SPATIAL INTERPOLATION TECHNIQUES (1)

Interpolation refers to the process of estimating the unknown data values for specific locations using the known data values for other points.

In many instances we may wish to model a feature as a continuous field (i.e. a 'surface'), yet we only have data values for a finite number of points. It therefore becomes necessary to interpolate (i.e. estimate) the values for the intervening points. For example, we may have measurements of the depth of a particular geological stratum from a number of bore holes, but if we want to model the stratum in 3-dimensions then we need to estimate its depth for places where we do not have bore hole information.

Interpolation may also be required in other situations. For example, we might want to convert from a raster with a particular cell size or orientation to a raster with a different cell size or orientation. This procedure is known as convolution. Alternatively we might want to convert from a raster data model to a TIN, or vice versa. Interpolation would again be required.

When interpolating from a sample of points we would normally express the estimated values of the intervening points in the same units as were used for the measurements at the sample points, but sometimes we may be more interested in the probability that a certain value is exceeded or that the interpolated value is within a certain range.

Most interpolation methods can be divided into two main types called global and local. Global interpolators use all the available data to provide estimates for the points with unknown values; local interpolators use only the information in the vicinity of the point being estimated. Global interpolators are often used to remove the effects of major trends before using local interpolators to analyse the residuals. Kriging is a particular type of local interpolation using more advanced geostatistical techniques.

Interpolation methods may also be classified as exact or inexact. Using exact interpolation, the predicted values at the points for which the data values are known will be the known values; inexact interpolation methods remove this constraint (i.e. the observed data values and the interpolated values for a given point are not necessarily the same). Inexact interpolation may produce a smoother (and arguably more plausible) surface.

Interpolation methods may be either deterministic or stochastic. Deterministic methods provide no indication of the extent of possible errors, whereas stochastic methods provide probabilistic estimates.

We will begin by looking at some of the more common forms of spatial sampling; then we will look at global and then local interpolators, paying particular attention to a geostatistical technique known as kriging (a local, exact and stochastic method).

To make the discussion more concrete we will make occasional references to the example discussed by Burrough and McDonnell in which the objective is to estimate the concentration of zinc in the soils in a region in the southern Netherlands, using data measured at 98 sample points. Apart from zinc concentrations, data were recorded at the sample points on other easily measured variables such as elevation, distance from the river Maas, frequency of flooding, and soil type. The data are listed in Appendix 3 of Burrough and McDonnell. Burrough and McDonnell compare the results of different techniques using these data. If you wish to experiment with the data yourself, they may be downloaded in Excel, dBase and SPSS formats using the Gis09.bat option from the icon for this module on the desktop.

SPATIAL SAMPLING

Interpolation involves making estimates of the value of an attribute variable at points for which we have no information using the data for a sample of points where we do have information. In general the more sample points we have the better. Likewise it is best to have a good spread of sample points across the study area. The spatial arrangement of these points can take different forms (see figure overleaf).

Regular sampling guarantees a good spread of points, but it can result in biases if the attribute to be mapped has regular fluctuations (e.g. soil moisture where there are regular spaced drains). It is therefore preferable to have some form of random sampling, where the co-ordinates of the points are selected using random numbers. Pure random sampling tends to produce a pattern with little clusters in some areas and a sparse coverage in other areas, so a
stratified random sample in which points are randomly allocated within a regular spaced lattice provides a good compromise. The other three types shown overleaf are for special purposes: cluster (or nested) sampling provides detailed information on selected areas and is sometimes used to examine spatial variation at different scales (e.g. you could compare within the amount of variation within a cluster with the variation between clusters); transect sampling is used to survey profiles; and contour sampling is sometimes used to sample printed maps to make a DEM (although, as we shall see next day, this can be problematic).

The area or volume of the sample at each sample point upon which measurements are made is referred to as the support. For example, if measuring the mineral content of soils, the support would be the amount of soil (e.g. 10 x 10 x 5 cm) which is taken for analysis at each location. In human geography, the support might be an irregular sized and shaped area (e.g. an ED).

GLOBAL INTERPOLATORS

There are two broad approaches to global interpolation. One uses classification techniques to infer the values of one variable attribute based upon a knowledge of the values of another attribute. The other uses regression techniques to infer the value of the variable of interest based upon a knowledge of attributes that are easy to measure.

