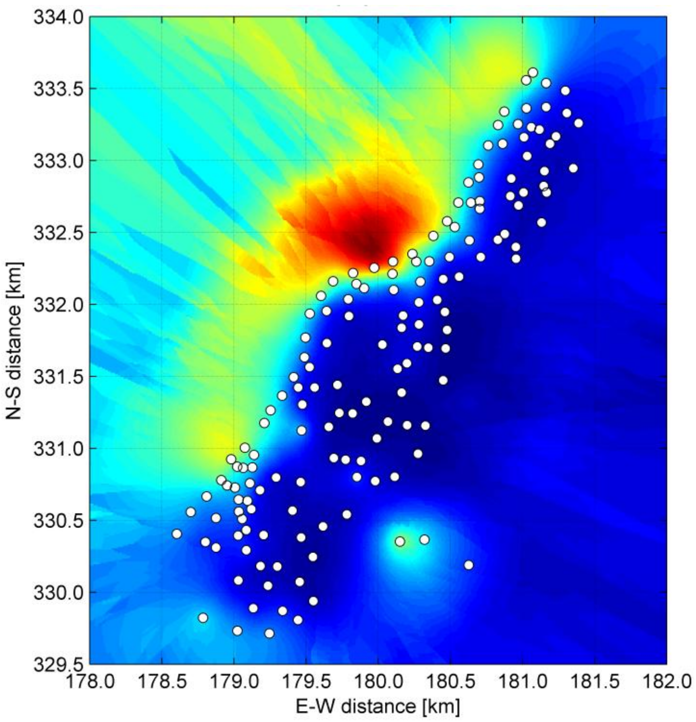
Interpolation and Kriging for Earth Scientists

self-tuition module



UNIVERSITEIT VAN AMSTERDAM

Author: Jurriaan H. Spaaks
Date: November 2006
Credits: 10 EC

1 Introduction

1.1 Course content

Earth scientists often deal with problems for which only limited sampling of the process under consideration can be executed due to limited time and resources; from this limited set of observations, a spatially continuous estimation must be inferred. This requires a scheme to interpolate the observations. Many methods are available, all of them having their own (dis)advantages. This self-tuition module covers the basic theory of the most common interpolation methods, both deterministic and (geo)statistical. Special attention is paid to the statistical method of kriging. In addition to the theoretical background of both the deterministic and statistical interpolation methods, a number of exercises have been included for practical application of kriging concepts using a selection of m-files from the *mGstat* toolbox constructed by T.M. Hansen, slightly modified for educational purposes.

1.2 Intended audience

It is assumed that the student has some background in concepts from univariate statistics, such as (auto)covariance, and probability distributions. Moreover, some of the exercises in this course require that the student is familiar with the programming language MATLAB. If this is not the case, an introductory course is available. It is strongly recommended to have this document at your disposal while making the MATLAB exercises.

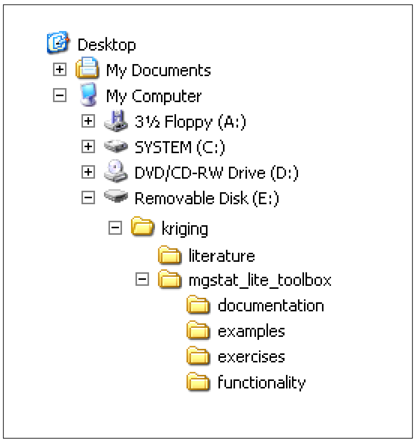
1.3 Digital resources

The necessary files for this course have been compressed into the *.zip-file 'environmetrics_kriging.zip'. You must extract the contents of this file to your network drive or USB storage device. After successful extraction, a directory 'kriging' will be present on your disk, containing 2 folders (see Figure 1) and a digital version of this document. The first folder ('literature') contains digital copies of Chapter 5 of *Meijerink et al.* [1994], Chapter 6 from *Burrough and McDonnell* [1998], *De Gruijter et al.* [2005], and Chapter 5.3 of *Mathews and Fink* [1992], based on which this course has been compiled. These files have been included for your convenience and future reference.

The second folder ('mgstat_lite_toolbox') has 4 subfolders containing:

'documentation'	: help documentation of the toolbox
'examples'	: exemplary scripts and functions
'exercises'	: information necessary to complete the exercises
'functionality'	: Gstat executable and m-files constituting the toolbox

Please note that changing the folder structure or names is not recommended.



```
Desktop
  ⊞ My Documents
  ⊟ My Computer
      ⊞ 3½ Floppy (A:)
      ⊞ SYSTEM (C:)
      ⊞ DVD/CD-RW Drive (D:)
      ⊟ Removable Disk (E:)
          ⊟ kriging
                literature
              ⊟ mgstat_lite_toolbox
                    documentation
                    examples
                    exercises
                    functionality
```

Figure 1. Folder structure.

2 Deterministic and statistical interpolation

The principle of interpolation is that points close to one another are more likely to be similar than points farther away. For example, in Figure 2 we have the observed values of variable y at locations x_a and x_b. The estimated value of y at location x_0 will be more like y_a than y_b because of the relative proximity of x_0 to x_a. Interpolation is all about weighing how the estimated value of y depends on the value of nearby observation points.

Interpolation methods can be divided into 2 groups based on how the observations are used to estimate values of the variable under consideration in between observation locations:

(1) deterministic or non-stochastic schemes;

(2) statistical or stochastic schemes.

In ***deterministic interpolation*** (see section 2.1), a mathematical function is fitted through the observed data. Deterministic methods are associated with so-called ***exact*** or ***pure*** interpolation schemes, meaning that the fitted function accurately describes the observations, as depicted in Figure 2. This implies that deterministic methods treat the observed value as being the true value of the variable of interest, in other words, the observations are assumed to be error-free. Variations in the observed values are thus fully attributed to the deterministic process, as opposed to a stochastic process. ***Polynomials*** (such as $y = ax^2 + bx + c$) are a practical choice of interpolation method, because they are straightforward to interpret, and their derivatives and integrals can be determined more easily compared to other methods, such as Fourier or Taylor series.

In contrast to deterministic methods of interpolation, ***statistical interpolation*** methods (see section 2.2) treat observations as being uncertain, meaning that the value of each observation can be thought of as being drawn from that variable's probability distribution. Many types of probability distribution are available, but the most frequently applied method is the Gaussian distribution.

2.1 Deterministic interpolation: basic concepts of polynomial fitting

2.1.1 *Exact interpolation using a 1st-order polynomial (linear interpolation)*

Perhaps the most widely used form of exact interpolation using polynomials is ***linear interpolation***, more formally referred to as piecewise 1st-order polynomial interpolation. Thanks to its speed and simplicity, this method is frequently used in explorative data analysis or when little additional information is available about the behavior of the variable in between reference points (i.e. observations).

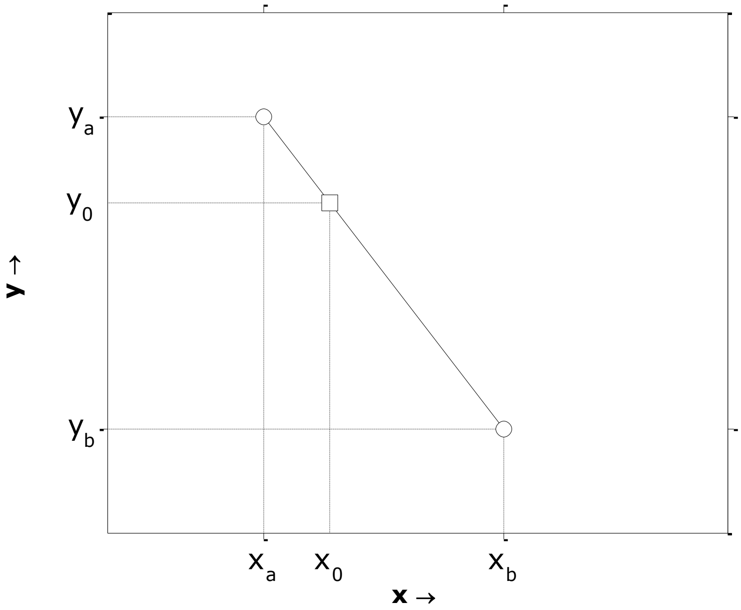


Figure 2. Univariate example of linear interpolation. Circles indicate measured value of y_i at location x_i . Square represents interpolated value of y at x_0 calculated with the fitted 1st-order polynomial on the interval [x_a , x_b] represented by the black solid line.

For example, if variable y has been recorded at locations x_a and x_b, as depicted in Figure 2, the value of y at an intermediate point x_0 can be estimated by first calculating the change of y with increasing x:

$$\alpha = \frac{y_b - y_a}{x_b - x_a} \qquad [1]$$

Multiplying α with the actual distance from x_a to x_0 thus represents the change in y over the distance x_a to x_0; y_0 is then simply y_a corrected by the change in y:

$$y_0 = y_a + \alpha \cdot (x_0 - x_a) \qquad [2]$$

If all observed values are located along a line (for the univariate case) or a plane (for the bivariate case), all observations are represented accurately by the fitted linear polynomial and the interpolation is exact. In most cases, however, the observations will vary more

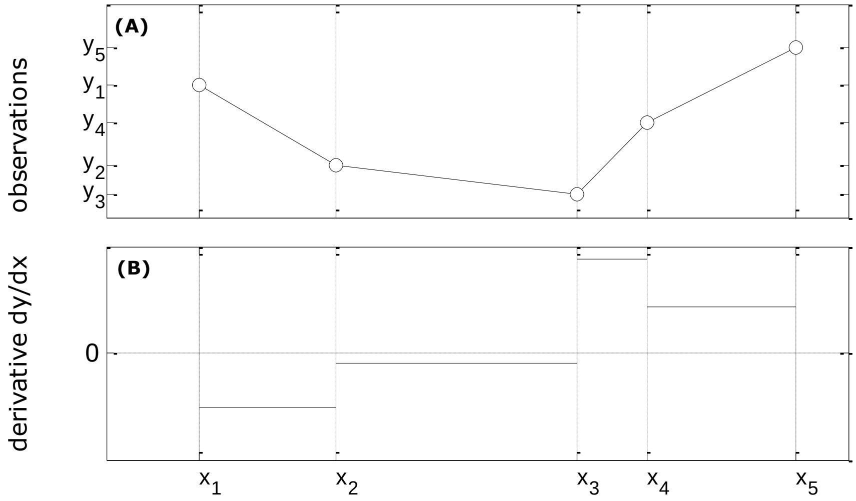


Figure 3. (A) Observations of y_i at locations x_i (circles), interpolated by a piecewise 1st-order polynomial (solid black line). **(B)** Derivative of (A), showing discontinuity at the joints.

irregularly and exact linear interpolation of the reference points can only be accomplished using a ***piecewise*** 1st-order polynomial, meaning that for each interval between 2 adjacent reference points, different coefficients need to be calculated (see Figure 3). The connection between interpolated lines (known as the ***joint*** or ***knot***) is then located precisely at the reference point. Consequentially, the first derivative of y to x must be discontinuous (see Figure 3B). When all measurements are used to establish the interpolant, the fitting procedure is said to be ***global***, as opposed to ***local*** schemes that use merely a subset of the observations.

2.1.2 Exact interpolation using a 2nd-order polynomial (quadratic interpolation)

As an alternative to using a piecewise 1st-order polynomial, higher-order polynomials may be applied, the general equation of which is:

$$f(x) = \sum_{i=0}^{n} c_i x^i \qquad [3]$$

with n representing the order of the equation and c_i the polynomial coefficient of x^i. The advantage of applying higher-order polynomials is that they allow for an increasing number of slope reversals: a polynomial of order n can have n-1 slope reversals at maximum (see Figure 4). In order to achieve an exact fit for any polynomial of order n, n+1 non-redundant observations are needed. Figure 5 depicts the observations that were previously interpolated using a linear interpolation scheme, but this time they have been interpolated by a piecewise 2nd-order method, each 'piece' using 3 reference points to establish an exact fit♣.

