The Use of Optimal Estimation for Gross-error Detection in Databases of Spatially Correlated Data.

by

C.C. Tscherning, Geophysical Institute,
University of Copenhagen, Haraldsgade 6,
Dk-2200 Copenhagen N., Denmark

Abstract: When establishing and updating a data base it is necessary to identify - and possibly correct - gross errors. For large numerical data bases, semi-automatic methods, which identify suspected gross errors, are used, followed by a more detailed analysis of the individual values.
For data, which are associated with a spatial position (location), it is very often so that data are spatially correlated. The distance plays the role time does in time series, while the directional dependence often is small or may be disregarded.
This may be used to detect gross errors, using tools developed for optimal estimation in stohastic processes. A new item to be loaded into the data base is first predicted from the data associated with the e.g. the 10 closest points stored in the data base. Here methods like optimal linear prediction (sometimes denoted least squares collocation or Kriging) makes it possible also to estimate the error of prediction. A comparison of the difference between the observed and the predicted value with the error estimate, may then be used to identify a possible gross error.
The success of this procedure depends on whether the statistical properties are homogeneous for the geographical area being considered. If this is not the case, the data must be preprocessed, removing trends and the physical factors causing the inhomogeneity. This has been used for the detection of gross errors in gravity field related data, which in the paper is used as an example to illustrate the method.

1. Introduction

When establishing a (scientific) data base it is important to eliminate or at least flag gross errors. The size of many modern data bases now makes it virtually impossible to check and correct a suspected error. A remeasurement of quantities obtained e.g. by a satellite 10 years ago will in most cases be impossible due to the costs involved.

Data bases seems to obey the same law as computer programs: there is always an error left. Unfortunately, many data sets, in the Earth sciences at least, seems to contain up to 1 % erroneous data, see e.g. Tscherning (1990), Remmer (1984). Despite these errors, excellent scientific results are generally obtained because one single data entity is not used alone, but together with other entities. A simple example is from surveying, where one always will try to observe all three angles in a triangle, even if two are sufficient when determining the shape of the triangle. The data bases will therefore contain redundent data, even if this generally is one of the characteristics used to describe a file system which is not a data base. An important requirement for data bases in many fields of earth sciences is therefore that the data contained <u>are</u> redundant. This is then used when applying least-squares procedures for the determination of associated parameters like the positions of the vertices of a triangle. Even if data are not as clearly redundant as in the case of the triangle, data may be strongly correlated, simply because they are associated with points spatially close to each other. The gravity in two points a few meters apart will be only slightly different, because the attracting masses, as seen from the points, will be nearly identical.

This spatial correlation may be used to detect gross-errors, in the same manner as it is done for time series. If a value differs more than usual from its neighbouring values, it may be a gross-error.

For time series, the similarity of two values $x(t_1)$, $x(t_2)$ is expressed through the auto-covariance function, $C(t)$, which for stationary time series is a function of time difference (t_1-t_2) only. For spatially distributed data, the correlation may be expressed through a covariance function generally only dependent on the distance and the altitude of two points. We will explain this in section 2.

When a covariance function is given, we may qualitatively express what we mean by stating that a value differ more than usual from other values. This is expressed by a comparison of the observed value with the value obtained by optimal prediction (interpolating or extrapolation) from the neighbouring values. In section 3 we give the necessary algorithms.

The method is routinely used to detect gross errors in data bases containing gravity field related data, and in section 4 two examples illustrate this.

2. Covariance function for spatially distributed data

We will in the following discuss spatially distributed data. These data may also vary as a function of time, but we will here keep the time fixed.
For such data the (horizontal) distance plays (or as we shall see may be forced to play) the role time difference does for a time series. The directional (azimuthal) dependence is often small or may be removed, but (as with gravity data) the altitude does play a role, which should not be disregarded.
For a time series, we generally have many repetitions of the same phenomena. This has been used to justify the description of the signal as a stohastic process or random function. For repeated occurrences the covariance may be computed in the usual manner as the mean value of products of observations, observed at the same time by two observers. If only one observer is "active", the products are formed for observations, observed with a constant time difference. The time series is supposed to be stationary.
For spatially distributed data, especially data observed on or outside the surface of the Earth, we have no possibility to have repeated occurrences. We only have one Earth! But as with time-series, repetitions may be introduced artificially. The Earth(or a planet) rotated around its center, is regarded as a new Earth. (It may be enough to rotate around a certain axis, but we will not discuss this here). Data located in the same altitude may therefore be used when estimating the covariance, as a function of the (spherical) distance between pairs of points where observations have taken place.
The estimation of the covariance C, as a function of distance d, is in practice simply done by grouping pairs of data in classes according to distance intervals, and then forming the mean of the products of the pairs,

