Geochemistry Geophysics Geosystems G3
AN ELECTRONIC JOURNAL OF THE EARTH SCIENCES
Published by AGU and the Geochemical Society

Article
Volume 7, Number 6
27 June 2006
Q06017, doi:10.1029/2005GC001092
ISSN: 1525-2027

Tectonic discrimination diagrams revisited

Pieter Vermeesch
Institute for Isotope Geology and Mineral Resources, ETH-Zurich, CH-8092 Zurich, Switzerland (pvermees@erdw.ethz.ch)

[1] The decision boundaries of most tectonic discrimination diagrams are drawn by eye. Discriminant analysis is a statistically more rigorous way to determine the tectonic affinity of oceanic basalts based on their bulk-rock chemistry. This method was applied to a database of 756 oceanic basalts of known tectonic affinity (ocean island, mid-ocean ridge, or island arc). For each of these training data, up to 45 major, minor, and trace elements were measured. Discriminant analysis assumes multivariate normality. If the same covariance structure is shared by all the classes (i.e., tectonic affinities), the decision boundaries are linear, hence the term linear discriminant analysis (LDA). In contrast with this, quadratic discriminant analysis (QDA) allows the classes to have different covariance structures. To solve the statistical problems associated with the constant-sum constraint of geochemical data, the training data must be transformed to log-ratio space before performing a discriminant analysis. The results can be mapped back to the compositional data space using the inverse log-ratio transformation. An exhaustive exploration of 14,190 possible ternary discrimination diagrams yields the Ti-Si-Sr system as the best linear discrimination diagram and the Na-Nb-Sr system as the best quadratic discrimination diagram. The best linear and quadratic discrimination diagrams using only immobile elements are Ti-V-Sc and Ti-V-Sm, respectively. As little as 5% of the training data are misclassified by these discrimination diagrams. Testing them on a second database of 182 samples that were not part of the training data yields a more reliable estimate of future performance. Although QDA misclassifies fewer training data than LDA, the opposite is generally true for the test data. Therefore LDA is a cruder but more robust classifier than QDA. Another advantage of LDA is that it provides a powerful way to reduce the dimensionality of the multivariate geochemical data in a similar way to principal component analysis. This procedure yields a small number of "discriminant functions," which are linear combinations of the original variables that maximize the between-class variance relative to the within-class variance.

Components: 9425 words, 46 figures, 7 tables, 1 dataset.

Keywords: basalt; classification; discriminant analysis; discrimination diagrams; statistics.

Index Terms: 1021 Geochemistry: Composition of the oceanic crust; 1065 Geochemistry: Major and trace element geochemistry; 1094 Geochemistry: Instruments and techniques.

Received 28 July 2005; **Revised** 20 November 2005; **Accepted** 9 February 2006; **Published** 27 June 2006.

Vermeesch, P. (2006), Tectonic discrimination diagrams revisited, *Geochem. Geophys. Geosyst.*, *7*, Q06017, doi:10.1029/2005GC001092.

1. Introduction

[2] Recovering the tectonic affinity of ancient ophiolites is a problem of great scientific interest. In addition to field data, basalt geochemistry is another way to address this problem. Tectonic discrimination diagrams have been a popular technique for doing this since the publication of landmark papers by *Pearce and Cann* [1971, 1973]. This paper revisits some of the popular discrimination diagrams that have been in use since then. Nearly all discrimination diagrams that are currently in use were drawn by eye. The present paper revisits these diagrams in a statistically more rigorous way.

[3] First, a short introduction will be given to the discriminant analysis method. The fundamental difference between the reduction in dimensionality achieved by principal components and by linear discriminant analysis will be explained. Then, the consequences of the constant-sum constraint of geochemical data for discriminant analysis will be discussed. In section 4, *Aitchison*'s [1982, 1986] solution to this problem will be briefly explained. Section 5 revisits some of the historically most important and popular discrimination diagrams, based on a new database of oceanic basalts of known tectonic affinity. The effect of data-closure will be taken into account and a statistically rigorous reevaluation of these diagrams will be made in both the linear and the quadratic case.

[4] This paper does not restrict itself to only those geochemical features that have already been used by previous workers. Section 6 gives an exhaustive exploration of all possible bivariate and ternary discrimination diagrams using a set of 45 major, minor, and trace elements. This will result in a list of the 100 best linear and quadratic ternary discriminators, ranked according to their success in classifying the training data. Finally, section 7 tests the most important discrimination diagrams discussed elsewhere in the paper on a second database of oceanic basalts that were not part of the training data. This provides a more objective estimator of misclassification risk on future data than the misclassification rate of the training data. Section 7 also contains a formal comparison of the new decision boundaries with the old ones of *Pearce and Cann* [1973], *Shervais* [1982], *Meschede* [1986], and *Wood* [1980]. It will be shown that the new decision boundaries perform at least as well as the old ones.

2. Discriminant Analysis

[5] Consider a data set of a large number of N-dimensional data X, which belong to one of K classes. For example, X might be a set of geochemical data (e.g., SiO_2, Al_2O_3, etc.) from basaltic rocks of K tectonic affinities (e.g., mid-ocean ridge, ocean island, island arc). We might ask ourselves which of these classes an unknown sample x belongs to. This question is answered by Bayes' Rule: the decision d is the class G ($1 \leq G \leq K$) that has the highest posterior probability given the data x:

$$d = \underset{k=1,\ldots,K}{argmax}\, Pr(G = k|X = x) \tag{1}$$

where *argmax* stands for "argument of the maximum," i.e., when f(k) reaches a maximum when k = d, then $\underset{k=1,\ldots,K}{argmax}$ f(k) = d. This posterior distribution can be calculated according to Bayes' Theorem:

$$Pr(G|X) \propto Pr(X|G)Pr(G) \tag{2}$$

where Pr(X|G) is the probability density of the data in a given class, and Pr(G) the prior probability of the class, which we will consider uniformly distributed (i.e., Pr(G = 1) = Pr(G = 2) = ... = Pr(G = K) = 1/K) in this paper. Therefore plugging equation (2) into equation (1) reduces Bayes' Rule to a comparison of probability density estimates. We now make the simplifying assumption of multivariate normality:

$$Pr(X = x|G = k) = \frac{\exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right)}{(2\pi)^{N/2}\sqrt{|\Sigma_k|}} \tag{3}$$

where μ_k and Σ_k are the mean and covariance of the kth class and $(x - \mu_k)^T$ indicates the transpose of the matrix $(x - \mu_k)$. Using equation (3), and taking logarithms, equation (1) becomes

$$d = \underset{k=1,\ldots,K}{argmax} -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) \tag{4}$$

[6] Equation (4) is the basis for quadratic discriminant analysis (QDA). Usually, μ_k and Σ_k are not known, and must be estimated from the training data. If we make the additional assumption that all

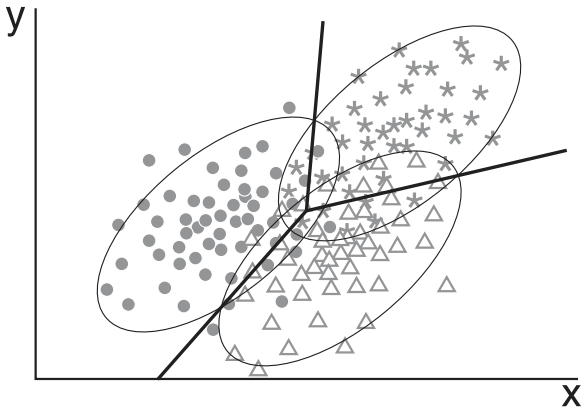

Figure 1. Discriminant analysis of three classes with equal covariance matrices leads to linear discriminant boundaries. The ellipses mark arbitrary (e.g., 95%) confidence levels for the underlying populations.

the classes share the same covariance structure (i.e., $\Sigma_k = \Sigma \; \forall \; k$), then equation (1) simplifies to

$$d = \underset{k=1,\ldots,K}{argmax}\, x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k \qquad (5)$$

[7] This is the basis of linear discriminant analysis (LDA), which has some desirable properties. For example, because equation (5) is linear in x, the decision boundaries between the different classes are straight lines (Figure 1). Furthermore, LDA can lead to a significant reduction in dimensionality, in a similar way to principal component analysis (PCA). PCA finds an orthogonal transformation B (i.e., a rotation) that transforms the centered data (X) to orthogonality, so that the elements of the vector BX are uncorrelated. B can be calculated by an eigenvalue decomposition of the covariance matrix Σ. The eigenvectors are orthogonal linear combinations of the original variables, and the eigenvalues give their variances. The first few principal components generally account for most of the variability of the data, constituting a significant reduction of dimensionality (Figure 2).

[8] Like PCA, LDA also finds linear combinations of the original variables. However, this time, we do not want to maximize the overall variability, but find the orthogonal transformation Z = BX that maximizes the between class variance S_b relative to the within class variance S_w, where S_b is the variance of the class means of Z, and S_w is the pooled variance about the means (Figure 2).

3. The Compositional Data Problem

[9] One of the assumptions of discriminant analysis is that the elements of X are statistically independent from each other, apart from the covariance structure contained in their multivariate normality. However, geochemical data are generally expressed as parts of a whole (percent or ppm) and therefore are not free to vary independently from each other. For example, in a three-component system (A + B + C = 100%), increasing one

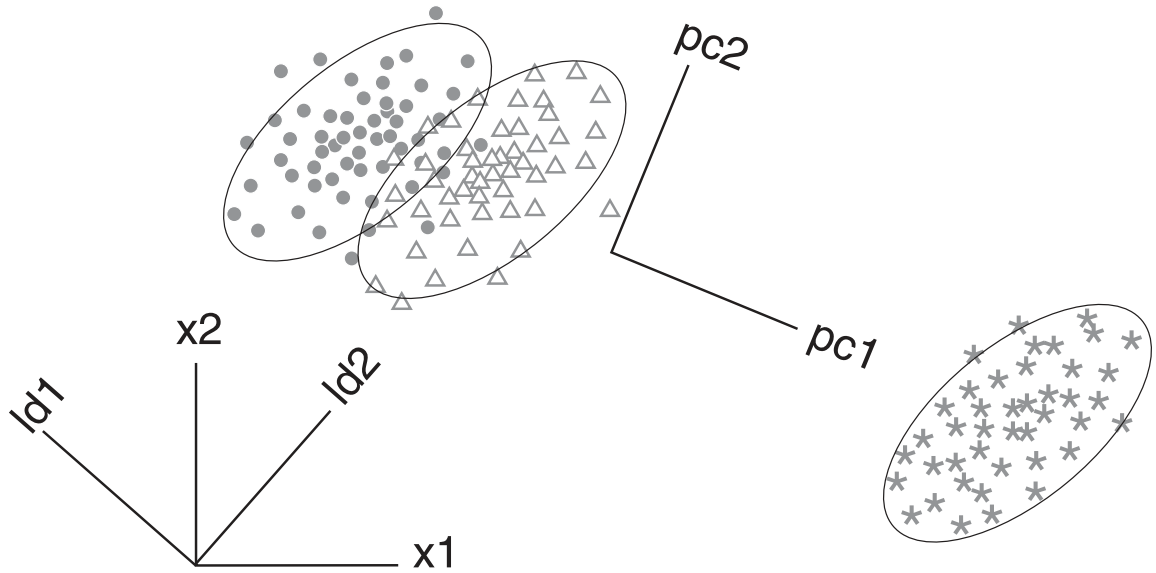

Figure 2. Similarities and differences between linear discriminant and principal component analysis. x1 and x2 are the original variables, pc1 and pc2 are the principal components, and ld1 and ld2 are the linear discriminant functions.

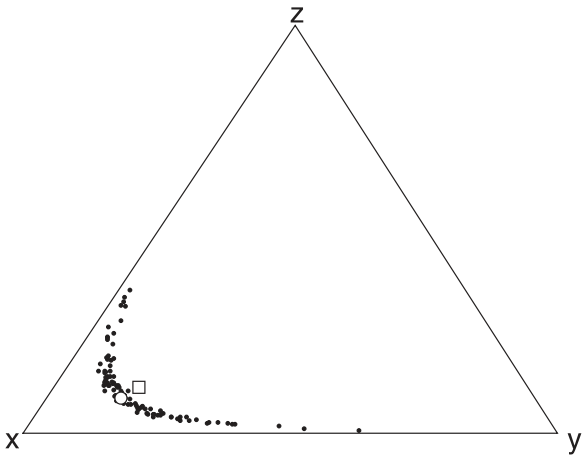

Figure 3. One of the consequences of the constant-sum constraint of compositional data is that the arithmetic mean (marked by the open square) of populations (black dots) has no physical meaning. Instead, the geometric mean should be used (open circle).

component (e.g., A) causes a decrease in the two other components (B and C). The constant-sum constraint has several consequences, besides introducing a negative bias into correlations between components. One of these consequences is that the arithmetic mean of compositional data has no physical meaning (Figure 3). This is very unfortunate because some popular discrimination diagrams [e.g., *Pearce and Cann*, 1973] are based on the arithmetic means of multiple samples, and it is these averages that are published in the literature. Therefore the discriminant analyses discussed in this paper will not be based on these historic data sets, but will use a newly compiled database of individual analyses.

[10] Another statistical issue that deserves to be mentioned is spurious correlation. Bivariate plots of the form X vs. X/Y, X vs. Y/X or X/Z vs. Y/Z can show some degree of correlation, even when X, Y and Z are completely independent from each other (Figure 4). This effect was first discussed more than a century ago by *Pearson* [1897], and was brought to the attention of geologists more than half a century ago by *Chayes* [1949]. Spurious correlation is an effect that should be borne in mind when interpreting discrimination diagrams like the Zr/Y-Ti/Y diagram [*Pearce and Gale*, 1977], the Zr/Y-Zr diagram [*Pearce and Norry*, 1979], or the Ti/Y-Nb/Y and K_2O/Yb-Ta/Yb diagrams [*Pearce*, 1982]. Note that whereas in Figure 4, X, Y and Z are completely independent, this is never the case for compositional data, due to the constant-

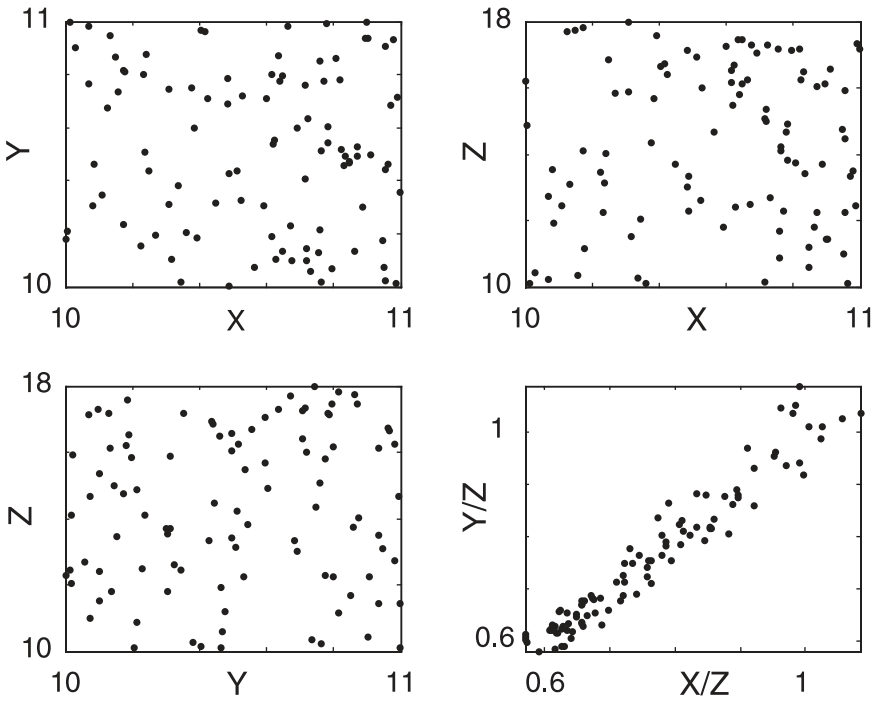

Figure 4. X, Y, and Z are uncorrelated, uniform random numbers. The strong spurious correlation of the ratios Y/Z and X/Z is an artifact of the relatively large variance of Z relative to X, Y, and Z.

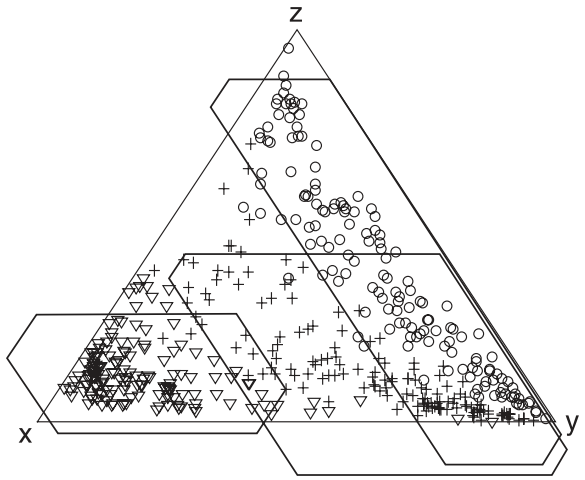

Figure 5. The 95% normal confidence regions [e.g., *Weltje*, 2002] for synthetic trivariate compositional data partly fall outside the ternary diagram, a nonsense result illustrating the dangers of performing "traditional" statistics on the simplex.

sum constraint described before. This only aggravates the problem of spurious correlation.

4. Aitchison's Solution to the Compositional Data Problem

[11] Although *Chayes* [1949, 1960, 1971] made significant contributions to the compositional data problem, the real breakthrough was made by *Aitchison* [1982, 1986]. Aitchison argues that N-variate data constrained to a constant sum form an N − 1 dimensional sample space or simplex. An example of a simplex for N = 3 is the ternary diagram [e.g., *Weltje*, 2002]. The very fact that it is possible to plot ternary data on a two-dimensional sheet of paper tells us that the sample space really has only two, and not three dimensions. The "traditional" statistics of real space ($\mathbb{R}^N$) do no longer work on the simplex (Δ_{N-1}). Figure 5 shows the breakdown of the calculation of $100(1-\alpha)\%$ confidence intervals on Δ_2. Treating Δ_2 the same way as $\mathbb{R}^3$ yields 95% confidence polygons that partly fall outside the ternary diagram, corresponding to meaningless negative values of x, y and z.

[12] As a solution to this problem, Aitchison suggested to transform the data from Δ_{N-1} to $\mathbb{R}^{N-1}$ using the log-ratio transformation (Figure 6). After performing the desired ("traditional") statistical analysis on the transformed data in $\mathbb{R}^{N-1}$, the results can then be transformed back to Δ_{N-1} using the inverse log-ratio transformation. For example, in the ternary system (X + Y + Z = 1), we could use the transformed values V = log(X/Z) and W = log(Y/Z). Alternatively, we could also use V = log(X/Y) and W = log(Z/Y), or V = log(Y/X) and W = log(Z/X). The inverse log-ratio transformation is given by

$$X = \frac{e^V}{e^V + e^W + 1}, Y = \frac{e^W}{e^V + e^W + 1}, Z = \frac{1}{e^V + e^W + 1} \quad (6)$$

[13] The back-transformed confidence regions of Figure 6 are no longer elliptical, but completely

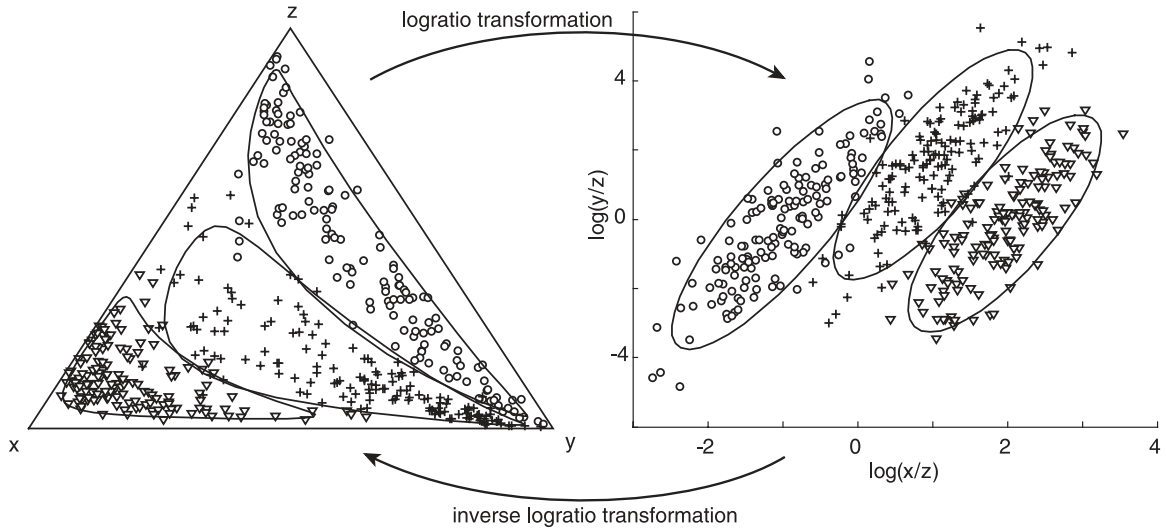

Figure 6. Following *Aitchison* [1986], the statistical problems of Figure 5 can be avoided by mapping the data from the simplex Δ_2 to $\mathbb{R}^2$ using the logratio transformation.

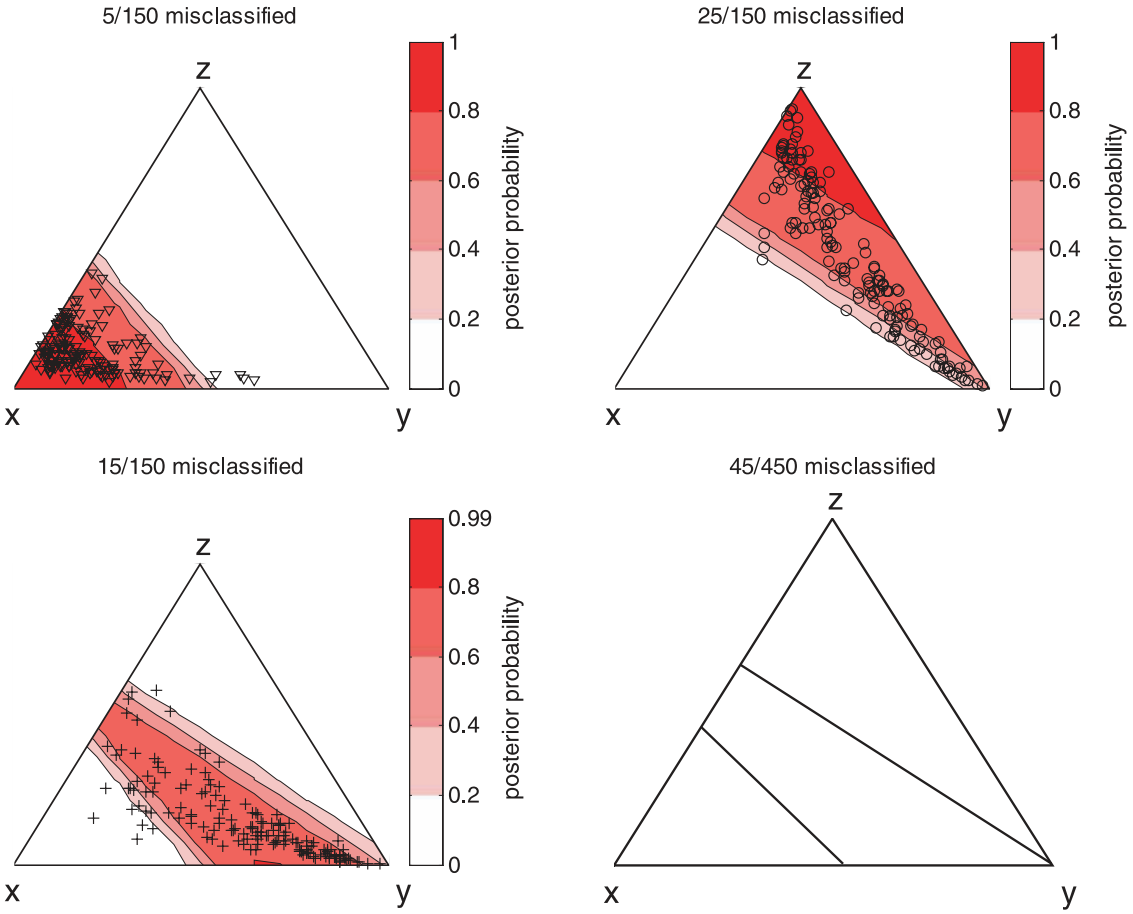

Figure 7. Linear discriminant analysis using the crude covariance approach of Figure 5. The red-shaded contours of the first three ternary diagrams represent the posterior probabilities for the three classes. The last diagram shows the linear decision boundaries. Ten percent of the training data are misclassified.

