9

Spatial interpolation

9.1 Introduction

This chapter introduces a variety of approaches for the generation of surfaces, including topographic surfaces and other properties that can be treated as surfaces, such as precipitation amount or airborne pollutants. The term 'spatial interpolation' refers to the prediction of values at locations where no sample is available; a common objective in GIS contexts is to predict values on a regular grid using irregularly distributed point data, and this is the principal focus here.

There is a wide range of applications that depend on tools for predicting values on a regular grid from a set of samples irregularly distributed in space. An example is precipitation mapping. Precipitation amount may be measured using a set of rain gauges but if values are required elsewhere then it may be necessary to predict these values using the sample data. Another class of techniques that fall within the remit of this chapter are those that are used to transfer counts from one set of zones (e.g. census areas) to another set of zones or from zones to a grid. The term 'areal interpolation' describes such approaches.

9.2 Interpolation

Spatial interpolation approaches can be divided into those that are global and those that are local. Global approaches make simultaneous use of all sample data in the prediction process. Local approaches use only a subset of data (in a moving window) to make predictions. Another division into which spatial interpolation methods can be divided is exact and approximate methods. Exact methods 'honour' data locations—that is, observed values are not replaced and the predicted value at a location where there is a sample is the same as the sample value. With approximate methods, there is no guarantee that this is the case.

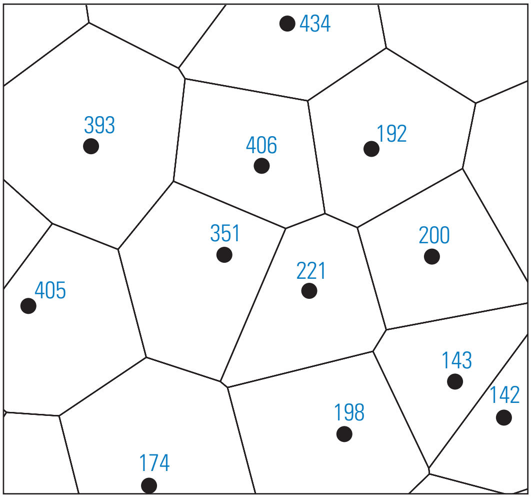

Figure 9.1 Thiessen polygons.

There are many approaches to spatial interpolation. A simple global approach is to fit a trend surface to all of the data, and values of the fitted surface can then be read off at any location, whether there is a sample at the location or not. Trend surface analysis entails fitting a plane or a curved surface through the data that represents the general trend of values (e.g. a 'gradual' increase in values from south to north). Trend surface analysis is simply multiple regression whereby the dependent variable is the variable of interest (e.g. precipitation) and the independent variables are the data coordinates or some function of them. Conceptually, the most simple interpolation method is Thiessen polygons. With this approach, the values assigned to unsampled locations are those of the nearest observation. Figure 9.1 gives an example of Thiessen polygons—the point value is assigned to the area which is closer to that point than any other.

The focus in this section is on methods that have a raster grid as their output. One interpolation framework that does not is triangulation. The output of triangulation is a triangulated irregular network (TIN), which is essentially a surface comprising triangular facets that connect the observations (see Heywood *et al.* (2006) for an example).

The following sections introduce five of the most widely used spatial interpolation methods: regression (introduced in Section 3.3, and expanded on in the previous chapter), TINs, inverse distance weighting (IDW), thin plate splines (TPS), and ordinary kriging (OK). There is a whole literature on TPS and geostatistical interpolation (kriging), and the accounts provided below are necessarily brief.

9.3 Triangulated irregular networks

The TIN is a vector-based representation of a surface. In essence it comprises a set of vertices joined by arcs, which together form triangles. TINs are more efficient than

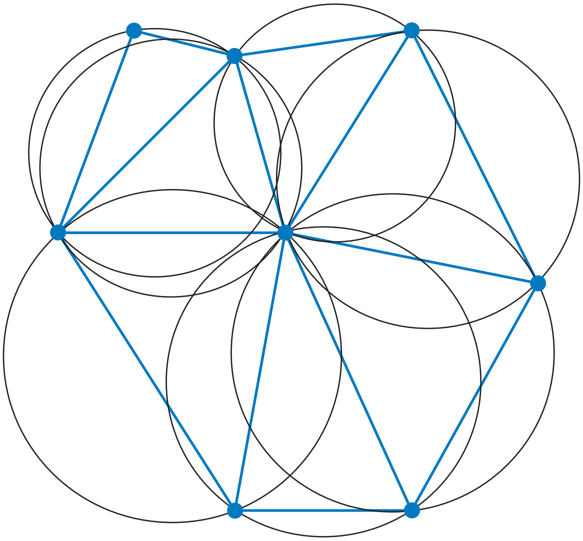

Figure 9.2 TIN subset with circumcircles superimposed.

raster-based digital elevation models (DEMs) as, in the latter case, elevations are stored at all locations on a regular grid while with TINs the sampling density can be varied as a function of the nature of the topography. Generally, marked breaks of slope are represented and there are usually more samples in areas with more variable elevations and fewer samples in relatively flat areas. Various approaches exist for the selection of observations with the objective of representing the surface as precisely as possible with the minimum of redundancy. One widely used approach to select points from a regular grid is the very important points (VIP) algorithm. With this approach, points are assigned a significance that is a function of the difference between each point and its neighbours. Following this procedure a specific number of the most significant points can be retained or points could be retained such that the loss of accuracy is minimized (Li *et al.*, 2004).

A TIN can be constructed from known point values using a process called Delaunay triangulation. Peuker *et al.* (1978) provide a detailed account of TINs. With Delaunay triangulation, the triangles are formed such that the circumcircle of each triangle contains no vertices other than those that make up the triangle. The circumcircles for a set of triangles are shown in Figure 9.2. As can be seen in the figure, the circumcircle runs through the three vertices that belong to a given triangle and in no case does a circumcircle for a given triangle contain any vertices other than those that belong to that triangle. With Delaunay triangulation, two vertices are connected if their Thiessen polygons share an edge, and this is illustrated in Figure 9.3. Delaunay triangulations can be derived from Thiessen polygons. Li *et al.* (2004) outline a variety of alternative approaches to selecting the starting point for, and conducting, Delaunay triangulation.

A TIN is illustrated using the Walker Lake sample of 470 point observations. These data come from the book by Isaaks and Srivastava (1989), and are provided through the AI-Geostats website (see http://www.ai-geostats.org/index.php?id=data).

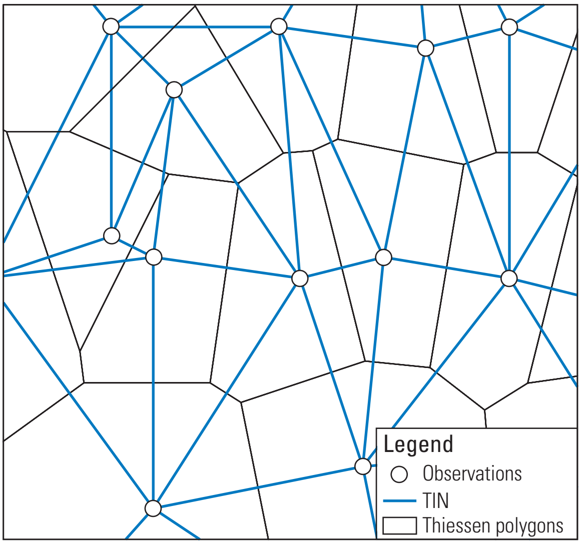

Figure 9.3 TIN subset with Thiessen polygons superimposed.

The illustrations here are based on the *V* variable derived from elevation data and described by Isaaks and Srivastava (1989). For the purposes of their example, Isaaks and Srivastava refer to the variable *V* as concentrations of some material in parts per million (ppm), here the variable is expressed in the same way, although the data are treated as elevation values for illustrative purposes. The data are used as they are preferentially sampled and allow the illustration of the TIN and some of its potential benefits in terms of smaller data storage requirements relative to an altitude matrix.

Figure 9.4 shows the point data, a raster grid generated with IDW (with an exponent of 2 and using eight nearest neighbours), and the edges of triangular facets derived using Delaunay triangulation. The TIN was generated using ArcGIS™ 3D Analyst.

Figure 9.5 shows a '2.5D' visualization of the TIN. In this case the 'elevations' are multiplied by 0.04 with respect to the *x* and *y* coordinate values, which has the effect of compressing the 'elevation' values. The ridge of large values running along the west of the region from north to south is apparent in Figure 9.5.

