THEORY AND APPLICATION OF COLLOCATION IN SURVEYING

A. Dermanis

Department of Geodesy and Surveying
University of Thessaloniki
1. INTRODUCTION

In the last decade new methods for handling observations related to the gravity field of the earth have found wide application. These methods have been indiscreetly labeled with the collective name "collocation" (Krarup, 1969), which includes two different but strongly related aspects:

The first is the adjustment of observations when the model parameters include quantities related to the gravity field, the so-called signals. This adjustment, although it is really nothing new from the viewpoint of the classical least squares adjustment theory, has been called "least squares collocation" by Moritz (1972). The second aspect is the prediction of new gravity-related quantities from already known ones. The known signals may be either directly observed without errors (exact collocation) or be estimated from least squares adjustment. The last case includes the prediction of new signals from observations of old signals affected by observational errors (smoothing collocation).

The essential characteristic of the prediction method is the discrete character of the original information which justifies to a certain degree the use of the term collocation, in analogy with the inner collocation methods for numerically solving boundary value problems in partial differential equations, from boundary values on a set of discrete points instead of the whole of the boundary (Collatz, 1966).

The prediction by collocation consists of two separate steps, although their separation cannot be clearly realized by just looking at the algorithms used. The first step is the determination of the gravity potential (in practice its disturbing part is only used as a consequence of linearization) from known gravity related signals at discrete points. This step is essentially an interpolation procedure although of a somewhat generalized type. A straightforward interpolation would appear in case the known signals were just values of the unknown potential function at discrete points.

The insufficiency of the original information (less data than unknowns) leads to an infinity of solutions for the potential function, in the same way that one can interpolate in infinite many ways from discrete function values. The heart of the collocation methods is exactly how one potential
function is selected out of an infinite number of candidates.

The second step is more or less trivial. Once a potential function is determined, any new gravity related quantity can be easily computed (predicted) in a straightforward way.

These techniques are not particular to the analysis of gravity related data. They can be applied, and have in fact been applied in other disciplines since a long time, for the analysis of observations which involve quantities related to an underlying unknown function (Dermanis, 1978). In fact the case where the unknown function is the gravity potential of the earth is not the most appropriate for the understanding of the method. The relation of the signals (gravity anomalies, deflections of the vertical, geoid heights, etc.) to the potential function and the peculiarities of the spherical domain of the surface of the earth introduce complications thus hiding the simple essential principles of the method which are independent of its application to any particular problem.

Here we shall try to reveal the principles of the collocation method with the help of some very simple applications in surveying problems, emphasizing mainly its interpolation aspects. A lot of mathematical rigor has been conciously sacrificed for the shake of simplicity.

2. INTERPOLATION

The need to interpolate is familiar to all branches of applied science. Looking in textbooks on approximation theory or numerical analysis, one finds quite a few techniques for interpolating functions of one variable whose domain is the real line. Surveyors deal with data on the two dimensional horizontal plane, and we shall therefore discuss interpolation of functions of two variables, i.e., the plane coordinates $x$ and $y$.

To give an example where the need for interpolation arises, consider the situation where heights have been determined at discrete points of an area and contours must be drawn. The contours are a graphical representation of a height function $h(x,y)$ and interpolation is actually carried out
not analytically but by graphical methods; a network of adjacent triangles is formed and linear interpolations lead to broken contour lines which are later on smoothed to become the usual curved contours of topographic maps. If contours are to be drawn automatically by an automatic plotter, heights must be known on a regular grid of points. Since the original points where heights are known have an irregular distribution, an interpolation must take place in order to arrive at height values on the grid or any other points.

For another example suppose that heights \( h \) have been measured at the vertical control points of a certain area. In a later epoch heights \( h' \) are measured at a limited number of vertical control points, e.g., the higher order points, which differ from the previously measured heights \( h \) due to vertical crustal motion described by the uplift function

\[
(1) \quad u(x,y) = h'(x,y) - h(x,y)
\]

The problem is to determine by means of interpolation the uplift function \( u(x,y) \) from its values \( u(x_i,y_i) \) at the remeasured points \( P_i \) with horizontal coordinates \( (x_i,y_i) \), \( i=1,2,\ldots,n \). Once \( u(x,y) \) is obtained, its values at the secondary points, or any other desired point, may be used for the correction of the old heights \( h \).

Avoiding mathematical descriptions for the time being, the interpolation procedure has the following steps:

a. Construct a collection of functions, among which the required function is to be sought.

b. Pick up from the whole collection only those functions whose values at the given points agree with the known values of the unknown function. There might be none, only one, or many such functions in the collection.

c1. If there is only one such function the interpolation problem has been already solved.

c2. If there are many such functions (even infinite), pick out the one which looks more appropriate. To do this, one must a priori decide about a way to measure the appropriateness of each function. This measure is called the optimality criterion and leads to the optimal function.

c3. If there is no such function, choose among the whole collection the one which best fits the known values. Here again a criterion is needed which measures how well each function fits the data. In this case we do not
have an exact interpolation, but rather what is called a smoothing interpolation.

The first problem in the above procedure is the construction of the collection of functions where the interpolating function is to be sought. The simpler way to do this is the so called linear interpolation; a family of known functions \( \varphi_i(x, y) \), \( i=1, 2, \ldots, m \) is rather arbitrarily chosen and the required collection of functions is the set of all the linear combinations

\[
(2) \quad f(x, y) = a_1 \varphi_1(x, y) + a_2 \varphi_2(x, y) + \ldots + a_m \varphi_m(x, y)
\]

where the \( a_i \) are real coefficients. To each set of numerical values of the coefficients \( a_i \) corresponds a different function from the collection. The problem of selecting an interpolating function is equivalent to selecting a set of numerical values for the coefficients \( a_i \), \( i=1, 2, \ldots, m \).

