A poly-Gaussian regression model to locate and quantify multiple anomalies in 3D geochemical exploration data

A. I. Burago[1], V. A. Burago[1], N. G. Vlasov[2] and S. Henley*[3]

A non-linear regression method has been developed to fit multiple anomalies in geochemical data obtained from sampling in three dimensions from drill holes, trenches and surface geochemical surveys. A heuristic description of the method is given. Related to trend surface analysis, it uses a set of Gaussian functions (negative exponentials of quadratic functions of coordinates) instead of the polynomial or Fourier series used in classical trend surface analysis. Isosurfaces of Gaussian functions are ellipsoids. Their combinations are capable of giving adequate approximations of much more complicated surfaces. Location, size and form of the anomalies are modelled by including the ellipsoid axes and orientation angles as variables to be fitted. Using lithogeochemical data, the method further allows quantitative estimation of the resources within each anomaly or in other limits of interest, by integration of the fitted function for response values above a defined cutoff concentration. A case study is provided, from exploration geochemical data in the area around the Pokrovsky mine in the Amur region of far eastern Russia.

Keywords: Geochemistry, Exploration, Modelling, Regression, Poly-Gaussian, Non-linear, Three-dimensional

Introduction

Trend surface analysis was one of the earliest methods developed for modelling spatial data. In its original version, a surface defined by a simple polynomial function of the form

$$z(x,y)=a_0+a_1x+a_2y+a_3x^2+a_4xy+a_5y^2+\ldots$$

was fitted to a two-dimensional dataset where x and y are spatial coordinates and z is the geochemical element or other variable being studied. Fitting was carried out by a standard multiple linear regression procedure using the set of functions of x and y (x, y, x^2, xy, y^2, …) as the independent variables. Because of the simplicity of the functions used, trend surface analysis yielded simple surfaces with few inflection points, but nevertheless became widely used, especially for structural modelling, but also in geochemistry and other fields.[8,13,15]

Other functions have also been used: in particular, Fourier series, which allowed more complex surfaces to be modelled for modest additional computing power. However, trend surface analysis was rapidly supplanted by other spatial modelling methods as computing power increased, and as alternative methods were developed. In particular, the development of regionalised variable theory by Matheron[9,10] and its application in the form of kriging,[7] became standard methodology in mineral resource modelling and in many other fields where the spatial distribution of observed variables is studied. Discussion of this approach and further references can readily be found in many publications on spatial statistics.[4,6,11]

However, the applicability of kriging to spatial statistical modelling is restricted by the assumptions on which the technique is based. The standard form of kriging model is described by the equation

$$z^*(\mathbf{r})=m(\mathbf{r})+\sum_i w_i(\mathbf{r})[z(\mathbf{r}_i)-m(\mathbf{r}_i)]$$

where $z^*(\mathbf{r})$ is an estimator of the variable z at the point $\mathbf{r}$, $m(\mathbf{r})$ is the mathematical expectation of $z(\mathbf{r})$, i.e. $m(\mathbf{r})=E[z(\mathbf{r})]$ and $w_i(\mathbf{r})$ are weight coefficients, which should be obtained such that the estimate $z^*(\mathbf{r})$ is unbiased and has minimal variance. There are a number of modifications of such models, each of which demands different degrees of knowledge (or assumptions) about $m(\mathbf{r})$. For example, in simple kriging, $m(\mathbf{r})=m$ where m is known. In the case of ordinary kriging, $m(\mathbf{r})$ is an unknown constant, and in universal kriging

$$m(\mathbf{r})=\sum_i c_i f_i(\mathbf{r})$$

In this last case, the value of $m(\mathbf{r})$ is controlled by a known set of functions $f(\mathbf{r})$ and unknown coefficients c_i. None of these particular models are entirely appropriate to the list of basic assumptions under which the fitting of anomalies in geochemical data is to be carried out: these assumptions are given below in the next section. More complex models for $m(\mathbf{r})$ make the problem computationally

[1]OOO MIF Ecocentre, Vladivostok, Russia
[2]OAO Pokrovsky Rudnik, Blagoveschensk, Russia
[3]Resources Computing International Ltd, Matlock, UK

*Corresponding author, email stephen.henley@resourcescomputing.com

Published by Maney on behalf of the Institute and The AusIMM
Received 5 January 2009; accepted 10 April 2009
DOI 10.1179/174327509X434939

intractable and do not guarantee the optimality of the geostatistical estimates.

It is important to note here that the fact that distribution of an observed variable may be different at different locations, is the principal characteristic of a geochemical anomaly, and suggests how it could be quantified. However, to apply kriging techniques, spatial and temporal homogeneities are mandatory. Homogeneity of the random field means stationarity of increments of the modelled variable. This feature makes it possible to use the variogram. Strictly speaking, if the problem is not homogeneous (stationary), then kriging is not applicable unless it can be made so by special treatment: such as through data transformation, de-trending or other means.