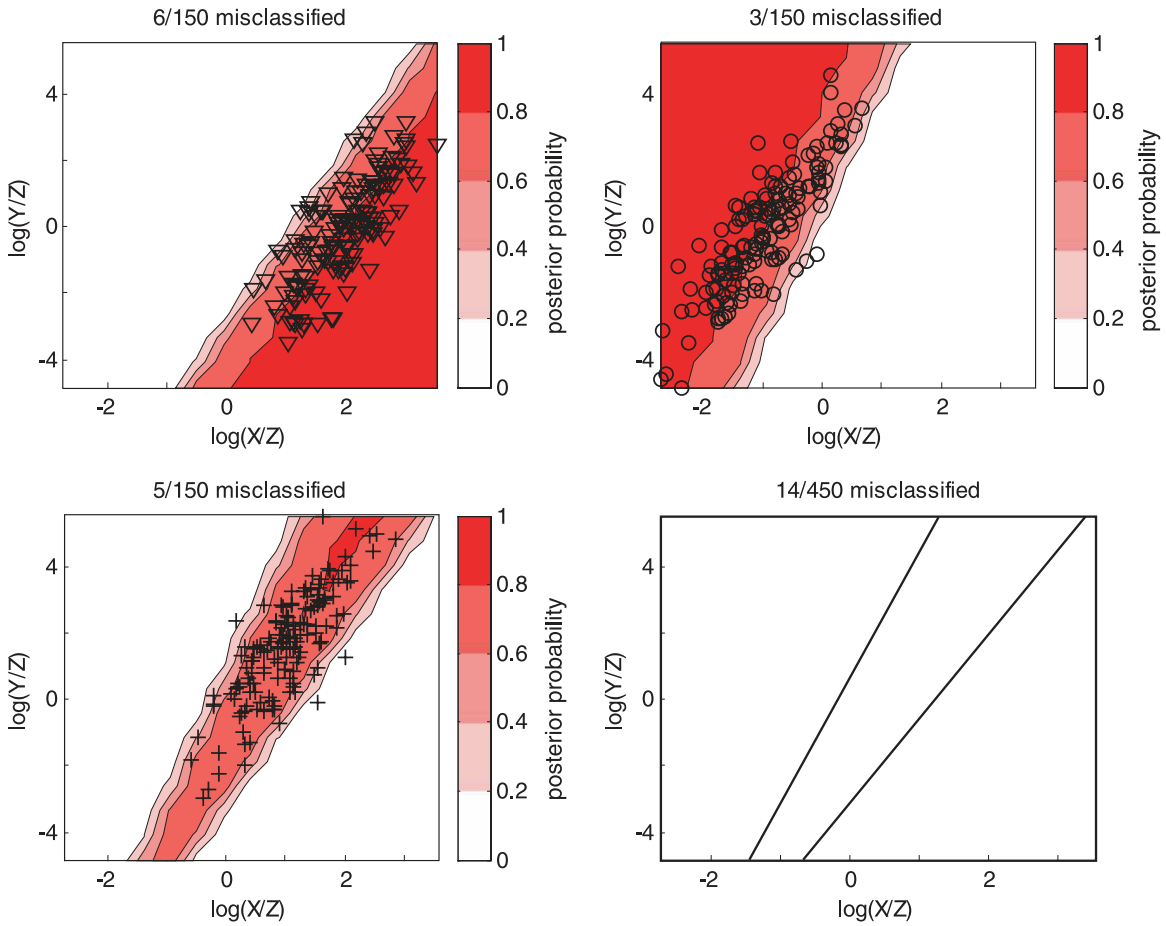

Figure 8. The same data of Figure 7, mapped to logratio space using the approach illustrated by Figure 6. Linear discriminant analysis of these bivariate data misclassifies only 3% of the training data.

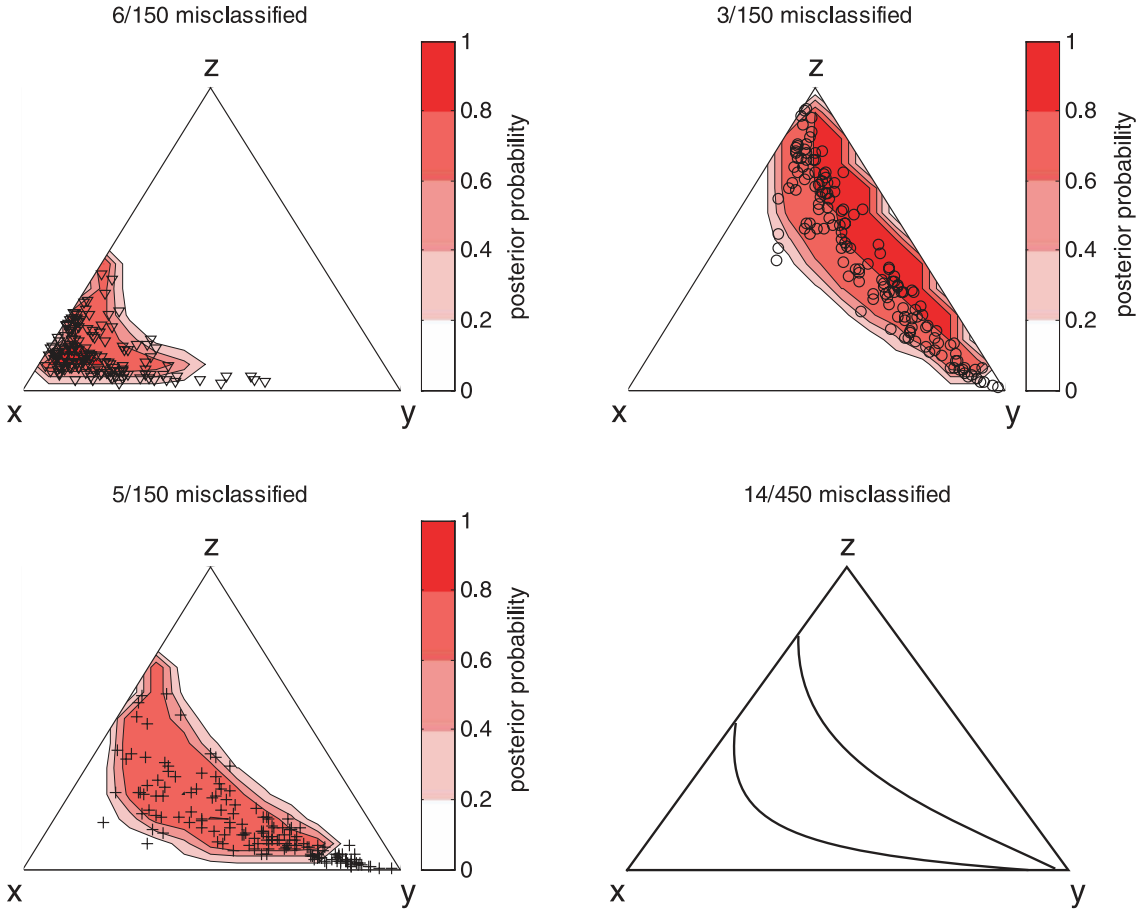

Figure 9. Mapping the results of Figure 8 back to the ternary diagram with the inverse logratio transformation shown in Figure 6 yields curved posterior densities and decision boundaries.

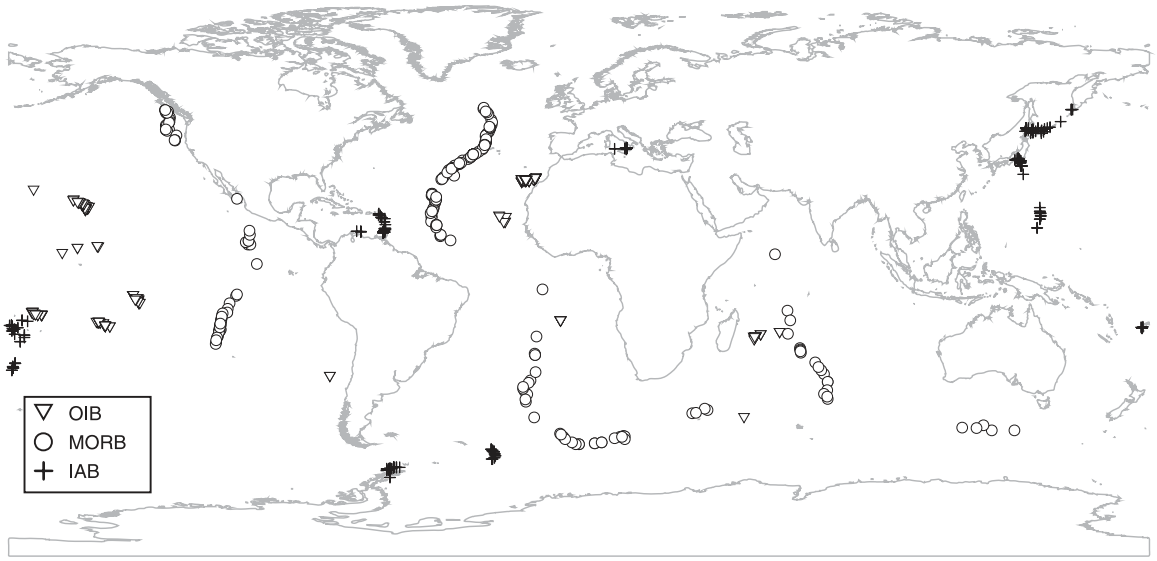

Figure 10. Locations of the training data: 756 island arc (IAB), mid-ocean ridge (MORB), and ocean island (OIB) basalts.

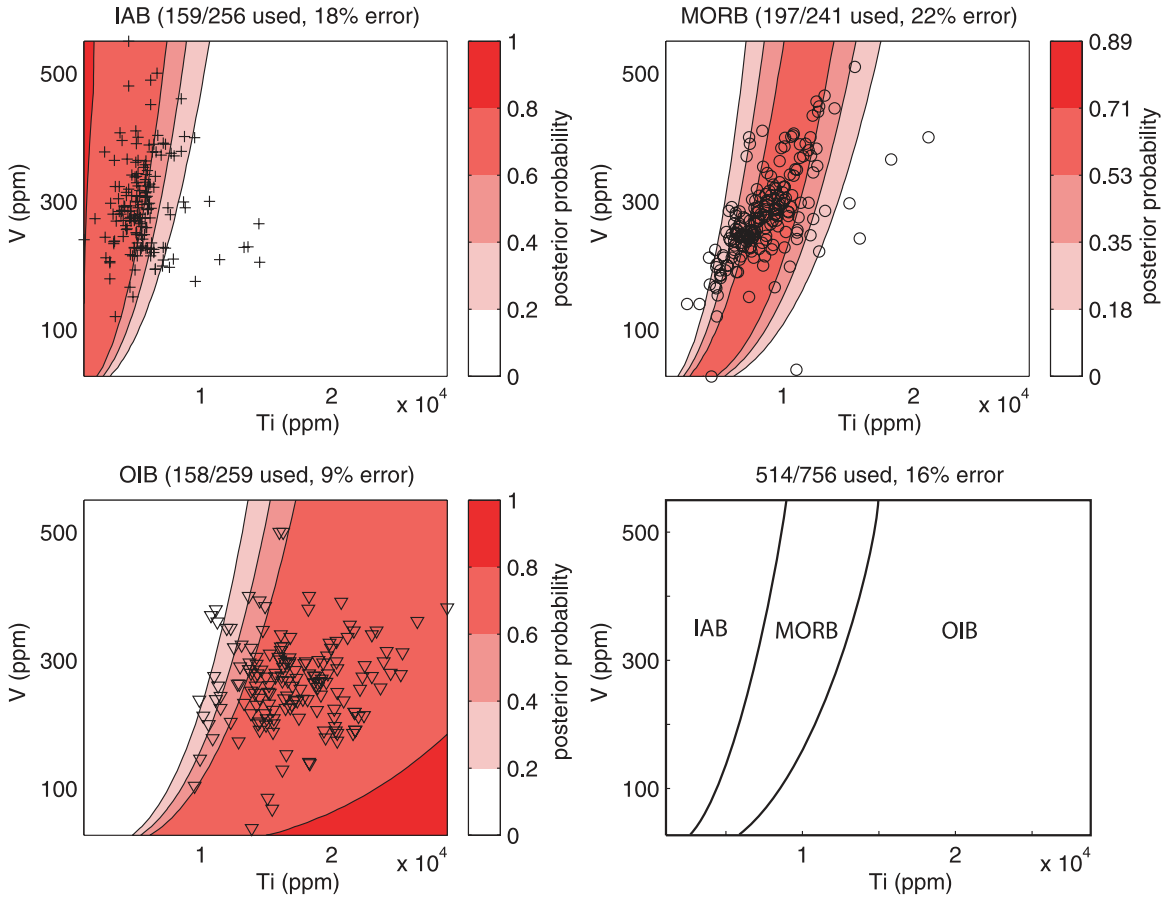

Figure 11. Linear discriminant analysis (LDA) of the Ti-V system of *Shervais* [1982]. The red-shaded contours on the first three subplots show the posterior probability of a particular "class" (IAB, MORB, or OIB) given the training set of 756 basalt samples and a uniform prior. The last subplot (lower right) shows the new decision boundaries. The number of training data used and a resubstitution error estimate are given for each of the tectonic affinities. The overall resubstitution error is shown above the lower right subplot.

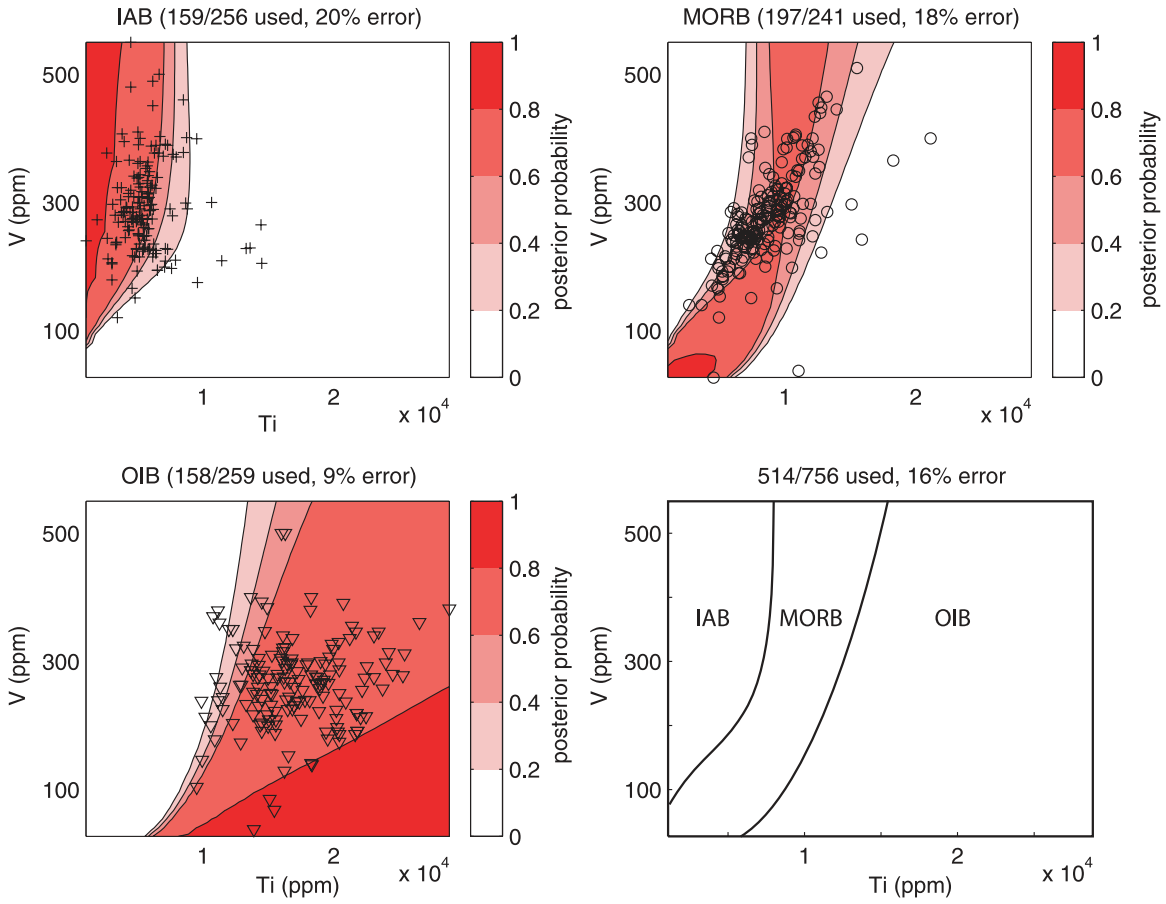

Figure 12. Quadratic discriminant analysis (QDA) of the Ti-V system. In contrast with the LDA of Figure 11, each tectonic "class" was allowed to have a different covariance matrix, resulting in slightly different decision boundaries.

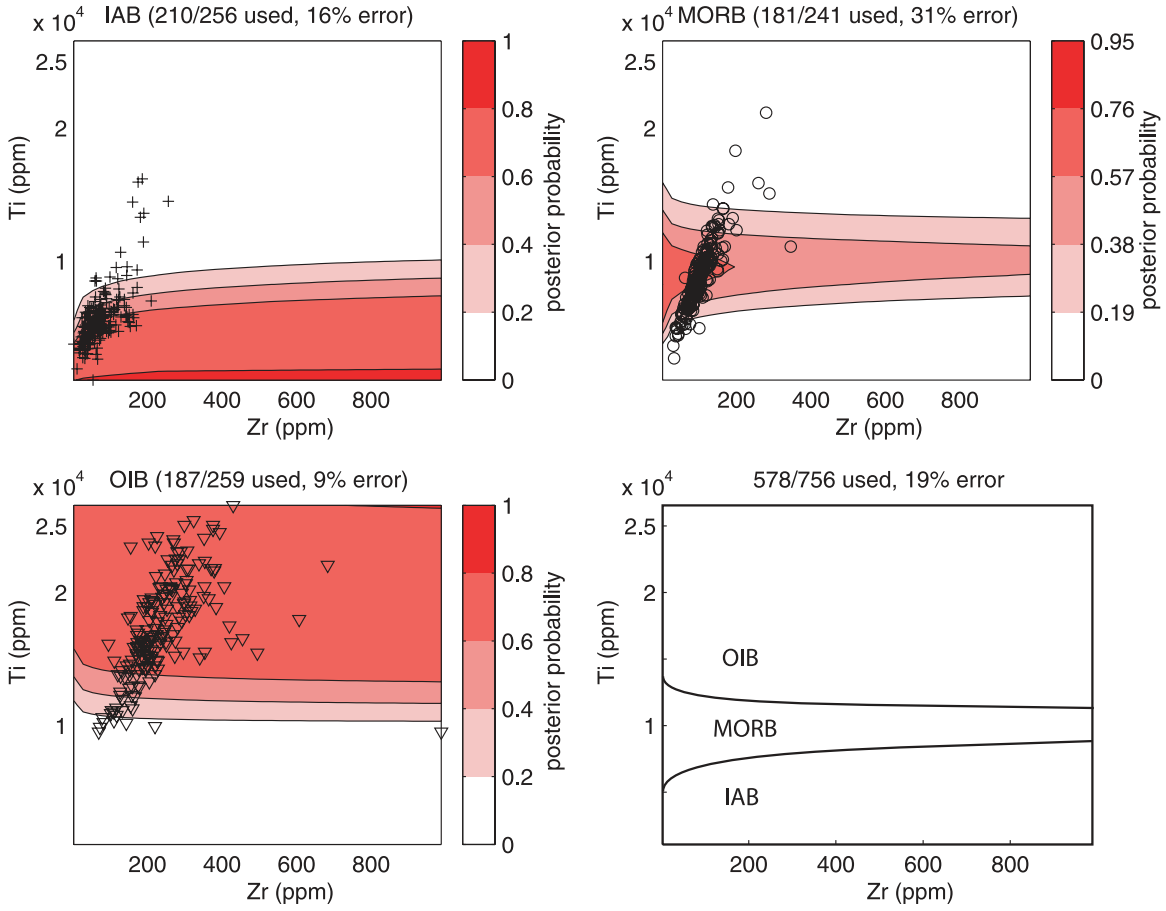

Figure 13. Linear discriminant analysis of the Ti-Zr system of *Pearce and Cann* [1973].