♣ See section 7.2 in the Appendix for an example of the derivation of the coefficients of a 2nd-order polynomial using three observation points [x , y].

The application of the higher-order polynomial clearly yields a smoother fit on the intervals [x_1,x_3] and [x_3,x_5], but the discontinuity in the first derivative at x_3 remains. However, the number of discontinuities in the first derivative is lower than with linear interpolation, resulting in a smoother appearance of the interpolation. Whether we choose a quadratic or linear interpolation method depends upon the purpose of the interpolation and upon the (assumed) nature of the variation underlying the observations.

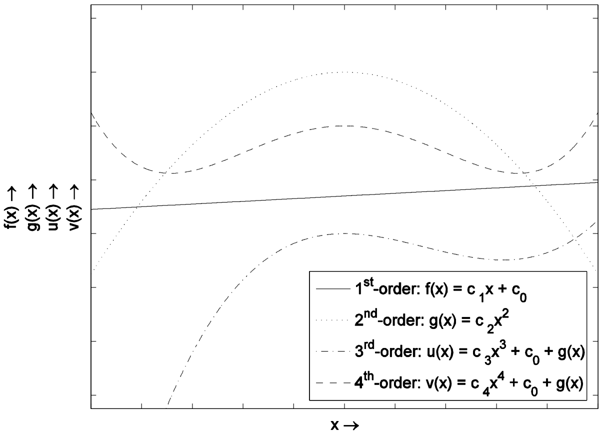


Figure 4. Polynomials *f(x)*, *g(x)*, *u(x)*, and *v(x)* of 1st, 2nd, 3rd, and 4th order, respectively. Subscripts of coefficients in legend refer to the order of the corresponding x_i (see equation [3] above); note that c_i may have different values for the different polynomials.

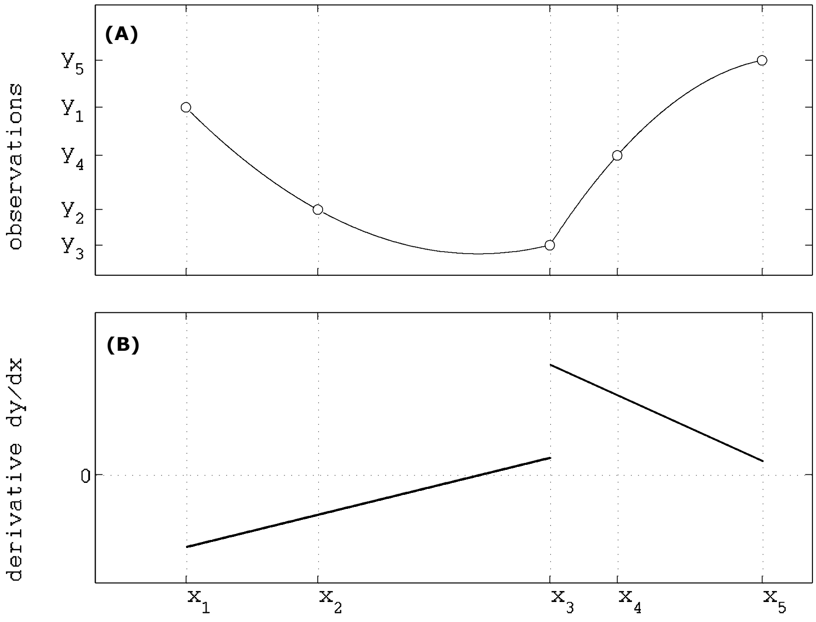


Figure 5. **(A)** Observations of y_i at locations x_i (circles), interpolated by a piecewise 2nd-order polynomial (solid black line). **(B)** Derivative of (A), showing discontinuity at the joint located at x_3; total number of joints lower than with piecewise linear interpolation.

Exercise 1: Discontinuities do also occur in the landscape. Placing your observation points thoughtfully can greatly increase interpolation accuracy. Figure 6 displays two types of landscape, each with its own topographical characteristics.

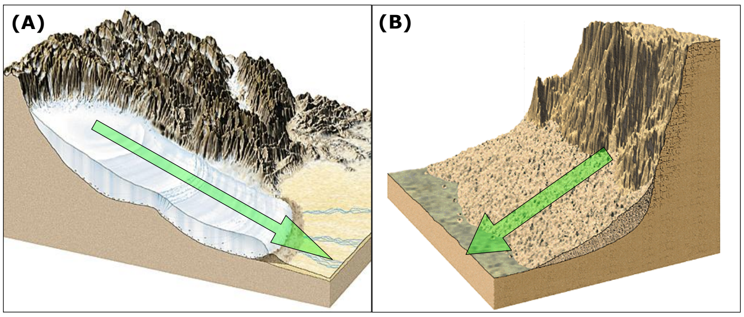


Figure 6. **(A)** Glacier with end moraine; **(B)** talus slope. Pictures from: *Lutgens and Tarbuck* [2003].

a) Given that you can only perform 7 observations of elevation per transect, indicate whether you would rather use a global linear, piecewise linear, or piecewise quadratic method for interpolating the topography of these two geomorphologic features. The transects are positioned parallel to the direction of steepest descent, as indicated by the arrows.

b) Where would you place the observations?

Exercise 2: Figure 7 shows water levels recorded by a tidal gauge on Mustang Island, Texas.

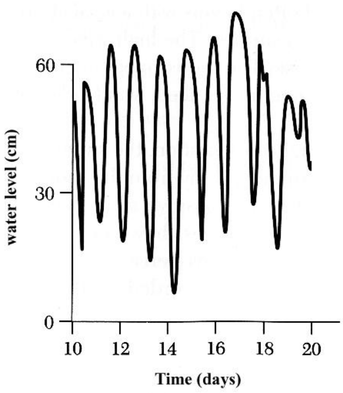


Figure 7. Tidal gauge record from Mustang Island, Texas. Picture from *Davis Jr* [1994].

a) Assuming that the wavelength λ equals 24 hours, indicate which of the following subsampling schemes A through C would yield the most accurate results when interpolated using a piecewise first-order polynomial:

(A) random sampling of 120 samples/30 days

(B) random sampling of 4 samples per λ

(C) random sampling of 2 samples per 0.5 λ

b) Explain the choice made under a)

c) What is the difference between (B) and (C)?

2.1.3 *Exact interpolation using a 3rd-order polynomial (cubic interpolation)*

With increasing order of the piecewise polynomial, each piece consists of increasingly more points, thus resulting in a smaller number of piecewise polynomials needed for interpolating the entire dataset, which in turn reduces the number of discontinuities.

Exercise 3: Why are discontinuities in the first order derivative generally conceived of as "bad"?

A useful method for smooth interpolation is the so-called ***cubic spline***[♣]. Splines are actually a subclass of cubic interpolation methods, because of their additional conditions of having a continuous first-order and second-order derivative. The continuity of the first-order derivative guarantees a smooth fit; the continuity of the second-order derivative is necessary to define the so-called ***radius of curvature*** at each point which is useful in e.g. turbulent flow analysis and defining curvature of objects. The 4 points needed to establish an exact fit for a 3rd-order polynomial divide the range of the measurements into 3 intervals (also known as ***spans***). However, as we have seen earlier with linear and quadratic polynomials, discontinuities at the joints will remain if we are simply connecting each cubic polynomial to the next (see Figure 8A). We can improve the smoothness by using only the middle span of each cubic polynomial, such as depicted in Figure 8B, but still, the first (and second) derivatives of (B) can be discontinuous (see Figure 8C).

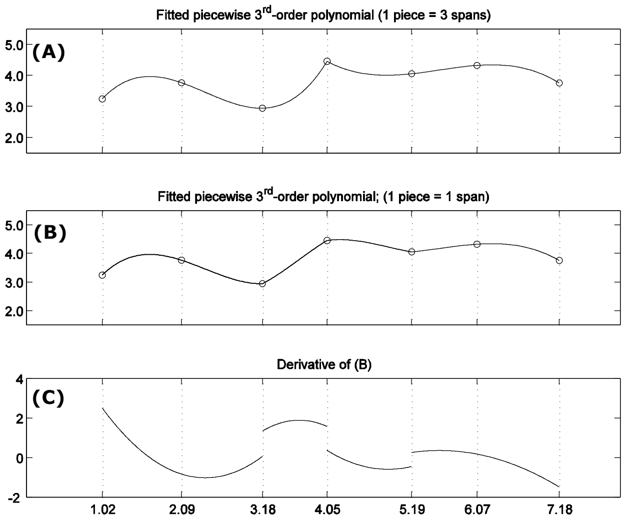


Figure 8. **(A)** Piecewise 3rd-order polynomial, each piece consisting of 3 spans; **(B)** Piecewise 3rd-order polynomial constructed by linking up the middle spans of consecutive 3rd-order polynomials. **(C)** Derivative of (B), showing that discontinuities remain.

[♣] The word spline comes from its physical counterpart, i.e. a flexible piece of wood which has long been used for drawing smooth, aerodynamic curves in the design of airplanes before computer simulation took over.

In order to create a piecewise cubic interpolation *S(x)* that yields continuous functions upon derivation, we must apply some inverse reasoning: instead of focusing immediately on piecewise 3rd–order polynomials, we should start by constructing piecewise linear functions that are in fact the second derivatives of the cubic spline function we are looking for. Provided there is only one value of *S''(x)* at every location *x*, such a piecewise linear function must be continuous and can be described by the so-called Lagrange linear interpolation method (refer to Figure 9 for the meaning of the variables):

$$S''_k = S''(x_k)\,\frac{x - x_{k+1}}{x_k - x_{k+1}} + S''(x_{k+1})\,\frac{x - x_k}{x_{k+1} - x_k} \qquad [4]$$

(Note that this equation is valid over multiple spans, which makes it different from Equation [2]).

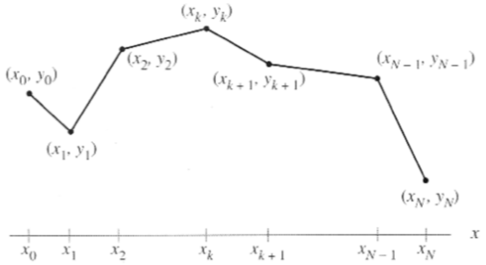


Figure 9. Definition of variables used in Equation [4]. Figure from *Mathews and Fink* [1992].

Having established this continuous, piecewise first order function, we can infer the piecewise cubic spline equation by integrating the Lagrange equation twice. This introduces two integration constants that must be assigned a value later on, in order to describe the precise appearance of our cubic spline curve. Figure 10 further illustrates this concept.

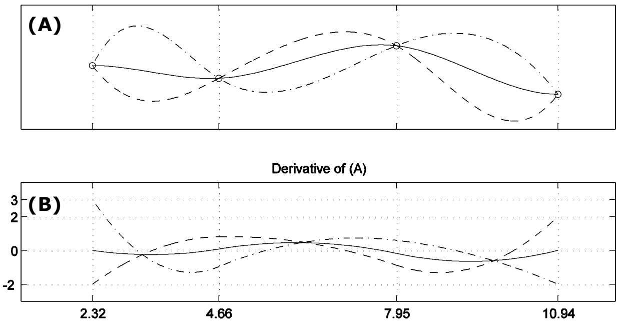


Figure 10. **(A)** Different cubic splines fitted through the observations; solid line represents a cubic spline with natural end-point constraints, dashed line represents a cubic spline with clamped end-point constraints: $S'(x_0)$ equal to -2 and $S'(x_N)$ equal to +2; dash-dotted line represents a cubic spline with clamped end-point constraints: $S'(x_0)$ equal to +3 and $S'(x_N)$ equal to -2. **(B)** First-order derivatives of (A).