Nowadays, there is much greater computing power available than when trend surface analysis was first developed, and it is now possible to define more complex and useful functions for use with this technique. One such approach[1–3] has been developed by Ecocentre, a consulting firm based in Vladivostok, in the far east of the Russian Federation. The purpose of their approach is to model geochemical dispersion haloes around multiple concentrated mineralisation sources (which may be considered effectively as point sources). Of the infinitely many possible functions which could be used for this, Ecocentre selected linear combinations of Gaussian functions, for two reasons: because of its mathematical tractability, and because of its close relationship to the expected dispersion curves for such geochemical data.

Nevertheless, even with present-day computing power, the large number of unknowns for multi-centre models, together with the unknown number of point sources, makes a direct attack on the problem, using simple non-linear regression algorithms, infeasible. Instead an indirect iterative solution must be sought.

The problem statement

Basic assumptions

Initial information on distribution of the ore target element (e.g. Sn, Pb, Zn, Ag, W, etc.) in a defined region is (or is derived from) the set of results of chemical analysis of bedrock samples and can include both direct surface assaying and assays of samples from trenches or drill holes. The more representative data of such kind which have been collected, the better. The problem itself is that of identifying the areas where concentrations of the target element or associated pathfinder elements are anomalously high, i.e. noticeably exceed regional background values. The goal is not only to locate such points but to link them in a geologically reasonable way. This will help us to quantify anomalies and interpret them in geometrical terms. To this end, the aim is to fit data by a rational and flexible model under a certain 'theoretical' framework so that a number of basic assumptions should be fulfilled. Those are the following:

(i) the geochemical field is stationary (temporally homogeneous, not to be confused with spatially stationary as required by geostatistical methods)
(ii) the spatial distribution of the ore element is continuous and its isocontours are closed
(iii) the modelled variables are additive
(iv) the field contains an unknown number of geochemical anomalies
(v) the modelled variables are superimposed on a random background
(vi) at sufficient distance from any anomaly centre, concentrations tend towards the mean regional background value, so that integration above background concentration over all coordinate space (and any its subset as well) is finite
(vii) anomalous geochemical fields of different hierarchical levels are geometrically similar (genetically homogeneous).

Stationarity means that there are no temporal variations in the geochemical field, and thus no need to include temporal terms into the model. This is trivially true for geological survey data. Nevertheless, it is explicitly pointed out here that in this instance the goal of the mathematical modelling is to explain the final result of a geochemical process but not the process itself.

The assumption that the observed distribution is continuous should not cause objections, even nugget-type distributions can be approximated, with accuracy sufficient for practical purposes, by continuous functions. Closeness of the isolines above the background level is a consequence of the continuity.

It may be supposed that the investigated geochemical field may result from several geological processes of the same type. Every single process leads to occurrence of one or more separate anomalies. Generally speaking, their total number is unknown. Additivity means that the resulting concentration field is an ordinary sum (i.e. sum without interaction) of partial fields generated by each of these processes.

It should be understood that the regional background in the sixth assumption and random background in the previous one are not the same. The 'random background' refers to an unknown additive term (noise) that is absolutely non-systematic and cannot so easily be excluded from the data.

The 'regional background' is an additive constant term which must be estimated beforehand in one way or another. Such techniques will not be discussed in this paper. We just assume that it has been carried out somehow and $C(\mathbf{r})$ is the net concentration, concentration above the background level. This assumption does not seem unfeasible: normally geochemical modelling is undertaken after a preliminary statistical description of the territory has been obtained and the analyst (geochemist) has come to a decision on what value should be treated as a background concentration.

Geological uniformity of geochemically similar objects leads to uniformity of the mathematical description of corresponding concentration fields. Description of complex geochemical fields may be achieved extensionally by superimposing simpler models of the same kind.

The model

In view of the basic assumptions formulated above, the geochemical field of concentration can be presented by additive (normal)

$$C(\mathbf{r}) = \mathrm{F}(\mathbf{r}) + \varepsilon = \sum_{k=1}^{L} \mathrm{f}_k(\mathbf{r}) + \varepsilon \tag{1a}$$

or multiplicative (lognormal)

$$C(\mathbf{r}) = \mathrm{F}(\mathbf{r}) \cdot \varepsilon = \sum_{k=1}^{L} \mathrm{f}_k(\mathbf{r}) \cdot \varepsilon \tag{1b}$$

models, where $\mathbf{r}$ is the vector of spatial coordinates defining the location of any point, $C(\mathbf{r})$ is the concentration at point $\mathbf{r}$ after subtraction of the mean regional background, $f_k(\mathbf{r})$ is a model of a separate anomaly, $F(\mathbf{r})$ is an additive model of observed concentration field, L is the (initially unknown) number of anomalies being modelled and ε is a random background.

A separate anomaly is modelled by the Gaussian function that can be written down as (omitting bottom indices)

$$f(\mathbf{r}) = g(\mathbf{r};\ C_0, \boldsymbol{\mu}, \mathbf{A}) = C_0 \exp\left[-(\mathbf{r}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{A}(\mathbf{r}-\boldsymbol{\mu})\right] \qquad (2)$$

where $g(\mathbf{r};\ C_0, \boldsymbol{\mu}, \mathbf{A})$ denotes a Gaussian function with the tuple of defining parameters (C_0, $\boldsymbol{\mu}$, $\mathbf{A}$): $\boldsymbol{\mu}$ is the vector defining the location of maximum concentration, i.e. $C(\boldsymbol{\mu}) = C_0$, C_0 is the maximum concentration given by the model (i.e. concentration at $\mathbf{r}=\boldsymbol{\mu}$) and $\mathbf{A}$ is the matrix of coefficients defining the field shape and spatial orientation.

