Spatial Interpolation

Helena Mitasova, NCSU, lecture notes for MEA592 Geospatial Analysis and modeling

The problem is formulated as follows:

Given the m values of a studied phenomenon $z_j,\ j = 1,\ldots,m$ measured at discrete points $\mathbf{r}_j = (x_j, y_j),\ j = 1,\ldots,m$ within a certain region of 2-dimensional space, find a function $F(\mathbf{r})$ which fulfils the following condition:

$$F(\mathbf{r}_j) = z_j, \quad j = 1,\ldots,m \tag{1}$$

The problem does not have a unique solution so additional conditions are used.

However, most interpolation methods can be expressed as a sum of two components

$$F(\mathbf{r}) = T(\mathbf{r}) + \sum_{j=1}^{m} \lambda_j R(\mathbf{r}, \mathbf{r}_j) \tag{2}$$

where $T(\mathbf{r})$ is a 'trend' function, λ_j are weights (coefficients) and $R(\mathbf{r}, \mathbf{r}_j)$ is a function of distance between an unsampled and a measured point. The explicit form of $R(.)$ depends on the choice of the additional conditions (locality, ad hoc, geostatistical -based on variogram, physics - minimization of energy, etc.)

Voronoi polygons

First, Voronoi polygons are created from the measured data (each measured point will be at the center of a Voronoi polygon) . Then, z-value at a point $\mathbf{r}$ will be the same as the $z(\mathbf{r_j})$-value at the center of the Voronoi polygon V_j within which the point $\mathbf{r}$ is located. That means for our general equation, that $T(\mathbf{r}) = 0$ and $\lambda = 1, \mathbf{r} \in V_j$ and $\lambda = 0, \mathbf{r} \notin V_j$ leading to

$$z(\mathbf{r}) = \lambda z(\mathbf{r_j}) \tag{3}$$

The resulting surface is not continuous and it includes only the measured values.

TIN-based linear interpolation

First, TIN is created from the measured data using Delaunay triangulation (each measured point will be a triangle vertex). Then, z-value at an unsampled point $\mathbf{r}$ (we denote it as point K within a triangle ABC) will be computed as a function of the z-values at the vertices of the triangle within which the point $\mathbf{r}$ is located. The function is a weighted average, where the weights are function of areas of triangles ABK, ACK, BCK. That means for our general equation, that $T(\mathbf{r}) = 0$ and $\lambda_k = p_k$ where $p_k, k = 1,2,3$ are areas of sub-triangles ABK, ACK, BCK leading to

$$z(\mathbf{r}) = \frac{p_{BCK} z(\mathbf{r_A}) + p_{ACK} z(\mathbf{r_B}) + p_{ABK} z(\mathbf{r_C})}{p_{ABC}}$$

or

$$z(\mathbf{r}) = \frac{\sum_{k=1}^{3} p_k z(\mathbf{r}_k)}{p_{ABC}(4)}$$

It can be also interpreted as point on a plane defined by the vertices of the triangle ABC. The resulting surface is continuous, but its first derivatives are not (C^0).

There are many non-linear interpolation methods for TINs that lead to C^1 or C^2 continuous surface.

Inverse distance weighted interpolation (IDW)

The method is based on an assumption that the value at an unsampled point can be approximated as a weighted average of values at measured points within a certain cut-off distance, or from a given number m of the closest points (typically 10 to 30). Weights are usually inversely proportional to a power of distance leading to an estimator:

$$z(\mathbf{r}) = \sum_{i=1}^{m} w_i z(\mathbf{r}_i) = \frac{\sum_{i=1}^{m} z(\mathbf{r}_i)/|\mathbf{r} - \mathbf{r}_i|^p}{\sum_{j=1}^{m} 1/|\mathbf{r} - \mathbf{r}_j|^p} \tag{5}$$

where p is a parameter (typically $p = 2$, lower p gives smoother surface - similar to lower tension), and $|\mathbf{r} - \mathbf{r}_i|^2 = (x - x_i)^2 + (y - y_i)^2$ is the distance between the unsampled location $\mathbf{r}$ and a given point $\mathbf{r}_i$.

Smoothing can be introduced by adding a parameter β to the weight term $|\mathbf{r} - \mathbf{r}_j|^2 + \beta$, leading to approximation function.

Note: GRASS modules use $p = 2$ and $m = 12$ as default values.