$$C(d_k) = \left(\sum_{i=1}^{n} x(P_i)\, x(Q_i) \right)/n. \qquad (1)$$

Here P_i and Q_i are points with distance d_i,

$$d_k - \tfrac{1}{2}v \le d_i < d_k + \tfrac{1}{2}v,$$

where v is the magnitude of the interval within which the products are sampled, and d_k the k'th interval midpoints. We have in eq (1) supposed that the mean value has been subtracted.

An example of a gravity autocovariance function is given in Figure 1, computed using data in a local areas (Ohio). Note that the function has both positive and negative values.

The applied procedure is statistically correct, if we may regard the physical phenomena as an isotropic random function (see e.g. Sanso (1986)). But even if it is not correct, we are able to evaluate numerically eq. (1) and obtain a sample of values $C(d_k)$, k = 1, ... m. If data are distributed globally, the covariance function may be expressed as the sum of a Legendre series,

$$C(d) = \sum_{i=1}^{\infty} \sigma_i \, P_i(\cos d) \qquad (2)$$

where d is the distance in radians, P_i are the Legendre-polynomials and σ_i are positive constants called degree-variances. The σ_i express the variation per spherical harmonic degree of the observed quantity. These quantities will not be known, but their values may be expressed as simple functions of the degree. For gravity values simple exponential "rules" have been found, see e.g.Kaula (1959), Rapp (1990). Some of these rules makes it possible to express C(d) by a closed formula. If data are given only in a limited area, a Fourier analysis may be used to obtain the coefficients of the Fourier series. The square of the coefficients (the power-spectrum) are then the coefficients in the planar auto-covariance function similar to eq. (2).

3. Gross-error detection for spatially distributed data

Simple statistical tools, such as the formation of a histogram, are extremely useful when trying to detect gross-errors. Here it may be useful to first low-pass filter the data (trend removal), and then inspect the histogram of the filtered data. For globally distributed data, a spherical harmonic analysis, will produce a function representing the global trend. For regional or local data, the first coefficients in a Fourier expansion or a low degree polynomial in plane coordinates (x, y) will represent the trend.

The filtered data may then be contoured, and gross-errors may then show up as "chimneys" or volcanos on a contour map, see Fig. 2. However, a contour map

may show several chimneys, and it may be very difficult to judge which are important, and which are "normal".
Here the method of optimal linear prediction (Krigning, least-squares collocation) has been used with success, as will be described in the next section. The idea is to interpolate or extrapolate (predict) an observed value from neighbouring values, and then compare the predicted and the observed value. If the difference is larger than for example 3 times the error of the prediction, then the observative is flagged as a suspected gross-error.
Let us denote the observation to be analyzed by y, and the surrounding observations by x_i, i = 1, ..., n. Then linear prediction determines an estimate $\tilde{y}$ by

$$\tilde{y} = \sum_{i=1}^{n} a_i \, x_i , \tag{3}$$

where a_i are unknown constants. We may decide to determine $\tilde{y}$ so the mean square error is as small as possible, where the mean is take over all point configurations, which may be created on a sphere by rotations of the observations points Q, P_i, i = 1,..., n associated with y_{obs}, x_i, i = 1, ..., n, respectively. Then it is easily shown (Moritz, 1980) that

$$\{a_i\} = \{C_{qi}\}^T \{\bar{C}_{ij}\}^{-1}$$
$$\bar{C}_{ij} = C_{ij} + D_{ij} \tag{4}$$

where C_{qi} is the covariance between y and x_i, C_{ij} is the covariance of the observations and D_{ij} the covariance of the observation error associated with x_i and x_j. The mean square error of $\tilde{y}$ will be