Two familiar choices for the functions \( \varphi_i(x, y) \) are polynomials

\[
(3) \quad f(x, y) = \sum_{i,j=1}^{N} b_{ij} x^i y^j = b_{00} + b_{10} x + b_{01} y + b_{20} x^2 + b_{11} x y + b_{02} y^2 + \ldots + b_{0N} y^N.
\]

and trigonometric functions

\[
(4) \quad f(x, y) = \sum_{i=1}^{N} \{ a_i \cos^2 \frac{\pi x}{T} \cos^2 \frac{\pi y}{T} + b_i \cos^2 \frac{\pi x}{T} \sin^2 \frac{\pi y}{T} + c_i \sin^2 \frac{\pi x}{T} \cos^2 \frac{\pi y}{T} + d_i \sin^2 \frac{\pi x}{T} \sin^2 \frac{\pi y}{T} \}
\]

Both choices have shortcomings that will be pointed out later.

If

\[
(5) \quad s_k = f(x_k, y_k)
\]

are the known values of the function at \( n \) points \( P_k \), \( k=1, 2, \ldots, n \), with coordinates \( (x_k, y_k) \), our problem is to solve the system of \( n \) equations with \( m \) unknowns
\[ f(x_1, y_1) = \varphi_1(x_1, y_1) a_1 + \varphi_2(x_1, y_1) a_2 + \ldots + \varphi_m(x_1, y_1) a_m = s_1 \]
\[ f(x_2, y_2) = \varphi_1(x_2, y_2) a_1 + \varphi_2(x_2, y_2) a_2 + \ldots + \varphi_m(x_2, y_2) a_m = s_2 \]
\[ \vdots \]
\[ f(x_n, y_n) = \varphi_1(x_n, y_n) a_1 + \varphi_2(x_n, y_n) a_2 + \ldots + \varphi_m(x_n, y_n) a_m = s_n \]

These equations are written in matrix form

\[
\begin{bmatrix}
\varphi_1(x_1, y_1) & \varphi_2(x_1, y_1) & \ldots & \varphi_m(x_1, y_1) \\
\varphi_1(x_2, y_2) & \varphi_2(x_2, y_2) & \ldots & \varphi_m(x_2, y_2) \\
\vdots & \vdots & \ddots & \vdots \\
\varphi_1(x_n, y_n) & \varphi_2(x_n, y_n) & \ldots & \varphi_m(x_n, y_n)
\end{bmatrix}
\begin{bmatrix}
a_1 \\
a_2 \\
\vdots \\
a_m
\end{bmatrix}
= 
\begin{bmatrix}
s_1 \\
s_2 \\
\vdots \\
s_n
\end{bmatrix}
\]

or simply

\[ F \ a = s \]

where the element of the \(i^{th}\) row and \(j^{th}\) column of \(F\) is

\[ F_{ij} = \varphi_j(x_i, y_i) \]

In the classical interpolation techniques found in textbooks of numerical analysis it is always \(n=m\). In this case there is only one solution of the system (6), i.e., there is only one function out of the whole collection which takes on all the known values \(s_i\) at the points \(P_i\). This unique function corresponds to coefficient values \(\hat{a}_i\) obtained by

\[ \hat{a} = F^{-1} s \]

provided the determinant of \(F\) does not vanish.

In case \(n>m\) (more equations than unknowns) there is generally no
solution for the system (6). For any function \( f(x, y) \) out of the collection described by equation (2), its values at the points \( P_i \) differ from the given ones by quantities

\[
(11) \quad e_i = s_i - f(x_i, y_i)
\]

in which case equation (8) must be replaced by

\[
(12) \quad s = Fa + e
\]

In this case we have not an (exact) interpolation but rather a smoothing interpolation. In order to obtain a solution which best fits the known values \( s \), the residuals \( e_i \) must become as small as possible. Introducing the quantity

\[
(13) \quad \sum_{i,j=1}^{n} W_{ij} e_i e_j = e^T W e
\]

as a measure of the smallness of the residuals, where \( W \) is a symmetric positive-definite weight matrix, we arrive at a typical least squares problem

\[
(14) \quad s = Fa + e, \quad e^T W e = \min
\]

similar to the one in network adjustment by the method of observation equations. The well known solution is

\[
(15) \quad \hat{a} = (F^T W F)^{-1} F^T W s
\]

The case of more interest to us, since it is connected as we shall see to the collocation methods of physical geodesy, is the case with \( n < m \) (more unknowns than equations). The system (6) or (8) has now an infinite number of solutions. These solutions a give rise to those functions out of the original collection (2) which take on the known values \( s_i \) at the points \( P_i \). We must next pick up the most appropriate out of these functions.

For this purpose we introduce, for the time being without physical justification, a quantity

\[
(16) \quad \sum_{i,j=1}^{m} R_{ij} a_i a_j = a^T R a
\]
as a measure of "appropriateness", where $R$ is a symmetric positive-definite weight matrix. Minimization of the quantity (16) leads to a least squares problem

$$\text{(17)} \quad F a = s , \quad a^\top R a = \min$$

similar to the one in network adjustment with the method of condition equations. The well-known solution is

$$\hat{a} = R^{-1}F^\top (F R^{-1}F^\top)^{-1} s$$

A mathematician would have called the quantity

$$\text{(19)} \quad \|f\| = (a^\top R a)^{\frac{1}{2}}$$

the norm of the function $f$, and this interpolation method takes on the name minimum norm interpolation.