Geostatistical approach: Kriging

Kriging is based on a concept of random functions: the surface or volume is assumed to be one realization of a random function with a certain spatial covariance. Using the given data $z(\mathbf{r}_i)$ and an assumption of stationarity one can estimate a semivariogram $\gamma(\mathbf{h})$ (Getis 1997 P3.2), defined as

$$\gamma(\mathbf{h}) = \frac{1}{2} Var\left[\{z(\mathbf{r} + \mathbf{h}) - z(\mathbf{r})\}\right] \approx \frac{1}{2N_h} \sum_{(ij)}^{N_h} [z(\mathbf{r}_i) - z(\mathbf{r}_j)]^2 \tag{6}$$

which is related to the spatial covariance $C(\mathbf{h})$ as

$$\gamma(\mathbf{h}) = C(0) - C(\mathbf{h}) \tag{7}$$

where $C(0)$ is the semivariogram value at infinity (sill). The summation in Eq. (6) runs over the number N_h of pairs of points which are separated by the vector $\mathbf{h}$ within a small tolerance $\Delta\mathbf{h}$ (size of a histogram bin). For isotropic data, the semivariogram can be simplified into a radial function dependent on

$|\mathbf{h}|$. The kriging literature provides a choice of functions which can be used as theoretical semivariograms (spherical, exponential, Gaussian, Bessel, etc.) The parameters of these functions are then optimized for the best fit of the experimental semivariogram.

The interpolated surface is then constructed using statistical conditions of unbiasedness and minimum variance. In its dual form (Matheron 1971, Hutchinson and Gessler 1993) the universal kriging interpolation function can be written as

$$F(\mathbf{r}) = T(\mathbf{r}) + \sum_{j=1}^{N} \lambda_j C(\mathbf{r} - \mathbf{r}_j) \tag{5}$$

where $T(\mathbf{r})$ represents its non-random component (drift) expressed as a linear combination of low-order monomials. The monomial and $\{\lambda_j\}$ coefficients are found by solving a system of linear equations (Hutchinson and Gessler 1993).

Radial basis functions and splines

Variational approach to interpolation and approximation is based on the assumption that the interpolation function should pass through (or closely to) the data points and, at the same time, should be as smooth as possible. These two requirements are combined into a single condition of minimizing the sum of the deviations from the measured points and the smoothness seminorm of the spline function:

$$\sum_{j=1}^{N} |z_j - F(\mathbf{r}_j)|^2 w_j + w_0 I(F) = \mathit{minimum} \tag{6}$$

where w_j, w_0 are positive weights and $I(F)$ denotes the smoothness seminorm (Table 1). The solution of Eq. (6) can be expressed as a sum of two components (Talmi and Gilat 1977, Wahba 1990)

$$F(\mathbf{r}) = T(\mathbf{r}) + \sum_{j=1}^{N} \lambda_j R(\mathbf{r}, \mathbf{r}_j) \tag{7}$$

where $T(\mathbf{r})$ is a 'trend' function and $R(\mathbf{r}, \mathbf{r}_j)$ is a basis function which has a form dependent on the choice of $I(F)$. Bivariate smoothness seminorm with squares of second derivatives (Table 1) leads to a **thin plate spline (TPS)** function, when first derivatives are added we get **thin plate spline with tension**, when all derivatives are used with decreasing weight we get **completely regularized spline** or **regularized spline with tension**.

Thin plate spline with tension is then

$$F(\mathbf{r}) = a_0 + \sum_{j=1}^{N} \lambda_j \left[\mathrm{K}_0(\varphi r/2) + \ln(\varphi r/2) + C \right]$$

where $K_0(.)$ is the Bessel? function and φ is tension parameter, C is constant.

Regularized spline with tension is given as

$$F(\mathbf{r}) = a_0 - \sum_{j=1}^{N} \lambda_j \left[\mathrm{E}_1(\varrho) + \ln \varrho + C_E\right]$$

where $C_E = 0.577215...$ is the Euler constant, $\mathrm{E}_1(\varrho)$ is the exponential integral function, while the trend function is a constant and $r = |\mathbf{r} - \mathbf{r}_j|$, $\varrho = (\varphi r/2)^2$ and φ is a tension parameter.

The coefficients $a_1, \{\lambda_j\}$ are obtained by solving the following system of linear equations:

$$a_1 + \sum_{j=1}^{N} \lambda_j \left[R(\mathbf{r}^{[i]}, \mathbf{r}^{[j]}) + \delta_{ji} w_0/w_j\right] = z^{[i]}\,, \quad i = 1, \ldots, N$$

$$\sum_{j=1}^{N} \lambda_j = 0\,.$$

where w_0/w_j are positive smoothing weights.

While not obtained by variational approach, similar in formulation and performance are **multiquadrics** (Hardy 1990, Fogel 1996, Kansa and Carlson 1992, Franke 1982a, Foley 1987, Nielson 1993) with $R_d(r) = (r^2 + b)^{\frac{1}{2}}$ or $R_d(r) = (r^2 + b)^{-\frac{1}{2}}$, offering high accuracy, differentiability, d-dimensional formulation, and, with segmentation, applicability to large data sets. Originally ad hoc proposed multiquadrics were later put on a more solid theoretical ground. Good performance of multiquadrics, especially in 3D, is not surprising, considering that for $d = 3$ in the limit of $b \to 0$ the basis functions $(r^2 + b)^{\frac{1}{2}}$ and $(r^2 + b)^{-\frac{1}{2}}$ are solutions of biharmonic and harmonic equations, respectively (Hardy 1990).