Ternary sandstone composition and provenance: an evaluation of the 'Dickinson model'

G. J. WELTJE

Delft University of Technology, Faculty of Civil Engineering and Geosciences,
Department of Geotechnology, Applied Geology Section, PO Box 5028,
NL-2600GA Delft, The Netherlands (e-mail: g.j.weltje@citg.tudelft.nl)

Abstract: A popular model proposed by W. R. Dickinson and co-workers in the early 1980s relates the composition of sandstones to the plate-tectonic setting of the sedimentary basins in which they were deposited. The present study is devoted to revision and testing of the 'Dickinson model' based on the original data which comprise 11 000 thin sections point-counted by hundreds of different operators over a period of three decades. Statistical analyses based on Aitchison's additive log-ratio transformation are used to obtain an optimal partitioning of ternary compositional spaces into 'provenance fields' and combined with stochastic simulation to assess the success ratio of the optimized 'Dickinson model'. Results indicate that differences between the grand means of each of the three major provenance associations (continental block, magmatic arc and recycled orogen) are highly significant, whereas overall inferential success ratios range from 64% to 78% in the four ternary systems studied. Current methods of dealing with sands of mixed provenance are unsatisfactory. To improve provenance models, the use of ternary subcompositions should be replaced by analyses of the full six-part (Qm, Qp, P, K, Lv, Ls) composition, and their covariance structure could be employed to 'unmix' samples into end-member provenance types.

The idea that sand(stone) composition reflects the nature of the rocks exposed in a source area, as well as the climatic and physiographic regime in which the sand was generated from these rocks, forms the basic premise of sediment provenance studies (Haughton *et al.* 1991; Johnsson 1993; Basu 2003; Weltje & Von Eynatten 2004). The empirical relation between ternary compositions of sands and the plate-tectonic setting of the sedimentary basins in which they were deposited, first explored by Crook (1974) and Schwab (1975), was formally presented by Dickinson and co-workers (Dickinson & Suczek 1979; Dickinson 1982, 1985; Dickinson *et al.* 1983). The 'Dickinson model' (DM) is the first quantitative representation of this key concept in sand provenance studies, whose origins may be traced as far back as the late nineteenth century (Weltje & Von Eynatten 2004). The DM consists of four ternary diagrams subdivided into different provenance fields in which subcompositions of sands may be plotted to infer their most likely plate-tectonic environment (see Table 1 for definitions of compositional variables, ternary subcompositions and plate-tectonic settings used in the DM). The apparently straightforward DM enjoyed great popularity from its inception and has been regarded as a benchmark by at least two generations of sediment petrographers.

In the light of this popularity, it is remarkable that little attention has been paid to tests of its predictive power. The main reason for the lack of such tests is that problems associated with statistical analysis of 'closed-sum' data (i.e. compositions) have been recognized widely but no solution was available until the log-ratio transformation of Aitchison (1982, 1986) became known outside of the field of mathematical statistics. Although the log-ratio transformation is discussed in some detail in recent textbooks on geological data analysis (Rollinson 1993; Swan & Sandilands 1993; Davis 1997; Pawlowsky-Glahn & Olea 2004), one can hardly call it a standard technique, as witnessed by the majority of geological papers in which compositional data are statistically analysed without much regard for their inherent limitations.

Molinaroli *et al.* (1991) attempted to test the DM by means of discriminant function analysis of the ternary QFL and QmFLt data of Dickinson *et al.* (1983) without applying a log-ratio transformation. They concluded that the DM correctly classifies 85% of the data at most. However, this conclusion is difficult to justify from a methodological point of view.

- The intrinsic limitations of compositional data caused by the constant-sum and non-negativity constraints ('closure effects'), which are known to affect the results of discriminant function analysis (e.g. Butler 1982), were not taken into account. It implies that the DM may

From: Bucclanti, A., Mateu-Figueras, G. & Pawlowsky-Glahn, V. (eds) *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Geological Society, London, Special Publications, **264**, 79–99.
0305-8719/06/$15.00 © The Geological Society of London 2006.

Table 1. *Compositional variables, ternary systems and provenance associations referred to in this study*

Grain categories	
Total quartzose grains: Q = Qm + Qp	
Qm = monocrystalline quartz	
Qp = polycrystalline quartz	
Total feldspar grains: F = P + K	
P = plagioclase grains	
K = alkali feldspar grains	
Total unstable lithic fragments: L = Lv + Ls	
Lv = (meta)volcanic lithic fragments	
Ls = (meta)sedimentary lithic fragments	
Total lithic fragments: Lt = L + Qp	
Ternary systems	
QFL	Framework (emphasis on maturity)
QmFLt	Framework (emphasis on parent rock)
QmPK	Subcomposition: monomineralic grains
QpLvLs	Subcomposition: lithic fragments
Provenance associations	
A	Continental block provenance
B	Magmatic arc provenance
C	Recycled orogen provenance
M	Mixed provenance

actually be more powerful if examined in the light of an appropriate statistical model.

- The data used to calculate the success ratio of the empirical classification procedure were also used to establish the discriminant functions. This tends to flatter the results and overestimate the success ratio of the DM, because sampling variability is not taken into account. Tests with independent data are more likely to provide a reasonable assessment of the efficiency of empirical classification schemes, which quite often is disappointingly low (e.g. Armstrong-Altrin & Verma 2005).

In other words, the actual performance of the DM may be better or worse than suggested by the analysis of Molinaroli *et al.* (1991).

In this study, Aitchison's log-ratio approach will be used to analyse the DM database and to obtain the optimal partitioning of ternary compositional space into 'provenance fields'. The final step in the analysis is quantification of the discriminatory power of the DM. The revised DM employs an alternative graphic representation of ternary data, which will be introduced under the term 'log-ratio diagram', but the results have also been transferred to the familiar ternary space to permit a direct comparison with the provenance fields of the original DM (Dickinson 1985).

Statistical analysis of ternary compositions

Many classification schemes developed for sediments employ ternary diagrams (Klein 1963; Okada 1971). The popularity of the ternary diagram, which appears to have been invented in the late nineteenth century (Becke 1897), is most likely attributable to its intuitive appeal. It allows the display of three-part compositions $\mathbf{x} = (x_1, x_2, x_3)$ in a way that treats all components equally, even though one component is redundant because the x-values are non-negative and their sum equals unity (or 100%). The non-negativity and constant-sum constraints represent two fundamental properties of compositional data which are equally relevant to the study of compositions with more than three parts and have frustrated many attempts at statistical analysis, as illustrated by Chayes (1960), Butler (1979), Aitchison (1986), Rollinson (1993) and many others.

The additive log-ratio transformation introduced by Aitchison (1982) is a powerful tool that removes the non-negativity and constant-sum constraints on compositional variables, and permits the use of standard multivariate statistical methods based on the assumption of multivariate normality. It is defined as follows. Let x_i represent the relative abundances of components in a composition made up of k constituents ($1 \leq i \leq k$). The kth component x_k, whose value is fully specified by the sum of the other $k-1$ values, is used as a common denominator (or numerator) to form a series of $k-1$ ratios of component abundances. The logarithms of these ratios are defined as the set of additive log-ratios y_i:

$$y_i = \log\left(\frac{x_i}{x_k}\right) = \log x_i - \log x_k, \quad \text{where } i = 1, 2, \ldots, k-1 \qquad (1)$$

or, alternatively $\mathbf{y} = \mathrm{alr}(\mathbf{x})$

Log-ratios are amenable to rigorous statistical analysis, unlike the constrained compositional variables. They are unconstrained in the sense that they can take on any value, and their values can be modified without automatically forcing a response from the other log-ratios formed from the same composition. Moreover, the outcomes of log-ratio statistical analysis are permutation invariant, i.e. unaffected by the choice of common denominator or numerator. Compositional data follow an additive logistic normal distribution if their log-ratios are multivariate normally distributed. The requirement of additive logistic normality appears to be fulfilled by many types of compositional data. Statistical models for ternary compositions (x_1, x_2, x_3) are thus preferentially constructed under the assumption of a bivariate normal distribution of the corresponding set of log-ratios (y_1, y_2). The results of log-ratio statistical analysis may be mapped back onto the compositional plane for display in a ternary diagram. Mapping is accomplished by the inverse log-ratio transformation, which comprises the following steps. The logistic transformation reimposes the non-negativity constraint:

$$z_i = \begin{cases} e^{y_i} & \text{for } i = 1, 2, \ldots, k-1 \\ 1 & \text{for } i = k \end{cases} \qquad (2)$$

After which the constant sum C is restored:

$$x_i = \frac{C \cdot z_i}{\sum_{i=1}^{k} z_i} \qquad (3)$$

It should be pointed out that the additive log-ratio transformation alr(.) is but one way of approaching the problem. Two drawbacks of alr(.) are its lack of symmetry and orthogonality, which are attributable to the use of a common numerator or denominator. Different forms of log-ratio transformation have been developed to alleviate these problems and to accommodate the ever-widening range of applications in compositional data analysis (Aitchison & Egozcue 2005). The centred log-ratio transformation clr(.) provides a symmetrical treatment of all parts of a composition (Aitchison 1986), whereas the isometric log-ratio transformation ilr(.) was developed to enable statistical analyses on orthonormal coordinates (Egozcue *et al.* 2003). In the present study, alr(.) was used because it leads to a representation of compositional data that is more similar to conventional ratios used in sedimentary petrology than clr(.) and ilr(.), and therefore easier to understand.

Weltje (2002) discussed the construction of confidence regions and predictive regions in ternary diagrams by means of the alr transformation in detail, and illustrated their superiority over the conventional hexagonal fields of variation employed in sedimentary petrology. The term confidence region is reserved for regions of ternary compositional space (or its binary alr-transformed equivalent) in which the fixed population mean is expected to be located with some probability, generally referred to as the confidence level. The term predictive region refers to the population as a whole, i.e. to the region of compositional space in which future observations are expected to be located. The probability associated with a predictive region is termed the content.

Figure 1a shows a set of ternary compositions and the corresponding hexagonal fields of variation calculated from univariate summary statistics. The hexagonal fields are clearly inadequate, because they fail to capture the curvature of the dataset and extend beyond the boundaries of the diagram, which implies the prediction of negative percentage values of one or more of the components. The reason for this unrealistic result is the underlying statistical model which erroneously assumes independent normal distributions of ternary percentage values and does not incorporate unit-sum and non-negativity constraints. Figure 1b illustrates the calculation of true confidence and predictive regions from the alr-transformed data. Such regions are, by definition, elliptical in log-ratio space if the data follow an additive logistic normal distribution. Figure 1c shows the same regions projected onto a ternary diagram by application of the inverse alr-transform (equations (2) and (3)). Note that the curvature of the data points is adequately captured and the region is physically meaningful, because it does not extend beyond the boundaries of ternary compositional space.