The value of the integration constants may be calculated from the desired value of $S'(x)$ or $S''(x)$ at the extremes of the cubic spline curve (i.e. locations x_0 and x_N). Because of their location and their necessity for pinning down the exact shape of the cubic curve, these additional conditions are usually referred to as ***end-point constraints***. Some of the most common end-point constraints are the ***clamped*** (user-specified $S'(x)$ at $x = x_0$ and $x = x_N$) and ***natural*** ($S'(x)$ equal to zero at $x = x_0$ and $x = x_N$) methods, although a number of other methods exists. These methods differ in how they impose a value on the first or second derivatives of the piecewise cubic spline at the end-points. A more elaborate description can be found in Section 5.3 "*Interpolation by spline functions*" of *Mathews and Fink* [1992] that has been included in this module's 'literature' folder.

Exercise 4: Tipping buckets are used to measure precipitation. In contrast to normal rain gauges that record the volume of rain collected at preset time intervals, tipping buckets record the time needed to collect a preset volume of rain, after which the collected volume is discarded by a tipping motion of the bucket, hence the name.

a) Interpolation of precipitation data using higher-order polynomials is more sensitive to oscillation if the data were collected using tipping buckets compared to normal rain gauges. Why?

b) Under what circumstances do the largest differences between actual and interpolated data occur?

Exercise 5: Given the nature of variability of certain variables, some interpolation methods are better than others. For example, groundwater levels in the Dutch dunes vary more or less smoothly.

a) Would you rather use piecewise linear or cubic spline interpolation to estimate groundwater levels at locations in between wells?

b) Why?

2.1.4 Exact interpolation using a zero-order polynomial

A special case of n^{th}-order polynomial interpolation is ***nearest-neighbor interpolation***. This method actually uses a zero-order polynomial to interpolate the observations. As an example, consider the time series depicted in Figure 11.

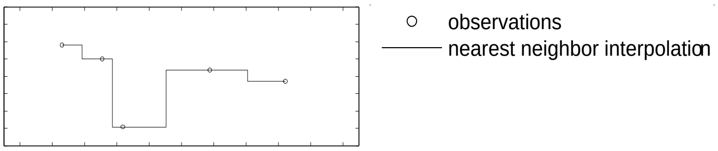


Figure 11. 1D example of nearest-neighbor interpolation.

Here, the interpolation is established by assigning the value of an observation point $z(s_i)$ to all possible points $z(s_0)$ that are closer to s_i than to any other observation point $z(s_j)$:

$$z(s_0) = z(s_i) \qquad \{ d_{0,i} < d_{0,j} \} \tag{5}$$

With $d_{0,i}$ the distance between s_0 and s_i, and $d_{0,j}$ the distance between s_0 and s_j. In 1D, such as the time series above, only the two neighboring observations need to be considered, but if the observation locations are specified in 2D, such as in a map, the problem of choosing which observations should be assessed, is less straightforward.

Given a set of points in 2D space, we can determine the nearest neighbors of a point (a in Figure 12) by application of a neat mathematical method known as ***Delaunay triangulation***. With this method, two of the other observation points within the set must be chosen (for example, points b and c). A circle can be drawn connecting points b and c with the point of interest a (see Figure 12A). If the circle contains no other points, b and c are so-called ***natural neighbors*** of point a. However, point d is present in the circle connecting points a, b, and c, meaning that the circle must be discarded and replaced with a smaller circle connecting point a, b, and d (see Figure 12B). This circle does not contain any other observation points, therefore d (and b) is a natural neighbor and the triangle abd is called a ***Delaunay triangle***. In order to construct the nearest neighbor interpolation later on, we must store the location of the center of the circle connecting a, b, and d (solid square in Figure 12B), or in fact of all circles connecting the cornerpoints of Delaunay triangles. After all Delaunay triangles have been determined, connecting the circle midpoints by a straight line (see Figure 12G) yields a so-called ***Thiessen polygon***, also termed ***Dirichlet*** or ***Voronoi*** polygon. For each point within the Thiessen polygon, Equation [5] is valid.

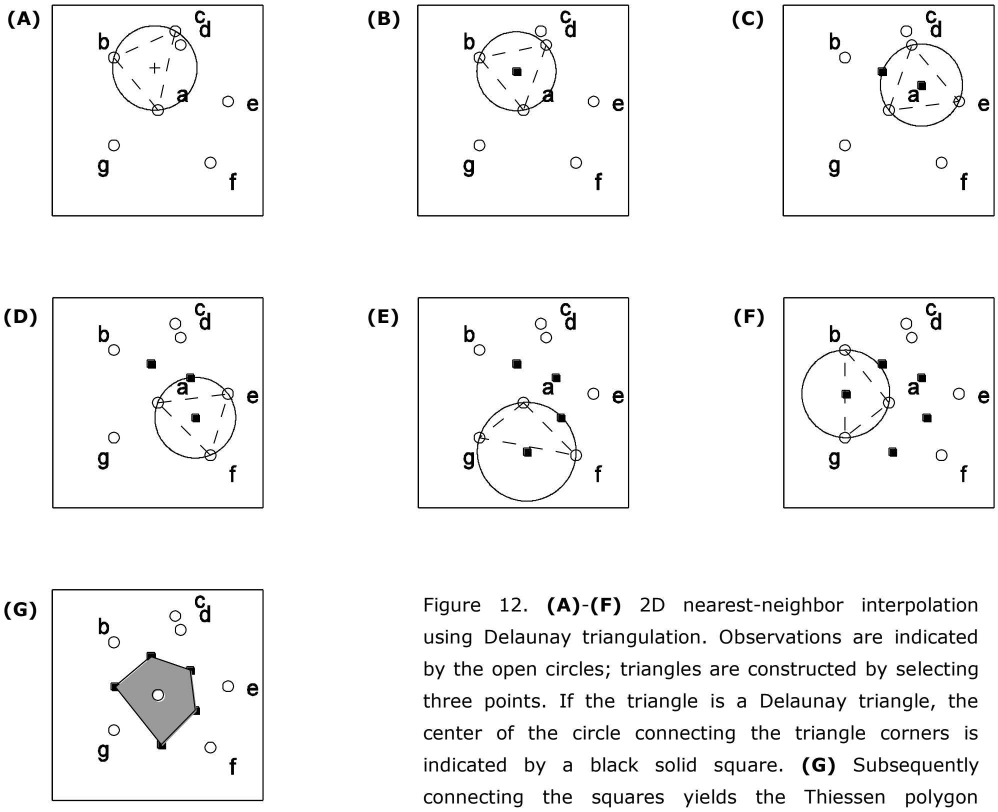


Figure 12. **(A)**-**(F)** 2D nearest-neighbor interpolation using Delaunay triangulation. Observations are indicated by the open circles; triangles are constructed by selecting three points. If the triangle is a Delaunay triangle, the center of the circle connecting the triangle corners is indicated by a black solid square. **(G)** Subsequently connecting the squares yields the Thiessen polygon (shaded area).

2.2 Statistical interpolation: uncertain observations

From the previous section, it is clear that establishing an exact global fit on a large number of data points usually requires a polynomial of high order. The calculation of its coefficients quickly becomes a burdensome task with increasing variation, as may be seen from the example in the Appendix, which is merely the calculation of the three parameters of a 2nd-order polynomial. However, even more important is the belief we have in our data representing the true value of the variable.

Although deterministic methods covered in the previous section can provide an estimate of the variable of interest at unobserved locations, there are a number of drawbacks besides choosing the order of the interpolant.

For example, if the data in Figure 13 is assumed to be error-free, a polynomial of order 15 can provide an exact interpolation of the data. However, we may be inclined to choose lower-order polynomials because a polynomial of order 15 seems unlikely; part of the variation in our data is apparently caused by error. An interpolation method that is frequently applied to smooth out this error, is the so-called ***moving average***. In this approach, all observations within a preset distance d_0 away from the location of the interpolation are averaged to obtain the interpolated value.

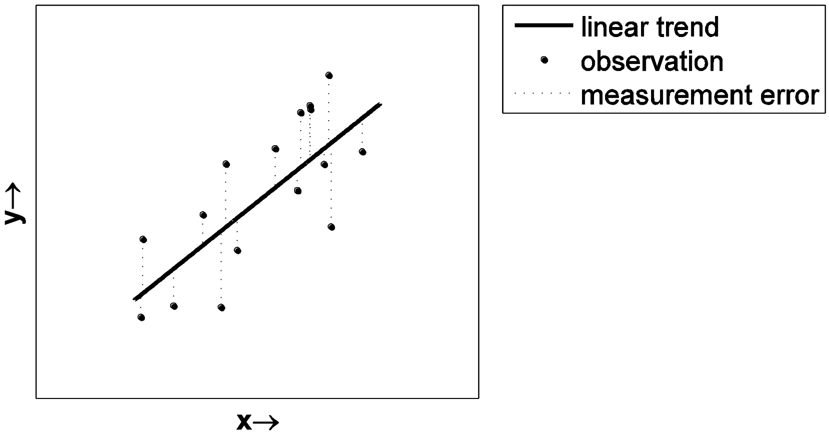


Figure 13. Example of observational error complicating the deterministic fitting process.

In the simplest case, all observations within the search window ($2d_0$) are assigned equal weights, but more intricate weighting schemes are also used, such as assigning increasingly smaller weights to observations that are farther away from the interpolation location. Because of the averaging involved, this method does not oscillate like n^{th}-order polynomials would; the larger the search window, the smoother the interpolation will be.

In summary, the methods discussed so far do not:

(1) provide direct estimates of the uncertainty of the interpolation

(2) provide *a priori* information on whether the size of the search window is appropriate, or whether the best values have been chosen in weighing the observations.

These disadvantages are avoided with statistical interpolation, more commonly referred to as ***kriging***, the methodology of which is covered in the next chapter.

3 Kriging

The application of geostatistical methods for interpolating observations starts with the recognition that the spatial variation of any continuous attribute is often too irregular to be modeled by a smooth polynomial. Instead, the observed variation in the measurements as well as the hypothesized variation at unobserved locations can be considered as being the result of 3 components, each having its own characteristics such as mean, probability distribution, and spatial correlation. In mathematical terms, this can be written as:

$$Z(s) = m(s) + \varepsilon'(s) + \varepsilon''(s) \qquad [6]$$

where s denotes the position (not necessarily restricted to one dimension only) of the ***prediction*** (i.e., the interpolation), $Z(s)$ is the prediction at location s, $m(s)$ represents a spatially varying, deterministic component describing the structural variation of $Z(s)$, $\varepsilon'(s)$ represents a stochastic, but spatially dependent term known as the ***regionalized variable***, and finally, a spatially uncorrelated residual term $\varepsilon''(s)$, that varies randomly according to the Gaussian probability distribution. Please note that truly random variation is absent from the earth sciences; the variation in $\varepsilon'(s)$ and $\varepsilon''(s)$ is actually the net result of structural variation for which $m(s)$ does not account (such as small-scale variation), and error introduced by the measurement itself. These processes are in fact not stochastic at all, but they vary in a complex way, making them difficult to explain physically, therefore they are treated in statistic terms.