The problem to be now solved is how to choose the matrix $R$, or equivalently the norm (19) in a meaningful way. This is the so-called norm choice problem. The most widely admissible suggestion in the direction of solving this problem is to choose $R$ or $\|f\|$ in a way that the function

$$\text{(20)} \quad \hat{f}(x,y) = \hat{a}_1 \varphi_1(x,y) + \hat{a}_2 \varphi_2(x,y) + \ldots + \hat{a}_m \varphi_m(x,y)$$

resulting from (18) is as smooth as possible. The reasons behind this suggestion are well grounded. A lot of natural processes described by functions have indeed a smooth variation. A good example is the uplift function (1) describing vertical crustal motion as we have already seen. Even when dealing with processes which are not of such a smooth character, it simply does not make sense to interpolate by means of functions which are not smooth. If we do so, we introduce fluctuations in function values between observation points which have extremely little chance to agree with the similar fluctuations of the "true" function. Some choices where the smoothness criteria are easy to see are the minimization of the mean square value of the function $f$ in a certain area $A$ of the plane.
\[ (21) \quad \| f \| ^2 = \frac{1}{A} \int _A \left[ f(x,y) \right] ^2 \, dx \, dy \]

which leads to weights

\[ (22) \quad R_{ij} = \frac{1}{A} \int _A \varphi _i (x,y) \varphi _j (x,y) \, dx \, dy \]

and the minimization of

\[ (23) \quad \| f \| ^2 = \frac{1}{A} \int _A \left[ f^2 + \left( \frac{\partial f}{\partial x} \right)^2 + \left( \frac{\partial f}{\partial y} \right)^2 \right] \, dx \, dy \]

which leads to weights

\[ (24) \quad R_{ij} = \frac{1}{A} \int _A \left[ \varphi _i \varphi _j + \frac{\partial \varphi _i}{\partial x} \frac{\partial \varphi _j}{\partial x} + \frac{\partial \varphi _i}{\partial y} \frac{\partial \varphi _j}{\partial y} \right] \, dx \, dy \]

Equation (23) poses even stronger smoothness restrictions on the function \( f \) since it contains the gradient of the function which is smaller for smoother functions (Dermanis, 1982).

3. PREDICTION

The next step after interpolation is the prediction of function values

\[ (25) \quad s_k' = \hat{f}(x_k', y_k') \]

at \( q \) new points \( P_k' \) with corresponding coordinates \((x_k', y_k')\). From equation (20) the predicted values are
\[ s'_1 = \varphi_1(x'_1, y'_1) \hat{a}_1 + \varphi_2(x'_1, y'_1) \hat{a}_2 + \ldots + \varphi_m(x'_1, y'_1) \hat{a}_m \]

\[ s'_2 = \varphi_1(x'_2, y'_2) \hat{a}_1 + \varphi_2(x'_2, y'_2) \hat{a}_2 + \ldots + \varphi_m(x'_2, y'_2) \hat{a}_m \]

\[ \vdots \]

\[ \vdots \]

\[ s'_q = \varphi_1(x'_q, y'_q) \hat{a}_1 + \varphi_2(x'_q, y'_q) \hat{a}_2 + \ldots + \varphi_m(x'_q, y'_q) \hat{a}_m \]

or in matrix form

\[ s' = F' \hat{a} \]

where

\[ F'_{ij} = \varphi_j(x'_i, y'_i) \]

Combination of equation (18) with equation (27) leads to the direct prediction formula

\[ s' = F'R^{-1}F^T (F R^{-1}F^T)^{-1} s \]

In order to throw more light on this formula we shall consider the simple case where the matrix \( R \) is diagonal. The use of non-diagonal \( R \) leads to the same results but makes the derivations more lengthy and tedious.

Since

\[ R_{ij} = \begin{cases} 0 & i \neq j \\ R_{ii} & i = j \end{cases} \]

and taking into account equations (9) and (28) we arrive by direct computations to the \( i^{th} \) row and \( j^{th} \) column elements of the matrices \( FR^{-1}F^T \) and \( F'R^{-1}F^T \)

\[ (F R^{-1}F^T)_{ij} = \sum_{k=1}^{m} \frac{1}{R_{kk}} \varphi_k(x'_i, y'_i) \varphi_k(x'_j, y'_j) \]

\[ (F'R^{-1}F^T)_{ij} = \sum_{k=1}^{m} \frac{1}{R_{kk}} \varphi_k(x'_i, y'_i) \varphi_k(x'_j, y'_j) \]
Introducing the auxiliary two-point function,

\[(33) \quad k(P, Q) = k(x_P, y_P; x_Q, y_Q) = \sum_{k=1}^{m} \frac{1}{R_{kk}} \varphi_k(x_P, y_P) \varphi_k(x_Q, y_Q) = \sum_{k=1}^{m} \frac{1}{R_{kk}} \varphi_k(P) \varphi_k(Q)\]

the prediction formula (29) can be written

\[(34) \quad s' = K_s's \quad K_{ss}^{-1} s\]

where the matrices $K_{ss}$, $K_s's$ have elements

\[(35) \quad (K_{ss})_{ij} = k(p_i, p_j)\]
\[(36) \quad (K_s's)_{ij} = k(p_i', p_j')\]

In order to give the function $k(P, Q)$ a name, we note that for diagonal $R$ the least squares principle of equation (17) becomes

\[(37) \quad a^T R a = \sum_{k=1}^{m} R_{kk} a_k^2 = \min\]

The quantity $R_{kk}$ is the weight of the coefficient $a_k$ and in analogy with the weights of observational errors in least squares adjustment of observations, we may set

\[(38) \quad R_{kk} = \frac{1}{\sigma_k^2}\]

and call $\sigma_k^2$ the "variance" of the coefficient $a_k$. This is only a formal name for $\sigma_k^2$ and it does not necessarily imply that the coefficient $a_k$ is a random variable as in the case of observational errors. Writing

\[(39) \quad f(P) = \sum_k \varphi_k(P) a_k\]

it follows easily that
\( (40) \quad \varphi_k(p) = \frac{af(p)}{a_k} \)

Replacing from (38) and (40) into equation (33) we have

\( (41) \quad k(p,q) = \sum_{i=1}^{m} \frac{af(p)}{a_{k_i}} \frac{af(q)}{a_{k_i}} \sigma^2_k \)

Equation (41) has a striking resemblance to the well known law of propagation of covariances if \( k(p,q) \) is seen as the covariance between the function values \( f(p) \) and \( f(q) \). For this reason we formally call \( k(p,q) \) covariance function. Then the matrix \( K_{SS} \) contains the covariances between the known function values at the observation points \( P_i \), while the matrix \( K_{S1S} \) contains covariances between values at prediction points \( P_i \) and observation points \( P_j \).