9.4 Regression for prediction

Regression (whether a global or local variant, like geographically weighted regression) can be used for spatial interpolation if values of the independent variable are available at all locations where predictions are required. In the previously outlined case of elevation and precipitation amount, if we have a DEM with elevation values at all locations in the study area then we can use the regression equation (either globally or locally) to predict precipitation amounts for all grid cell locations in the DEM (see, for example, Lloyd, 2005).

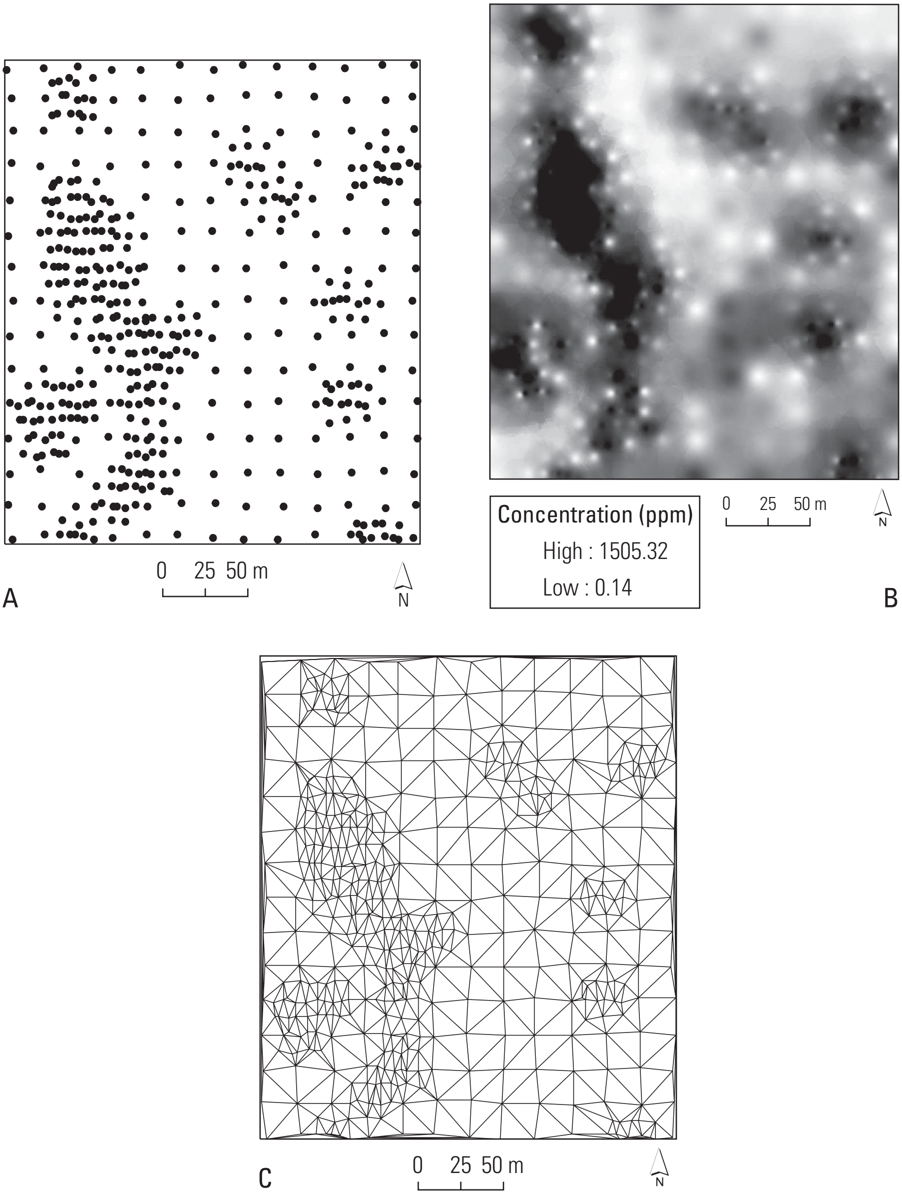

Figure 9.4 (A) Walker Lake sample data locations, (B) map derived using IDW with eight nearest neighbours, and (C) TIN.

9.4.1 Trend surface analysis

As summarized in Section 9.2, a trend surface is fitted using regression, but instead of regressing different variables, values are predicted using a regression of the dependent variable (e.g. elevation) against the coordinates or some function of them.

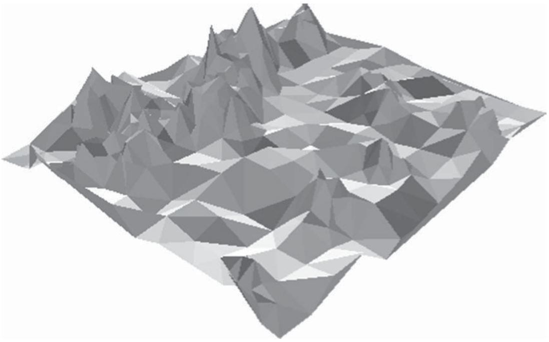

Figure 9.5 Shaded '2.5D' visualization of the TIN in Figure 9.4(C), viewed from the south east. Based on a *z* conversion factor of 0.04.

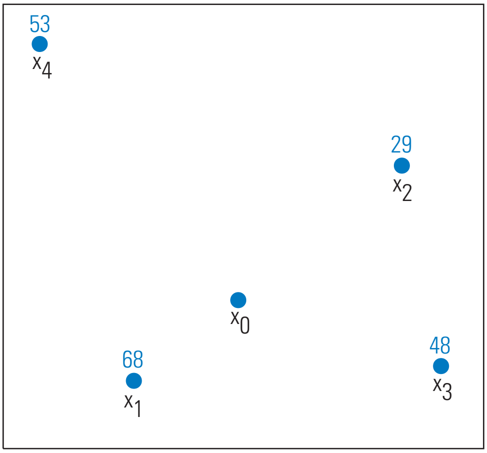

Figure 9.6 Location of prediction location and observations. Values and ID codes are given for the four samples and the value at location x_0 is treated as unknown.

For example, where the coordinates are given by *x* and *y*, the independent variables may be, for example, just *x* and *y* (this is called a first-order polynomial trend) or *x*, *y*, *xy*, x^2 and y^2 (this is called a second-order polynomial trend). Such approaches are useful for depicting general trends, but are unlikely to be of much practical use for direct interpolation.

9.5 Inverse distance weighting

Weighted moving averaging is a widely used approach to interpolation. A variety of different weighing functions have been used but IDW is the most common form in GIS.

IDW is an exact interpolator, so the predicted values at locations where there are observations are the same as at the observed values. The IDW predictor can be given as:

$$\hat{z}(\mathbf{x}_0) = \frac{\sum_{i=1}^{n} z(\mathbf{x}_i) \cdot d_{i0}^{-k}}{\sum_{i=1}^{n} d_{i0}^{-k}} \tag{9.1}$$

where the prediction is made at the location $\mathbf{x}_0$ as a function of the n neighbouring observations, $z(\mathbf{x}_i)$, $i=1,\ldots,n$ (i.e. we feed only the n nearest neighbours into Equation 9.1), k is an exponent that determines the weight assigned to each of the observations, and d_{i0} is the distance by which the prediction location $\mathbf{x}_0$ and the observation location $\mathbf{x}_i$ are separated. As noted in Section 4.7, as the exponent becomes larger, the weight assigned to observations at large distances from the prediction location becomes smaller. Conversely, as was shown in Figure 4.6, for smaller values of the exponent, the weights are proportionately larger for more distant observations. The exponent is usually set to 2 (i.e. d_{i0}^{-2}) and the inverse squared distance (where $k=2$) is obtained with $^1/_{d^2}$. The inverse square of a distance of 6481.996 m, is therefore $^1/_{6481.996^2}=0.00000002380$, as shown in Table 9.1.

In this section, a worked example of IDW is given using four observations, with the objective of predicting at another location (Figure 9.6 shows the data configuration). Since an observation is available at the prediction location $\mathbf{x}_0$, but it has been removed for the present purpose, it is possible to assess the accuracy of the predictions. The data are given in Table 9.1. The same data set is used to illustrate other interpolation methods in this chapter.