The log-ratio diagram

The elliptical shape of confidence regions in log-ratio space offers an attractive alternative to the use of hexagons in ternary space. In many

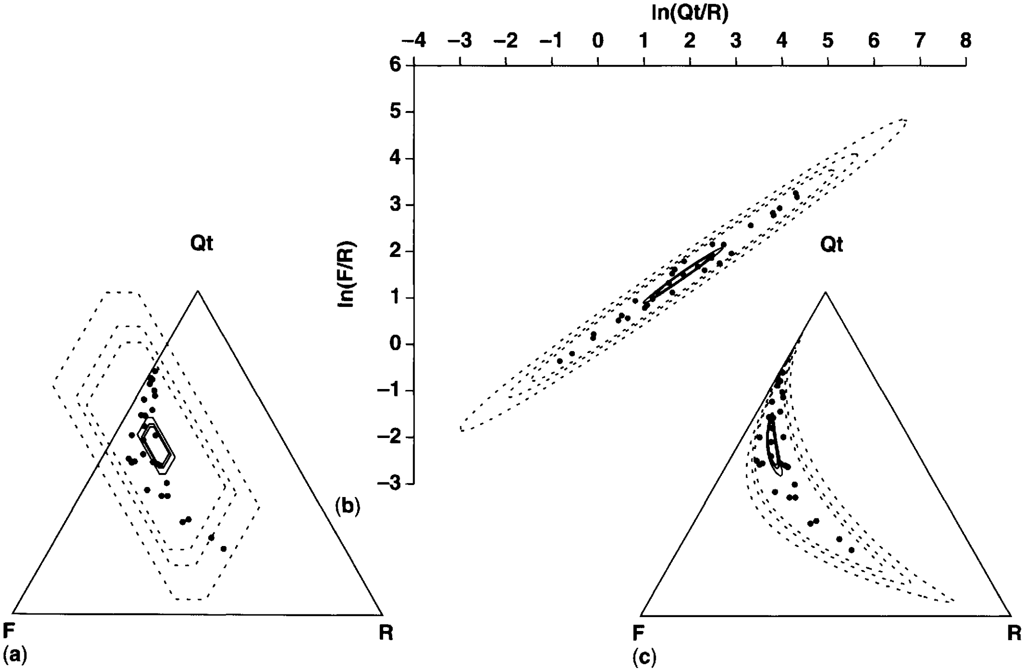

Fig. 1. Hexagonal fields of variation versus alr-based regions. Solid lines: confidence regions of population mean. Dashed line: predictive regions of population. Confidence limits are 90%, 95% and 99%. (**a**) Hexagonal region constructed from intersections of univariate normal approximations; (**b**) alr-based regions in log-ratio space; (**c**) alr-based regions transformed to ternary compositional space (after Weltje 2002).

applications of compositional data analysis, the results of log-ratio statistical analysis are retransformed to percentage data and displayed in a ternary diagram (cf. Figs 1b, c). This inverse transformation has two opposite effects. On the one hand, transformation to the original (conventional) units of measurement makes the results easier to understand. On the other hand, the elegant elliptical shape of confidence regions is lost after transformation to ternary percentages and results may be more difficult to interpret if several partly overlapping regions are plotted in the same ternary diagram. An alternative approach, examined in this study, is to 'map' the log-ratio space that corresponds to the ternary diagram and display the results as ellipses in that space. If researchers become accustomed to this method of display, it could eventually replace the ternary diagram. An ilr-based log-ratio diagram could also be developed as an alternative representation of the alr-based diagram presented in this study.

The ternary diagram of Figure 2a contains three lines from each of the vertices towards the middle of the opposite sides, i.e. lines along which the abundance of one component equals that of another. Because these lines represent constant (log-)ratios, they are also straight lines in log-ratio space (Fig. 2b). This does not apply to fixed-percentage triangles, i.e. lines along which one of the components has a constant value (Fig. 2c). Such triangles are represented by convex, roughly hexagonal shapes in the log-ratio diagram (Fig. 2d). The transformation of fixed-percentage lines reveals another property of the log-ratio diagram: the distance between two lines with values 0.1% and 1% is the same as the distance between the 1% and 10% lines. This geometric scale is a natural result of the logarithmic transformation, and indicates that the log-ratio diagram is much more sensitive to compositional differences in the areas near the edges of the ternary diagram. The opposite holds for areas in the centre of the ternary diagram, as demonstrated by the distances between the 10%, 20% and 30% fixed-percentage lines (Figs 2c and d).

Figure 2e shows a subdivision of the ternary QFL diagram into six equal fields. Each field has been labelled according to the most abundant component (uppercase) and the second-most abundant component (lowercase). This straightforward classification of sands comprises the following types (clockwise from Q vertex): Quartzolithic (Ql), Lithoquartzose (Lq), Lithofeldspathic (Lf),

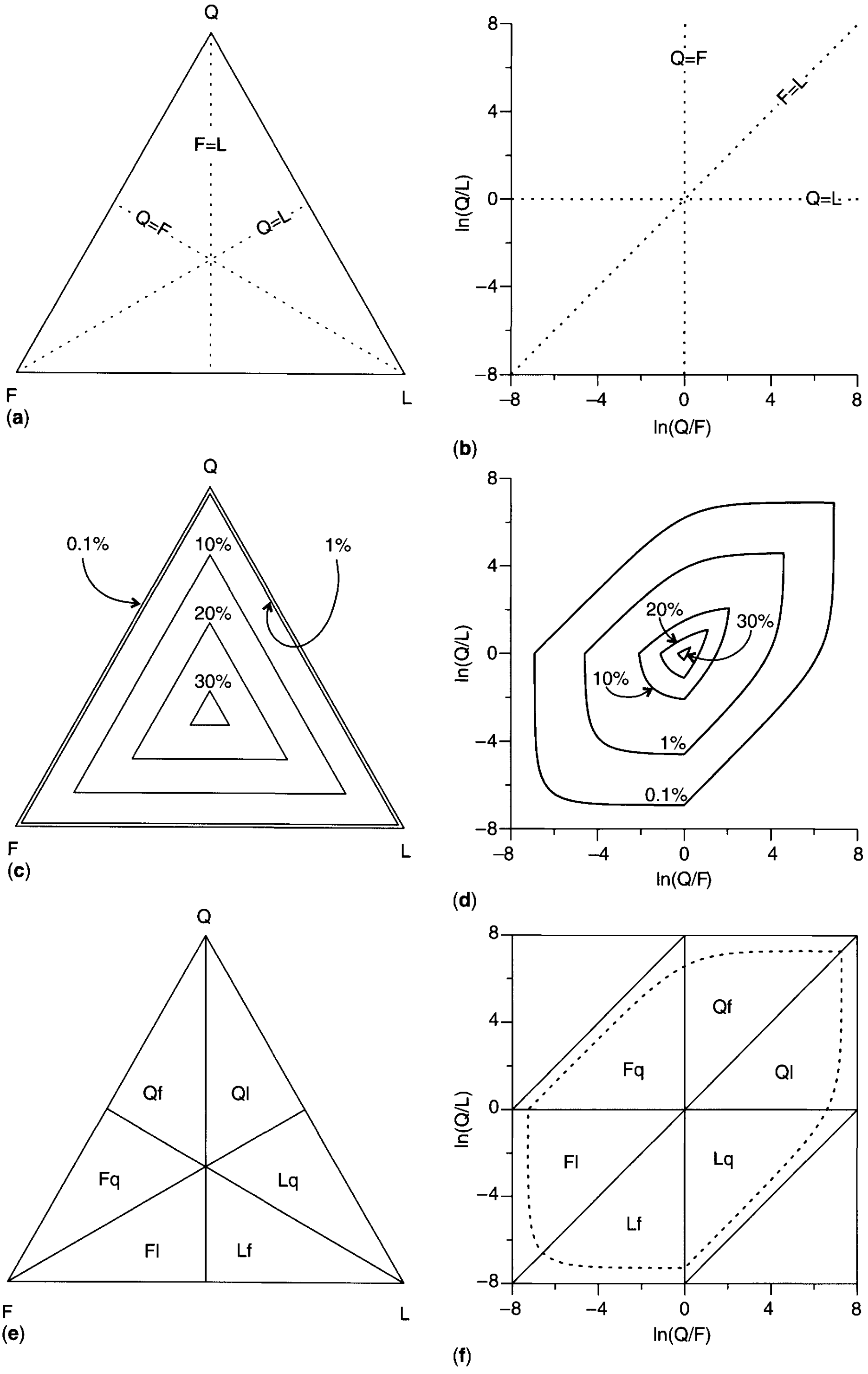

Fig. 2. Exploration of log-ratio space corresponding to the ternary diagram. (**a**, **b**) Straight lines corresponding to fixed (log-)ratios; (**c**, **d**) minimum-percentage contours; (**e**, **f**) sixfold subdivision of compositional space into Quartzolithic (Ql), Lithoquartzose (Lq), Lithofeldspathic (Lf), Feldspatholithic (Fl), Feldspathoquartzose (Fq) and Quartzofeldspathic (Qf) sands. Dashed line is 0.07% contour (see text for discussion).

Feldspatholithic (Fl), Feldspathoquartzose (Fq) and Quartzofeldspathic (Qf). These sand types also occupy similar-sized areas in the part of the log-ratio diagram enclosed by the dashed line (Fig. 2f). This line corresponds to the outer limit of log-ratio space one expects to be covered by composition estimates derived from point counting. The reason for this is that no zero component abundances are allowed in log-ratios (division by zero is not permitted and the logarithm of zero is undefined). If point counting results in zero abundance of one or two components, one therefore has to assume that these zeros reflect sampling error. In other words, some components are present in the population in trace amounts only and have not been observed during point counting. Such zeros must be replaced by statistically acceptable positive values before log-ratio transformation (Aitchison 1986; Weltje 2002; Martín-Fernández *et al.* 2003).

The fixed-percentage line in Figure 2f was calculated by replacing the binary compositions located on the edges of the ternary diagram, i.e. compositions with one zero value, by a ternary composition in which the zero value was replaced by $\delta = 0.07\%$. The non-zero components were multiplied by a factor $(100 - \delta)/100$. The value of δ was obtained from the binomial formula by solving for the case in which an analyst who counts 1000 points fails to record a rare component and assumes the probability of failure to be 50%. In other words, the analyst assumes that the unsampled component is so rare it will be recorded only in half of the point counts of this length, if the procedure was repeated many times. The number of points counted by this hypothetical analyst is much larger than customary in sedimentary petrology, which implies that the values of alr-transformed point-count results are expected to fall within the interval $[-8; +8]$. In addition, many compositions are likely to plot along the main diagonal of the log-ratio diagram in view of the intrinsic correlation between the two log-ratios, which share one component (in this case Q has been used as a common numerator).

