13

Terrain Analysis

13.1 THREE-DIMENSIONAL TRANSFORMATIONS

The cartographic transformations discussed so far have avoided instances where three-dimensional data are involved. In analytical and computer cartography, much digital data is three-dimensional, usually relating to the surface of the land or the bottom of bodies of water. These data are three-dimensional because we need eastings and northings *and* the elevation at a location to give the necessary level of information. The part of analytical cartography concerned with the analysis of terrain-type data, including any continuous surface, is called terrain analysis. For this chapter, we have broadened terrain analysis to include the symbolization of terrain using computer cartography. Both analytical and computer cartography must be concerned with the data structures and the special-purpose data structure conversions involved in terrain analysis.

The first of these data structure conversions is by far the most critical and can fully determine the level of accuracy and the error involved in terrain analysis and mapping. This transformation is that of point data to both the TIN and the grid three-dimensional formats. In only a few instances are terrain data collected at locations that are evenly spaced. Usually, elevations are measured at significant points on the landscape, such as the crests of ridges, the tops of hills and mountains, and the bottoms of depressions and lakes.

When data are collected on a grid, such as the elevations taken from stereophoto interpretation using an analytical stereoplotter, there is often a need to treat the data as point data for transformation into a particular map projection or spacing and then to reinterpolate to a grid. The point-to-TIN and point-to-grid conversion is of obvious importance. In a TIN, the irregularly spaced points can be used directly as part of the three-dimensional cartographic object definition, eliminating the interpolation problem. For the grid, however, there is no one simple solution to the interpolation problem, and we often are left making trade-offs to achieve terrain maps that suit a specific cartographic purpose.

3.2 INTERPOLATION TO A GRID

A general statement of the interpolation problem would be the following: Given a set of point elevations with coordinates (x, y, z), generate a new set of points at the nodes of a regular grid so that the interpolated surface is a reasonable representation of the surface sampled by the points.Units for x and y are often in meters, but can also be in degrees, feet, and so forth. Units for the z value are usually meters, rounded to the nearest meter, or feet. Ocean depths are often measured in meters, fathoms, feet, and other units.

In the following worked examples, a test set included on the companion disk is used. This is a data set of elevations digitized from a map of the area surrounding the summit of Mount Everest, a total of 5,112 points. Eastings and northings are in meters on an a local grid centered on the summit area, and elevations are in meters. Many applications use rectangular grids, and sometimes the data values are assumed to lie at the intersection points of the grid rather than at the grid squares. In this case, the grid squares are assumed to "contain" the data value.

To build a grid based on data at these points, we use the neighborhood property of the surface. We assume that the elevations are continuously distributed; that is, there are no sudden cliffs, caves, overhangs, or the equivalent, and that values at any point are most closely related to elevations that are in the immediate proximity. Furthermore, we assume that the influence of any point increases as distance to the point decreases, the so-called inverse-distance method. We can choose to make any interpolated surface fit exactly through the data points given, and we can choose whether to allow the highest and lowest points on the interpolated surface to fall beyond the ranges of the data at the points.

Many methods are available for interpolation. Some of them model the entire distribution of the surface and are considered in Section 13.4. The remainder are local operations; that is they work on small areas one at a time to achieve the interpolation. The first set of methods are the simplest and use inverse distance as their neighborhood model. The methods work by moving from grid cell to grid cell, each time computing an interpolated value for that grid cell. In the following discussion, this grid cell will be termed the *kernel*.

13.2.1 Weighting Methods

Weighting methods work by assigning weights to the elevation values found within a given neighborhood of a kernel. The neighborhood is determined in one of several ways discussed in the following section on search methods. In a computer program, the search method computes the distance of each kernel to each point for every kernel. For a 200 by 200 grid, with 1,000 data points, this means 40 million distance calculations, each of which involves taking a square root and squaring two values and then sorting each of the 1,000 distances to determine which is the closest. Hodgson (1989) called this the "brute force" method and suggested an alternative "learned search" approach in which the points are divided into coarser cells called *sorted cells*. The points are sorted within these

areas, and this allows the nearest point information to be retained as the algorithm moves from a kernel to a neighboring kernel, saving considerably on the amount of time required.

Once the points within a neighborhood of a kernel are found, the distances from the kernel to each point are computed and used to weight the elevations at the surrounding points. Mathematically, the formula used is

$$Z_{i,j} = \frac{\sum_{p=1}^{R} Z_p d_p^{-n}}{\sum_{p=1}^{R} d_p^{-n}}$$

where Z_p is the elevation at point p in the point neighborhood R, d is a distance from the kernel to point p, and $Z_{i,j}$ is the elevation at the kernel. The value n is the "friction of distance" that allows very distant points to be penalized with respect to closer points. As n increases, so also does the retention of breaks and extremes in the surface. When n is 1.0 and R is 3, the method is equivalent to linear interpolation over the TIN. Values of n for terrain have varied from 1.0 to 6.0, although many cartographers use a value of 2.0, in which case the technique is called *inverse-squared distance weighting*.

Refinements to the technique involve inserting barriers that the interpolator cannot cross, such as fault lines or coastlines, and using the cosine of the vertical angle between the point and the kernel in the weighting to eliminate the shadowing effect of closer points on the same bearing (Shepard, 1968).

A C language function that reads a file containing data points and creates a grid is included on the companion disk as `inverse_d.c`. The number of nearest neighbors is passed to this function as an argument from the calling program. The neighborhood search, however, is performed unlike those described above. As points are read in, they are assigned their place in the as-yet-empty grid. If two or more points fall into one cell, they are averaged in.

Each empty cell is then subjected to a systematic search. Square neighborhoods around the cell are examined until the required number of points are found. It is possible that more than the required number can be found, because one extra row and column are searched on each side of the kernel in each scan. This algorithm has been found to be highly effective and is far faster than the brute force method. It is implemented in the program `terrapin`, also on the companion disk.

13.2.2 Trend Projection Methods

Trend projection methods are designed to overcome the limitation that occurs because grid maxima and minima can lie only at data points. For a regularly sampled set of data points, it is rare that the grid will exactly coincide with the extreme highs and lows of the actual terrain. In trend projection, sets of points within a region are used. The region is

usually determined using one of the search methods discussed in Section 13.2.3. Fits to the surface are made locally, using a mathematical projection method such as a bicubic spline or a polynomial trend surface. This "best fit" surface is then used to estimate the elevation at the kernel. If simple linear trend estimates are taken, often for triplets of points in the neighborhood, the actual value given to the kernel can be the mean of the various estimates, as in Figure 13.1, or the value can be weighted in the general direction of the surface trend.

In rough terrain, the trend projection methods may generate an interpolated surface that is far more textured than the true surface, but when the data are sparse and the need for texture information is high, this method may be useful. Sampson (1978) described in detail the trend projection method integrated into the SURFACE II package, a software system used extensively in the geosciences.

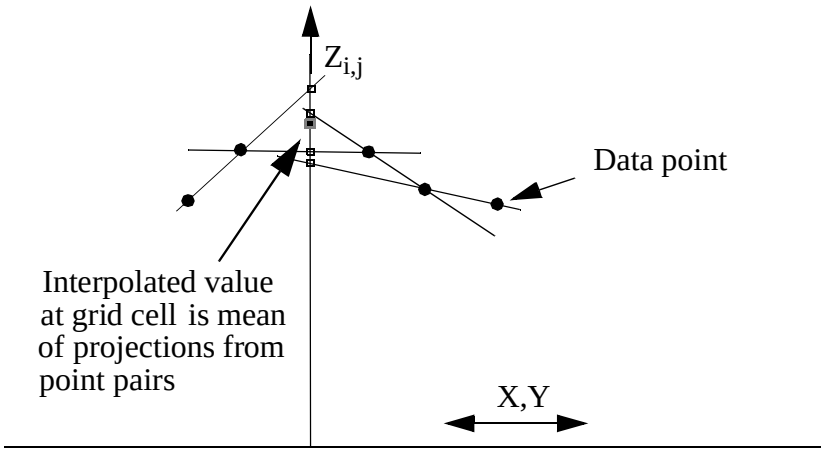


Figure 13.1 Trend projection interpolation.

13.2.3 Search Patterns

Two major variables have a great influence on the two approaches to interpolation discussed above. The first variable is how the neighborhood is chosen. A neighborhood can consist of either a given number of points according to some criterion or of all the points that satisfy certain conditions. The second influencing variable is the relationship between the spacing of the points and the spacing of the grid, and the related problems of the choice of the interpolation area, and the orientation of the grid.

The simplest way of determining which points to use is to include all points in the interpolation. This method, while avoiding the sorting by distance described above, violates the neighborhood property nature of terrain data. Given that we must select points, one approach is to limit the search to points within a certain coarser resolution cell, de-

termined by partitioning up the whole map. This leaves the problem of dealing with the discontinuities that result at the boundaries of cells. The brute force method sorts all the distances and takes for the neighborhood all points that fall less than a given distance or search radius away (Figure 13.2, upper left). This method can result in finding no or few points, especially at the map edges and corners.

An alternative method , which avoids this problem, is to take the nearest R points, regardless of distance or direction (Figure 13.2, upper right). Often, the nearest three, four, or eight points are selected. This method suffers when clusters of points exist, because any interpolation of kernels within the vicinity of the cluster will be overly influenced by the cluster. The problem of directional clustering can be overcome by choosing the nearest points within each of the four quadrants determined by the grid (Figure 13.2, lower left). Finally, each quadrant can be divided to assure equal representation of each octant (Figure 13.2, lower right). Both of these variations suffer from boundary effects; that is, at the edges and corners of the map grid, no points may be available within a quadrant of octant. This problem could be solved by moving to select the points for the neighborhood with another search method when no points are found in the original search area. Clearly, some very complex search strategies can be constructed using combinations of these approaches.