Following the IDW equation we first calculate the inverse square of the distances and then multiply these values by the value of the observations. Table 9.1 shows the inverse square distances and the observation values multiplied by the inverse square distances. The IDW prediction is given by (using the figures from Table 9.1) $^{0.000002526}/_{0.00000004592}=55.003$. The 'true' value at the prediction location is 61, so there is a prediction error of 5.997. In practice, assessment of prediction can be conducted using jackknifing or cross-validation. Jackknifing entails splitting the sample into two and using one set of data to make predictions at the locations represented by the second data set. The accuracy of these predictions can obviously be assessed directly. The basic idea of cross-validation was described in another context in Section 8.5.3. As detailed in that section, cross-validation entails removal of an observation, using the

Table 9.1 Precipitation (mm): IDW prediction using observations $\mathbf{x}_1$ to $\mathbf{x}_4$

i	x_i	y_i	$z(\mathbf{x}_i)$	d_{i0}	d_{i0}^{-2}	$z(\mathbf{x}_i) \cdot d_{i0}^{-2}$
1	292500	329100	68	6481.996	0.00000002380	0.000001618
2	305700	339700	29	10448.860	0.00000000916	0.000000266
3	307629	329826	48	10517.774	0.00000000904	0.000000434
4	287854	345702	53	15969.356	0.00000000392	0.000000208
				Sum	0.00000004592	0.000002526

remaining observations to predict the value of the removed value. The removed value is returned to the data set and the next observation in order is removed, after which the procedure is repeated for all observations. The prediction errors can then be assessed. The accuracy of prediction is commonly assessed using summaries such as the mean error, the mean absolute error, and the root mean square error (RMSE).

In terms of selecting a data subset for interpolation, several common strategies exist. The n (four in the example) nearest neighbours to a prediction location could be selected. Alternatively, all observations within a specified distance of the prediction location could be used. Another strategy is to divide the search neighbourhood into quadrants, for example north-east, south-east, north-west, and south-west of the prediction location. The weights could then be scaled according to the number of observations in each quadrant and this would help to overcome the effect of clustering of observations in particular areas.

IDW was used to generate a map of precipitation amount in July 2006 in Northern Ireland using the 16 nearest neighbours to each cell of the prediction grid. The data locations are shown in Figure 8.7 and the IDW-derived map is shown in Figure 9.7. The map in Figure 9.7 is very smooth in appearance and there are clear clusters of values around the sample locations—this is a common feature of maps generated using IDW. With IDW, there tend to be clusters of similar values around data points (see Lloyd (2005) for another example).

IDW is rapid and easy to implement, although it often performs less well than more sophisticated approaches (e.g. see Lloyd, 2005).

9.6 Thin plate splines

TPS are, like IDW, a very widely used approach to spatial interpolation. TPS functions are available in ArcGIS™, the GRASS GIS (Neteler and Mitásová, 2007), and in other software packages. TPS can be viewed as surfaces that are fitted to some local subset of the data. The spline can be fitted exactly to the data points or it can be smoothed—that is, if the spline is not forced to fit to the data points the resulting surface can be made smoother than if the surface runs through every point. In effect, the thin plate smoothing spline generated map is a map of local weighted averages. With TPS, the aim is to obtain a prediction of the unknown value $\hat{z}(\mathbf{x}_0)$ with a smooth function g by minimizing (as defined below):

$$\sum_{i=1}^{n}(z(\mathbf{x}_i)-g(\mathbf{x}_i))^2+\rho J_m(g) \tag{9.2}$$

where $J_m(g)$ is a measure of the roughness of the spline function, $\boldsymbol{m}$ is the degree of the polynomial used in the model, and ρ is the smoothing parameter. We seek to find the function g so that it is as close as possible to the observations (indicated by $z(\mathbf{x}_i)$), with the smoothing function determining if the fit of the function to the observations is

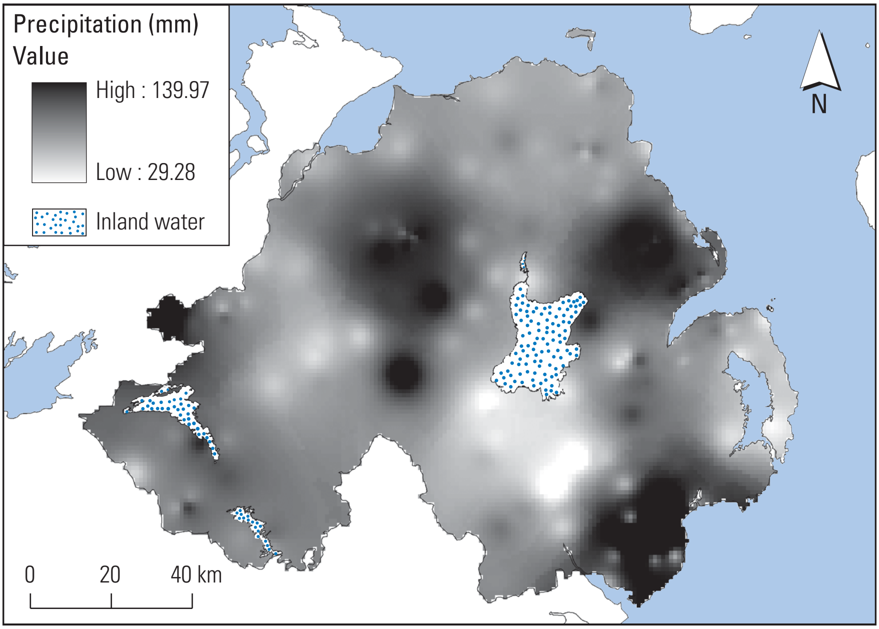

Figure 9.7 Precipitation in July 2006: IDW prediction using 16 nearest neighbours.

exact or approximate. If the smoothing parameter is zero then the spline passes through the data (it is an exact interpolator); if the value is greater than zero then the spline is not forced to fit to the data (it is an approximate interpolator). If the data are 'noisy' (e.g. there is notable measurement error), use of a smoothing parameter may be desirable. Large values of the smoothing parameter result in smoother maps. The TPS function g is made up of two parts (as defined after the equation):

$$g(\mathbf{x}) = a_0 + a_1 x + a_2 y + \sum_{i=1}^{n} \lambda_i R(\mathbf{x} - \mathbf{x}_i) \tag{9.3}$$

The aim is to find values of a_0, a_1, a_2, and λ_i to make a prediction, as we know the other terms (which will be detailed below) in advance. The left-hand side of the equation (i.e. $a_0 + a_1 x + a_2 y$) indicates the local trend in the data. The introduction to this chapter, as well as Section 9.4, mentioned trend surface analysis, in which a surface (either a flat plane or a more complex surface) is fitted to the data in the same way that a line is fitted to a scatter plot using regression. In the case of splines, a surface of this kind is fitted to the data, but only to some local subset (e.g. the 16 nearest neighbours to the prediction location). In the same way as a slope value is found in (bivariate) linear regression and multiplied by the independent variable, we must find values for a_1 and a_2, and these will be multiplied by x and y. The term $R(\mathbf{x} - \mathbf{x}_i)$ is called a basis function and for TPS it is given by:

$$d_i^2 \log d_i \tag{9.4}$$

The distance d is that between the location prediction $\mathbf{x}$ and the location $\mathbf{x}_i$, so $R(\mathbf{x}-\mathbf{x}_i)$ is, in this case, the distance between those two locations fed into Equation 9.4. As an example, for a distance of 16.9292646 units:

$$d_i^2 \log d_i = 16.9292646^2 \log 16.9292646 \ = 352.128$$

In short, the TPS function comprises the local trend and weights (λ_i) by which the basis function values are multiplied (the process is illustrated below). Using matrix notation, the coefficients a_k and λ_i are the solution of:

$$\mathbf{R}\lambda = \mathbf{z} \tag{9.5}$$

Appendix F shows how to solve such equations (i.e. how to find the unknown values, which in this case are the coefficients a_k and λ_i).

R is a matrix obtained by feeding the distances between local observations into the equation $d^2 \log d$. As above, a given distance is squared and multiplied by the log of the distances. The matrix **R** is given by:

$$\mathbf{R} = \begin{bmatrix} R(\mathbf{x}_1-\mathbf{x}_1) & \cdots & R(\mathbf{x}_1-\mathbf{x}_n) & 1 & x_1 & y_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ R(\mathbf{x}_n-\mathbf{x}_1) & \cdots & R(\mathbf{x}_n-\mathbf{x}_n) & 1 & x_n & y_n \\ 1 & \cdots & 1 & 0 & 0 & 0 \\ x_1 & \cdots & x_n & 0 & 0 & 0 \\ y_1 & \cdots & y_n & 0 & 0 & 0 \end{bmatrix}$$

λ are the TPS weights and **z** are the observations:

$$\lambda = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ a_0 \\ a_1 \\ a_2 \end{bmatrix} \qquad \mathbf{z} = \begin{bmatrix} z(\mathbf{x}_1) \\ \vdots \\ z(\mathbf{x}_n) \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

The matrix **R** has the weights and three rows and columns corresponding to a constant trend component (the 1s) and the x and y coordinates of each location, the vector (a matrix with only one row or column) λ has three extra rows, which are values of a_k for the constant (a_0) and for the x and y coordinates of each location (i.e. a_1 and a_2), and the vector **z** includes three zeros, corresponding to the three trend components.