The left-hand side of Figure 3 shows the ternary diagrams of the DM with the first-order provenance fields proposed by Dickinson *et al.* (1983). The QFL diagram (Fig. 3a) contains three fields in which sands of continental block (A), magmatic arc (B) and recycled orogen (C) provenance are expected to plot. The QmFLt diagram (Fig. 3c) contains an additional field reserved for sands of mixed provenance (M). The diagrams on the right-hand side (Figs 3b, d) show the same fields in log-ratio space. Dickinson (1985) referred to these fields as 'provisional' and 'nominal' and stated they correspond to 'actual reported distributions of mean detrital modes'. The criteria used to establish these provenance fields were not explicitly stated and it is not clear why a mixed-provenance field is only present in the QmFLt diagram. The main purpose of this study is to define an optimal subdivision of compositional space according to statistical criteria, based on the same data that were used to establish the DM, and compare the resulting provenance fields to those proposed by Dickinson *et al.* (1983). In addition, the inferential success ratio of the optimized DM will be determined.

Material

Acquisition

The database used in this study was assembled from three datasets compiled by Dickinson and coworkers (Dickinson & Suczek 1979; Dickinson 1982; Dickinson *et al.* 1983) that represent the foundation of the DM (Dickinson 1985, 1988). The paper copies were scanned and digitized by means of OCR software and carefully checked for digitization errors. This resulted in an initial (raw) database of 385 records, each of which comprised the following fields: (a) the sample code; (b) up to four mean compositions of ternary subsets of compositional variables (see Table 1); (c) the sample size *n*, i.e. the number of observations used to calculate the means (samples are often referred to as 'suites' by petrographers); (d) the inferred plate-tectonic setting of the sedimentary basin (see Table 1); (e) a short description of the lithostratigraphic unit, location and/or age of the deposit; (f) the data source (author, year of publication).

A few typographic errors were detected in the ternary compositions, which were corrected by checking their internal consistency (using the interdependence of some ternary compositions, see Table 1) and by comparing the tabulated compositions with the corresponding ternary graphs.

Inspection of the raw database showed that not all of the records were unique. Several records appeared in more than one study, either as exact replicates or with modifications to calculated ternary compositions or inferred plate-tectonic setting. Such replicates were identified by comparing the data sources and the descriptions of the records from each of the three datasets. They were treated in various ways, depending on the nature of the redundancy between records.

- If records were identical, the oldest was retained.
- The most complete version of two fully overlapping records was retained.
- The latest version of a record was retained if corresponding ternary compositions appeared to

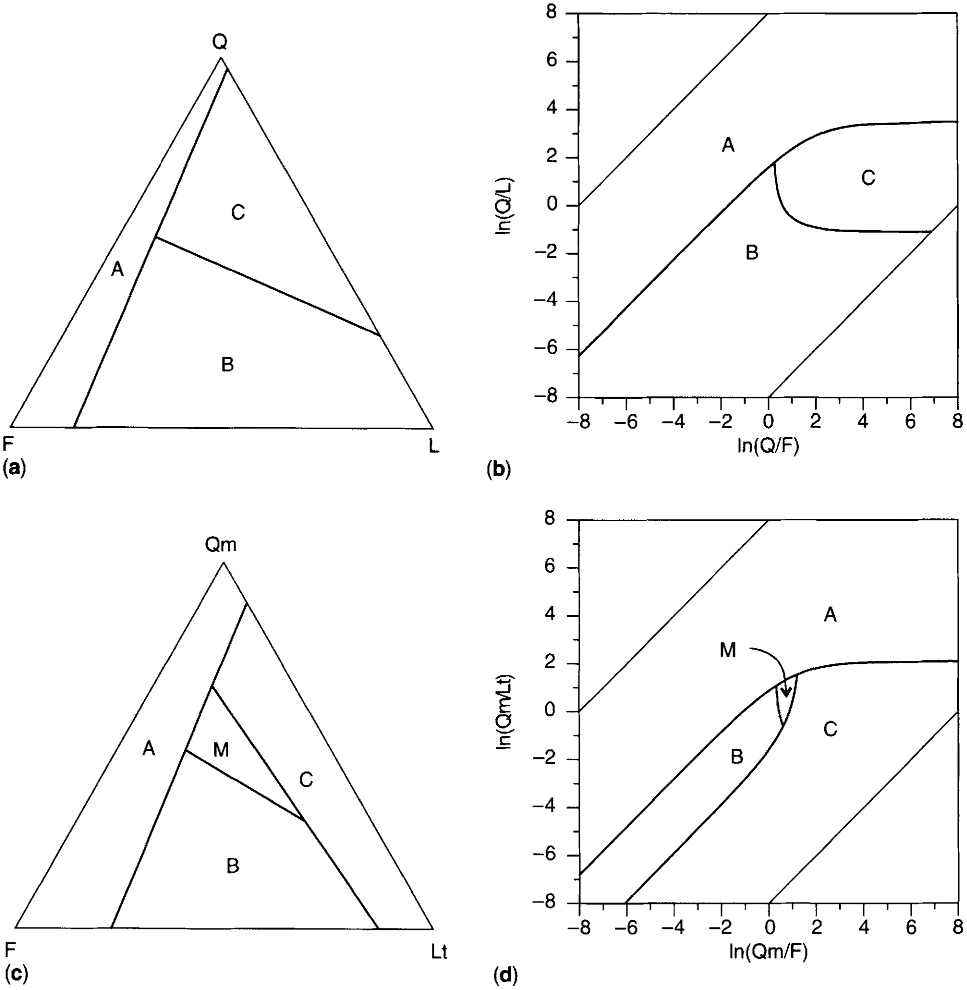

Fig. 3. Subdivision of ternary spaces into provenance fields according to Dickinson *et al.* (1983). For legend to provenance associations and ternary systems see Table 1. (**a**, **b**) QFL system; (**c**, **d**) QmFLt system.

have been recalculated or if the inferred provenance type had changed.

- The latest version of a set of records was retained in the case where different records referred to a common data source (for instance if a sample as well as its subsets were reported in different studies).
- The redundancy in records with a common data source, but containing compositions of different ternary subsets of variables, was removed by deleting overlapping subsets.

This operation reduced the total number of records by 54, leaving a database of 331 records in which the four ternary subcompositions are represented by 309 (QFL), 267 (QmFLt), 101 (QpLvLs) and 100 (QmPK) records, respectively. Together these records represent 11 000 thin sections point-counted by hundreds of operators over a period of three decades. Because very few records are complete, statistical analyses are limited to studies of the variation within each of the four ternary systems separately.

Pre-processing

Statistical analysis of the DM database requires some pre-processing to replace missing or truncated data by appropriate values. Where sample size n was missing from the records, its most likely value was estimated from the overall distribution of sample size (Fig. 4), which is approximately log-normal. The median value of sample size ($n_{50} = 18$) was substituted in 14 records.

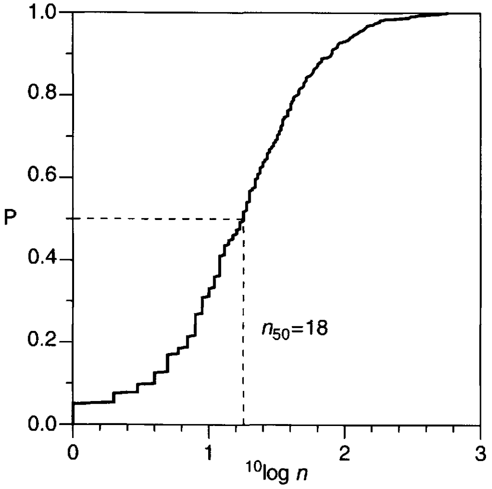

Fig. 4. Distribution of sample size in the database is effectively log-normal and spans almost three orders of magnitude. Median sample size equals 18.

The problem of dealing with zeros in log-ratio analysis, discussed above, also applies to the DM database. However, since each record in the DM database is an average of a series of point counts of unknown length that has been rounded to the nearest integer, it is not immediately clear which zero-replacement value to choose. An upper limit on the replacement value may be derived from the notion that it should not exceed the smallest positive number actually recorded (the 'detection limit'). This upper limit, which was set at 0.9%, represents the replacement value that transforms the ternary compositions (99, 1, 0) into the composition (98.11, 0.99, 0.90). It seemed appropriate that a zero in an average composition calculated from a large sample should be replaced by a different value than a zero that occurs in a single point count, which prompted the introduction of a weighting scheme for zero replacement based on n. In view of the several-orders-of-magnitude range of n (Fig. 4) the following definition of the replacement value δ (in %) was adopted:

$$\delta = \frac{9.9}{10 + \sqrt{n}}. \tag{4}$$

Equation (4) provides the desired upper limit of $\delta = 0.9\%$ for the case $n = 1$ and smaller values for larger samples. Compositions with one or two zeros were recalculated to 100% by replacement of zero component(s) by δ and multiplication of non-zero component(s) by a factor $(100 - N\delta)/100$, where N equals the number of zeros in the original composition. This zero-replacement strategy is consistent with the multiplicative method advocated by Martín-Fernández *et al.* (2003).

Analysis of the DM database

Sources of uncertainty

Statistical analyses should take into account the level of noise in the data as well as possible effects of systematic deviations from 'true' values. Many sources of error can be distinguished in the multistage data-acquisition procedure involved in the construction of the DM database. Each record in the DM database is made up of one or more ternary subcompositions of sandstone calculated by arithmetically averaging of a series of n specimens collected by a single analyst. No information is provided on the spread of values about the means or their covariance, or on parameters of the data-acquisition scheme such as the spatial extent of the sampling programme, the volumes of the samples, pre-measurement laboratory treatments of samples, the point-counting conventions, the number of grains counted and the (spread in) grain size of the sands analysed. Although all of these factors influence the degree to which the mean composition may be considered representative of the lithosome studied (Weltje 2002, 2004), they cannot be taken into account without going back to the original data sources. Other sources of uncertainty play a role when it comes to assessing the integrity of the database as a whole. Potential sources of bias are the uneven spread of data in a geographical and/or stratigraphical sense and errors in assignment of inferred plate-tectonic setting. The lack of standardized data-acquisition methodology could introduce all sorts of bias into the results of statistical analyses of such heterogenous data, but the net result of all these systematic deviations from an unknown 'truth' could equally well be indistinguishable from random error. The following assumptions appear to be reasonable in the absence of any other information:

- geographical and stratigraphical coverage of the DM database are sufficiently representative to allow inferences about sand(stone) composition in relation to global tectonics;
- no significant bias is introduced by possible errors in assignment of plate-tectonic settings;
- no significant bias is introduced by failures to recognize post-depositional (diagenetic) modifications to detrital framework grains;
- possible systematic errors do not invalidate the results of the statistical analysis, because they are indistinguishable from random error;

- the uncertainty of all data-acquisition parameters being equal, the magnitude of random errors in composition estimates is proportional inversely to the square root of sample size *n*.

