ABSTRACT. Global change research has placed new demands on methods of spatial analysis. In particular, spherical methods for spatial interpolation are required when spatial analyses are performed over large areas of the Earth’s surface. In this article, spherical spatial interpolation procedures are reviewed, compared, and evaluated. Three classes of spherical interpolants are evaluated in detail: distance weighting, functional minimization, and tessellation. The strengths and weaknesses of a method from each of these classes—inverse-distance weighting, thin-plate splines, and surfaces fit to triangulated patches—are evaluated using a hypothetical mathematical surface and a global scale representation of topography. For smooth functions, such as the hypothetical mathematical surface, thin-plate splines produce a visually pleasing surface and have low interpolation error. For non-smooth surfaces, such as global topography, inverse-distance weighting, interpolating thin-plate splines, and triangulated \( C^0 \) patches appear to handle rapid surface changes well. When choosing a spherical interpolant, the properties of the data being analyzed (e.g., smoothness, spatial coherence, etc.) must be taken into account. In addition, multivariate interpolation should be considered when related, ancillary data are available at higher spatial resolution than the original data.

Introduction

Increasing awareness of global-scale processes and phenomena—particularly global environmental change—has made spherical methods an increasingly important component of spatial analysis. Spherical methods for spatial interpolation of point data, in particular, are widely used in global change research for both mapping and analysis. For example, air temperature anomalies measured at irregularly distributed points can be interpolated to a latitude/longitude grid for visualization and then averaged to obtain estimates of “global warming” (e.g., Jones et al. 1986; Jones and Briffa 1992).

Spatial interpolation is a common procedure and there are a variety of methods available (Bennett et al. 1984; Burrough 1986; Franke 1982; Lam 1983; Schumaker 1976; Thiébaux and Pedder 1987; Watson 1992; Weber and Englund 1992), but a large majority of interpolants treat the spatial dimension as planar. When spatial data are located on the surface of the Earth, interpolating on the surface of a sphere—or an ellipsoid or geoid—is a more geometrically consistent and accurate approach. When interpolating over large areas of the Earth, planar interpolation methods—i.e., interpolation within a cartographic projection—can produce large interpolation errors. In comparing a spherical and planar interpolant, Willmott et al. (1985) found that planar interpolation methods can produce interpolation errors as large as 10°C in air temperature fields. Since spherical and planar distances (and angles) diverge with greater distance between control points, data-sparse regions produce larger interpolation errors than data-dense regions. In addition, not only does planar interpolation of global- or continental-scale data produce interpolation errors, but also each cartographic projection produces its own characteristic distribution of interpolation error.

The importance of using spherical interpolants can be illustrated by two classic cartographic errors:

1. isolines that exit a global map at one latitudinal position and re-enter at another position (e.g., different values at 180°E and 180°W), and
2. multiple isolines through the poles (Figure 1).

When calculating distances within planar space, locations along opposite sides of a global map are very distant; therefore, values along identical meridians may not be equivalent unless constrained to be so (this constraint is often in place when interpolating manually). Similarly, the North and South Poles are not represented as points within many projections. As a result, multiple values can be assigned to a unique geographic
location. Both errors are particularly common when global-scale rectangular projections are used and both have extremely important implications for the interpretation and analysis of global-scale data and processes (Figure 1 on page 6).

**Background**

In a general mathematical framework, spatial interpolation of point data can be viewed as an estimation problem that typically uses linear combinations of known values at known locations, or control points, to estimate values at unsampled locations. Often, the control points are irregularly distributed, although another common procedure is to interpolate from one gridded field to another. For spherical interpolation, values of a variable $z$, sampled at latitudes $\phi$ and longitudes $\lambda$, are combined to produce estimates of $z(\hat{z})$ at unsampled locations $(\hat{\phi},\hat{\lambda})$. Schematically, the process is:

$$
\hat{z}_j = f(z_1, z_2, ..., z_i, ..., z_n), \quad i \neq j
$$

where $z_i = (\phi_i, \lambda_i, z_i)$, $f$ represents a spherical interpolation method, and $n_j$ is the number of control points used to obtain $\hat{z}_j$.

The spatial surface estimated from the values at the control points may provide an exact fit (i.e., passing through the values at the control points), or an approximation (or smoothing) of the values. Whether a fit is exact or approximate often depends on both the intended use of the estimated field and the accuracy of the original data. If the values at control points are known to contain errors, an exact fit may not be desirable. Whether it is better to apply smoothing procedures within an interpolation algorithm or to correct the individual values known to be in error partly depends on how much is known about the errors. When the data are reliable and the observed spatial variability is of interest, true interpolation—passing through the values at the control points—is usually used. Another consideration is whether the interpolation method should use a subset of the data at a time (i.e., a local method), or all of the data at once (i.e., a global method). Thus, what is commonly called "spatial interpolation" is actually a group of distinct mathematical approaches that fall within four categories: global interpolation, local interpolation, global approximation, and local approximation (Schumaker 1976).

