A Statistical Approach to Geological Mapping[1]

J. K. Ghosh,[2] J. Bhanja,[2] S. Purkayastha,[3] T. Samanta,[4] and S. Sengupta[5]

A geological map is the representation, on a two-dimensional plane, of the disposition of three-dimensional rock bodies exposed on the earth's surface. The problem of mapping is essentially that of dividing an area into "homogeneous" subregions on the basis of the exposed rock types. Automatic Bayesian methods of model selection using default Bayes factors have been employed to solve the problem of choosing a set of boundaries between "homogeneous" subregions, assuming no complication excepting low-angle tilting affected rock bodies. The method is tested on two data sets. A sampling scheme for optimum allocation of observation points is also presented.

KEY WORDS: Bayes factor, Bayesian model selection, fractional Bayes factor, intrinsic Bayes factor, minimal training sample, noninformative prior.

INTRODUCTION

The conventional technique of geological mapping in an area having sedimentary rock formations involves (a) identificaiton of rock bodies (beds) outcropping on the earth's surface, (b) representation, on a two-dimensional plane, of the locations of these three-dimensional beds, their inclinations (dips) and trends, and (c) tracing or interpolating the boundaries between the contrasting rock types taking into account their trends (in a simplified case, the strike directions). Location of the boundaries drawn in the same area by different geologists may, however, vary considerably, due to personal perceptions of the geologist concerned.

For a statistician the problem concerned is one of defining boundaries between several "homogeneous" subregions. The present work aims at providing an

[1]Received 18 March 1997; accepted 29 June 2001.

[2]Division of Theoretical Statistics and Mathematics, Indian Statistical Institute, 203 B. T. Road, Kolkata 700035, India.

[3]Division of Theoretical Statistics and Mathematics, Indian Statistical Institute, 203 B. T. Road, Kolkata 700035, India; e-mail: sumitra@isical.ac.in

[4]Applied Statistics Division, Indian Statistical Institute, 203 B. T. Road, Kolkata 700035, India; e-mail: tapas@isical.ac.in

[5]INSA Honorary Scientist, Indian Institute of Technology, Kharagpur 721302, India.

automatic Bayesian solution to the problem with a view to defining the boundaries as objectively as possible.

The boundaries are drawn between outcrops of rocks that are composed of mineral grains of varying sizes. Hence, information on mineral proportion, provided by modal analysis of rock samples (by point counting in thin sections), as well as that on grain size, provided by granulometric analysis, are essential inputs for our work.

Starting with an area having sedimentary rock formations where the boundary between two adjacent rock units is planar, and the topography is flat, we develop our method that is general enough to be applicable to any data set involving mineral composition and mean grain-size. In the simple case that we consider, the rock bodies are *gently* tilted, but not deformed. The strike directions running normal to the axis of tilt are taken into consideration.

We apply our technique firstly to an area where all these information are available, in a limited quantity, from an earlier study (Sengupta, 1970). At a later stage a larger and relatively more complicated case involving observed as well as "generated" data is considered, the latter being well within the limits of geological acceptability.

Our methods, being automatic, are easy to implement in a PC and may help a geologist in fine tuning his first impressions gathered from a field trip. We try to use as much of the geological information as possible. We do not want to replace a geologist's expertize, only process the information he has or at least the part of it that is easy to quantify to start with. In an interactive mode, the geologist can spend time saved in this way to make further subdivisions based on any additional knowledge or information beyond the composition of rocks and strike directions.

The statistical formulation of the problem involving theory as well as motivation is presented in the next two sections. A model is set up for the available data, given the subregions into which a given set of boundaries partitions the whole area. Thus the problem of choosing a set of boundaries corresponds to selecting one of several possible models. Readers interested only in the application of our method may skip the next two sections and go directly to the section on Numerical Calculations giving an algorithm as an aid to application.

As briefly discussed in the text, we may also use our formulation to answer two very important questions faced by geologists, namely the optimum number of samples to be collected in the field and their locations. We hope to discuss the sampling problem in detail elsewhere.

STATISTICAL FORMULATION AND MODELLING

Our problem is to identify a set of boundaries (between geological formations) that partitions the region into a number of subregions that are homogeneous with

respect to certain characteristics of the rock such as mineral composition, mean grain-size, etc.

For each possible subregion we postulate a lognormal distribution for the mean grain-size (Sengupta, Ghosh, and Mazumder, 1991) and a logistic normal distribution (Aitchison, 1986) for the mineral proportions. A partition then corresponds to a model that assumes distributions for different subregions with different sets of parameters. Thus the problem of choosing a set of boundaries or a partition corresponds to selecting one of several possible models. Although many different models (partitions) are possible theoretically, our technique, discussed in the following section, allows us to limit the choice of plausible models.

We first consider the case with a single variable. In the specific case that we consider first, spatial variation of mineral composition of a two phase aggregate (feldspar + quartz or feldspar + grain size) is involved. In an earlier study the feldspar proportion was found to have the highest discriminating power among all the variables (Ghosh, Saha, and Sengupta, 1981), so we begin with the univariate data on feldspar. Let X denote the feldspar proportion. As mentioned above, an appropriate distribution for X would be a logistic normal distribution that assumes normality of the transformed variable $\log(X/(1-X))$. We cannot exactly assume this distribution, however, because feldspar proportion may be zero for some of the rock samples. This technical problem is solved by considering a mixture with a distribution degenerate at zero. Thus the distribution of X is assumed to be of the form

$$\alpha P(\mu, \sigma) + (1-\alpha)\delta \tag{1}$$

where $P(\mu, \sigma)$ is a logistic normal distribution with parameters μ and σ, α is the mixing constant, and δ is a degenerate measure putting all its mass at zero. This distribution has a density with respect to a dominating measure, which is the sum of the Lebesgue measure and δ, so we can calculate density and likelihood. The density is given by

$$f(x \mid \alpha, \mu, \sigma) = \alpha \frac{1}{x(1-x)} \phi\left(\log \frac{x}{1-x} \mid \mu, \sigma\right) + (1-\alpha)\delta'(x) \tag{2}$$

where $\phi(\cdot \mid \mu, \sigma)$ is the density of $N(\mu, \sigma^2)$, α is the mixing constant and

$$\delta'(x) = \begin{cases} 1, & \text{if } x = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

If we wish to use a bivariate distribution, the variables suggested are feldspar proportion and mean grain-size. Let X = feldspar proportion, and Y = log of mean grain-size. Set

$$\Omega = \{(x, y) : 0 \le x < 1, \ -\infty < y < \infty\},$$

$$\Omega_1 = \{(x, y) : 0 < x < 1, \ -\infty < y < \infty\},$$

and

$$\Omega_2 = \{(x, y) : x = 0,\ -\infty < y < \infty\}.$$

On Ω_1, we assume $(\log(X/(1-X)), Y) \sim N_2(\mu, \Sigma)$ where Σ is positive definite and on Ω_2, we assume $Y \sim N(\xi, \lambda^2)$. The pair of random variables (X, Y) is assumed to have a distribution on Ω that is a mixture of the above two distributions on Ω_1 and Ω_2. If we consider the sum of the Lebesgue measure on $\mathbb{R}^2$ and a one-dimensional Lebesgue measure on Ω_2 as the dominating measure, the density of the joint distribution of X and Y is given by

$$f(x, y \mid \alpha, \mu, \Sigma, \xi, \lambda) = \alpha g(x, y \mid \mu, \Sigma) + (1-\alpha)h(x, y \mid \xi, \lambda), \quad (x, y) \in \Omega \tag{4}$$

where

$$g(x, y \mid \mu, \Sigma) = \frac{1}{x(1-x)}\phi_2\left(\log\frac{x}{1-x}, y \mid \mu, \Sigma\right), \quad \text{if } (x, y) \in \Omega_1$$
$$= 0, \qquad \text{otherwise}$$

and

$$h(x, y \mid \xi, \lambda) = \phi_1(y \mid \xi, \lambda), \quad \text{if } (x, y) \in \Omega_2$$
$$= 0, \qquad \text{otherwise.}$$

Here $\phi_2(\cdot \mid \mu, \Sigma)$ is the density of a bivariate normal distribution with mean μ and dispersion Σ, $\phi_1(\cdot \mid \xi, \lambda)$ is the density of $N(\xi, \lambda^2)$, and $0 \le \alpha \le 1$ is the mixing constant.