To obtain the TPS weights (λ_i) and the values of a_0, a_1, and a_2, the inverse (see Section 3.3 and Appendices E and F for a discussion about matrix inversion) of the matrix **R** is multiplied by the vector of data values, **z**:

$$\boldsymbol{\lambda} = \mathbf{R}^{-1}\mathbf{z}$$

Using the same data as for the IDW example in Section 9.5, following Equation 9.5 the TPS system is given as:

$$\begin{bmatrix} 0 & 352.128 & 270.779 & 367.512 & 1 & 292.500 & 329.100 \\ 352.128 & 0 & 101.483 & 451.925 & 1 & 305.700 & 339.700 \\ 270.779 & 101.483 & 0 & 902.999 & 1 & 307.629 & 329.826 \\ 367.512 & 451.925 & 902.999 & 0 & 1 & 287.854 & 345.702 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 292.500 & 305.700 & 307.629 & 287.854 & 0 & 0 & 0 \\ 329.100 & 339.700 & 329.826 & 345.702 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 68 \\ 29 \\ 48 \\ 53 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Note that the distances were divided by 1000 prior to calculating the values for **R**. This gives the same results but reduces the size of the values obtained for $R(\mathbf{x}-\mathbf{x}_i)$ and makes the process more manageable. The diagonals in the matrix **R** are all 0 and they indicate the distance between an observation and itself (obviously 0); where a smoothing parameter is used (and the spline function is not forced to fit to the data), the smoothing parameter value is added to the diagonals (for this example, the first four 0 components, reading from the left), as described by Lloyd (2006).

Solving the TPS system, the weights are as follows: $\lambda_1=-0.0319$, $\lambda_2=-0.0493$, $\lambda_3=0.0520$, $\lambda_4=0.0292$, $a_0=1078.474$, $a_1=-1.5906$, and $a_2=-1.6794$.

We then put the distances between each observation and the prediction location into the equation $d^2 \log d$. For each observation this gives the following values: $\mathbf{x}_1=34.105$, $\mathbf{x}_2=111.261$, $\mathbf{x}_3=113.049$, and $\mathbf{x}_4=306.863$.

The predicted value is then given by multiplying the weights (λ_i) by the basis function values ($R(\mathbf{x}-\mathbf{x}_i)$), adding a_0 (the constant, note that the intercept is also called the constant), and multiplying the trend coefficients (a_1 and a_2) by the coordinates (x and y). In this case this leads to: $(34.105\times-0.0319)+(111.261\times-0.0493)+(113.049\times 0.0520)+(306.863\times 0.0292)+1078.474+(297.624\times-1.5906)+(333.070\times-1.6794)=60.569$.

The 'true' value is 61 and the TPS prediction error is smaller than the IDW prediction error (with an IDW prediction of 55.003; see Section 9.5).

Figure 9.8 shows a map of precipitation amount generated using TPS with the same data used to illustrate the application of IDW (see Figure 9.7). Lloyd (2006) gives a summary account of variants of the TPS approach, which may provide more robust and more accurate predictions than standard TPS in some circumstances.

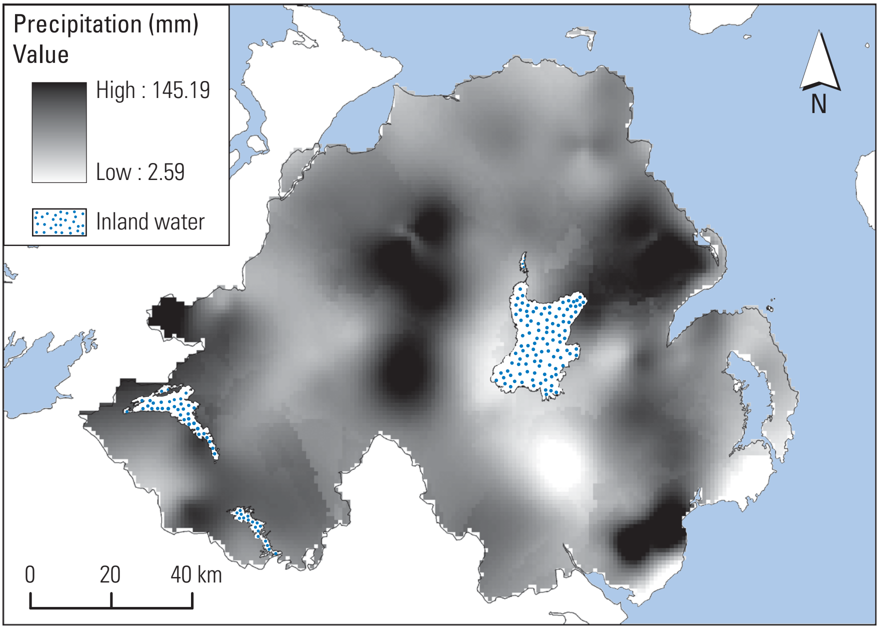

Figure 9.8 Precipitation in July 2006: TPS prediction using 16 nearest neighbours.

9.7 Ordinary kriging

Ordinary kriging is one of a family of kriging methods. Kriging falls within the remit of a field known as geostatistics. Burrough and McDonnell (1998) provide a short introduction to geostatistics in the context of GIS. The basis of geostatistics is the theory of regionalized variables. Geostatistics entails a conceptual division of spatial variation (at a location $\mathbf{x}$) into two distinct parts: a deterministic component ($\mu(\mathbf{x})$) (representing 'gradual' change over the study area) and a stochastic (or 'random') component ($R(\mathbf{x})$):

$$Z(\mathbf{x}) = \mu(\mathbf{x}) + R(\mathbf{x}) \tag{9.6}$$

This is termed a 'random function' (RF) model. The random part reflects our uncertainty about spatial variables—what seems random to us is a function of a multiplicity of factors that may be impossible to model directly (and this does not mean that we really think variation is random; Isaaks and Srivastava, 1989). In geostatistics, a spatially referenced variable, $z(\mathbf{x})$, is treated as an outcome of an RF, $Z(\mathbf{x})$. In other words, we effectively consider an observation to have been generated by the RF model and this gives us a framework to work with these data. A realization of an RF is called a regionalized variable (ReV; i.e. an observation). The theory of regionalized variables (Matheron, 1971) is the fundamental framework on which geostatistics is based.

The discussion on first- and second-order effects in Section 7.1 is relevant in this context as it deals with the distinction between variation in the mean and spatial dependence. This section also has links with the introduction to TPS, whereby the TPS function (Equation 9.3) is shown to comprise a trend component and the component (here termed the 'random' part) modelled by the basis function $R(\mathbf{x}-\mathbf{x}_i)$ (in the present case, the variogram is used instead to model this component, as described below).

In practical terms, as we estimate parameters, namely the mean and variance, of a distribution, we estimate parameters of the RF model using the data. These parameters, like the mean and variance, summarize the variable. The mean and variance of a distribution are useful only if the distribution is approximately normal and, similarly, the parameters of the RF model are only meaningful in certain conditions. Where the properties of the variable of interest are the same, or at least similar in some sense, across the region of interest we can employ a stationary model. In other words, we can use the same model parameters at all locations. If the properties of the variable are clearly spatially variable then a standard RF model may not be appropriate. There are different degrees of stationarity, but for present purposes we will only consider one, intrinsic stationarity. There are two requirements of intrinsic stationarity. Firstly, the mean is constant across the region of interest. In other words, the expected value of the variable does not depend on the location, **x**:

$$E\{Z(\mathbf{x})\} = \mu(\mathbf{x}) \text{ for all } \mathbf{x} \tag{9.7}$$

The mean is therefore assumed to be the same for all locations. Secondly, the expected squared difference between paired RFs (i.e. the observations) (summarized by the variogram, $\gamma(\mathbf{h})$) should depend only on the separation distance and direction (the lag **h**) between the RFs and not on the location of the RFs:

$$\gamma(\mathbf{h}) = \frac{1}{2}\mathrm{E}[\{Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})\}^2] \text{ for all } \mathbf{h} \tag{9.8}$$

where $\mathbf{x}+\mathbf{h}$ indicates a distance (and direction) **h** from location **x**.