In the simplest case, where we may assume that no trend (or ***drift***) is present, the expected difference between Z at location s and Z at a distance h (called the ***lag***) away from location s must equal zero:

$$U = Z(s) - Z(s+h) \qquad [7]$$

$$E(U) = 0 \qquad [8]$$

If equation [8] is valid, the dataset is said to exhibit ***stationarity of difference***. Further assuming that the variance of U depends only on the distance between sites h yields:

$$E(U^2) = E\left(\left\{\varepsilon'(s) - \varepsilon'(s+h)\right\}^2\right) = 2\gamma(h) \qquad [9]$$

where γ (pronounced 'gamma') is called the semivariance. From the above, it may be inferred that once structural effects have been accounted for, the variance of U does not depend on absolute location but rather on h only. This is known as ***stationarity of variance***.

Exercise 6: Why does equation [8] apply under conditions where no trend is present?

Exercise 7: Why does $\varepsilon''(s)$ not occur in equation [9]?

If the stationarity conditions are fulfilled, the semivariance can be estimated from sample data:

$$\hat{\gamma}(h) = \frac{1}{2n}\sum_{i=1}^{n}\left\{z(s_i) - z(s_i + h)\right\}^2 \qquad [10]$$

where n is the number of pairs of observations and $z(s_i)$ the observed value of Z at location s_i. Please note the accent circonflex above γ to indicate that it is actually an estimate determined from observed values.

3.1 The semivariogram and its properties

A plot of γ vs. h is referred to as the ***semivariogram*** or, more informally, ***variogram***. If the relation of γ with h has been determined based on observations, it is called the ***experimental*** or ***empirical*** semivariogram. Figure 14 visualizes a typical experimental semivariogram, along with a few important features. This semivariogram has been constructed by calculating the average of all semivariances in each of the 12 so-called 'bins'. The labels below the '+'-signs indicate the number of paired observations in a particular bin. A line has been fitted through the observations to provide a continuous function of γ. It may be seen that at large values of h, the semivariance becomes constant. This horizontal part is known as the ***sill***; the value of h at which γ reaches the sill is known as the ***range*** (commonly denoted by a). Please note that the fitted line does not go through the origin [0,0]; the intercept of the line is known as the ***nugget***.

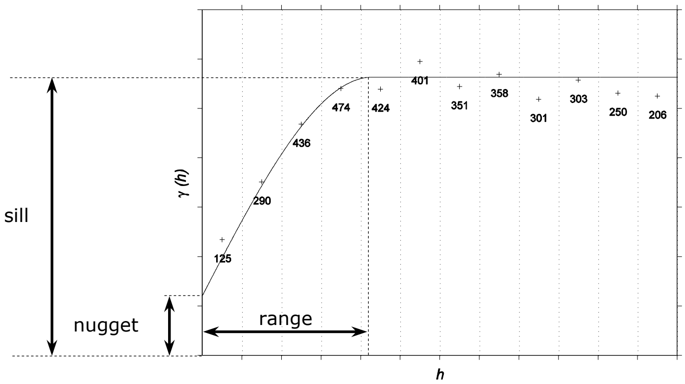


Figure 14. Typical semivariogram, with 12 lag bins. Solid line represents fitted nugget + spherical model. Labels indicate the number of pairs of observations in a bin.

Exercise 8: The nugget value is the sum of variation from 2 sources, the first of which is deterministic variation at a scale smaller than the measurement interval. What is the second source? (Hint: what is the difference between deterministic and statistical interpolation?)

Exercise 9: Why do we need to divide by $2n$ in equation [10]?

3.1.1 Choosing the appropriate semivariogram model

Because the semivariogram is based on a limited number of observations, it is a discrete function, meaning that γ has been calculated only at values of h present in the data set. To make spatially continuous estimates, however, a continuous semivariance function $\gamma(h)$ is needed. Therefore, we need to fit a semivariogram model through the calculated points $[h,\gamma]$. Only models are permitted which ensure that the estimates of the prediction variance are non-negative. Frequently applied models that do not violate this condition are the spherical, exponential, Gaussian, nugget, and linear models. Spatial predictions of Z at location s, made with the semivariogram model, are so-called ***Best Linear Unbiased Estimations*** (**BLUE**; or **BLUP**, 'P' for 'predictions') of Z at location s.

Below is a brief discussion of the most common semivariogram models along with their equations.

3.1.2 Semivariogram models: spherical model

The ***spherical*** model is used when there is a clear range and sill. The presence of a clear transition point at $h = a$ is indicative of having one dominant process. The spherical model is calculated according to:

$$\gamma(h) = \begin{cases} c \cdot \left[\frac{3}{2}\left(\frac{h}{a}\right) - \frac{1}{2}\left(\frac{h}{a}\right)^3 \right] & \text{if } h < a \\ c & \text{if } h \geq a \end{cases} \qquad [11]$$

Where parameter c may be considered as the semivariance model 'amplitude'. See Figure 16A.

3.1.3 Semivariogram models: exponential model

In contrast to the spherical model, the ***exponential*** model can be indicative of several processes interfering. It is used when there is only a gradual approach to the distinct sill. It is modeled according to:

$$\gamma(h) = c \cdot \left[1 - \exp(-h/a) \right] \qquad [12]$$

Of course, parameter a can only be interpreted as being *proportional* to the range, since the equation is asymptotic. See Figure 16B.

3.1.4 *Semivariogram models: Gaussian model*

If the spatial field is hypothesized to vary smoothly, the ***Gaussian*** semivariance model is often the best choice:

$$\gamma(h) = c \cdot \left[1 - \exp(-[h/a]^2) \right] \qquad [13]$$

As with the exponential model, parameter a can only be interpreted as being *proportional* to the range, since the equation is asymptotic. See Figure 16C.

3.1.5 *Semivariogram models: nugget model*

As may be observed from Figure 14, the best fit is accomplished by combining one of the above models with a ***nugget*** model (see Figure 16D):

$$\gamma(h) = \begin{cases} 0 & \text{if } h = 0 \\ c & \text{if } h < a \end{cases} \qquad [14]$$

3.1.6 *Semivariogram models: linear model*

When the data varies at all scales, no sill may be reached at the observed scale. In such cases, the ***linear*** semivariogram model may be used; they are by definition ***non-transitive***, meaning that they do not approach or reach a sill within the sampled area. Linear models are calculated according to equation [15]. Note the absence of the intercept, because that is modeled separately by combining the linear model with the nugget model [14].

$$\gamma(h) = \alpha \cdot h \qquad [15]$$

For the linear model, the model coefficient α represents the slope of the line per unit of h.

Exercise 10: Indicate which model would be best suited to represent the semivariogram of:

a) the topography of a peneplain landscape

b) a random walk time series (as an example of random walk, consider a time series that is constructed by randomly choosing to go up or down relative to the last position)

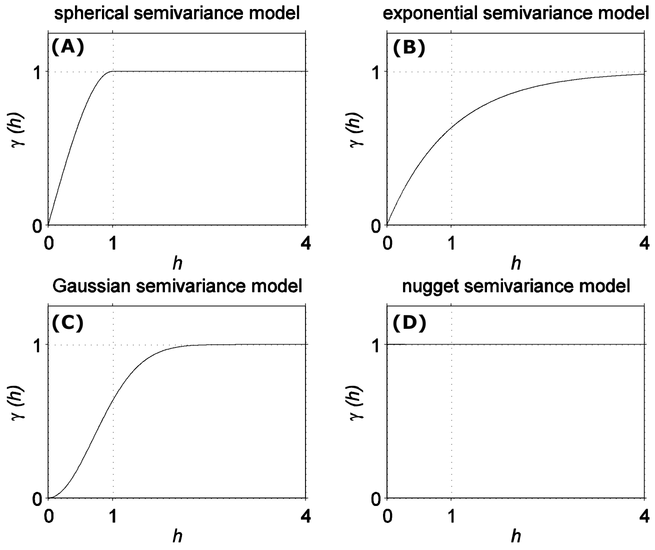


Figure 16. Transitive semivariogram models; both *a* and *c* have been set to 1.

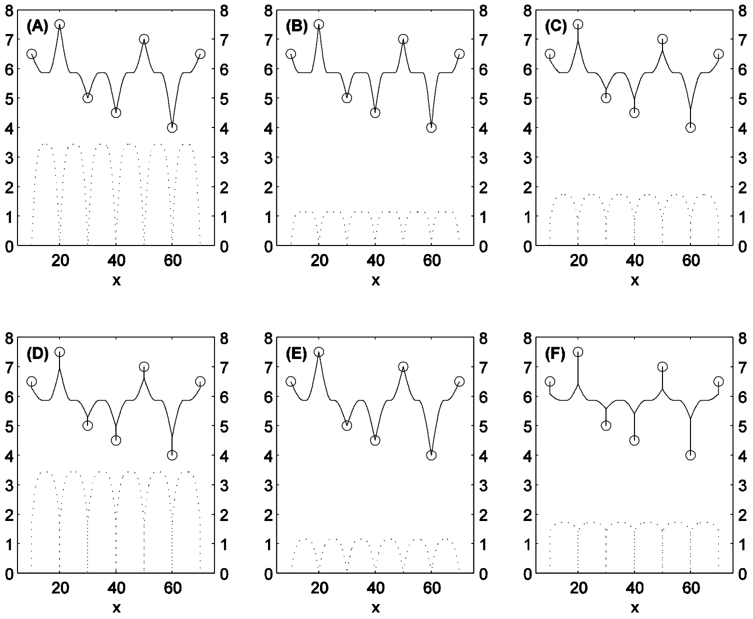


Figure 16. 1-D ordinary kriging examples; spherical model. Solid lines represent predictions, dashed lines represent prediction variances.

Exercise 11: The plots of Figure 16 all represent the same measurements (indicated by the circles) that have been interpolated using a spherical semivariogram model. However, in each plot, the semivariogram model has different parameters. For each combination of plots A-D, B-E, and C-F, indicate which parameter is responsible for the difference in appearance between the two plots that are being compared. Note: C-F differ in two parameters!

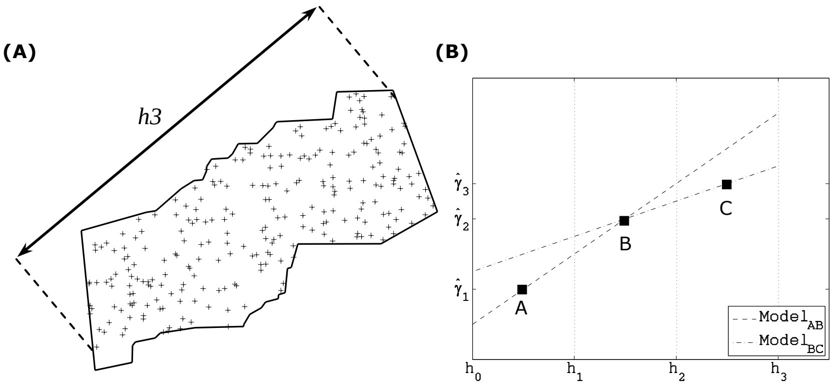


Figure 17. Study area and semivariogram with two models.

Exercise 12: Figure 17A shows a study area in which 250 measurements have been performed of a variable X at the locations indicated by the '+'-signs. Figure 17B visualizes an experimental variogram with 3 bins $[h_0, h_1)$, $[h_1, h_2)$, and $[h_2, h_3]$. The dots represent average semivariance values for each of the bins. A linear model (Model_{AB}) has been fitted through points A and B. In addition to Model_{AB}, Model_{BC} has been fitted through points B and C.

a) Should a linear model based on all 3 points be more like Model_{AB} or Model_{BC}?

b) Why?

Exercise 13: At which value(s) of h will a semivariogram of the periodical data presented in Figure 7 reach its maximum?