The case with three variables, namely feldspar proportion, quartz proportion, and mean grain-size can be handled in a similar way.

Remark 1. If feldspar proportions for all the rock samples are positive, we need not consider the mixture distributions as given by (2) or (4). The densities in this case are obtained from (2) or (4) putting $\alpha = 1$.

Our object is to make inference about both the number of boundaries and their locations. Note that each set of $(k-1)$ boundaries partitions all the data points into k groups. For samples within a group (subregion) we assume the random variables (or vectors) to be independent and identically distributed, whereas across the groups we assume independence and homoscedasticity. Thus in our model, we assume that data for different groups have the same dispersion parameters but other

parameters (e.g., means) differ according to the group. For example, if we wish to use only feldspar proportion, we use for the data in the ith group the univariate distribution (2) for feldspar where μ and α are replaced by μ_i and α_i but σ does not depend on i. It is clear that different boundaries leading to the same partition will have the same likelihood and it is easy to translate inference about partitions into inference about boundaries. So we concentrate on partitions that are slightly easier to deal with. In passing we note that in more complicated cases than that treated here, where there are k rock types and more than $(k-1)$ boundaries because the same rock types are repeated in space, our exercise may be applied more than once in different parts of the map.

LIKELIHOOD AND MODEL CHOICE

We first deal with the univariate case. For any fixed $k \geq 1$, a set of $(k-1)$ boundaries or a partition with k groups corresponds to a model M that states that the observations in the ith group are i.i.d. with a common distribution having density

$$f(x \mid \alpha_i, \mu_i, \sigma) \tag{5}$$

as given in (2) with $0 \leq \alpha_i \leq 1$, $-\infty < \mu_i < \infty$, $\sigma > 0$, $i = 1, 2, \ldots, k$, and the k groups of observations are also independent. The likelihood of such a partition or model is given by

$$L(\alpha_i, \mu_i, \sigma, i = 1, 2, \ldots, k \mid \boldsymbol{X}) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} f(X_{ij} \mid \alpha_i, \mu_i, \sigma)$$

where

n_i = number of observations in the ith group, $n = \Sigma n_i$
X_{ij} = jth observation in the ith group, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, k$.

and $\boldsymbol{X}$ denotes the whole data set.

Note that the above likelihood involves the unknown model parameters. In order to make a choice of a partition (that gives a map) we now find the likelihoods of the possible partitions by integrating out the parameters with respect to a suitable prior. We call this the integrated likelihood. Since subjective specification of prior distributions for all parameters of all the models is not feasible here, we use standard noninformative priors to compute the integrated likelihoods. Indeed we consider uniform prior for α_i and standard noninformative prior $\frac{1}{\sigma}\, d\mu_i\, d\sigma$ for μ_i and σ and calculate the integrated likelihood

$$m(\boldsymbol{X}) = \int L(\mu_i, \alpha_i, \sigma, i = 1, 2, \ldots, k \mid \boldsymbol{X}) \prod_{i=1}^{k} d\alpha_i \prod_{i=1}^{k} d\mu_i \, \frac{d\sigma}{\sigma}.$$

Alternatively, we can eliminate the unknown parameters by maximizing the likelihood with respect to them keeping the partition fixed and thus obtain a sort of "profile" likelihoods of the partitions. It turns out that the integrated and profile likelihoods can be expressed in closed form. The expressions are given in the Appendix.

We choose the partition with the highest integrated likelihood. If all partitions are equally likely a priori, this would correspond to choosing the partition with the highest posterior probability. It is also the partition that minimizes the Bayes risk with 0–1 *loss.*

Alternatively, we can choose the partition with the maximum profile likelihood. This would be the maximum likelihood choice, which will usually be close to the Bayesian choice described in the previous paragraph.

We now consider the bivariate data on feldspar proportion (X) and logarithm of mean grain-size (Y). We assume a distribution with density.

$$f(x, y \mid \alpha_i, \boldsymbol{\mu}_i, \Sigma, \xi_i, \lambda) \tag{6}$$

for the observations (X_{ij}, Y_{ij}), $j = 1, 2, \ldots, n_i$, in the ith group ($i = 1, 2, \ldots, k$) where f is as given in (4). As in the univariate case we consider uniform prior for α_i, and standard noninformative priors $\frac{1}{\lambda}\, d\boldsymbol{\mu}_i\, d\xi_i\, d\lambda$ for $\boldsymbol{\mu}_i, \xi_i$, and λ. The standard noninformative prior considered for Σ is $|\Sigma|^{-(3/2)}$. We calculate integrated likelihoods of possible partitions. We may also calculate "profile" likelihoods of the partitions by maximizing the likelihood with respect to the model parameters. The expressions of the integrated and profile likelihoods of a partition are given in the Appendix.

In actual practice we do not calculate the integrated or profile likelihood of all partitions. This is primarily because it would have taken a lot of computer time. Moreover, most of this time would have been spent on improbable partitions. So we choose a set of partitions using a heuristic principle that is based on the available strike directions. The strike directions clearly indicate a trend of the bed, which give some idea as to which groupings of points will be more probable. Although the boundaries may not be strictly linear, they can be, at least approximately, be imagined to be "piecewise linear" with the directions given by the strike directions.

For the data sets considered in this paper, we have identified a number of partitions that we need to choose from corresponding to

Case (i) No boundary assumed.
Case (ii) One boundary assumed.
Case (iii) Two boundaries assumed.

The methods described so far are only adequate for making a choice from a set of models (partitions) of the same dimension. An explicit justification of this for location-scale models (like ours) as well as more general group models appears

in Berger, Pericchi, and Varshavsky (1998). In the actual problem of simultaneous determination of the number of boundaries and their locations, it is not appropriate to make a choice based on just the integrated (or maximized) likelihoods, since the competing models that we have considered corresponding to Cases (i), (ii), and (iii) above, do not have the same number of parameters. (It is to be noted here that, due to the presence of some zero feldspar proportions, even some models with the same number of boundaries may not have the same number of parameters.) In such cases the principle of parsimony requires that a model with more parameters should be penalized for its complexity. Use of maximized likelihood tends to favour complex models. Integrating the likelihood with respect to a noninformative prior has a different problem when comparing models of different dimensions. A solution lies in the use of recently developed methods due to Berger and Pericchi (1996), O'Hagan (1995), and others. We describe two of them below.

Let M_1 and M_2 be two competing models with the data $\boldsymbol{X}$ having density $f_j(x \mid \boldsymbol{\theta}_j)$ under model M_j. The unknown parameter vectors $\boldsymbol{\theta}_j$ are of dimension p_j and have prior distributions $\pi_j(\boldsymbol{\theta}_j)$, $j = 1, 2$. The *Bayes factor* (BF) of M_2 to M_1 is defined as

$$B_{21} = \frac{m_2(\boldsymbol{X})}{m_1(\boldsymbol{X})} = \frac{\int f_2(\boldsymbol{X} \mid \boldsymbol{\theta}_2)\pi_2(\boldsymbol{\theta}_2)\, d\boldsymbol{\theta}_2}{\int f_1(\boldsymbol{X} \mid \boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_1} \tag{7}$$

where $m_j(x) = \int f_j(x \mid \boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j)\, d\boldsymbol{\theta}_j$ is the marginal or predictive density of $\boldsymbol{X}$ under M_j. Note that $m_j(\boldsymbol{X})$ is the integrated likelihood with respect to prior π_j under M_j.