The second of the problems facing the interpolation search strategy is the following: How should the grid relate to the points? The orientation of the grid is usually the same as the coordinate system in use, although this is not a requirement, and for spherical data, where data should be interpolated across the poles, for example, some subtle refinements are necessary. The size of the grid is first determined by the spacing of the grid cells, that is, the grid resolution, and also by the area mapped. The total map area for interpolation is usually rectangular, and the map either is buffered by a data-point void strip around the edges, or it extends over an area for which points lie beyond the map. Clearly, the latter is preferable, because values at the edges of the grid can be interpolated using real data outside the map area, giving the map reliability across the edges. When a void exists at the edge, some interpolator, especially the trend projection methods, can introduce artificial features into the terrain, called *edge effects*.

The grid spacing is also critical. At one extreme, a very large number of evenly distributed points and a coarse grid could allow the grid to be superimposed on the data and all cells to be assigned simply by averaging elevations within cells. At the other extreme, an extremely fine grid could ensure that every data point falls at a grid cell center exactly, leaving the problem of filling in the blanks. In fact, most irregular point distributions contain data-rich and data-poor areas. Statistics like the nearest-neighbor value may be helpful in determining grid spacing; the mean, minimum, and maximum point separation are also useful.

If the source of point data is field measurement, then an effort to get both a uniform spread of points and to collect the data extremes is critical. A map of the distance from each kernel to its nearest data point is a useful aid in determining where to draw the map boundaries and what grid spacing to use. No “best” spacing exists, because the usual limitation is the collection of the data. The sampling theorem states that once a grid spacing is chosen, all features with a spatial size less than twice the spacing are essentially eliminated from the grid, at best becoming random variation or noise. Thus when a grid spac-

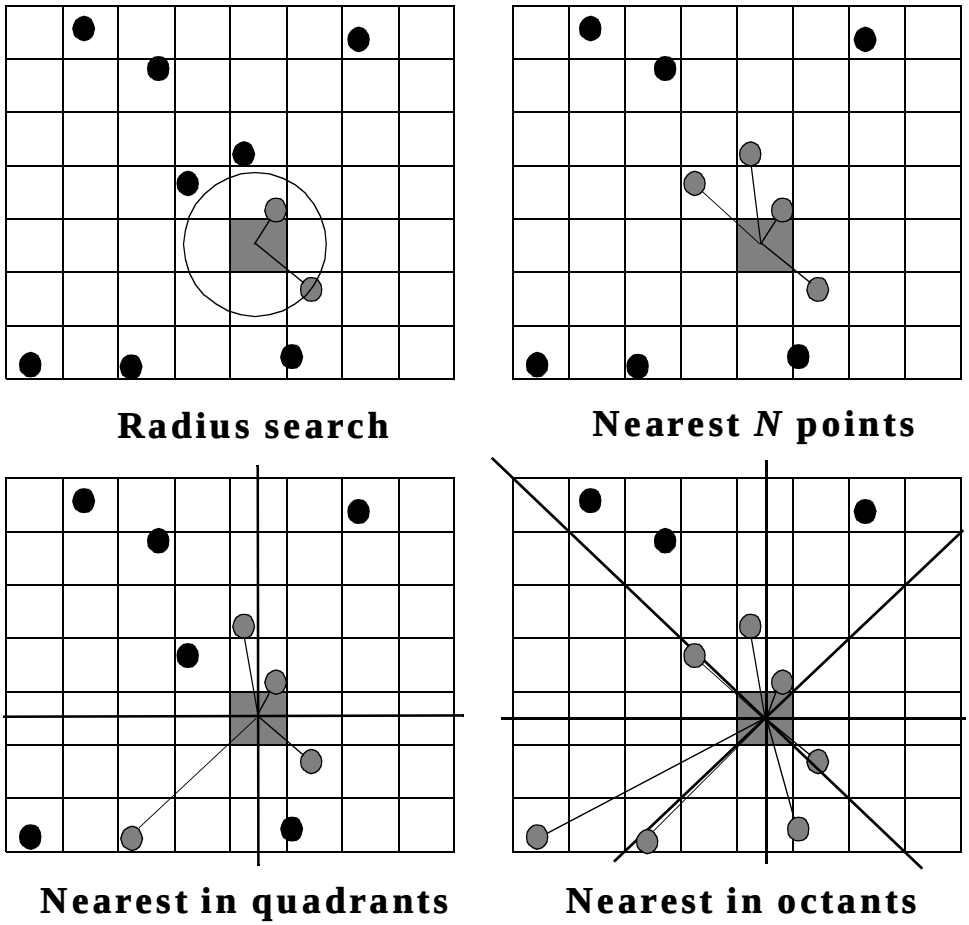


Figure 13.2 Interpolation point search strategies.

ing is chosen, the purpose of the intended symbolization of the terrain should be carefully considered so that features are not eliminated by smoothing.

13.2.4 Kriging

A very large number of methods have been used in the interpolation of point data to a grid. Only one method, however, uses statistical theory to optimize the interpolation. This method, called *kriging*, advanced by Matheron and named for the originator, D. G. Krige, was originally applied to ore bodies in gold mining. Kriging is based on the mathematical theory of the *regionalized variable*. Regionalized variable theory breaks spatial variation down into a drift or structure, a random but spatially correlated part, and random noise.

Thus while hiking up a mountain, the elevation drift is up along a line between the trailhead and the summit, even though we may find local drops to traverse along the trail (random correlated elevations) and boulders to step over while doing so (elevation noise).

Kriging involves a multistep approach to interpolation. First, the drift is estimated using a mathematical function. If none exists, then a good estimate of any map elevation is the mean elevation of the data points. With drift, however, the expected elevation difference with a given separation between the kernel and a point is given by the semivariogram. The semivariogram is computed within regions that are determined using one of the search strategies discussed above.

The semivariogram is then used to statistically fit a model, usually an exponential but a linear model when the semivariogram has no obvious sill, to the distribution. This allows the estimation of semivariance and also the estimation of the weights to be used in computing the local moving average. The weights are chosen so that the best unbiased values are used and so that the estimation variance is minimized. A more detailed discussion, with numerical examples, is contained in Burrough (1986).

Because kriging yields a surface that passes directly though the data points, and because the technique also yields the estimated variance at each interpolated point, the technique is statistically superior to the interpolation methods discussed above. The method is available in several computer packages, even on microcomputers, for example on the SURFER package. The technique of universal kriging works best for data with well-defined local trends. When it is difficult to use the points in a neighborhood to estimate the form of the semivariogram, the model used is not entirely appropriate, and the interpolation may be no better than another method.

While kriging is optimal in statistical terms, it is very computationally intensive and can take up a great deal of computer time, especially for large numbers of data points and large grids. Burrough (1986) has noted few comparative studies of the results of kriging compared with other methods. Figure 13.3 shows the results of gridding the Everest point data set from the companion disk: first using inverse distance squared weighting with a four-point, nearest-neighbor search; and second using universal kriging. In the figure, areas of discrepancy between the two methods are shown in the difference image at the bottom. Kriging estimates are significantly higher than those of the distance weighting in the data-poor zone of the map such as the corners. In other areas, the discrepancies in the distance weighting are over- and underestimates, with the regions of highest slope predominating. Clearly, especially in small grids, boundary effects are major and can hold the balance between differences in the two interpolations.

Research into interpolation continues, and algorithms for gridding are becoming increasingly sophisticated. The major source of data to be interpolated, in addition to field or survey data, is data digitized from contours. Contours are a special case, because they are already graphic interpolations of the point data. In many cases, contour lines contain more information about the terrain than simply the elevations, because when drafting the map, the cartographer often traced features detected on the ground or drew the contours to show streams. Digitizing points from contour maps, therefore, is complex. Simply scanning the contour separation of a map is usually a poor way of generating digital terrain. A more effective approach is to use surface-specific point sampling and then to use either a TIN or an interpolated grid with known interpolation properties to map the ter-

rain.

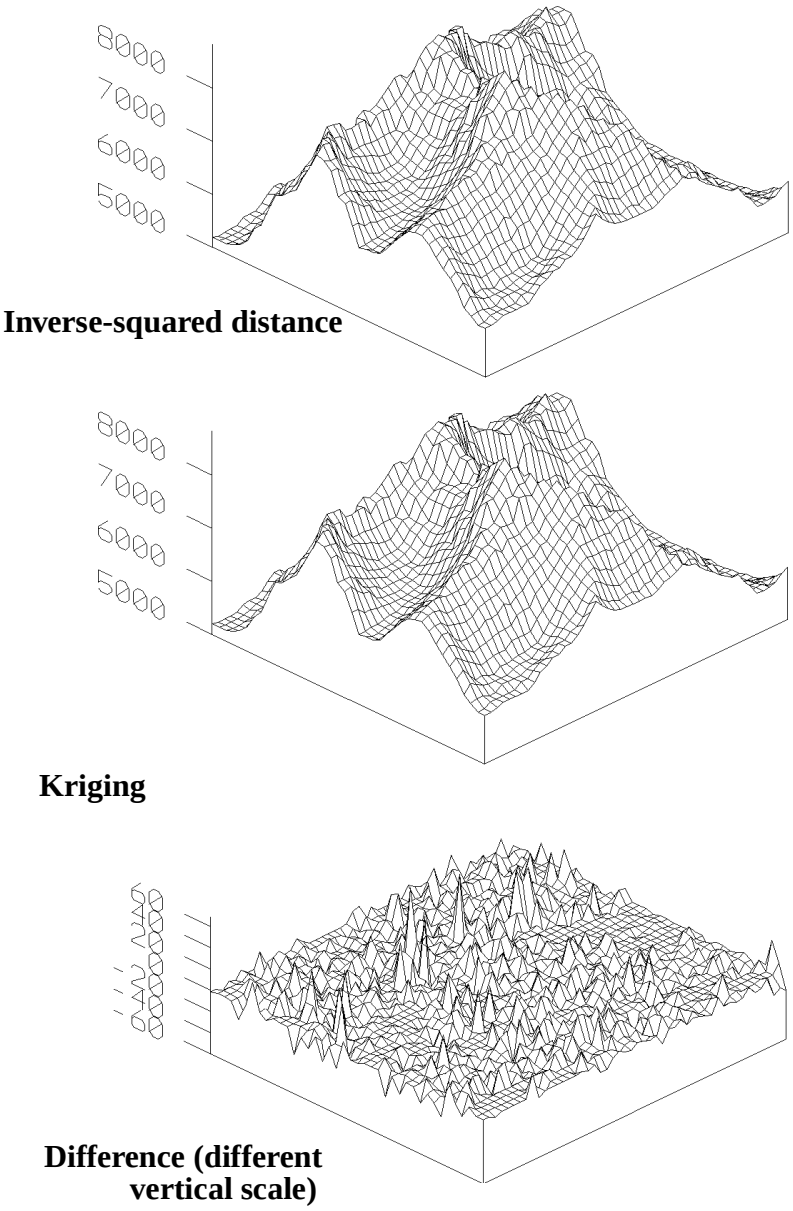


Figure 13.3 Comparison of two grid interpolation methods.