In the general case, there will be multiple anomalies. The model will be represented by an additive mixture of partial single-anomaly models

$$F(\mathbf{r}) = \sum_{k=1}^{L} C_{0k} \exp\left[-(\mathbf{r}-\boldsymbol{\mu}_k)^{\mathrm{T}}\mathbf{A}_k(\mathbf{r}-\boldsymbol{\mu}_k)\right] \qquad (3)$$

It is easy to see that the Gaussian functions and models (1)–(3) conform to the basic assumptions listed above.

It is worthy of note that it does not always make sense to interpret each summand in the poly-Gaussian mixture (3) as a separate anomaly. In the simplest case, a single Gaussian function may really give a self-sufficient description of a separate anomaly, but sometimes it is reasonable to use several Gaussian functions to describe one anomaly. So, generally speaking, it is better to designate function (2) as a component of the model (3) (not as an independent anomaly).

The mathematical problem consists in estimation of parameters of the model that are optimal in the sense of some appropriate criterion. Actually, various criteria are admissible for this purpose. When assumption of normality of a random background ε is acceptable, it is pertinent to use the standard least squares fitting (which is then identical to the maximum-likelihood estimation)

$$\sum_{j=1}^{M}\sum_{k=1}^{L}\left[C(\mathbf{r}_j) - F(\mathbf{r}_j)\right]^2 \to \min \qquad (4a)$$

Often a lognormal distribution is postulated instead. In that case, the least squares technique with logarithmic residuals should be used

$$\sum_{j=1}^{M}\sum_{k=1}^{L}\left(\log\left[C(\mathbf{r}_j)/F(\mathbf{r}_j)\right]\right)^2 \to \min \qquad (4b)$$

Basically both problems (with normal or lognormal distributions) may be resolved with essentially the same computational method.

The method

The approach offered below is based on an iterative numerical procedure. Generally speaking, there are two principal aspects in such procedures that are usually discussed: choice of initial approximation and iterative improvement of the current solution. While the latter could be adopt a standard method (there are straightforward methods of local optimisation that could be adapted for this task), a good choice of the initial solution almost always depends on special features of the problem. To show the basic ideas, we preface the formulation of the method with some simple definitions.

Initial approximation: local estimates

Consider a product of two Gaussian exponentials $f_1(\mathbf{r}) = g(\mathbf{r};\ C_1,\ \boldsymbol{\mu}_1,\ \mathbf{A}_1)$ and $f_2(\mathbf{r}) = g(\mathbf{r};\ C_2,\ \boldsymbol{\mu}_2,\ \mathbf{A}_2)$. Obviously, the result $f_3(\mathbf{r}) = f_1(\mathbf{r})\cdot f_2(\mathbf{r})$ turns out to be another Gaussian function with some C_3, $\boldsymbol{\mu}_3$ and $\mathbf{A}_3$. It is clear that knowing of any pair of tuples out of the set $\{C_i, \boldsymbol{\mu}_i, \mathbf{A}_i | i=1, 2, 3\}$ allows computation of the third one (which is unique provided that the former two are correctly defined). The respective set of equations is

$$C_3 \exp\left[-\boldsymbol{\mu}_3^{\mathrm{T}}\mathbf{A}_3\boldsymbol{\mu}_3\right] = C_1 \exp\left[-\boldsymbol{\mu}_1{}^{\mathrm{T}}\mathbf{A}_1\boldsymbol{\mu}_1\right]\cdot C_2 \exp\left[-\boldsymbol{\mu}_2^{\mathrm{T}}\mathbf{A}_2\boldsymbol{\mu}_2\right] \qquad (5a)$$

$$\mathbf{A}_3\boldsymbol{\mu}_3 = \mathbf{A}_1\boldsymbol{\mu}_1 + \mathbf{A}_2\boldsymbol{\mu}_2 \qquad (5b)$$

$$\mathbf{A}_3 = \mathbf{A}_1 + \mathbf{A}_2 \qquad (5c)$$

Let now consider a weighting function $\Psi(\mathbf{r}) = g(\mathbf{r};\ D,\ \mathbf{m},\ \mathbf{Q})$, where D, $\mathbf{m}$ and $\mathbf{Q}$ are some parameters (a scalar, a vector and a matrix of required sizes). Function Ψ quantifies the concept of 'locality' and works as a 'filter'. Values $\Psi(\mathbf{r}_k)$ can be thought of as statistical weights: multiplication of $C(\mathbf{r}_k)$ by $\Psi(\mathbf{r}_k)$ suppresses assays taken far (in terms of the function Ψ) from $\mathbf{r}_k$ and filters out samples for the purposes of local estimation. Choosing $\mathbf{m}$ in the vicinity of a certain Gaussian component (let denote its index as j) and tuning other parameters of the filter function so as to suppress other components in equation (3), the authors get an approximate equality

$$F(\mathbf{r})\Psi(\mathbf{r}) = \sum_{k=1}^{L} g(\mathbf{r};\ C_{0k}, \boldsymbol{\mu}_k, \mathbf{A}_k) g(\mathbf{r};\ D, \mathbf{m}, \mathbf{Q}) \approx g(\mathbf{r};\ C_{0j}, \boldsymbol{\mu}_j, \mathbf{A}_j) g(\mathbf{r};\ D, \mathbf{m}, \mathbf{Q}) = g(\mathbf{r}, C^*, \mathbf{m}^*, \mathbf{A}^*) \qquad (6)$$