The Intrinsic Bayes Factor

A solution to the problem with improper priors is to use part of the data as a *training sample*. The idea is to use the training sample to obtain proper posterior distributions for the parameters that can then be used as priors to compute a Bayes factor with the remainder of the data. Let $\boldsymbol{X}(l)$ denote a part of the entire sample so that the posterior $\pi_j(\boldsymbol{\theta}_j \mid \boldsymbol{X}(l)) = f_j(\boldsymbol{X}(l) \mid \boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j)/m_j(\boldsymbol{X}(l))$ is proper. Here $m_j(x(l)) = \int f_j(x(l) \mid \boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j)\, d\boldsymbol{\theta}_j$ is the marginal density of $\boldsymbol{X}(l)$. The (conditional) Bayes factor with the remainder of the data, using $\pi_j(\boldsymbol{\theta}_j \mid \boldsymbol{X}(l)), i = 1, 2$ as priors is given by

$$B_{21}(l) = B_{21}\frac{m_1(\boldsymbol{X}(l))}{m_2(\boldsymbol{X}(l))} = B_{21}B_{12}(\boldsymbol{X}(l)), \quad \text{(say)} \tag{8}$$

corresponding to a training sample $\boldsymbol{X}(l)$.

The idea of a training sample has also been used in Atkinson (1978), Geisser and Eddy (1979), Spiegelhalter and Smith (1982), San Martini and Spezzaferri

(1984), and Gelfand, Dey, and Chang (1992). Berger and Pericchi (1996) suggested using only *minimal* training samples $\boldsymbol{X}(l)$ for which the marginals are finite and then taking an average of all the corresponding $B_{21}(l)$s to obtain what is called the intrinsic Bayes factor (IBF). For example, the average can be done arithmetically, leading to

$$\text{AIBF}_{21} = \frac{1}{L}\sum_{l=1}^{L} B_{21}(l) = B_{21}\frac{1}{L}\sum_{l=1}^{L} B_{12}(\boldsymbol{X}(l)) \tag{9}$$

where L denotes the number of minimal training samples used. If the $B_{21}(l)$s for different training sample $\boldsymbol{X}(l)$s vary much, taking an arithmetic average does not seem reasonable and in that case Berger and Pericchi (1996) suggested using the trimmed averages or even the median (complete trimming) of the $B_{21}(l)$s.

The Fractional Bayes Factor

O'Hagan (1995) proposed a solution using a fractional part of the full likelihood in place of using training samples and averaging over them. The resulting "partial" Bayes factor, called the fractional Bayes factor (FBF) is given by

$$\text{FBF}_{21} = \frac{m_2(\boldsymbol{X})}{m_1(\boldsymbol{X})} \cdot \frac{m_1(\boldsymbol{X}, b)}{m_2(\boldsymbol{X}, b)} \tag{10}$$

where

$$m_j(x, b) = \int [f_j(x \mid \boldsymbol{\theta}_j)]^b \pi_j(\boldsymbol{\theta}_j)\, d\boldsymbol{\theta}_j, \tag{11}$$

$m_j(x)$ are as defined in (7) and $0 < b < 1$ is a suitably chosen fraction such that $m_j(\boldsymbol{X}, b)$, $j = 1, 2$ are finite.

In the present situation a model is represented as a partition of the entire sample into a number of groups with distributions given by the density (5) or (6) for the ith group. The expressions for $m(\boldsymbol{X})$ under a model (partition) M are as given in (A1), (A2), (A4), or (A5) of the Appendix depending on the case. The expressions for $m(\boldsymbol{X}, b)$ are given in (A7), (A8), and (A9).

A SAMPLING SCHEME

The observations made during a reconnaissance (first stage) study provides only a broad idea regarding the possible boundaries between rock formations (Fig. 2). For more precise delineation of the boundary, observations on a large

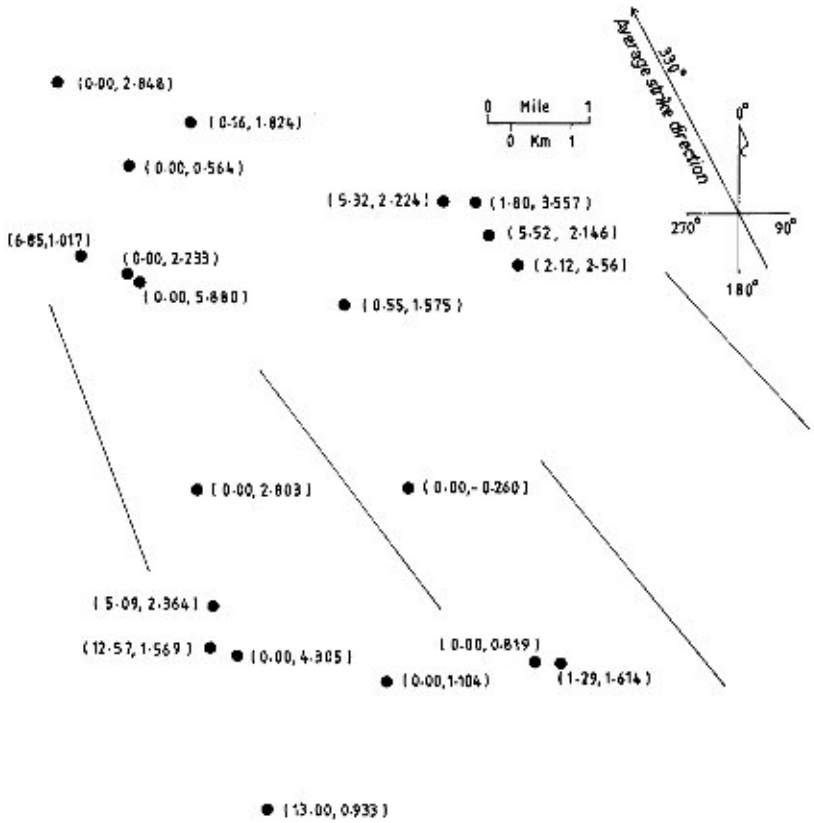

Figure 1. Data set 1: data on (feldspar percentage, mean grain-size) at different location points. Mean grain-size is expressed in ϕ units, where $\phi = -\log_2 d$, d being the grain diameter in mm.

number of locations are required. The sampling scheme, discussed below, allow us to answer the very important questions—(1) how many new samples are to be collected, and (2) where these sampling points are to be located. Such a scheme cannot be discussed or derived without a statistical formulation as presented earlier in the paper.

Our sampling scheme is illustrated with the help of the information on feldspar proportions in the western part of the map (Fig. 1 and Table 1). Suppose that we have already obtained a band for the boundary in the western part of the map from our first stage observations (Fig. 2). We consider a finite set of uniformly spaced possible boundaries, say, $\{B_1, B_2, \ldots, B_k\}$ (from which we are to make a choice) in the neighbourhood of this band and also a prior probability distribution on this set. In the present situation, utilizing the strike directions, it is possible to choose a set of nonintersecting boundaries.