$$\sigma^2(y-\tilde{y}) = C_o - \{C_{qi}\}^T \bar{C}^{-1} \{C_{qj}\}, \tag{5}$$

where C_o is the variance of y. If this quantity and the observation error of $\tilde{y}$, σ_y are both χ^2 - distributed, then

$$\sigma^2(y_{obs}-\tilde{y}) = \sigma_y^2 + \sigma^2(y-\tilde{y}) \qquad (6)$$

If the actual difference

$$|y_{obs}-\tilde{y}| > k \cdot \sigma(y_{obs}-\tilde{y}) \qquad (7)$$

then we may decide to regard y as a possible gross-error.
The success of this procedure depends on the statistical homogenity of the data. A good (and actual example) is the situation where data consist of two subsets each associated with a separate unknown linear parameter: Satellite observations, from two different time periods from the same area, affected by different satellite orbit errors, may form two such sets. The above described linear estimation method may be extended to take account of this. Also data do not need to be of the same type, as long as a cross-covariance function is also known, see Moritz (1980).
Other inhomogenieties may have physical causes, such as varying geology or topography. These physical causes may be taken into account (removed), see Forsberg (1986).
The above described procedure is not without pitfalls. A gross-error may "contaminate" neighbouring values. So an iterative procedure is recommended, where the largest suspected gross-errors are removed, and the prediction is repeated for smaller suspected errors.

4. Gross-error detection in gravity field data

A data base with global gravity coverage is managed by Bureau Gravimetrique International (BGI) in Toulouse, France. Here several million data are stored. Data originate from both scientific and commercial sources, and date 100 years back. Many national institutions also administrate gravity data bases.
Gravity is the modules of the gradient of the gravity potential. The Mean sea-level coinsides with a surface where the potential is constant called the geoid. The height of this surface above a reference ellipsoide, plus height variations due to tides, currents etc. may be observed by measuring the distance from a satellite to the sea surface by a radar altimeter. Such data has been collected by several satellites at a rate of 1 per second, and new

satellites will be launched in the comming years (ERS-1, Topex-Poseidon). So, huge gravity field data sets are and will be collected, which are spatially correlated. But the data contain many errors, which must be eliminated or flagged. At BGI, optimal linear prediction is used to mark suspected gross-errors on a CRT (see Fig. 2), which is used to display a contour map of the data. Interactively the suspected errors may be removed, and if smooth contour lines occur after the removal, then the observation is flagged as an error. An experienced analyst can validate 5000-7000 points per day under optimal conditions, (BGI, 1989, p. 110).

The method is also used when analyzing gravity and satellite radar altimeter data, see Tscherning (1990). Here the possibility of also removing biases were used.

The method of least-squares collocation is not restricted to be used with only one datatype. The combined use of altimeter an gravity data improved the power of the method, so that 0.3% further gross errors were detected in addition to the 0.7% error detected using the altimeter data alone.

5. Conclusion

Spatially distributed data are often spatially correlated. A covariance function which primarily is a function of distance may be estimated, by forming mean values of products of data having the same spherical distance. Using optimal linear prediction, the values may be computed from neighbouring values, and compared with the observed value, thereby indicating an error if the difference is larger than a factor k times the prediction error.

The use of the method requires that the observations are "homogenized", by low-pass filterning, and removal of an-isotropies, if possible. It has been used with success for gravity field data, but it should be possible to use the method for many other types of spatial data.

References:

Bureau Gravimetrique International: Workshop on Gravity Data Validation (Review), October 17-19, 1989. Bulletin d'Information, No. 65, pp. 89-142, 1989.

Forsberg, R.: Spectral Properties of the Gravity Field in the Nordic Countries. Boll. Geodesia e Sc. Aff., Vol. XLV, pp. 361-384, 1986.

Moritz, H.: Advanced Physical Geodesy. H. Wichmann Verlag, Karlsruhe, 1980.

Rapp, R.H.: The decay of the spectrum of the gravitational potential and the topography for the Earth. Geophys. J. Int., Vol. 99, pp. 449-455, 1989.

Remmer, O.: Modelling errors in geometric levelling. Proceedings Control Survey Networks, Meeting of (FIG) Study Group 5B, 7-9 July, 1982, Aalborg University Center, pp. 355-372, Hochschule der Bundeswehr Muenchen, 1982.

