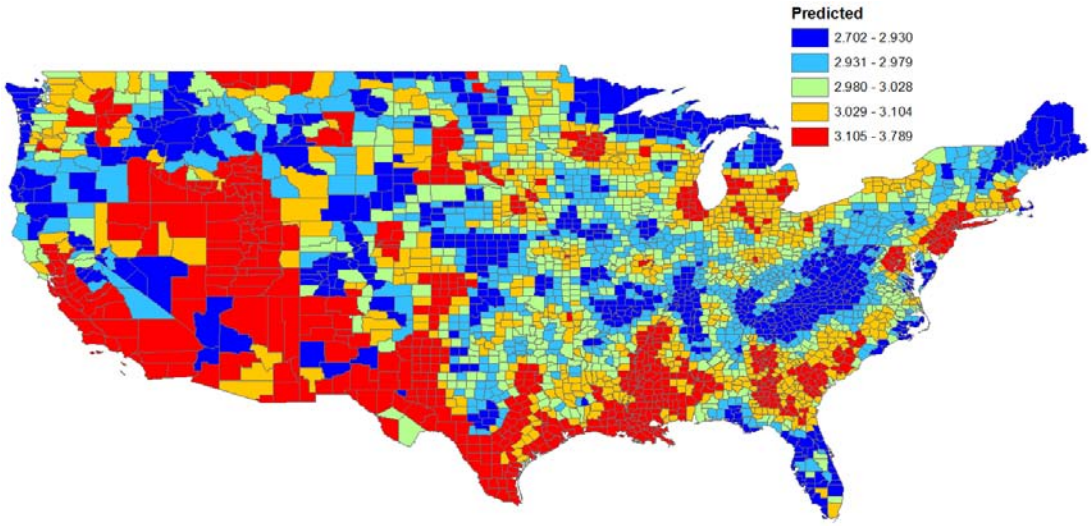


Figure 1.18

But the modeling job is not over until a determination is made of how much uncertainty is associated with the smoothing just accomplished. Figure 1.19 shows that the uncertainty associated with data smoothing is small in counties with little migration and a homogeneous population, shown in blue. This estimated uncertainty substitutes for the unknown standard deviation of the family size in U.S. counties.

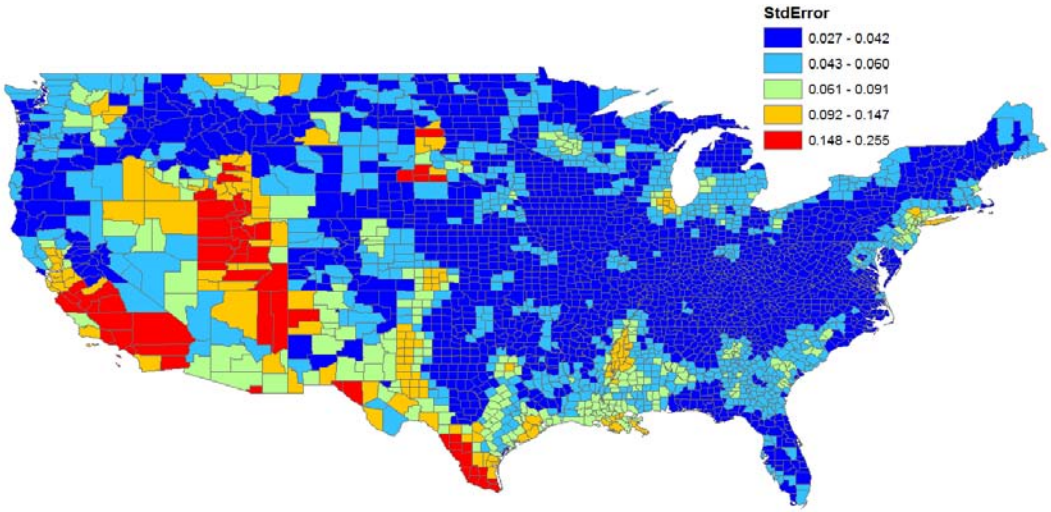


Figure 1.19

The 95-percent prediction interval for the displayed values of the family size in figures 1.18 and 1.19 can be estimated as

Prediction plus or minus two prediction errors

For example, in the southern states, the family size value is approximately 3.0 ± 0.4.

This estimation (it is no more than that) of data uncertainty will have to be taken into account when family size is used in further calculations. Each operation on family size values, whether overlay or buffering or other, will increase the uncertainty.