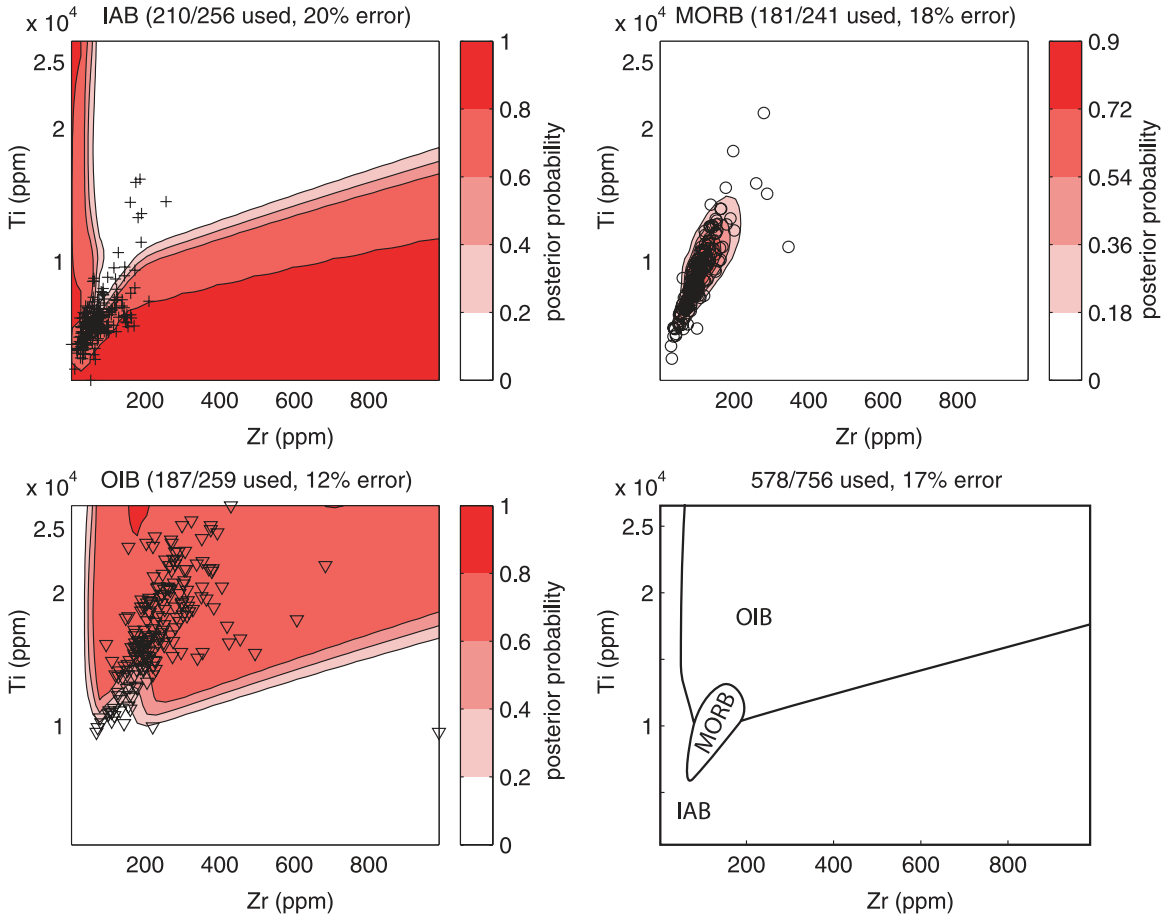

Figure 14. Quadratic discriminant analysis of the Ti-Zr system.

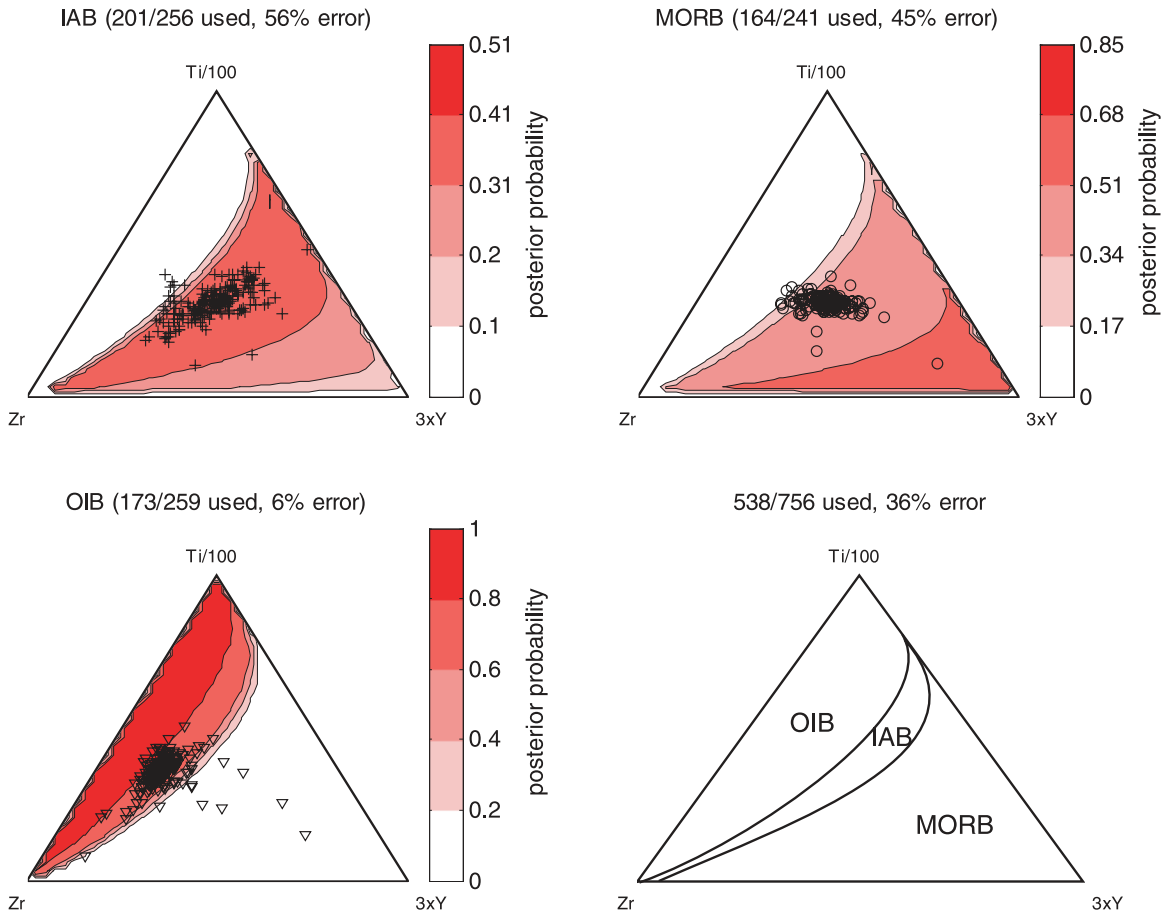

Figure 15. Linear discriminant analysis of the Ti-Zr-Y system of *Pearce and Cann* [1973]. The posterior probabilities of nearly all the IAB and MORB training data are low (<0.4), resulting in large misclassification rates for these affinities. As noted by *Pearce and Cann* [1973], the Ti-Zr-Y diagram can be used to separate OIBs from IAB/MORBs but cannot be used to distinguish between IAB and MORB. For this purpose, the Ti-Zr diagram (Figure 13) can be used.

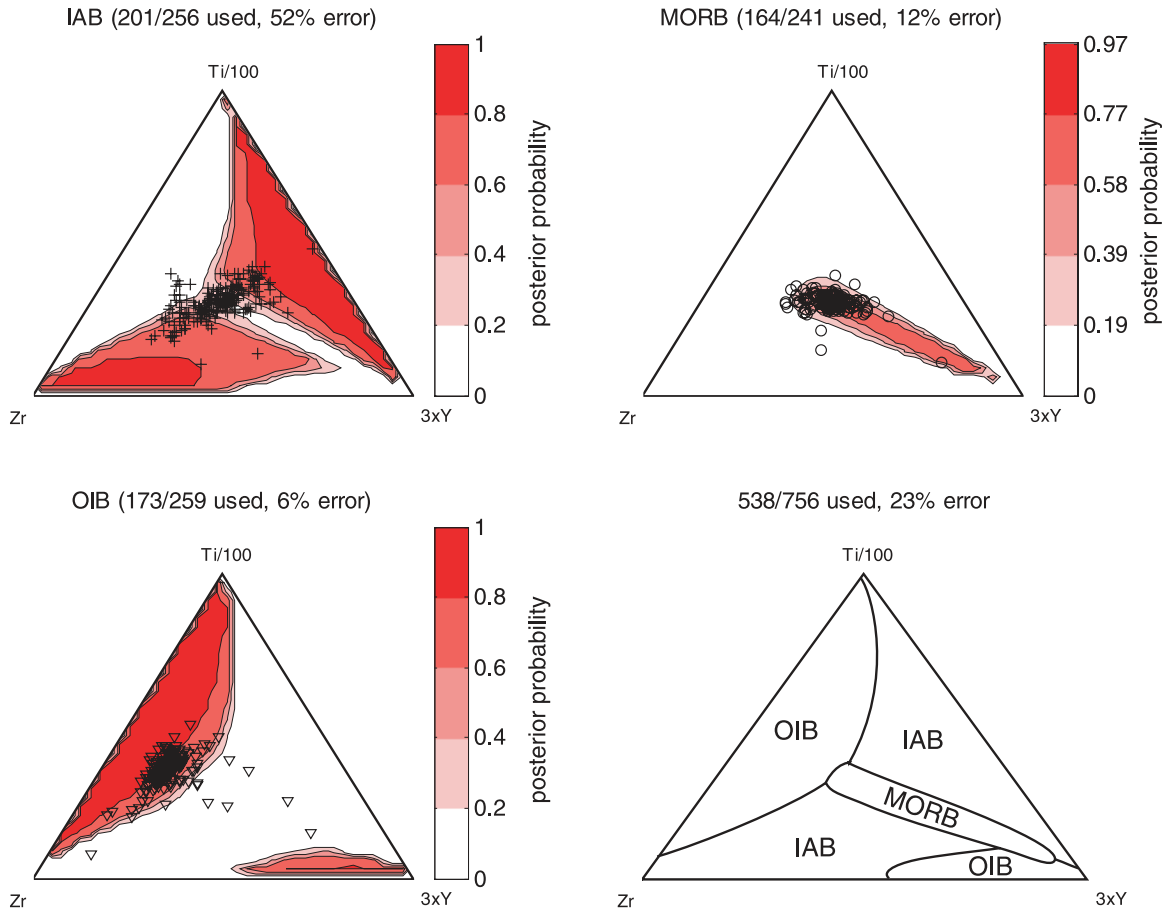

Figure 16. Quadratic discriminant analysis of the Ti-Zr-Y system. The OIB/IAB decision boundary (at low Y) is nearly identical to that of Figure 15, whereas there is a lot more (unstable) structure at higher Y concentrations.

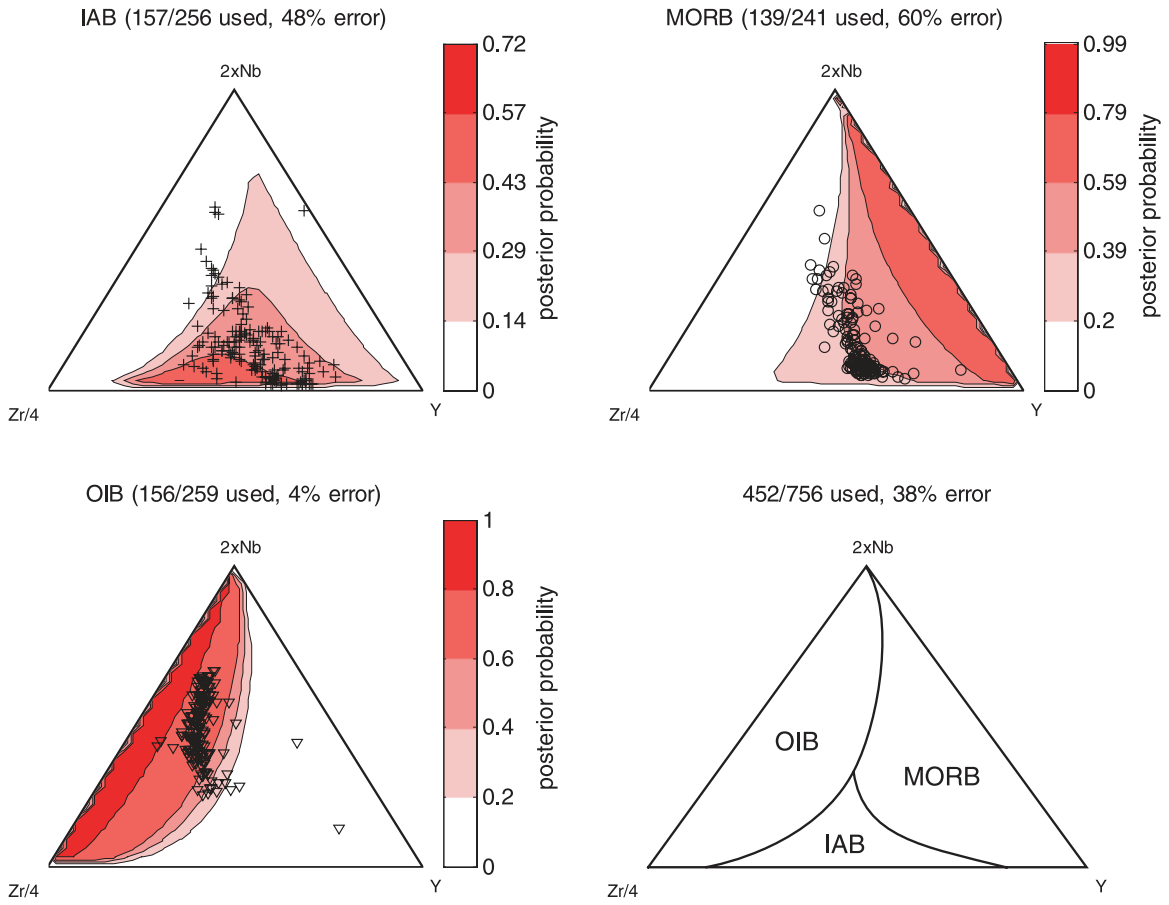

Figure 17. Linear discriminant analysis of the Zr-Y-Nb system of *Meschede* [1986]. Like in Figure 15, posterior IAB and MORB probabilities are low, resulting in high misclassification rates.

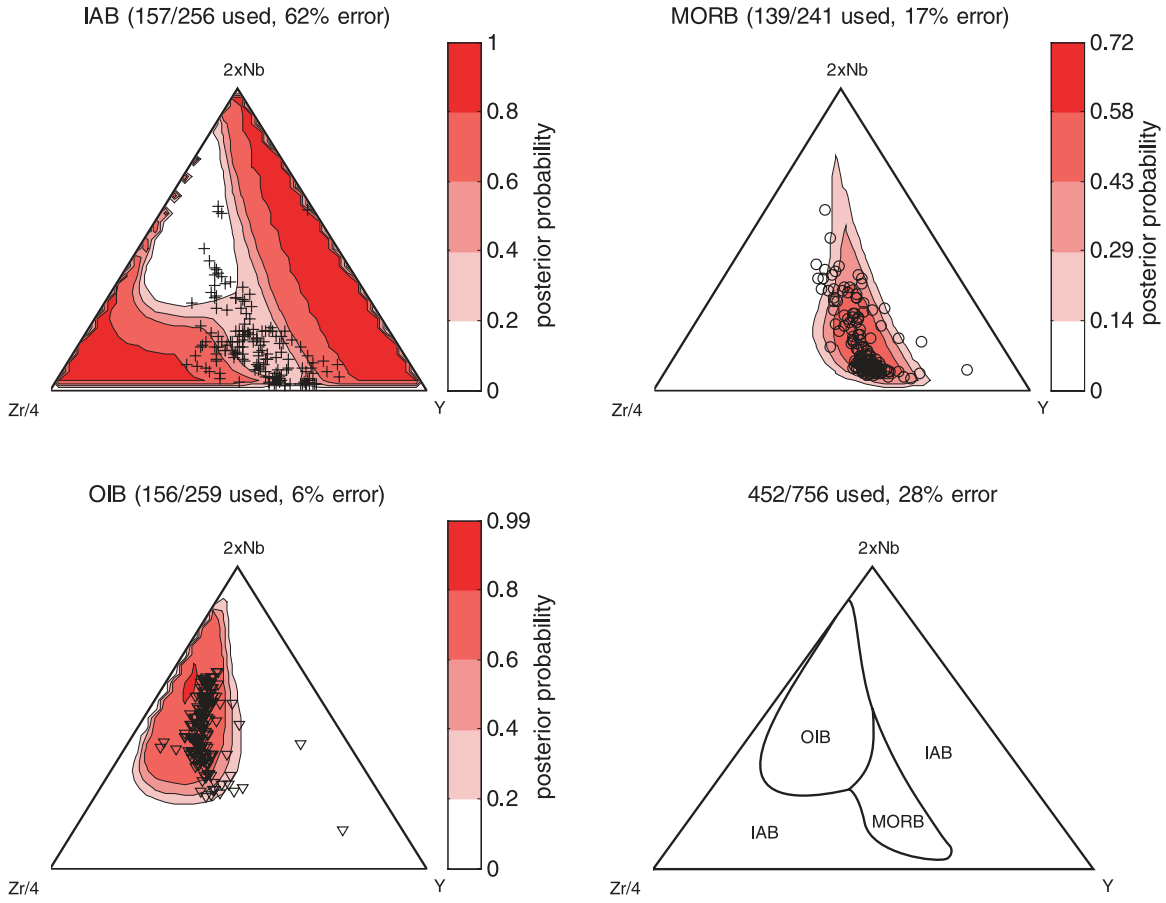

Figure 18. Quadratic discriminant analysis of the Zr-Y-Nb system.

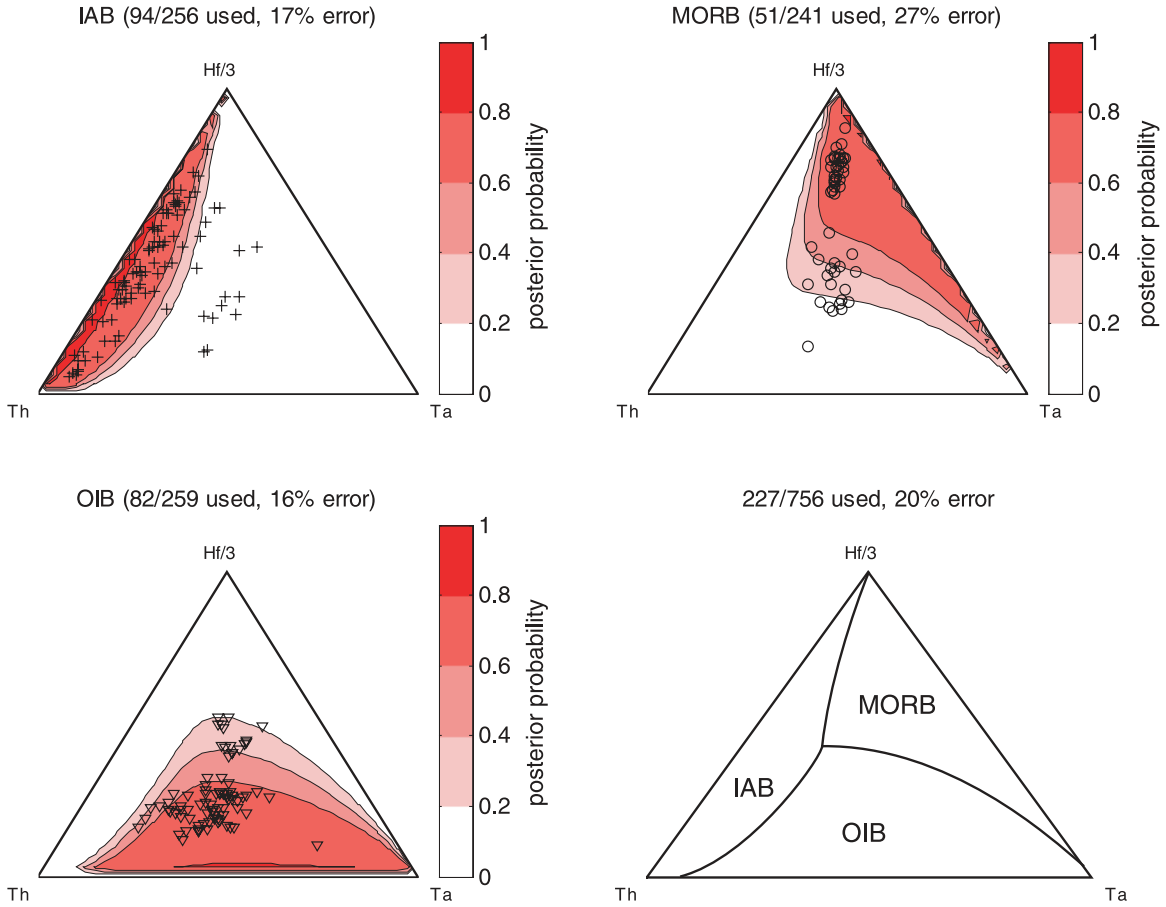

Figure 19. Linear discriminant analysis of the Th-Ta-Hf system of *Wood* [1980].

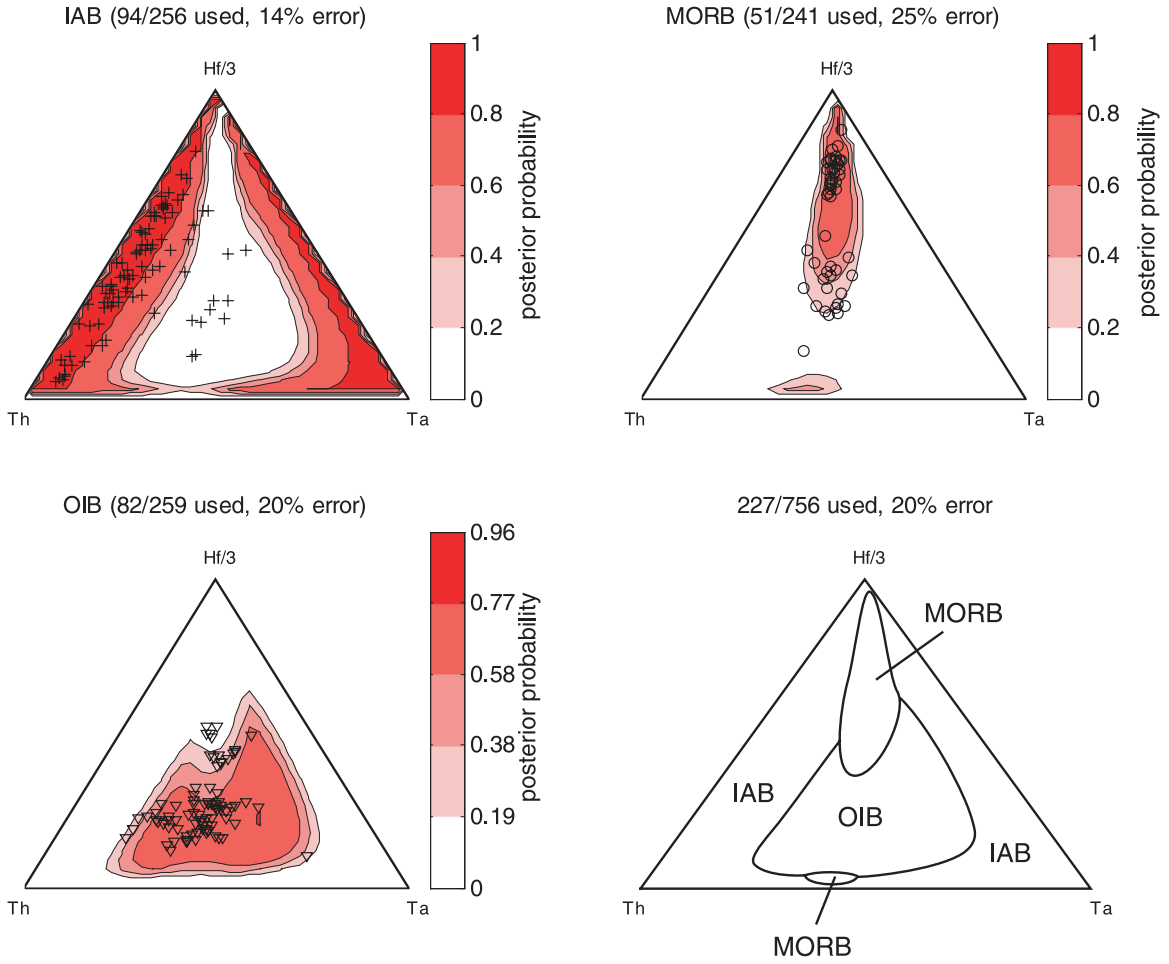

Figure 20. Quadratic discriminant analysis of the Th-Ta-Hf system.