3.2 Ordinary kriging

3.2.1 Theory

Although many kriging methods exist, ***ordinary kriging*** is perhaps the most frequently applied method. From the ordinary kriging theory outlined in this section and the example (see section 3.2.2), other methods such as simple kriging, co-kriging and block kriging may be derived.

First of all, it can be shown that interpolation methods vary only in their calculation of the weights λ (pronounced 'lambda'), and that the general form of any interpolation equation is:

$$\hat{z}(s_0) = \sum_{i=1}^{n} \lambda_i \cdot z(s_i) \tag{16}$$

with:

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{17}$$

where n represents the number of observations of z at locations s_i (which may have more than one dimension) that are used to estimate the value of z at location s_0. The procedure for calculating the weights is similar to that used in moving average interpolation, except that with the latter, we *impose* a (possibly inappropriate) inverse distance model to calculate the weights from the lag. With kriging, the weights are *derived* from the observed data by means of a geostatistical analysis. The weights λ_i must be chosen in such a way that the prediction of z at location s_0 is unbiased, and that the estimation variance $\hat{\sigma}_p^2$ of the prediction is less than for any other combination of λ_i. In mathematical terms, the above is equivalent to finding the minimum of:

$$\hat{\sigma}_p^2 = \sum_{i=1}^{n} \lambda_i \cdot \gamma(s_i, s_0) + \phi \tag{18}$$

which is obtained when[♣]:

$$\sum_{i=1}^{n} \lambda_i \cdot \gamma(s_i, s_j) + \phi = \gamma(s_j, s_0) \qquad \text{for } j = 1, \ldots, \ldots, \; j = n \tag{19}$$

where ϕ ('phi') is the so-called Lagrange multiplier that is necessary for the minimalization.

[♣] The derivation of the minimand function is beyond the scope of this text.

Defining $\mathbf{A}$ as the semivariance matrix between locations s_i and s_j, and vector $\mathbf{b}$ as the semivariance matrix between location s_j and the location where the spatial prediction is needed s_0 allows for rewriting Equation [19] using matrix notation:

$$\mathbf{A} = \gamma(s_i, s_j) \qquad \text{for } i = 1, \ldots, \ldots,\ i = n; \text{for } j = 1, \ldots, \ldots,\ j = n \tag{20}$$

$$\mathbf{b} = \gamma(s_j, s_0) \qquad \text{for } j = 1, \ldots, \ldots,\ j = n \tag{21}$$

$$\mathbf{A} \cdot \begin{bmatrix} \lambda \\ \phi \end{bmatrix} = \mathbf{b} \tag{22}$$

where λ represents the n interpolation weights. The semivariance can be calculated from the lag between points (be it observation-observation lag or observation-prediction lag) using the chosen semivariogram model. Recalling that we are attempting to calculate the vector of interpolation weights λ, we must rearrange [22] in order to bring $\mathbf{A}$ to the right:

$$\mathbf{A} \cdot \mathbf{A}^{-1} \cdot \begin{bmatrix} \lambda \\ \phi \end{bmatrix} = \mathbf{b} \cdot \mathbf{A}^{-1} \tag{23}$$

$$\begin{bmatrix} \lambda \\ \phi \end{bmatrix} = \mathbf{b} \cdot \mathbf{A}^{-1} \tag{24}$$

The right-hand side may now be solved relatively easily (see numerical example below).

3.2.2 *Ordinary kriging: 2-D example*

Let the spatial variation of variable z be modeled by a combined nugget-spherical semivariogram model with c_{nugget} = 2.1, $c_{spherical}$ = 6.3 and a = 7.0. We have n=4 observations s_i at locations [x_i , y_i]:

Table 1. Observations of z at locations [x_i , y_i].

i	x_i	y_i	$z(s_i)$
1	1.9186	1.0440	4
2	1.3365	7.1722	2
3	7.3299	2.9922	6
n	7.4003	5.8449	8

and we want to interpolate to x_0 = 5, y_0 = 5, such as depicted in Figure 18:

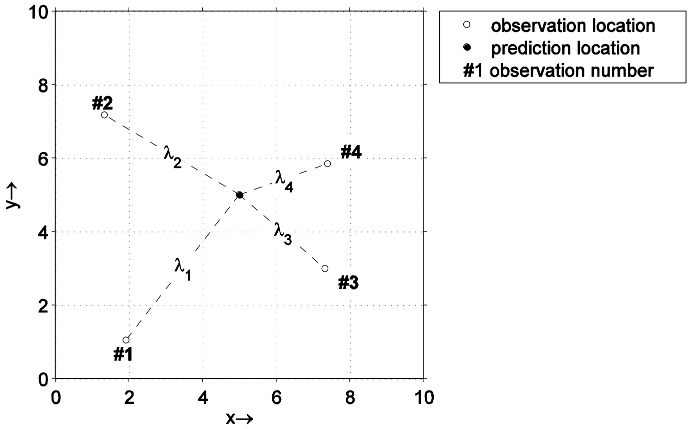


Figure 18. Locations of observed values $z(s_i)$ and interpolation location. λ_i represents the interpolation weight. See also Table 1.

The first step is to calculate the lag matrix between observation locations:

$$h(s_i, s_j) = \begin{bmatrix} 0 & 6.1558 & 5.7513 & 7.2868 \\ 6.1558 & 0 & 7.3071 & 6.2074 \\ 5.7513 & 7.3071 & 0 & 2.8536 \\ 7.2868 & 6.2074 & 2.8563 & 0 \end{bmatrix} \qquad [25]$$

From the semivariogram model, $\gamma(s_i, s_j)$ may be calculated:

$$\gamma(s_i, s_j) = \begin{bmatrix} 0 & 8.2681 & 8.1172 & 8.4000 \\ 8.2681 & 0 & 8.4000 & 8.2834 \\ 8.1172 & 8.4000 & 0 & 5.7389 \\ 8.4000 & 8.2834 & 5.7389 & 0 \end{bmatrix} \qquad [26]$$

A similar procedure must be followed to calculate the lag vector between the observed locations and the interpolation location:

$$h(s_j, s_0) = \begin{bmatrix} 5.0145 \\ 4.2591 \\ 3.0757 \\ 2.5447 \end{bmatrix} \qquad [27]$$

By substituting the observed lag values into the semivariogram model, we get:

$$\gamma(s_j, s_0) = \begin{bmatrix} 7.7116 \\ 7.1402 \\ 5.9849 \\ 5.3840 \end{bmatrix} \qquad [28]$$

To be able to rearrange equation [22] and solve for λ, we must apply a mathematical trick because the number of rows and columns on the left-hand side must be the same as that of the right-hand side. Adding a row and a column of ones to [26] and a single '1' to [28] solves the problem:

$$\mathbf{A} = \begin{bmatrix} 0 & 8.2681 & 8.1172 & 8.4000 & 1 \\ 8.2681 & 0 & 8.4000 & 8.2834 & 1 \\ 8.1172 & 8.4 & 0 & 5.7389 & 1 \\ 8.4000 & 8.2834 & 5.7389 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \qquad [29]$$

$$\mathbf{b} = \begin{bmatrix} 7.7116 \\ 7.1402 \\ 5.9849 \\ 5.3840 \\ 1 \end{bmatrix} \qquad [30]$$

Using a computer (see for example the MATLAB documentation of the function `inv`) to calculate the inverse of **A** yields:

$$\mathbf{A}^{-1} = \begin{bmatrix} -0.0871 & 0.0345 & 0.0288 & 0.0238 & 0.2809 \\ 0.0345 & -0.0860 & 0.0241 & 0.0275 & 0.2850 \\ 0.0288 & 0.0241 & -0.1140 & 0.0612 & 0.2130 \\ 0.0238 & 0.0275 & 0.0612 & -0.1125 & 0.2211 \\ 0.2809 & 0.2850 & 0.2130 & 0.2211 & -5.9429 \end{bmatrix} \qquad [31]$$

So that we can finally obtain the weights:

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \phi \end{bmatrix} = \begin{bmatrix} 0.1559 \\ 0.2286 \\ 0.2538 \\ 0.3616 \\ 0.7235 \end{bmatrix} \qquad [32]$$

The prediction may now be calculated according to:

$$\hat{z}(s_0) = \sum_{i=1}^{n} \{ \lambda_i \cdot z(s_i) \} = 5.4968 \qquad [33]$$

and the estimation variance as:

$$\sigma_p^2(s_0) = \mathbf{b}^{\mathrm{T}} \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \phi \end{bmatrix} = 7.0245 \qquad [34]$$

Exercise 14: You can answer part a) of this question without actually calculating.

a) Will the value of $z(s_0)$ with s_0 equal to x=10, y=10 be lower than, higher than, or equal to the value of observation #4?

b) What will the value be?

Exercise 15: Observation #2 has a value of 2 (see Table 1).

a) Indicate whether the value of Z at a distance 1/∞ away from observation #2 is

(A) approximately the same as the value of Z at the location of observation #2;

(B) higher than the value of Z at the location of observation #2;

(C) lower than the value of Z at the location of observation #2.

b) Why?

Exercise 16:

a) Analogous to the example, calculate $z(s_0)$ with s_0 equal to x=3, y=4.

b) Calculate the kriging variance.

3.2.3 Anisotropy

So far, we have constructed semivariograms by calculating γ as a function of separation distance h, thereby implicitly assuming that the data is ***isotropic***, meaning that the data do not exhibit any directional effects. However, if we have reason to believe that the appearance of the semivariogram is not only a function of lag, but also of direction β, we may construct an ***anisotropic*** semivariogram, in which paired observations are divided over a number of bins according to β as well as h: $\gamma = \mathrm{f}(h, \beta)$, i.e. the semivariogram is now bivariate. The directional dependency can be caused by the semivariogram model parameters being different in different directions, as could happen in observations of sand contents parallel and perpendicular to a straight section of a river (see Figure 19).

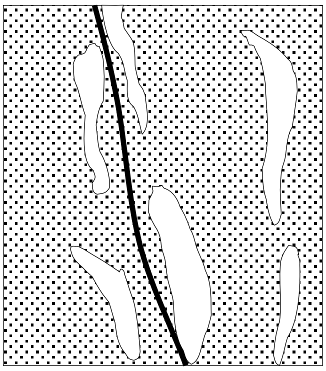


Figure 19. Geometric anisotropy in sediment composition parallel and perpendicular to the river (black solid line).

If the directionally dependent parameter changes with direction according to a continuous function, the anisotropy in the data is ***geometric***, whereas a discontinuous directional dependency function results from ***zonal*** anisotropic data.

Exercise 17: Indicate whether you would expect isotropy, zonal anisotropy, or geometric anisotropy in:

a) Water content in the plough layer of a field

b) Hydraulic conductivity of a jointed rock

3.3 Kriging using extra information

3.3.1 Stratified kriging

When there is enough soft information to classify the area into meaningful sub-areas and there are enough data points to compute semivariograms for each sub-area, the interpolation can be carried out for each sub-area separately. This procedure is known as ***stratified kriging***. The kriging equations may (optionally) be adjusted to avoid discontinuities at the class boundaries. Stratified kriging can reduce the kriging variance and can help to preserve narrow features such as river channels, that would otherwise be lost during the interpolation.

Exercise 18: Thoughtfully dividing the river valley area of Figure 20 before using anisotropic variograms would enable more accurate predictions relative to using just 1 anisotropic semivariogram. Indicate how you would divide the area in two parts.