ASSIGNMENT

INVESTIGATE HOW A SEMIVARIOGRAM CAN CAPTURE SPATIAL DEPENDENCE.

Statistical data interpolation, also called kriging, is based on the statistical description of the data using a semivariogram. Combinations of playing cards can be used to illustrate how a semivariogram captures spatial dependence (see glossary for statistical terminology explanations). The initial order of a new partial deck of cards is shown in figure 1.20.

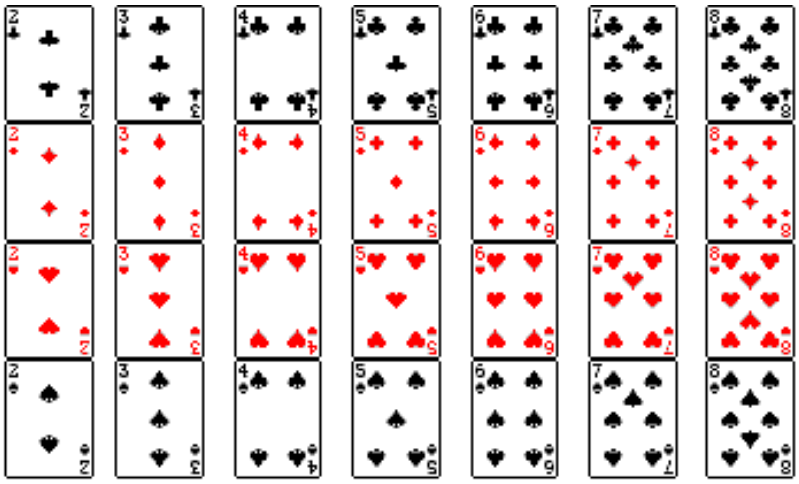


Figure 1.20

The suits can be ignored and only the value of each card considered. The semivariogram can be constructed assuming that the center of a card represents its location and that distances between neighboring cards in horizontal and vertical directions are equal to one, as shown in the graph in the top left part of the dialog box in figure 1.21.

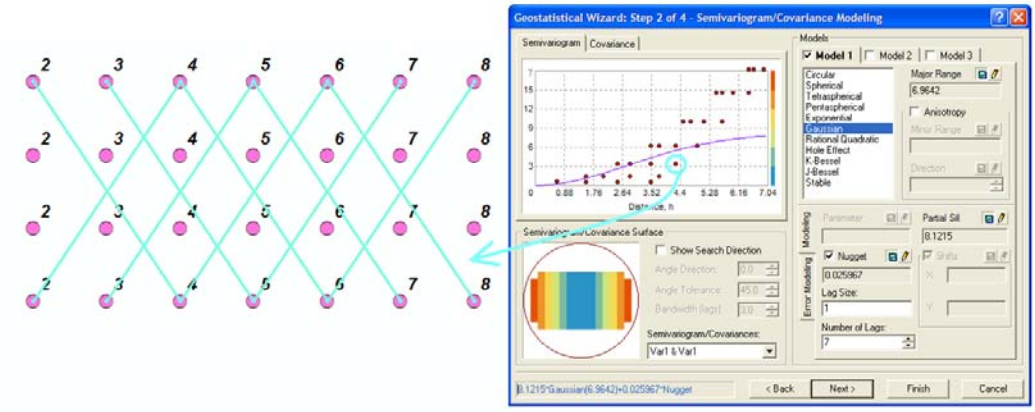


Figure 1.21

The semivariogram γ(h) is constructed by calculating half the average squared difference of the values $Z(s_i)$ of all the pairs of cards separated by a given distance. The resulting quantity is gamma (h) (in Greek). Gamma (h) is plotted on the y-axis against the separation distance, h, known as the lag distance. In mathematical notation

$$\hat{\gamma}(h) = \frac{1}{2 \cdot (\text{number of pairs})} \cdot \sum_{\text{set of pairs (i, j) with similar lag h}} [Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]^2$$

For example, the semivariogram value in blue circle in figure 1.21 is calculated using the 10 pairs connected by the light blue lines.

The differences between the values of points separated by small distances are expected to be smaller than the differences between points separated by greater distances if data are spatially correlated. The pink line is an approximation of the semivariogram values for the distances between pairs of locations. Many functions can be used as semivariogram models. One of them, Gaussian, is used in figure 1.21.