From equation (33) follows that the number \( m \) of the coefficients \( a_k \) and the functions \( \varphi_k \) can increase to infinity, provided that the series

\( (42) \quad \sum_{k=1}^{\infty} \sigma_k^2 \varphi_k(p) \varphi_k(q) \)

converges for all pairs of points \( P, Q \), at least within a certain plane area of interest.

A significant property that must be imposed on minimum norm interpolation and prediction is invariance with respect to the choice of the reference frame to which the cartesian coordinates \( x, y \) refer. Since the choice of reference frame is quite arbitrary, the predicted values must remain unchanged when the frame undergoes arbitrary translations and rotations. The desirable properties of prediction are homogeneity and isotropy referring to invariance under translations and rotations respectively.

A different way of looking upon these properties is to hold the reference frame fixed and let the whole configuration of observation and prediction points undergo translations and rotations. Thus, the predicted values must, in addition to the original data, depend only on the configuration of the network of observation and prediction points and not on its position and orientation.

Homogeneity and isotropy are ensured when the elements of the matrices
\( K_{SS}, K_{S,S} \) or equivalently the covariance function \( k(P,Q) \), are invariant under translations and rotations, i.e., when they depend only on the distances between points. In this case

\[
(43) \quad k(F,Q) = k(r_{PQ})
\]

\( r_{PQ} \) being the distance

\[
(44) \quad r_{PQ} = \sqrt{(x_p-x_Q)^2 + (y_p-y_Q)^2}
\]

between the points \( P \) and \( Q \).

Another property that the function \( k(P,Q) \) must have is to be positive-definite. This means that for any set of observation points \( P_i, i=1,2,..,n \) the symmetric matrix \( K_{SS} \) must always be positive-definite and thus invertible.

It is extremely difficult to find functions \( \phi_k \) and constants \( \sigma_k^2 \) such that the finite sum in equation (33) or the infinite sum (42) has the above properties of invariance and positive-definiteness. Note that the choices of functions in equations (3) and (4) fail in this respect. Instead, one can directly choose a two-point covariance function \( k(r) \) having the required properties. This function corresponds to a set of infinite functions \( \phi_k \) and constants \( \sigma_k^2 \) which remain unknown. A typical choice is the use of exponential functions

\[
(45) \quad k(r) = k_0 e^{-\sigma^2 r^2} = k_0 2^{-(r^2/R^2)}
\]

where

\[
(46) \quad k_0 = k(0)
\]

is the variance and \( R \) the correlation length, such that

\[
(47) \quad k(R) = \frac{1}{2} k(0)
\]
The prediction algorithm of equation (34) can also be seen from a completely different philosophy than the strictly deterministic approach given here. According to this philosophy which we shall call stochastic, the unknown function $f(x,y)$ is viewed as a random function (the term stochastic process is more widely used). Its values at various points are considered as random variables with zero mean and known variances and covariances. The term covariance function for $k(P,Q)$ is taken literally, while in the previous deterministic approach it serves only as a convenient label attached to it.

Equation (34) can be derived by trying to express $s'$ as a linear function of $s$ in such a way, that the variance of any predicted value, i.e., of any element of $s'$ is minimized. This linear minimum variance prediction of the outcomes of random variables $s'$ from the known outcomes of random variables $s$ correlated to the former, is called least squares prediction within the theory of stochastic processes.

In addition for this particular stochastic interpretation, the covariance matrix $K_{s's'}$ of $s'$ can be computed applying the law of propagation of covariances on equation (34)

\[(48) \quad K_{s's'} = (K_{s's} K_{ss}^{-1}) K_{ss} (K_{s's} K_{ss}^{-1})^T = K_{s's} K_{ss}^{-1} K_{s's}^T\]

The opponents of this stochastic interpretation argue that the "true" function $f(x,y)$ is just one concrete function and not an outcome out of a collection of possible functions as the classical probabilistic interpretation of random variables and random functions requires.

A counterargument is based on the concept of ergodicity: Although there is only one function, its values at different parts of the $(x,y)$-plane may take the place of the many outcomes. Expectation, i.e., average over all possible outcomes, is replaced by the process of taking the average over the domain of definition of the function, in our case the plane. The homogenous and isotropic covariance function is now denoted by $C(r)$ and is defined as

\[(49) \quad C(r) = M\left\{ \int_{\theta=0}^{2\pi} f(x,y) f(x+r \cos \theta, y+r \sin \theta) \, d\theta \right\} \]
where \( M(\cdot) \) denotes the process of taking the average over the whole plane.

\[
(50) \quad M(\cdot) = \lim_{a \to \infty} \frac{1}{a^2} \int_{x=-\frac{a}{2}}^{\frac{a}{2}} \int_{y=-\frac{a}{2}}^{\frac{a}{2}} \{ \cdot \} \, dx \, dy
\]

The average in equation (49) is taken over all pairs of points on the plane having the given distance \( r \).

The covariance function (49) can be used in the prediction algorithm (34) instead of any arbitrary homogenous, isotropic and positive definite function \( k(r) \), without resorting to a stochastic interpretation (Sansó, 1978). For this purpose in addition to the properties of linearity and invariance under translations and rotations the additional property of minimum mean square error is imposed on the prediction. If \( \hat{f}(x,y) \) is the predicted function value at some point \((x,y)\) and \( f(x,y) \) the unknown true value, the prediction must be such that the prediction error

\[
(51) \quad \varepsilon(x,y) = \hat{f}(x,y) - f(x,y)
\]

has an average of its square on the whole \((x,y)\)-plane which is as small as possible.

\[
(52) \quad M[\varepsilon(x,y)^2] = \min
\]

In addition to the problems arising from the appearance of the infinite sign in equation (50), there is a great difficulty concerning the use of the covariance function \( C(r) \). Equation (49) tells us that in order to compute the covariance function \( C(r) \) one must in fact know the function \( f(x,y) \). On the other hand if the function \( f(x,y) \) is known one does not need to predict, and much less to compute the covariance function!