Often, it is desirable to use ancillary variables that are correlated with the variable of interest to aid the interpolation—i.e., a multivariate interpolation. An example would be to use elevation data to improve the interpolation of precipitation or air temperature data (Daly et al. 1994; Hutchinson and Bischof 1983; Willmott and Matsuura 1995). Sometimes, a higher-resolution realization of the same variable (e.g., a long-term mean) is available to aid the interpolation of data observed at a lower resolution (Willmott and Robeson 1995). Since multivariate interpolation procedures usually are ad hoc and require considerable insight into the processes influencing a particular variable, all the interpolation methods discussed in this article are univariate. It should be kept in mind, though, that nearly all univariate methods can be improved by incorporating information from other variables. All of the methods discussed here also apply only to point data, although methods for interpolation of areal data are available (e.g., Goodchild and Lam 1980; Tobler 1979).

**Spherical Methods for Spatial Interpolation**

A handful of methods is available for interpolating data in a spherical domain. Several widely used planar methods (e.g., Akima 1978, as implemented by IMSL 1991) have not, however, been adapted to incorporate spherical geometry. The most commonly used methods for spherical spatial interpolation are reviewed below.

**Inverse-distance Weighting**

Inverse-distance weighting schemes (Figure 2a on page 7) are perhaps the most commonly used interpolation method (e.g., Jones et al. 1986; Bussières and Hogg 1989; Legates and Willmott 1990). A local approach or global approach may be used, depending on how the distance-weighting function is constructed. For a local procedure, a value at point $f(\hat{z}_j)$ is obtained as a weighted sum of values at $n_j$ nearby control points ($z_i$):

$$
\hat{z}_j = \sum_{i=1}^{n_j} w_{ij} z_i / \sum_{i=1}^{n_j} w_{ij}
$$

The number of nearby control points to be used ($n_j$) may be fixed (for example, $n_j = 7$ is sometimes used), or determined by a search radius. Weighting functions ($w_{ij}$) are inverse functions of distance, usually taking the form

$$
w_{ij} = d_{ij}^{-\gamma}
$$

4 Cartography and Geographic Information Systems
where

d_{ij} \text{ is the distance between control point } i \text{ and estimation point } j.

A value of one or two is common for the exponent \( \gamma \), but an optimal value for \( \gamma \) can (and should) be estimated from the data (Legates 1987). By calculating \( d_{ij} \) as a great-circle distance:

\[
d_{ij} = R \cos^{-1} \left[ \sin \phi_i \sin \phi_j + \cos \phi_i \cos \phi_j \cos (\lambda_i - \lambda_j) \right]
\]

(4)

where \( R \) is the radius of the Earth. Equation 2 becomes a spherical interpolation method.

Interpolation schemes using Equations 2 and 3 are non-optimal in the sense that the weighting function does not depend on the spatial covariance of the data. Further, without appropriate modifications, Equation 3 does not eliminate directional bias (e.g., Figure 2b), nor are spatial gradients incorporated (i.e., directional derivatives are zero at control points and \( \min z_i \leq z_j \leq \max z_i \).

Shepard (1968) developed a local interpolant that improved on simple inverse-distance weighting. In Shepard's method, weighting is a three-step process that accounts not only for distance, but for angular distribution of control points as well as for spatial gradients within the data. The angular correction is made for directional isolation of each control point relative to all other control points (e.g., Figure 2b). Spatial gradients allow extrapolation beyond the range of data values (e.g., Figure 2c).

Adapting Shepard's algorithm for interpolation on a spherical surface, Willmott et al. (1985) modified the distance, angle, and gradient calculations to account for spherical geometry. The algorithm of Willmott et al. (1985) has been used extensively to interpolate continental, terrestrial, and global climate fields (e.g., Huffman et al. 1995; Legates and Willmott 1990; Robeson 1995; Willmott et al. 1994). Although the distance-weighting function is not derived from the data, the inclusion of angular relationships and spatial gradients represents an improvement on simply weighting control points according to their great circle distances.

Optimal Statistical Objective Analysis

A method for optimizing Equation 3 entails estimating spatial covariance in order to determine the appropriate covariance parameters of the distance-weighting function as well as an optimal search radius (to determine \( n_i \) rationally). Such a function may be both anisotropic (Thiébaux 1976) and heterogeneous—i.e., the weighting function may vary with both direction from a estimation point and with location over the domain. An optimal modified version of distance-weighting interpolation considered is that of Gandin (1963): optimal statistical objective analysis (OSOA). As described by Thiébaux and Pedder (1987), OSOA is a local interpolation procedure that estimates values as a linear combination of deviations \( z_i' \) from a "first guess" field \( z_f \) plus the first guess at the estimation point \( z_j' \):

\[
\hat{z}_j = \sum_{i=1}^{n_i} w_{ij} z_i' + z_f' = z_j' + z_f.
\]

(5)

OSOA is most often used in the atmospheric and oceanic sciences where "first guess" fields are usually a long-term mean or a forecast (Thiébaux and Pedder 1987), although, in the absence of other information, \( z_f' = 0 \) is assumed. The important difference between OSOA and inverse-distance weighting, however, is that the weights \( w_{ij} \) are derived from the data. Again, for spherical interpolation, the distances are computed along great circles.