There is a good visual fit between the model and the semivariogram points for the distances smaller than 4. However, the Gaussian semivariogram model cannot describe data with a strong trend at greater distances.

The amount of change in the value with movement from the value at a known point to the unsampled location is governed by the distance dependency explicitly shown by the semivariogram model. Weights of the measurements separated from the prediction location by a distance greater than 3 are small, less than 1 percent (in Geostatistical Analyst, weights are provided by the Searching Neighborhood dialog box next to the Semivariogram Modeling dialog box), and reliable prediction in the extent of the cards with the Gaussian semivariogram shown in figure 1.21 is possible.

In the example shown in figure 1.22, four cards are removed from the array, and the estimated Gaussian semivariogram is used to predict the values at those locations. The data values are displayed in black, and predicted values are given in pink. The resulting surface indicates that the predictions are very good.

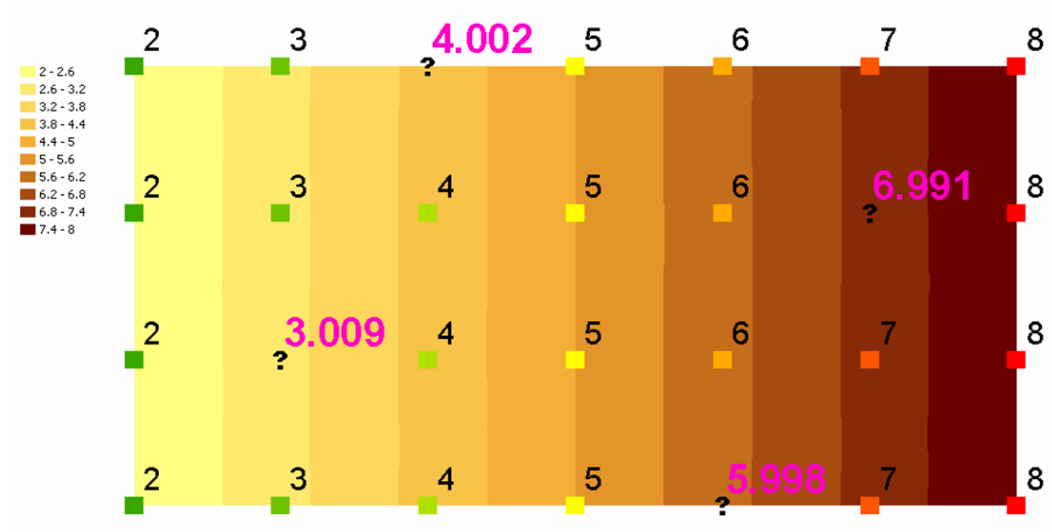


Figure 1.22

Removing an arbitrary number of cards from the middle of the deck and putting them at the end results in the following sequence, figure 1.23.

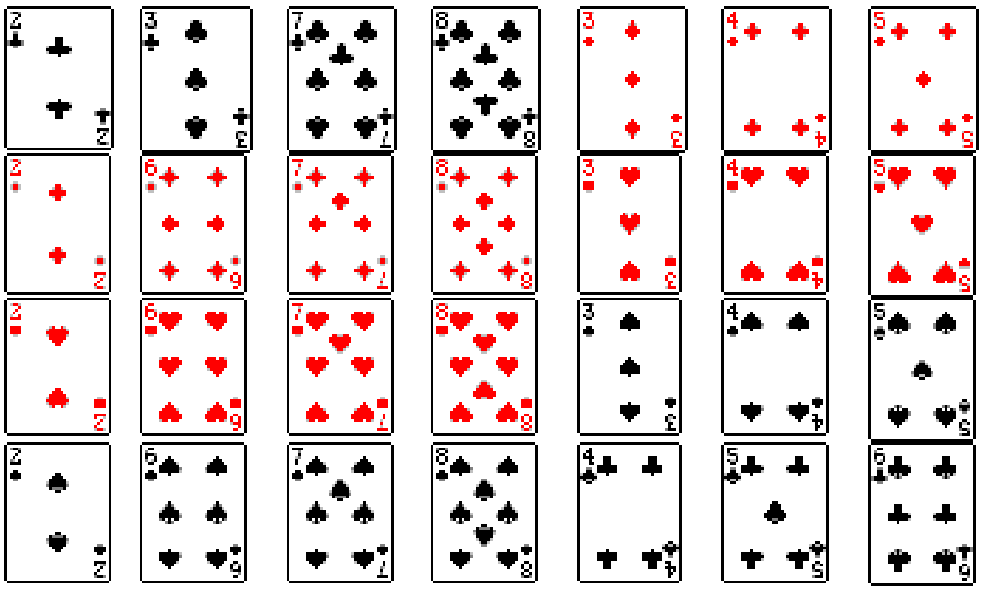


Figure 1.23

The semivariogram based on the shifted data in figure 1.24 still shows spatial dependence, and small data semivariogram values are typically at small distances between the pairs of points rather than at greater distances, but not as much as in the case of the cards prior to reordering.