Classification Techniques

This approach may be used if spatial data are very sparse, although regression techniques (see below) would generally be preferred if there is sufficient data. Classification techniques are global, inexact and deterministic. The basic assumption is that the value of the variable of interest is strongly influenced by another variable which can be used to classify the study area into zones. Given that main source of the heavy metals in the soils in our case study area is from flooding by the river Maas, it seems plausible that the zinc levels are a function of the frequency of flooding. The study area can be divided into three zones based upon the frequency of flooding. For each point in the study area, the zinc concentrations can be thought of as comprising three components:

\[ z(x_0) = \mu + \alpha_k + \varepsilon \]
where \( z \) is the zinc concentration at location \( x_0 \), \( \mu \) is the overall mean zinc concentration, \( \alpha_k \) is mean zinc concentration in flooding class \( k \) (relative to the overall mean), and \( \varepsilon \) is a random noise factor (which governs the variations within each class).

The best estimate of the zinc concentrations for each point in flood class \( k \) (i.e. \( \mu + \alpha_k \)) can be calculated by taking the mean of the values for the sample points falling within flood class \( k \). The actual (but unknown) value at each point to be extrapolated will vary from the estimated value by a random amount determined by \( \varepsilon \).

Standard analysis of variance tests can be conducted to gauge the usefulness of the classification. If the F ratio (defined as the ratio of the between class variance to the pooled within class variance) based on the information for the sample points is low, then the classification is weak.

The approach makes a number of basic assumptions:
1. Variations in the value of \( z \) within a flooding class are random (i.e. there is no spatial autocorrelation).
2. The variance of the noise is the same for all flooding classes.
3. The zinc concentrations are normally distributed.
4. Spatial changes in zinc concentrations occur in steps at the boundaries of the flooding classes.

Most of these assumptions probably do not hold. If the data are non-normal, they can be transformed to make them more normal, but there is little that can be done about the other assumptions.

**Trend Surface Analysis**

Trend surface analysis is **global, inexact** and **deterministic**. A trend surface can be thought of as a high order three dimensional regression surface. To understand this, let us begin with a simple situation in which we can model the data values along a transect using a simple regression model:

\[
    z(x) = b_0 + b_1x + \varepsilon
\]

where \( z(x) \) is the zinc concentration at location \( x \), \( b_0 \) is the intercept (i.e. value of \( z \) when \( x=0 \)), \( b_1 \) is the slope or gradient, and \( \varepsilon \) is the residual (error term or noise).

In some cases the data values cannot be adequately summarised as a linear function, in which case a higher order polynomial may provide a better summary. For example, a second order polynomial (or quadratic) equation might provide a better fit:

\[
    z(x) = b_0 + b_1x + b_2x^2 + \varepsilon
\]

Trend surfaces are similar except, instead of having data values along a transect, the sample points would be in two dimensions (measured by \( x \) and \( y \) co-ordinates) with the attribute values \( z \) modelled as a third dimension. A first order trend surface (analogous to a simple regression line) is an inclined plane with the formula:

\[
    z(x, y) = b_0 + b_1x + b_2y + \varepsilon
\]

A second order trend surface would be an undulating surface with the formula:

\[
    z(x, y) = b_0 + b_1x + b_2y + b_3x^2 + b_4xy + b_5y^2 + \varepsilon
\]

It should be noted that this model includes a **cross product** term (i.e. \( b_{xy} \)). Higher order trend surfaces not only include even higher powers of \( x \) and \( y \), but also more cross product terms. A third order trend surface has a total of 10 terms. Higher order trend surfaces are more convoluted than lower order trend surfaces, and provide closer fits to the observed data values. However, this does not necessarily result in more accurate predictions for the points in between. In fact, trend surfaces higher than third order tend to become counter productive. The objective in many instances, consequently, is not to get a ‘perfect’ fit of the observed data values using a higher order trend, but to identify areas where there are spatially autocorrelated residuals from a low order trend surface as this may indicate the presence of locally important influences upon the variable of interest.
The values of the b terms can be easily determined using the standard regression options available in most statistical packages. The significance of a trend surface can be tested using an analysis of variance test. There is also an analysis of variance test to test whether a trend surface of a given order represents a significant improvement on a trend surface one order lower. (See Burrough and McDonnell, p111, for details).