By means of multiplicative 'filtering' of such kind, it is possible to reduce the original estimation problem to a series of essentially simpler single-component problems (i.e. to problems in which $L=1$). After obtaining C^*, $\mathbf{m}^*$ and $\mathbf{A}^*$, it is possible to calculate corresponding parameters (C_0, $\boldsymbol{\mu}$, $\mathbf{A}$) which will define a component to include into the mixture (3): they are the solution of the system of equations (5a)–(5c) where $f_2(\mathbf{r}) = \Psi(\mathbf{r})$, $f_3(\mathbf{r}) = g(\mathbf{r},\ C^*,\ \mathbf{m}^*,\ \mathbf{A}^*)$, relatively to $f_1(\mathbf{r})$ which now is the sought partial component of the model.

Thus to obtain reasonable estimates for all partial components, it is sufficient to have L suitable tuples of parameters for function $\Psi(\mathbf{r})$ defined above (value of L also needs estimation). Each such tuple allows a modified subset of samples to be derived where the effect of a separate Gaussian function becomes maximally apparent. However, finding such parameters remains problematic even if the optimal number of anomalies is known.

Generally, different positions of the filtering function (i.e. different $\mathbf{m}$) lead to different local estimates. It is expedient to successively substitute the sample coordinate vectors $\mathbf{r}_k$ for $\mathbf{m}$ in equation (6). Coefficient D can

be arbitrary (so let $D=1$), and matrix $\mathbf{A}$ expresses a guess about the form and scale of anomalies to be revealed. Other (not necessarily deterministic) choices of filtering function are also possible.

As a rule, local estimates are not convenient for direct use in model construction. The reason for this is obvious. In practice, parameters of the weighting function cannot be ideally fitted to the data and unknown components; hence a mutual influence of partial components is unavoidable. This may make approximation (6) too approximate and lead to a situation where not all local estimates really correspond to partial components that should be involved in the poly-Gaussian mixture (3). Nevertheless, at least in the case where all components of the model are spatially separated, it is natural to expect that in the tuple-structured parametric space, the local estimates will be grouped in well-distinguished clusters, and each cluster will correspond to a partial component in the model (3). So it is reasonable to couple each separate cluster with a certain partial component and substitute the cluster average for its parameters.

Success of this approach in principle depends on the quality of single-component modelling and efficiency of subsequent clustering of the local estimates in parametric space. Thus the authors come to a procedure in which the emphasis is moved from a complex numerical optimisation technique to a cluster analysis problem.

Formulation of the algorithm

The principles of the computational procedure to produce a multi-anomaly model and on its basis to estimate resources are essentially simple. Its basic stages are local estimation, clustering, calibration and final adjustment, applying as follows:

(i) first, an initial approximation of the sought solution must be determined. To this end, select initial values for parameters D_k, $\mathbf{m}_k$ and $\mathbf{Q}_k$, for each k ($k=1, 2, \ldots, N$) in turn, on a regular grid or a random basis. A large number N ($N \gg L$, generally, the greater the value of N, the better) of different initial parameters (location and scale parameters) is used in separate runs to fit the single-anomaly function (1). Each component of the model is identified by carrying out a search for a local fit. Using the weighing function $\Psi(\mathbf{r})=g(\mathbf{r};\ D_k,\ \mathbf{m}_k,\ \mathbf{Q}_k)$, solve least squares fitting problem for a single-component model. This solution is likely to use an iterative non-linear regression algorithm. Repeat this to obtain a tuple-structured set of a large number of locally optimal parameters $\{C_{0k},\ \boldsymbol{\mu}_k,\ \mathbf{A}_k\}$

(ii) examine the results by applying cluster analysis to the set of triples $(C_{0k},\ \boldsymbol{\mu}_k,\ \mathbf{A}_k)$, to determine the number L of anomalies and initial approximations for the parameters of each (being the averages of the parameters for each cluster). This step is necessary because, out of the initial set of N starting points, many will lead to essentially the same L solutions, while some of them may give geologically infeasible combinations of parameters. The former should be merged together, while the latter should be excluded from consideration

(iii) keep the cluster centres as model components and calibrate the model, i.e. fit the optimal values of C_{0k}. Because of (near) multiple collinearity, some of the components may be found to be redundant, so exclude these and fit the rest of them again

(iv) once the number L of model components is determined, and an initial set of parameters is obtained for each, these can be used as the starting point for the final step, to obtain a solution simultaneously for all parameters, for the multiple-anomaly system. The optimisation problem (4) must be solved. This will also use a similar iterative non-linear regression algorithm to that of step (i), but the existence of good estimates for the parameters for each of the anomalies will help greatly to reduce the number of iterations required to reach a better model

(v) compute estimates of the resources by numerical integration of the fitted function for response values in the spatial limits of interest, above a given cutoff value.