Nevertheless, equation (49) can still be usefull. It leads to the computation of sample covariances utilizing the known values of \( f(x,y) \) at the observation points. Separating the reals in intervals of equal length

\[
(53) \quad k \Delta r \leq r < (k+1) \Delta r , \quad k=0,1,...
\]
and computing the average value of all products $f(P_i) f(P_j)$ for points whose distance $r_{ij}$ lies within each particular interval, a "covariance histogram" is constructed. A model covariance such as the exponential one of equation (45) may then be used by selecting the parameters $k_0$, $R$ in a way that it best fits the corresponding histogram.

Such a sample covariance gives an approximation to the minimum mean square error prediction of Sansó (1978), or to the minimum variance (least squares) prediction, if one prefers the stochastic over the deterministic interpretation.

All interpretations lead to the same results when the same covariance function is used. In some cases the choice of covariance function, within reasonable limits of course, has very little influence on the results of the prediction.

Perhaps there is a much simpler way of looking upon these things, especially when the point of view of the everyday practitioner takes over the point of view of the reluctant and trouble-minded theoreteician. To do this we return to the original idea about the smoothness of the interpolated function. For those natural phaenomena that have a smooth variation, we can express smoothness by saying that close points have highly correlated values, while distant points have values with little or no correlation at all.

Suppose that the height at a certain point is given and one asks about the possible values at a point 100 m away, as well as, another point 100 km away. For the nearby point it is easy to answer that the height is very similar to the known height, while for the distant point no guess can be made at all. For some other phaenomena, such as temperature, one may expect similar values even at points several kilometers apart.

This a priori knowledge of the degree of interdependence of function values at different points comes from previous experience about the variation and smoothness of the relevant function. Within probability theory this a priori knowledge is expressed by the concept covariance and correlation. In simple terms, the correlation function

\begin{equation}
-1 \leq \rho(r) = \frac{C(r)}{C(0)} \leq +1
\end{equation}
expresses the likeliness of function values as a function of the distance of the points. The variance \( C(0) \) expresses the variation of the function. Covariance is a combination of correlation and variance.

Here correlation and variance do not refer to means over an ensemble of functions, but over function values at different points of the plane.

5. ADJUSTMENT

Up to now we have assumed that the original quantities \( s \) depending on the unknown function \( f(x,y) \), the so-called signals, are exactly known, e.g., from error-free observations. In many situations one observes instead a vector of quantities \( y \) which are functions not only of signals \( s \) but also of some individual unknown parameters \( x \). Furthermore the observations are corrupted by observational errors \( v \). After linearization the observation equations take the form

\[
(55) \quad y = Ax + Gs + v
\]

where \( A \) and \( G \) are known coefficient matrices.

There are two ways of obtaining estimates for \( x, s \) and \( v \). The first is to ignore the dependance of the signals \( s \) on an underlying function and to treat them in a similar way as the parameters \( x \). In this case we have a classical adjustment problem

\[
(56) \quad y = \begin{bmatrix} A & G \end{bmatrix} \begin{bmatrix} x \\ s \end{bmatrix} + v, \quad v^T P v = \min
\]

where \( P = \Sigma^{-1} \) is the weight matrix of the observations, \( \Sigma \) the corresponding covariance matrix and the solution is

\[
(57) \quad \begin{bmatrix} \hat{x} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} A^T P A & A^T P G \\ G^T P A & G^T P G \end{bmatrix}^{-1} \begin{bmatrix} A^T P y \\ G^T P y \end{bmatrix}
\]

This approach is possible only when the number and the informational content
of the observations $y$ allows the estimation of both $x$ and $s$.

The second approach is to take into account the dependance of the signals on the underlying function. Since the signals have a covariance matrix $K$ just like the errors $v$ have a covariance matrix $\Sigma$, we can treat signals in a similar way as the errors (for simplicity we write $K$ instead of $K_{ss}$). In this case we have an adjustment problem

\[
y = A x + \begin{bmatrix} G & I \end{bmatrix} \begin{bmatrix} s \\ v \end{bmatrix}, \quad s^T K^{-1} s + v^T \Sigma^{-1} v = \min
\]

with solution

\[
M = G K G^T + \Sigma
\]

\[
\hat{x} = (A^T M^{-1} A)^{-1} A^T M^{-1} y
\]

\[
\hat{s} = K G^T M^{-1} (y - A \hat{x})
\]

\[
\hat{v} = \Sigma M^{-1} (y - A \hat{x})
\]

The adjusted signals $s$ can be used in both cases of equations (56) or (58), as the known signals in the prediction algorithm

\[
\hat{s}' = K_{ss}^{-1} \hat{s} = K_{ss}^{-1} K_{ss} \hat{s}
\]

The most simple case of the type of equation (58) is the case where signals are directly observed in the presence of observational noise

\[
y = s + v
\]

The adjusted signals are

\[
\hat{s} = K (K + \Sigma)^{-1} y
\]

and prediction according to equation (63) gives
(66) \[ \hat{s}' = K_{s's} (K_{ss} + \Sigma)^{-1} y \]

This algorithm is usually referred to as smoothing collocation. In fact it is a combination of adjustment and prediction by exact collocation.

The algorithm of smoothing collocation can also be derived in a purely deterministic way. Consider again the interpolation of the type of equation (2) and allow for non-exact solutions also, independently of the relative magnitude of n and m. Instead of the minimization problem of equation (17) or (14), a combination can be expressed as

(67) \[ s = F a + v , \quad v^T P v + \lambda (a^T R a) \]

This hybrid approach relates directly to the regularization method for solving improperly posed problems (Tikhonov and Arsenin, 1977). \( \lambda \) is a positive scalar called the regularization parameter. The solution is

(68) \[
\hat{a} = (\lambda R)^{-1} F^T \left[ F (\lambda R)^{-1} F^T + P^{-1} \right]^{-1} s = \\
= R^{-1} F^T (F R^{-1} F^T + \lambda \Sigma)^{-1} s
\]

and the prediction formula becomes

(69) \[ s' = F' \hat{a} = F' R^{-1} F^T (F R^{-1} F^T + \lambda \Sigma)^{-1} s \]

Introducing the symbols

(70) \[ K_{ss} = F R^{-1} F^T \]

(71) \[ K_{s's} = F' R^{-1} F^T \]

the prediction in this case of smoothing interpolation is given by

(72) \[ s' = K_{s's} (K_{ss} + \lambda \Sigma)^{-1} s \]

For \( \lambda=1 \) equation (72) becomes similar to equation (66) used in smoothing collocation. For \( m>n \) and \( \lambda \to 0 \) it degenerates to the case of exact
interpolation and prediction.