In terms of the data, the expected semivariance should be the same for all observations separated by a particular lag, irrespective of where the paired observations are located. In practical terms, the geostatistical approach can be applied irrespective of these conditions, but the results will clearly be suboptimal if the data depart markedly from the conditions. In some cases the mean is allowed to vary from place to place, but stay constant within a moving window. This is known as quasi stationarity (Webster and Oliver, 2007).

9.7.1 Variogram

Analysis of the degree to which values differ according to how far apart they are can be conducted by computing the variogram (or semivariogram). With reference to the variogram, the term 'lag' is used to describe the distance and direction by which observations are separated. For example, two observations may be 5 km apart and one may be directly north of the other. In simple terms, the variogram is estimated by calculating the squared differences between all the available paired observations and obtaining half the average for all observations separated by that lag (or within a lag tolerance (e.g. 5±2.5 km) where the observations are not on a regular grid). Semivariance refers to half the squared difference between data values. An example of variogram estimation is given below. Figure 9.9 gives a simple example of a transect along which observations have been made at regular intervals. Lags (**h**) of 1 and 2 are indicated. In this case, therefore, half the average squared difference between observations separated by a lag of 1 is calculated and the process is repeated for a lag of 2 and so on. In many cases the distance between observations will not be regular, so ranges of distances are grouped. The selection of the bin size (e.g. 0–5 km, >5–10 km, >10–15 km, … or 0–10 km, >10–20 km, …) is important. Smaller bin sizes will result in more noisy variograms, while a bin size that is too large will smooth out too much spatial structure and it will not be possible to capture spatial variation of interest. In other words, the plotted values in a variogram with too small a bin size will appear to be widely scattered, while the values in a variogram with a larger bin size will tend to be more similar to neighbouring values on the plot. Finding an appropriate bin size is important in characterizing spatial structure and in guiding the selection and fitting of a model, as detailed below.

The variogram can be estimated for different directions to enable the identification of directional variation (anisotropy). In other words, rather than consider all observations 5 km from a given observation, we may consider only observations that are directly north or south (for example) of the observation of interest within a particular angular tolerance (e.g. north or south ±45 degrees). An example of an anisotropic phenomenon is temperature—Hudson and Wackernagel (1994) showed that average January temperature in Scotland decreased systematically from west to east (a function of the warming effect of the Gulf Stream), but there was no systematic trend in a north to south direction.

In summary, the variogram characterizes the degree of difference in values as a function of the distance by which they are separated. The experimental variogram, $\hat{\gamma}(\mathbf{h})$, relates semivariances to distances (and directions)—it has distance and direction (the lag) on the x axis and semivariance on the y axis. If a property is spatially autocorrelated, we would expect the semivariance to increase as the distance between observations increases.

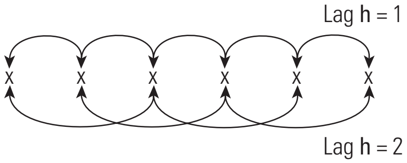

Figure 9.9 Transect with paired points selected for lags of 1 and 2 units.

As an example, if we take a distance range of 1000 to 2000 m and there are 346 pairs of observations separated by a distance within that band then $p(\mathbf{h})$, the number of paired observations, is 346. Note that for each pair (e.g. observations 23 and 37), the semivariance is calculated twice: once with respect to the first location and once with respect to the second. We then calculate the squared difference between each of these paired values. The first value in each pair is given by $z(\mathbf{x}_i)$ and the value separated from it by the specified lag $\mathbf{h}$ (in this example the distance is 1000 to 2000 m and we are concerned with all directions) is given by $z(\mathbf{x}_i+\mathbf{h})$. Their squared difference is therefore given by $\{z(\mathbf{x}_i)-z(\mathbf{x}_i+\mathbf{h})\}^2$. The summed values are then divided by two, hence the term 'semivariance'. Putting this together, the experimental variogram for lag $\mathbf{h}$ is computed from:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2p(\mathbf{h})}\sum_{i=1}^{p(\mathbf{h})}\{z(\mathbf{x}_i)-z(\mathbf{x}_i+\mathbf{h})\}^2 \tag{9.9}$$

As an example, if our lag is 5±2.5 km (i.e. 2.5 to 7.5 km) and we have two values separated by 6.2 km, then these paired observations qualify and we compute the squared difference. If the two values are 26.2 and 43.3 then their squared difference is:

$$\{z(\mathbf{x}_i)-z(\mathbf{x}_i+\mathbf{h})\}^2 = \{26.3-43.4\}^2 = \{-17.1\}^2 = 292.41$$

In the same way we compute the squared difference for all other pairs separated by 2.5 to 7.5 km and at each stage add the computed value to the previous values computed for that lag. Once this is done, we multiply the summed values by $^1/_{(2p(\mathbf{h}))}$.

Figure 9.10 gives an example of an experimental variogram estimated from the precipitation data introduced in Section 9.5. The Gstat software (Pebesma and Wesseling, 1998; Pebesma, 2004) was used to estimate the variogram. The lags are 0–5000 m, 5000–10,000 m and so on in groups of 5000 up to 60,000 m. In this case, data values are

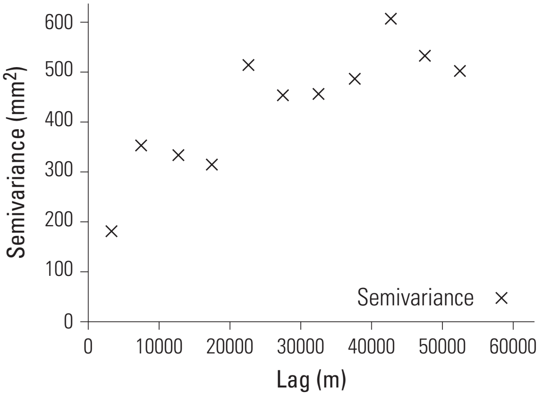

Figure 9.10 Omnidirectional variogram of July 2006 precipitation amount in Northern Ireland.

compared irrespective of the direction in which they are aligned—that is, whether they are aligned (approximately or absolutely) on a line north–south or east–west, etc. of one another is irrelevant. A variogram computed from data in all directions is termed 'omnidirectional'.

In Figure 9.10, the semivariance values tend to be smaller for small lags and they generally increase with an increase in lag size until perhaps 25,000 m, where the values tend to level out (this is demonstrated below). This indicates that values are positively spatially autocorrelated up to approximately this distance. At distances larger than this, there is no spatial structure. The variogram provides a useful means of summarizing how values change with separation distance. Using topography as an example, data representing a 'smooth' surface like a flood plain will have a very different variogram to data representing a 'rough' surface like a mountain range.

A mathematical model may be fitted to the experimental variogram and the coefficients of this model can be used for spatial prediction using kriging or for conditional simulation (defined below). A model can be fitted 'by eye' or by using some fitting procedure such as ordinary least squares (see Sections 3.3 and 8.5, and Appendix E) or weighted least squares. A model is usually selected from one of a set of 'authorized' models. Webster and Oliver (2007) provide a review of some of the most widely used authorized models.

There are two principal classes of variogram model. Transitive (bounded) models have a sill (finite variance)—that is, the variogram levels out as it reaches a particular lag. Unbounded models do not reach an upper bound. Figure 9.11 shows the components of a bounded variogram model. These will be defined and then practical examples given. The nugget effect, c_0, represents unresolved variation (a mixture of spatial variation at a finer scale than the sample spacing and measurement error). The structured component, c, represents the spatially correlated variation. The sill (or sill variance), $c_0 + c$, is the *a priori* variance. The range, a, represents the scale (or frequency) of spatial variation. For example, if a region is mountainous and elevation varies markedly over quite small distances, then the elevation can be said to have a high frequency of spatial variation (a short range), while if the elevation is quite similar over much of the area (e.g. it is a river flood plain) and varies markedly only at the extremes of the

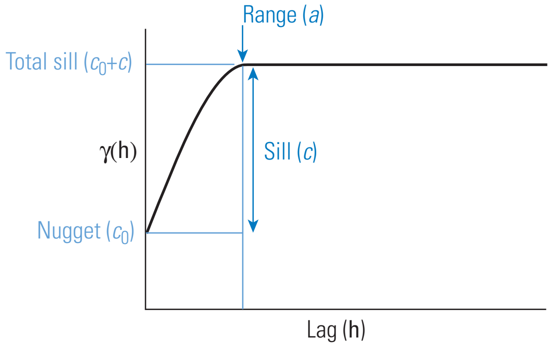

Figure 9.11 Bounded variogram model.

site (i.e. at large separation distances), then the elevation can be said to have a low frequency of spatial variation (a long range).