It is often desirable to obtain a more compact model (with as few anomalies as possible). Provided that the underlying distribution of the observed variable is known (e.g. normal or lognormal), the last stage of the algorithm may be followed (or accompanied) by standard statistical testing which may allow the number of anomalies to be reduced, subject to a required level of statistical significance within the model.

A demonstration problem

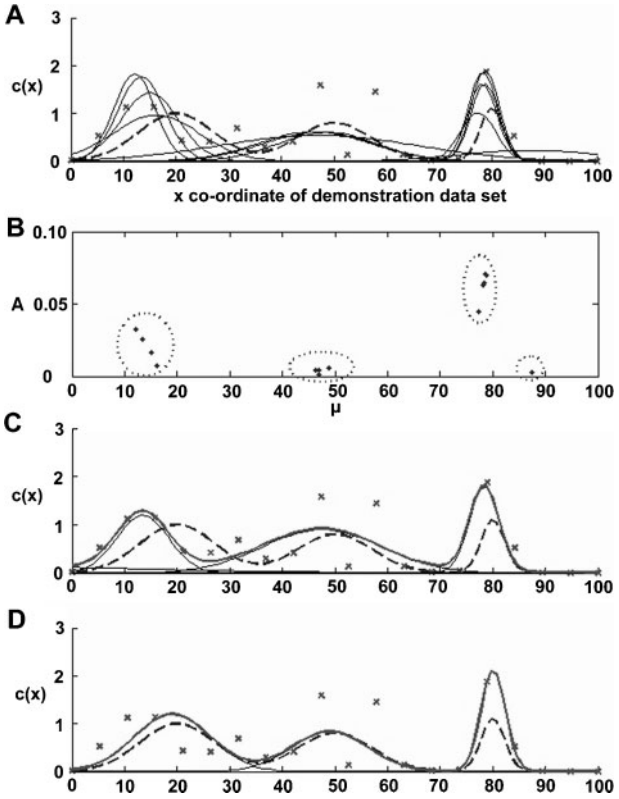
To get an idea of how the whole algorithm works it is best to examine the solution of a one-dimensional demonstration problem with small (to make graphic illustrations simpler) values of N and L.

Consider a computer-generated one-dimensional dataset with the coordinate x randomly chosen in some interval, let it be the interval (0, 100) in some relative units, and $c(x)$ calculated in accordance with equation (3) where $L=3$, $C_{01}=1$, $\boldsymbol{\mu}_1=20$, $\mathbf{A}_1=0{\cdot}01$; $C_{02}=0{\cdot}8$, $\boldsymbol{\mu}_2=50$, $\mathbf{A}_2=0{\cdot}01$; $C_{03}=1{\cdot}1$, $\boldsymbol{\mu}_3=80$, $\mathbf{A}_3=0{\cdot}1$ (all values in any compatible units). To lessen loading of diagrams, a rather small set of samples $N=20$ was chosen. Uniform sampling was simulated. To represent possible measurement/modelling error, the generated data were multiplied by a pseudorandom coefficient $\exp[\sigma \cdot \xi]$, where ξ has the standard normal distribution while σ defines the resulting variance. Diagrams below corresponds to the case $\sigma=1$, which is a case of medium–high level of noise pollution, in comparison with the maximal values of $c(x_k)$ which are expected to be about 1.

Figure 1 shows the results of each stage of the algorithm. Some comments on the diagrams are provided to aid the reader in following the selection procedure:

Figure 1*A* shows initial data (indicated by cross signs) along with the exact (unknown) fitting curve (dashed line); solid lines show components from the set of local estimates. The total number of data points is 20. All of them were used as points for local estimation. Parameters of the locality function are $m_k=x_k$ and $A=0{\cdot}01$. The number of meaningful model components is 14 (other local estimates were deleted as physically infeasible).

Figure 1*B* displays the feature space of the clustering problem solved after local estimates were obtained. Dots



A initial data (indicated by cross signs) along with the exact (unknown) fitting curve (dashed line); *B* feature space of the clustering problem solved after local estimates were obtained; *C* calibrated model [step (iii) as described above]; *D* model after a final optimisation

1 Solution of a simplified problem by the poly-Gaussian modelling procedure

are image points of the local estimates in coordinates ($\boldsymbol{\mu}$, $\mathbf{A}$). The clusters are shown with dotted ellipses; the total number of clusters is 4.

Figure 1*C* represents the calibrated model [step (iii) as described above]; now there are only three components left, centred approximately at the locations $x=14$, 47 and 79.

Figure 1*D* shows the model after a final optimisation: notice that two components are revealed with a quite good accuracy. The third one also is identified, though the precision is not so good. Notice that it is represented by an extremely small number of samples: there are only a couple of sample values in the vicinity of $x=\boldsymbol{\mu}_3=80$. In practice, this would mean that additional verification of the model in such areas is needed.