Remmer, O.: New Error Parameters in Levelling. Report of SSG 1.53 at the IAG General Assembly, Hamburg, August 15-26, 1983.
Traveaux de l'Association Internationale de Geodesie, Tome 27, pp. 116-122, 1984.

Sanso', F.: Statistical methods in physical geodesy. In: Suenkel, H.: Mathematical and Numerical Techniques in Physical Geodesy. Lecture Notes in Earth Sciences, Vol. 7, pp. 49-155, Springer-Verlag, 1986.

Tscherning, C.C.: A strategy for gross-error detection in satellite altimeter data applied in the Baltic-Sea area for enhanced geoid and gravity determination. Presented Nordic geodetic Commission XI meeting, Copenhagen, May, 1990b.

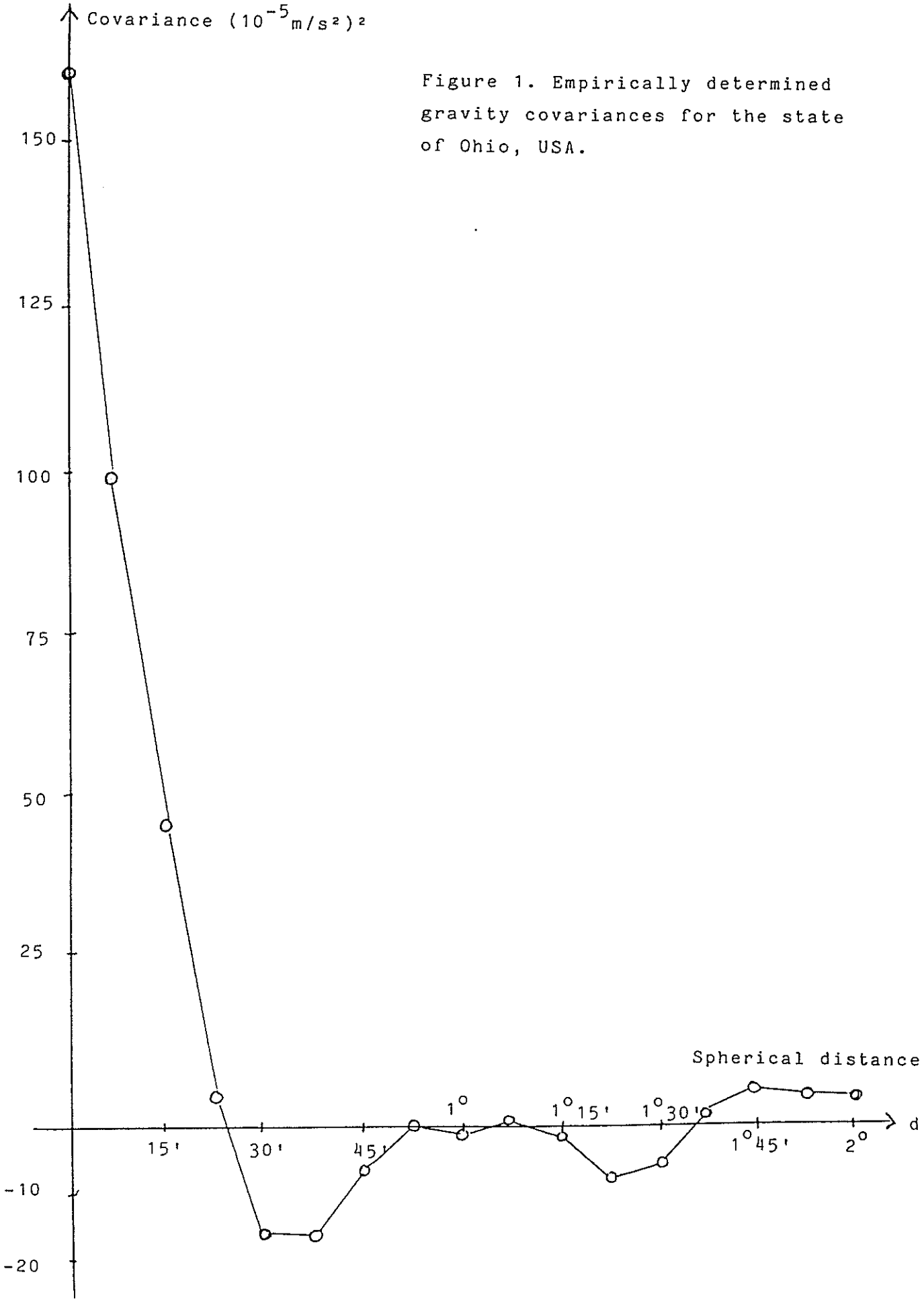

Figure 1. Empirically determined gravity covariances for the state of Ohio, USA.