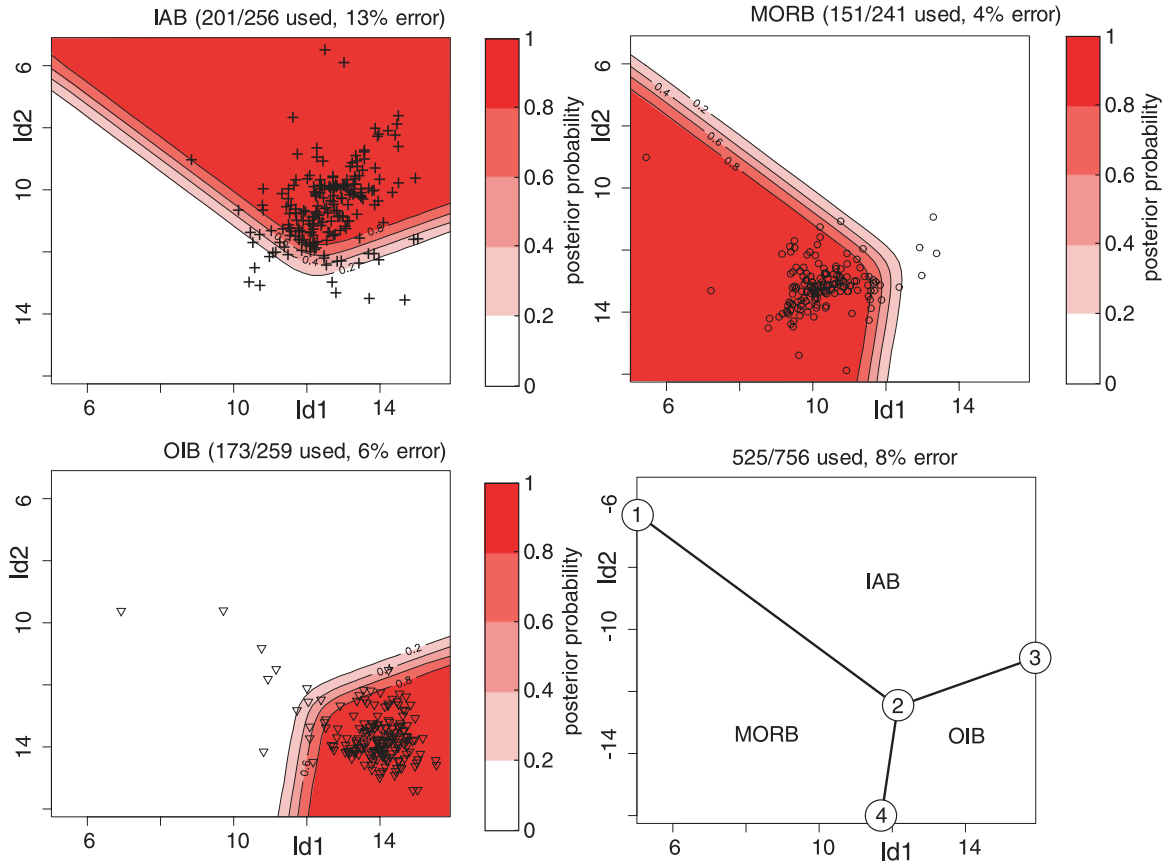

Figure 21. Linear discriminant analysis of the Ti-Zr-Y-Sr system. ld1 and ld2 are the two linear discriminant functions, given by equation (7). They represent two projection planes that optimally separate the three tectonic affinities (IAB, MORB, and OIB) (see also Figure 2). The encircled numbers on the lower right subplot are "anchor points" that can be used by the user to reconstruct the decision boundaries in logratio space. The ld1/ld2 coordinates of these anchor points are given in Table 6.

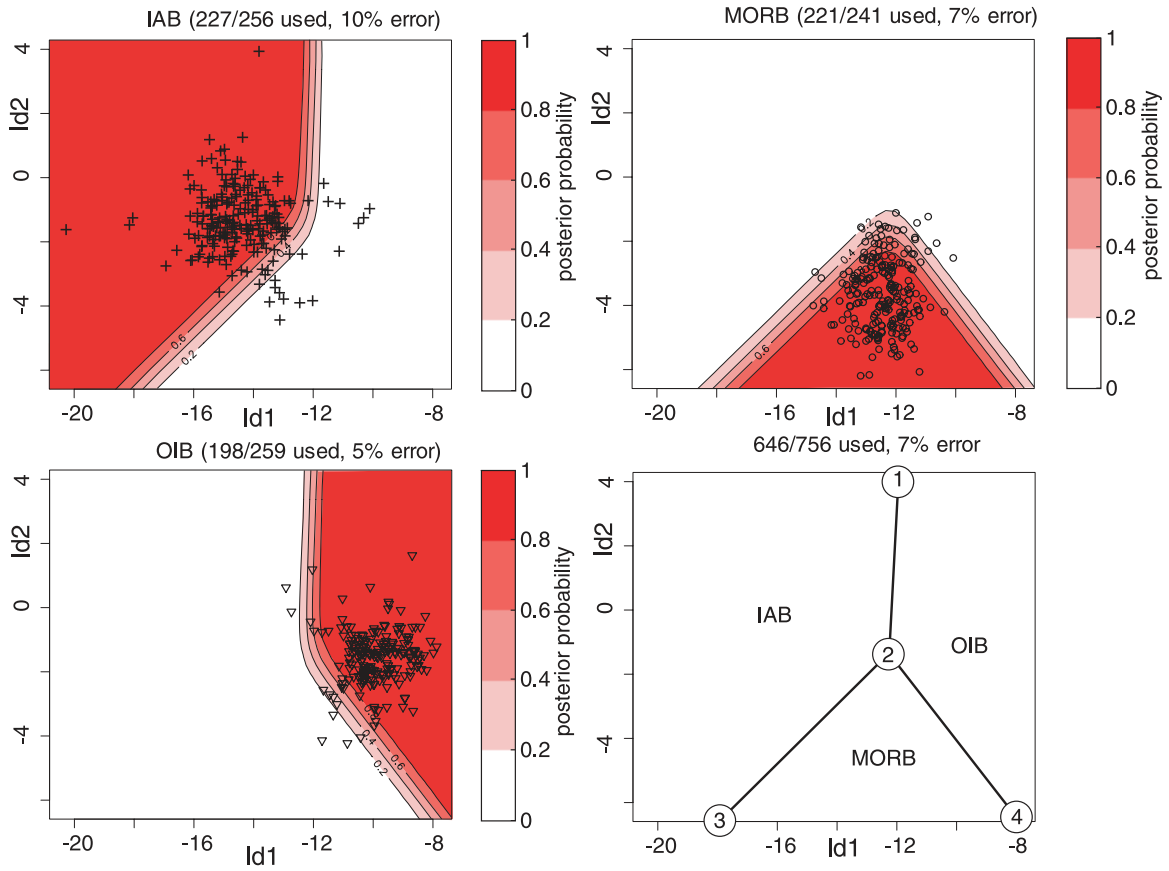

Figure 22. Linear discriminant analysis of major element data (SiO_2, Al_2O_3, TiO_2, CaO, MgO, MnO, K_2O, Na_2O), mapped to $\mathbb{R}^2$ using the logratio transformation. ld1 and ld2 are given by equation (8). Anchor points are given in Table 6.

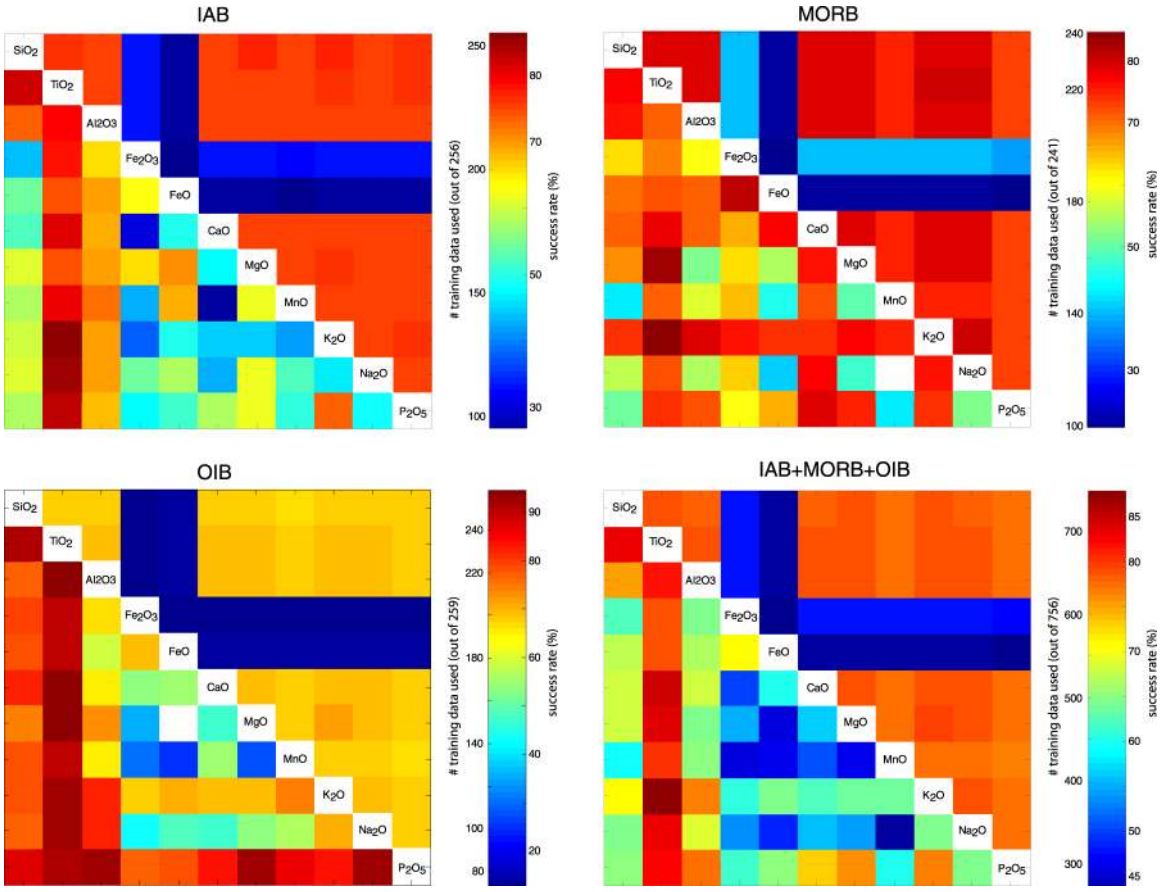

Figure 23. Visual representation of the performance of all possible bivariate linear discriminant analyses using the major element data of the training set of 756 oceanic basalts. The upper right triangular section of each matrix shows the number of samples that contained both variables. The lower left sections color-code the fraction of successfully classified training data.

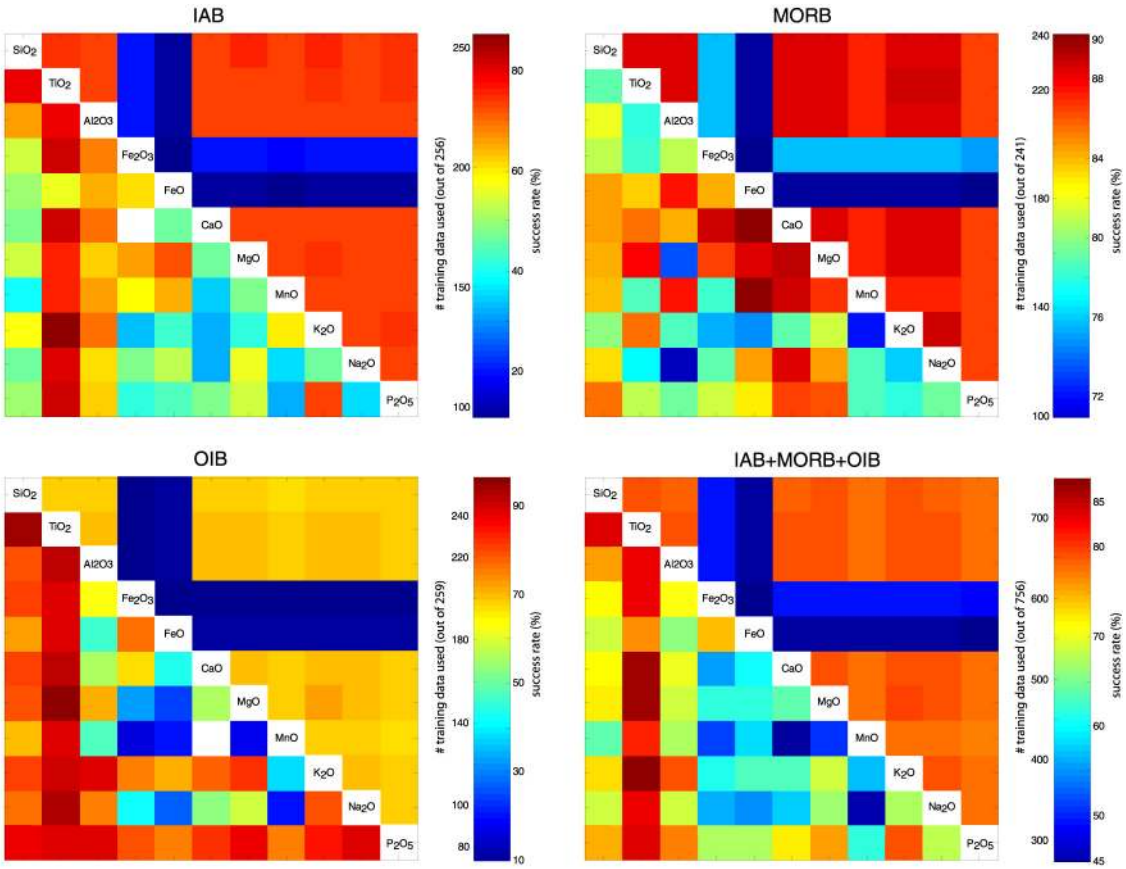

Figure 24. Same as Figure 23, but for quadratic discriminant analysis.

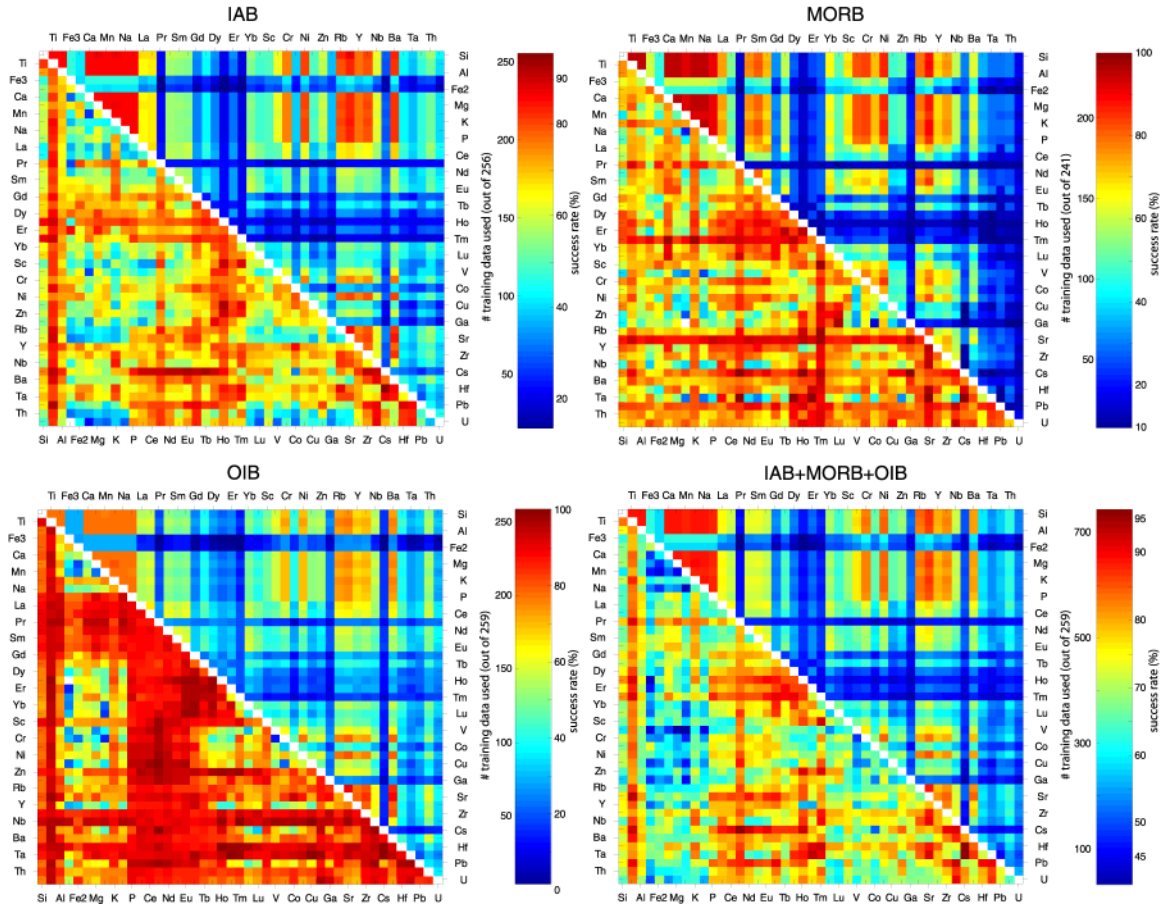

Figure 25. Matrices showing the performance of all possible bivariate linear discriminant analyses using combinations of 45 elements.

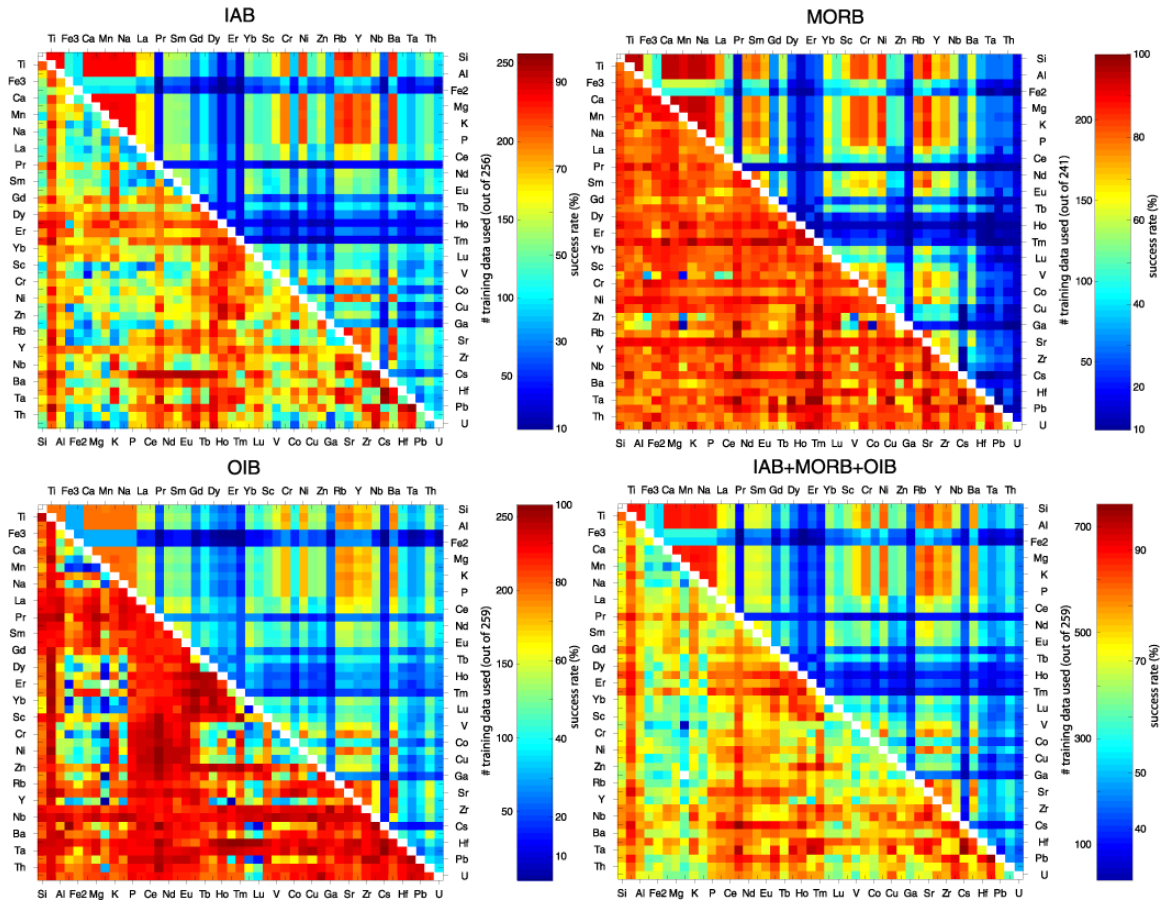

Figure 26. Same as Figure 25, but for quadratic discriminant analysis.