3.3 SURFACE-SPECIFIC POINT SAMPLING

The logic behind surface-specific point sampling is that contour maps, and even field elevation data collected by topographic surveys, contain information about the terrain surface that would be lost by placing a grid over the map and digitizing elevations at grid cells. Significant features that dominate the form of the terrain are streams, ridges, summits, saddle points, and the bottoms of depressions (Douglas, 1986). In addition, actual point elevations are often available on the map as benchmarks, data collection points, or spot heights..

Ignoring the structure or "skeleton" of the terrain, one approach to converting a contour map into either a grid or a TIN is to digitize points along the contours with the attribute of the elevation of the contour. This may work well in only two circumstances: when the point density along the lines is about the same as the map spacing between the contours and in terrain that is very rough. In most cases, more points are digitized along the contours than between them, as shown in Figure 13.4.

When this is the case, interpolation, especially when only a few points are used in the point search, will result in artificial plateaus within the loops of the contours. Since contours usually move back and forth much like a stream, the result is a series of flat terrain steps or plateaus along the contour line, often called the "wedding cake effect" as shown in Figure 13.5. The wedding cake effect is often invisible if the digitized data are only proofed by contouring. It becomes obvious when the resultant grid is hill-shaded.

These steps are artificial and are the result of poor digitizing technique associated with the weaknesses of certain interpolation methods. Avoidance of this problem means taking into account the skeleton of the terrain. Digitizing software which permits an automatic increment by the contour interval with each point digitized is particularly valuable when digitizing streams and ridges. An effective strategy is to start at the low point on a stream or ridge, and to continue up until the high point is reached. Saddle points, summits, and pits (depressions) are then places where the stream and ridge lines converge (Figure 13.6). This terrain structure is sometimes used in environmental modeling to partition the terrain into slopes and watersheds divided by streams and ridge lines, plus features like peaks, depressions and other significant points. Algorithms exist to detect these features in grids, and to vectorize the features for use in GISs.

Points along the skeleton are the targets of terrain data capture. Additional ridge lines can follow significant features up slopes, such as the ridge edges formed by a stream incision. The significant points, such as summits themselves, and other benchmarks, should also be entered. Finally, the intermediate empty spaces can be filled with the occasional point digitized along contours, although because of the wedding cake effect, following a single contour should be avoided.

The result is a set of points which best represent the surface of the terrain. If the TIN were generated from these points, the TIN would be a best-fit model of the surface, and the triangles generated would be good representations of the actual slope facets on the ground. This set of points is also optimal for interpolation, since the interpolator is often better at assigning intermediate elevations from the skeleton than the original manual cartographer was at drafting contours.

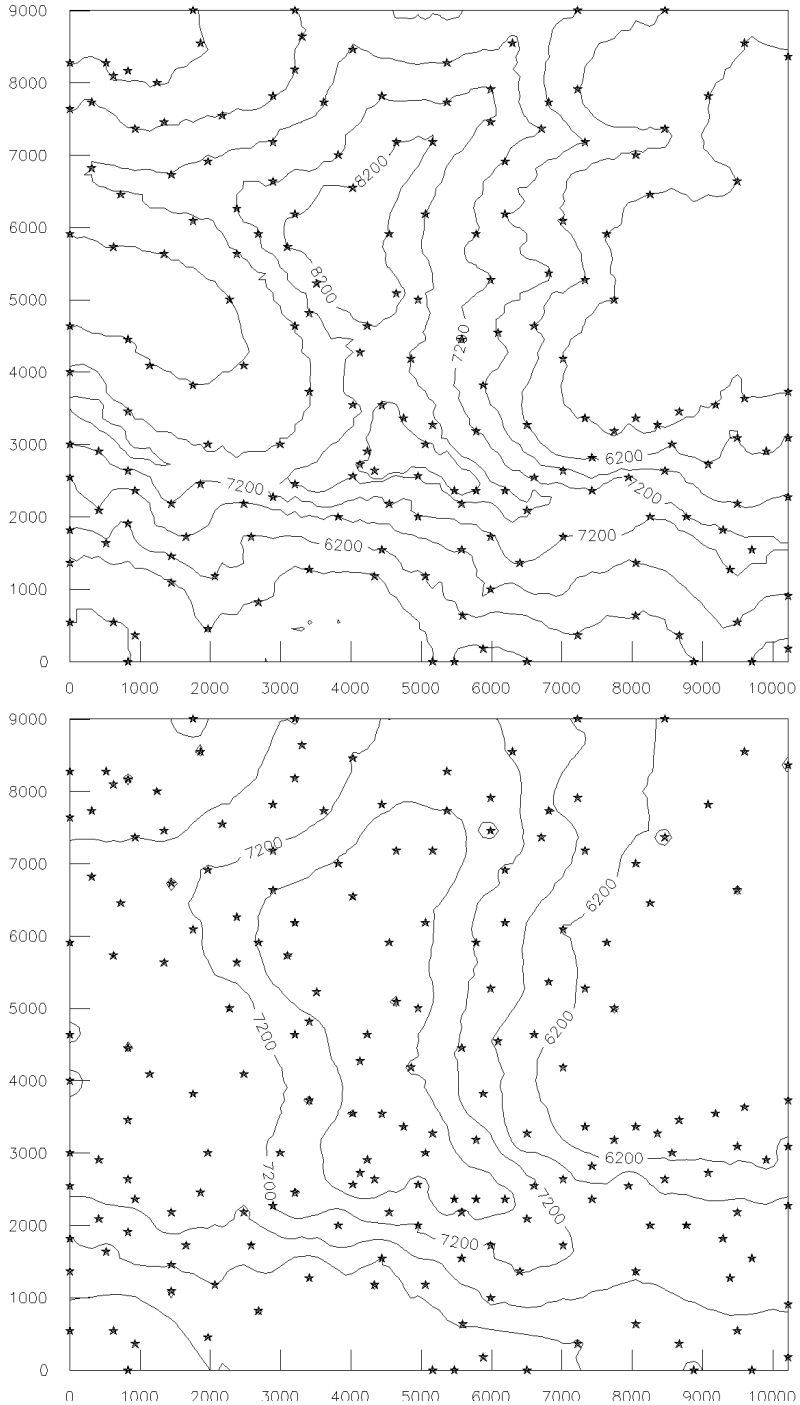


Figure 13.4 Errors in gridding from digitized contours. Top: Gridded from all points. Bottom: Gridded only from the points shown along contours.

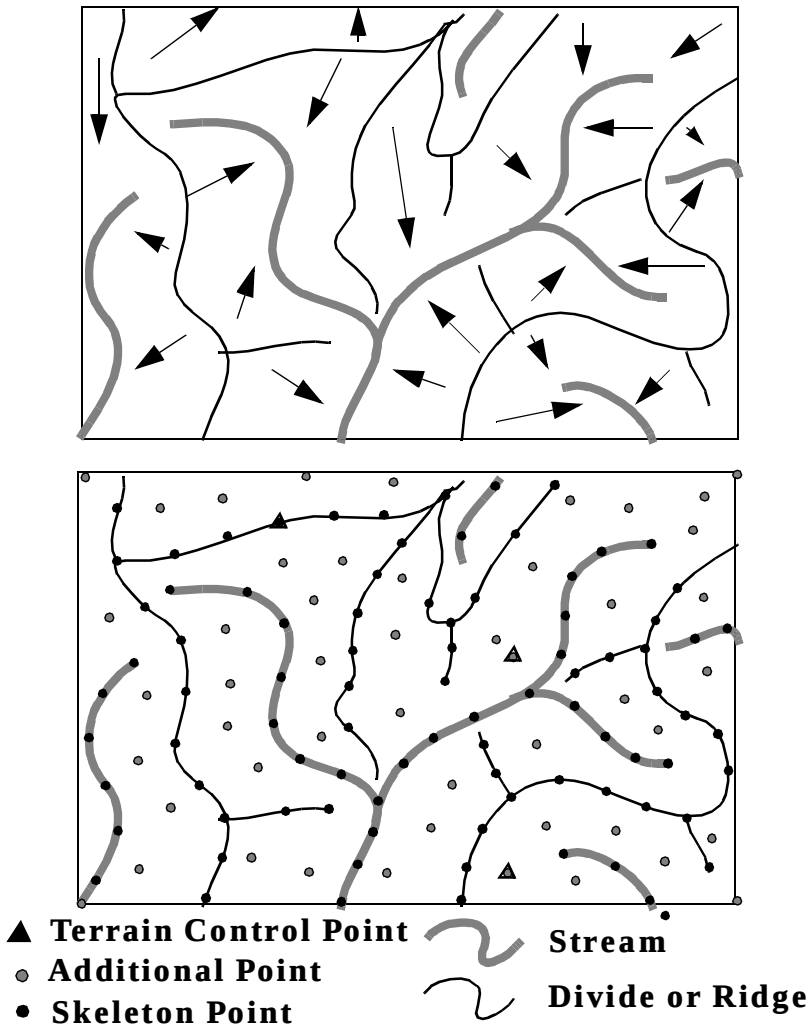


Figure 13.6 Surface-specific point selection.