6. EXTENSIONS

In the previous discussions we have limited ourselves without loss of generallity to the case where both the originally known signals s as well as the predicted signals s' are values of the underlying function at certain points. In practice signals of a more complicated dependance on the unknown function arise.

For example, consider the case where the unknown function is the uplift function in repeated levellings. If a tiltmeter has been placed at some point of the area along the x (east-west) direction, the quantity observed is the change between the original terrain inclination $\frac{\partial h}{\partial x}$ along x, and the same quantity $\frac{\partial h'}{\partial x}$ at the second epoch. The difference of these inclinations is the derivative of the uplift function of equation (1) according to

$$\frac{\partial u}{\partial x} = \frac{\partial h'}{\partial x} - \frac{\partial h}{\partial x}$$

(73)

The coefficients $a_i$ of the interpolating function of equation (2) must also fulfill the equation

$$\frac{\partial f}{\partial x}(P) = \sum_{i=1}^{m} \frac{\partial \varphi_i}{\partial x}(P) a_i$$

(74)

where P is the tiltmeter point. The row of matrix F which corresponds to the above equation, say the $k^{th}$ row, has the form

$$\left[ \frac{\partial \varphi_1}{\partial x}(P), \frac{\partial \varphi_2}{\partial x}(P), \ldots, \frac{\partial \varphi_m}{\partial x}(P) \right]$$

(75)

Let the $l^{th}$ row of the same matrix correspond to a simple uplift observation at a point Q. Its form will be

$$\left[ \varphi_1(Q), \varphi_2(Q), \ldots, \varphi_m(Q) \right]$$

(76)

For diagonal matrix R with elements as in equation (38) we have that
\[(77) \quad (K_{ss})_{k1} = (F R^{-1} F^T)_{k1} = \frac{\sigma_i^2}{\sigma_i} \varphi_i(Q) \varphi_i(P) = \frac{\partial}{\partial x} \left[ \sum_{i=1}^{m} \sigma_i^2 \varphi_i(P) \varphi_i(Q) \right] = \frac{\partial}{\partial x} k(P,Q) \]

Since the function \(k(P,Q)\) has been called the covariance between the values \(f(P)\) and \(f(Q)\), we can use the symbol \(\sigma\) for covariance and to express this as

\[(78) \quad k(P,Q) = \sigma(f(P),f(Q)) \]

We denote by \((x,y)\) the coordinates of \(P\) and by \((x',y')\) those of \(Q\). Thus, the element of the matrix \(K_{ss}\) in question is written

\[(79) \quad (K_{ss})_{k1} = \sigma(\frac{\partial f}{\partial x}(P),f(Q)) = \frac{\partial}{\partial x} \sigma(f(P),f(Q)) \]

according to the so-called covariance propagation law.

If the observation at point \(Q\) is also a tiltmeter observation along the \(y\) (north-east) direction, the \(l^{th}\) row of \(F\) will be of the form

\[(80) \quad \left[ \frac{\partial \varphi_1}{\partial y}(Q) \quad \frac{\partial \varphi_2}{\partial y}(Q) \quad \ldots \quad \frac{\partial \varphi_l}{\partial y}(Q) \right] \]

and the corresponding element of the covariance matrix \(K_{ss}\) will be

\[(81) \quad (K_{ss})_{k1} = \frac{\sigma_i^2}{\sigma_i} \frac{\partial \varphi_i}{\partial x}(P) \frac{\partial \varphi_i}{\partial y}(Q) = \frac{\partial}{\partial x} \frac{\partial}{\partial y} \sigma(f(P),f(Q)) = \sigma(\frac{\partial f}{\partial x}(P),\frac{\partial f}{\partial y}(Q)) \]

This covariance propagation law allows us to compute covariances for various types of signals from the original covariance function \(k(P,Q)\). In order to express this law in a more general way, suppose that two signals \(s_i, s_j\) depend on the unknown function \(f\) according to

\[(82) \quad s_i = L_i(f) \]

\[(83) \quad s_j = L_j(f) \]
\( \sigma(s_i, s_j) = L_i \{ L_j \sigma(f^\top, f(Q)) \} = L_i \{ L_j \ k(P, Q) \} \)

7. TREND REMOVAL

The deterministic minimization condition \( a^T R a = \text{minimum imposed on} \) the interpolating function \( f(x,y) \) tends to make the coefficients \( a \) and the function itself, as close to zero as possible. However the real function may not be a "small" function. If an approximation \( f_o(x,y) \) of \( f(x,y) \) is known one should try to make the interpolation \( \hat{f}(x,y) \) as close as possible to \( f_o(x,y) \). This is achieved by working with the reduced function

\( \delta f(x,y) = f(x,y) - f_o(x,y) \)

instead of \( f(x,y) \) and the reduced signals

\( \delta s_i = \delta f(x_i, y_i) = s_i - f_o(x_i, y_i) \)

instead of the original signals \( s_i \).

When a model function \( f_o(x,y) \) is not a priori known an approximation to it can be obtained by a process called trend removal.

One way is to estimate and remove the trend a priori by means of an interpolation of the type of equation (2) with a very small number of terms \( m \) in which case \( m < n \). This leads to a least squares problem like in eq. (14)

\( s = Fa + e, \quad e^T We = \min \)
with solution

\[ (88) \quad \hat{a} = (F^T W F)^{-1} F^T W s \]

The interpolation is next carried out using the residual signals

\[ (89) \quad \delta s = s - F \hat{a} \]

Application of the prediction algorithm using a covariance function for the residual function gives the predicted residual signals

\[ (90) \quad \delta s' = K_{s'} s K_{s}^{-1} \delta s \]

in case of exact collocation, or

\[ (91) \quad \delta s' = K_{s'} s \left( K_{s} + \Sigma \right)^{-1} \delta s \]

in case of smoothing collocation. The total predicted signals are

\[ (92) \quad s' = F' \hat{a} + \delta s' \]

The most simple detrending of this type is the removal of the mean of the available function values. The next choice is to fit a linear function to the available values, thus determining a best fitting plane.