Inherent problems

It is unreasonable to expect that any particular method is capable of producing a perfect model. Ore estimates (nearly) always are carried out under uncertainty and lack of data. In such conditions, modelling is inevitably associated with a certain set of problems, which cannot be resolved completely in all aspects.

Generally speaking, possible mis-descriptions of concentration fields modelled by additive poly-Gaussian mixtures may lead to the model missing (true) anomalies, apparent recognition of extra (false) anomalies, excessive agglomeration of anomalies (one large instead of several small), unreasonable fragmentation of anomalies (several small instead of one large) and misrepresentation of the scale or form of anomalies. It is unreasonable that we should only blame the method for such drawbacks and unwanted features. They are objective by their nature and caused by a number of possible reasons including:

(i) lack of data, or information deficit in general
(ii) sparse and irregular sampling
(iii) low quality of laboratory analysis and poor topographic control
(iv) blind ore bodies and weak erosion of aureoles
(v) planar topography and lack of drill samples
(vi) presence of elongated veins and blankets that are difficult to interpret geometrically as a superposition of ellipsoids.

Each of these factors can create difficulties in the solution and there is no universal prescription for overcoming them. Nevertheless, certain manipulations of the original data and adjustment of the algorithms may, to some degree, improve the situation. First and foremost among such actions is de-clustering: a preliminary averaging of exploration data over a regular grid of rectangular areas ('bins'). This is useful because it makes the problem 'better conditioned' because as a consequence, binned and averaged data are smoother and more uniformly distributed. As a result, binning often leads to more robust computing and helps to obtain better estimates, especially when data contain extremely large or extremely small values (which is usually the case for geochemical survey data). Choice of the bin size will have an effect on the resolving power of the method. As a rule, smaller bins will allow more compact anomalies to be identified, but may increase the fragmentation effect. Enlarging of bins makes data smoother and lessens the influence of outliers but with possible loss of detail of the model.

Another important part of the work is parametric adjustment. Each algorithm (except the simplest and primitive) has a number of control parameters that define its behaviour. In this case, it is possible to modify the weighting function, binning scheme, clustering algorithm and method of computation of local estimates for the Gaussian components of the model, so the method as a whole is very flexible. Parameters of the weighting function help partially to adapt the model components to the expected shape and intensity of sought anomalies. Clustering could be carried out with various possible metrics representing the distances between the model components, and various admissible ranges for the number of clusters. Thus it becomes possible to choose parameters to reduce unwanted fragmentation or agglomeration of the modelled anomalies and to some extent to take into account geomorphological properties of the explored area.

A large mathematical difficulty in this (and other similar) problems is created by the multimodality of the target function that results in non-uniqueness of models proposed by the method. Therefore, it may be necessary to browse a set of alternative models and choose the 'best' from among them. In these circumstances, a rational approach to performing the computations implies multiple runs with a variety of starting conditions for the algorithm parameters.

Of course, these measures should not be considered as universal tools to resolve all the problems listed above in all cases. However, well-conditioned data and a well-adjusted

2 Location of Pokrovsky Rudnik within the Amur Region of Russia

procedure give stable qualitative results and lead to reasonable conclusions.

Case study: exploration area around Pokrovsky mine

The method described above is illustrated here using a case study based on the area around the Pokrovsky Rudnik gold mine in the Amur region of far eastern Russia (Fig. 2). This area is characterised by a large number of zones of mesothermal gold mineralisation, and has been extensively sampled,[5] from shallow and deep drill holes as well as in surface trenches. A dataset was compiled consisting of 16 521 gold assays (approximately 80% from drill holes) from these exploration samples, over an area of 38 km^2 within the exploration licence (but excluding the very large number of samples from the operating mine and its immediate vicinity).

The method described above was applied to this dataset, and 136 components of the model were identified and fitted by equation (3), with 1360 unknowns: 10 for each component: these are the maximum Au concentration, the three x, y and z coordinates, and six ellipsoid scale/orientation parameters (matrix **A**). Many of the computed components turned out to be non-essential in that sense that they had small (or very small) coefficients C_0 (a few hundredth or first tenth of a gram per tonne). They were retained in the model to describe the total aureole, but they did not significantly influence determination of the modelled resources.

In total there are 67 components with C_0 above the indicated level in the fitted model. Such (essential) components are shown in Fig. 3. This is a horizontal section of the three-dimensional model of the explored region, and each independent component of the model is represented by an ellipsoidal volume, i.e. elliptical in plan, while their sums may form more complicated contours. Because it is the concentration of gold that is modelled, each can be outlined at the level of a defined cutoff grade, and furthermore, numerical integration can be applied in order to estimate the tonnage of gold above this cutoff within each anomaly. It is therefore possible to make preliminary estimates of the economic significance of each modelled anomaly.