Methods

As discussed above, very few records are complete, which implies that the four ternary subcompositions had to be analysed separately because the full six-part compositions (Qm, Qp, P, K, Lv, Ls) could not be reconstructed. The following analysis was performed for each of the four alr-transformed ternary systems (each step will be discussed in more detail below).

1. Predictive regions of the population were constructed for each provenance association by a weighted version of the method outlined in Weltje (2002).
2. The compositional space was partitioned by constructing iso-density lines for each pair of predictive distributions in log-ratio space.
3. The grand mean of each provenance association and its 99% confidence region was estimated to provide reference compositions of sands with A, B and C provenance, as well as sand of mixed provenance (corresponding to the iso-density point of the three predictive distributions).
4. Stochastic simulation of compositions from each of the three predictive distributions was carried out to assess the efficiency of the iso-density partitioning.

The first part of the analysis closely follows the method outlined by Weltje (2002) for the construction of predictive regions based on the assumption of additive logistic normality. The main difference between the standard case of constructing a predictive distribution from a set of data points and the present application is that each ternary composition is itself a sample (average) of *n* observations. The mean vector and sample covariance matrix of each alr-transformed set belonging to a provenance association must therefore be calculated by a weighted method. If the original data had been available instead of a series of averages, each observation would have had equal weight (assuming that other data-acquisition parameters do not differ much between observations), indicating that the averaging effect should be modelled by assigning a weight of *n* to each sample. However, indiscriminate use of this linear weighting scheme may cause problems since values of *n* vary by more than two orders of magnitude, so that the estimated parameters of predictive distributions would be heavily influenced by the compositions of a few large samples, which is not desirable, given the possibility of systematic errors in the data. In addition, the smallest samples ($n = 1$) are all from the river mouths of major river systems whose sands have been thoroughly mixed in large drainage basins, indicating that their influence on the parameters of the predictive distribution should be larger than sample size suggests (cf. Ingersoll 1990). In view of these considerations, it was decided to employ a scheme in which the weights assigned to each record equal $\sqrt{n}$. The sum of the weights within each provenance group was used as an estimate of the number of degrees of freedom used in the calculation of the parameters of the predictive distribution. One can think of these degrees of freedom as an effective sample size that encompasses all the sources of uncertainty listed above. The predictive distributions constructed in this way are considered faithful representations of the heterogeneous dataset.

In the second stage of the analysis, the three predictive distributions (of the A, B and C associations) were plotted together and iso-density lines were constructed for each pair of distributions to provide an optimal partitioning of log-ratio space into provenance fields. The rationale behind this partitioning method is the notion that probability densities relative to each of the provenance associations A, B and C vary continuously in compositional space. In each of the three fields that correspond to provenance association A, B or C, the probability density relative to the parent distribution should always exceed the probability densities relative to the other two distributions. The boundary between two provenance fields is thus an iso-density line, i.e. a set of compositions at which the probability densities relative to both distributions are identical (but not constant). The three iso-density lines coincide at the point in compositional space where the probability densities relative to each of the three parent distributions are identical: an iso-density point. This partitioning maximizes the probability that a sample mean of a series of sandstones with unknown provenance is classified correctly.

The results of the analysis were summarized in terms of the vector means of the provenance associations and their associated 99% confidence regions in each of the ternary systems. The composition corresponding to the iso-density point relative to the A, B and C associations was also calculated and presented as a typical sand of mixed provenance.

The final step in the analysis was a stochastic simulation exercise designed to quantify the overall probabilities associated with the empirical classification. A series of 10 000 pseudorandom numbers was generated from each predictive distribution with the Box–Muller algorithm (Press *et al.* 1994) and the probability densities of these data points with respect to each of the A, B and C distributions were calculated. Each data point was then classified in terms of its actual parent distribution and the distribution for which it has the highest probability density. Probability densities were calculated by a method derived from Barceló *et al.* (1996). The result of this stochastic simulation was a 3 × 3 frequency table for every ternary system, based on 30 000 synthetic data points, from which the desired probabilities of (mis)classification were calculated.

Results

Figures 5–8 show the distributions of provenance associations A, B and C in the four ternary systems in the form of predictive regions of 50%, 90% and 99% content plotted together with the records of the DM database. The log-ratio diagrams are shown on the left-hand side of these figures, the corresponding ternary diagrams on the right-hand side. The log-ratio diagrams suggest that the additive logistic normal distribution fits most datasets reasonably well, especially the QFL and QmFLt compositions of provenance associations A and C (Figs 5a, e, 6a, e). The QmPK and QpLvLs compositions of these associations (Figs 7a, e, 8a, e) are not well constrained, owing to the limited number of data points. However, the additive logistic normal distribution shows a lack of fit to the data points of provenance association B (Figs 5c, 6c, 7c, 8c), which suggests the presence of multiple outliers and/or several distinct sub-populations within this association. Formal evaluation of the goodness-of-fit of the predictive distributions by means of appropriate normality tests (Aitchison 1986; Pawlowsky-Glahn & Buccianti 2002) was not attempted.

The moderate lack of fit displayed by the data of provenance association B merits further investigation, but is not considered a matter of great concern in the present study, which is mainly devoted to the construction of provenance fields according to a reproducible method. Geologically sound explanations for the apparent deviation of association B from the simple additive logistic normal model require detailed examination of the original data, which is beyond the scope of this study. A purely statistical approach to the lack-of-fit problem involving operations such as outlier detection and removal, alternative data transformations (cf. Barceló *et al.* 1996) and/or the invocation of other classes of distributions would not improve the geological viability of the DM. Furthermore, the choice of an alternative model for the distribution of data points belonging to provenance association B will affect the partitioning of compositional space into provenance fields, but such shifts in the location of provenance fields are likely to be quite small.

Figures 9–12 illustrate the construction of the three provenance fields in each of the four ternary systems. The upper rows of Figures 9–12 show the predictive regions of 50% and 90% content for each provenance association. The iso-density boundaries between each pair of partly overlapping distributions are displayed in the bottom rows. Triple junctions of iso-density boundaries correspond to compositions with equal probabilities of belonging to one of the provenance associations. Such compositions are regarded as typical examples of mixed provenance. Also plotted in the bottom rows of Figures 9–12 are the 99% confidence regions of the grand means of each of the provenance associations. The small size of these confidence regions indicates that the grand means are well constrained by the large amount of data. The fact that they do not overlap implies that compositional differences between grand means are highly significant. Table 2 summarizes the four characteristic compositions in each of the four ternary systems studied.

The results of the stochastic simulation (Table 3) are a set of probabilities associated with the inference of provenance from a ternary (sub)composition of sand(stone). For instance, if a sample mean plots in field A of the QFL diagram, the probability that its actual provenance is A is 79%. The probability that its actual provenance is C is 20%, and the probability that it is B is only 1%. The overall probability of correct inference in a given ternary system (its success ratio) may be calculated as the average of the probabilities of correct identification of each of the three provenance associations. These numbers equal 76% for QFL, 74% for QmFLt, 64% for QmPK and 78% for QpLvLs. They apply to the iso-density classification presented above; the original subdivision of ternary space in the DM as presented by Dickinson *et al.* (1983) would have given less favourable results.

Discussion and conclusions

Most of the conclusions and points of discussion that emerge from this study are methodological as well as geological. However, one aspect of the compositional data analysis presented in this study is of purely methodological interest. The iso-density partitioning is based on the notion that each point in compositional space may be associated with a vector of relative probability densities, which can itself be regarded as a composition. In the examples presented, both compositional spaces are of the same dimensionality, but this need not be the case. An efficient method to capture the relation

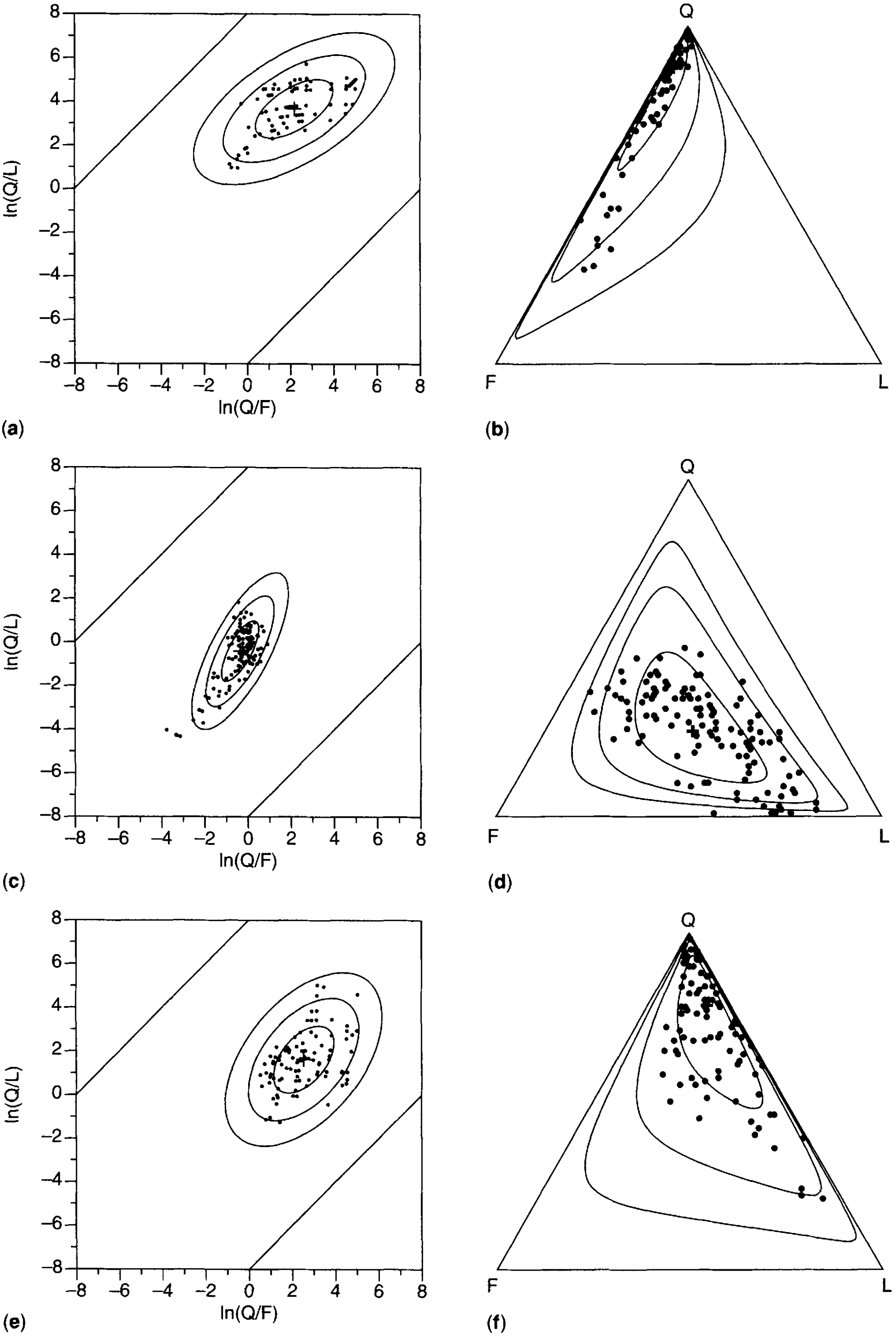

Fig. 5. Predictive distributions for each provenance association in QFL space, represented by regions of 50%, 90% and 99% content. (**a**, **b**) Continental block provenance; (**c**, **d**) magmatic arc provenance; (**e**, **f**) recycled orogen provenance.