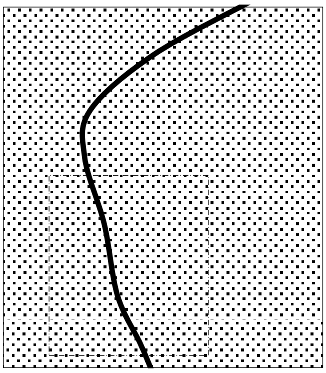


Figure 20. Larger view of the river valley; dash-dotted line represents the extent of Figure 19.

3.3.2 Co-kriging

If the spatial structure of the variable of interest (also called the ***target variable***) is thought to be geostatistically related to that of another variable (termed the ***co-variable***), it may be profitable to apply so-called ***co-kriging***. With this approach, only a limited number of observations is available for the target variable, whereas some other, related variable has been sampled much more intensively. If the relation between the two variables can be quantified in geostatistical terms, the spatial structure of the co-variable can be used to improve prediction of the target variable. For example, if we want to use kriging to estimate spatially distributed snow cover thickness on a mountain top, we may have only a small number of observations. This means that there is a considerable chance that the location where a prediction is wanted, will be located more than a distance a away from any observations.

Exercise 19: What will be the best estimate of snow cover thickness in this case?

Introducing a co-variable, for example elevation, can improve our estimation; suppose our measurements are located on the mountain side. A prediction point located more than a distance a away from any observation location will probably have a thinner snow cover if it is located on a relatively low elevation.

In order to use co-kriging, it is not sufficient to construct semivariograms for the target variable and co-variable only. Additional information on their ***joint distribution*** must also be provided in the form of the so-called ***cross-variogram*** (The name ***cross-semivariogram*** is actually more correct, but simply too long). Fortunately, the procedure for constructing a cross-variogram is not so different from that of a 'normal' semivariogram. Let U be the target variable and V the co-variable. The cross-semivariance is then defined as:

$$2 \cdot \gamma_{UV}(h) = E\left\{ \left[Z_U(s) - Z_U(s+h) \right] \cdot \left[Z_V(s) - Z_V(s+h) \right] \right\} \qquad [35]$$

and can be estimated from observations according to:

$$\hat{\gamma}_{UV}(h) = \frac{1}{2n} \sum_{i=1}^{n} \left\{ \left[z_U(s_i) - z_U(s_i+h) \right] \cdot \left[z_V(s_i) - z_V(s_i+h) \right] \right\} \qquad [36]$$

When fitting cross-variogram models, a condition known as the ***Cauchy-Schwartz relation*** must be met:

$$\left| \gamma_{UV}(h) \right| \quad \leq \quad \sqrt{\left(\gamma_U(h) \cdot \gamma_V(h) \right)} \qquad \text{for all h} > 0 \qquad [37]$$

Provided Equation [37] holds for a particular combination of variograms, prediction variances will always be positive (as they should be).

For the case of a single co-variable V, the ordinary co-kriging estimator of $z_U(s_0)$ is:

$$\hat{z}_U(s_0) = \sum_{i=1}^{n_U} \lambda_{U_i} \cdot z(s_{U_i}) \; + \; \sum_{i=1}^{n_V} \lambda_{V_i} \cdot z(s_{V_i}) \qquad [38]$$

To avoid bias, the weights of the co-variable must sum to zero, whereas the weights of the target variable must sum to unity:

$$\begin{cases} \sum_{i=1}^{n_U} \lambda_{U_i} = 1 \\ \sum_{i=1}^{n_V} \lambda_{V_i} = 0 \end{cases} \qquad [39]$$

Minimization of the variance is obtained when:

$$\sum_{i=1}^{n_U} \lambda_{U_i} \cdot \gamma_{U_i}(s_{U_i}, s_{U_g}) + \phi_U + \qquad [40]$$

$$\sum_{i=1}^{n_V} \lambda_{V_i} \cdot \gamma_{V_i}(s_{V_i}, s_{V_g}) + \phi_V = \gamma_{UV}(s_0, s_g) \qquad \text{for } g = 1,\ldots, n$$

Exercise 20: Do the units of U and V have to be the same in order to calculate the estimator of z at s_0? Why (not)?

Block kriging

With ordinary kriging, we use point observations to estimate the value of the variable of interest at other, unobserved points. However, in some cases, we may want to use our point observations to estimate values representative of a greater length (1D), area (2D), or volume (3D); in other words, we want to estimate a variable of larger ***support***, meaning that it is representative of a larger ***domain***. This is the case if we are estimating temporal or spatial averages, densities, etc. These entities can be predicted using ***block kriging***. Please note that each form of point kriging also has a block kriging counterpart, so we have ordinary point kriging and ordinary block kriging, point co-kriging and block co-kriging, etc.

In practice, the ordinary block kriging system of equations is very similar to that of ordinary point kriging covered in section 3.2. With ordinary point kriging, the observations $z(s_i)$ can be considered as realizations of a spatially autocorrelated random function (Z) and because Z is random, the average of Z over the domain D must also be a random function. Similar to ordinary point kriging, the unknown, block-averaged value of Z (denoted $\overline{Z}$) can be estimated as a weighted combination of nearby observations at locations s_i:

$$\hat{\overline{\mathrm{Z}}} = \sum_{\mathrm{i}=1}^{n} \lambda_i \cdot \mathrm{Z}(\mathrm{s_i}) \qquad [41]$$

Analogous to equation [19], the prediction must be unbiased and the variance minimal, which is achieved when:

$$\sum_{j=1}^{n} \lambda_j \cdot \gamma(s_i, s_j) + \phi = \bar{\gamma}(s_i, D) \qquad \text{for i} = 1,\ldots, \text{i} = \text{n} \qquad [42]$$

and

$$\sum_{i=1}^{n} \lambda_i = 1 \qquad [43]$$

In contrast to equation [19], the right-hand term represents the semivariance between the point observation and all possible points in the domain D, which can be calculated by:

$$\bar{\gamma}\,(s_i, D) = \frac{1}{|D|} \cdot \int_{s \in D} \gamma(s_i\,, s)\, ds \qquad [44]$$

Usually however, the integral in equation [44] is not solved analytically, but rather by dicretizing the domain D into N nodes s_j, after which the block-average semivariance is calculated according to:

$$\bar{\gamma}\,(s_i, D) \approx \frac{1}{N} \cdot \sum_{j=1}^{N} \gamma(s_i\,, s_j\,) \qquad [45]$$

The block kriging variance is given by:

$$E\left[\hat{\bar{Z}} - \bar{Z}\right]^2 \;=\; \sum_{i=1}^{n} \lambda_i \cdot \bar{\gamma}\,(s_i, D) + \phi - \bar{\gamma}\,(D, D) \qquad [46]$$

where $\bar{\gamma}\,(\mathrm{D}, \mathrm{D})$ represents the average semivariance within the domain D:

$$\bar{\gamma}\,(D, D) = \frac{1}{|D|^2} \int_{s_2 \in D} \int_{s_1 \in D} \gamma\,(s_1, s_2) \quad ds_1\, ds_2 \qquad [47]$$

which is also approximated by numerical integration:

$$\bar{\gamma}\,(D, D) \approx \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \gamma\,(s_i, s_j) \qquad [48]$$

3.4 Other forms of kriging

3.4.1 Simple kriging

In ordinary kriging, uncertainty about the prediction is introduced by two factors, the obvious one being the autocorrelation; however, some of the total uncertainty is introduced to account for the population mean μ of the spatially varying function $Z(s_0)$ being uncertain. This extra uncertainty is reflected in the addition of the Lagrange parameter ϕ in equation [18]. Under a stricter set of conditions, known as ***second-order stationarity***, when μ is known without error (for example from theoretical considerations), and the variance is finite, ***simple kriging*** may be applied. Because the population mean is known, it is unnecessary to implicitly estimate its value using observations (as is the case with ordinary kriging), which in turn makes simple kriging predictions less uncertain.

3.4.2 Probabilistic kriging/ Indicator kriging

For certain purposes, a researcher may not be primarily interested in the best unbiased prediction of a certain variable, but rather in the probability of whether or not a certain threshold is exceeded. This may be the case with, for example, making a decision to mine a certain area, or to study a sub-area of the study site more intensively in search of a specific species. In such cases, ***indicator kriging*** is a useful method to assist in the decision-making. With this method, the original data are transformed from the continuous scale to a binary scale; depending on whether or not a certain threshold value is exceeded, observations are assigned a one or a zero. Semivariograms and spatial predictions can subsequently be calculated for the binary data in the usual way; spatial patterns of predictions will display values between 0 and 1, representing the desired probability that the threshold is exceeded at that location. To assess the effect of varying the threshold value, multiple indicator kriging maps of different threshold value may be compared.

3.4.3 Simulation

So far, we have used kriging to predict the value of Z at unobserved locations. However, application of kriging is not limited to merely interpolating the observations; it can also be used to generate ***stochastic stimulations***. With stochastic simulation, intermediate points are not assigned the best estimate, but instead the value is drawn from that point's probability distribution, therefore the result of a stochastic simulation is referred to as a ***realization***. Two types of simulation exist; if the observations will be reproduced in all realizations, the stochastic simulation is ***conditional***, as opposed to ***unconditional simulation*** where the appearance of the semivariogram is the only contraint.

Exercise 21: Explain why realizations show more erratic patterns compared to interpolations.

Exercise 22: Why is unconditional simulation different from generating a stochastic field with, for example, the MATLAB `rand` function?

To illustrate the importance of choosing the appropriate method for representing the spatial variation, consider the figures below: Figure 21A depicts observations of elevation on an Alpine slope. The observations have been interpolated using a Gaussian semivariogram model, yielding a smoothly varying slope surface (solid line). A conditional simulation with the same semivariogram model has yielded the realization of the slope surface shown in Figure 21B.

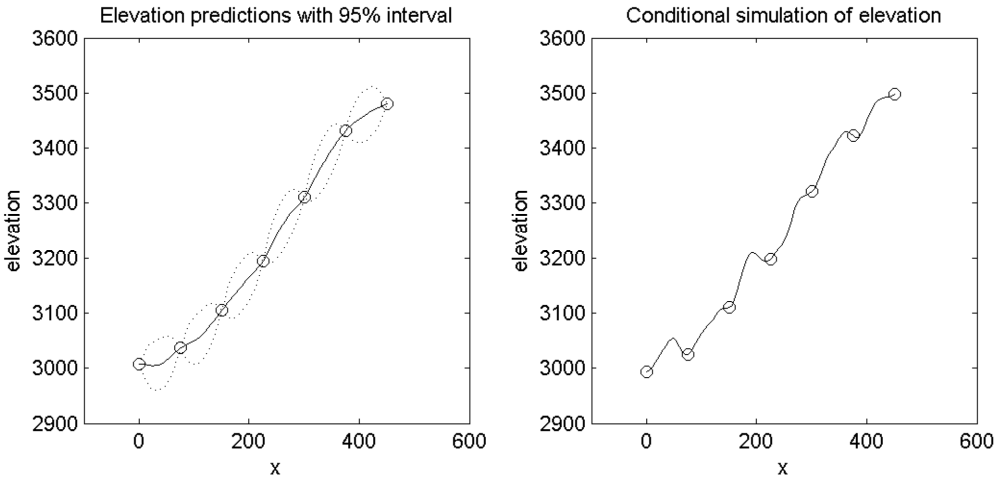


Figure 21. **(A)** Observations of elevation (circles), interpolated using a combined nugget/ Gaussian semivariogram model (solid line). Parameters of the semivariogram: c_{nugget} = 0.001; $c_{Gaussian}$ = 500; a = 25; dashed lines indicate 95% probability interval. Before the interpolation, the trend line $m(s)$ has been subtracted from the observations. Trend line has been calculated according to $m(s)$ = 3250 + 250 · sin(2 · π · (x-250)/1000). **(B)** Realization of elevation observations, constructed using the same semivariance model and deterministic component as in (A).