Since both the \( z_i' \) and \( z_f' \) are known, the only unknowns on the right side of Equation 5 are the weights \( w_{ij} \). Weights are determined by minimizing the squared deviations of the estimated value from the true value (ideally, in the ensemble mean). The weights for estimation point \( j \) are derived from the spatial covariance of surrounding control points with

1. the estimation point, and
2. every other control point.

In matrix form, the weights are

\[
\begin{pmatrix}
w_{1j} \\
\vdots \\
w_{nj}
\end{pmatrix} = \begin{pmatrix}
\sigma_{11} & \cdots & \sigma_{1n_j} \\
\vdots & \ddots & \vdots \\
\sigma_{nj} & \cdots & \sigma_{n_jn_j}
\end{pmatrix}^{-1} \begin{pmatrix}
\sigma_{1j} \\
\vdots \\
\sigma_{nj}
\end{pmatrix}
\]

or

\[
w = \Sigma^{-1} \sigma_j
\]

(6)

(7)

where the variance-covariance matrix \( \Sigma \) contains the spatial covariance of each control point with every other control point. The column vector \( \sigma_j \) contains the spatial covariances of the estimation point with the control points. In practice, \( \Sigma \) will always be symmetric positive definite (the degenerate case is the trivial one of perfect correlation over an entire region); therefore, Cholesky decomposition of \( \Sigma \) into \( LL^T \) (where \( L \) is lower triangular) yields an efficient matrix inversion.
Figure 1. Maps of annually averaged air temperature anomalies (°C) during 1978 generated by (a) a planar interpolation procedure that utilizes bicubic splines (Sandwell, 1987) and (b) a spherical procedure that utilizes thin-plate splines (Wahba, 1981). While both have a similar overall pattern (although the planar method appears to overestimate spatial gradients), the planar procedure has multiple isolines through the poles and isolines that do not adjoin across 180°E and 180°W.
observations vary with increasing separation distance:

where

\( y_h \) is the semi-variogram of variable \( z \) at separation distance \( h \).

\[ y_h = \frac{1}{n} \sum (z_i - z_{i+h})^2 / 2n \]  \( (8) \)

\( y_h \) is the semi-variogram of variable \( z \) at separation distance \( h \).

For stationary data, the semi-variogram and the spatial autocovariance function are mirror images (Davis 1986). When applying Equation 8 to irregularly-spaced data, stations that are separated by similar distances (e.g., 0-5 km, 5-10 km, etc.) are grouped together to calculate the variogram (since only regularly-spaced data have a number of observations that are exactly \( h \) apart). A variety of functional forms can be fit to the variogram; linear, exponential, and spherical models are the most common (Clark 1979). Choice of variogram model can be quite important and should be dependent on the variable being modeled (i.e., similar to the choice of exponent in inverse-distance weighting; Weber and Englund 1994).

Interpolation weights for kriging are determined by a procedure similar to that of OSOA, except that a Lagrangian multiplier (\( \lambda \)) is introduced to ensure that the weights sum to one:
This matrix equation must be solved at every point where an interpolated value is desired. One particularly advantageous feature of kriging is that estimates of the error variance of the interpolated values can be obtained (since this is precisely what the semi-variogram measures). The error variance is equal to twice the semivariance at a given separation distance (Davis 1986).

Since both kriging and OSOA model spatial fields by fitting a low-order function to some measure of spatial covariance, they are conceptually extremely similar. In practice, however, there are important differences in the way in which the two methods are implemented. When using kriging, geostatisticians typically have one realization of a spatial field (e.g., a spatial sample of ore grades) and, therefore, the variogram must be estimated over the entire field, assuming homogeneity. In the geostatistical literature, this is known as the “intrinsic hypothesis”—it is equivalent to the assumption of stationarity of spatial mean, variance, and covariance. Methods such as “universal kriging” attempt to overcome the limitations of assuming stationarity by modeling the “drift” of the field (Journel and Huijbregts 1978). Typically, this is done by fitting (and removing) a low-order polynomial in the neighborhood of each control point.

OSOA is more commonly used in meteorology and oceanography where a series of realizations of the spatial fields are available through the added dimension of time. Temporal stationarity is assumed in order to calculate correlations between spatial locations, allowing a spatial covariance function to be specified at each location, if necessary. Removing the constraint of the intrinsic hypothesis can make OSOA more adaptive, but careful attention to problems associated with temporal nonstationarities is needed (i.e., spatial autocorrelation functions may be biased; Gunst 1995).