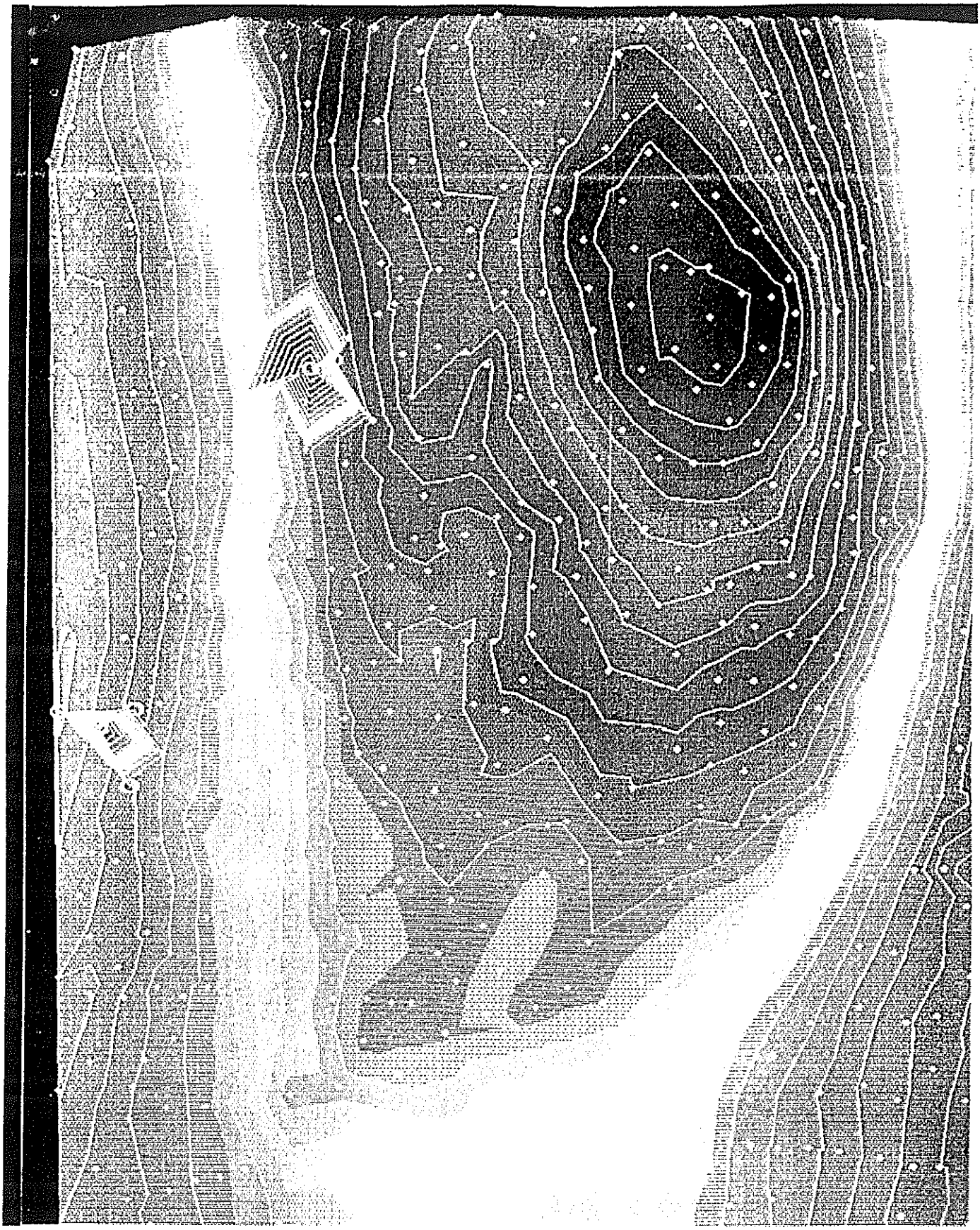

Figure 2: Gravity map with six measurements predicted as suspected grosserrors using collocation (From BGI, 1989, p. 120).