13.4 SURFACE MODELS

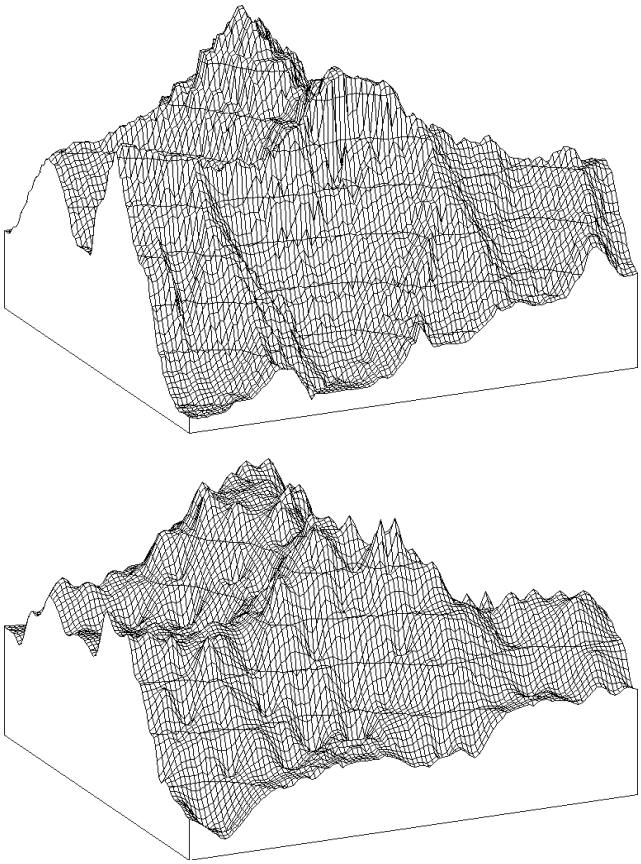


Figure 13.5 The "wedding cake" effect. (Same data as Figure 13.4.)
Top: All points. Bottom: Points on contours only

3.4 SURFACE MODELS

The interpolation methods discussed so far all fit the interpolated surface by local adjustments. Other methods, especially those which for analytical or modeling purposes seek to generalize the surface, use a global or entire surface approach. The *surface models* fall into categories based on the mathematics that describe the surface. The two major surface models, which are global in scope, are polynomial series and Fourier series. Each has analytical powers beyond generalization, and each has been used extensively in analytical cartography.

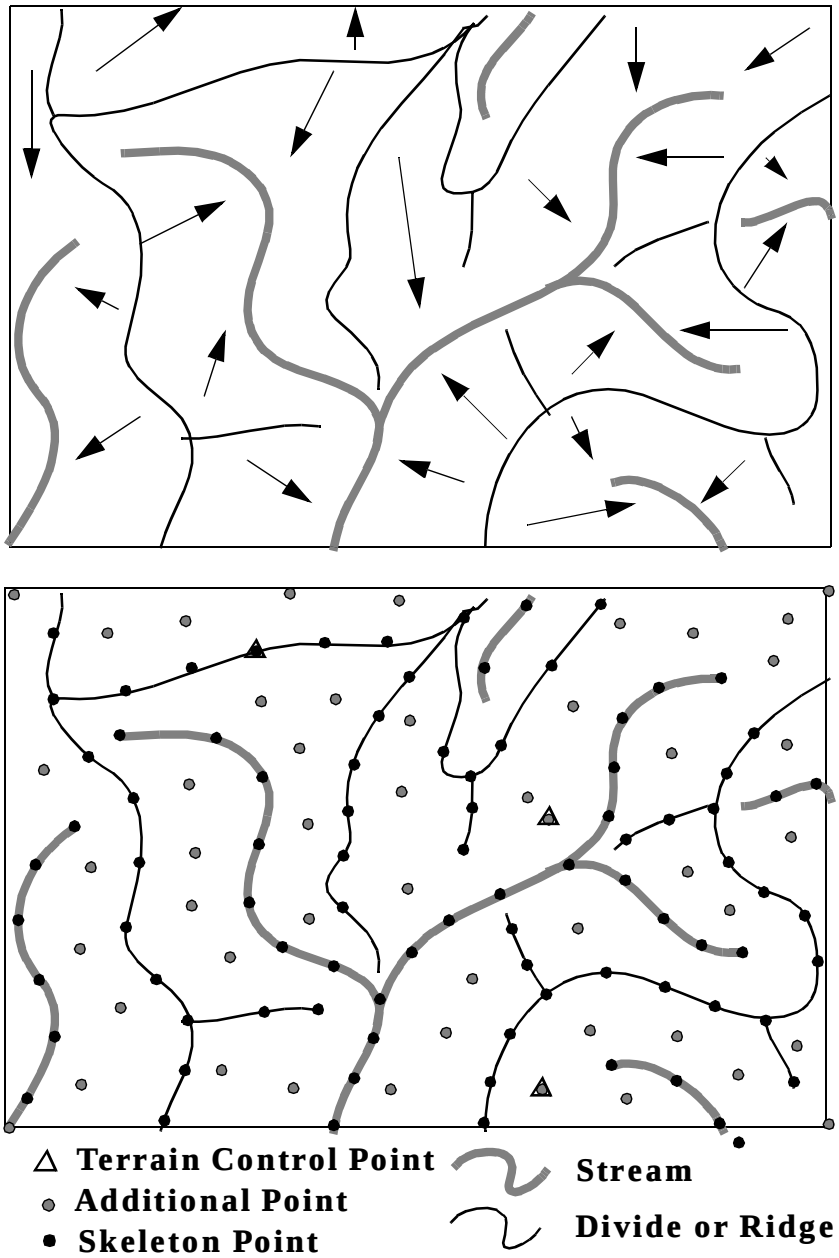


Figure 13.6 Significant points in the terrain skeleton.

13.4.1 Polynomial Series

Polynomial series are power series in that they are series summations of increasing powers of eastings and northings and their cross products. In the simplest form, a linear surface can be represented mathematically as

$$Z_{x,y} = \beta_0 + \beta_1 x + \beta_2 y$$

As the first constant increases, so the overall average height of the surface increases. As each of the others increases, so the linear plane surface represented by the equation dips or tilts down more and more up to the west or south, respectively. Because a flat tilting plane is rare as terrain, real data seldom fall on such a plane but overshoot and undershoot it. Figure 13.7 shows an artificial surface generated by a computer program. This surface is a cubic surface function, where elevation is a function of *x*, *y*, *x* and *y* squared, and *y* cubed. Such a surface can be thought of as having two components, the strongest of which is the overall trend in the data, the drift or trailhead-to-summit path discussed above.

On top of this trend are the local variations, the overshoots and undershoots. These are called residuals: overshoots are positive residuals and undershoots are negative. Statistically, the residuals can be used to fit the linear equation above to a set of points or a surface using least squares. The technique computes the β coefficients for the surface equation which minimize the total of the residuals squared, because some are positive and some negative. The `trend.c` function in the program `terrapin` on the companion disk performs this computation. In the case of Figure 13.7, the β coefficients were:

b_0 = 177.189765
b_1 = 0.871156
b_2 = -3.313856

The b_1 applies to the *x* direction, and the b_2 to the *y* direction. Thus elevation increases with *x*, but decreases with *y*, giving the viewing angle of Figure 13.7 as from the north west.

Because trend surfaces are computed by least squares, correlation statistics can be determined. Again, in the case of Figure 13.7, the percent goodness of fit was 99.7 with an R^2 of 0.994. The resultant trend therefore has an associated "goodness of fit" measure, that gives the percentage of the variance in the elevation values accounted for by the linear trend model, which misses only 0.01% of the overall variance.

The fitted trend surface looks almost identical to the original data except that when the pattern of residuals is examined, rounding error (grid cell values were rounded to the nearest whole meter) and the cubic trend become obvious. If the analysis were repeated using a square or cubic trend surface, the fit would have increased to close to one, with the integer rounding error remaining as the sole source of error. This rounding error is not independent of the data, and so it would continue to have spatial pattern and to be related to the relief of the map, that is, the numerical range of the *z* values. The mapping of residuals has found many applications in geology and geography (Davis, 1973).

When residuals cluster, usually a larger scale spatial process is at work. Increasing the order of the polynomial is one way to capture this variation in the trend surface model. The least squares principle is applied to higher-order polynomials.

For example, a quadratic surface would be given by

$$Z_{x,y} = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy$$

As the surface becomes more complex the number of β terms increases. A cubic surface has 10 terms. At some point, when the number of terms becomes high, the value of the trend surface as a generalization of the original data becomes questionable, because the model has more parameters than we have data.

For cartographic purposes, such as generalizing terrain for mapping or for simplifying slope and aspect computations, a trend surface of appropriate order is adequate. As a theoretical model of terrain, however, the trend surface may introduce unacceptable error during generalization. Davis (1973) provided an excellent discussion of the use and theory of trend surface analysis in geology and published a FORTRAN computer program for its implementation.

13.4.2 Fourier Series

An alternative method for the generalization of terrain is to use a theory developed by the French mathematician Fourier. Fourier showed that any sequence of data could be represented by the sum of a set of trigonometric functions, sines and cosines, over a long enough range of wavelengths. Using space rather than time, because Fourier methods are most commonly used for time series analysis, the wavelength is the distance at which the wave repeats itself. The amplitude of the wave for terrain is the elevation, and the phase angle, the point at which the wave starts, is the elevation at distance zero.