As noted above, there are many different models that can be fitted to variograms. The variogram illustrated above was fitted with a nugget effect and a spherical component. The nugget effect (nugget variance) is given as:

$$\gamma(h) = \begin{cases} 0 \text{ if } h = 0 \\ c_0 \text{ if } h > 0 \end{cases} \tag{9.10}$$

In other words, the modelled semivariance has a value of 0 for a lag of 0, but is equal to c_0 for all positive values of the lag. In Figure 9.11, the nugget effect is indicated on the y axis of the graph.

The spherical model, a bounded model (i.e. it reaches a sill) is defined as:

$$\gamma(h) = \begin{cases} c \cdot [1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3] & \text{if } h \leq a \\ c & \text{if } h > a \end{cases} \tag{9.11}$$

where ***c*** is called, as noted above, the structured component. In other words, the modelled semivariance is computed using the top line for all lag values up to and including the range. For lag values larger than the range the modelled semivariance is equal to *c*. Authorized models may be used in combination where a single model is insufficient to properly represent the form of the variogram. For example, if the spatial structure is complex and does not simply increase and level out (as in the example in Figure 9.11) then models may be combined to take this complexity into account (e.g. a model could comprise a nugget effect and two spherical components, thus there would be two breaks of slope, rather than just one). Figure 9.12 shows an omnidirectional variogram of July 2006 precipitation amount in Northern Ireland with a fitted model comprising

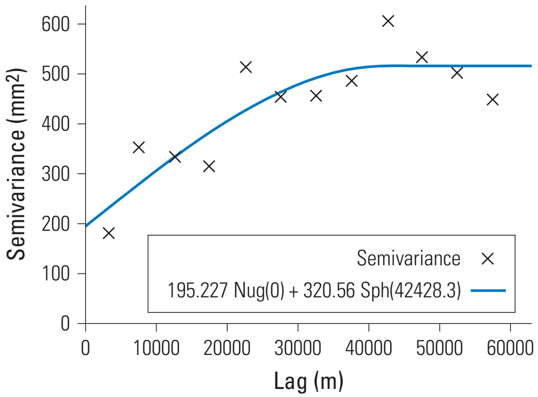

Figure 9.12 Omnidirectional variogram of July 2006 precipitation amount in Northern Ireland, with fitted model.

a nugget effect (with a value of 195.227) and a spherical component (with a structured component of 320.56 and a range of 42428.3 m).

The model fitted to the variogram can be used to determine the weights assigned to observations using a geostatistical prediction procedure (or family of procedures) called kriging. In kriging, the variogram model is used to obtain values of the semivariance for the lags by which observations are separated and for the lags that separate the prediction location from the observation.

For the variogram model in Figure 9.12 there is a nugget effect and a spherical component. Combining Equations 9.10 and 9.11 this gives:

$$\gamma(h) = \begin{cases} c_0 + c \cdot [1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3] & \text{if } h \le a \\ c_0 + c & \text{if } h > a \end{cases}$$

For a lag of 6481.996 m (which is less than the range value of 42428.3 m) the modelled semivariance is obtained from:

$$\gamma(6481.996) = 195.227 + 320.560 \cdot \left[1.5\frac{6481.996}{42428.3} - 0.5\left(\frac{6481.996}{42428.3}\right)^3\right] = 268.116\text{ m}$$

This will be confirmed by examining Figure 9.12 and reading upwards from a lag of 6482 m to the variogram model and then left to read off the semivariance value. Mulla (1988) used variograms, along with other measures, to characterize landforms. As noted previously, in terms of landforms, a mountainous area would have a short range, since there are large changes in elevation over small distances. In contrast, a river flood plain would have a long range as elevations tend to be similar over quite large distances. The variogram is, therefore, a useful tool for measuring the scale of spatial variation in a property. Prediction using kriging, which makes use of the variogram model, is the subject of the following section.

9.7.2 Kriging

There are many varieties of kriging. Its simplest form is called simple kriging (SK). To use SK it is necessary to know the mean of the property of interest and this must be constant across the region of interest. In practice this is rarely the case. The most widely used variant of kriging, ordinary kriging (OK), allows the mean to vary and the mean is estimated for each prediction neighbourhood. OK predictions are weighted averages of the n available data (i.e. the predictions are based on the n nearest neighbours of the prediction location). The OK prediction, $\hat{z}(\mathbf{x}_0)$, is defined as:

$$\hat{z}(\mathbf{x}_0) = \sum_{i=1}^{n} \lambda_i z(\mathbf{x}_i) \qquad (9.12)$$

with the constraint that the weights, λ_i, sum to 1 (this is to ensure an unbiased prediction):

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{9.13}$$

The objective of the kriging system is to find appropriate weights by which the available observations will be multiplied before summing them to obtain the predicted value. These weights are determined using the coefficients of a model fitted to the variogram (or another function such as the covariance function). This is in contrast to IDW (Section 9.5), where the weights are selected arbitrarily (i.e. not using information about the spatial variation in the data).

The weights are obtained by solving (i.e. finding the values of unknown coefficients in) the OK system:

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j) + \psi & = \gamma(\mathbf{x}_i - \mathbf{x}_0) \quad i = 1,\ldots,n \\ \sum_{j=1}^{n} \lambda_j & = 1 \end{cases} \tag{9.14}$$

where ψ is the Lagrange multiplier. This equation may seem at first sight complicated. In words, it says that the sum of the weights multiplied by the modelled semivariance for the lag separating locations $\mathbf{x}_i$ and $\mathbf{x}_j$ plus the Lagrange multiplier equals the semivariance between locations $\mathbf{x}_i$ and the prediction location $\mathbf{x}_0$ with the constraint that the weights must sum to 1. The way we find the weights and the Lagrange multiplier is outlined below.

Computing the weights and a value of the Lagrange multiplier, ψ, allows us to obtain the prediction variance of OK, a by-product of OK, which can be given as:

$$\hat{\sigma}^2_{\mathrm{OK}} = \sum_{i=1}^{n} \lambda_i \, \gamma(\mathbf{x}_i - \mathbf{x}_0) + \psi \tag{9.15}$$

The kriging variance is a measure of confidence in predictions and is a function of the form of the variogram, the sample configuration, and the sample support (the area over which an observation is made, which may be approximated as a point or may be an area) (Journel and Huijbregts, 1978). If the variogram model range is short then the kriging variance will increase markedly with distance from the nearest samples. There are two varieties of OK: punctual OK and block OK. With punctual OK the predictions cover the same area (the support, V) as the observations. In block OK, the predictions are made to a larger support than the observations (e.g. prediction from points to areas of 2 m by 2 m). The system presented here is for the more commonly used form, punctual OK.

Returning to Equation 9.14, using matrix notation, the OK system can be written as:

$$\mathbf{K}\boldsymbol{\lambda} = \mathbf{k} \tag{9.16}$$

where K is the $n+1 \times n+1$ (with n nearest neighbours used for prediction) matrix of semivariances between each of the observations:

$$\mathbf{K} = \begin{bmatrix} \gamma(\mathbf{x}_1 - \mathbf{x}_1) & \cdots & \gamma(\mathbf{x}_1 - \mathbf{x}_n) & 1 \\ \vdots & \vdots & \vdots & \vdots \\ \gamma(\mathbf{x}_n - \mathbf{x}_1) & \cdots & \gamma(\mathbf{x}_n - \mathbf{x}_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}$$

λ are the OK weights and **k** are semivariances for the observations to the prediction location (with one placed in the bottom position):

$$\lambda = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \psi \end{bmatrix} \qquad \mathbf{k} = \begin{bmatrix} \gamma(\mathbf{x}_1 - \mathbf{x}_0) \\ \vdots \\ \gamma(\mathbf{x}_n - \mathbf{x}_0) \\ 1 \end{bmatrix}$$

To obtain the OK weights, the inverse of the data semivariance matrix is multiplied by the vector of data to prediction semivariances:

$$\lambda = \mathbf{K}^{-1}\mathbf{k} \tag{9.17}$$

The OK variance is then obtained from:

$$\sigma^2_{\mathrm{OK}} = \mathbf{k}^T \lambda \tag{9.18}$$

Using the same data as for the example in Sections 9.5 (IDW) and 9.6 (TPS), the OK system is given as:

$$\begin{bmatrix} 0 & 376.905 & 359.589 & 379.853 & 1 \\ 376.905 & 0 & 307.108 & 394.601 & 1 \\ 359.589 & 307.108 & 0 & 448.401 & 1 \\ 379.853 & 394.601 & 448.401 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \psi \end{bmatrix} = \begin{bmatrix} 268.116 \\ 311.250 \\ 311.983 \\ 367.662 \\ 1 \end{bmatrix}$$

Note that the semivariance between a given location and itself is set to 0.