**Regression Techniques**

The trend surface techniques discussed in the previous section use only information measured on the variable of interest (i.e. zinc concentrations) and the location of the sample points, whereas the classification methods discussed in the previous section make use of 'external' information (i.e. frequency of flooding). A third strategy is to make use of external information using regression techniques. Regression techniques are **global**, **inexact** and **stochastic**.

There is no restriction on the type of external information that may be used, provided that the regression model is intuitively plausible. However, it obviously makes sense to use information which can be readily obtained.

The example provided by Burrough and McDonnell models zinc concentrations using the model:

\[
\hat{z}(x) = b_0 + b_1P_1 + b_2P_2 + \varepsilon
\]

where \(P_1\) is distance from the river, and \(P_2\) is elevation.\(^1\) Information on distance from the river and elevation can be readily calculated for any given point using a GIS. In the absence of a DEM from another source, the elevation data collected for the sample points can be used to construct one. Both variables can be assumed to exercise an influence upon the likelihood of flooding and therefore of zinc deposition. The regression model determines the relative importance of the two variables using other information available for the sample points. The parameters of the regression model can obviously be estimated using standard regression techniques. Likewise the significance of the regression can be tested using standard techniques. This type of regression which empirically estimates the values of a variable using external information is sometimes referred to as a **transfer function**.

**Other Global Interpolators**

More complex mathematical techniques, such as **spectral analysis** or **Fourier analysis**, can be used to model the surface in a manner analogous to trend surface analysis.\(^2\) They generally require large amounts of data at different scales of resolution.

**LOCAL INTERPOLATORS**

When using global interpolators local variations tend to dismissed as random noise. However, intuitively this does not make sense as the data values for each point often tend to be very similar to those for neighbouring points. Local interpolators therefore attempt to estimate the data values for unknown points using the known data values for points nearby.

The general procedure is to identify a lattice of points for which data values are to be estimated. For each point, the procedure involves the following steps:

1. A search area (neighbourhood) is defined around the point;
2. The sample points within the search area are identified;
3. A mathematical function is selected to model the local variation between these points;
4. The data value for the point is estimated from the function.

\(^1\) The dependent variable could be represented as \(z(x,y)\) to maintain consistency with the previous section. However, the handout follows the notation used by Burrough and McDonnell - i.e. the dependent variable is represented as \(z(x)\), where \(x\) refers to the set of cartesian co-ordinates for a particular point. Other texts ignore the locational qualifiers completely, and simply refer to the dependent variable as \(z\).

\(^2\) The main difference is that, instead of using polynomials, the surface is modelled as the sum of a number of sinusoidal functions with different wavelengths.
Different results will tend to arise depending upon the size of the search area and the type of mathematical function selected. We will discuss a few of the more common functions.

**Thiessen Polygons**

In this approach, Thiessen polygons (also known as Dirichlet or Voronoi diagrams) are constructed around each sample point. All points within a polygon are assumed to have the same data value as the sample point around which the polygon is constructed. This is equivalent to saying that each point has the same data value as its nearest sample point. Interpolation using Thiessen polygons would be classed as local, exact and deterministic.

Thiessen polygons are constructed by drawing lines between neighbouring points - in our case sample points. These lines form the sides of Delaunay triangles. A Delaunay triangle has the property that a circle drawn through its three corners will never contain any other sample point. If the circle contains a sample point, then the triangles need to be redrawn. To construct the Thiessen polygons, a second set of lines are then constructed to bisect the first set of lines (i.e. the edges of the Delaunay triangles) at right angles at their mid-points. The second set of lines form the boundaries of the Thiessen polygons whilst their intersections form the corners of the polygons. The key property of a Thiessen polygon is that all points within a polygon lie closer to the point around which it was constructed than to any other point.

Using Thiessen polygons to interpolate results in sharp jumps in data values as you move from one polygon to the next, so a technique known as pycnophylactic interpolation is sometimes used to smooth the transition (see diagram below). This technique was developed by Waldo Tobler to 'blur' the boundaries in choropleth maps showing features like population density for administrative regions. The technique preserves 'volume' (i.e. the total number of people per area), but 'moves' them around within the areas to form a continuous surface. Burrough and McDonnell (pp.116-117) provide details of the mathematics.