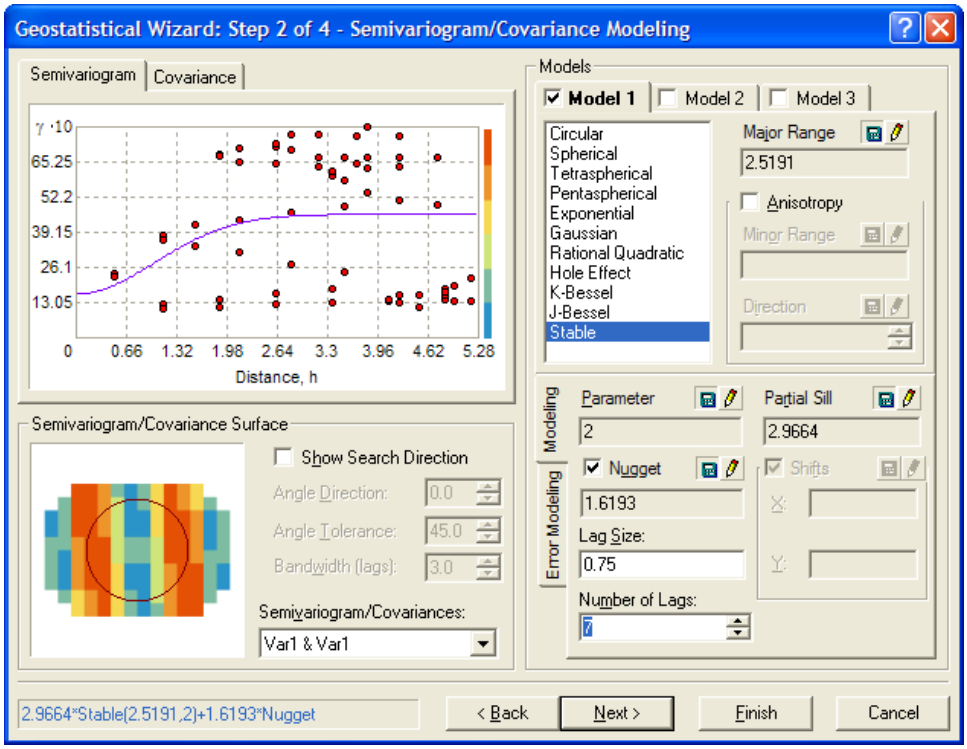


Figure 1.24

The kriging predictions (figure 1.25) using the semivariogram in figure 1.24 indicate that two of the four predictions *(4.28* and *5.28)* are not accurate.