Table 1. Observations on Feldspar Proportion, Labelled According to the Trend of the Bed (Data Set 1)

Serial number	Feldspar proportion	Serial number	Feldspar proportion
1	0.1300	11	0.0
2	0.1257	12	0.0
3	0.0685	13	0.0
4	0.0509	14	0.0016
5	0.0	15	0.0055
6	0.0	16	0.0129
7	0.0	17	0.0532
8	0.0	18	0.0180
9	0.0	19	0.0552
10	0.0	20	0.0212

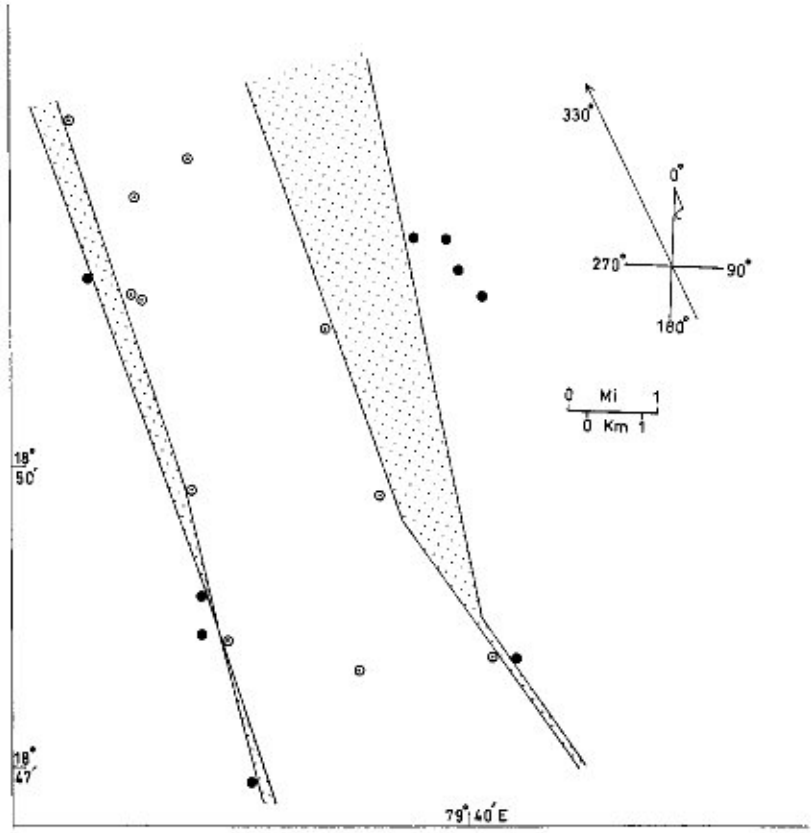

Figure 2. Bands for the boundaries corresponding to Data set 1.

Let $\boldsymbol{X}$ denote the data at the first stage. Given a boundary B_i, the data points $\boldsymbol{X}$ are divided into two groups. As described in the Appendix [see (A1)], we can find the integrated likelihood $m(\boldsymbol{X} \mid B_i)$ (say) for the partition corresponding to each B_i, $i = 1, \ldots, k$. From these one can find the marginal density $m(x)$ and the posterior probabilities for each of the boundaries. For a uniform prior over the set of boundaries (which will be used here) we have

$$m(x) = \frac{1}{k} \sum_{j=1}^{k} m(x \mid B_j)$$

and the posterior probability of B_i is

$$p_i = p_i(\boldsymbol{X}) = \frac{m(\boldsymbol{X} \mid B_i)}{\sum_{j=1}^{k} m(\boldsymbol{X} \mid B_j)}, \quad i = 1, 2, \ldots, k. \tag{12}$$

We consider a loss $L(B, \hat{B}) = \|B - \hat{B}\|$, the area between a choice $\hat{B}$ of the boundary and the true boundary B (within the area to be mapped). This is similar to the loss considered in Switzer (1967). The posterior risk of $\hat{B}$ is then given by

$$\sum_{i=1}^{k} p_i \|B_i - \hat{B}\|.$$

Thus the posterior risk of $\hat{B}$ is the expected area that is misclassified when $\hat{B}$ is chosen to be the boundary.

A (Posterior) Bayes choice of the boundary is one for which the above posterior risk is minimum with respect to $\hat{B}$ and this minimum value of the posterior risk is called the (posterior) Bayes risk. In case of nonintersecting boundaries, if $B_1, B_2, \ldots, B_k$ denote the consecutive possible boundaries starting from the west, then the (posterior) Bayes choice is a sort of median B_r where r is a positive integer such that $\sum_{i=1}^{r-1} p_i < 1/2$ and $\sum_{i=1}^{r} p_i \geq 1/2$. The (posterior) Bayes risk is then given by

$$\mathrm{BR}(\boldsymbol{X}) = \sum_{i=1}^{k} p_i \|B_i - B_r\|. \tag{13}$$

Note that the area of the band for the boundary (within the area to be mapped) obtained with the first stage observations is proportional to the corresponding Bayes risk. For example, if we take five possible boundaries (as taken in our illustration), this Bayes risk is approximately 0.3 times the area of the band.

We now consider the problem of finding optimally the number of second stage location points and their location given the first stage observations $\boldsymbol{X}$. For a particular choice of s (say) locations for the second stage, let $Y_1, \ldots, Y_s$ denote the corresponding feldspar proportions and also let $\boldsymbol{X}' = (\boldsymbol{X}, Y_1, \ldots, Y_s)$. Using the expressions for the integrated likelihoods $m(\boldsymbol{X}' \mid B_i)$s given in (A1) we can find the predictive density $m(y_1, \ldots, y_s \mid \boldsymbol{X})$ of the second stage (future) observations $Y_1, \ldots, Y_s$ given the first stage observations $\boldsymbol{x}$:

$$m(\boldsymbol{y} \mid \boldsymbol{X}) = \frac{\sum_{i=1}^{k} m(\boldsymbol{X}, \boldsymbol{y} \mid B_i)}{\sum_{i=1}^{k} m(\boldsymbol{X} \mid B_i)}.$$

We can also calculate the posterior probabilities $p_i(\boldsymbol{X}')$ of the boundaries given $\boldsymbol{X}'$ and the (posterior) Bayes risk $\mathrm{BR}(\boldsymbol{X}')$ (given $\boldsymbol{X}'$) that is given by (13) with p_i replaced by $p_i(\boldsymbol{X}')$. Since the Bayes risk $\mathrm{BR}(\boldsymbol{X}')$ depends on $\boldsymbol{Y}$ which is yet unobserved, we find the average Bayes risk, averaged with respect to the predictive distribution $m(\boldsymbol{y} \mid \boldsymbol{X})$. It is, however, difficult to integrate with respect to the distribution $m(\boldsymbol{y} \mid \boldsymbol{X})$. So, we proceed as follows. We draw samples $\boldsymbol{y}^{(j)} = (y_1^{(j)}, \ldots, y_s^{(j)})$ from the distribution $m(\boldsymbol{y} \mid \boldsymbol{X})$ and calculate $\mathrm{BR}(\boldsymbol{X}, \boldsymbol{y}^{(j)})$ for $j = 1, 2, \ldots, N$ and then find $\frac{1}{N}\sum_{j=1}^{N} \mathrm{BR}(\boldsymbol{X}, \boldsymbol{y}^{(j)})$, which serves as an approximation to the average Bayes risk for large enough N.

Simulation from the distribution $m(\boldsymbol{y} \mid \boldsymbol{X})$ is possible. We can draw samples $y_1, y_2, \ldots, y_s$ one by one. The function $m(y_1 \mid \boldsymbol{X})$ will be of the form.

$$\begin{aligned} m(y_1 \mid \boldsymbol{X}) &= p && \text{if } y_1 = 0 \\ &= \sum_{i=1}^{k} w_i \frac{1}{y_1(1-y_1)} \frac{1}{\left(Q_i\left(\log \frac{y_1}{1-y_1}\right)\right)^{\lambda_i}} && \text{if } y_1 > 0, \end{aligned} \tag{14}$$

where $Q_i(z)$ are quadratic expressions in z, the constants p_i, w_i, λ_i, and the constants in the quadtratic Q_i depend on $\boldsymbol{X}$. For an appropriate linear function Z of $\log(y_1/(1-y_1))$, the density $m(y_1 \mid \boldsymbol{X})$ on $y_1 > 0$ can be written as

$$\sum_{i=1}^{k} w_i \frac{c(\lambda_i)}{(1+z^2)^{\lambda_i}}, \quad \lambda_i \geq 1,$$

and then sampling can be done, for example, by Acceptance–Rejection method (see, e.g., Fishman, 1978, p. 399).