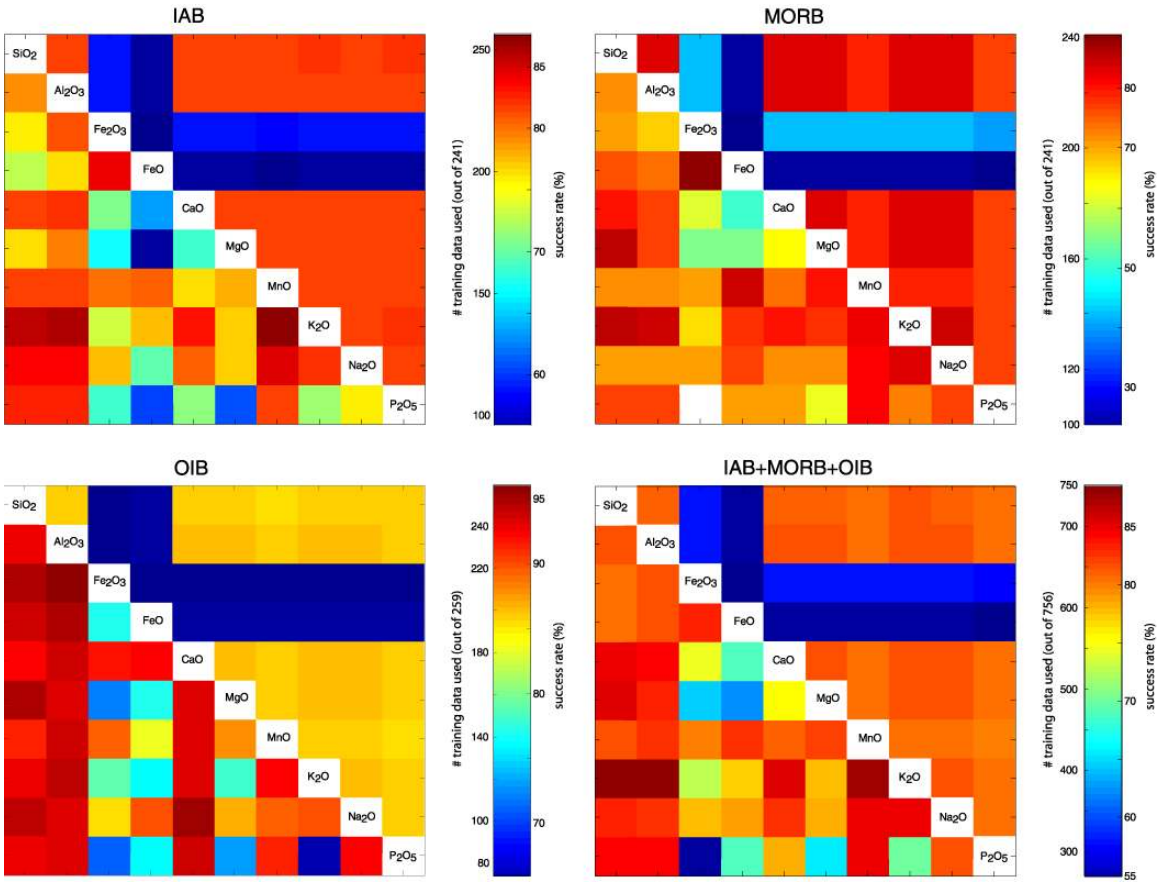

Figure 27. Performance analysis of all possible ternary linear discriminant analyses using TiO_2 and other major element oxides.

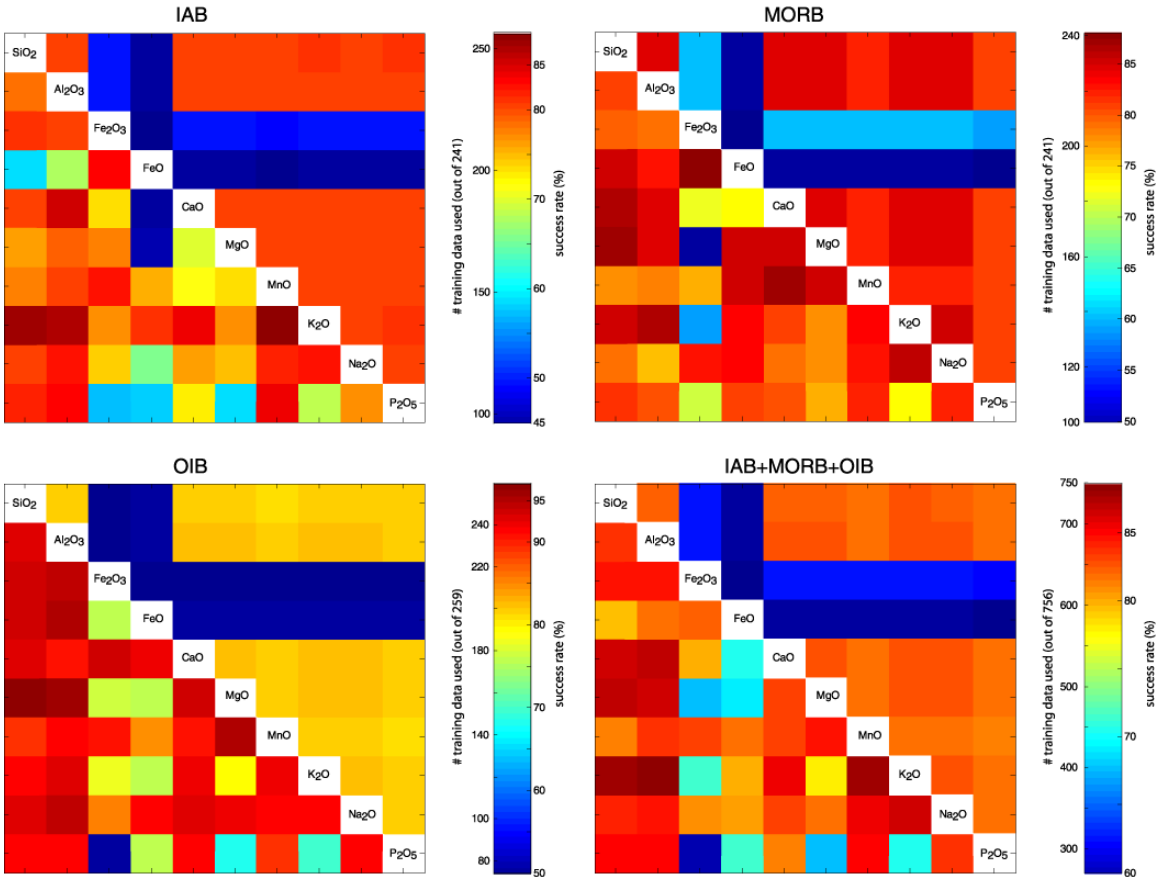

Figure 28. Same as Figure 27, but using quadratic discriminant analysis.

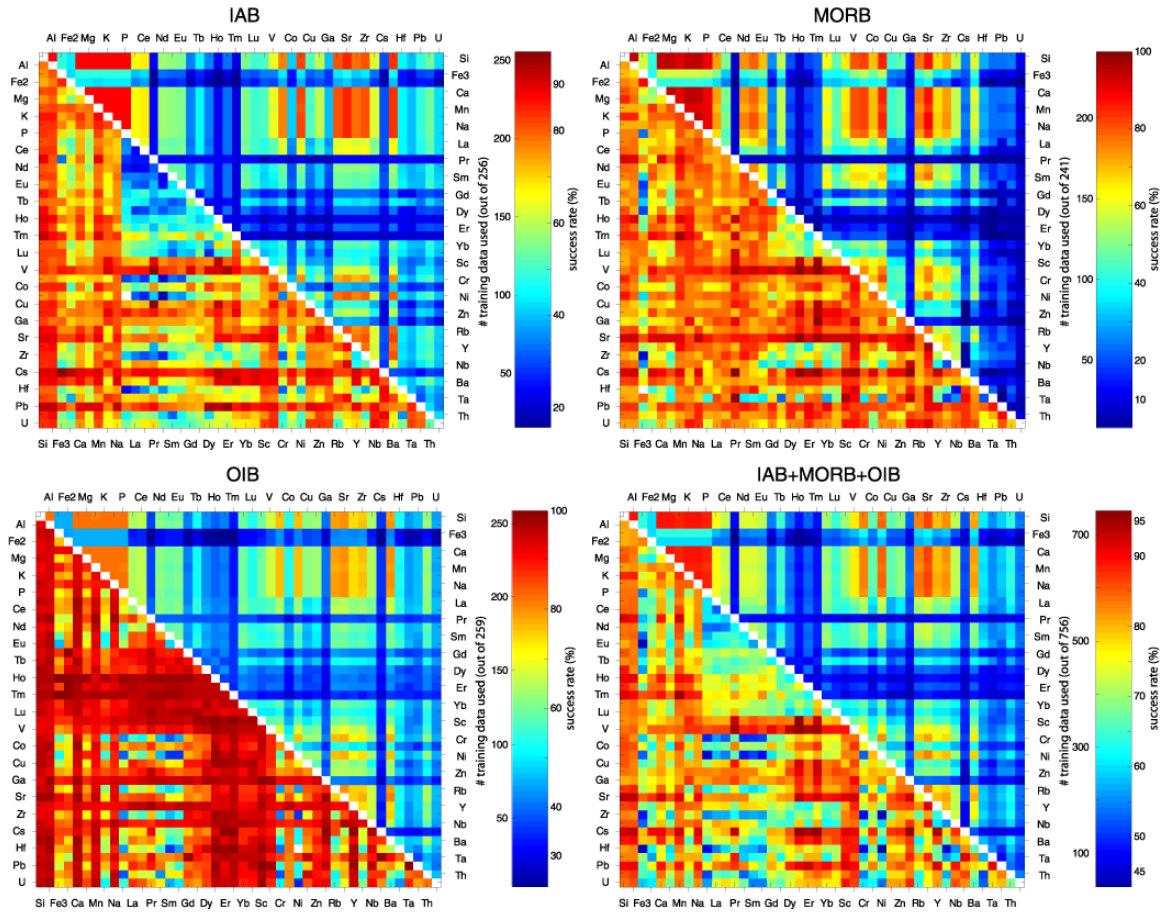

Figure 29. Performance analysis of all possible ternary linear discriminant analyses using Ti and two of 45 other elements.

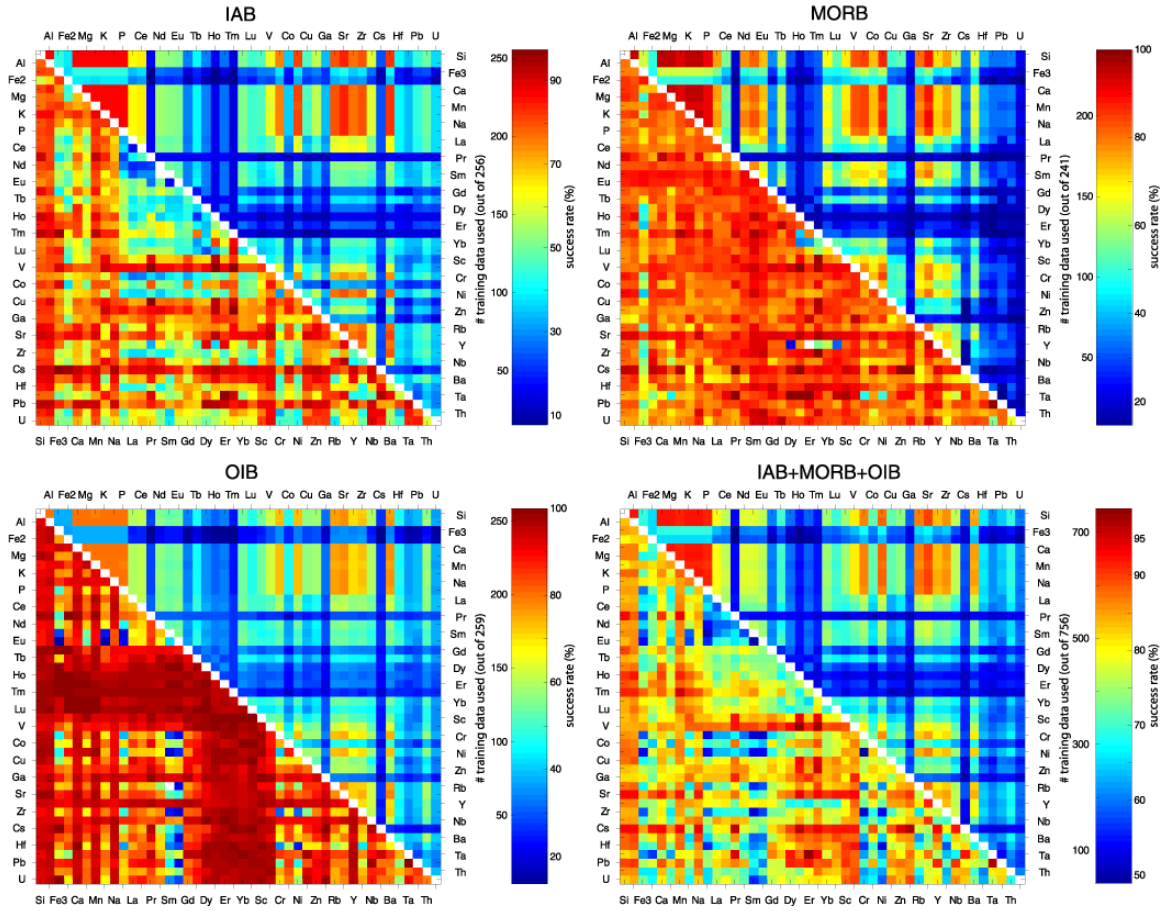

Figure 30. Same as Figure 29, but using quadratic discriminant analysis.

Table 1. The 100 Best Ternary Linear Discrimination Diagrams

	Elements			Resubstitution Error, %				# IAB	# MORB	# OIB
Rank	1	2	3	Overall	IAB	MORB	OIB	(/256)	(/241)	(/259)
1	Si	Sr	Ti	6.2	10.0	6.6	2.1	221	211	192
2	Ti	Sr	Al	6.5	10.0	7.6	2.1	220	211	194
3	Eu	Sr	Lu	6.6	10.5	6.0	3.3	124	117	120
4	Sr	Nb	Y	6.6	13.4	3.9	2.5	157	127	160
5	Ca	Nb	Sr	7.6	16.6	4.8	1.4	157	126	142
6	Ti	Sr	V	7.7	9.5	7.2	6.4	158	180	156
7	Eu	Y	Sr	7.8	16.1	4.7	2.5	124	106	121
8	Ti	Sr	Ca	7.8	12.3	9.0	2.1	220	211	194
9	Ti	Sr	Sc	7.9	12.6	7.1	3.9	119	155	128
10	Al	Nb	Sr	8.1	20.4	3.2	0.7	157	126	142
11	Ti	Sr	Mn	8.1	11.9	8.3	4.2	219	204	191
12	Ti	Y	Sr	8.4	12.9	3.9	8.5	202	153	177
13	Eu	Sr	Yb	8.4	15.2	5.1	5.0	138	157	141
14	Si	Nb	Sr	8.8	19.5	4.8	2.1	159	126	142
15	Ti	Sr	Na	8.8	13.6	6.1	6.7	220	213	194
16	Na	Nb	Sr	9.0	22.3	4.0	0.7	157	126	142
17	Tb	Sr	Lu	9.2	11.0	9.8	6.7	100	102	105
18	Ti	Nb	Sr	9.3	10.1	14.3	3.4	158	126	149
19	Mn	Nb	Sr	9.4	19.7	6.3	2.1	157	126	140
20	Nd	Y	Sr	9.5	19.6	6.7	2.4	138	135	127
21	Ti	Ba	Al	9.5	11.1	16.0	1.6	217	144	192
22	Na	Zr	Sr	9.6	18.8	5.5	4.5	208	165	177
23	Ti	Sr	Lu	9.6	11.5	6.2	11.1	113	113	108
24	Al	Sr	Nd	10.1	20.9	4.5	4.8	139	177	125
25	Al	Zr	Sr	10.1	20.2	5.5	4.5	208	163	177
26	V	Nb	Sr	10.2	18.9	9.4	2.4	122	117	124
27	Tb	Sr	Yb	10.3	14.7	6.4	9.9	102	125	111
28	Ti	V	Sc	10.4	15.2	10.1	5.8	105	148	121
29	Ti	Ba	Na	10.4	9.7	15.8	5.7	217	146	192
30	V	Nb	Rb	10.4	10.7	14.2	6.5	122	113	123
31	K	Nb	V	10.6	10.7	14.0	7.0	121	129	114
32	Ti	V	Sm	10.6	17.3	6.8	7.6	104	162	105
33	Sr	Zr	Y	10.6	21.6	3.9	6.4	204	155	203
34	Ti	Sr	Yb	10.6	13.4	6.7	11.8	127	150	127
35	Na	Sr	Nd	10.7	22.3	7.3	2.4	139	179	125
36	Si	Ba	Ti	10.8	12.0	17.4	3.2	217	144	190
37	Ca	Sr	Nd	10.9	20.9	6.2	5.6	139	177	125
38	Nd	Sr	Yb	10.9	19.4	9.7	3.8	134	145	133
39	Sm	Y	Sr	11.0	21.9	5.1	5.9	128	137	119
40	Al	Sr	Eu	11.0	20.9	7.1	5.0	129	154	120
41	Yb	Zr	Sr	11.0	19.8	6.5	6.7	126	107	134
42	Ti	Ba	Sc	11.2	13.1	16.5	3.9	122	115	129
43	Sc	Zr	Sr	11.2	21.0	6.8	5.7	119	118	122
44	Si	Sr	Nd	11.3	21.9	6.2	5.6	146	177	124
45	Si	Sr	Eu	11.3	18.5	7.8	7.6	135	154	118
46	Sm	Sr	Lu	11.3	23.0	6.0	5.0	122	116	119
47	Ti	Ba	Mn	11.4	11.1	18.2	4.8	216	137	189
48	Mn	Zr	Sr	11.4	21.7	5.0	7.5	207	160	174
49	Si	Zr	Sr	11.4	21.8	6.1	6.3	211	163	175
50	Ti	K	Al	11.5	13.6	15.4	5.4	228	228	203
51	Nd	Sr	Lu	11.5	20.7	9.5	4.5	121	105	112
52	Ti	Y	V	11.6	19.6	8.4	6.8	153	155	147
53	Ti	Sc	K	11.6	10.7	15.4	8.7	122	162	126
54	Ti	Rb	Al	11.6	12.4	18.2	4.2	209	187	189
55	Ti	Ba	Ca	11.6	12.0	20.8	2.1	217	144	192
56	Si	K	Ti	11.7	14.0	14.0	7.0	229	228	201
57	Ti	Ba	V	11.7	10.8	16.0	8.4	158	125	155
58	Ti	Sr	Zn	11.7	12.8	11.9	10.6	149	109	142
59	Ti	V	Nd	11.9	16.8	9.0	9.7	113	155	113
60	Na	Sr	Ce	11.9	26.1	6.7	2.9	165	119	140
61	Ca	Zr	Sr	11.9	24.0	6.1	5.6	208	163	177

Table 1. (continued)

	Elements			Resubstitution Error, %				# IAB	# MORB	# OIB
Rank	1	2	3	Overall	IAB	MORB	OIB	(/256)	(/241)	(/259)
62	Eu	Sr	V	12.0	18.8	7.6	9.4	101	131	106
63	Ca	Nb	K	12.0	14.6	14.5	7.0	157	138	143
64	Mn	Sr	Nd	12.1	21.6	10.6	4.1	139	170	123
65	K	Nb	Y	12.2	14.2	16.2	6.1	155	136	147
66	Al	Sr	Ce	12.3	24.2	6.8	5.7	165	117	140
67	Ti	V	K	12.3	9.4	14.2	13.2	159	197	151
68	Si	Rb	Ti	12.3	12.4	19.3	5.3	210	187	189
69	Ti	V	Na	12.3	14.5	15.2	7.3	159	197	151
70	Al	Nb	K	12.4	16.6	13.8	7.0	157	138	143
71	Al	Nb	Rb	12.5	14.7	16.4	6.3	156	122	142
72	Ti	K	Mn	12.5	12.3	17.6	7.5	227	221	200
73	Mg	Nb	Sr	12.6	19.0	7.1	11.6	158	126	147
74	Sr	Nb	Zr	12.7	15.6	20.5	1.9	160	132	157
75	Ca	Sr	Ce	12.7	23.0	8.5	6.4	165	117	140
76	Nd	Sr	V	12.7	22.8	9.8	5.4	114	153	111
77	K	Nb	Na	12.7	16.6	15.2	6.3	157	138	143
78	Mn	Sr	Eu	12.8	18.6	9.5	10.2	129	147	118
79	Eu	Sr	Tb	12.8	23.6	7.6	7.1	106	131	113
80	Si	Sr	Ce	12.8	24.7	8.5	5.1	170	117	138
81	Na	Sr	P	12.8	27.3	6.4	4.7	220	202	192
82	K	Zr	Yb	12.8	18.4	11.3	8.7	125	106	115
83	Al	Sr	P	12.8	27.3	5.4	5.7	220	202	192
84	K	Lu	Eu	12.8	16.8	17.2	4.5	119	116	112
85	Ce	Sr	Lu	12.8	24.2	6.0	8.3	124	100	120
86	Ti	Y	Al	12.9	12.9	18.3	7.4	201	164	175
87	Si	Nb	K	12.9	15.7	15.9	7.0	159	138	143
88	Ti	V	Eu	13.0	21.0	9.6	8.3	100	146	108
89	Ca	Nb	Rb	13.0	14.7	17.2	7.0	156	122	142
90	Ce	Sr	Yb	13.0	23.2	8.7	7.1	138	126	141
91	Zn	Zr	Sr	13.1	21.8	8.3	9.2	147	109	152
92	P	Sr	Sc	13.1	26.1	6.6	6.7	119	151	120
93	Ti	V	Ce	13.1	13.5	14.4	11.5	126	104	122
94	Ti	Nd	Mn	13.2	21.1	13.8	4.6	142	174	131
95	Sm	Sr	Yb	13.3	25.0	7.7	7.3	136	156	137
96	Ca	Sr	Eu	13.3	23.3	8.4	8.3	129	154	120
97	V	Zr	Sr	13.4	21.7	8.2	10.3	157	147	156
98	Ti	Ce	Mn	13.4	19.9	14.9	5.5	171	114	145
99	Mn	Nb	K	13.5	15.3	17.4	7.8	157	138	141
100	Ti	Cu	V	13.5	13.1	15.7	11.7	107	108	120

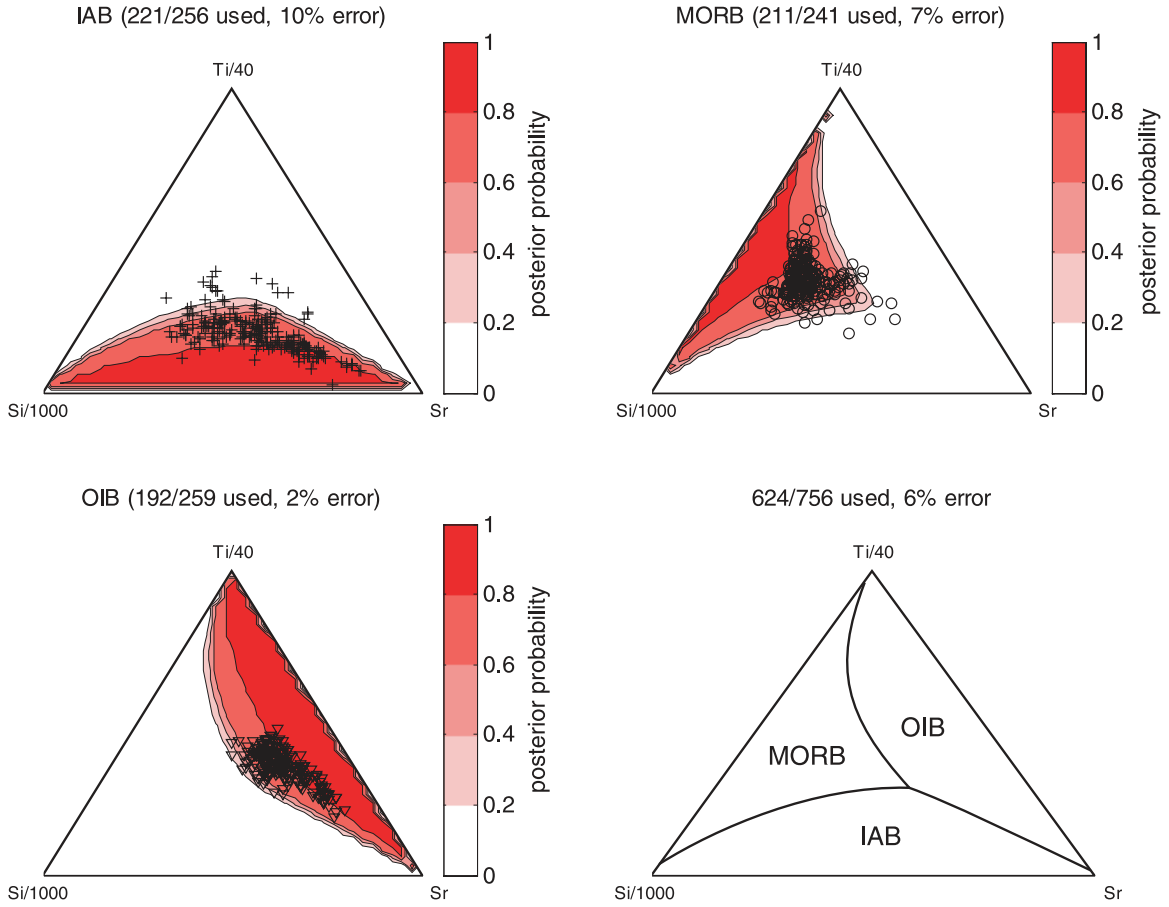

Figure 31. The best ternary linear discriminant analysis, using Si, Ti, and Sr.