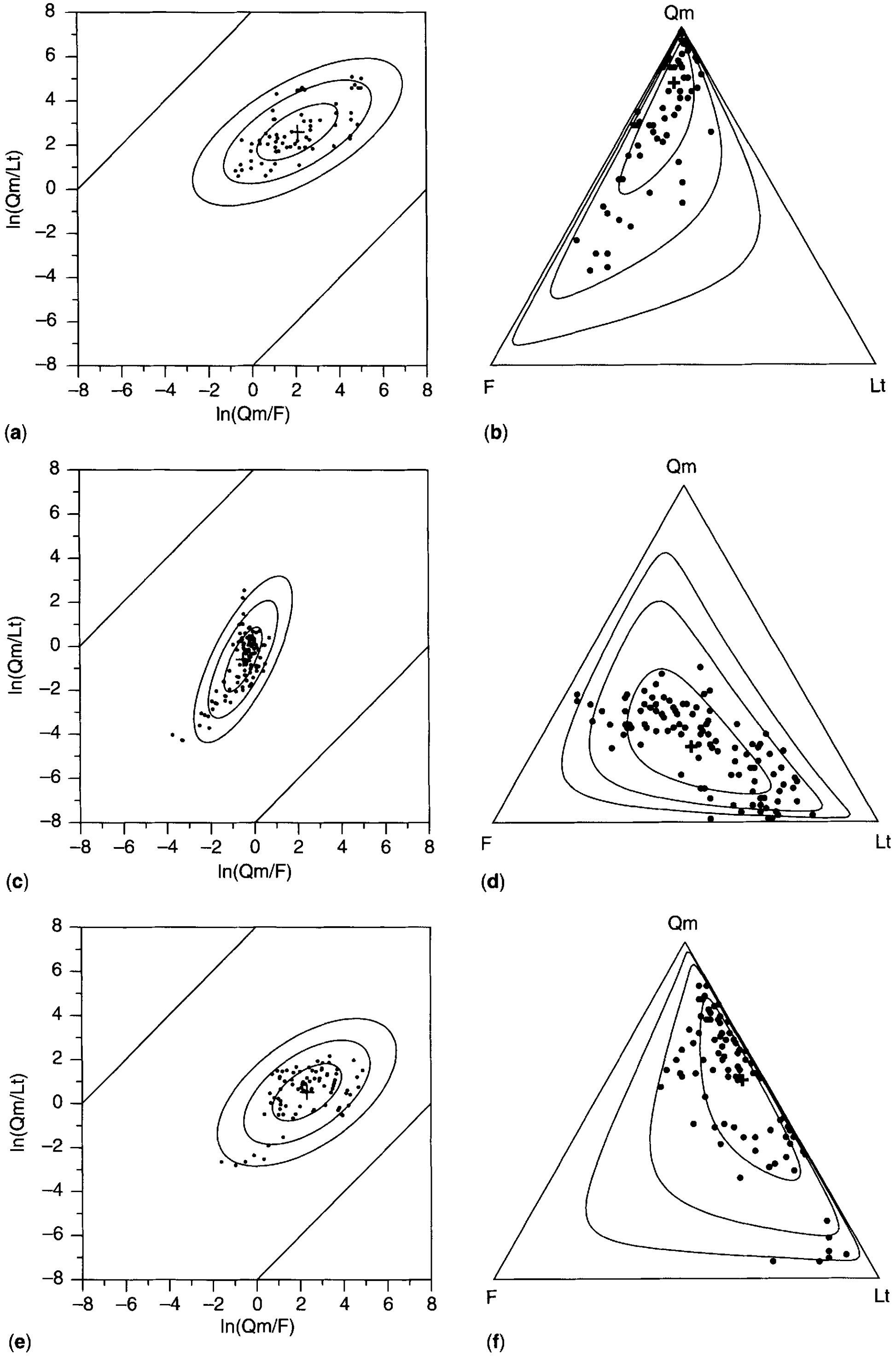

Fig. 6. Predictive distributions for each provenance association in QmFLt space, represented by regions of 50%, 90% and 99% content. (**a**, **b**) Continental block provenance; (**c**, **d**) magmatic arc provenance; (**e**, **f**) recycled orogen provenance.

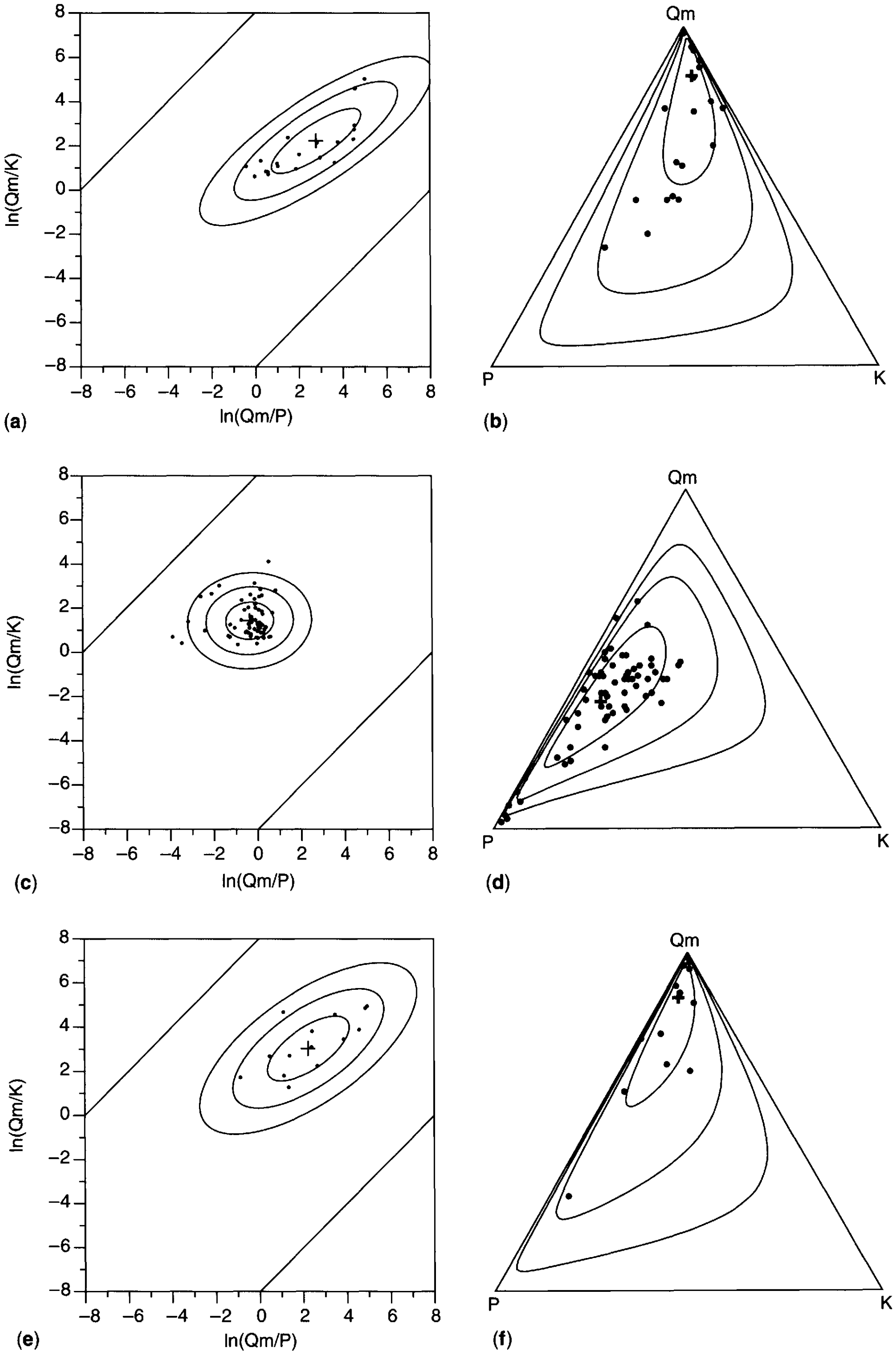

Fig. 7. Predictive distributions for each provenance association in QmPK space, represented by regions of 50%, 90% and 99% content. (**a**, **b**) Continental block provenance; (**c**, **d**) magmatic arc provenance; (**e**, **f**) recycled orogen provenance.