For several possible choices of the second stage location points we now find the average Bayes risk on the basis of which we can make our final choice of the second stage locations. For example, given a threshold we find the number of

location points (s) and their location such that the average Bayes risk is less than or equal to the given threshold. Alternatively, given a threshold we find the second stage location points such that the probability of the posterior Bayes risk (not averaged) being less than or equal to the threshold is at least 0.99 or so. It can be shown that as the number of second stage units, suitably distributed, increases, the average Bayes risk can be made as small as we wish. So in principle our methods will work.

For location of the boundary in the western part of the map (Fig. 2) we first choose a set of possible boundaries suggested by the corresponding band and the local strike directions. Indeed, we have chosen five (nonintersecting) uniformly spaced piecewise linear curves in the neighbourhood of the band. Let $B_1, \ldots, B_5$ denote these five possible choices of the unknown boundary (starting from the west). Given $B_1, \ldots, B_5$, the whole region is divided into six subregions $R_1, \ldots, R_6$ where R_1 and R_6 denote respectively the regions to the left of B_1 and to the right of B_5 and $R_2, \ldots, R_5$ are the regions in between consecutive boundaries. We represent a particular choice of the second stage observation points by the six-tuple $(s_1, \ldots, s_6)$, s_i being the number of points in R_i, because the average Bayes risk for a particular choice is the same as long as $(s_1, \ldots, s_6)$ remains the same. Corresponding to different values of the total number of second stage observation points s, we calculate the average Bayes risk for all possible allocations $(s_1, \ldots, s_6)$ and can find the (optimum) allocation with minimum average Bayes risk.

NUMERICAL CALCULATIONS

We begin this section by presenting a description of our method in the form of an algorithm.

Our method is applicable when candidate boundaries may be approximated by piecewise linear curves with slopes determined by local strike directions. The data consist of mineral proportions, like those of feldspar and quartz, and also the mean grain-sizes in rock samples collected at different location points. Below we use either the univariate data on feldspar proportion or the bivariate data on feldspar proportion and mean grain-size. The successive steps of our method are as follows.

1. The first step of our method consists of labelling the data along the trend of the bed. If we have data for n rock samples, we assign labels $1, 2, \ldots, n$ to them. To do this we think of a continuum of (nonintersecting) piecewise linear curves, fixed by local strike directions, We then proceed along a line traversing all of them, and number the data points as they appear one after another on the successive piecewise linear curves.

A partition of the data into k groups may then be specified by assigning n_1 of these points to the first group, n_2 to the second group, etc., with $\sum_{l=1}^{k} n_l = n$.

Usually, but not necessarily, the n_1 points of the first group will be the "first" n_1 points with labels $1, 2, \ldots, n_1$, the second group will have points with labels $n_1 + 1, \ldots, n_1 + n_2$ and so on. In this case we specify the corresponding partition by the k-tuple $(n_1, n_2, \ldots, n_k)$ with $0 \leq n_i \leq n$, $i = 1, 2, \ldots, k$ and $\sum_{i=1}^{k} n_i = n$. In other cases, like the one appearing in the fourth row of Table 6, we specify each group of a partition by mentioning the labels of the observations falling in that group.

Each partition specifies the distribution of the observed data as random variables, that is, it specifies a "model."

2. We consider next all possible partitions of the form $(n_1, n_2, \ldots, n_k)$ for different values of k, where each group consists of data points having consecutive labels. Each of these partitions corresponds to a model M for which we calculate the value of the integrated likelihood using one of the expressions (A1), (A2), (A4), and (A5) depending on the case. If we have zero feldspar proportions for some of the rock samples, we use (A1) for univariate data and (A4) for bivariate data; otherwise we use (A2) and (A5) for univariate and bivariate data, respectively.

3. The next step consists of identifying the partitions having the highest values of the integrated likelihood for each k, $k = 1, 2, \ldots, k_0$ for some appropriate k_0 that is the geologist's choice. For the data we have analyzed, k_0 is taken to be 3. These give us the *best* partitions, and accordingly boundaries, for each k.

It should be noted here that our mode of grouping the data depends heavily on the way the candidate boundaries are fixed by local strike directions. It is quite possible that actual boundary may show a moderate deviation from strike direction. To cope with this possibility we allow a few more partitions as follows. For each k, we first consider all the partitions of the form $(n_1, n_2, \ldots, n_k)$ and choose top few partitions from them. Each of these chosen partitions $(n_1, n_2, \ldots, n_k)$ suggests approximate location of $k - 1$ boundaries. For each of these $k - 1$ boundaries, we reallocate the location points, close to it, to anyone of the subregions on both sides of it and thus obtain a number of new partitions. We calculate integrated likelihood for all these partitions and choose those with highest values of the integrated likelihood.

4. We address next the issue of choosing the *best* partitions over the choices of k we have made. This involves pairwise comparison of partitions that correspond to models of differing dimensions. Suppose that we want to compare two partitions P_1 and P_2 that correspond to two models M_1 and M_2 where the number of groups in P_1 is less than that in P_2. We calculate the (adjusted) ratio of likelihood given by default Bayes factors such as IBF or FBF. *If a Bayes factor of M_2 to M_1 is bigger than* 1, *we accept M_2 (P_2), otherwise we accept M_1 (P_1).* In the present context, the FBF is easier to calculate because it does not involve choices of "training samples" as in the IBF approach. The readers interested only in FBF can omit the next three paragraphs.

The IBF Approach

This approach involves the notion of minimal training samples. These are minimal subsets of the whole data set for which the integrated likelihoods are finite. For each minimal training sample $\boldsymbol{X}(l)$, we calculate the conditional Bayes factor (CBF) $B_{21}(l)$ using (7) and (8) noting that $m_j(\boldsymbol{X})$ and $m_j(\boldsymbol{X}(l))$ of (7) and (8) are respectively the integrated likelihoods for the whole data $\boldsymbol{X}$ and for a training sample $\boldsymbol{X}(l)$ under M_j; the integrated likelihoods are calculated as described in Step 2 with model M_j used as model M. We then calculate the arithmetic average (AIBF), trimmed arithmetic averages and the median (Median IBF) of the CBFs $B_{21}(l)$.

To compute the IBFs, we need to choose minimal subsets $\boldsymbol{X}(l)$ of the whole data $\boldsymbol{X}$ for which the integrated likelihoods $m_j(\boldsymbol{X}(l))$ are finite. Consider, for example, the univariate data on feldspar. Let k_0 (k_{0j}) denote the number of groups with at least one positive observation under a model M (M_j). From the expression (A1) of the integrated likelihood $m(\boldsymbol{X})$ under a model M, we note that for this to be finite, the total number of positive observations must be greater than k_0. *Thus for comparing two models M_1 and M_2 a training sample $\boldsymbol{X}(l)$ must have* $\max(k_{01}, k_{02}) + 1$ *positive observations and should represent all the groups (corresponding to both the models) containing positive observations.*

In particular, if M_1 is nested in M_2 (i.e., if P_2 is obtained from P_1 by splitting some of its groups), a minimal training sample may consist of two positive observations from one of the groups corresponding to M_2 that contain (at least one) positive observations, and one positive observation from each of the remaining groups containing positive observations.

The FBF Approach

FBF is defined in (10). The values of FBF can be obtained using (10), (11), one of the expressions (A1), (A2), (A4), and (A5), and one of the expressions (A7), (A8), and (A9) depending on the case. FBF involves a fraction b. For the case with univariate data, we choose $b = (\max(k_{01}, k_{02}) + 1)/m$ where k_{0i} is the number of groups with at least one positive observation for model M_i and m is the total number of positive observations. For bivariate data we choose $b = (\max(k_{11}, k_{12}) + 2)/m$ where k_{1i} is the number of groups with at least one positive feldspar proportion for model M_i and m is the total number of rock samples with positive feldspar proportions.