A more traditional cartographic technique, known as dasymetric mapping uses a somewhat similar approach, except that the location of different shading categories within a polygon is based upon external information. For example, if the mean population density in an administrative area is 500 people per square mile, but it is known that exactly half the area is an uninhabited bog, then the population density in the bog area could be allocated a shading

---

3 Figure 5.7 in Burrough and McDonnell, purporting to show Thiessen polygons (and replicated here), is poorly constructed, resulting in some areas being allocated to the wrong polygon.
indicating 0 people per square mile, and the rest of the area could be allocated a shading indicating 1,000 people per square mile.

**Weighted Moving Average Methods**

This family of techniques estimates the data value for each point by calculating a distance weighted average of the points within the search radius. These techniques are **local**, **exact** and **deterministic**.

The general formula is:

\[
\hat{z}(x_0) = \sum_{i=1}^{n} \lambda_i z(x_i)
\]

where \(z(x_i)\) are the data values for the \(n\) points \((x_1…x_n)\) within the search radius, and \(\lambda_i\) are weights to be applied to the data values for each point. One constraint is that the weights must add up to 1.0. The weights are usually some function of the distance between the point for which the estimate is being made and the sample points. The most common function is the **inverse distance weighting** (IDW) predictor. The above formula then becomes:

\[
\hat{z}(x_0) = \sum_{i=1}^{n} \left( \frac{d_{ij}^{-r}}{\sum_{j=1}^{n} d_{ij}^{-r}} \right) z(x_i)
\]

where \(j\) represents the point whose value is being interpolated, \(d_{ij}\) is the distance from point \(j\) to sample point \(i\) and \(r\) is an arbitrary value which can be selected by the investigator. If \(r\) is set equal to 1, then this becomes a simple **linear interpolator**. The \(r\) value is frequently set to 2, in which case the influence of each sample point is in proportion to the square root of its distance from the point to be interpolated. However, \(r\) can be set to higher values if required. High values of \(r\) give even more weight to the nearer sample points.

If the point to be interpreted corresponds exactly with one of the sample points, \(d_{ij}\) would be zero and the sample point would be assigned an infinite weight, producing an infinite value for the estimate. This is obviously undesirable, so in such situations the point to be interpolated is simply assigned the same value as the sample point.

One problem with this particular technique is that isolated data points tend to produce ‘duck-egg’ patterns. One solution might be to increase the search radius, but this may have the undesirable effect of over-smoothing other parts of the surface. Another variant is to base the interpolation at each point on a fixed number of sample points. The search radius will then be small in areas where there is a high density of sample points, but larger where there is a lower density of points. If it is believed that there may be a directional bias, the search window could even be divided into, say, 4 quadrants and its radius might be increased until there was a minimum of a specified number of points in each quadrant.

**Splines**

Before computers cartographers used flexible rulers called splines to fit smooth curves through a number of fixed points. Spline functions are the mathematical equivalents of these flexible rulers. Splines are **piece-wise** functions - i.e. they consist of a number of sections, each of which is exactly fitted to a small number of points, in such a way that each of the sections join up at points referred to as **break points**. One advantage of piece-wise functions is that if there is a change in the data value at a point, then it is only necessary to make a local adjustment (i.e. to the relevant section) whereas, in contrast, the whole surface would need to be completely recalculated if using a trend surface. Splines may be classes as **local**, **exact** and **deterministic**.

The splines are normally fitted using low order polynomials (i.e. second or third order) constrained to join up. The splines may be two dimensional (e.g. if smoothing a contour line) or three dimensional (if modelling a surface). Third order three dimensional splines are frequently used. These are sometimes referred to as a **bicubic spline**.
Most practical applications use a special type of spline called a **B-spline**. These are described by Burrough and McDonnell as ‘themselves the sums of other splines that by definition have the value of zero outside the interval of interest’ (p.119).

The results of applying a spline are dependent upon various decisions taken by the investigator. These include the order of polynomial selected, the number of break-points, and also whether the break points are selected to correspond with sample points or at points in between. It is also possible to relax the requirement that the spline should fit the sample points exactly. Some exact splines can produce excessively high or low values. **Thin plate splines** (an inexact alternative) are therefore sometimes used to smooth the surface, subject to the constraint that the difference between the observed data values and those given by the function are minimised.