This account does not show how the weights are obtained. To see how this is done (i.e. to see how the OK system is solved) go to Appendix F, where the same example is given (but exactly how the system is solved is shown).

Solving the OK system, the weights are as follows: $\lambda_1 = 0.368$, $\lambda_2 = 0.227$, $\lambda_3 = 0.234$, $\lambda_4 = 0.171$, and $\psi = 33.332$.

The predicted value is then given by: $(0.368 \times 68) + (0.227 \times 29) + (0.234 \times 48) + (0.171 \times 53) = 51.889$.

The kriging variance is given by: $(0.368 \times 268.116) + (0.227 \times 311.250) + (0.234 \times 311.983) + (0.171 \times 367.662) + (33.332 \times 1) = 338.537$.

The kriging variance is a useful by-product which, as detailed above, provides a guide to uncertainty in predicted values.

The 'true' precipitation amount value is 61 mm, so there is a prediction error of 9.111. In this case, the IDW prediction of 55.003 (see Section 9.5) is closer to the true value, as is the TPS prediction of 60.569 (see Section 9.6). There is, of course, no guarantee that OK will provide more accurate predictions than IDW, despite the use of arbitrary weights in the latter case, but many real-world case studies have shown how techniques like OK often provide an increase in prediction accuracy over simpler methods like IDW (see Lloyd (2005) for an example). Lloyd (2006) and Chang (2008) provide worked examples of IDW, TPS, OK, and other approaches.

Figure 9.13 shows a map of precipitation in July 2006 generated using OK with 16 nearest neighbours.

Comparison of Figure 9.13 with Figures 9.7 (IDW derived map) and 9.8 (TPS derived map) shows quite large differences in the range of values. This demonstrates the large variations that can result from the application of different interpolation procedures. This issue is discussed further below.

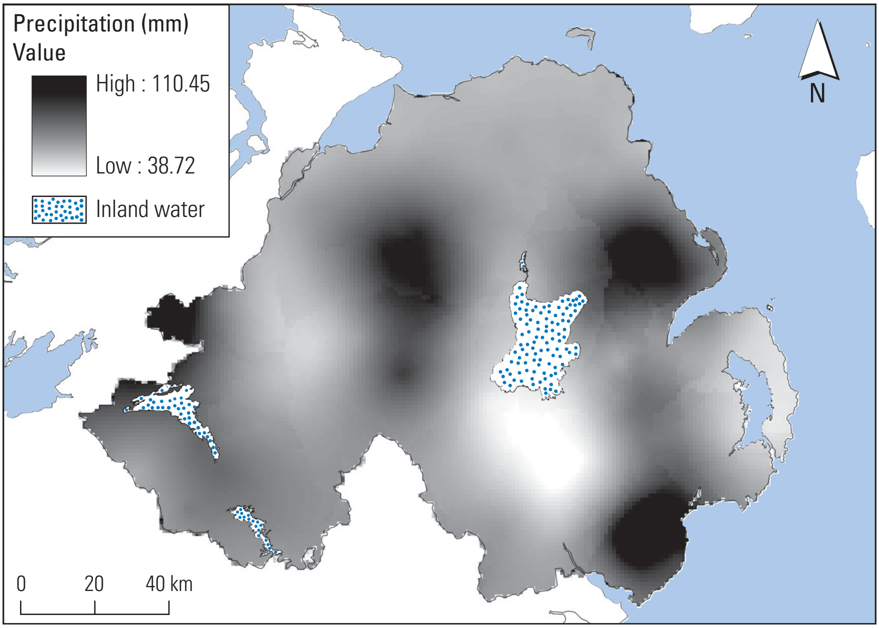

Figure 9.13 Precipitation in July 2006: OK prediction using 16 nearest neighbours.

9.7.3 Cokriging

There are several other forms of kriging; cokriging, for example, allows the integration of information about secondary variables. In cases where we have a secondary variable (or variables) that is cross-correlated with the primary variable, both (or all) variables may be used simultaneously to make predictions using cokriging. With cokriging, the variograms (which can be termed 'autovariograms') of both (or all) variables and the cross-variogram (describing the spatial dependence between the two variables) must be estimated and models fitted to all of these. Cokriging is based on the linear model of coregionalization (see Webster and Oliver, 2007). For cokriging to be beneficial, the secondary variable should be cheaper to obtain or more readily available than the primary variable (i.e. the variable that will be mapped). If the variables are strongly related linearly then cokriging may provide more accurate predictions than OK.

9.8 Other approaches and issues

There are many other widely used spatial interpolation approaches in addition to variants of TPS and kriging. Several approaches are summarized by Mitás and Mitásová (1999). Specialist routines have been written for some applications. For example, the routine of Hutchinson (1989) is used specifically for generating DEMs. Clearly, the selection of an interpolation method impacts on the final results, and researchers have assessed variations in results following application of different interpolation methods (see Chapter 10 of Burrough and McDonnell (1998) for a review of related topics). The performance of different spatial interpolation procedures will vary as a function of sampling density and spatial variation. For example, if the sampling density is low (there are large distances between samples) and there is short-range spatial variation (values differ a great deal over short distances), then we would expect there to be larger differences between results obtained using different procedures than in cases where the sampling density is high and there is long-range spatial variation. Lloyd and Atkinson (2002) show how differences in predictions (in terms of their accuracy) increase as the sampling density decreases, and the benefits of more sophisticated approaches are shown to be more apparent where the sampling density is low.

9.9 Areal interpolation

The focus so far in this section has been on point interpolation—that is, prediction from a point sample to a regular grid. Often, there is a need to transfer between different sets of zones or transfer, for example, counts from zones (such as census reporting areas) to grids (Martin *et al.*, 2002). Many techniques exist for solving such problems. In the case of transferring values between different sets of zones, overlay procedures (as detailed in Chapter 5) provide a partial solution. Counts could be reassigned to new zones

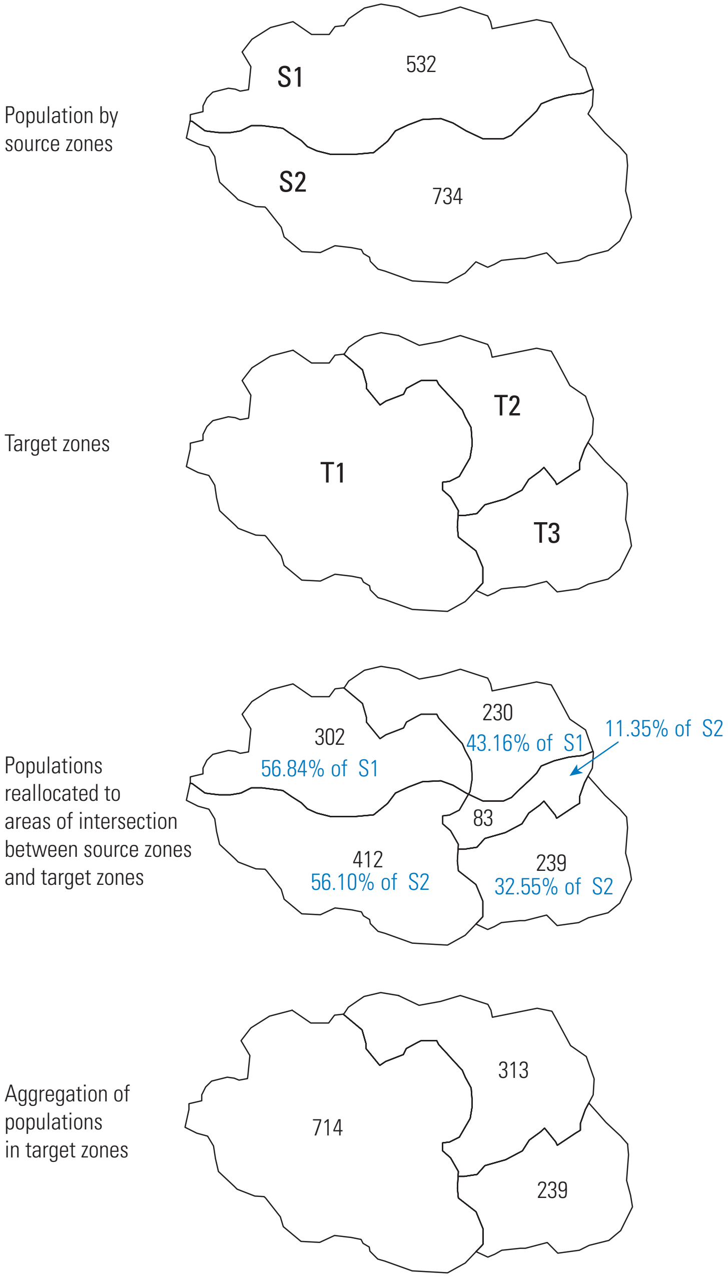
Population by source zones
S1
532
S2
734
Target zones
T1
T2
T3
Populations reallocated to areas of intersection between source zones and target zones
302
56.84% of S1
230
43.16% of S1
11.35% of S2
83
412
56.10% of S2
239
32.55% of S2
Aggregation of populations in target zones
714
313
239

Figure 9.14 Areal weighting example.

according to the size of the overlapping areas between the old zone set and the new zone set. Such an approach is called areal weighting and an example is given in Figure 9.14.