5. Comparison of the best partitions for different values of k on the basis of the IBF or FBF yields a best partition from which we obtain approximate location of the boundaries and also the corresponding bands for the boundaries using the available strike directions.

One can write computer programs for the calculations involved in the steps mentioned above. A few FORTRAN programs are available from the authors (SP or TS).

As examples of how our methods work in practice, we now present some of the numerical calculations we have done with two sets of data.

Example 1. We first consider a real data set collected by one of us (SS) from the Godavari Valley, India. Figure 1 gives a plot of the 20 location points at which rock samples were collected, with data on feldspar percentage and mean grain-size (after a logarithmic transformation) shown against each of these points. Since feldspar proportions were found to have the highest discriminating power, we show our calculations with feldspar data only; the idea is to use the other variables for fine tuning. Table 1 presents the observations on feldspar labelled according to the trend of the bed.

We did our calculations with feldspar data following the steps mentioned above. Moreover, we calculated also the profile likelihood for all the partitions considered using the expressions given in (A3). We choose the partition with the highest integrated likelihood or profile likelihood. It was expected that the methods with profile likelihood and integrated likelihood would yield similar results since in a sense one is an approximation to the other. Interestingly, the partitions are ordered exactly in the same order of preference on the basis of both the integrated likelihoods and the profile likelihoods calculated from the observed data.

Table 2 shows the top two preferred partitions, for both the case with one boundary and the case with two boundaries, together with the corresponding values of the integrated likelihood and logarithm of the profile likelihood.

Note that all models (partitions) corresponding to the case with one boundary that we considered are of the same dimension, and therefore, comparison in this case can be made on the basis of the ratio of the integrated likelihoods (BF). Thus we select the partition (13,7) if only one boundary is assumed.

On the other hand the models corresponding to the case with two boundaries are not all of the same dimension. Indeed, the model (4,11,5) is of higher dimension

Table 2. The Integrated Likelihood and the Logarithm of the Profile Likelihood for the Feldspar Data (Data Set 1)

Partitions	Profile loglikelihood	Integrated likelihood
All observations in a group (no boundary)	7.52	290.60
Case with one boundary		
(13,7)	16.63	480755.49
(4,16)	13.69	39591.63
Case with two boundaries		
(4,11,5)	25.69	45.43×10^7
(4,9,7)	24.65	9.6247×10^7

Table 3. BFs, IBFs, and FBFs of (4,11,5) Relative to the Other Models Considered (Data Set 1)

Models	(4,11,5) to (4,9,7)	(4,11,5) to (13,7)	(4,11,5) to (4,16)
BF	4.72	944.97	11,474.64
IBF—AM	0.33	65.13	790.92
IBF—10% trimmed mean	1.08	215.85	2620.98
IBF—20% trimmed mean	1.14	228.32	2772.45
IBF—median	1.15	231.02	2805.23
FBF	5.38	285.74	1687.77

than the model (4,9,7). Thus in order to make a final choice we are to calculate the IBFs of (4,11,5) and (4,9,7) relative to (13,7) and (4,16) and of (4,11,5) relative to (4,9,7). Now it is intuitively clear that comparison of the model (4,9,7) with (13,7) or (4,16) may be based on the ratio of the integrated likelihoods (BF) because the former has only one additional α-parameter for which proper prior is used. Indeed, it can be easily seen that the IBF of (4,9,7) relative to (13,7) or (4,16) is the same as the corresponding BF.

From Table 2 we can conclude that (4,9,7) is clearly preferred over (13,7) and (4,16), the corresponding BF (or IBF) being 200.2 and 2431.0, respectively. Table 3 presents the values of IBFs and FBFs of (4,11,5) relative to (4,9,7), (13,7), and (4,16). We have also presented 10 and 20% trimmed means of the $B_{21}(l)$s and also the median (complete trimming). The models (4,11,5) and (4,9,7) are preferred over the other models we have considered.

Remark 2. Choice of the training samples for calculation of IBF in the general case with special mention to the nested case is described above in Step 4 of the algorithm. When we compare (4,11,5) with (4,9,7) or (13,7) we are in a nonnested case and we form training samples in such a way that all the groups of (4,11,5) (and hence the groups of (4,9,7) or (13,7) with positive observations) are represented. Here a training sample must contain four positive observations and we take, in a training sample, one or two (positive) observations from the first group of (4,11,5), one of the two positive observations from the second group, and two or one (positive) observations from the third group.

As an illustration of the use of our sampling scheme proposed in the previous section, we consider, as an example, the problem of locating a boundary in the western portion of the map (Fig. 2) for the feldspar data of Table 1.

As mentioned in the last paragraph of the previous section, we have calculated the average Bayes risk for all possible allocations of the second stage observation points corresponding to different values of s (the total number of second stage observation points). Table 4 presents the best possible allocations of s second stage observation points and the corresponding average Bayes risks (in sq. km) for different values of s. For example, if we want to sample from six more location

Table 4. Best Possible Allocations of s Second Stage Sample Points and Average Bayes Risks (in sq. km) Corresponding to Diferent Values of s (Data Set 1)

s	Best possible allocation	Average Bayes risk for the best allocation
5	(0, 0, 3, 2, 0, 0)	0.8601
6	(0, 1, 3, 2, 0, 0)	0.7750
7	(0, 1, 3, 2, 1, 0)	0.7014
8	(1, 1, 3, 2, 1, 0)	0.6394
9	(1, 1, 3, 3, 1, 0)	0.5883
10	(1, 1, 4, 3, 1, 0)	0.5469
15	(3, 0, 5, 5, 2, 0)	0.4014
30	(4, 1, 11, 10, 4, 0)	0.2215

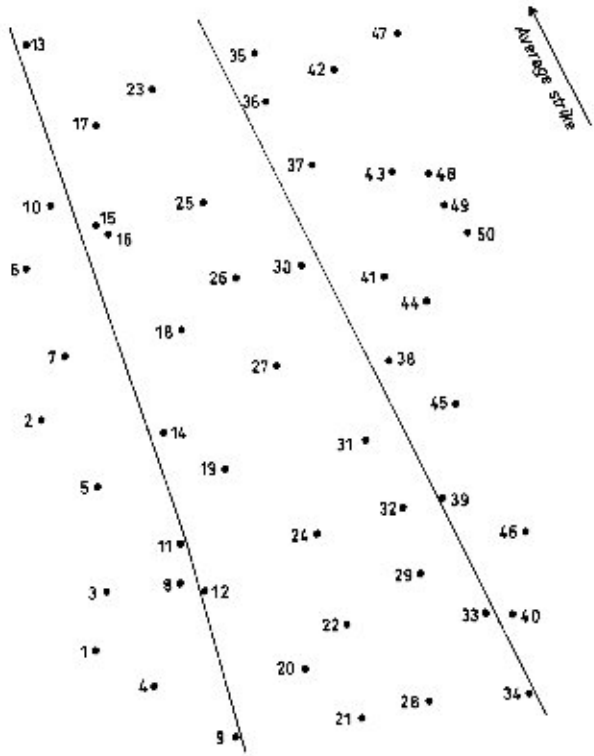

Figure 3. Fifty location points and boundaries corresponding to Data set 2.

points, the optimum allocation is (0,1,3,2,0,0,), that is, we collect one rock sample from region R_2, three from R_3, and two from R_4 where R_i are as defined in the last paragraph of the previous section. If we require the expected area of misclassified region to be less than 0.3 sq. km, we need to sample 30 points (see last row of Table 4).