**Kriging**

Kriging is a form of local interpolation using geostatistical methods developed by a French geostatistician called Georges Matheron and a South African mining engineer called D.G. Krige. Kriging is **local** and **exact** and **stochastic**. Whilst much more complex than the methods discussed above, it provides a number of advantages:

1. Given sufficient data, kriging produces better estimates than the other methods because the method takes explicit account of the effects of random noise.
2. Although the investigator can choose between different types of kriging, kriging is less susceptible to arbitrary decisions taken by the investigator (e.g. search distance, number of sample points to use, location of break points, etc.). Kriging identifies the optimal interpolation weights and search radius.
3. Kriging provides an indication of the reliability of the estimates.

Kriging begins with the recognition that the spatial variation of any continuous attribute is usually too complex to be modelled by a simple, smooth mathematical function. So, instead, it is modelled as a stochastic surface or random field. Regionalized variable theory assumes that the spatial variation of any variable can be expressed as the sum of three components:

1. A structural component having a constant mean or trend;
2. A random, but spatially correlated component, known as the **regionalized variable**; and
3. Spatially uncorrelated random noise i.e. residual component.

The value of a random variable \( Z \) at \( x \) is given as:

\[
Z(x) = m(x) + \epsilon'(x) + \epsilon''
\]

where \( m(x) \) is a structural function describing the structural component, \( \epsilon'(x) \) is the stochastic but spatially autocorrelated residuals from \( m(x) \) (i.e. the regionalized variable), and \( \epsilon'' \) is random noise having a normal distribution with a mean of 0 and a variance \( \sigma^2 \).

The first step is to decide on a suitable function for \( m(x) \). In the simplest case this can be thought of as a flat surface with no trend. The mean value of \( m(x) \) is the mean value within the sample area. This means the expected difference in the values for two points \( x \) and \( x+h \) (where \( h \) is the distance between the points) is zero. i.e.

\[
E[Z(x) - Z(x + h)] = 0
\]

It is also assumed that the variance of the differences is a function of the distance between the points (i.e. they are spatially autocorrelated, meaning that near points are more likely to have similar values, or smaller differences, than distant points) - i.e.

\[
E[(Z(x) - Z(x + h))^2] = E[(\epsilon'(x) - \epsilon'(x + h))^2] = 2\gamma(h)
\]

where \( \gamma(h) \) is known as the **semivariance**. Under these two assumptions (i.e. stationarity of difference and stationarity in the variance of differences), the original model can be expressed as:

\[
Z(x) = m(x) + \gamma(h) + \epsilon''
\]

The semivariance can be estimated from the sample data using the formula:
\[ \gamma(h) = \frac{1}{2n} \sum_{i=1}^{n} \left[ z(x_i) - z(x_i + h) \right]^2 \]

where \( n \) is the number of pairs of sample points separated by distance \( h \).

The semivariance can be calculated for different values of \( h \). A plot of the calculated semivariance values against \( h \) is referred to as an experimental variogram.

Experimental variograms typically have a number of characteristic features, as illustrated in the diagram:

1. The variance is small for low values of \( h \), but increases as the value of \( h \) gets larger. However, beyond a certain point the graph levels off to form what is known as the sill.
2. The distance at which the graph levels off is known as the range. At distances less than the range, points closer together are more likely to have similar values than points further apart. At distance larger than the range, points do not exert an influence upon one another. The range therefore provides a indication of how large the search radius needs to be when doing a distance weighted interpolation.
3. The fitted model does not pass through the origin - i.e. according to the graph the semivariance when \( h \) is zero has a positive value referred to as the nugget. However, one would expect the variance to be zero (i.e. one would expect the difference between points and themselves to be zero). The nugget provides an indication of the amount of non-spatially autocorrelated noise (i.e. \( \sigma^2 \)).

The semivariance depicted in an experimental variogram must be modelled by a mathematical function taking the form \( \gamma(x) = \ldots \). Different models may be used, each with a different formula. The choice between these models is determined by the shape of the experimental variogram. The figure below shows some of the more commonly used models. A spherical model (a) is used when the variogram has the 'classic' shape, an exponential model (b) is used when the approach to the sill is more gradual, a Gaussian model (d) may provide a good fit when the nugget is small and the variation is very smooth, and a linear model (c) may be the most appropriate when there is no sill within the study area. Fitting the most appropriate model to the data requires a lot of skill.