A Fourier generalization of terrain works in two dimensions, *x* and *y*, and gives values for elevation. To compute the parameters of the trigonometric functions, Fourier coefficients are computed for all ranges possible within a data set. Because the Fourier analysis of irregular data is complex, the usual data used for Fourier analysis is the grid. Analysis proceeds starting at the longest-wavelength wave that will fit the data, that is, with a wavelength of half the map length in *x* and half the map length in *y*. The response or "power" of each pair of *x* and *y* wavelengths is computed, a value similar to the percentage of variance explained in least squares. The wavelengths used are the map length *L* over 2, 3, 4, 5, and so on, until the distance associated with the wave approaches twice the grid spacing, a level at which the Gibbs phenomenon introduces seemingly random errors. This gives a new grid of Fourier coefficients, which are normalized to give the "power" of each pair of wavelengths or "harmonics."

The generalization involves selecting particular pairs of harmonics that are major contributors to the structure of the data. In most phenomena, including linear phenomena such as voice and radio transmission, just a few of the wavelengths carry almost all the structure of the data. These harmonic pairs can be extracted and their Fourier coefficients used to reconstruct a data grid based on their values. This is another example of an inverse

transformation.

.

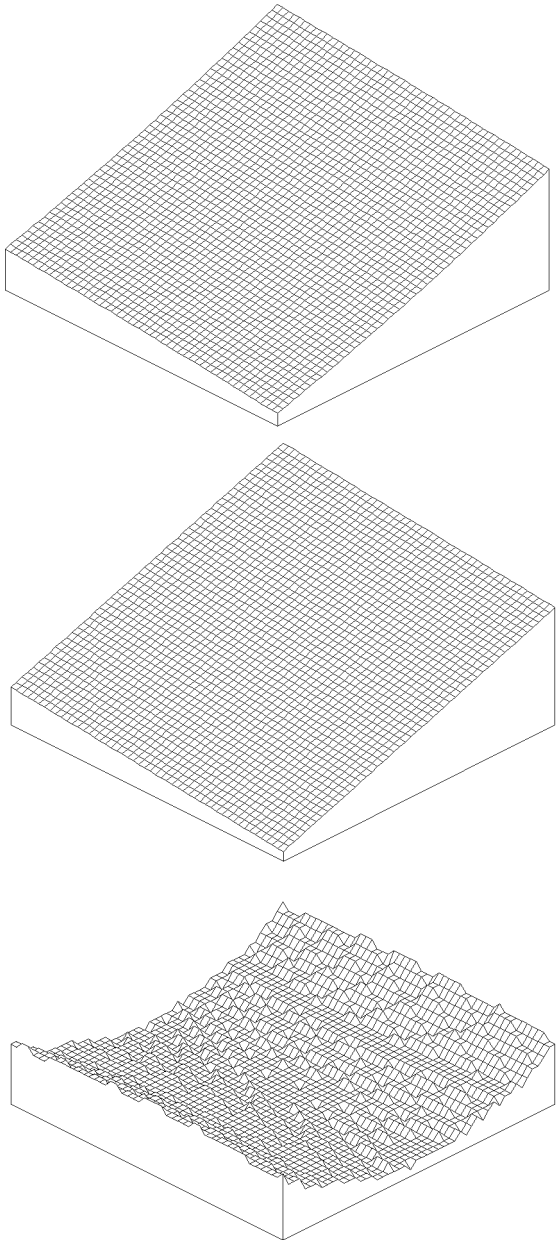


Figure 13.7 Cubic function, linear trend surface and residual (scaled).

The conversion from the spatial to the frequency or wavelength domain is known as the *Fourier transform*. The inverse Fourier transform can be used to reconstruct either the entire original data grid (less some errors due to boundary problems) or a part of it. A common use of the Fourier transform and its inverse is to eliminate the higher frequencies, that is the variations at small distances. The forward and inverse Fourier transforms give an objective method for this form of generalization.

Davis (1973) gave a FORTRAN program for performing two-dimensional Fourier analysis. Davis's method used the discrete Fourier transform, a grid cell–by–grid cell method. Far faster is the fast Fourier transform. This technique, however, requires data sets to be of specific sizes. Algorithms, equations, and C language computer programs for both forms of the Fourier transform are contained in Press et al. (1988).

Application of the discrete Fourier technique to terrain is discussed in Clarke (1988). The byproduct of Fourier analysis of terrain is a precise account of which spatial scales are "active" in a particular piece of terrain. This assists in the choice of a sampling grid as well as in analyzing the processes which have formed the terrain surface.

13.4.3 Surface Filtering

An effect virtually identical to Fourier generalization is possible using local operators. This method is called *spatial filtering*. Filtering is usually performed exclusively on gridded data, and the TIN will therefore be left out of the following discussion. In spatial filtering, the entire grid is processed cell by cell to generate a new or filtered grid. This is done by using a smaller grid, and moving the filter grid step by step, kernel by kernel, over the original data grid. In each case, the filter grid is used to compute a moving average.

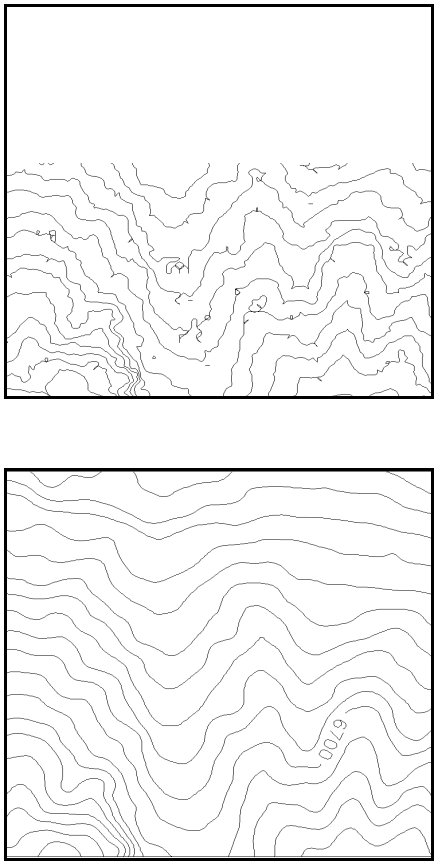
Filter grids are centered on a kernel, and must therefore have an odd number of rows and columns. Filters of 3 by 3, 5 by 5, and 7 by 7 are common. Larger filters lead to problems, because in each application of the filter the original grid gets smaller. Three-by-three filters lose one row from each edge of the grid with each application, five-by-five lose two, and so forth. The filter is placed over the original data grid, and its values are multiplied by the originals to give new values, which are then summed.

The new filtered value for the grid is then the sum of the values over the filter (Figure 13.8). A requirement for the filter is that the weights or values within the filter grid sum to one. Otherwise, the filter has the effect of damping (< 1.0) or amplifying (> 1.0) the terrain. Filter weights can be tailor-made to yield different effects. Equal weights in each of the cells is simply a moving average or smoothing filter. The weights can be distance weighted by weighting the cells in proportion to their distance from the kernel. A common filter is the Hanning filter, a two-dimensional version of the binomial distribution.

Special-purpose filters can be designed to enhance features with a specific size, orientation, or characteristic. A horizontal linear filter, for example, can be used to amplify erroneous scan lines on satellite images. A filter with the shape of a ship, for example, can be used to scan ocean images for ships.

Another commonly used filter is the "lint-picker," a three by three with a weight of

minus one in each cell except the center, which has a weight of nine. Note that the weights



Original Data

125	124	123	121	
121	125	125	126	
118	120	128	129	

1/16	1/8	1/16
1/8	1/4	1/8
1/16	1/8	1/16

Hanning filter

Filter × Original

7.81	15.5	7.69		
15.1	31.2	15.6		
7.38	15.0	8.0		

Summation for kernel

*	*	*	*	
*	123→			
*				

Figure 13.8
Section of the Everest map filtered using a Hanning filter. Top: original data; middle: filtered three times; bottom: filtered six times.

in this filter still sum to one. The effect of this filter is to enhance features that are one grid cell in size and are very different from their neighbors. This is a good definition of random noise, so a lint-picker is often used to generate an image or map that is subtracted from the original to remove random noise, the so-called "computer-enhanced image." Filtering an image is fairly easy to program, but can be computationally intensive because the whole grid must be used and a second version of the grid must be saved in RAM during processing.

The function `filter.c` in the `terrapin` directory on the companion disk allows the selection of one of several filters to apply to a grid in the GRID structure. The weights are stored in a vector, and can be adjusted for any size or type of filter with a little modification. The function shown assumes a three by three window and a Hanning filter. With very little modification, the user can place any filter of any size into the `filter.c` function.

13.5 VOLUMETRIC CARTOGRAPHIC TRANSFORMATIONS

Surface models and filtering are transformations of the three-dimensional cartographic data. So far, the motive for transformations discussed have been to transform between data structures (interpolation) and to transform between scales (generalization). Many other transformations are commonly applied to digital elevation data. The purpose of these transformations is purely analytical; that is, the result is a map with enhanced meaning for a specific cartographic problem. Four aspects of analytical transformations will be discussed in this section: the transformation of elevations to slope and aspect, the automatic delineation of terrain-significant points from gridded data, the simulation of terrain data, and the transformations necessary to provide visibility maps.

13.5.1 Slope and Aspect

Evans (1980) noted that slope is defined by a plane tangent to a surface at a specific point and is specified in terms of the maximum rate of change of altitude (slope) and the compass direction associated with the maximum (aspect). Slope, therefore, also has a local neighborhood over which it is computed, usually a three by three region. The maximum slope is familiar to skiers as the fall line, and the aspect is the direction in which the fall line trends. When slope is zero, the terrain is flat and the aspect is undefined. Slope is computed by solving a best fit surface through the points in the neighborhood and by measuring the change in elevation per unit distance in this neighborhood and the direction. These values can be assigned as data to a new grid. Using the TIN, each triangle has a uniform slope within the triangle and also a single aspect. The slope and aspect values in a TIN are discontinuous at the triangle boundaries. Figure 13.9 shows values of slope and aspect computed for an elevation grid covering the summit of Mount Everest. The horizontal resolution is 18 meters per grid cell. The slope map shows a limited set of slopes with extremely steep slopes in white. The aspect image shows with gray tones the eight major compass directions, starting with black for south and ending with white for southwest.