Example 2. We now demonstrate applicability of our method on a larger and more complicated data set. For this second data set we assume the same strike directions (i.e., the same trend of the bed) as in the preceding example, and 50 location points (as opposed to 20 in the previous case) and generate data on feldspar proportion and mean grain-size using information available from the real data set we have. Figure 3 gives a plot of these 50 location points scattered over the area. The data on feldspar proportion and mean grain-size, after a logarithmic transformation, for 50 location points are presented in Table 5. We identified the probable partitions as

Table 5. Observations on Feldspar Proportion and Mean Grain-Size, Labelled According to the Trend of the Bed (Data Set 2)

Serial number	Feldspar proportion	Mean grain-size	Serial number	Feldspar proportion	Mean grain-size
1	0.1121	1.141	26	0.0760	1.926
2	0.1014	1.134	27	0.0703	2.024
3	0.0955	1.145	28	0.0725	2.134
4	0.0930	1.102	29	0.0674	1.673
5	0.0922	1.108	30	0.0812	1.935
6	0.0884	1.047	31	0.0811	1.910
7	0.0862	0.963	32	0.0821	1.772
8	0.0778	1.118	33	0.0831	1.876
9	0.0790	1.014	34	0.0819	1.863
10	0.0829	1.127	35	0.0879	1.249
11	0.0829	1.117	36	0.0877	1.255
12	0.0811	1.743	37	0.0894	1.367
13	0.0813	1.736	38	0.0826	0.741
14	0.0821	1.789	39	0.0797	1.124
15	0.0887	1.982	40	0.0819	1.116
16	0.0788	1.765	41	0.0938	1.022
17	0.0744	1.477	42	0.0921	1.288
18	0.0864	1.938	43	0.0889	0.834
19	0.0725	1.843	44	0.0926	1.257
20	0.0703	1.732	45	0.0847	0.819
21	0.0689	1.653	46	0.0879	1.065
22	0.0739	2.047	47	0.1042	1.262
23	0.0657	1.650	48	0.1107	1.454
24	0.0642	1.432	49	0.0979	1.339
25	0.0877	2.031	50	0.0935	1.282

Note. Mean grain-size is expressed in ϕ units, where $\phi = -\log_2 d$, d being the grain diameter in mm.

Table 6. Integrated Likelihoods of Most Preferred Partitions for Data Set 2

Partitions	Integrated likelihood
Case with one boundary	
(34,16)	1.487×10^{60}
(1, . . . , 34, 39; 35, . . . , 38, 40, . . . , 50)	1.217×10^{60}
Case with two boundaries	
(11,23,16)	8.205×10^{77}
(1, . . . , 11; 12, . . . , 34, 37; 35, 36, 38, . . . , 50)	1.599×10^{75}

above and calculated the corresponding integrated likelihoods using the expression given in (A5). Table 6 shows the top two preferred partitions, for the cases with one boundary and two boundaries, together with the corresponding values of the integrated likelihood. In order to compare the best three-group partition (11,23,16) with the best two-group partition (34,16) we calculated the AIBF and FBF of (11,23,16) relative to (34,16), which were found to be 3.1286×10^{16} and 5.2397×10^{16}, respectively. Thus the most preferred partition is (11,23,16).

ACKNOWLEDGMENTS

The idea of applying statistical techniques to problems of geological mapping stems from a notion maintained by Professor P. C. Mahalanobis. We thank Professor James O. Berger, Dr Dilip Saha, and the anonymous referees of an earlier version for their very helpful comments and suggestions. We also thank Mr. Ajay Kumar Das, Geological Studies Unit, ISI, for drawing the figures. The work is supported by the Council of Scientific and Industrial Research, Extra Mural Research Division, India.

REFERENCES

Aitchison, J., 1986, The statistical analysis of compositional data: Chapman and Hall, London, 416 p.

Atkinson, A. C., 1978, Posterior probabilities for choosing a regression model: Biometrika, v. 65, p. 39–48.

Berger, J. O., and Pericchi, L. R., 1996, The intrinsic Bayes factor for model selection and prediction: J. Am. Stat. Assoc., v. 91, p. 109–122.

Berger, J. O., Pericchi, L. R., and Varshavsky, J. A., 1998, Bayes factors and marginal distributions in invariant situations: Sankhya, Ser. A, v. 60, p. 307–321.

Fishman, G. S., 1978, Principles of discrete event simulation: John Wiley and Sons, New York, 514 p.

Geisser, S., and Eddy, W. F., 1979, A predictive approach to model selection: J. Am. Stat. Assoc., v. 74, p. 153–160.

Gelfand, A. E., Dey, D. K., and Chang, H., 1992, Model determination using predictive distributions with implementation via sampling-based methods (with discussion), *in* Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., eds., Bayesian statistics, Vol. 4: Oxford University Press, London, p. 147–167.

Ghosh, J. K., Saha, M. R., and Sengupta, S., 1981, Gondwana stratigraphic classification by statistical method, *in* Merriam, D. F., ed., Down-to-earth statistics: Solutions looking for geological problems: Syracuse University Geology Contributions, Syracuse, New York, p. 47–62.

O'Hagan, A., 1995, Fractional Bayes factor for model comparisons: J. R. Stat. Soc., Ser. B, v. 57, p. 99–138.

San Martini, A., and Spezzaferri, F., 1984, A Predictive model selection criterion: J. R. Stat. Soc., Ser. B, v. 46, p. 296–303.

Sengupta, S., 1970, Gondwana sedimentation around Bheemaram (Bhimaram), Pranhita–Godavari valley, India: J. Sed. Pet., v. 40, p. 140–170.

Sengupta, S., Ghosh, J. K., and Mazumder, B. S., 1991, Experimental-theoretical approach to interpretation of grain size frequency distributions, *in* Syvitski, J. P. M., ed., Principles, methods, and application of particle size analysis: Cambridge University Press, Cambridge, UK, p. 264–279.

Spiegelhalter, D. J., and Smith, A. F. M., 1982, Bayes factor for linear and log-linear models with vague prior information: J. R. Stat. Soc., Ser. B, v. 44, p. 377–387.

Switzer, P., 1967, Reconstructing patterns from sample data: Ann. Math. Stat., v. 38, p. 138–154.

APPENDIX: EXPRESSIONS FOR THE LIKELIHOODS

In order to make a choice of a partition or model we need to find the integrated or profile likelihoods of the possible partitions. Expressions for these likelihoods are given below.

Univariate Data on Feldspar

Let us consider a partition with n_i observations X_{ij}, $j = 1, 2, \ldots, n_i$, in the ith group, $i = 1, 2, \ldots, k$ and let M be the corresponding model.

Let m_i be the number of positive observations in the ith group, I be the set of all i $(i = 1, 2, \ldots, k)$ for which the ith group contains at least one positive observation, k_o = number of elements in the set I, $m = \sum m_i$, $n = \sum n_i$, and $\boldsymbol{X}$ denote the whole data set.

Integrated Likelihood

The expressions for the integrated likelihood under model M, denoted by $m(\boldsymbol{X})$, are given below.

Case 1. Some of the observations are zero.

$$m(\boldsymbol{X}) = \prod_{i=1}^{k} \frac{(m_i)!(n_i - m_i)!}{(n_i + 1)!} \frac{1}{\prod_{i,j:X_{ij}>0}(1 - X_{ij})X_{ij}} \times \prod_{i \in I} \frac{1}{\sqrt{m_i}} \frac{1}{(\pi)^{\frac{m-k_o}{2}}} \frac{\Gamma\left(\frac{m-k_o}{2}\right)}{2(\mathrm{TCSS})^{\frac{m-k_o}{2}}} \tag{A1}$$

where

$$\mathrm{TCSS} = \sum_{i \in I} \mathrm{TCSS}(i) = \sum_{i \in I} \sum_{\substack{j=1 \\ X_{ij}>0}}^{n_i} \left(\log \frac{X_{ij}}{1-X_{ij}} - \frac{1}{m_i} \sum_{\substack{j=1 \\ X_{ij}>0}}^{n_i} \log \frac{X_{ij}}{1-X_{ij}} \right)^2 .$$

We note that the expression in (A1) is finite only if $m > k_o$.