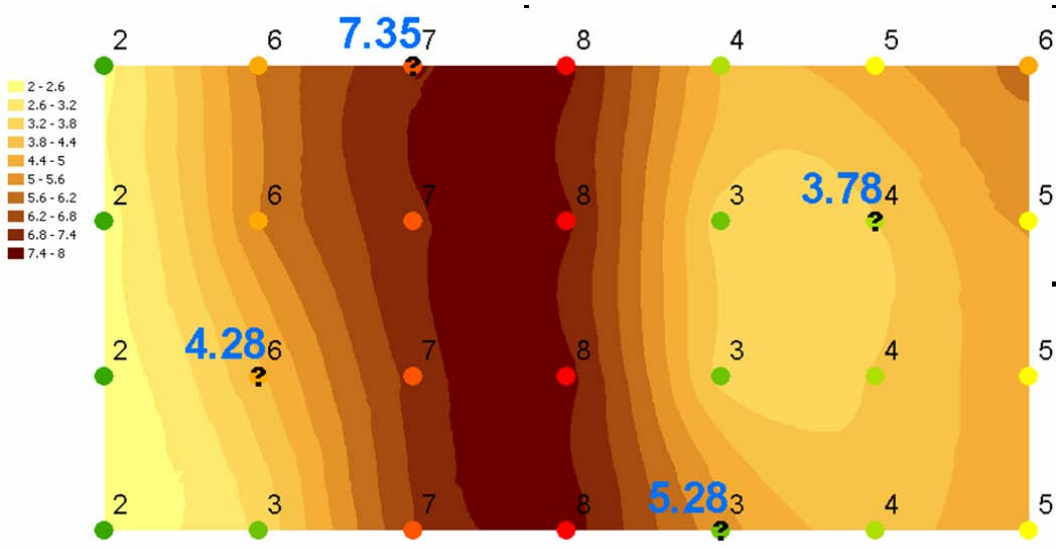


Figure 1.25

Continuing the exercise, the cards are shuffled twice more, and a new semivariogram is calculated. The data in the left part of figure 1.26 were used to create the semivariogram in the right part. It can be seen that the spatial dependence has almost disappeared: the semivariogram model is close to the horizontal line with almost the same data variability at all distances between data pairs.

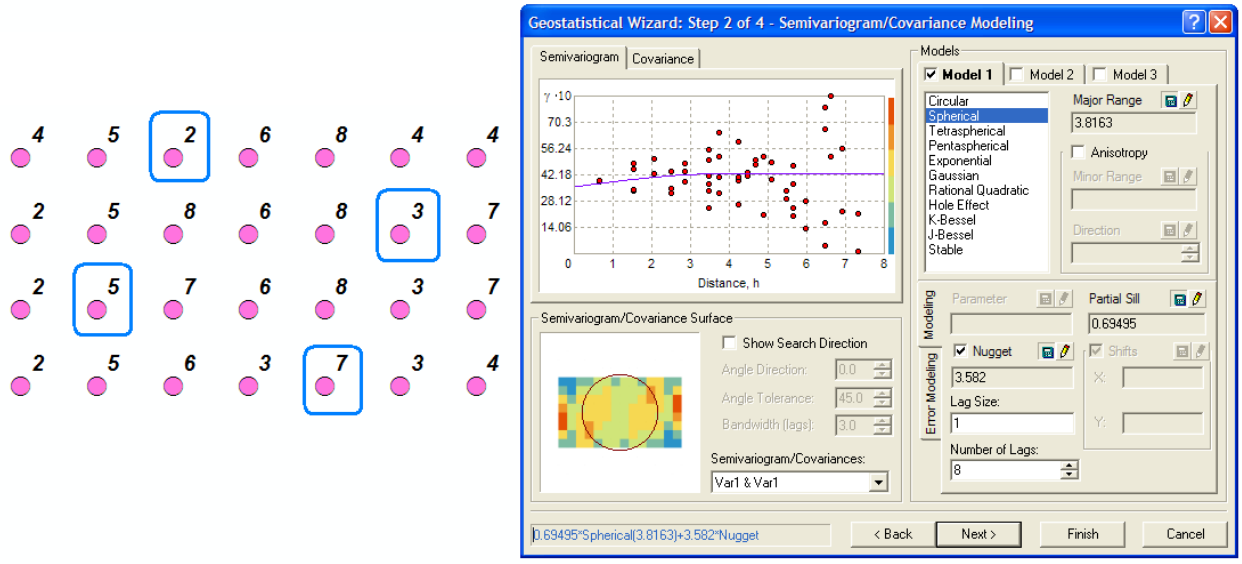


Figure 1.26

Kriging using the semivariogram shown in figure 1.26 results in weak predictions at three out of four locations (figure 1.27).