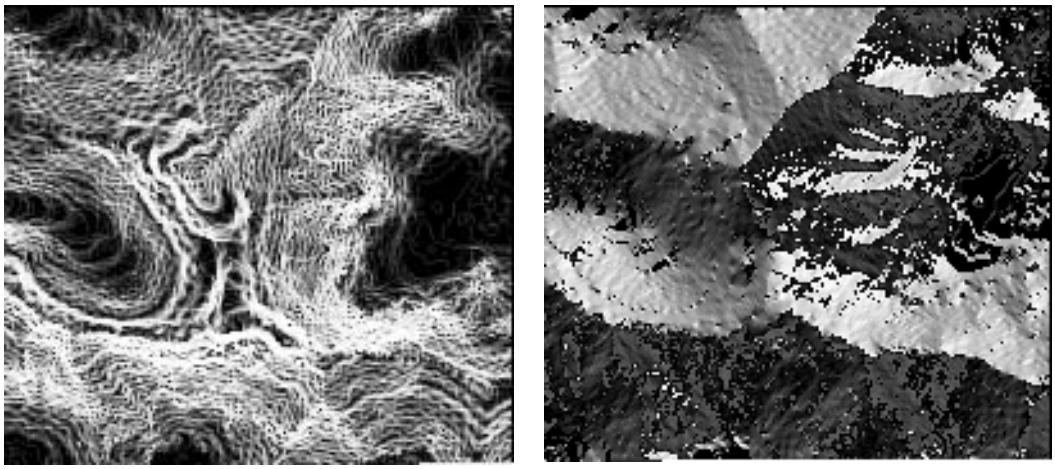


Figure 13.9 Slope and aspect maps for the Everest summit area.

13.5.2 Terrain Partitioning

The primary data structure conversion for three-dimensional data, with the exception of the point-to-grid or point-to-TIN conversion, is the TIN-to-grid conversion, and vice versa. A TIN-to-grid conversion is comparatively simple. Either the points in the TIN can be used in a grid interpolation using one of the methods discussed in this chapter or the elevation values at every point in the grid can be computed using a linear, polynomial, or spline fit to the data in the triangle that includes the point.

The grid-to-TIN conversion, however, is more difficult. There are distinct advantages to performing this conversion. TINs are very compact, are suited to symbolization using polygon rendering hardware and software, and are computationally less demanding than grids. This makes TIN the preferred data structure for solid modeling, for terrain analysis programs, and for surface visualization applications. Essential to the advantages of the conversion, however, is the ability of an algorithm to detect critical points in the landscape. These are the same points discussed as surface-specific in the preceding section. A factor of reduction of 18 times was achieved for conversion of triangles formed by grid points to a TIN based on critical points (Scarlatos, 1989).

The detection of stream and ridge lines in gridded terrain data was discussed by Douglas (1986). Peaks, saddles, and pits are detectable by filtering. The advantages of detecting the significant points is not only for creating the TIN, but also for automatically detecting river channels, for selecting ridges for hill shading and intervisibility, and for dividing the surface into sloping facets for the modeling of hydrology. As yet, however, no simple algorithm to convert a grid to a TIN exists, other than the worst-case solution of dividing every grid cell into two triangles.

13.5.3 Terrain Simulation

Interest has recently been focused on the simulation of artificial terrain with realistic characteristics. Such terrain is used in flight simulators, movies, video games, and models of natural processes. The most common means by which such terrain is produced is using the mathematical methods of fractal geometry. Fractal geometry defines a property called self-similarity, in which as the scale gets larger and larger, geometric form is simply repeated. Two methods for the generation of fractal terrain are commonly used. The first is the midpoint displacement method, in which an original square is drawn, its four corners are displaced either up or down at random, and then the square is divided into four quarters whose corners are displaced by the same amounts, and so on. In each case, the center cell is the computed average of the four corner cells. This method was first used by Fournier, Fussell, and Carpenter (1982) for animations and is fast, simple to program, and gives satisfactory results. For many divisions, however, the technique leaves creases at the edges of the larger cells, for which Jeffery (1987) provided a solution.

A second algorithm, given in Jeffery and attributed to Voss, adds the displacements to all points at each scale instead of just the midpoints. Because this allows steps down of more than a half at a time, the creasing problem disappears. Jeffery (1987) published pseudo-code for his algorithms, and Pascal versions of the programs are available. Information on accessing the programs is contained in an editor's note in the article. Although many fractal surfaces are perfectly adequate for simulations, cartographically incorrect characteristics arise. For example, as many pits as peaks are generated. Often, additional texture algorithms or filtering are used to produce the final "fractal forgery."

13.5.4 Intervisibility

The intervisibility problem can be stated as follows: Given a digital cartographic representation of a terrain surface and a single point on or above the surface, determine the set of regions on the surface that are visible from that point. This set of regions is the opposite of the set of invisible regions. A map of either visible or invisible regions has immense value. Intervisibility maps are used in siting radar and television transmitters, in locating fire towers, in planning ski resorts and housing developments, in highway planning, and in military planning. The solutions to several related problems, such as planning sets of points within view of each other, hiding environmentally obtrusive buildings and land uses in the terrain, and computing the regions where an aircraft is visible to radar, are of direct use in many areas.

The simplest approach to determining intervisibility is to connect a viewing location to each possible target and to follow the line back looking for points that are higher, a method known as *ray tracing*. Any intervening higher point along the ray would screen the target from the viewer (Figure 13.10). Because this involves a very large number of computations, the screening out of invisible areas along rays is advantageous. Intervisibility is easier using the TIN than using a grid. Sutherland et al. (1974) reviewed several algorithms, and DeFloriani et al. (1986) provided a TIN algorithm and Anderson (1982) a grid algorithm. An improved method for the grid was published by Dozier et al. (1981).

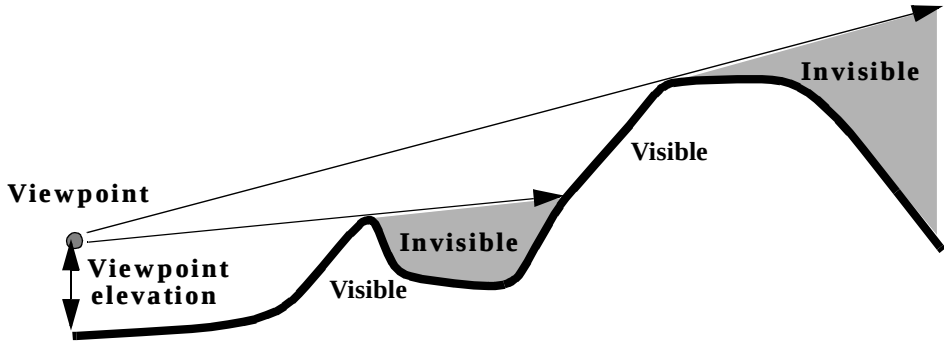


Figure 13.10 Intervisibility using ray tracing.

A primary use of intervisibility algorithms is in selecting viewing locations for gridded and realistic perspectives. The algorithms are also used in calculating the hidden sections of perspective views and in eliminating hidden lines from gridded perspective views. Figure 13.11 shows an application where rays are traced from a viewpoint and are used to assist the cartographer in the selection of a viewing position for generating a realistic perspective view. This step can act as a preprocessor for this computationally intensive task.

13.6 TERRAIN SYMBOLIZATION

The final three-dimensional transformation for cartographic data is that which produces a map. Maps of three dimensional data have problems similar to those with map projections, that is, how to produce on a flat two-dimensional plane a map depicting a volume in three-dimensions. Several cartographic techniques have been devised in recent history for the cartographic symbolization of three-dimensional data. The definitive study of manual methods is Imhof (1982). These methods include contouring and hill shading as well as block diagrams. With computer cartography, many additional methods have arisen, and many methods that required much labor by hand have become available to all.

13.6.1 Automated Contouring

The oldest nonpictorial method of representing three-dimensional data is to use the isoline, first used to show magnetic variation by Sir Edmund Halley. When applied to terrain, the isoline is called a contour and is a line joining points with equal elevation. Reference contours are drawn thicker and numbered at short breaks on the straighter segments. Closed depressions are annotated with hatch marks to distinguish them from hills. Many computer contouring programs automate these features.