**Spline Methods**

In their traditional cartographic and mathematical representation, splines are a local interpolation procedure. Considering a one-dimensional spline, for a set of \( n \) nodes \( x_0, x_1, \ldots, x_n \) with associated values \( z_{x_0}, z_{x_1}, \ldots, z_{x_n} \), a spline function \( s(x) \) of degree \( m \) is defined in each subinterval \( (x_{i-1}, x_i) \) to be a polynomial of degree \( m \) (or less), having its first \( m-1 \) derivatives piecewise continuous (De Boor 1978). Hence, \( s(x) \) is a smooth function that passes through all control points. Extending splines to two-dimensional data is not straightforward because they are not simple cross-products of one-dimensional splines (Lam 1983). Since the two-dimensional control points may be irregularly-spaced, the surface is often subdivided into “patches” where the splines are to be fitted. There are a number of different criteria for determining the subdivision, and each may produce a different interpolating surface (see below). Further, “sewing” the patches together in order to create an entire surface must be performed in such a way that spurious variations are not incorporated at or near the patch boundaries (Burrough 1986).

Splines may also be used as a basis for global minimization of an objective function. A particular class of splines commonly used as basis functions for linear interpolation and approximation is “thin-plate splines” (TPS). The name is derived from the functions that solve a differential equation describing the bending energy of an infinite, thin-plate constrained by point loads at the control points (Wahba 1990). One implementation of thin-plate splines, a generalization of one-dimensional smoothing polynomial splines (Reinsch 1967), allows the resulting surface to smooth or approximate the data (Wahba 1981). As an approximation method, TPS is particularly advantageous when data are known to contain errors or when small-scale features need to be eliminated (e.g., low-pass filtering of topographic variability). Methods for TPS interpolation and approximation are well-developed (Wahba and Wendelberger 1980; Wahba 1981; Dubrule 1983, 1984; Wahba 1990) and a version of the spherical procedure of Wahba (1981) has been implemented (Burrough 1991).

As described by Wahba (1981), the objective of TPS approximation on the sphere is to find the global function \( u \) and smoothing parameter \( \lambda \) that minimize:

\[
\|u(\phi, \lambda) - z_i\|^2 + \lambda J_u(u)
\]  

where \( \|u(\phi, \lambda) - z_i\| \) is often the \( L_2 \)-norm (or least-squares minimization) of the difference between \( u \) and \( z_i \), and \( J_u(u) \) is a penalty function that measures the smoothness of the spherical function \( u \) (Wahba 1981). The subscript \( m \) refers to the order of partial derivatives in the penalty function (Courant and Hilbert 1953).
Spherical Harmonics

When using spherical harmonics to represent a functional surface, they can be expanded with many kinds of spatial data. For instance, the method of McMahon and Franke (1992) assumes that:

$$\|u(\phi, \lambda) - z_i\|$$

Constraining $\theta$ to equal zero gives thin-plate spline interpolation, i.e., an exact fit through the control data. Typically, a linear combination of thin-plate spline basis functions approximates $u$ (Wahba 1990), while generalized cross validation (GCV) determines $\theta$ and $m$ (although $m=2$ is commonly used). In contrast to ordinary cross validation, GCV implicitly removes one control point at a time and estimates the error in predicting the removed point (Golub et al. 1979).

Computational requirements for TPS are demanding. Very large design matrices (incorporating the evaluation of $u$, the penalty function, and the GCV function) are subjected to a series of decomposition methods (e.g., Cholesky, QR, and singular value decompositions). As a result, both large computer memory and fast processing capability are needed. Two alternatives are available to reduce computing requirements: a partial spline model and partitioning of the dataset.

A partial spline model uses a subset of the control points when forming the basis functions (Bates et al. 1986). All control values, however, enter into the design matrix, giving an overdetermined system of equations. By using a smaller set of basis functions, computational requirements are reduced while an attempt is made to maintain the representation of the surface that would result from using all the control points. There are a variety of methods for choosing which control points (sometimes called knots in the spline literature) should be used to determine the basis functions (e.g., Schiro and Williams 1984; Bozzini et al. 1986; LeMéhauté and Lafranche 1989; McMahon and Franke 1992; Lyche 1993). Some of these methods may not be appropriate with many kinds of spatial data. For instance, the method of McMahon and Franke (1992) assumes that:

1. a high density of control points means that the surface is changing rapidly and a low control point density means that the surface is changing slowly, and
2. each point is equally important in defining the surface.

While these assumptions may be appropriate in a geophysical modeling context (e.g., in finite-element modeling), they certainly are not appropriate for many types of observed data, especially "samples of convenience" (Freedman et al. 1978). Networks of climatic data, for instance, often have control point distributions that are functions of human population distributions and levels of industrialization rather than the spatial variability of any climatic variable. In many contexts, therefore, a stratified random sample of control points would be appropriate in a partial spline model.