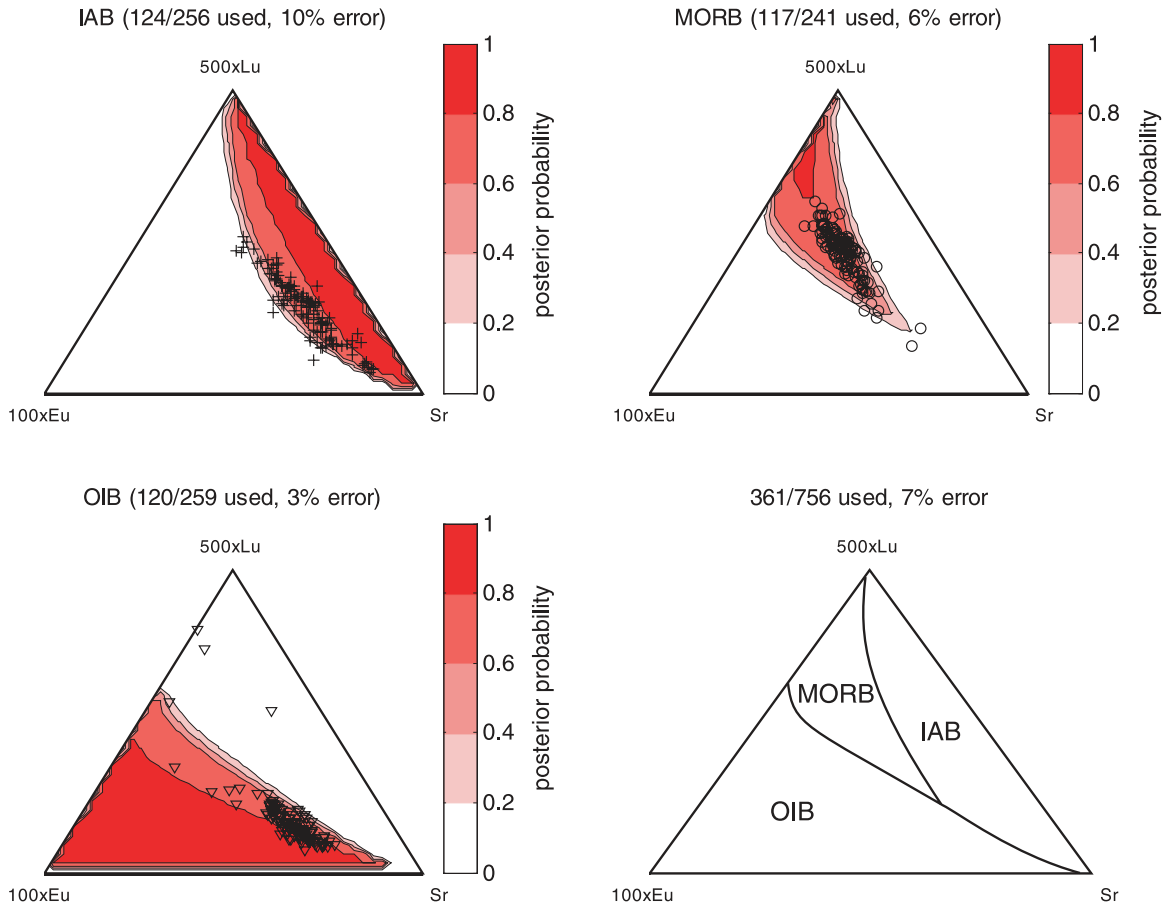

Figure 32. Linear discriminant analysis using Eu, Lu, and Sr.

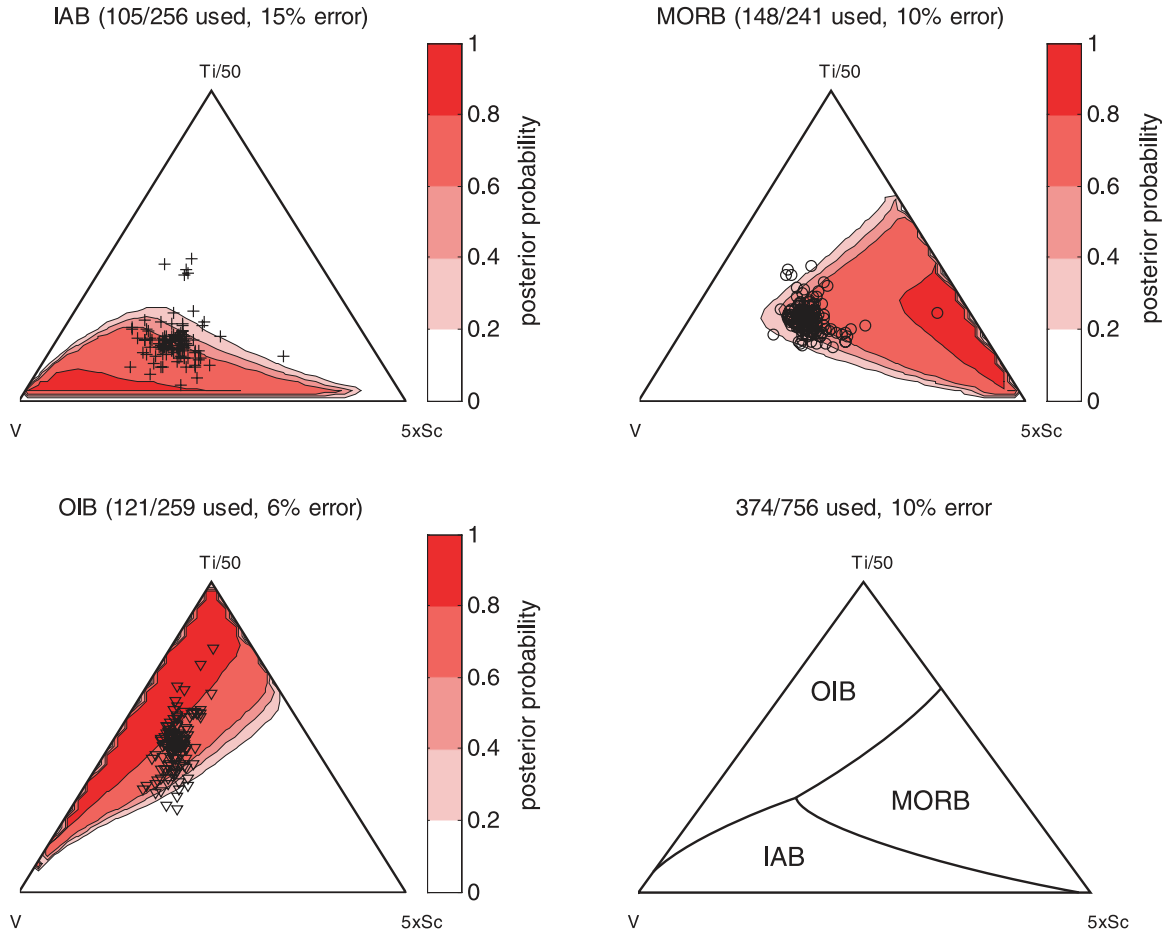

Figure 33. The best performing linear discriminant analysis using only incompatible elements (Ti, V, and Sc).

Table 2. The Best Ternary Linear Discrimination Diagrams Using Only Incompatible Elements

	Elements			Resubstitution Error, %				# IAB	# MORB	# OIB
Rank	1	2	3	Overall	IAB	MORB	OIB	(/256)	(/241)	(/259)
28	Ti	V	Sc	10.4	15.2	10.1	5.8	105	148	121
32	Ti	V	Sm	10.6	17.3	6.8	7.6	104	162	105
52	Ti	Y	V	11.6	19.6	8.4	6.8	153	155	147
59	Ti	V	Nd	11.9	16.8	9.0	9.7	113	155	113
93	Ti	V	Ce	13.1	13.5	14.4	11.5	126	104	122
101	Ti	V	La	13.5	13.6	17.5	9.5	125	143	116
108	Ti	Zr	V	13.9	19.2	11.7	10.8	156	162	148
159	V	Zr	Y	15.3	26.8	9.7	9.4	153	155	149
182	Ti	Nb	V	15.9	23.1	17.1	7.4	121	129	121
189	Ti	Cr	Sc	15.9	26.4	16.7	4.7	121	162	127
208	Ti	Cr	V	16.2	23.4	14.3	11.0	158	182	154
249	La	Nb	Zr	17.1	21.4	19.8	9.9	140	101	131
251	Nd	Nb	Y	17.1	35.7	11.5	4.1	129	113	123
252	Ti	Zr	Sc	17.1	22.9	22.5	5.9	118	120	119
267	V	Nb	Y	17.3	30.8	18.6	2.4	120	129	124
277	Nd	Y	V	17.5	38.1	9.6	4.9	113	125	103
324	Sm	Nb	Y	18.2	35.2	15.0	4.3	122	113	115
380	La	Y	V	19.1	31.5	16.0	9.8	124	106	112
385	La	Zr	V	19.2	32.0	17.3	8.2	125	110	110
393	Ti	Y	Sc	19.3	28.6	25.9	3.4	112	112	118
423	Ti	Sc	La	19.8	28.3	26.5	4.5	113	117	112
426	Nd	Zr	V	19.8	35.4	15.2	8.7	113	125	103
538	Sc	Zr	V	21.5	32.4	23.5	8.7	105	115	115
543	Ti	Y	Nd	21.6	42.6	16.4	5.7	136	134	122
551	Ti	Y	Sm	21.7	42.9	16.2	6.1	126	136	114
584	Ce	V	Yb	22.2	43.0	19.0	4.6	100	105	108
587	Sm	Zr	Y	22.3	50.4	8.7	7.8	127	138	116
591	Sc	Zr	Y	22.3	33.0	27.0	7.0	112	115	115
612	La	Zr	Yb	22.6	38.9	21.0	7.9	126	105	127
624	Ti	Y	La	22.7	38.6	25.0	4.5	153	116	134
633	Nd	Zr	Yb	22.8	48.0	13.9	6.6	123	101	122
639	Ti	Nb	Nd	22.9	34.1	25.9	8.7	129	112	115
646	Sc	Zr	Cr	23.0	35.5	23.4	10.1	121	124	119
658	Nd	Zr	Y	23.1	48.9	14.0	6.5	137	136	124
659	Sc	Y	V	23.1	28.2	19.8	21.4	103	111	117
679	Nd	Cr	Sc	23.4	43.0	22.2	4.9	100	135	103
766	La	Zr	Y	24.3	41.9	23.7	7.3	155	118	150
777	La	V	Yb	24.5	44.0	22.6	6.8	100	124	103
787	Ti	Cr	Lu	24.6	40.9	25.2	7.7	110	115	104
791	Ce	Cr	V	24.7	37.3	21.9	14.8	126	105	122
806	Nd	Nb	Zr	24.9	37.4	29.7	7.6	131	118	119
817	V	Zr	Cr	25.0	32.7	24.2	18.2	156	149	154
836	La	Cr	V	25.3	41.6	21.4	12.8	125	131	117
861	Ti	Nb	Sm	25.5	35.0	33.0	8.4	123	112	107
869	Ti	Cr	Yb	25.6	42.5	26.0	8.3	120	131	121
894	Ti	Yb	Ce	25.8	47.4	26.2	3.8	133	122	130
899	La	Cr	Sc	25.9	42.1	23.8	11.7	121	122	111
900	V	Nb	Zr	25.9	39.3	33.3	5.0	122	129	120
908	Sm	Zr	Yb	26.0	64.6	5.6	7.9	127	108	126
957	Sc	Y	Cr	26.6	27.7	24.1	28.0	112	116	118
966	Ti	Yb	Nd	26.6	53.2	22.7	4.1	126	141	123
976	Ti	Yb	La	26.8	44.6	30.2	5.6	130	149	126
977	Nd	Cr	V	26.8	41.6	25.4	13.5	113	142	111
984	Sm	Cr	Lu	26.9	56.6	20.2	3.8	113	114	104

Table 3. The 100 Best Ternary Quadratic Discrimination Diagrams

	Elements			Resubstitution Error, %				# IAB	# MORB	# OIB
Rank	1	2	3	Overall	IAB	MORB	OIB	(/256)	(/241)	(/259)
1	Na	Nb	Sr	5.0	8.3	4.0	2.8	157	126	142
2	Al	Nb	Sr	5.7	10.2	4.0	2.8	157	126	142
3	Si	Nb	Sr	5.9	10.1	4.0	3.5	159	126	142
4	Ca	Nb	Sr	6.0	9.6	5.6	2.8	157	126	142
5	Sr	Nb	Y	6.1	7.0	3.9	7.5	157	127	160
6	Eu	Sr	Lu	6.3	9.7	7.7	1.7	124	117	120
7	Ti	Sr	Al	6.7	10.0	8.1	2.1	220	211	194
8	Si	Sr	Ti	6.7	9.5	8.1	2.6	221	211	192
9	Mn	Nb	Sr	6.9	10.2	6.3	4.3	157	126	140
10	Ti	Sr	V	7.0	7.6	8.9	4.5	158	180	156
11	Ti	Sr	Na	7.9	10.9	6.6	6.2	220	213	194
12	Eu	Sr	Yb	7.9	13.0	5.7	5.0	138	157	141
13	Ti	Sr	Lu	8.0	11.5	8.0	4.6	113	113	108
14	Ti	Sr	Sc	8.0	12.6	8.4	3.1	119	155	128
15	Na	Zr	Sr	8.1	14.4	4.8	5.1	208	165	177
16	Ti	Sr	Ca	8.1	11.8	9.5	3.1	220	211	194
17	Ti	Sr	Mn	8.2	10.5	10.3	3.7	219	204	191
18	Eu	Y	Sr	8.4	16.9	5.7	2.5	124	106	121
19	Al	Sr	Eu	8.6	14.7	8.4	2.5	129	154	120
20	K	Nb	V	8.6	9.1	12.4	4.4	121	129	114
21	V	Nb	Rb	8.8	9.0	13.3	4.1	122	113	123
22	Ti	Y	Sr	8.9	11.9	5.2	9.6	202	153	177
23	Na	Sr	Eu	9.0	16.3	5.8	5.0	129	156	120
24	Al	Zr	Sr	9.1	14.9	6.7	5.6	208	163	177
25	V	Nb	Sr	9.2	12.3	12.0	3.2	122	117	124
26	Tb	Sr	Lu	9.2	13.0	9.8	4.8	100	102	105
27	Ti	Nb	Sr	9.2	5.7	15.9	6.0	158	126	149
28	Ti	Sr	Yb	9.2	13.4	8.0	6.3	127	150	127
29	Nd	Y	Sr	9.3	17.4	5.9	4.7	138	135	127
30	Al	Nb	K	10.0	12.7	10.9	6.3	157	138	143
31	K	Nb	Na	10.0	12.7	13.0	4.2	157	138	143
32	Ti	Ba	Al	10.0	10.6	13.2	6.3	217	144	192
33	Ti	V	Sm	10.0	12.5	6.2	11.4	104	162	105
34	Ti	V	Nd	10.1	12.4	9.0	8.8	113	155	113
35	Ti	Ba	Na	10.1	12.4	13.7	4.2	217	146	192
36	Mg	Nb	Sr	10.3	11.4	7.1	12.2	158	126	147
37	Si	Zr	Sr	10.4	17.5	6.7	6.9	211	163	175
38	Nd	Sr	Yb	10.4	19.4	9.7	2.3	134	145	133
39	Ca	Zr	Sr	10.5	20.2	6.1	5.1	208	163	177
40	Yb	Zr	Sr	10.5	18.3	6.5	6.7	126	107	134
41	Sr	Zr	Y	10.5	19.1	4.5	7.9	204	155	203
42	Si	Sr	Eu	10.5	15.6	8.4	7.6	135	154	118
43	Ca	Sr	Nd	10.5	20.9	6.8	4.0	139	177	125
44	Mn	Zr	Sr	10.6	19.3	6.3	6.3	207	160	174
45	Al	Sr	Nd	10.7	22.3	5.6	4.0	139	177	125
46	Ca	Sr	Eu	10.7	17.1	8.4	6.7	129	154	120
47	Ti	V	Sc	10.8	17.1	9.5	5.8	105	148	121
48	Al	Nb	Rb	10.8	11.5	13.9	7.0	156	122	142
49	Ca	Nb	K	10.9	12.1	13.0	7.7	157	138	143
50	Sm	Y	Sr	11.0	23.4	4.4	5.0	128	137	119
51	Na	Nb	Rb	11.0	11.5	16.4	4.9	156	122	142
52	V	Zr	Sr	11.0	17.2	7.5	8.3	157	147	156
53	Si	Sr	Nd	11.1	21.9	7.3	4.0	146	177	124
54	Si	Nb	K	11.1	12.6	13.8	7.0	159	138	143
55	Sc	Zr	Sr	11.2	19.3	6.8	7.4	119	118	122
56	Ti	Cu	Al	11.2	10.7	16.8	6.0	121	107	134
57	Nd	Sr	Lu	11.3	19.8	11.4	2.7	121	105	112
58	Ti	Sc	K	11.4	11.5	15.4	7.1	122	162	126
59	Sm	Sr	Lu	11.4	20.5	7.8	5.9	122	116	119
60	Ti	K	Al	11.4	13.2	13.6	7.4	228	228	203
61	Eu	Sr	V	11.4	18.8	6.9	8.5	101	131	106

Table 3. (continued)

	Elements			Resubstitution Error, %				# IAB	# MORB	# OIB
Rank	1	2	3	Overall	IAB	MORB	OIB	(/256)	(/241)	(/259)
62	Ti	V	Na	11.4	12.6	13.7	7.9	159	197	151
63	Rb	Nb	Y	11.4	11.5	14.6	8.1	156	123	160
64	Si	Ba	Ti	11.4	11.5	16.0	6.8	217	144	190
65	Ti	Sr	Zn	11.5	11.4	11.0	12.0	149	109	142
66	K	Nb	Y	11.5	11.6	16.2	6.8	155	136	147
67	Ti	Ba	Sc	11.7	14.8	15.7	4.7	122	115	129
68	Mn	Nb	K	11.7	12.7	18.1	4.3	157	138	141
69	Zn	Zr	Sr	11.7	17.7	8.3	9.2	147	109	152
70	Ti	Lu	Mn	11.7	20.3	13.9	1.0	118	115	105
71	Tb	Sr	Yb	11.8	16.7	9.6	9.0	102	125	111
72	Ti	V	K	11.8	8.2	15.2	11.9	159	197	151
73	Na	Sr	Nd	11.8	24.5	7.8	3.2	139	179	125
74	Si	Nb	Rb	11.8	11.4	16.4	7.7	158	122	142
75	Na	Sr	Ce	11.9	23.0	7.6	5.0	165	119	140
76	Mn	Sr	Nd	11.9	21.6	10.0	4.1	139	170	123
77	Sr	Nb	Zr	11.9	8.1	21.2	6.4	160	132	157
78	Ti	K	Mn	11.9	11.5	16.3	8.0	227	221	200
79	Ti	Rb	Na	12.0	15.3	14.8	5.8	209	189	189
80	Ti	Y	V	12.0	19.6	12.3	4.1	153	155	147
81	Si	K	Ti	12.0	12.7	14.9	8.5	229	228	201
82	Si	Ni	Ti	12.0	24.2	10.2	1.7	211	205	180
83	P	Y	Sr	12.1	23.3	4.7	8.2	202	149	170
84	Ca	Nb	Rb	12.1	11.5	16.4	8.5	156	122	142
85	Ti	Ba	V	12.1	12.0	16.0	8.4	158	125	155
86	Ti	Rb	Al	12.1	12.4	17.6	6.3	209	187	189
87	Ti	Ba	Mn	12.2	12.0	19.7	4.8	216	137	189
88	K	Yb	Nd	12.2	22.8	11.3	2.5	127	141	121
89	Al	Sr	Ce	12.3	24.2	6.8	5.7	165	117	140
90	K	Lu	Nd	12.3	19.5	13.6	3.8	113	103	104
91	Ti	Ba	Ca	12.3	12.0	18.8	6.3	217	144	192
92	Mn	Sr	Eu	12.3	16.3	8.8	11.9	129	147	118
93	Ti	V	P	12.3	13.2	11.1	12.8	159	190	149
94	Mn	Nb	Rb	12.3	11.5	20.5	5.0	156	122	140
95	Sm	Sr	V	12.4	19.0	7.5	10.7	105	160	103
96	Ca	Sr	Ce	12.4	23.0	8.5	5.7	165	117	140
97	Ti	V	La	12.5	13.6	16.1	7.8	125	143	116
98	Ti	Sr	Cu	12.6	10.8	9.8	17.0	120	102	141
99	Nd	Sr	V	12.6	21.9	10.5	5.4	114	153	111
100	Ti	Rb	V	12.6	10.5	16.8	10.7	153	161	150

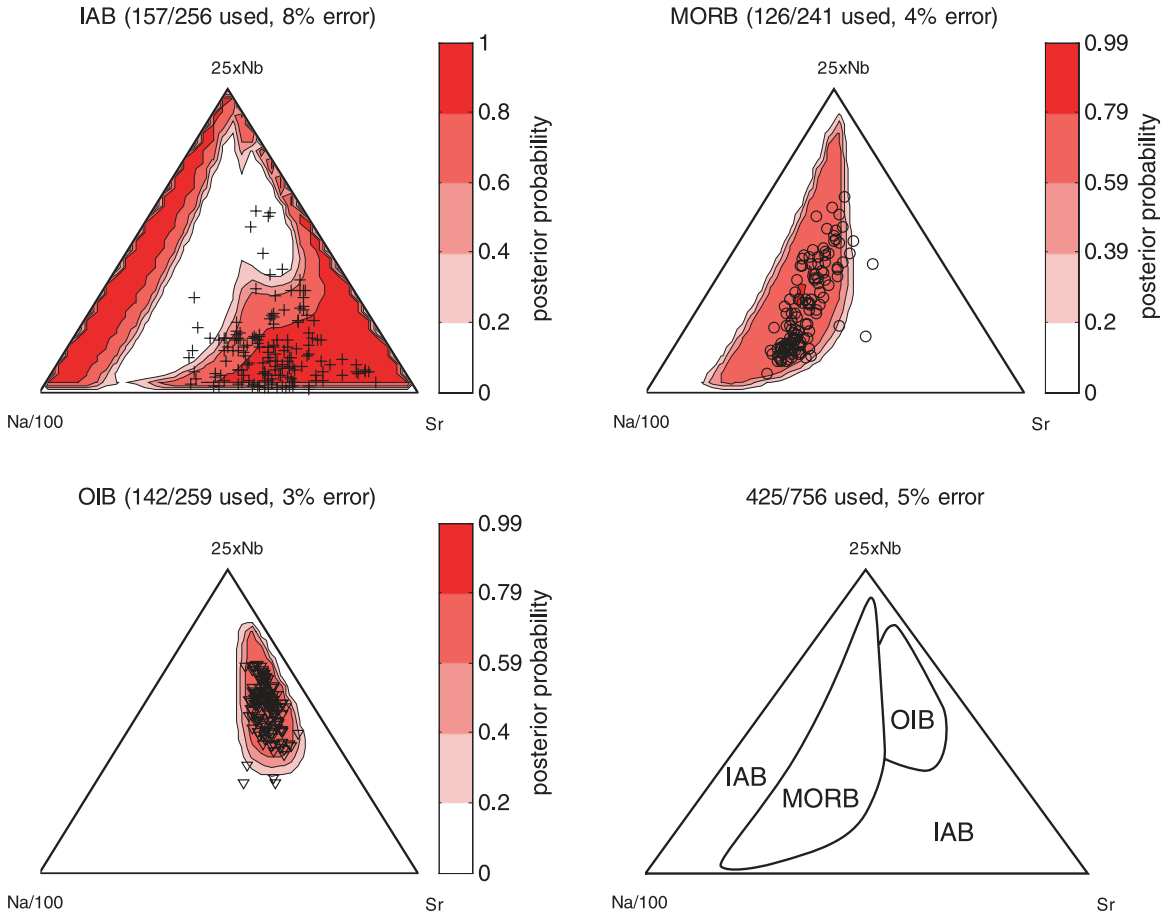

Figure 34. The best performing quadratic discriminant analysis, using Na, Nb, and Sr.