Case 2. All the observations are positive.

$$m(\boldsymbol{X}) = \frac{1}{\prod_{i,j}(1-X_{ij})X_{ij}} \prod_{i=1}^{k} \frac{1}{\sqrt{n_i}} \frac{1}{(\pi)^{\frac{n-k}{2}}} \frac{\Gamma\left(\frac{n-k}{2}\right)}{2(\mathrm{TCSS})^{\frac{n-k}{2}}} \tag{A2}$$

where TCSS is as in (A1) with $m_i = n_i$ and $I = \{1, 2, \ldots, k\}$.

Profile Likelihood

Maximizing the likelihood with respect to the parameters we have

$$\log L^*(\boldsymbol{X}) = \sum_{i=1}^{k} \left\{ m_i \log \frac{m_i}{n_i} + (n_i - m_i) \log \left(1 - \frac{m_i}{n_i}\right) \right\} - \frac{m}{2} \log \left(\frac{2\pi}{m}\right) - \sum_{i,j:X_{ij}>0} \log[X_{ij}(1-X_{ij})] - \frac{m}{2} - \frac{m}{2} \log[\mathrm{TCSS}] \tag{A3}$$

where L^* denotes the maximized likelihood.

Bivariate Data

We now consider the bivariate data on feldspar proportion (X) and logarithm of mean grain-size (Y).

Let us consider a partition with n_i pairs of observations (X_{ij}, Y_{ij}), $j = 1, 2, \ldots, n_i$, in the ith group, $i = 1, 2, \ldots, k$ and let M be the corresponding model.

Let m_i = number of pairs in the ith class with positive X-observation, I_1 is the set of all i ($i = 1, 2, \ldots, k$) for which $m_i > 0$, I_2 is the set of all i ($i = 1, 2, \ldots, k$) such that $m_i < n_i$, k_r = number of elements in the set I_r, $r = 1, 2, n = \sum n_i$, $m = \sum m_i$,

$$S_1 = \sum_{i \in I_2} \sum_{j:X_{ij}=0} (Y_{ij} - \hat{\xi}_i)^2,$$

$$S_2 = \sum_{i \in I_1} \sum_{j:X_{ij}>0} (\boldsymbol{U}_{ij} - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{U}_{ij} - \hat{\boldsymbol{\mu}}_i)',$$

$\hat{\xi}_i = \frac{1}{n_i - m_i} \sum_{j:X_{ij}=0} Y_{ij}$, $\boldsymbol{U}_{ij} = (\log(X_{ij}/1 - X_{ij}), Y_{ij})'$, and $\hat{\boldsymbol{\mu}}_i = \frac{1}{m_i} \sum_{j:x_{ij}>0} \boldsymbol{U}_{ij}$.

Integrated Likelihood

Case 1. Some of the feldspar proportions are zero. The integrated likelihood $m(\boldsymbol{X}, Y)$ under model M is given by

$$m(\boldsymbol{X}, Y) = \prod_{i=1}^{k} \frac{(m_i)!(n_i - m_i)!}{(n_i + 1)!} \prod_{i \in I_1} \left(\frac{1}{m_i}\right) \prod_{i \in I_2} \frac{1}{\sqrt{n_i - m_i}}$$
$$\times (\pi)^{-(n+m-2k_1-k_2-1)/2} \Gamma\left(\frac{n-m-k_2}{2}\right) \frac{1}{\prod_{i,j:X_{ij}>0}(1 - X_{ij})X_{ij}}$$
$$\times \frac{1}{2}\Gamma\left(\frac{m-k_1}{2}\right)\Gamma\left(\frac{m-k_1-1}{2}\right) S_1^{-(n-m-k_2)/2} |S_2|^{-(m-k_1)/2} \quad \text{(A4)}$$

Case 2. All the feldspar proportions are positive. Here

$$m(\boldsymbol{X}, Y) = \frac{1}{\prod_{i,j}(1 - X_{ij})X_{ij}} \frac{1}{\pi^{n-k-1/2}} \prod_{i=1}^{k} \left(\frac{1}{n_i}\right) \Gamma\left(\frac{n-k}{2}\right)$$
$$\times \Gamma\left(\frac{n-k-1}{2}\right) |S_2|^{-(n-k)/2} \quad \text{(A5)}$$

where n_i and S_2 are as in (A4) with $m_i = n_i$ and $I_1 = \{1, 2, \ldots, k\}$.

Profile Likelihood

Maximizing the likelihood with respect to the parameters we obtain the profile likelihood $L^*(\boldsymbol{X}, Y)$ that is given by

$$\log L^*(\boldsymbol{X}, Y) = \sum_{i=1}^{k} \left\{ m_i \log\left(\frac{m_i}{n_i}\right) + (n_i - m_i) \log\left(1 - \frac{m_i}{n_i}\right) \right\}$$
$$- \sum_{i \in I_1} \sum_{j:X_{ij}>0} \log(X_{ij}(1 - X_{ij})) - \frac{n+m}{2} \log 2\pi$$
$$- \frac{n+m}{2} - \frac{m}{2} \log\left|\frac{S_2}{m}\right| - \frac{(n-m)}{2} \log\left(\frac{S_1}{n-m}\right). \quad \text{(A6)}$$

where $m_i, n_i, m, n, I_1, S_1,$ and S_2 are as in (A4).

For calculation of FBF we also need to calculate integral of a fractional part of the likelihood denoted by $m(X, b)$ or $m(X, Y, b)$ defined in (11) where $0 < b < 1$. The expressions for different cases are given below.

Case 1. Univariate data with some zero observations.

$$m(\boldsymbol{X}, b) = \prod_{i=1}^{k} \frac{\Gamma(bm_i + 1)\Gamma(b(n_i - m_i) + 1)}{\Gamma(bn_i + 2)} \frac{1}{\prod_{i,j:X_{ij}>0}(1 - X_{ij})^b X_{ij}^b} \times \prod_{i \in I} \frac{1}{\sqrt{m_i}} \frac{1}{(\pi)^{\frac{bm - k_o}{2}}} b^{-bm/2} \frac{\Gamma\left(\frac{bm - k_o}{2}\right)}{2(\mathrm{TCSS})^{\frac{bm - k_o}{2}}} \tag{A7}$$

where $n_i, m_i, m, X_{ij}, k_o, I$, and TCSS are as defined in (A1).

Case 2. Univariate data with all positive observations.

$$m(\boldsymbol{X}, b) = \frac{1}{\prod_{i,j:X_{ij}>0}(1 - X_{ij})^b X_{ij}^b} \prod_{i=1}^{k} \frac{1}{\sqrt{n_i}} \frac{1}{(\pi)^{\frac{bn-k}{2}}} b^{-bn/2} \frac{\Gamma\left(\frac{bn-k}{2}\right)}{2(\mathrm{TCSS})^{\frac{bn-k}{2}}} \tag{A8}$$

where n_i, n, X_{ij}, k, and TCSS are as defined in (A1).

Case 3. Bivariate data.

For the bivariate data analysed in this paper, all the feldspar proportions are positive and we give below the expression of $m(\boldsymbol{X}, Y, b)$ only for this case.

$$m(\boldsymbol{X}, Y, b) = \frac{1}{\prod_{i,j}(1 - X_{ij})^b X_{ij}^b} \frac{1}{\pi^{bn-k-1/2}} \prod_{i=1}^{k} \left(\frac{1}{n_i}\right) \times \Gamma\left(\frac{bn-k}{2}\right) \Gamma\left(\frac{bn-k-1}{2}\right) b^{-k} |bS_2|^{-(bn-k)/2} \tag{A9}$$

where X_{ij}, n_i, n, k, and S_2 are as in (A4) with $m_i = n_i$ and $I_1 = \{1, 2, \ldots, k\}$.