An alternative to the partial spline model is to partition the control points into subsets with splines fit over each (overlapping) partition. The partitions subsequently are woven together to form the entire interpolating surface (Mitásova and Mitás 1993). While all of the control points are used in this approach, substantial overlap between partitions may be needed in order to ensure that the entire surface can be pieced together seamlessly. Also, partitioning the data does reduce computer memory requirements, but may take more processing time due to the overlap and additional post-processing.

Spherical Harmonics

Spherical harmonics have been used to model geophysical fields for over a century (Byerly 1893). They arise whenever a wave equation is solved by separation of variables in spherical coordinates. As a result, spherical harmonics are most widely used in the study of electromagnetism (Barralough 1978; Jones 1985) and geophysical fluid dynamics (Ellsaesser 1966; Swartztrauber 1979). Similar to Fourier series, spherical harmonics use trigonometric functions in the longitudinal direction; however, associated Legendre polynomials are used as basis functions in the latitudinal direction:

$$Y_n^m(\phi, \lambda) = P_n^m(\phi)[a_{n,m} \cos m \lambda + b_{n,m} \sin m \lambda]$$

where $Y_n^m(\phi, \lambda)$ is the spherical harmonic of order $m$ and degree $n$ ($m$ and $n$ determine the number of basis functions in the longitudinal and latitudinal directions)

$$P_n^m(\phi)$$

is a normalized associated Legendre polynomial, and

$a_{n,m}$ and $b_{n,m}$ are the Fourier-Legendre (spherical harmonic) coefficients.

When using spherical harmonics to represent a functional surface, they can be expanded as:

Vol. 24, No. 1
A complete discussion of methods for specifying $P_n^m(\phi)$ and estimating $a_{n,m}$ and $b_{n,m}$ can be found in Swarztrauber (1979).

Spherical harmonic representations of data may produce either an exact fit or an approximation of the data. Approximations are obtained by truncating the series to form a discrete basis for the spherical surface. Within numerical models of atmospheric circulation, two truncations are commonly used: triangular and rhomboidal, referring to the shape of the region of truncation in the $n,m$ plane (Washington and Parkinson 1986). For spatial interpolation, the truncation should be data-dependent, although computational considerations will constrain the order and degree of truncation.
As with most surface-fitting procedures, irregularly spaced data can pose problems for estimation of spherical harmonic coefficients. When control points are clustered, the matrices used to solve for the Fourier-Legendre coefficients can become ill-conditioned. Similar to the thin-plate spline methods discussed above, singular value decomposition can be used to reduce the effects of ill-conditioning—i.e., to produce a pseudo- or generalized inverse of the design matrix (Nashed 1976; Golub and Van Loan 1989).

Figure 4. Maps of a mathematical surface using a variety of spherical spatial interpolation methods: (a, top) the original surface with the 50 sample points shown, (b, above) inverse-distance weighting, (c, top, page 12) thin-plate spline approximation, (d, bottom, page 12) tessellation using $C^0$ surface patches, and (e, top, page 13) tessellation using $C^1$ surface patches.
Although they are widely used as solutions and approximations to differential equations in geophysics, spherical harmonics have not been widely used for spatial interpolation of spherical data. Spherical harmonics, nonetheless, can form the basis for spatial interpolation of a wide variety of spherical data and should provide a smooth representation of data fields when control points are well-distributed. For instance, a spherical harmonic series with 361 terms has been used to represent world population distribution (Tobler 1992).

Tesselation Methods

In order to interpolate between control points, the spatial domain may be decomposed into a number of polygons or surface "patches" (the set of polygons is known as a tesselation). An
interpolating surface can then be fit over each polygon, using the values at vertices as local control points (e.g., Figure 3). Tessellation-based methods have the advantage of automatically retaining the original resolution of the control points. In areas where the sampling is dense, there will be more polygons; in areas where sampling is sparse, there will be relatively fewer polygons. Although grid-based interpolation schemes usually use regular grids, polygon density provides a means for choosing data-dependent grid resolution (Tobler 1988).

One commonly used method for dividing an interpolation domain into patches involves a recursive subdivision of the convex hull of the data into triangles that are as nearly equiangular as possible (Figure 3). An equiangular (Delaunay) triangulation is the first step in creating Voronoi regions—also known as Thiessen or Dirichlet polygons—that can be used for a variety of spatial analyses (e.g., Thiessen 1911; Laurini and Thompson 1992; Okabe et al. 1992). Delaunay triangulations frequently are used for interpolation, but other triangulations may be preferable in certain situations (Quak and Schumaker 1990; Rippa 1992). For instance, there is evidence that long and thin triangles may produce a more accurate interpolation, depending on the magnitude of second directional derivatives (Rippa 1992). For global- or continental-scale geographic data, nonetheless, it is important that the triangulation be done on the surface of a sphere. If the triangulation is not performed spherically, each cartographic projection can produce different patches and therefore different interpolations as well (Figure 3).