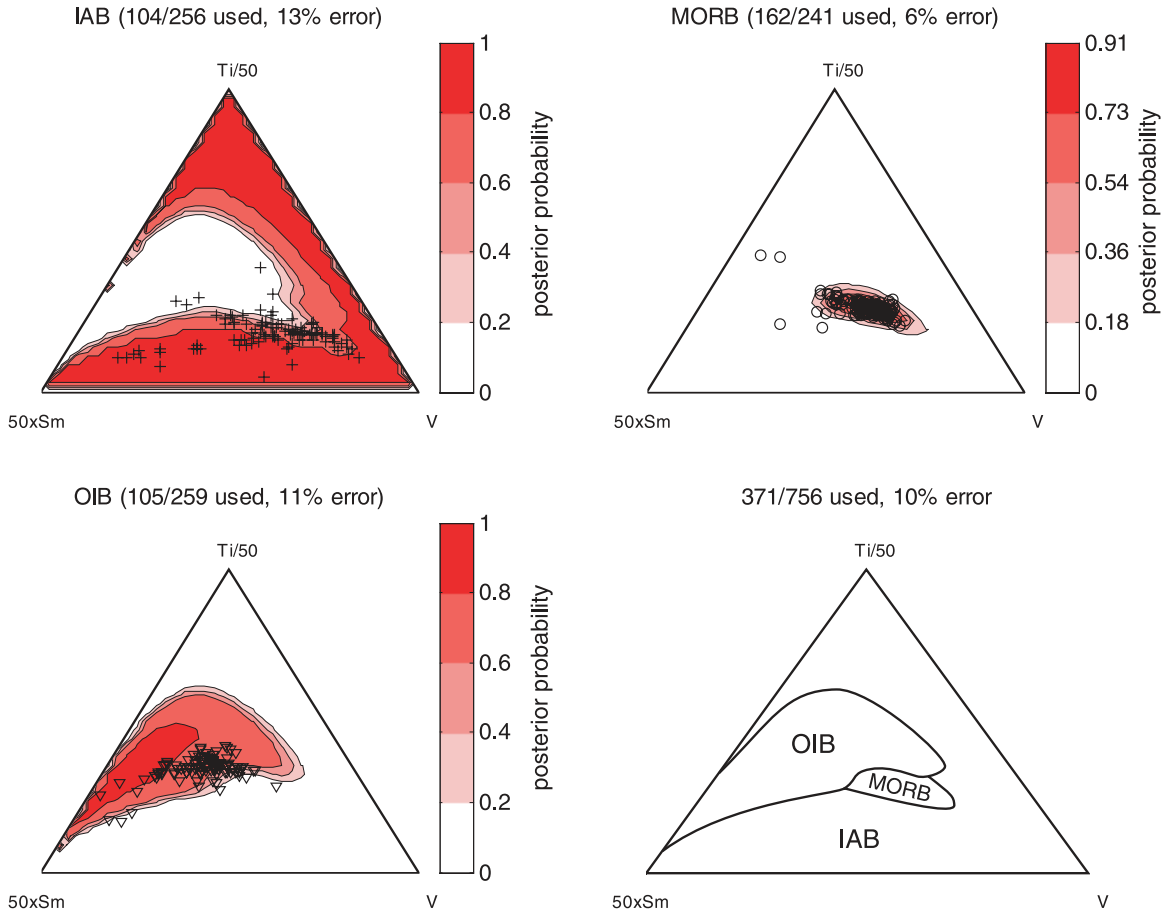

Figure 35. The best performing quadratic discriminant analysis using only incompatible elements (Ti, V, and Sm).

Table 4. The Best Ternary Quadratic Discrimination Diagrams Using Only Incompatible Elements

	Elements			Resubstitution Error, %				# IAB	# MORB	# OIB
Rank	1	2	3	Overall	IAB	MORB	OIB	(/256)	(/241)	(/259)
33	Ti	V	Sm	10.0	12.5	6.2	11.4	104	162	105
34	Ti	V	Nd	10.1	12.4	9.0	8.8	113	155	113
47	Ti	V	Sc	10.8	17.1	9.5	5.8	105	148	121
80	Ti	Y	V	12.0	19.6	12.3	4.1	153	155	147
97	Ti	V	La	12.5	13.6	16.1	7.8	125	143	116
118	Ti	V	Ce	13.1	14.3	14.4	10.7	126	104	122
123	Nd	Nb	Y	13.3	28.7	7.1	4.1	129	113	123
163	Ti	Zr	V	14.4	18.6	11.7	12.8	156	162	148
182	Ti	Nb	V	14.8	23.1	16.3	5.0	121	129	121
202	V	Zr	Y	15.1	23.5	12.9	8.7	153	155	149
232	Ti	Cr	Sc	15.5	31.4	13.6	1.6	121	162	127
252	La	Nb	Zr	15.9	15.7	19.8	12.2	140	101	131
283	Ti	Cr	V	16.3	25.3	12.6	11.0	158	182	154
310	Ti	Nb	Nd	16.8	19.4	17.0	13.9	129	112	115
340	V	Nb	Y	17.3	29.2	17.8	4.8	120	129	124
353	Nd	Zr	V	17.5	31.9	12.8	7.8	113	125	103
365	Nd	Y	V	17.6	37.2	8.8	6.8	113	125	103
405	La	Zr	V	18.2	27.2	17.3	10.0	125	110	110
407	Sm	Nb	Y	18.2	36.1	9.7	8.7	122	113	115
408	Ti	Lu	Sm	18.2	44.8	7.0	2.7	116	114	113
443	Ti	Y	Nd	18.6	37.5	12.7	5.7	136	134	122
448	La	Y	V	18.7	32.3	14.2	9.8	124	106	112
461	Nd	Zr	Y	18.9	40.1	11.8	4.8	137	136	124
485	Ti	Sc	La	19.3	38.1	15.4	4.5	113	117	112
494	Sc	Zr	Y	19.4	36.6	19.1	2.6	112	115	115
506	Nd	Zr	Yb	19.6	43.9	10.9	4.1	123	101	122
536	Ti	Zr	Sc	20.0	39.0	14.2	6.7	118	120	119
548	Nd	Cr	Sc	20.1	43.0	13.3	3.9	100	135	103
568	Nd	Nb	Zr	20.3	30.5	21.2	9.2	131	118	119
571	Ti	Y	Sm	20.4	47.6	7.4	6.1	126	136	114
573	Nd	Nb	Sm	20.4	33.9	15.4	11.9	124	117	109
590	Ti	Y	Sc	20.5	42.0	17.9	1.7	112	112	118
591	Sc	Zr	V	20.5	38.1	13.0	10.4	105	115	115
593	Sc	Zr	Cr	20.6	38.0	16.9	6.7	121	124	119
610	Ti	Yb	Sm	20.8	51.1	7.2	3.9	131	152	127
620	La	Zr	Yb	20.9	38.9	14.3	9.4	126	105	127
633	Ti	Lu	Nd	21.0	44.6	15.5	2.8	112	103	106
656	Ti	Y	La	21.2	44.4	14.7	4.5	153	116	134
670	Ti	Cr	Lu	21.4	50.9	10.4	2.9	110	115	104
690	Ce	Cr	V	21.7	36.5	16.2	12.3	126	105	122
759	Sm	Zr	Y	22.3	50.4	8.0	8.6	127	138	116
773	Ti	Yb	Nd	22.6	52.4	12.1	3.3	126	141	123
791	Sc	Y	V	22.8	31.1	12.6	24.8	103	111	117
792	Ce	V	Yb	22.8	44.0	16.2	8.3	100	105	108
797	Ti	Lu	La	22.9	50.0	16.1	2.7	118	112	113
803	Ti	Zr	Yb	23.0	48.8	15.1	5.1	125	106	117
818	La	Cr	V	23.2	42.4	16.0	11.1	125	131	117
824	Ti	Zr	Y	23.2	51.7	11.6	6.4	201	164	173
829	La	Cr	Sc	23.3	46.3	16.4	7.2	121	122	111
838	Sm	Zr	Yb	23.4	52.0	11.1	7.1	127	108	126
858	V	Zr	Cr	23.6	35.9	16.1	18.8	156	149	154
865	Sm	Cr	Lu	23.7	54.9	12.3	3.8	113	114	104
867	Nd	Cr	V	23.7	41.6	16.9	12.6	113	142	111
930	La	V	Yb	24.3	47.0	16.1	9.7	100	124	103
932	Ti	Cr	Yb	24.3	55.0	13.7	4.1	120	131	121
936	Sc	Y	Cr	24.3	29.5	15.5	28.0	112	116	118
941	Sm	Cr	Yb	24.4	53.7	13.4	6.0	123	134	116
947	La	Zr	Y	24.5	47.1	16.9	9.3	155	118	150
958	Ti	Yb	Ce	24.5	54.1	16.4	3.1	133	122	130

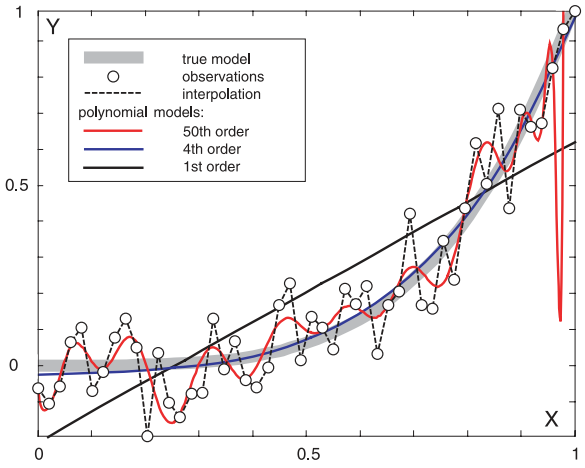

Figure 36. Illustration of the bias-variance tradeoff in a regression context. The thick gray line is the true model ($Y = X^4$). The white circles are 50 samples with random normal errors. The dashed line is the interpolator, which is one of infinitely many functions that go through all the data points and thus have zero bias. The solid black line is a linear regression model, which has a large bias but small variance. In this case, the fourth-order polynomial (blue) is the best predictor of future behavior. Although it has larger bias than the 50th-order polynomial (red) and larger variance than the first-order polynomial (straight black line), it minimizes the mean squared error (MSE = variance + bias^2).

fall within the ternary diagram, as they should. Figure 7 shows an LDA of the synthetic data of Figures 5 and 6, done the "wrong" way (i.e., treating the simplex as a regular data space). As explained in the previous section, such an analysis yields linear decision boundaries. 10% of the training data were misclassified. Figure 8 shows an LDA done the "correct" way (i.e., after mapping the data to log-ratio space). The decision boundaries are still linear, but this time only ~3% of the training data were misclassified. Because log(Y/Z) and log(X/Z) are rather hard quantities to interpret, it is a good idea to map the results back to the ternary diagram using the inverse log-ratio transformation (Figure 9). The transformed decision boundaries are no longer linear, but curved. However, the misclassification rate is still only 3%.

[14] Note that there are two different kinds of constant-sum constraint. The first is a physical one, resulting from the fact that all chemical concentrations add up to 100%. The second is a diagrammatic constraint caused by renormalizing three chosen elements to 100% on a ternary plot. Aitchison's logratio transform adequately deals with both types of constant sum constraint. The first type is discussed in sections 5.1 and 5.3; the second type is discussed in section 5.2.

5. Revisiting a Few Popular Discrimination Diagrams

[15] In this section, a few historically important and popular tectonic discrimination diagrams will be discussed. They are as follows:

[16] • Ti-V [*Shervais*, 1982]

[17] • Ti-Zr [*Pearce and Cann*, 1973]

[18] • Ti-Zr-Y [*Pearce and Cann*, 1973]

[19] • Zr-Y-Nb [*Meschede*, 1986]

[20] • Th-Ta-Hf [*Wood*, 1980]

[21] • SiO_2-Al_2O_3-TiO_2-CaO-MgO-MnO-K_2O-Na_2O [*Pearce*, 1976] (but without FeO)

[22] • Ti, Zr, Y and Sr [*Butler and Woronow*, 1986]

[23] The word "discrimination diagram" is used instead of "discriminant analysis," because most of these diagrams are only loosely based on the principles of discriminant analysis outlined in section 2 and the decision boundaries were drawn by eye. This section will revisit the combinations of elements used in these discrimination diagrams. An extensive data set of 756 samples (Figure 10) was compiled from the PETDB and GEOROC databases [*Lehnert et al.*, 2000]. It contains:

[24] • 256 Island arc basalts (IAB) from the Aeolian, Izu-Bonin, Kermadec, Kurile, Lesser Antilles, Mariana, Scotia and Tonga arcs.

[25] • 241 Mid-ocean ridge (MORB) samples from the East Pacific rise, Mid-Atlantic Ridge, Indian Ocean and Juan de Fuca ridge.

[26] • 259 Ocean island (OIB) samples from St. Helena, the Canary, Cape Verde, Caroline, Crozet, Hawaii-Emperor, Juan Fernandez, Marquesas, Mascarene, Samoan and Society islands.

[27] All the training data had SiO_2 concentrations between 45 and 53%. Duplicate analyses were excluded from the database to avoid potential bias toward overrepresented samples. From this database, two sets of training data were generated:

[28] • 11 major oxides (in weight percent): SiO_2, TiO_2, Al_2O_3, Fe_2O_3, FeO, CaO, MgO, MnO, K_2O, Na_2O and P_2O_5.

Table 5. Misclassification Estimates

	Predicted Affinity (Training)			Predicted Affinity (Test)		
True Affinity	IAB	MORB	OIB	IAB	MORB	OIB
			Linear Ti-V			
IAB	130	24	5	19	8	0
MORB	34	153	10	2	39	12
OIB	0	14	144	0	0	36
			Quadratic Ti-V			
IAB	127	27	5	19	8	0
MORB	26	161	10	4	37	12
OIB	0	14	144	0	0	36
			Linear Ti-Zr			
IAB	176	28	6	36	10	1
MORB	34	125	22	0	7	4
OIB	0	17	170	0	2	29
			Quadratic Ti-Zr			
IAB	167	37	6	32	14	1
MORB	20	148	13	3	5	3
OIB	5	18	164	0	4	27
			Linear Ti-Zr-Y			
IAB	89	101	11	33	9	3
MORB	67	91	6	0	11	0
OIB	6	5	162	3	2	24
			Quadratic Ti-Zr-Y			
IAB	97	93	11	20	22	3
MORB	11	145	8	6	5	0
OIB	8	3	162	3	2	24
			Linear Zr-Y-Nb			
IAB	81	57	19	16	5	2
MORB	73	55	11	2	6	0
OIB	1	6	149	0	4	23
			Quadratic Zr-Y-Nb			
IAB	60	79	18	6	17	0
MORB	12	115	12	0	8	0
OIB	5	5	146	2	2	23
			Linear Th-Ta-Hf			
IAB	78	6	10	12	12	2
MORB	0	37	14	0	0	0
OIB	0	13	69	0	0	10
			Quadratic Th-Ta-Hf			
IAB	81	3	10	14	10	2
MORB	1	38	12	0	0	0
OIB	4	12	66	0	0	10
			Linear Discriminant Function Analysis of SiO_2, Al_2O_3, TiO_2, CaO, MgO, MnO, K_2O, and Na_2O			
IAB	205	15	7	52	7	5
MORB	7	205	9	2	17	4
OIB	2	8	188	1	0	59
			Linear Discriminant Function Analysis of Ti, Zr, Y, and Sr			
IAB	175	13	13	41	1	2
MORB	3	145	3	0	10	0
OIB	5	5	163	0	4	25

Table 5. (continued)

	Predicted Affinity (Training)			Predicted Affinity (Test)		
True Affinity	IAB	MORB	OIB	IAB	MORB	OIB
			Linear Si-Ti-Sr			
IAB	199	15	7	45	9	7
MORB	7	197	7	0	45	1
OIB	0	4	188	0	0	57
			Linear Eu-Lu-Sr			
IAB	111	9	4	31	7	3
MORB	3	110	4	0	35	0
OIB	1	3	116	0	0	27
			Linear Ti-V-Sc			
IAB	89	11	5	19	0	0
MORB	10	133	5	9	28	4
OIB	0	7	114	0	0	12
			Quadratic Na-Nb-Sr			
IAB	144	6	7	21	0	0
MORB	2	121	3	5	7	0
OIB	4	0	138	0	0	28
			Quadratic Ti-V-Sm			
IAB	91	8	5	24	2	0
MORB	5	152	5	1	44	5
OIB	3	9	93	0	0	9

[29] • 45 major, minor and trace elements (in ppm): Si, Ti, Al, Fe(III), Fe(II), Ca, Mg, Mn, K, Na, P, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Sc, V, Cr, Co, Ni, Cu, Zn, Ga, Rb, Sr, Y, Zr, Nb, Cs, Ba, Hf, Ta, Pb, Th and U.

[30] The data are available as auxiliary material[1] Tables S1 and S2. Not all samples were analyzed for all the components. The data set of major oxides is redundant, but a rescaling from % to ppm is avoided by treating it separately. Being admitted to the GEOROC and PETDB databases, it was assumed that the training data are reliable. Each data point in the auxiliary material is associated with a unique ID that allows the user to recover the original publication source. Different normalization procedures were used for different data sets, but this is unlikely to have major consequences for the discriminant analysis. So many data sources are mixed that at most, this mixing of normalization and laboratory procedures would have induced some additional random uncertainty, with only minor effects on the actual decision boundaries. Mixing different data sources and normalization procedures in the training data has the positive side-effect that the user is more or less free to use whichever normalization procedure (s)he wishes.

[31] First, two simple bivariate discrimination diagrams will be discussed: the Ti-V diagram of *Shervais* [1982] and the Ti-Zr diagram of *Pearce and Cann* [1973]. Many of the problems that plague the study of compositional data and were discussed in section 3 are far less serious in the bivariate than the ternary case. Of course, Ti and V, or Ti and Zr are still subject to the (physical) constant-sum constraint, but considering they typically constitute less than a few percent of the total rock composition, a change in one element will have little effect on the other one when the raw measurement units are used on the axes of the bivariate discrimination diagrams. In contrast with this, all popular ternary discrimination diagrams have been rescaled to a (diagrammatic) constant sum of 100%, thus magnifying the effects of closure. For all of the following discriminant analyses, a uniform prior was used. Statistical analysis was done with a combination of Matlab© and R (http://www.r-project.org).

[1]Auxiliary material is available at ftp://agu.org/apend/gc/2005gc001092.