When the two representations of the slope are used as input to a computer model simulating rockfall, one can image the differences between model outcomes that can arise from the method used. For example, a falling rock might get stuck in the local depressions at x = 75, x = 225, or x = 375 when the realization is used, whereas the rock would end up near x = 0 if the prediction is used.

4 MATLAB’s kriging toolbox

4.1 Gstat

In order to be able to work with MATLAB’s toolbox later on, it is first necessary to read a little about how the MS-DOS program that is actually at the core of the toolbox should be operated. This program (called *Gstat*) was developed by E.J. Pebesma and others during the 1990s, when kriging techniques were starting to be employed in research fields outside of mining geostatistics, where its theoretical basis lies. The program is really a great tool for the modeling, prediction (i.e. interpolation), and simulation of geostatistical data with up to 3 dimensions. It offers great freedom in the methods to be used, ranging from simple kriging to co-kriging, and beyond. However, coming from the MS-DOS era, its graphical user interface is rather unforgiving and not very pleasant to work with. During this module, we will therefore not use Gstat in any direct way, but choose an approach in which Gstat is used only as an application running in the background, hence not visible for the user.

In its basic form, operating Gstat consists of 3 steps:

(1) Creating an ASCII (i.e. unformatted text) file, for example using NotePad. This file has a `.cmd` extension[♣], indicating that it is a so-called command file. This file should be saved in the directory where the Gstat program is located.

(2) In the command prompt window (In Windows XP usually to be found at Start>>All Programs>>Accessories), typing `gstat.exe` followed by a space and the name of the command file at the command prompt will invoke the program if the Gstat executable is located in the current directory.

(3) The output files generated by `gstat.exe` can be viewed using other software.

Action Take a look at the example command files in the '/examples/01_command_files' directory. Read the comments for example-specific information. A complete overview of Gstat options is provided by Pebesma [2001] that has been included in the '/literature' folder.

Exercise 23: Using Gstat to create an experimental variogram of 2D surface soil moisture measurements taken at different time steps (thus effectively constituting a 3D model) will not generate an error. Why are the results nonetheless of no use?

[♣] In Windows XP, you can set whether you want to show a file's extension in Explorer as follows: go to 'Control Panel'>>'Folder Options'>>select tab 'View' and (de)select the checkbox 'hide extensions for known file types'.

4.2 mgstat_lite

To avoid all the difficulties of working with command files in the DOS-environment, T.M. Hansen wrote a number of MATLAB m-files that together constitute the kriging toolbox known as *mGstat*[♣]. The files from this toolbox actually form a shell around Pebesma's Gstat program. This way, MATLAB is able to *automatically* compose the command file necessary to run Gstat, call the program `gstat.exe` from within MATLAB, let the data be processed with any desired Gstat options, and finally to import Gstat's output back into MATLAB, where it can be processed further or visualized just like any other data. This technique uses so-called file I/O to write command files based on the user's settings. This way, the user can not make any 'spelling' (syntaxis) or 'grammar' mistakes, as happens often when composing command files by hand.

Exercise 24: A script ('/exercises/01_file_io/ex_file_io.m') has been prepared that illustrates the operation of MATLAB functions used in file I/O such as `fopen`, `fprintf`, and `fclose`. Open the script in the MATLAB editor. When you understand how it works, alter it in such a way that it will precisely reproduce the contents of the file 'exercises/01_file_io/file_io_325.txt'.

Based on Hansen's approach, a MATLAB kriging toolbox ('mgstat_lite') has been composed with which one can perform the most common kriging operations, such as constructing and plotting (experimental) semivariograms, local and global ordinary kriging of 1-D, 2-D, or 3-D spatial data, co-kriging of 2 variables, block kriging, conditional simulation, unconditional simulation, and visualization of the output generated by the "gstat.exe" program.

Action Use Microsoft Windows Explorer to navigate to the folder 'mgstat_lite_toolbox' (see Figure 1).

Among its contents is a folder 'functionality' containing the m-files as well as Pebesma's `gstat.exe` MS-DOS program. Together, they constitute the backbone of the toolbox. For all m-files within the toolbox, a help file has been compiled with information on how the function should be used, what format the input should be in, etc. The help files have the same name as the m-files but with a different extension (*.html). They are located in the 'documentation' folder. The help files effectively form the toolbox manual, similar to the doc function that is available for the scripts and functions that come with the MATLAB software package. In order to access the manual, the 'functionality' folder must be added to the MATLAB path.

[♣] http://www.gfy.ku.dk/~tmh/mgstat

Action Start the MATLAB program and the m-file editor. Please note that m-files should only be opened from within MATLAB's m-file editor, which itself should be opened by typing the command `edit` at the prompt. Opening an m-file by clicking its filename in Windows Explorer will result in a reduced editor functionality; for example, debugging is not supported in this case.

Action Open the script 'add_manual.m' located in the folder 'exercises/02_add_manual' (see Figure 1) that has been supplied with this module. This script contains a few command lines that will add the 'functionality' folder to the MATLAB path upon execution.

Action Execute the script. Type `path` at the prompt to verify that the correct folder has been added to the MATLAB path. (The default path can be restored by typing the command `restoredefaultpath` at the prompt).

IMPORTANT NOTICE

Proper functioning of the manual depends on the 'functionality' and 'documentation' folders being located in the same directory.

Having added the toolbox documentation to the MATLAB path, we may now access the manual.

Action Type `manual` at the prompt. This will start MATLAB's help browser, providing an overview of the mgstat_lite toolbox. Click on the link 'manual' and read about how this function should be used to access the help documentation of the toolbox as well as that of the built-in MATLAB functions.

4.2.1 mgstat_lite: ordinary kriging of contaminants in the Meuse river valley

Exercise 25: The lower parts of the river Meuse (Maas) valley have long been subject to contamination with heavy metals due to mining activities upstream in Belgium. A large study has been executed, in which soil samples from more than 150 points were analyzed (*Rikken and Van Rijn* [1993]). The file 'X_Y_Zn.txt' in the folder '/exercises/03_meuse_variogram' contains data about the zinc (Zn) content of these samples as well as information about where each sample was taken. The first column in this file represents the Easting of the observation location in meters; the second column represents Northing, also in meters. The third and last column represents the observed value of Zn concentration in parts per million. During this exercise, you will determine the experimental semivariogram and choose an appropriate semivariogram model. In the next exercise, you must use this semivariogram model to make spatial predictions of Zn.

a) The first step in making a prediction is determining whether the data exhibits any spatial structure by constructing an experimental semivariogram. Such a semivariogram is really a vector of values h and a vector of values γ. If you had to write a MATLAB function that calculates the datapoints $[h, \gamma]$ for N bins given the input variables X, Y (i.e. the location of the measurement in 2 dimensions), and V (i.e. the value of the measurement), which major steps would your function have? Use no more than 5 steps. (Hint: taking another look at equation [10] may prove useful).

b) The function `svgram_exp` that is included in the folder 'functionality' calculates the data needed for plotting the semivariogram. Open the script 'meuse_zinc_svgram_exp.m' in the folder 'exercises/03_meuse_variogram'. Some parts have already been completed, for example the part where the data from 'X_Y_Zn.txt' is loaded into the workspace and the part that visualizes the experimental semivariogram. After consulting the manual on the use of `svgram_exp`, complete the script. Do not use any of the optional parameter/value pairs yet. Your result should look like Figure 22.

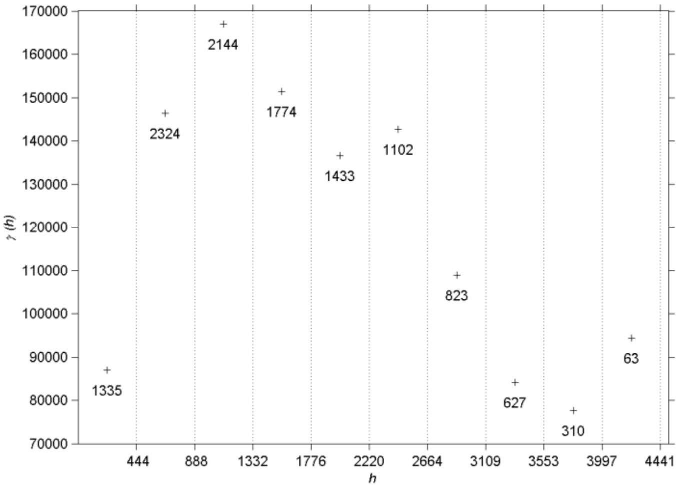


Figure 22. Preliminary result from 'svgram_exp.m'.

c) Generate a location map showing the locations of the observations. Use green dots for below-average values and red dots for above-average values. When fitting a semivariogram model to the zinc data, do you think it is sensible to include lags in excess of, say, 1200 meter? Why (not)?

d) Set the 'cutoff' parameter of `svgram_exp` to the appropriate value. Alter the code in order to set the y-axis minimum to 0.

e) Are you satisfied with the number of bins? Why?

f) From the experimental semivariogram, it may be inferred that the Zn concentration semivariance increases with distance between observation points. Do you think the Zn concentration also fluctuates with absolute location? Why?

g) Describe how you think the Zn concentration varies over the study area. What type of semivariogram model (or combination thereof) do you think best represents this variation? Why?

h) Does the data support your answer from (g)? Why (not)?

i) Consult the manual on the function calc_gamma and use a trial-and-error approach to fit a (combined) nugget/exponential semivariogram model to the $[h, \gamma]$ data points. Alter the script to simultaneously plot the data points $[h, \gamma]$ and the fitted model.

Exercise 26: With the semivariogram from the previous exercise, you must make a spatial prediction of the Zn concentration using 2D ordinary kriging. To help you, a script has been prepared. It is located in the '/exercises/04_meuse_2dok' folder.

a) Fill in your semivariogram parameters and the necessary input arguments to the function mgstat_lite (see manual) to make maps of spatial predictions and estimation errors. Use the variable names pre_inf and var_inf for the 2D arrays holding the kriging predictions and variances, respectively. **TIP**: It may be useful to take a look at the 1D ordinary kriging script in the '/examples/02_1d_ok' folder. Review the Chapter 'Debugging' in *Bouten et al.* [2006] if necessary.

b) Save the kriging predictions and kriging variances maps to a binary *.mat file.

c) Calculate the lag at which your semivariogram reaches 95% of its sill ($h_{95\%}$).

d) Observations at lags larger than $h_{95\%}$ do not significantly contribute to improving the prediction. Use the mgstat_lite option radius to include only observations within a distance $h_{95\%}$ of the prediction location. Use the variable names pre_95pct and var_95pct for the 2D arrays holding the kriging predictions and variances, respectively.

e) Save the kriging predictions and kriging variances maps to a binary *.mat file.

f) Start a new script in which you load the *.mat files and calculate the difference var_{∞} - $var_{95\%}$. Where on the map do the largest differences occur? **TIP**: use imagesc_plot for visualization.

g) Why do the largest differences occur in those areas?

h) When all observations within such a large radius as $h_{95\%}$ are included to generate a prediction, local effects may become smoothed out. Modify the script to calculate predictions based only on observations within 500 m. Compare the kriging prediction maps with pre_inf. Which of these do you think is more realistic? Why?

i) Because of the relatively small radius, predictions of Zn sometimes depend on the value of merely one observation. This effect can be avoided to by including the 'min' option of the function mgstat_lite. Set the minimum number of observations used for making a prediction to an appropriate value.