Methods that utilize spherical triangulation have been developed for both C₀ and C¹ interpolation (Lawson 1984; Nielson and Ramaraj 1987; Renka 1984a). A C¹ interpolant produces a surface that is continuous through first derivatives, while the C₀ method simply uses control points at the vertices of a spherical triangle to determine a planar surface over that triangle. Renka’s C¹ method utilizes either a local or a global gradient estimation to fit a cubic Hermite surface over the spherical triangle. For extrapolation outside the convex hull of the control points, interpolated values are designated as a linear extension of the nearest patch within the convex hull.

A widely-used and implemented (IMSL 1989) triangulation-based algorithm is that of Akima (1978). Comparison of the planar methods of Akima (1978) and Renka (1984a) suggests that Akima’s gradient estimation method may be more robust (within the convex hull). In its current form, however, Akima’s method only permits planar interpolation. “Natural neighbor interpolation” provides an interesting combination of tessellation and distance-weighting methods (Sibson 1981). Planar C¹ methods for natural neighbor interpolation are both computationally efficient and accurate (Philip and Watson 1987; Watson 1992). Modifications to the gradient estimation of both Akima’s method and natural neighbor interpolation would be needed for use with spherical data.
Evaluation of Spherical Interpolants

There are a variety of ways in which to compare spatial interpolants (e.g., Bennett et al. 1984; Bussièrè and Hogg 1989; Franke 1982; Gustavsson 1981; Lam 1983; Robeson 1994; Rhind 1975; S烘焙aker 1976; Watson 1992; Weber and Englund 1992). Visualization of the interpolated surface provides important information on the properties of the
interpolant (e.g., smoothness, spatial gradients, spurious values, etc.). Another useful comparison is to sample a known surface and to use the interpolants to reconstruct the surface. Differences between the surfaces as well as error estimates can be analyzed statistically and graphically. Finally, cross-validation (successively removing sample points and interpolating to the removed location) is a useful way in which to compare and evaluate spatial interpolants (Isaaks and Srivastava 1989; Robeson 1994; Weber and Englund 1992). When data are subject to error, however, cross-validation should be used with caution since the partitioning of error (between interpolation error and other types of error) may vary from one interpolation procedure to the next (Burt 1985; Davis 1987).

To illustrate the properties of the spherical interpolants discussed above, several methods are
compared visually and statistically, here. For brevity and generalization, the methods are grouped into three somewhat distinct categories: 1. distance-weighting, 2. functional minimization, and 3. tesselation. The distance-weighting category includes inverse-distance weighting, optimal statistical objective analysis, and kriging while functional minimization includes both thin-plate splines and spherical harmonics. Tesselation-based interpolation methods include a variety of strategies, including triangulation, proximal regions, and quadtrees. The methods of Willmott et al. (1985), Wahba (1981), and Renka (1984a) are used as representative approaches for the three categories, respectively. Although some of the methods within each category can produce somewhat different spatial representations of the data (e.g., spherical thin-plate splines and spherical harmonics), their properties are sufficiently similar for some generalizations to be made.

### A Mathematical Surface

To show the visual properties of the spherical interpolants for a smooth surface, maps are generated using a hypothetical mathematical function:

\[
\begin{align*}
f(x,y,z) &= \frac{6x^3 - 5x^2e^{yz} + 4xy^2 - 8y^3 + 3z^3 - 5e^x yz}{10} + \frac{1}{2} \left[ \frac{15x}{2} + \frac{15}{2} \cos y \right] \\
\end{align*}
\]

where \(x=\cos \lambda \cos \phi, y=\sin \lambda \cos \phi,\) and \(z=\sin \phi\) are the cartesian coordinates of a spherical surface.

The spherical surface was designed to have hills and valleys with gradients in both latitudinal and longitudinal directions. To examine how well the spherical interpolants handle irregularly spaced data, the surface was sampled at 50 randomly generated locations (Figure 4a). For each of the methods used, values of the mathematical function at the sample points were used to generate grid-point estimates and to produce global maps.

Most inverse-distance weighting methods are local interpolants; therefore, small-scale features typically are emphasized at the expense of maintaining a smooth surface (Figure 4b). Since the original mathematical surface (Figure 4a) is fairly smooth, the representation using inverse-distance weighting looks somewhat uneven, although the general pattern is maintained.

For surfaces that are less smooth (see below), the local nature of inverse-distance weighting may be advantageous. Thin-plate splines, on the other hand, are an inherently smooth interpolator. The major features of the sampled surface are reproduced and the map has a pleasing visual appearance (Figure 4c). Since the mathematical function is smooth, both thin-plate spline interpolation and approximation produce nearly identical parameter estimates and gridded fields (and therefore only one is shown). Triangulated \(C^0\) patches can produce faceted patterns when adjacent spherical triangles have very different (linear) surface slopes, which is inappropriate for an inherently smooth surface (Figure 4d). The \(C^1\) triangular patches (with local gradient estimation) also produce a fairly smooth field and reproduce the major features of the surface (Figure 4e). At times, though, the triangulations can create artifacts since the surface patches are sometimes fit over large areas (e.g., the large dark area near the equator from 100°E eastward is one such patch).