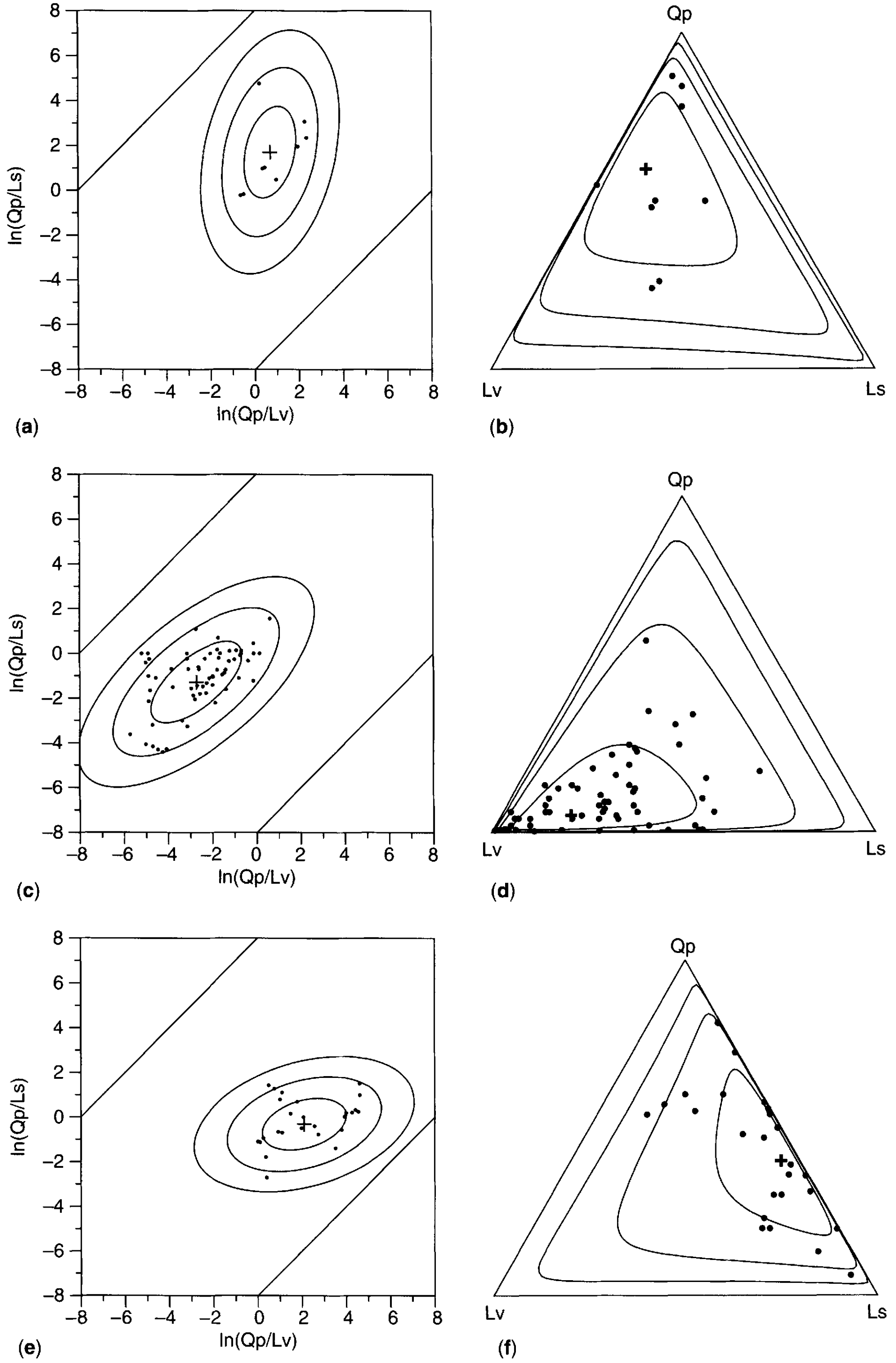

Fig. 8. Predictive distributions for each provenance association in QpLvLs space, represented by regions of 50%, 90% and 99% content. (**a**, **b**) Continental block provenance; (**c**, **d**) magmatic arc provenance; (**e**, **f**) recycled orogen provenance.

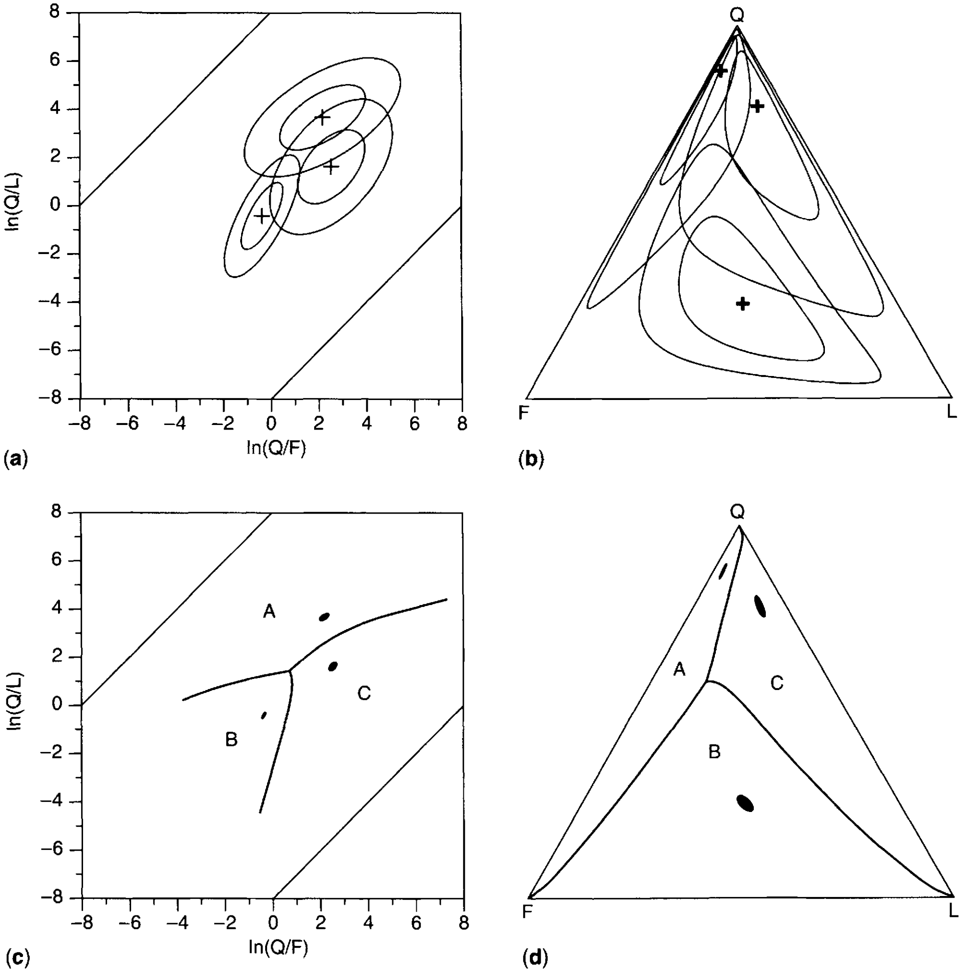

Fig. 9. Construction of provenance fields in QFL space. (**a**, **b**) Predictive distributions of the three provenance associations (regions of 50% and 90% content) plotted together; (**c**, **d**) iso-density partitioning of compositional space into provenance fields. Also shown are 99% confidence regions of population means of the three associations. See Table 1 for legend to provenance associations A, B and C, and Table 2 for ternary population means.

between these compositional spaces would be extremely useful in further iso-density partitioning experiments.

Overview of the optimized DM

Statistical analysis of the DM database has permitted an evaluation of the strengths and weaknesses of this popular plate-tectonic provenance model. The three fundamental provenance associations (continental block, magmatic arc and recycled orogen) are significantly different, as demonstrated by the 99% confidence regions of their grand means in each of the four ternary spaces studied. However, the predictive distributions of the populations display considerable overlap, which indicates that inference of the correct provenance from composition alone is not straightforward. The iso-density partitioning resulted in provenance fields that differ considerably from those proposed by Dickinson *et al.* (1983). The inferential success ratio associated with the optimized subdivision of compositional space into provenance fields is around 75%. The QpLvLs subcomposition has the highest overall success ratio (78%), followed closely by the QFL and QmFLt compositions, which are essentially equally powerful provenance tools with a success ratio of around 75%. The QmPK subcomposition, with its low overall success ratio of 64%, does not

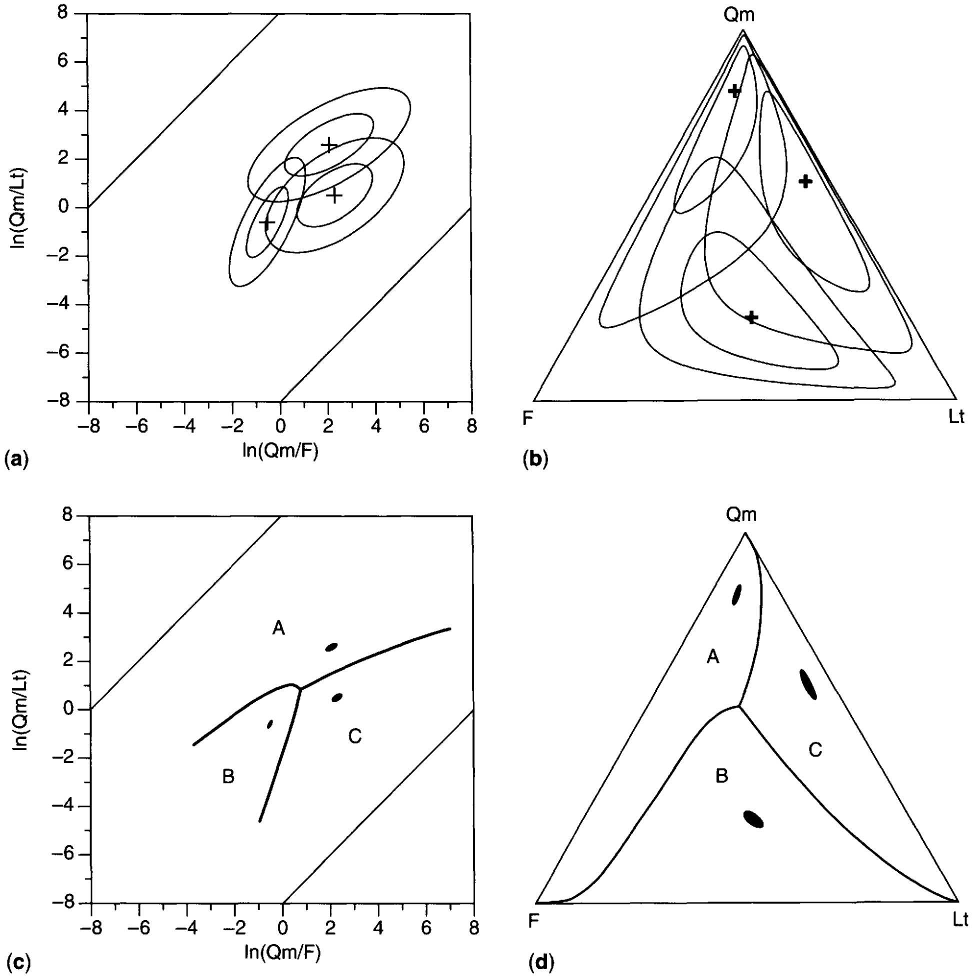

Fig. 10. Construction of provenance fields in QmFLt space. (**a**, **b**) Predictive distributions of the three provenance associations (regions of 50% and 90% content) plotted together; (**c**, **d**) iso-density partitioning of compositional space into provenance fields. Also shown are 99% confidence regions of population means of the three associations. See Table 1 for legend to provenance associations A, B and C, and Table 2 for ternary population means.

appear to hold much promise for provenance discrimination. The success ratios of the optimized DM suggest that Molinaroli *et al.* (1991) overestimated its discriminatory power by testing it with the same data used to establish their set of discriminant functions. Strategies to extend and further improve the DM are discussed below.