4.2.2 mgstat_lite: co-kriging contaminant concentration using distance to river Meuse

Exercise 27: In this exercise, you will use cokriging to make more accurate predictions of the spatial distributions of zinc. The secondary variable used to improve the prediction will be squared normalized distance to the river. A script has been prepared that reads the necessary data into the MATLAB workspace. It is located in the folder '/exercises/05_cokriging'. However, in order for gstat to yield the desired results, we first need to transform the data.

a) Alter the script to perform a log-transform on the zinc data. Use a natural logarithm. Call the new variable `LNzinc`.

b) Square the normalized distance-to-river data. Call the new variable `ndist2`.

c) Construct location maps of `LNzinc` and `ndist2`, displaying the above-average values as red dots and the below-average values as green dots.

d) Given the nature of the variation, what kind of semivariogram model do you think would best represent the variation of `ndist2`?

e) Construct experimental semivariograms of both `LNzinc` and `ndist2` and fit semivariogram models through the $[\gamma, h]$ points. Use a range of 1025 m for both semivariogram models.

f) Consult the manual on how to use `mgstat_lite` with co-kriging. Assume that the cross-variogram between the two variables `LNzinc` and `ndist2` is modeled by a semivariogram with parameters c_{nugget} = 0, $c_{spherical}$ = 0.125, and a = 1025. Fill in the parameters in the script.

g) Set 'min', 'max', and 'radius' to appropriate values and run `mgstat_lite` with the co-kriging setup.

h) Compare your co-kriging results with those of 'normal' kriging you made earlier. Explain the differences for both predictions and variances.

4.2.3 mgstat_lite: block kriging buzzard density over the Netherlands

Exercise 28: The common buzzard (*Buteo buteo*) can be a threat to aviation because of its relatively high body mass and because it is indeed fairly common in The Netherlands. It is therefore important to record its occurrence and derive spatially continuous predictions from these observations. One way of generating spatially continuous maps for the whole of The Netherlands is by using block kriging.

The script 'buzz_svgram_exp.m' located in the '/exercises/06_buzzards_bk' folder reads the observations of buzzards into the MATLAB workspace. Alter the script in such a way that it displays the buzzard semivariogram with 14 lag bins and 'cutoff' set to 42 km.

Exercise 29:

a) Copy the local ordinary point kriging script you made earlier to the folder pertaining to this exercise. Alter the script to make local ordinary point predictions of the occurrence of buzzards. In order to avoid long calculations, set both `min` and `max` properties to 20 and `radius` to 42 km.

b) Save the maps of local ordinary point kriging predictions and variances to a *.mat file.

c) Consult the manual on how to use the function `mgstat_lite` in order to make block predictions. Use blocks of size 5x5 km. In order to avoid long calculations, set both `min` and `max` properties to 20 and `radius` to 42 km.

d) Save the maps of local ordinary block kriging predictions and variances to a *.mat file.

e) Compare the maps and describe the differences between the methods for predictions and variances.

f) How does varying the blocksize affect the spatial prediction? And block kriging variance?

4.2.4 mgstat_lite: simulation of rockfall slope transect

Exercise 30:

a) The folder '07_rockfall_us' contains a script m-file that was used to generate the conditional simulation depicted in Figure 21. The appearance of the realization is very smooth; apart from the use of a Gaussian semivariogram model, what property of the statistical structure causes the smoothness?

b) Adapt the script in such a way that it performs an unconditional simulation upon execution.

c) Why do you need to specifiy the '`pop_mean`' parameter when using unconditional simulation?

d) Construct a semivariogram of your (unconditional) realization and compare it to the semivariogram that was used to generate the realization. Explain the difference.

4.2.5 mgstat_lite: conditional simulation

Exercise 31: Hydrological models often divide space into a number of grid cells, each of which having certain attributes, such as porosity, hydraulic conductivity and so on; the value assigned to each of the attributes is assumed to be representative of the entire grid cell. To investigate whether or not this assumption is violated, a Scandinavian hydrologist has collected soil hydraulic conductivity data for one of his model's grid cells by dividing the grid cell area (25 m^2) into smaller cells (0.01 m^2) and sub-sampling these. The grid cell happens to be located in a region dominated by patterned ground. Hydraulic conductivity in these areas is known to be orders of magnitude smaller for the soil inside of the polygons, compared to the edges (see Figure 23). When water is transported through such a soil, the inside of the polygon is therefore more or less by-passed, with most of the flow going around it. The hydrologist is aware of this, and wants to use realizations of the patterned ground to use as input to his model, instead of having to assume an average conductivity value.

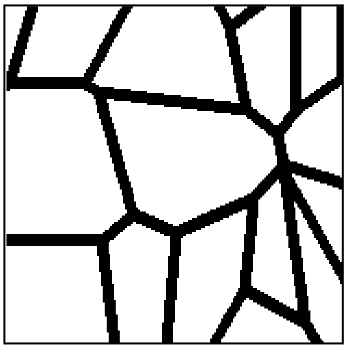


Figure 23. Soil hydraulic conductivity map showing the spatial distribution of conductive polygon edges (black) and the less permeable inside of the polygons (white).

a) The file 'conduct.mat' located in the folder '/exercises/08_conductivity_cs' contains the conductivity data. Randomly sample 500 points to construct the semivariogram with.

b) Use conditional simulation to generate a realization of conductivity. Convert the continuous-scale realization to binary values by dividing them into two classes:

class	*description*
0	smaller than the mean of your samples;
1	larger than the mean of your samples.

c) Would the realization pattern be more like Figure 23 when a zonal anisotropic semivariogram were used? Why (not)?

d) What property does the pattern in Figure 23 have that is not captured in the semivariogram, hence not represented in the realization?

e) Would you recommend using realizations as input to the model?

5 References

Bouten, W., Kraal, P., and Spaaks, J.H. (2006) "Programming in MATLAB" University of Amsterdam

Burrough, P.A. and McDonnell, R.A. (1998) "*Principles of Geographical Information Systems"* Oxford University Press

Burt, J.E. and Barber, G.M. (1996) "*Elementary statistics for geographers*" 2nd Edition, Guilford Press, London; 640 pp. ISBN: 0-89862-282-4

Davis Jr, R.A. (1994) "*The evolving coast*" Scientific American Library; ISBN: 0-7167-5042-2

De Gruijter, J.J., Brus, D.J., Bierkens, M.F.P., and Knotters, M. (2005) "*Sampling for natural resource monitoring*" Springer, Berlin

Mathews, J.H. and Fink, K.D. (1992) "*Numerical methods using MATLAB*" Prentice Hall, New Jersey, USA; ISBN: 0-13-270042-5

Meijerink, A.M.J., de Brouwer, J.A.M., Mannaerts, C.M. and Valenzuela, C.R. (1994) "*Introduction to the use of geographic information systems for practical hydrology*" IHP - IV M 2.3. Enschede, ITC, 1994. ITC Publication 23, 243 pp. ISBN: 90-6164-100-4.

Pebesma, E.J. (2001) "*Gstat user's manual*" Dept. of Physical Geography, Utrecht University, Utrecht, The Netherlands

Rikken, M.G.J. and Van Rijn, R.P.G. (1993) "*Soil pollution with heavy metals – An inquiry into spatial variation, cost of mapping and the risk evaluation of copper, cadmium, lead and zinc in the floodplains of the Meuse west of Stein, the Netherlands*" Doctoraalveldwerkverslag, Dept. of Physical Geography, Utrecht University

Lutgens, F.K., and Tarbuck, E.J. (2003) "*Foundations of Earth Science*" 3rd Edition Pearson Education Ltd., London; ISBN: 0-13-146103-6

6 Glossary of terms

7 Appendices

7.1 Brief summary of concepts from univariate statistics

7.1.1 Expected value

The expected value of a random variable X is defined as:

$$E(X) = \int P(X) \cdot X \quad dX \qquad [49]$$

With $P(X)$ the probability of X. (Note that the actual value of E may be *un*expected, for example with bimodal probability distributions).

7.1.2 Sample variance

$$Var(x) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} \qquad [50]$$

With n the number of samples.

7.1.3 Sample standard deviation

$$s(x) = \sqrt{Var(x)} \qquad [51]$$

7.1.4 Covariance

$$Cov(X,Y) = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} x_i y_j P(x_i, y_j) - E(X)E(Y) \qquad [52]$$

7.1.5 Correlation

$$\rho_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \qquad [53]$$

With σ the population standard deviation.

7.2 Fitting a 2nd-order polynomial through 3 observations – example

Suppose we have three observations (visualized in Figure 24) through which we want to fit a 2nd-order polynomial:

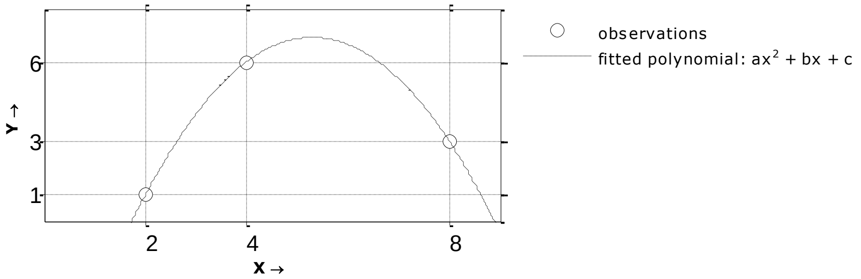


Figure 24. Three observations and the fitted 2nd-order polynomial.

With these three points, we can establish an exact fit for a 2nd-order polynomial (see equation [3]; for reasons of clarity, coefficients c_2, c_1, and c_0 have been substituted for a, b, and c, respectively):

$$ax^2 + bx + c = y \qquad [54]$$

In our example, filling in the observed values yields:

$$a \cdot 4 + b \cdot 2 + c = 1 \qquad [55]$$

$$a \cdot 16 + b \cdot 4 + c = 6 \qquad [56]$$

$$a \cdot 64 + b \cdot 8 + c = 3 \qquad [57]$$

Equating [55] to [56] yields:

$$a \cdot 4 + b \cdot 2 + c = a \cdot 16 + b \cdot 4 + c - 5 \qquad [58]$$

Bringing the first polynomial coefficient (a) to the left, and all others to the right yields:

$$a \cdot 4 = a \cdot 16 + b \cdot 2 - 5 \qquad [59]$$

$$a \cdot -12 = b \cdot 2 - 5 \qquad [60]$$

$$a = b \cdot -\tfrac{1}{6} + \tfrac{5}{12} \qquad [61]$$

Equating [56] to [57] yields:

$$a\cdot 16+b\cdot 4+c=a\cdot 64+b\cdot 8+c+3 \quad [62]$$

Bringing the second polynomial coefficient (b) to the left, and all others to the right yields:

$$b\cdot 4=a\cdot 48+b\cdot 8+3 \quad [63]$$

$$b\cdot -4=a\cdot 48+3 \quad [64]$$

$$b=a\cdot {}^{48}\!/_{-4}+{}^{3}\!/_{-4} \quad [65]$$

$$b=a\cdot -12-0.75 \quad [66]$$

Substituting equation [66] into [61] yields:

$$a=(a\cdot -12-0.75)\cdot -{}^{1}\!/_{6}+{}^{5}\!/_{12} \quad [67]$$

$$a=a\cdot 2+{}^{1}\!/_{8}+{}^{5}\!/_{12} \quad [68]$$

$$a=-{}^{1}\!/_{8}-{}^{5}\!/_{12} \quad [69]$$

Substituting equations [66] and [69] into [55], [56], or [57] and solving for c yields:

$$c=-8\,{}^{1}\!/_{3} \quad [70]$$

Substituting polynomial coefficients a, b, and c into equation [54] results in an exact fit through the observations shown in Figure 24.