Visual properties of spatial interpolants are important; in many applications, however, statistical properties can be even more critical. For instance, one spatial interpolation procedure may produce visually appealing maps that consistently overestimate the spatial mean or gradients. In order to evaluate and compare spatial interpolation procedures statistically, interpolation errors are estimated for the mathematical function used above. The mathematical function was sampled at 50 randomly generated points a total of ten times. Each of these random samples was interpolated to a spherical grid where the errors were estimated and averaged.

Two error statistics, mean bias error (MBE) and mean absolute error (MAE), are used to evaluate the spatial interpolation methods:

\[
\text{MBE} = \frac{\sum_{j} w_j (f_j - \hat{f}_j)}{\sum_{j} w_j} \\
\text{MAE} = \frac{\sum_{j} |f_j - \hat{f}_j|}{\sum_{j} w_j} \tag{15}
\]
Global-scale Topography

To evaluate the spherical interpolation methods using a nonsmooth surface, global-scale topography is used to account for differential areas associated with the spherical lattice of grid points.

All of the interpolants have low MBE for the mathematical surface (Table I). Thin-plate splines, however, have the smallest MAE while the \( C^4 \) tesselation method has the highest MAE. Since MBE simply is the difference between the (weighted) means of \( \hat{z}_i \) and \( z_p \), it represents any systematic overestimation or underestimation of the spatial mean. Unsystematic differences between \( \hat{z}_i \) and \( z_p \), however, can be averaged out to produce a small MBE even if an interpolation procedure produces a large error field. MAE, on the other hand, accounts for all spatial differences and, therefore, is a better overall indicator of spatial interpolation accuracy. Thin-plate splines, therefore, appear to have both small bias and small overall interpolation error. The \( C^4 \) tesselation method is unbiased but does not represent the mathematical surface as reliably as the other methods. Previous research using global air temperature data (Robeson 1994) has shown that the \( C^4 \) tesselation method can overestimate spatial gradients when control points are not well-distributed. Somewhat surprisingly, the \( C^0 \) tesselation method has low MBE and relatively low MAE. The linear surface patches that the \( C^0 \) method produces, however, are not appropriate for inherently smooth surfaces.

**Global-scale Topography**

To evaluate the spherical interpolation methods using a nonsmooth surface, global-scale topography and bathymetry are examined. Even when represented on a 0.5°-by-0.5° latitude/longitude grid, topography can have very strong gradients and is inherently nonsmooth (Figure 5a). The topographic reference field was sampled at 500 randomly generated locations and each of the interpolation methods was used to reproduce topography on the 0.5°-by-0.5° grid. A total of 30 random samples of global topography (each of size 500) were taken. As with the mathematical function, both visualization of the interpolated fields and error statistics are used to evaluate the interpolants.

Nearly all of the methods reproduce the major features of the continents and ocean basins; high-frequency information, however, such as high mountains and deep ocean trenches, are largely unresolved (Figures 5b-5d), mostly due to the low sampling density.

As a local interpolant, inverse-distance weighting is better-suited to nonsmooth surfaces than smooth ones and does appear to provide a realistic depiction of global-scale topography given the limited number of sample points (Figure 5b).

Smoothing thin-plate splines (with 500 basis functions) remove too much small-scale variability and probably are not appropriate for an uneven surface such as topography (Figure 5c). Interpolating thin-plate splines, however, pass through the data values but still provide a smooth, visually appealing interpolation. Nonetheless, the smoothness constraints within thin-plate splines may be unnecessary for a discontinuous field such as topography. In addition, the visual similarity of the fields generated by approximating and interpolating splines illustrates the importance of statistical analysis (see Table 2). The \( C^0 \) tesselation-based method produces a realistic representation of topography (Figure 5d) and, in some ways, the triangular patches match the variability of topography. The \( C^1 \) tesselation method (with local gradient estimation) appears to have problems when the surface changes rapidly (as in the southwestern Pacific and northern Atlantic oceans).

Statistical analysis of the (30 fields of) interpolated topography shows that inverse-distance weighting, interpolating thin-plate splines, and the \( C^0 \) tesselation method have the lowest interpolation errors (Table 2). When approximating thin-plate splines are used, the interpolation error (MAE) increases drastically, due to the nonsmooth nature of the topographic field. The \( C^1 \) tesselation method has considerably higher MAE than the other methods, demonstrating problems with gradient estimation when a surface changes rapidly. The accuracy of the \( C^0 \) method with topography is not completely unexpected since simple linear interpolation has been found to be better than inverse-distance weighting for local-scale representations of topography (Kumler 1994).