Computer contouring can be performed from either a TIN or a grid. Starting at one side of the map, the highest elevation, or the lowest elevation, a first step is to determine whether or not a contour line should appear in a given grid cell or triangle. If the answer

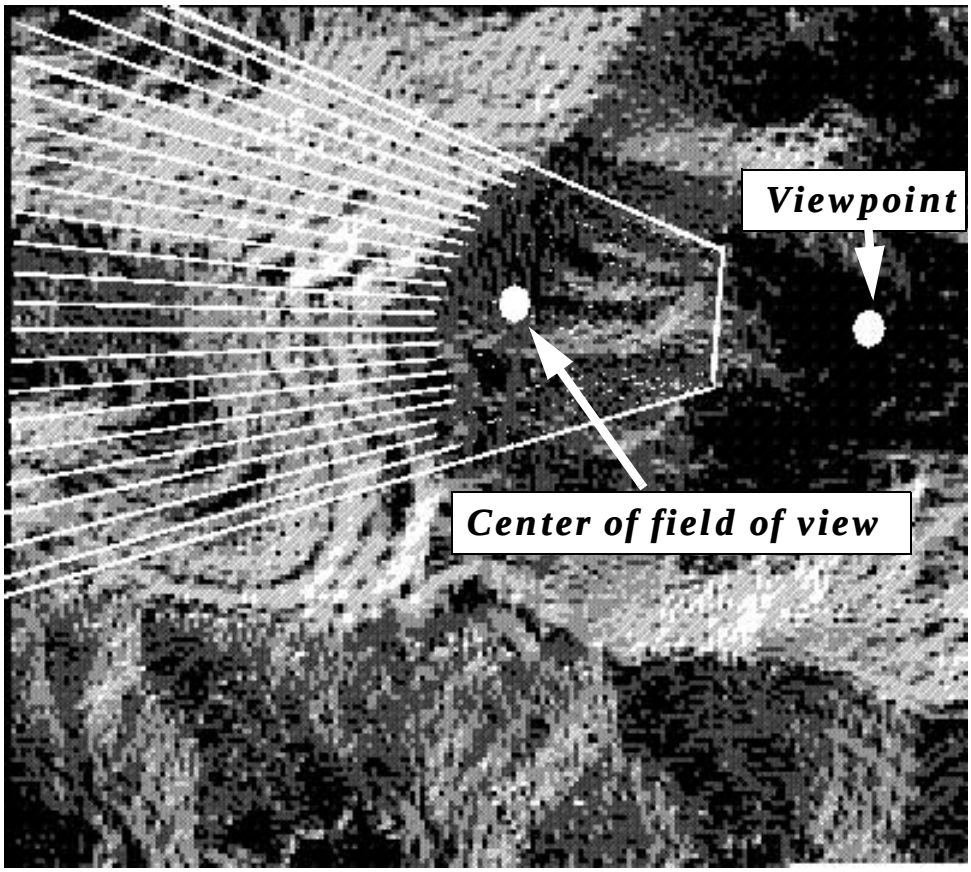


Figure 13.11 Ray tracing for viewpoint selection.

is yes, the next step is to determine points at the edge of the square or triangle where the contour enters and leaves. The next stage uses an interpolator—a quadratic trend surface, an average slope, or Lagrangian interpolation (Crain, 1970)—to generate points along a curve within either the square or the triangle.

An exception in the grid cell case, the case of a saddle point, complicates this process. In this case, a center point is computed as the average of the four corners and is used to move the contour to one side of the cell center. The final contour lines are then smoothed, using weighted averaging or spline functions. A final map consists of a set of points for each contour still structured by grid cell or TIN triangle. In an effort to assure continuity, the contour lines are often resorted so that each continuous loop is drawn without breaking the line, a step that improves plotter output considerably but is unnecessary for many output devices.

As an example, Figure 13.12 is an automated contour map of the Mount Everest summit area, with the exception that the generated contours were smoothed using a different tension factor for the spline smoothing. In some cases, using incorrect values for these coefficients can lead to erroneous contour lines. Some contours become dots or lines at

Figure 13.12 Automated contouring with different spline coefficients for smoothing of contours.

peaks, some cross, and some converge as lines. This last feature is visible on Figure 13.12. An extensive catalog of these line artifacts on automated contour maps is provided by Krajewski and Gibbs (1994). They list 17 types of artifacts commonly resulting from seven variations of six interpolation algorithms and include causes and suggested solutions for the artifacts.

Several inexpensive packages now permit rapid contouring even on microcomputers. It is important to remember, however, just how many parameters go into the production of a computer-generated contour map. The choices of these parameters, data structures, and methods, not to mention the type of output device, are strong determinants of the look of the final contour map. In some cases, the only difference between a manual and a computer contour map is the computer's ability to reproduce the same map given the original data and parameters used.

13.6.2 Analytical Hill Shading

A terrain representation method that has gained considerably in use with computer cartography is hill shading. Automated hill shading works by computing the three-dimensional vector normal to the surface at each point in a grid or for each triangle in a TIN. This normal vector is at right angles to the plane that defines the maximum slope and faces in the direction of the aspect. This vector is projected in three dimensions onto the vector given by a simulated sun angle. The sun, or light source, as multiple sources are possible, is located off the map, with a given zenith and azimuth. The zenith is the vertical angle of the light from the center of the map, and the azimuth is the compass bearing of the sun's direction. If the normal vector faces away from the sun, the surface at that point will be in shadow. If the normal vector points directly to the sun, the point will be fully illuminated. At other angles between the illumination and the normal vector, the point will be partially illuminated.

Analytical hill shading, first derived by Pinhas Yoeli, records the illumination value across the map, usually for every grid cell in a grid, but also over a TIN. Brassel (1974) noted that two values can be used, a reflectance value as discussed above or a density value, which is the logarithm of the inverse reflectance. Figure 13.13 shows the Everest summit elevation data with hill shading using four different solar illumination azimuth

directions (NE, SE, NW, SW) and solar zenith angles or elevations of 30 degrees. With control over these angles, it is possible to generate impossible illumination as far as nature is concerned. Hill shading is often used to modify color information on maps derived from remote sensing to present more striking images. Hill shading is also important for generating realistic perspective views.

13.6.3 Gridded Perspectives

The first computer cartographic equivalent of the block diagram used so commonly in geology was the gridded perspective map. These maps are produced exclusively from gridded data, because TINs viewed in perspective appear too complex. Tobler (1970) published a FORTRAN program to generate these views without hidden-line processing.

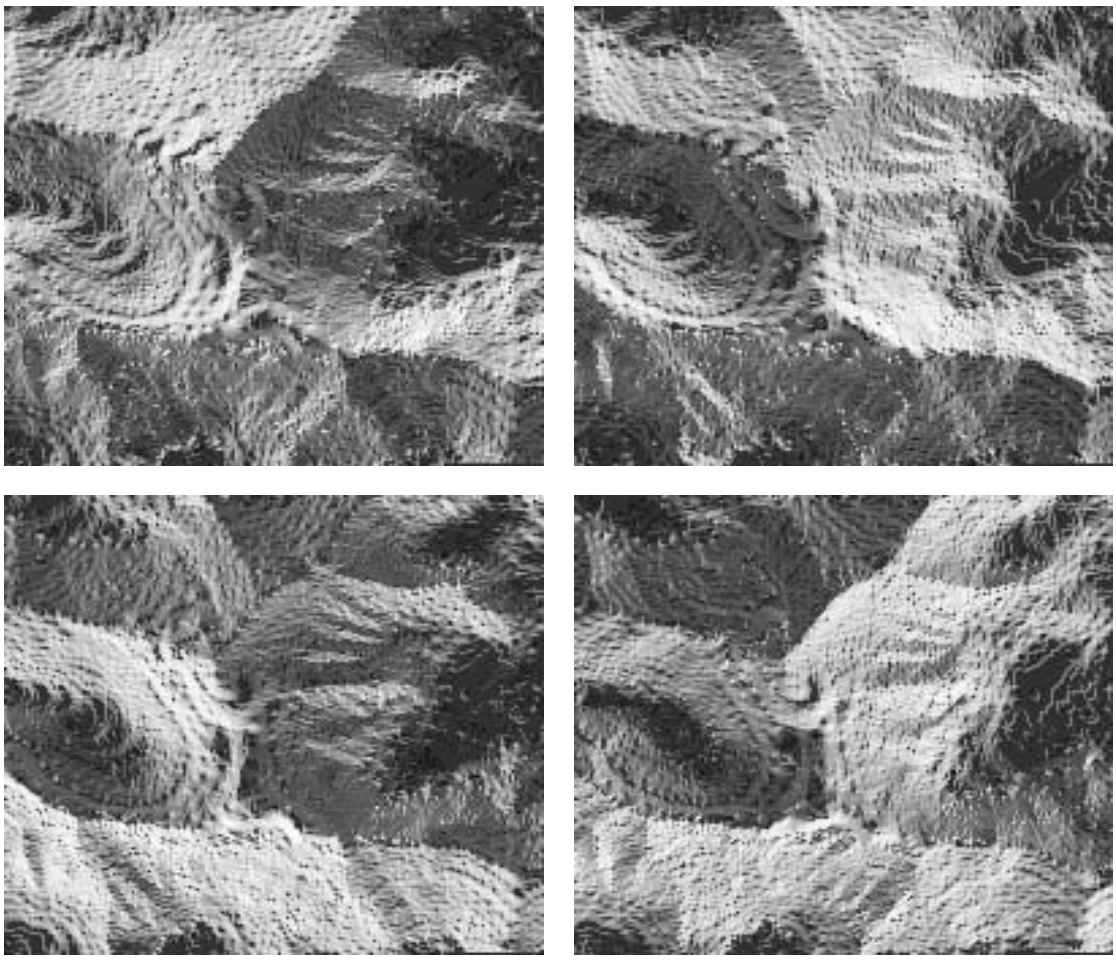


Figure 13.13 Reflectance hill shading with different illumination.

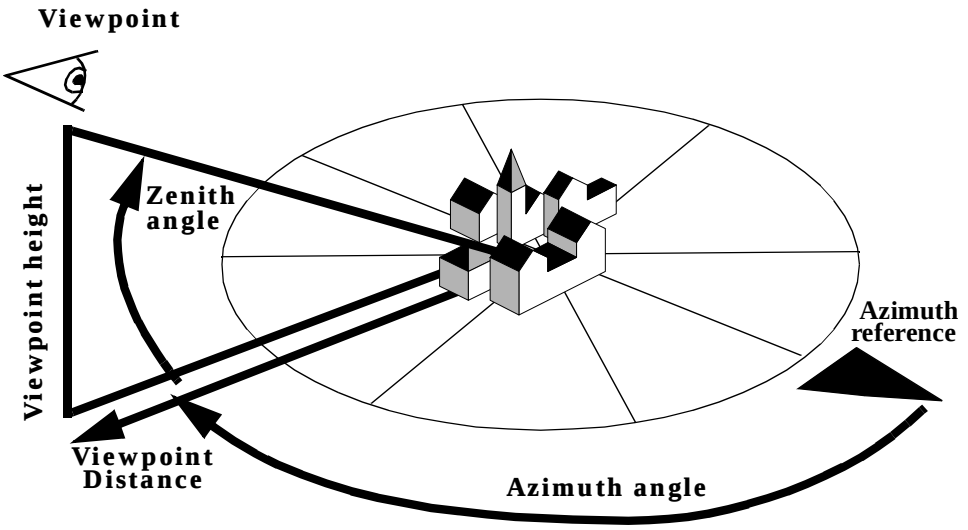


Figure 13.14 Viewing geometry for perspective views.