Increasing the dimensionality

One of the major shortcomings of the database underlying the DM is that very few records are complete, which limits statistical analysis to a separate description of compositional variation within the four ternary systems. The partial view on compositional variability obtained by analysing these amalgamations and subcompositions of the full six-part composition (Qm, Qp, P, K, Lv, Ls) may be insufficient to address the relevant geological problems at hand, as illustrated by the following example. In an early attempt to apply the log-ratio transformation to the DM database, Butler & Woronow (1986) analysed the dataset of Dickinson & Suczek (1979) for the presence of spurious correlations induced by the constant-sum constraint. Their results suggested that the compositional trend of decreasing L relative to Q + F within the magmatic-arc provenance field of Dickinson

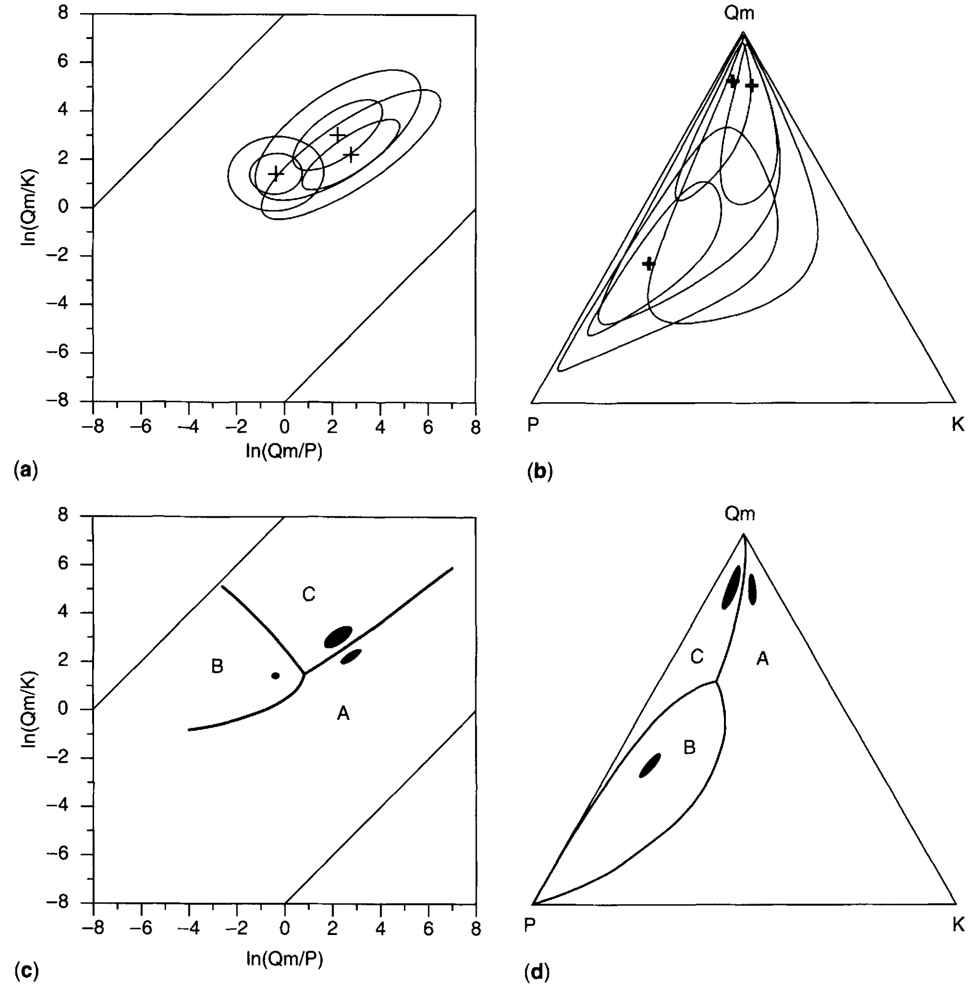

Fig. 11. Construction of provenance fields in QmPK space. (**a**, **b**) Predictive distributions of the three provenance associations (regions of 50% and 90% content) plotted together; (**c**, **d**) iso-density partitioning of compositional space into provenance fields. Also shown are 99% confidence regions of population means of the three associations. See Table 1 for legend to provenance associations A, B and C, and Table 2 for ternary population means.

& Suczek (1979) can be produced by imposing the constant-sum constraint on a set of independent variables, which may indicate that it is the sole result of percentage formation ('closure') and has no geological meaning. This purely statistical interpretation is not very likely, however, because the decease of L relative to Q + F, which defines the main trend of arc dissection by erosion, is usually accompanied by a decrease of P relative to K (W. R. Dickinson, pers. comm. 1994).

Answers to question of this kind require analyses of the relationships between log(P/K) on the one hand, and log(Q/L), log(F/L), or log{(Q + F)/L} on the other hand. This example indicates that an empirical provenance model built on a database of six-part compositions (Qm, Qp, P, K, Lv, Ls) is much more powerful than a series of models based on ternary (sub)compositions only. Full six-part compositions of individual specimens should be reported in future studies intended to contribute to a new database for a second-generation provenance model (and not only their mean six-part composition).

Sands of mixed provenance

The DM was designed for classification of means of sandstone suites only. Erroneous interpretations

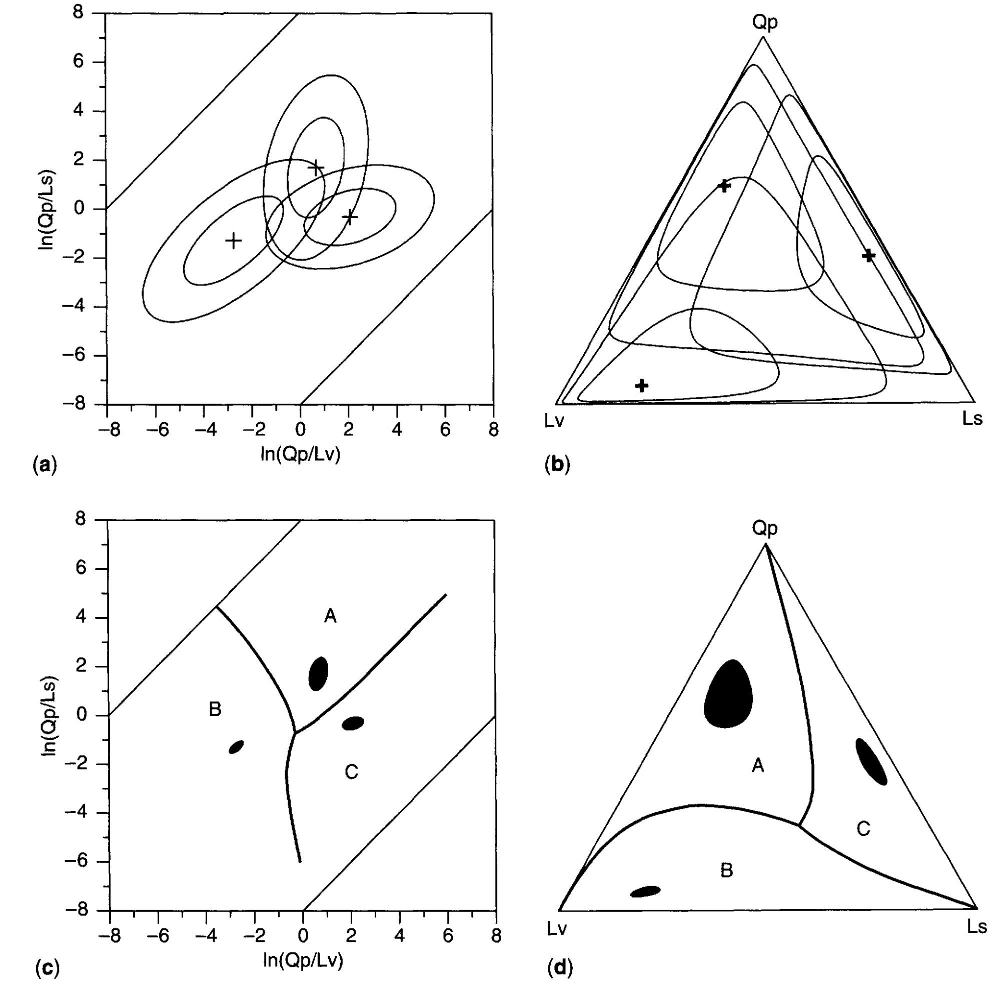

Fig. 12. Construction of provenance fields in QpLvLs space. (**a**, **b**) Predictive distributions of the three provenance associations (regions of 50% and 90% content) plotted together; (**c**, **d**) iso-density partitioning of compositional space into provenance fields. Also shown are 99% confidence regions of population means of the three associations. See Table 1 for legend to provenance associations A, B and C, and Table 2 for ternary population means.

may result if local provenance signals in the data have not been suppressed by spatial averaging of sandstone compositions (Ingersoll 1990; Ingersoll *et al.* 1993; Critelli *et al.* 1997). This averaging approach, which may be viewed as a way of artificially mixing local provenance signals, has the distinct advantage of robustness but implies a limited spatial and temporal resolution of the DM. It seems fair to state that many of the records in the DM database are not 'pure' sands of a single provenance, but contain varying admixtures of sands of different provenances. Analyses of modern deep-sea sands (Valloni 1985) and reviews of global dispersal systems (Dickinson 1988) indicate that sands of mixed provenance are extremely common. It is therefore not surprising that many sand suites plot in the mixed provenance field of the original QmFLt diagram (Dickinson 1985, 1988). Given these restrictions, success ratios of empirical provenance models based on averaging of ternary compositions are not likely to exceed those of the optimized DM presented in this study.

Differences between the original and the revised DM are not limited to the locations of the boundaries between provenance fields. In the present analysis, each of the ternary systems was treated in the same way, in contrast with the method of Dickinson *et al.* (1983), who introduced a separate

Table 2. *Grand means of provenance associations and typical compositions of mixed provenance (legend in Table 1)*

	Q (%)	F (%)	L (%)
A	88	10	2
B	25	37	38
C	78	6	16
M	57	29	14
	Qm (%)	F (%)	Lt (%)
A	84	10	6
B	22	37	41
C	59	6	35
M	52	25	23
	Qm (%)	P (%)	K (%)
A	86	5	9
B	38	53	9
C	87	9	4
M	60	27	13
	Qp (%)	Lv (%)	Ls (%)
A	59	30	11
B	5	77	18
C	40	5	55
M	22	28	50

field for sands of mixed provenance in the QmFLt system only (Fig. 3c). The provisional solution adopted in the present study is to regard the compositions at the triple junction of the three iso-density lines in each of the ternary systems as typical examples of mixed provenance.

It should be noted that any observation located in an area of compositional space where two or more distributions overlap one another is difficult to interpret without additional information. On the one hand, such sand could have been derived from one of these distributions exclusively (the point of view adopted in this study), but on the other hand, it could represent a mixture of sands from two or more of these distributions. The overlap between distributions causes the range of potential scenarios to be infinite and impossible to constrain without taking into account additional information about the palaeogeography of the area from which the sands were derived. Additional complications may arise from variability of sediment composition due to past climate change or the presence of diagenetic gradients across basins, both of which are essentially unrelated to provenance *sensu stricto*, i.e. the composition and texture of parent rocks (Johnsson 1993; Weltje & Von Eynatten 2004). The very fact that such information appears to be required contradicts the basic premise of the DM, i.e. provenance of sands may be inferred from composition alone.