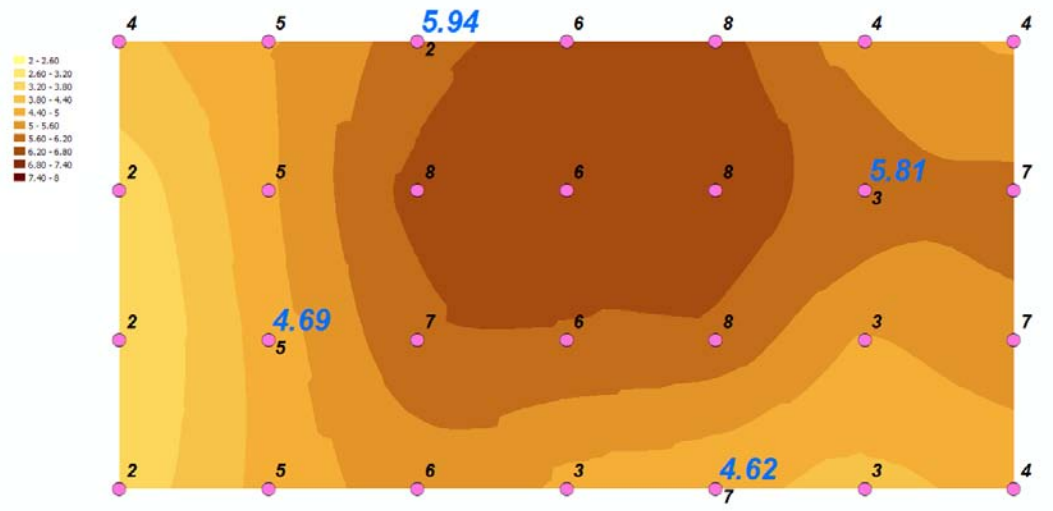


Figure 1.27

In summary, when strong spatial correlation exists, the semivariogram can help to reconstruct missed values, and kriging makes reliable predictions at the unsampled locations. If data are not spatially correlated, there is no way to predict the value between measurement locations other than assigning an arithmetical mean to all prediction locations. As the data correlation gets stronger, fewer samples are needed for reliable data prediction and interpolation.

With Geostatistical Analyst, repeat this exercise using a different number of initially ordered data, several reordering scenarios, and various locations of removed data.

FURTHER READING

1. De Veaux, R., P. Velleman, D. Bock (2007) *Stats: Data and Models.* Second Edition. Addison Wesley, 800 pp.

This book can be recommended for readers with no knowledge of statistics. The book focuses on statistical thinking, highlighting how statistics helps us to understand the world. It is organized into short chapters that focus on one topic at a time. Each chapter includes a discussion of common misuses, misapplications, and misunderstanding of statistics to help students recognize and avoid them.

2. Krivoruchko, K., and C. A. Gotway. 2002. "Expanding the 'S' in GIS: Incorporating Spatial Statistics in GIS."

It is available from Esri online at
`http://www.esri.com/software/arcgis/arcgisxtensions/geostatistical/research_papers.html`

This paper discusses the need for spatial statistics within a GIS and the power of the inferential methods it provides. The authors illustrate through several case studies the utility of spatial statistics within a GIS. Although Geostatistical Analyst is an excellent first step in integrating spatial statistics and GIS, there is also a need to include spatial statistics for other types of data for which the premises of geostatistical models do not apply.

3. Krivoruchko, K., and C. A. Gotway. 2003. "Using Spatial Statistics in GIS."

This paper was presented at the International Congress on Modeling and Simulation, Townsville, Australia, July 2003. It was published in the conference proceedings. It is available from Esri online at
`http://www.esri.com/software/arcgis/arcgisxtensions/geostatistical/research_papers.html`

The authors provide several examples that show the power of exploratory spatial data analysis within a GIS and how this can provide the foundation for more sophisticated probabilistic modeling. While Esri ArcGIS software now facilitates the integration of spatial data analysis and GIS functionality, more tools are needed for comprehensive spatial data analysis. The authors suggest how to implement additional spatial statistical methods within a GIS including methods for using non-Euclidean distances in the analysis of geostatistical, lattice, and point pattern data.

4. Krivoruchko K. and R. Bivand (2009) "GIS, Users, Developers, and Spatial Statistics: On Monarchs and Their Clothing." In *Interfacing Geostatistics and GIS*, pp. 209-228. Springer

This paper was presented at the StatGIS 2003, International Workshop on Interfacing Geostatistics, GIS, and Spatial Databases, September 29–October 1, 2003, Pörtschach, Austria. It is available from Esri online at
`http://www.esri.com/software/arcgis/arcgisxtensions/geostatistical/research_papers.html.`

The paper discusses the statistical tools and models that propagate between communities of users and the problems that arise with using statistical inference in inappropriate settings.

5. Bailey, T. C., and A. C. Gatrell. 1995. *Interactive Spatial Data Analysis.* Essex: Addison Wesley Longman Ltd.

This introductory book on spatial data analysis for undergraduate students describes statistical methods for discrete, regional, and continuous data. The Bailey and Gatrell book is recommended for additional reading in several other chapters.