As a particular example, zone T1 covers 56.84% of the area of zone S1. Given this information, the expected population of the area of overlap (i.e. intersection) between zones T1 and S1 is 56.84% of the zone S1 population of 532, thus ($^{532}/_{100}$)×56.84 = 302 (when rounded to a whole number). Once the populations of each of the areas of intersection have been obtained they can be summed within the target zones, as shown in the bottom part of Figure 9.14.

More sophisticated approaches to areal interpolation exist (see Lloyd (2006) for a summary). The main focus in this chapter is on the generation of surfaces rather than zones, and so of more immediate relevance here are approaches such as the pycnophylactic reallocation method of Tobler (1979) or the population surface modelling procedure of Martin (1989). Both of these approaches allow the transfer of zonal counts to regular grids. One benefit of such approaches is to enable direct comparison of values for different time periods even when the original zonal systems used at different periods are quite different.

9.10 Case studies

The following two case studies are based on the data introduced in Section 8.7. These case studies demonstrate (1) estimation of the variogram and (2) spatial interpolation using IDW, TPS, and OK.

9.10.1 Variogram estimation

Figure 9.15 shows an experimental variogram computed from precipitation data. Recall that the variogram tells us how different observations tend to be a function of how far apart the observations are. In this case, the semivariance values increase with increased distance and they level out at a distance of perhaps 75 km. A model can be fitted to the variogram, as detailed in Section 9.7.1, and the model used to inform prediction of precipitation amount at locations where there are no measurements

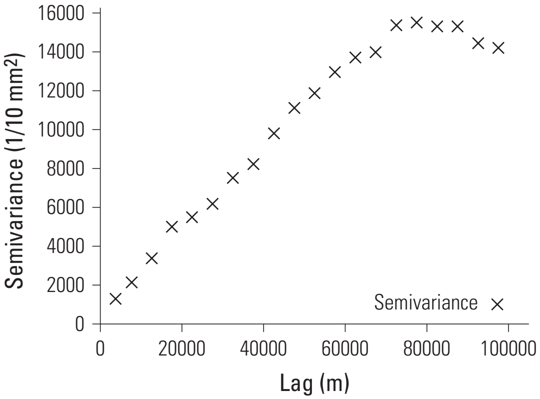

Figure 9.15 Experimental variogram computed from precipitation data for 8 May 1986 in Switzerland.

(see Section 9.7.2 for a discussion about this). The use of a fitted model for interpolation is detailed in the next subsection. In this case, the geostatistical software Gstat (Pebesma and Wesseling, 1998; Pebesma, 2004) was used for the analysis; ArcGIS™ Geostatistical Analyst could also be used (although the variogram will appear slightly different given the way the semivariances are computed in that package).

9.10.2 Interpolation

Precipitation amounts were predicted on a regular grid using the IDW, TPS, and OK functions of ArcGIS™ Geostatistical Analyst. For OK, the variogram was estimated using a lag size of 5000 and 20 lags, with the model (nugget effect = 403.9, spherical component (partial sill) = 14689, and range = 90653.3 m) fitted using Geostatistical Analyst. Figure 9.16 shows a grid generated using the 16 nearest neighbours with IDW.

Cross-validation was used to compare the three methods using 16 nearest neighbours. When using cross-validation to compare prediction approaches, it is usual to use summaries of the errors such as the mean error or the RMSE. As its name suggests, the RMSE is the root of the mean squared errors (i.e. take each error, square it, take the average of these squared errors, and then take the square root of this average); this is a widely used summary measure and is interpreted below. The mean error indicates bias: if it is negative then the procedure tends to under-predict values and when it is positive

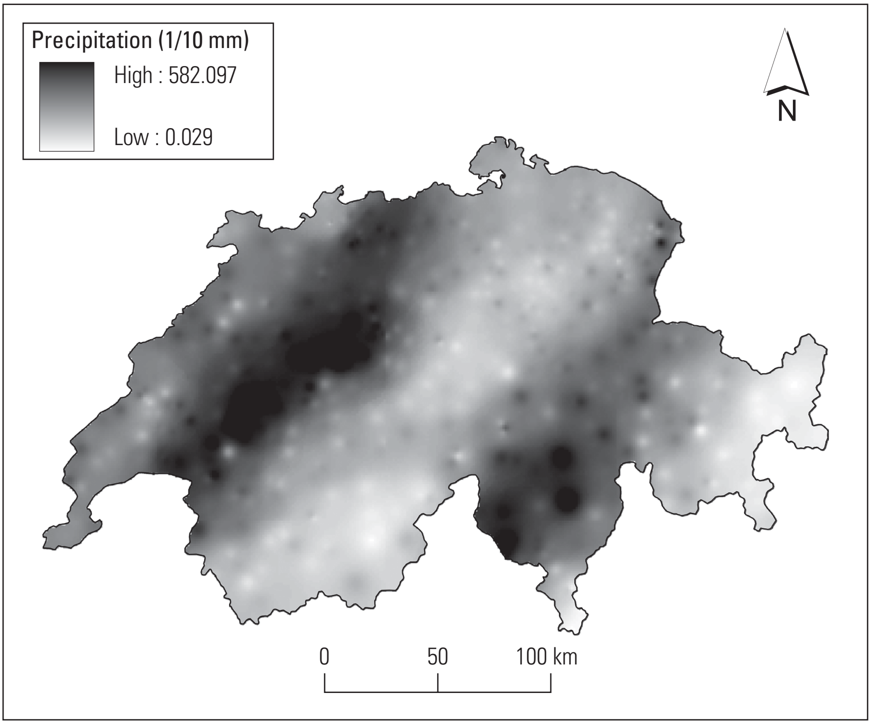

Figure 9.16 Precipitation on 8 May 1986: IDW prediction using 16 nearest neighbours.

it suggests over-prediction. Ideally, therefore, the mean error would be 0. The RMSE represents the magnitude of errors and a small RMSE value corresponds to accurate predictions. With IDW, the mean cross-validation error was –0.520 and the RMSE was 47.82. The figures for TPS were mean error –0.418 and RMSE 53.87. For OK the figures were mean error –0.189 and RMSE 47.81. In this case, the least biased predictions (mean error closest to 0) are provided by OK. The smallest errors overall (as measured by the RMSE) are also for OK, although the difference between the OK and IDW RMSE values is very small. Note that TPS predictions are the least accurate in this case, but that more accurate predictions are obtained when other variants of TPS (e.g. TPS with tension, which can be conducted in ArcGIS™, and is defined by Lloyd (2006)) are used.

Summary

This chapter provided an introduction to some of the most widely used approaches to the generation of surfaces from point data. In addition, a short outline of areal interpolation was given. In terms of selection of methods, it was noted that differences in prediction results are a function of sampling density and spatial variation. Approaches like cross-validation offer a way of assessing the performance of different approaches. However, such approaches should not be used blindly and other approaches, such as jackknifing (predicting to one set of locations (at which there are observed values) using a second data set and computing the errors of the predictions), are likely to be more robust.

Further reading

More information on spatial interpolation is provided by **Burrough and McDonnell (1998)**, **Lloyd (2006)**, and **Chang (2008)**, for example. There are many introductions to geostatistics (e.g. **Goovaerts, 1997**; **Armstrong, 1998**; **Webster and Oliver, 2007**; **Atkinson and Lloyd, 2009**). Many different applications of interpolation procedures can be found in the literature. Interpolation has been used to map elevation (**Lloyd and Atkinson, 2002**), precipitation amount (**Goovaerts 2000**; **Lloyd 2002**, **2005**, **2009**, **2010**), and airborne pollutants (**Lloyd and Atkinson, 2004**), amongst many other variables.

➜ The next chapter is concerned with the analysis of gridded data and the latter part of the chapter has a particular focus on the analysis of DEMs.