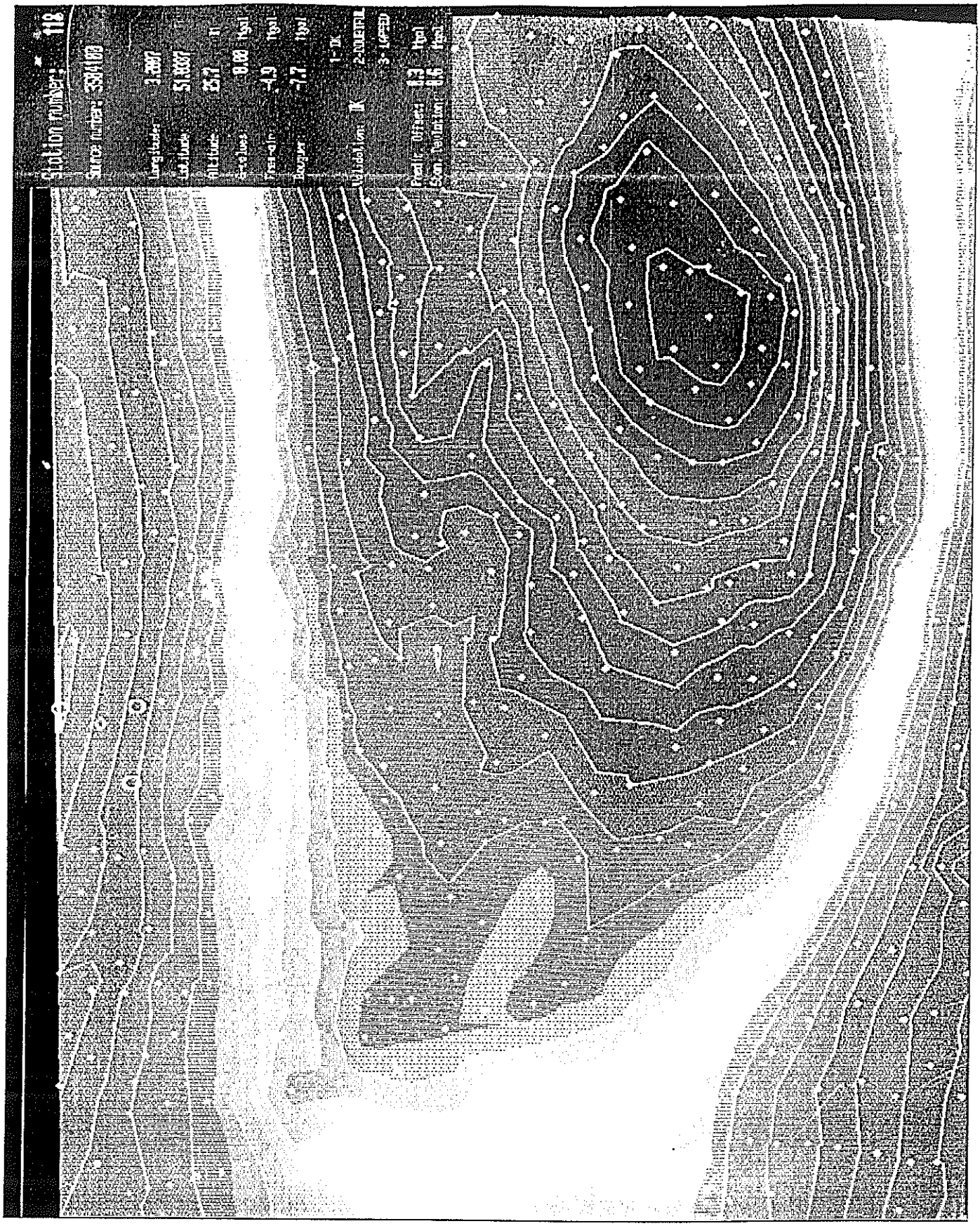

Figure 3: Gravity map, same as Figure 2, but with two largest errors removed (From BGI, 1989, p. 121).