Table 3. *Probabilites of inferring provenance from ternary composition (legend in Table 1)*

	Actual provenance		
QFL	A (%)	B (%)	C (%)
inferred A	79	1	20
inferred B	1	87	12
Inferred C	16	21	63
QmFLt	A (%)	B (%)	C (%)
inferred A	79	4	17
inferred B	2	81	17
Inferred C	13	24	63
QmPK	A (%)	B (%)	C (%)
inferred A	59	12	29
inferred B	16	72	12
Inferred C	34	6	60
QpLvLs	A (%)	B (%)	C (%)
inferred A	75	13	12
inferred B	7	81	12
Inferred C	17	4	79

The above considerations imply that sands of mixed provenance cannot be interpreted by 'averaging out' all variability. On the contrary, attention must be devoted to the development of methods to exploit the information contained within the covariance structure of compositional data. Ingersoll (1990), Ingersoll *et al.* (1993) and Critelli *et al.* (1997) noted a systematic decrease in the variance of mean sand composition with increasing spatial scales of dispersal systems, a phenomenon further explored by Weltje (2004). The covariance structure of compositional data is an essential tool of quantitative provenance analysis, which permits sands of mixed provenance to be statistically 'unmixed' into end-member provenance associations (Weltje 1997). The end-member mixing model allows one to address the issue of mixed provenance in a systematic and quantitative way – thereby providing insights that could never have been obtained by plotting arithmetic means in ternary diagrams. This approach also requires full six-part compositions of individual specimens rather than sets of disjointed three-part means. Weltje (1995) provides an example of end-member modelling of a suite of mixed-provenance sands from the Italian Alps and Apennines.

Conclusions

The DM in its present form – as a series of four separate ternary diagrams – should be recognized for what it is: an exploration tool designed to infer the large-scale tectonic setting of sediment-dispersal systems in the distant past and/or any remaining frontier areas of our planet. The DM deliberately bypasses all the details of the sediment-forming processes. This approach guarantees robustness but does not permit a meaningful analysis of mixing, identified as a factor of overriding importance in sediment generation. The DM is a successful exploration tool, but it does not lend itself easily to other applications, such as regional studies of multi-sourced basin fills.

Traditional provenance models which are aimed at inferring source-area characteristics from sediment properties could be improved greatly by incorporation of quantified knowledge about the processes that govern sediment generation. Prediction of sediment composition from properties of drainage basins (cf. Ibbeken & Schleyer 1991) through development of modelling tools to address sediment generation is an area of active research (Basu 2003; Weltje & Von Eynatten 2004). The capability to integrate modelling efforts, measurements of process rates under laboratory and field conditions, and analyses of comprehensive and well-documented compositional datasets will ultimately determine the rate of progress in provenance analysis.

The author thanks J. J. Egozcue for his perceptive review of the manuscript.

References

AITCHISON, J. 1982. The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society, Series B*, **44**, 139–177.

AITCHISON, J. 1986. *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.

AITCHISON, J. & EGOZCUE, J. J. 2005. Compositional data analysis: where are we and where should we be heading? *Mathematical Geology*, **37**, 829–850.

ARMSTRONG-ALTRIN, J. S. & VERMA, S. P. 2005. Critical evaluation of six tectonic setting discrimination diagrams using geochemical data of Neogene sediments from known tectonic settings. *Sedimentary Geology*, **177**, 115–129.

BARCELÓ, C., PAWLOWSKY, V. & GRUNSKY, E. 1996. Some aspects of transformations of compositional data and the identification of outliers. *Mathematical Geology*, **28**, 501–518.

BASU, A. 2003. A perspective on quantitative provenance analysis. *In*: VALLONI, R. & BASU, A. (eds) *Quantitative Provenance Studies in Italy*. Memorie Descrittive della Carta Geologica dell' Italia, **61**, 11–22.

BECKE, F. 1897. Gesteine des Columbretes. *Tschermak's Mineralogische und Petrographische Mitteilungen*, **16**, 308–336.

BUTLER, J. C. 1979. Trends in ternary petrologic variation diagrams – fact or fantasy? *American Mineralogist*, **64**, 1115–1121.

BUTLER, J. C. 1982. The closure problem as reflected in discriminant function analysis. *Chemical Geology*, **37**, 367–375.

BUTLER, J. C. & WORONOW, A. 1986. Extracting genetic information from coarse clastic modes. *Computers & Geosciences*, **12**, 643–652.

CHAYES, F. 1960. On correlation between variables of constant sum. *Journal of Geophysical Research*, **65**, 4185–4193.

CRITELLI, S., LE PERA, E. & INGERSOLL, R. V. 1997. The effects of source lithology, transport, deposition and sampling scale on the composition of southern California sand. *Sedimentology*, **44**, 653–671.

CROOK, K. A. W. 1974. Lithogenesis and geotectonics: the significance of compositional variation in flysch arenites (graywackes). *In*: DOTT, R. H. JR. & SHAVER, R. H. (eds) *Modern and Ancient Geosynclinal Sedimentation*. Society of Economic Paleontologists and Mineralogists, Special Publications, **19**, 304–310.

DAVIS, J. C. 1997. *Statistics and Data Analysis in Geology*, 3rd edn. Wiley & Sons, New York.

DICKINSON, W. R. 1982. Compositions of sandstones in Circum-Pacific subduction complexes and forearc basins. *American Association of Petroleum Geologists Bulletin*, **66**, 121–137.

DICKINSON, W. R. 1985. Interpreting provenance relations from detrital modes of sandstones. *In*: ZUFFA, G. G. (ed.) *Provenance of Arenites*. North Atlantic Treaty Organization – Advanced Study Institutes (NATO-ASI), Series C, **148**, 333–361.

DICKINSON, W. R. 1988. Provenance and sediment dispersal in relation to paleotectonics and paleogeography of sedimentary basins. *In*: KLEINSPEHN, K. L. & PAOLA, C. (eds) *New Perspectives in Basin Analysis*. Springer-Verlag, New York, 3–25.

DICKINSON, W. R. & SUCZEK, C. 1979. Plate tectonics and sandstone compositions. *American Association of Petroleum Geologists Bulletin*, **63**, 2164–2182.

DICKINSON, W. R., BEARD, L. S., BRAKENRIDGE, G. R. ET AL. 1983. Provenance of North American Phanerozoic sandstones in relation to tectonic setting. *Geological Society of America Bulletin*, **94**, 222–235, with Supplement (GSA data repository item # 8302).

EGOZCUE, J. J., PAWLOWSKY-GLAHN, V., MATEU-FIGUERAS, G. & BARCELÓ-VIDAL, C. 2003. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, **35**, 279–300.

HAUGHTON, P. D. W., TODD, S. P. & MORTON, A. C. 1991. Sedimentary provenance studies. *In*: MORTON, A. C., TODD, S. P. & HAUGHTON, P. D. W. (eds) *Developments in Sedimentary Provenance Studies*. Geological Society of London, Special Publications, **57**, 1–11.

IBBEKEN, H. & SCHLEYER, R. 1991. *Source and Sediment: A Case Study of Provenance and Mass Balance at an Active Plate Margin (Calabria, Southern Italy)*. Springer-Verlag, Berlin.

INGERSOLL, R. V. 1990. Actualistic sandstone petrofacies: discriminating modern and ancient source rocks. *Geology*, **18**, 733–736.

INGERSOLL, R. V., KRETCHMER, A. G. & VALLES, P. K. 1993. The effect of sampling scale on actualistic sandstone petrofacies. *Sedimentology*, **40**, 937–953.

Johnsson, M. J. 1993. The system controlling the composition of clastic sediments. *In*: Johnsson, M. J. & Basu, A. (eds) *Processes Controlling the Composition of Clastic Sediments*. Geological Society of America, Special Paper, **284**, 1–19.

Klein, G. D. 1963. Analysis and review of sandstone classifications in the North American geological literature, 1940–1960. *Geological Society of America Bulletin*, **74**, 555–576.

Martín-Fernández, J. A., Barceló-Vidal, C. & Pawlowsky-Glahn, V. 2003. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, **35**, 253–278.

Molinaroli, E., Blom, M. & Basu, A. 1991. Methods of provenance determination tested with discriminant function analysis. *Journal of Sedimentary Petrology*, **61**, 900–908.

Okada, H. 1971. Classification of sandstone: analysis and proposal. *Journal of Geology*, **79**, 509–525.

Pawlowsky-Glahn, V. & Buccianti, A. 2002. Visualization and modeling of sub-populations of compositional data: statistical methods illustrated by means of geochemical data from fumarolic fluids. *International Journal of Earth Sciences (Geologische Rundschau)*, **91**, 357–368.

Pawlowsky-Glahn, V. & Olea, R. A. 2004. *Geostatistical Analysis of Compositional Data*. Oxford University Press, New York.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. 1994. *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, 2nd edn. Cambridge University Press, Cambridge.

Rollinson, H. R. 1993. *Using Geochemical Data: Evaluation, Presentation, Interpretation*. Longman, Harlow.

Schwab, F. L. 1975. Framework mineralogy and chemical composition of continental margin-type sandstones. *Geology*, **3**, 487–490.

Swan, A. R. H. & Sandilands, M. 1993. *Introduction to Geological Data Analysis*. Blackwell Science, Oxford.

Valloni, R. 1985. Reading provenance from modern marine sands. *In*: Zuffa, G. G. (ed.) *Provenance of Arenites*. North Atlantic Treaty Organization – Advanced Study Institutes (NATO-ASI), Series C, **148**, 309–332.

Weltje, G. J. 1995. Unravelling mixed provenance of coastal sands: the Po Delta and adjacent beaches of the northern Adriatic Sea as a test case. *In*: Oti, M. A. & Postma, G. (eds) *Geology of Deltas*. Balkema, Rotterdam, 181–202.

Weltje, G. J. 1997. End-member modeling of compositional data: numerical–statistical algorithms for solving the explicit mixing problem. *Mathematical Geology*, **29**, 503–549.

Weltje, G. J. 2002. Quantitative analysis of detrital modes: statistically rigorous confidence regions in ternary diagrams and their use in sedimentary petrology. *Earth-Science Reviews*, **57**, 211–253.

Weltje, G. J. 2004. A quantitative approach to capturing the compositional variability of modern sands. *Sedimentary Geology*, **171**, 59–77.

Weltje, G. J. & Von Eynatten, H. 2004. Quantitative provenance analysis of sediments: review and outlook. *Sedimentary Geology*, **171**, 1–11.