Many software vendors now sell computer programs to produce gridded perspectives, even on microcomputers with menu control. The critical values in the generation of these images are the viewpoint, the vertical exaggeration, the skirt, the alignment of the lines, and the addition of scales. The viewpoint establishes several things, including the perspective. Views too close have a "wide-angle" perspective, while distant views produce the orthographic perspective favored in block diagrams. Viewpoint can be specified by azimuth, zenith, and distance, by locating a center of the field of view, or by giving a precise triplet of coordinates. The geometry is usually with respect to the center of the volume represented by the image.

The vertical exaggeration is important to the look of the perspective. Textured terrain when exaggerated too much looks chaotic, while too little terrain exaggeration cloaks any actual relief. The skirt is the plain base of the figure. Typically, control over the elevation of the base, as well as whether the surface lines will be drawn over the base, is available. Without a skirt, gridded perspectives seem to "float," but if parts of the underside are visible, added information is gained, especially if the underside is colored differently. Lines on the perspective are usually square to the *x* and *y* axes, but variants are to make the lines parallel to the line of sight, aligned to the *z* axis (raised contours), or sometimes any combination, perhaps in different colors.

The geometry of three dimensional viewing is shown in Figure 13.14.Two variants of fishnet perspective views are stereo plots and anaglyphs, both means by which stereo views can be simulated. In a stereo plot, two images are generated, separated by a 2-degree viewing difference and at a spacing suitable a stereo viewer (Figure 13.15). The separation of 2 degrees between the azimuth angles in the left and right images allow Figure 13.15 to be viewed in stereo because each eye receives a slightly different view, which the brain assembles as a three-dimensional image.

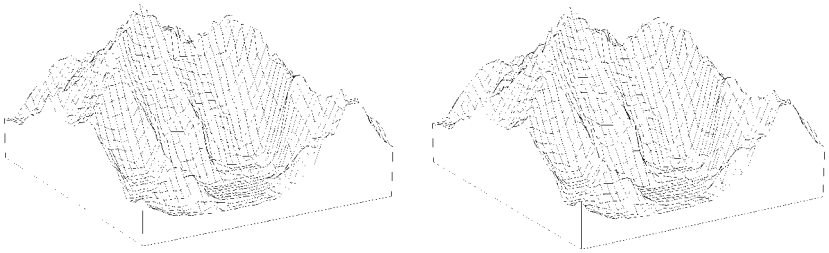


Figure 13.15 Stereoperspective view of Everest summit data.
To be viewed with pocket stereoscope.

Finally, the anaglyph plot is identical, with the exception that the two different views, with a 2-degree viewpoint separation, are plotted on top of each other. One image is plotted in green or blue and the other in red. Viewing the image through anaglyphic glasses, which have red and blue or red and green lenses, produces the stereo effect. The stereo image, combining two color opposites, should appear as black.

13.6.4 Realistic Perspectives

A refinement of the gridded perspective is the realistic perspective (Figure 13.16). Dubayah and Dozier (1986) presented a summary of work on this method, and discussed algorithms. These images can be generated from both grids and TINs, but employing very different methods. TINs are usually rendered using special-purpose hardware or display software, whereas grids are usually processed in batch mode and the image displayed after completion. The more powerful workstations are capable of almost real-time generation of these images, but even powerful microcomputers may take hours of processing to produce a single image. Hours of supercomputer time have been used to produce sets of these images, which can then be played back in sequence to simulate flight and motion. Most of the same parameters as gridded perspectives apply to realistic perspectives. One major problem is the enlarged effect of grid cells very close to the observer, which can appear blocky. Color for these images is often natural color derived from satellite data.

13.7 TRANSFORMATIONS IN REVIEW

This chapter began with a discussion of cartographic transformations. We have seen that a transformational view of cartography is broad enough to encompass both computer and analytical cartography. Although analytical cartography represents the intellectual challenge to the cartographer, the act of producing the maps often takes up much of the digital cartographer's time. In Part IV, an approach is presented that ensures that this expenditure of time is efficient. To incorporate cartographic transformations as algorithms into mapping systems demands that the cartographer understand both the art of map display and the science of computer programming. Both analytical and computer cartography have much to gain from effective computer programs, just as cartography as a whole can benefit from better and more available maps.

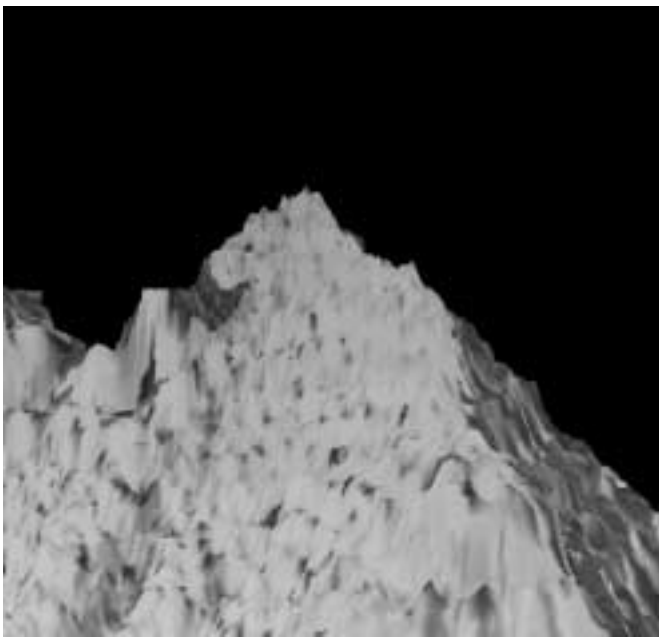


Figure 13.16 Realistic perspective view of Mount Everest.

13.8 REFERENCES

Anderson, D. P. (1982). "Hidden Line Elimination in Projected Grid Surfaces." *ACM Transactions, Graphics,* vol. 1, no. 4, pp. 274–291.

Brassel, K. E. (1974). "A Model for Automatic Hill Shading." *American Cartographer,* vol. 1, no. 1, pp. 15–27.

Burrough, P. A. (1986). *Principles of Geographical Informations Systems for Land Resources Assessment.* Oxford: Clarendon Press.

Clarke, K. C. (1988). "Scale-Based Simulation of Topographic Relief." *American Cartographer,* vol. 15, no. 2, pp. 173–181.

Crain, I. K. (1970). "Computer Interpolation and Contouring of Two-Dimensional Data: A Review." *Geoexploration,* vol. 8, pp. 71–86.

Davis, J. C. (1973). *Statistics and Data Analysis in Geology.* New York*:* Wiley.

Defloriani, L., B. Falcidieno, and C. Pienovi (1986). "A Visibility-Based Model for Terrain Features." *Proceedings,* Second International Symposium on Spatial Data Handling, IGU Commission on Geographic Data Sensing and Processing and the International Cartographic Association, Seattle, July 5–10, pp. 235–250.

Douglas, D. H. (1986). "Experiments to Locate Ridges and Channels to Create a New Type of Digital Elevation Model." *Cartographica*, vol. 23, no. 4, pp. 29–61.

Dozier, J., J. Bruno, and P. Downey (1981). "A Faster Solution to the Horizon Problem." *Computers and Geosciences,* vol. 7, no. 2, pp 145–151.

Dubayah, R. O., and J. Dozier (1986). "Orthographic TerrainViews Using Data Derived from Digital Elevation Models." *Photogrammetric Engineering and Remote Sensing*, vol. 52, no. 4, pp. 509–518.

Evans, I. S. (1980). "An Integrated System of Terrain Analysis and Slope Mapping." *Zeitschrift fur Geomorphologie,* Supplement-B.d. 36, pp. 274–295.

Fournier, A., D. Fussell, and L. Carpenter (1982). "Computer Rendering of Stochastic Models." *Communications of the ACM,* vol. 25, no. 6, pp. 371–384.

Hodgeson, M. E. (1989). "Searching Methods for Rapid Grid Interpolation." *Professional Geographer,* vol. 41, no. 1, pp. 51–61.

Imhof, E. (1982*). Cartographic Relief Presentation*. New York*:* DeGruyter.

Jeffery, T. (1987). "Mimicking Mountains." *Byte,* December, pp. 337–344.

Krajewski, S. A., and B. L. Gibbs (1994). "Computer Contouring Generates Artifacts." *Geotimes,* April 1994, pp. 15–19.

Press, W. H., B. P. Flannery, S. A. Teukolsky and W. T. Vetterling (1988) . *Numerical Recipes in C: The Art of Scientific Computing*. New York: Cambridge University Press.

Sampson, D. (1978). *Surface II Graphics System.* Lawrence: Kansas Geological Survey.

Scarlatos, L. L. (1989). "A Compact Terrain Model Based on Critical Topographic Features." *Proceedings, AUTOCARTO 9,* Ninth International Symposium on Computer-Assisted Cartography, Baltimore, April 2–7, pp. 146–155.

Shepard, D. (1968). "A Two-Dimensional Interpolation Function for Irregularly Spaced Data." *Proceedings, Twenty-third National Conference, ACM,* pp. 517–524.

Sutherland, I. E., R. F. Sproull, and R. A. Scumacker (1974). "A Characterization of Ten Hidden-Surface Algorithms." *Computing Surveys,* vol. 6, no. 1, pp 1–55.

Tobler, W. R. (1970). *Selected Computer Programs*. Michigan Geographical Publications, Department of Geography, University of Michigan, Ann Arbor.