Other variogram shapes may indicate a different course of action is required. For example,

1. A variogram that becomes increasingly steep with larger values of \( h \) indicates that there is a trend in the data that should be modelled separately.
2. If the nugget variance is large and the variogram shows no tendency to diminish with smaller values of \( h \), then interpolation is not really sensible. The best estimate of \( z(x) \) is simply the overall mean of the sample points.
3. A noisy variogram showing no particular pattern may indicate that there are too few sample points.
4. If the range is smaller than the distance between sample points, then the sample points are too far apart to influence one another. The best estimate of \( z(x) \) is again the overall mean of the sample points.
5. If the variogram dips at distances further than the range to create a 'hole effect', then the study area may be too small to capture some long wave-length variation in the data.
More complex variogram models can be developed where it is felt they are required. For example:

1. **Anisotropic** models may be developed if it is believed that there may be directional effects (i.e. that the relationship between the semivariance and distance is different in different directions). This essentially involves calculating a different experimental variogram for each direction. These can be shown as a contour 'map' rather than a graph.

2. A **complex** variogram may be required if it is believed that the total variation is the sum of the effects of two or more regionalized variables.

3. Different variograms may be required for different **cover classes** (e.g. rock types, land covers, etc.). In such situations a separate variogram should be calculated for each cover class.

Once the variogram has been modelled, the next step is to determine the weights \( \lambda_i \) required for local interpolation. The weights \( \lambda_i \) for each point should sum to 1.0. They are also selected so that the estimated data value for the point is unbiased, and the variance of the estimation is minimised - i.e. the estimate is a best linear unbiased estimate (BLUE). It can be shown that the estimation variance \( \sigma_e^2 \) is minimised when

\[
\sum_{i=1}^{n} \lambda_i \gamma(x_i, x_0) + \phi = \gamma(x_j, x_0)
\]

where \( n \) is the number of samples points used to interpolate the data value for point \( x_0 \) and \( \phi \) is a **Lagrange multiplier** (which ensures that the \( \lambda_i \) values add up to 1.0). This equation can be solved by substituting the estimated values of \( \gamma \) from the variogram model. Burrough and McDonnell provide a worked example on pp. 140-141.

Finally, the data value for each point can be estimated by inputting the weights into the formula:

\[
\hat{z}(x_0) = \sum_{i=1}^{n} \lambda_i z(x_i)
\]

The estimation variance \( \sigma_e^2 \) (also known as the **kriging variance**) for each point can be calculated using the formula:

\[
\hat{\sigma}_e^2 = \sum_{i=1}^{n} \lambda_i \gamma(x_i, x_0) + \phi
\]

The estimation variance \( \sigma_e^2 \) (or more usually the **kriging standard deviation** \( \sigma_e \)) can be mapped. This provides an indication of the reliability of the estimate at each location.
The validity of the variogram model may also be tested for consistency by comparing the actual values for a data point with the estimated value calculated using a variogram calculated using all the other data points. This process, known as cross validation, is repeated omitting each point in turn. If the mean difference between the predicted values and actual values is close to zero with a small variance then the variogram can be assumed to unbiased.

The above method, known as ordinary kriging, is the most common form of kriging. However, there are other variants. Block kriging is used to interpolate data values to areas larger than the support, simple kriging can be used under an assumption of second order stationarity and a known mean, but as these conditions rarely apply it is seldom used on its own; non-linear kriging can be used if the data are non-normally distributed (e.g. log-normal data), indicator kriging can be used to interpolate binary data, stratified kriging is used to interpolate data stratified by different cover classes, co-kriging can be used to make use of data different from, but correlated with, the variable to be estimated, whilst universal kriging is used to incorporate information about trends. Finally conditional simulation techniques (or stochastic imaging) may be used to identify the most likely data value for each place (but at the expense of continuity between adjacent areas). Burrough and McDonnell provide further details on most of these methods.

REFERENCES
Burrough and McDonnell, Chapters 5 and 6.