Note that, since only 500 sample points were used to represent a topographic surface that has 259,200 original data values, the interpolation errors generated here are extremely large. The intention here is to illustrate the general properties of the spherical interpolants for a nonsmooth surface.

\[
\text{MAE} = \frac{\sum w_j |\hat{z}_j - z_j|}{\sum w_j}
\]

where \( w_j = \cos \phi_j \) is the weight associated with grid-point \( j \) (located at latitude \( \phi_j \)) and \( \hat{z}_j \) is the interpolated value of \( z \) at \( \phi_j, \lambda_j \).
Discussion

The two examples shown above—a mathematical surface and global-scale topography—are meant to illustrate the properties of the spherical interpolation methods for smooth and nonsmooth surfaces. Although a number of mathematical functions were tested and the results were similar in each case, the choice of a particular type of surface and the number of sample points naturally influence the results. Given that the mathematical function is composed of analytical functions that are relatively smooth, the low interpolation error of thin-plate splines are not unexpected (spherical harmonics should have similar errors). Given the wealth of research that has used thin-plate splines (e.g., Hutchinson and Gessler 1994; Mitásová and Hofierka 1995; Wahba and Wendelberger 1980; Wahba 1981; and others), they are an excellent choice for estimating smooth surfaces. For nonsmooth surfaces, the choice is less clear. Inverse-distance weighting and related methods (such as OSOA and kriging) produce realistic surfaces with low interpolation error (Bussières and Hogg 1989; Philip and Watson 1987; Robeson 1994; Weber and Englund 1992). Both C⁰ surface patches and interpolating thin-plate splines have similar errors to distance-weighting for the topographic surface, but the inherent properties of the surface likely will dictate which method is most appropriate for other types of nonsmooth data. In addition, if the nonsmooth nature of the data is the result of errors (observational, random, etc.), then thin-plate spline approximation or a truncated spherical harmonic series may be more appropriate.

As with all statistical procedures, adding information from other related variables should improve estimation and reduce interpolation error. Multivariate interpolation procedures can be incorporated into the methods discussed above; however, they must be guided by the experience and expertise of scientists in a given area of study. A variety of physical constraints, for instance, is frequently built into interpolation procedures in the atmospheric sciences (Daley 1991), making the analysis both more accurate and physically consistent. In many applications where the strength of the relationship between the variable of interest and a related variable is less clear, careful analysis of interpolation accuracy will determine whether multivariate interpolation is necessary or feasible. Similar arguments regarding accuracy can be made for special-purpose univariate interpolation methods. For instance, interpolants that preserve statistical moments such as spatial mean or variance (e.g., Harzallah 1995) are useful for certain applications, but their accuracy also should be evaluated.

Summary and Conclusions

When applied to large portions of the Earth's surface, planar methods of spatial analysis can result in substantial errors. Spherical methods for spatial interpolation are particularly important since most global data are observed at irregularly spaced locations and must be interpolated in some sense (e.g., air temperature measurements, population estimates, etc.). In this article, spherical spatial interpolation procedures are reviewed, compared, and evaluated. While numerous methods are available for planar interpolation, only a handful of procedures have been modified to account for spherical geometry. Three methods—inverse-distance weighting, thin-plate splines, and triangulation-based methods—do have spherical analogs and are evaluated in detail using a smooth mathematical function and global-scale topography.

For smooth surfaces such as the mathematical function, thin-plate splines produce both a visually pleasing surface and low interpolation errors. For nonsmooth surfaces, such as topography, distance-weighting, C⁰ surface patches, and interpolating thin-plate splines produce low interpolation error, although the thin-plate spline surface is necessarily smooth, whereas topography is not. With appropriate modifications to spatial gradient estimation procedures, the C¹ triangulation-based methods of Renka (1984b) also may produce realistic surfaces with low interpolation error. Other widely-used planar interpolants, such as Akima's method and natural neighbor interpolation, should provide useful alternatives if modified for spherical geometry. In addition, methods not discussed here, such as multiquadric interpolation (Foley 1990; Pottmann and Eck 1990), need to be evaluated in the context of geographic data.
Clearly, the type of surface being interpolated must be considered when choosing a spatial interpolation procedure. In particular, a key factor is whether or not the surface is inherently smooth. Sometimes, however, it is meaningful to smooth an inherently nonsmooth surface. High-frequency information in raw topography, for instance, can cause numerical instabilities in atmospheric circulation models and must be removed.

The interpolation methods discussed here play an important role in spherical spatial analysis, but most univariate methods can be improved if ancillary data are incorporated. Elevation data, for example, can greatly improve the interpolation of both air temperature and precipitation data. The greatest improvements in spatial interpolation usually result from the creative use of related variables that are observed at a higher resolution.

ACKNOWLEDGEMENTS

I thank Cort Willmott (University of Delaware) for introducing me to many of the concepts of spherical spatial interpolation. I also thank James Burt (University of Wisconsin) for helpful discussions on thin-plate splines and estimating interpolation error. Michael Janis (Indiana University) provided useful comments on a draft of this article.

REFERENCES


Byerly, W. E. 1893. An elementary treatise on Fourier’s Series and spherical, cylindrical, and ellipsoidal harmonics, with applications to problems in Mathematical Physics. Boston, Mass.: Ginn.