A somewhat different approach to the problem of trend removal is to recognize from the beginning the fact that the deviations \(\bar{e}\) in equation (87) are the sum of signals \(\delta s\) and observational errors \(v\) with covariance matrices \(K\) and \(\Sigma = P^{-1}\) respectively. In this respect we formulate the minimization problem

\[ (93) \quad s = F a + \delta s + v \quad , \quad \delta s^T K^{-1} \delta s + v^T \Sigma^{-1} v = \min \]

This means that the observed function values \(s\) are the sum of a trend \(Fa\), a term of residual signals \(\delta s\) and observational errors \(v\). We have thus an adjustment problem similar to the one formulated in equation (58). The solution is
\[ M = K + \Sigma \]  
\[ \hat{a} = (F^T M^{-1} F)^{-1} F^T M^{-1} s \]  
\[ \hat{\delta} s = K M^{-1} (s - F \hat{a}) \]  
\[ \hat{\nu} = \Sigma M^{-1} (s - F \hat{a}) \]

Prediction of new residual signals gives

\[ \hat{\delta} s' = K_{s'ss} K_{ss}^{-1} \hat{\delta} s \]

and the total predicted new signals are

\[ \hat{s}' = F' \hat{a} + \hat{\delta} s' \]

In order to compare this result with that of equation (92) we combine equations (88), (89), (91) and (92) to obtain (recall that \( K = K_{ss} \))

\[ s' = \left[ F' - K_{s'ss}(K+\Sigma)^{-1} F \right] (F^T W F)^{-1} F^T W s + K_{s'ss}(K+\Sigma)^{-1} s \]

We also combine equations (94), (95), (96), (98) and (99) to obtain

\[ \hat{s}' = \left[ F' - K_{s'ss}(K+\Sigma)^{-1} F \right] (F^T M^{-1} F)^{-1} F^T M^{-1} s + K_{s'ss}(K+\Sigma)^{-1} s \]

The results in both cases will be identical if instead of an arbitrary weight matrix \( W \) in the first case we choose

\[ W = M^{-1} = (K + \Sigma)^{-1} \]

or if we use as covariance matrix of \( e \) the covariance matrix

\[ W^{-1} = M = K + \Sigma \]

which follows from straightforward covariance propagation in the relation
(102) \( \mathbf{e} = \bar{\mathbf{d}} \mathbf{s} + \mathbf{v} \)

8. APPLICATIONS

We shall give two surveying applications of the above mentioned collocation techniques.

The first one refers to the comparison of two configurations of the same network at two different epochs for the detection of crustal deformation parameters.

Let \( (x_i, y_i) \) and \( (x'_i, y'_i) \) denote the coordinates of network point \( P_i \) at the two epochs, while

\[
(103) \quad u_i = x'_i - x_i
\]
\[
(104) \quad v_i = y'_i - y_i
\]

are the displacement components at the same point.

In case some network points (at least two) are considered to be motionless either on the basis of geophysical evidence or as a result of statistical tests (Koch and Fritsch, 1981), the network configurations are said to be connected in the sense that the second one cannot undergo any translations, rotations and scale changes independently of the first. If this is not the case, as a consequence of the arbitrary frame of reference definitions at the two epochs, the second configuration can undergo rigid transformations when both epochs include distance measurements, or similarity transformations otherwise, with respect to the first configuration.

It is necessary to connect the two configurations in an optimal and clearly defined way if the derived crustal deformation parameters are to be of any intrinsic value. This problem of connection can be seen from the viewpoint of trend removal from the displacements \( u_i, v_i \). In this respect one determines the values of the parameters of the rigid or similarity transformation which transforms the coordinates \( (x'_i, y'_i) \) into coordinates
(x_i'''', y_i''') in such a way that the remaining displacements x_i''' - x_i, y_i''' - y_i satisfy the best fitting principle

\[(105) \quad \sum_i \{(x_i''' - x_i)^2 + (y_i''' - y_i)^2\} = \text{min}\]

We assume from now on that the displacement components at hand u_i, v_i refer either to connected configurations or they have their rigid or similarity transformation trend removed as described above.

The quantities of most interest for the study of crustal motion are the dilatation

\[(106) \quad \Delta = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\]

and especially the maximum shear strain

\[(107) \quad \nu = \sqrt{\nu_1^2 + \nu_2^2}\]

where

\[(108) \quad \nu_1 = \frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}\]

\[(109) \quad \nu_2 = \frac{\partial u}{\partial y} - \frac{\partial v}{\partial x}\]

For the computation of the above parameters at any desired point the values of the four derivatives \(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial v}{\partial x}, \frac{\partial v}{\partial y}\), at the same point must be predicted. Denoting

\[(110) \quad \mathbf{u} = [u_1 \ u_2 \ \ldots \ u_n]^T\]

\[(111) \quad \mathbf{v} = [v_1 \ v_2 \ \ldots \ v_n]^T\]

and assuming that the u and v displacements at the same or different points are uncorrelated, the prediction formulas are

\[(112) \quad \frac{\partial u}{\partial x}(p) = K_{ux} K_u^{-1} u\]
\[
\frac{\partial u}{\partial y}(p) = K_{uy} K_u^{-1} u \\
\frac{\partial v}{\partial x}(p) = K_{vx} K_v^{-1} v \\
\frac{\partial v}{\partial y}(p) = K_{vy} K_v^{-1} v
\]

The \( n \times n \) matrices \( K_u, K_v \) have elements

\[
(K_u)_{ij} = k_u(P_i, P_j) = \sigma(u(P_i), u(P_j))
\]

\[
(K_v)_{ij} = k_v(P_i, P_j) = \sigma(v(P_i), v(P_j))
\]

\( k_u, k_v \) being selected covariance functions (usually \( k_u = k_v \)).