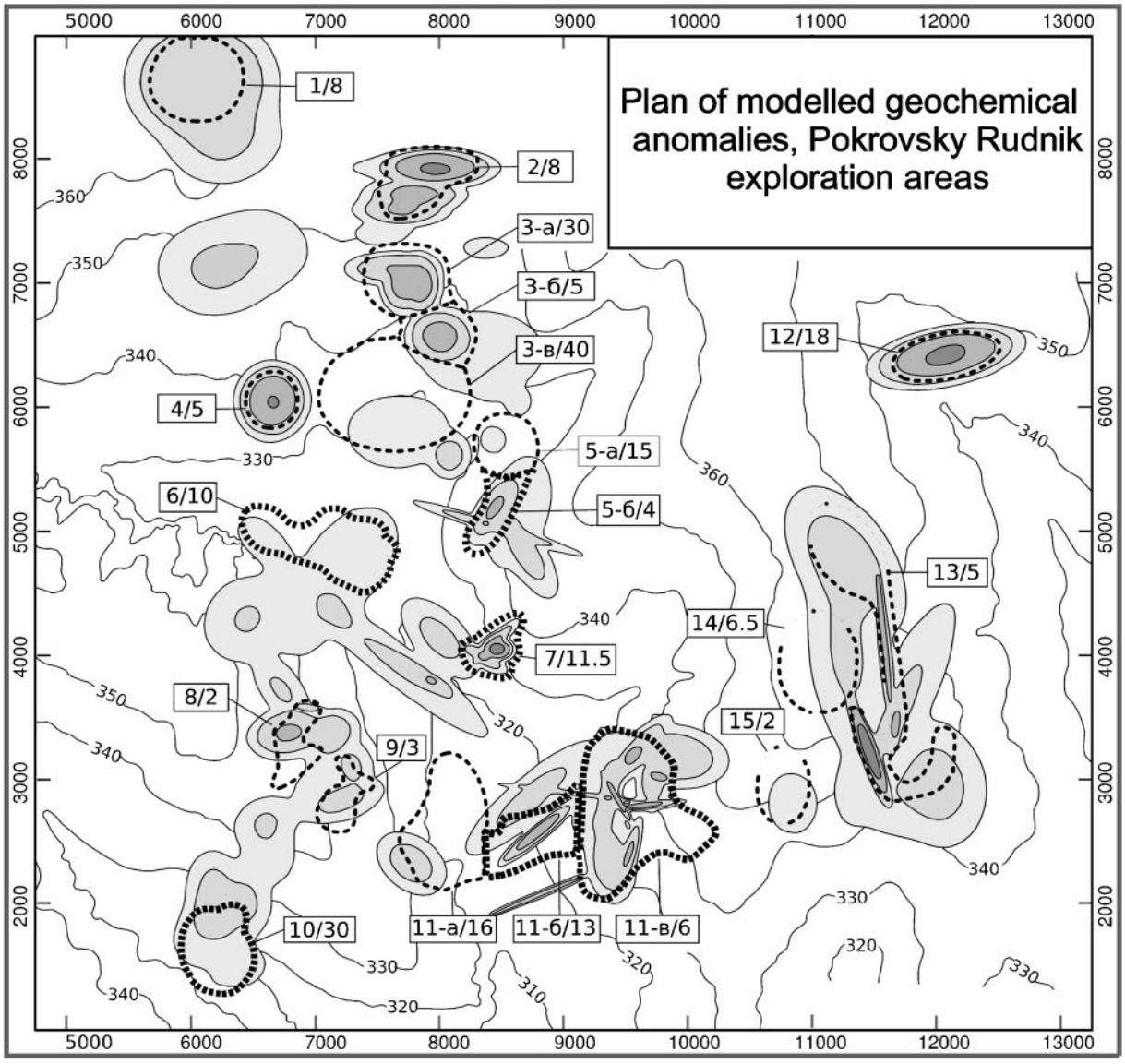
For three areas, Pokrovka-II, Pokrovka-III and Pokrovka-IV, these estimates can be compared with previous estimates from resource modelling carried out by conventional methods, based on trench and drill hole data (Table 1). Comparison of the Ecocentre model with conventionally modelled resources as approved by the Russian State Commission on Resources (GKZ) is given in the last two rows of the table (the latter was unknown at the stage of modelling). Please note that there is currently no formal equivalence between GKZ and international (e.g. JORC) classifications, but these numbers are broadly equivalent to JORC inferred or higher resource categories.

Figures are the estimated tonnes of contained gold above a 1 g t^{-1} cutoff grade.

Some additional information on empirical data and the model is also given in the table: the number of 'essential' components (with C_0 greater than cutoff grade) and the number of 'dominant' assays ('dominant' are points where the measured concentration is above the regional background value and the sum of components related to the given zone is higher then analogous sums of components related to any other zone). The coefficient of determination R^2 is also included, though it is not as meaningful here as it is in linear regression problems. Its value ranges from a rather poor or moderate for Pokrovka-II and Pokroovka-III to quite high for Pokrovka-IV. Note that in all cases, agreement between the model forecast and the GKZ approved value is excellent.

Conclusions

The method described here is different in principle from the usual methods of handling geochemical exploration data, which effectively use descriptive rather than predictive modelling approaches: for example, simple geochemical mapping and visual identification of



3 Fitted poly-Gaussian geochemical model from exploration data around Pokrovsky Rudnik. Map coordinates are in metres (eastings on *x* axis and northings on *y* axis)

Table 1 Some characteristics of three zones in the bounds of the Pokrovsky Rudnik gold mine: Pokrovka-II, Pokrovka-III and Pokrovka-IV

Characteristics	Pokrovka-II	Pokrovka-III	Pokrovka-IV
Anomaly no. in Fig. 3	11-б and 11-в	7	5-б
No. of 'essential' components	18	6	5
No. of 'dominant' assays	3215	1592	967
Coefficient R^2	0·27	0·46	0·73
Estimated resources in tonnes of gold above 1 g t^{-1} cutoff grade			
Ecocentre model forecast	19	11·5	4
GKZ approved resource	19	10	4

anomalies. For this reason, it is difficult to develop objective quantitative comparisons.

The characteristic property of the method is that the model structure is not specified *a priori*, as it is assumed to be in parametric regression models. The shape of anomalies is not fixed in advance and is very flexible. It is determined by the number, location and scale of Gaussian components used to approximate each anomaly. All of these are determined from the data. From this point of view, poly-Gaussian modelling may be thought of as a compromise between parametric and non-parametric statistical methodology.

This method developed by Ecocentre (and coded as program 'Locator') uses essentially a simple set of equations to model multi-centre geochemical anomalies or other data wherever a poly-Gaussian model may similarly be appropriate. In the theory and practice of geochemical exploration, Gaussian functions have long been known and used for the description of secondary dispersion aureoles.[12,14] In the problem under consideration, the model is based on bedrock assaying and so relates to primary (not secondary) aureoles. The nature of primary aureoles is explained by physicochemical processes which differ completely from those of secondary aureoles, nevertheless Gaussian functions are equally suitable for modelling of either primary or secondary aureoles.

Though the precision of this modelling technique, as of any other, depends on quantity and quality of available information, nevertheless it allows preliminary estimation of the likely resource even at an early exploration stage before there are enough data to obtain estimates of measured and indicated resources (as defined by JORC or other international reporting systems). Additive mixtures of Gaussian functions have repeatedly been used to characterise ore deposits and to quantify mineral resources in Primorski Krai (Russia) since the early 1980s.[3] This method gives stable and satisfactory results. It has also been used on a number of exploration projects in other parts of Russia, in particular at Pokrovsky Rudnik and other operations of Peter Hambro Mining plc, where the predictive power of this method has already helped to define exploration programmes and to identify new deposits.

References

1. A. I. Burago: 'Formalizovannaya model anomalnogo geokhimicheskogo polya', in 'Teoriya i praktika geokhimicheskikh poiskov' (in Russian), 50–68; 1990, Moscow, Nauka.
2. A. I. Burago and V. A. Burago: 'Teoria i metodi geokhimicheskoy tomografii v zadachakh poiskovoy geokhimii' (in Russian), in 'Prikladnaya geokhimiya', Vol. 3, 49–85; 2002, Moscow, IMGRE.
3. A. I. Burago and L. B. Reznik: 'Obyomnoye modelirovaniye anomalnogo geokhimicheskogo polya (metodicheskaya osnova, algoritm)' (in Russian), in 'Teoriya i praktika geokhimicheskikh poiskov v sovremennikh usloviyakh', Vol. 2, 21–22; 1988, Moscow, IMGRE.
4. M. David: 'Geostatistical ore reserve estimation'; 1977, Amsterdam, Elsevier.
5. A. I. Dementienko and N. G. Vlasov: 'Pokrovskoye zolotorudnoye mestorozhdeniye (prospekt ekskursii po programme Mezhdunarodnoy nauchnoy conferentsii ''Genezis mestorozhdeniy i metodi dobichi blagorodnikh metallov'')' (in Russian); 2000, Blagoveschensk, AmurKNII.
6. A. Journel and C. J. Huijbregts: 'Mining geostatistics'; 1978, London, Academic Press.
7. D. Krige: 'A statistical approach to some basic mine valuation problems on the Witwatersrand', *J. Chem. Metall. Min. Soc. S. Afr.*, 1951, **52**, 119–139.
8. W. C. Krumbein and F. A. Graybill: 'An introduction to statistical models in geology', Chapter 13, 317–356; 1965, New York, McGraw-Hill.
9. G. Matheron: 'Principles of geostatistics', *Econ. Geol.*, 1963, **58**, 1246–1266.
10. G. Matheron: 'Traité de geostatistique appliquée', Vols. 1 and 2; 1962, Paris, Technip.
11. B. D. Ripley: 'Spatial statistics'; 2004, New York, John Wiley & Sons.
12. N. I. Safronov: 'K voprosu ob 'oreolakh rasseyaniya' mestorozhdeniq polezhykh iskopaemykh i ikh ispolzovanii pri poiskakh i razvedke' (in Russian), in 'Geokhimicheskie metody poiskov rudnykh mestorozhdeniy', Part 1, 4–21; 1981, Novosibirsk, Nauka.
13. R. Sibson: 'A brief description of natural neighbour interpolation', in 'Interpreting multivariate data', (ed. V. Barnett), 21–36; 1981, New York, John Wiley & Sons.
14. A. P. Solovov: 'Geochemical prospecting for mineral deposit'; 1987, Moscow, Mir.
15. D. F. Watson: 'Contouring – a guide to the analysis and display of spatial data'; 1992, Oxford, Elsevier.