The \( 1 \times n \) matrices \( K_{ux}, K_{uy}, K_{vx}, K_{vy} \) have elements

\[
(K_{ux})_i = \frac{\partial}{\partial x} k_u(P_i, P_i) = \sigma(\frac{\partial u}{\partial x}, u(P_i))
\]

\[
(K_{uy})_i = \frac{\partial}{\partial y} k_u(P_i, P_i) = \sigma(\frac{\partial u}{\partial x}, u(P_i))
\]

\[
(K_{vx})_i = \frac{\partial}{\partial x} k_v(P_i, P_i) = \sigma(\frac{\partial v}{\partial x}, v(P_i))
\]

\[
(K_{vy})_i = \frac{\partial}{\partial y} k_v(P_i, P_i) = \sigma(\frac{\partial v}{\partial y}, v(P_i))
\]

In this approach collocation is simply used as a means for exact interpolation. Errors in the displacements \( u_i, v_i \) described by covariance matrices \( \Sigma_u, \Sigma_v, \Sigma_{uv} \) which follow from the covariances of the independently adjusted \((x_i, y_i), (x'_i, y'_i)\) coordinates are not taken into account. The displacements \( u_i, v_i \) are determined as differences of the adjusted coordinates at the two epochs. They are equivalent to those that will be obtained if the observations at the two epochs were adjusted simultaneously using again \((x_i, y_i)\) but \((u_i, v_i)\) instead of \((x'_i, y'_i)\) as parameters. Then the signals \((u_i, v_i)\) are treated as free parameters according to equation (56).

Let the observation equations at the two epochs be
\[ l_1 = A_1 x + v_1 \]
\[ l_2 = A_2 x' + v_2 \]

where \( x \) contains the \((x_i, y_i)\) coordinates and \( x' \) the \((x'_i, y'_i)\) ones. Setting
\[ x' = x + s \]

where \( s \) contains the displacements \((u_i, v_i)\), equation (56) becomes in this case
\[
\begin{bmatrix}
  l_1 \\
  l_2
\end{bmatrix}
\begin{bmatrix}
  A_1 & 0 \\
  A_2 & A_2
\end{bmatrix}
\begin{bmatrix}
  x \\
  s
\end{bmatrix}
\begin{bmatrix}
  v_1 \\
  v_2
\end{bmatrix}
\begin{bmatrix}
  P_1 & v_1 \\
  P_2 & v_2
\end{bmatrix}
\begin{bmatrix}
  P_1 & v_1 \\
  P_2 & v_2
\end{bmatrix} = \min
\]

When the covariances functions \( k_u(P, Q) \) and \( k_v(P, Q) \) have a stochastic interpretation, the signals \( s \) can be treated according to equation (58) using the covariance matrix \( K \) of \( s \). In this case equation (58) becomes
\[
\begin{bmatrix}
  l_1 \\
  l_2
\end{bmatrix}
\begin{bmatrix}
  A_1 & 0 \\
  A_2 & A_2
\end{bmatrix}
\begin{bmatrix}
  x \\
  s
\end{bmatrix}
\begin{bmatrix}
  v_1 \\
  v_2
\end{bmatrix}
\begin{bmatrix}
  P_1 & v_1 \\
  P_2 & v_2
\end{bmatrix}
\begin{bmatrix}
  P_1 & v_1 \\
  P_2 & v_2
\end{bmatrix} = \min
\]

The results in this approach are different not only for the adjusted displacements \( s \), but also for the coordinates \( x \) and \( x' = x + s \), when they are compared to those coming from independent network adjustments at the two epochs.

The term \( s^TK^{-1}s \) in the quantity to be minimized in equation (126) does not allow very large displacements \( s \), even if these give small errors \( v_1 \), \( v_2 \). The observational inconsistencies are not absorbed by the error estimates only, but they are distributed between them and the displacements.

Many people would strongly object to permitting the ambiguous covariance functions \( k_u(P, Q) \), \( k_v(P, Q) \) influence the results of the independent adjustments at the two epochs.
A second application quite similar to the first one, refers to the case of network densification.

An independently adjusted network includes points common with an existing network. For these points there exist two sets of coordinates with differences \((u_i, v_i)\). A similarity or rigid transformation on the densification network coordinates removes much of the inconsistencies of the common points, but still coordinate differences remain.

One solution is to readjust the new network holding the coordinates of all common points fixed to their old values.

An alternative method, which is much easier from the computational point of view, is to consider the remaining coordinate differences \((u_i, v_i)\) as systematic distortions and predict similar distortions for all the other points of the densification network with the help of the collocation prediction algorithm of equation (34).

In this way, the vectors \((u_i, v_i)\) at the common points are interpolated in a smooth way at any other point, so that distortions between neighbouring points are very small. The prediction can be done not only for the remaining points of the densification network, but also for any point in the area, such as secondary points established by traverses, tacheometry, etc.

This technique is extremely useful in a situation where the higher order network coordinates are updated, e.g., with the help of new observations with refined instrumentation. The readjustment of existing densification networks would also require the recomputation of traverses, tacheometry, etc. The prediction of a smooth field of coordinate corrections is much easier since it mainly requires the inversion of the covariance matrix \(K_{ss}\) of the coordinate differences at the common points. Usually the number of common points is small, the dimensions of \(K_{ss}\) are consequently small and it can be therefore easily inverted.
REFERENCES


Moritz, H. (1972): Advanced Least-Squares Methods. The Ohio State University, Dept. of Geodetic Science, Report no. 175, Columbus, Ohio.


TECNICHE MODERNE DI ANALISI DEI DATI GEODETICI CON PARTICOLARE RIGUARDO ALLA COLLOCAZIONE

Contributi di:

R. Berzaghi — F. Sansò — A. Dermanis
D. Vlachos — E. Livieratos

A cura di M. Unguendoli

Seminario dell’Istituto di Topografia, Geodesia e Geofisica Mineraria – Università di Bologna
(20-21 settembre 1983)