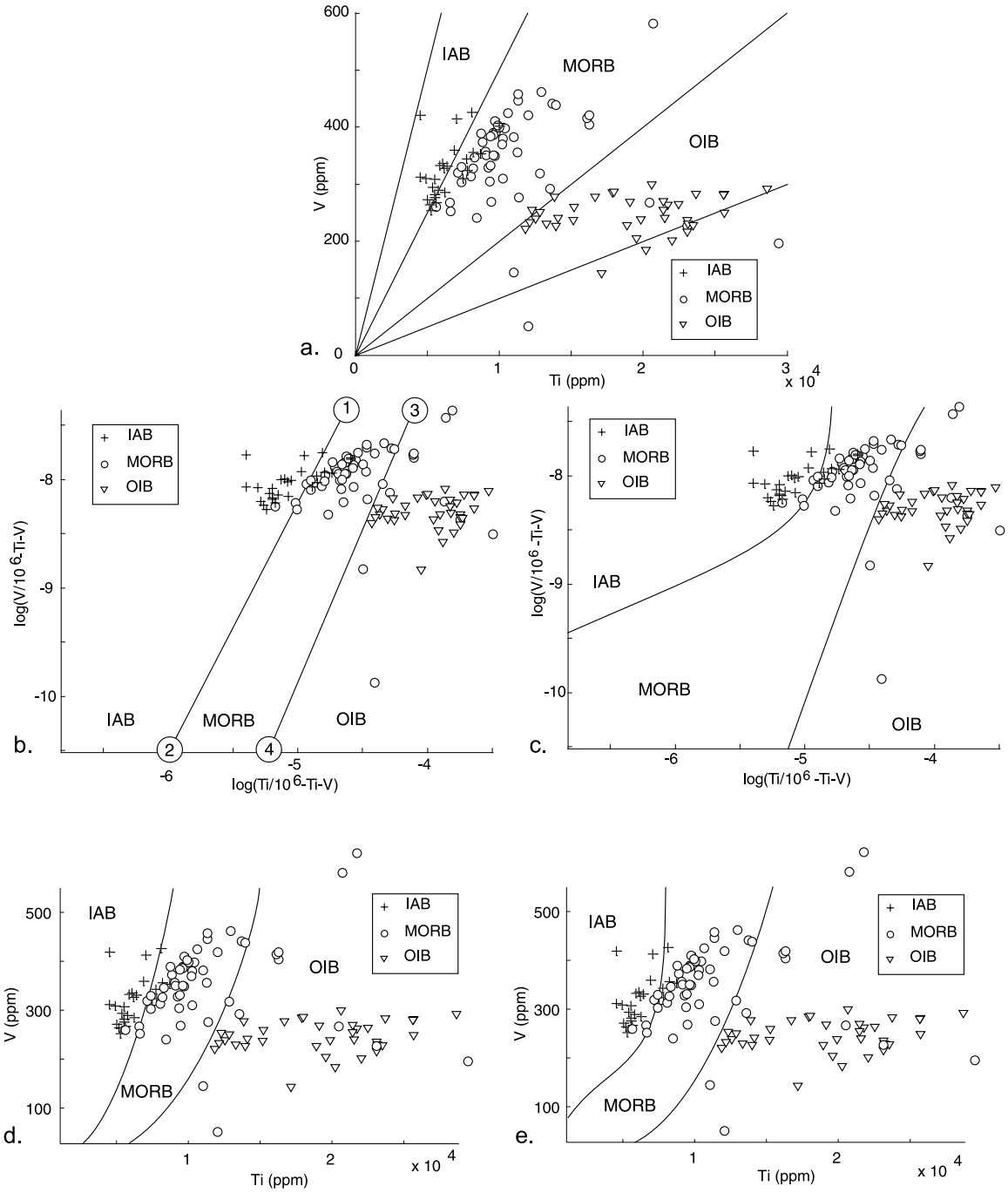

Figure 37. The test data (116/182 used) plotted on various versions of the Ti-V diagram with (a) the original decision boundaries of *Shervais* [1982], drawn by eye; (b) LDA on the logratio plot, with anchor points 1–4 given in Table 6; (c) QDA on the logratio plot; (d) the same LDA as in Figure 37b, but this time mapped back to the "traditional" compositional data space; and (e) the QDA of Figure 37c mapped back to Ti-V space. An error analysis of these and subsequent diagrams is given in Tables 5 and 7.

5.1. Binary Discrimination Diagrams

[32] For the Ti-V system, the data were transformed to the simplex by the log-ratio transformation. Thus two new variables were created: log(Ti/(10^6-Ti-V)) and log(V/(10^6-Ti-V)), where 10^6 is the constant sum of 1 million ppm. The discriminant analysis then proceeds as described in section 2. The results were mapped back to bivar-

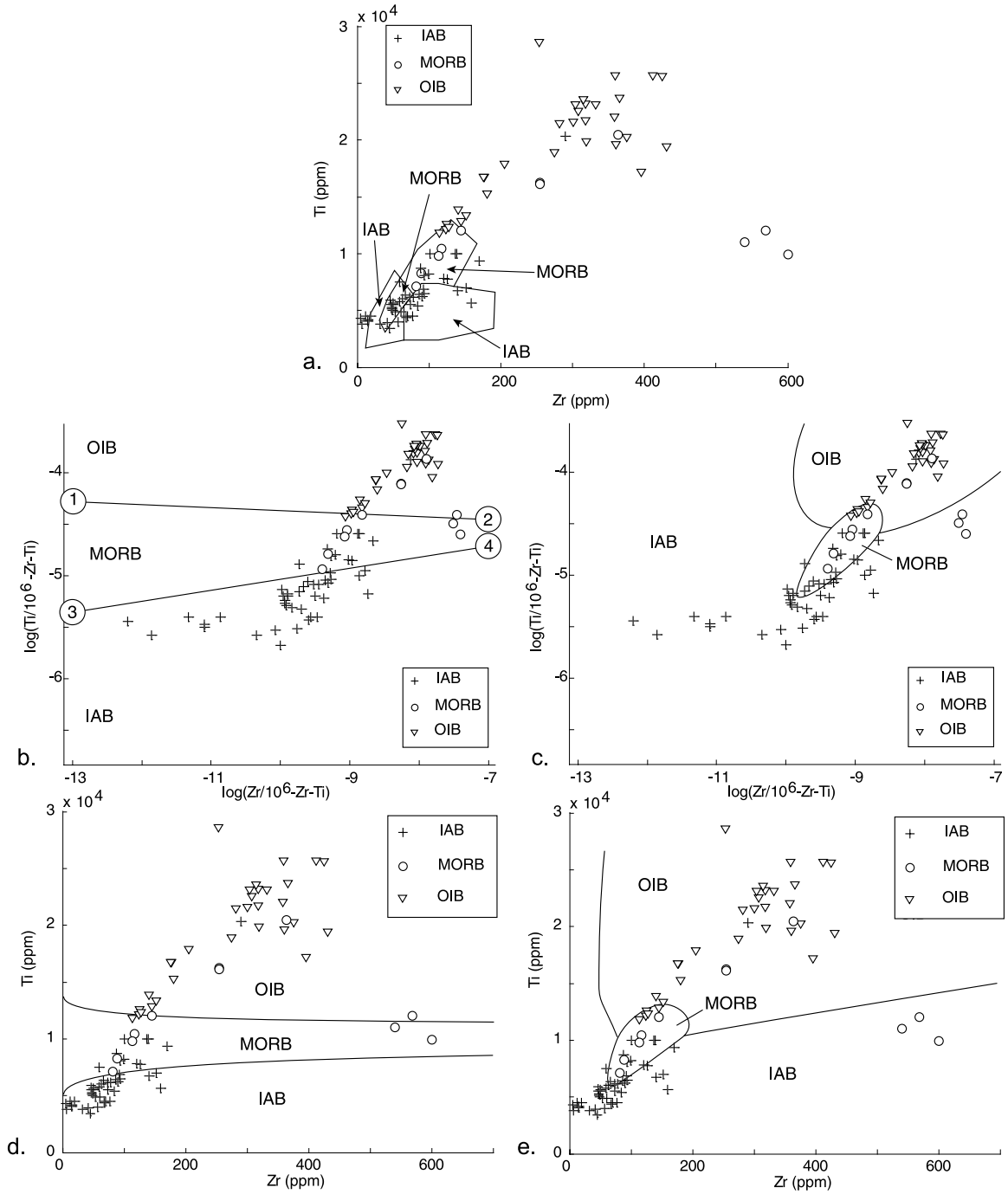

Figure 38. The test data (89/182 used) plotted on the Ti-Zr diagram with (a) the original decision boundaries of *Pearce and Cann* [1973] and (b–e) as in Figure 37.

iate Ti-V space using the inverse log-ratio transformation (equation (6)). Figure 11 shows the results of the LDA of the Ti-V system, whereas Figure 12 shows the QDA results. The decision boundaries look almost identical for both cases. Besides the decision boundaries, Figures 11 and 12 and subsequent figures also show the training data as well as the posterior probabilities. One of the properties of many data mining algorithms, including discriminant analysis, is the "garbage in, garbage out" principle: any rock that was analyzed for the required elements will be classified as either IAB, MORB or OIB, even continental basalts, granites or sandstones! Therefore it is recommen-

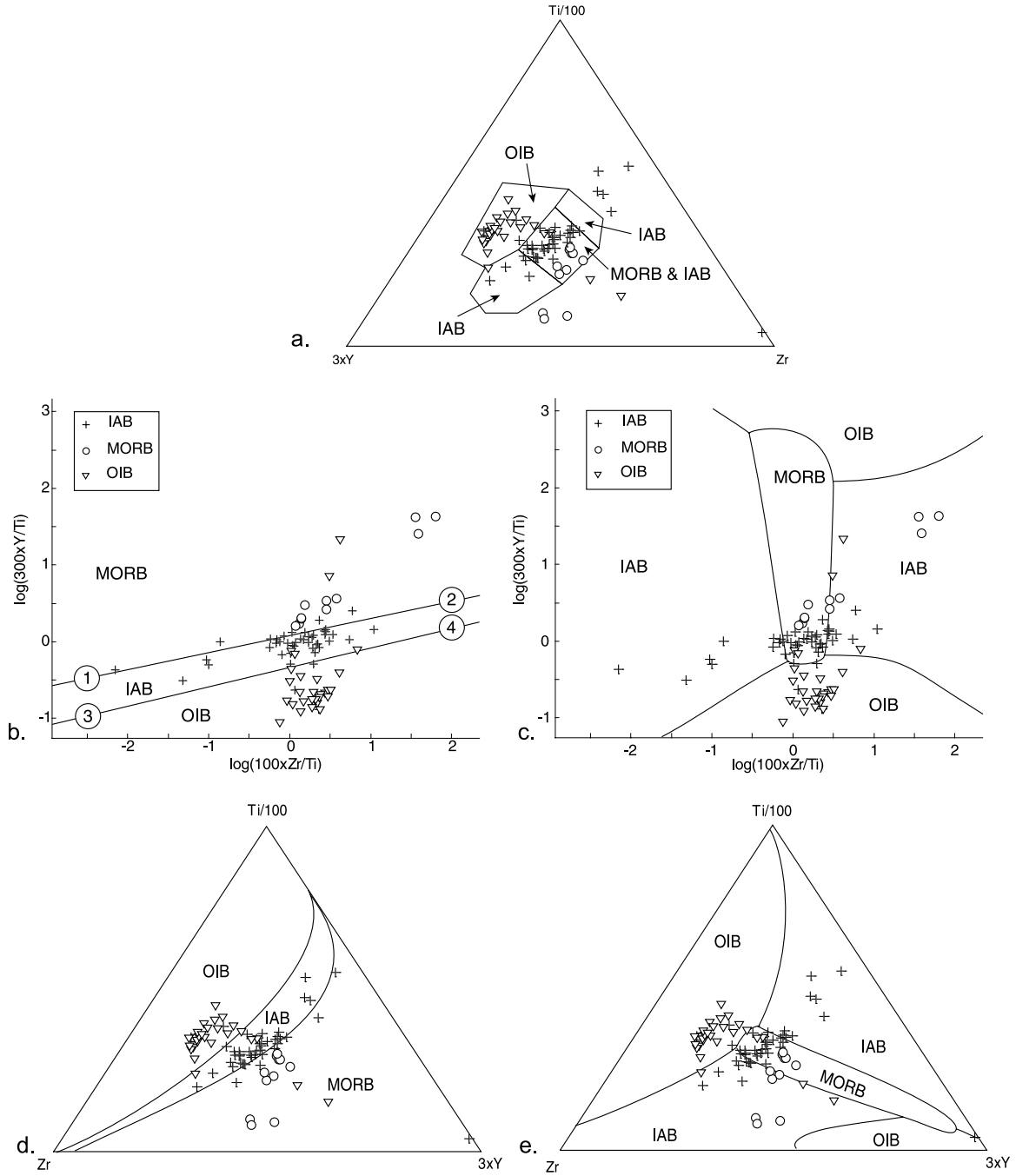

Figure 39. The test data (85/182 used) plotted on the Ti-Zr-Y diagram with (a) the original decision boundaries of *Pearce and Cann* [1973] and (b–e) as in Figure 37.

ded to treat the classification of samples plotting far outside the range of the training data with caution.

[33] In contrast with the Ti-V diagram, the decision boundaries of the Ti-Zr system look quite different between LDA (Figure 13) and QDA (Figure 14). The misclassification risk of the training data (i.e., the resubstitution error) of QDA is always less than that of LDA, because the former uses more parameters than the latter. However, this does not necessarily mean that QDA will perform better on future data sets. This problem will be discussed in section 7. For now, suffice it to say that the resubstitution error can be used to compare two binary or two ternary diagrams with each other,

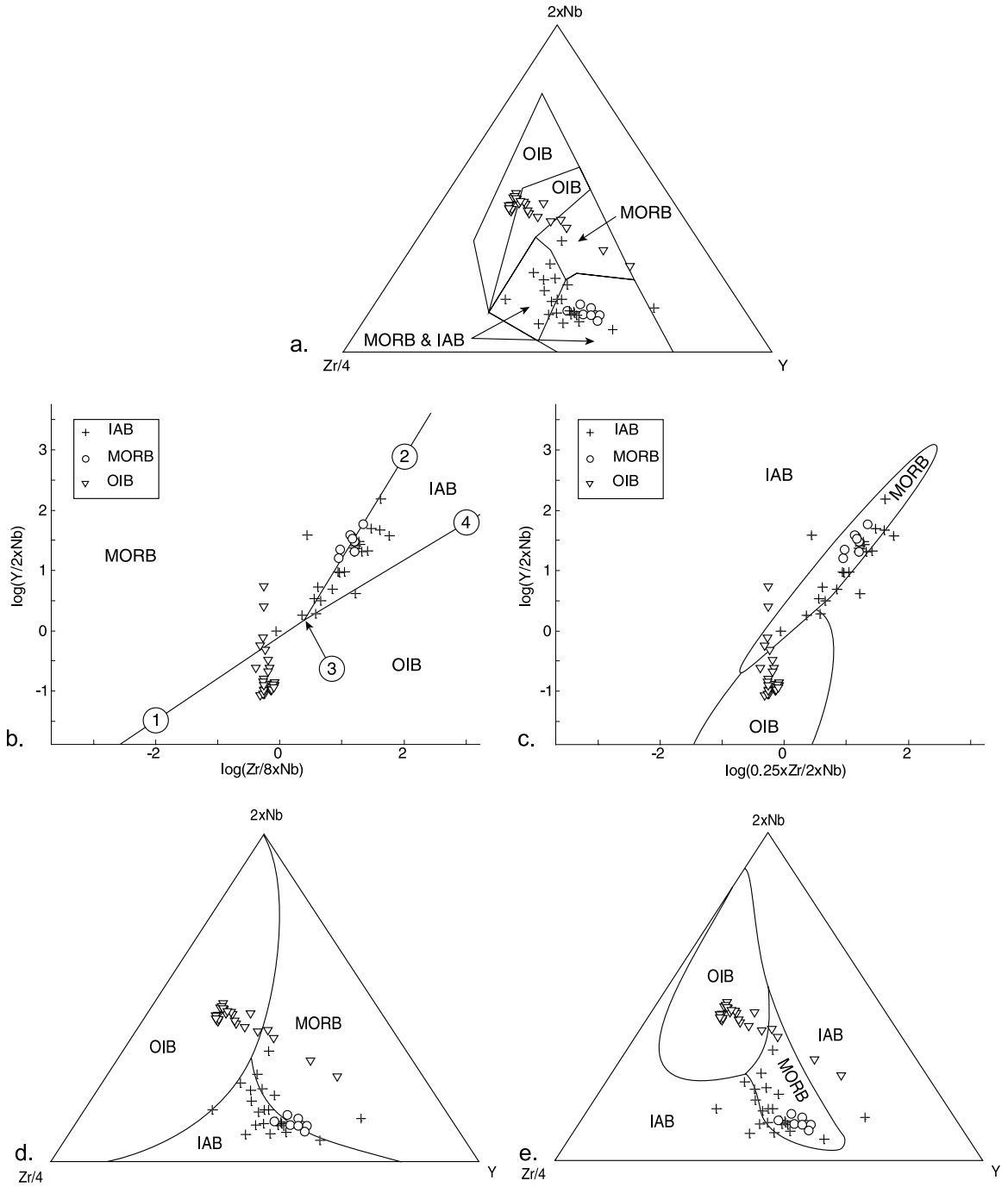

Figure 40. The test data (58/182 used) plotted on the Nb-Zr-Y diagram with (a) the original decision boundaries of *Meschede* [1986] and (b–e) as in Figure 37.

but not to compare the performance of QDA with LDA or of a binary with a ternary diagram.

5.2. Ternary Discrimination Diagrams

[34] The procedure for performing a discriminant analysis for ternary systems is very similar to the binary case. For example, for the Ti-Zr-Y system of *Pearce and Cann* [1973], we first impose the constant sum constraint: x = Y/(Ti + Zr + Y), y = Zr/(Ti + Zr + Y) and z = Ti/(Ti + Zr + Y). The log-ratio transformed variables are V = log(x/z) and W = log(y/z). Note that this transformation

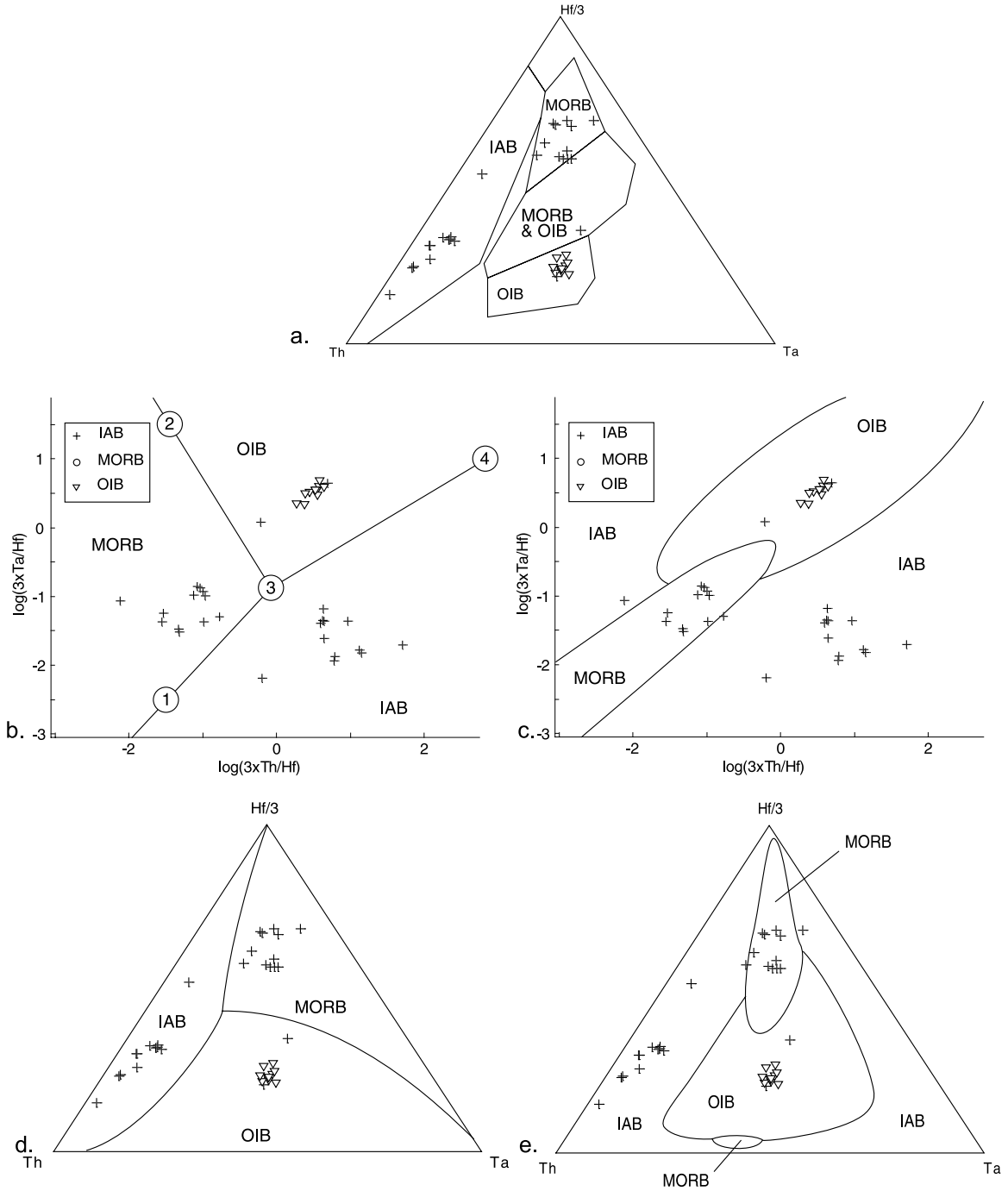

Figure 41. The test data (36/182 used, but no MORBs!) plotted on the Th-Ta-Hf diagram with (a) the original decision boundaries of *Wood* [1980] and (b–e) as in Figure 37.

only takes care of the diagrammatic constraint x + y + z = 1. Strictly speaking, it does not account for the physical constraint Ti + Zr + Y + (all other elements) = 100%. However, Ti + Zr + Y only amount to at most a few percent of typical basalt compositions, thereby greatly reducing the impact of this second type of constant sum. It would be possible to correct for the physical constraint, for example by performing a discriminant analysis on the following three variables: $\log(Ti/(10^6\text{-Ti-Zr-Y}))$, $\log(Zr/(10^6\text{-Ti-Zr-Y}))$, and $\log(Y/(10^6\text{-Ti-Zr-Y}))$. However, the results of such an analysis can no longer be plotted on a ternary diagram. In practice, neglecting the physical constant sum

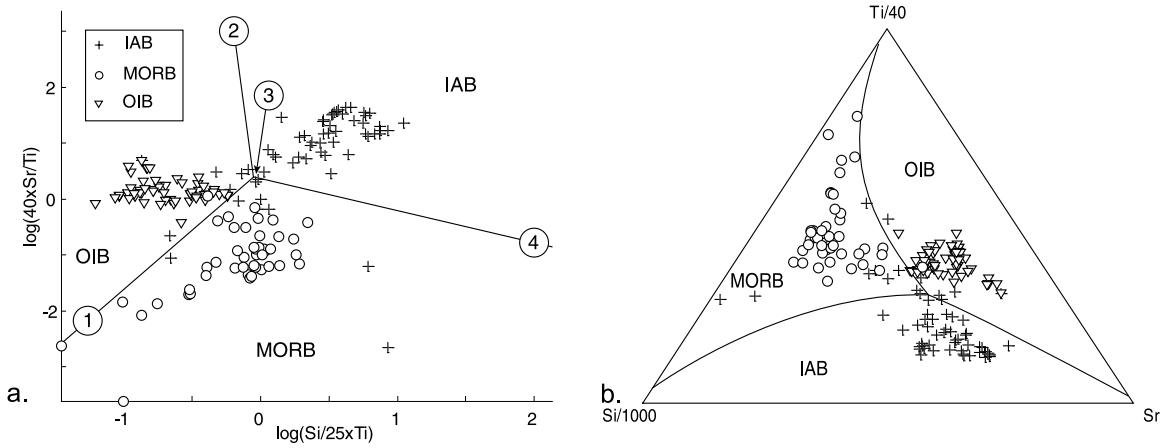

Figure 42. The test data (164/182 used) plotted on the Si-Ti-Sr LDA diagram with (a) the decision boundaries and anchor points (see Table 6) in log-ratio space and (b) the decision boundaries mapped back to the simplex.

constraint does not severely affect the performance of the classification in this case.

[35] Figures 15 and 16 show the results of both LDA and QDA transformed back to the Ti-Zr-Y ternary diagram. The raw variables of many discrimination diagrams are multiplied by constants to improve the spread of the data. This is equivalent to adding constants to the log-ratio transformed variables. Either transformation does not affect the discriminant analysis. As noted by *Pearce and Cann* [1973], the Ti-Zr-Y diagram is quite good at identifying OIBs, but cannot distinguish MORBs from IABs. The training data of the latter substantially overlap and their resubstitution errors are quite high. The posterior probabilities of the training data are low (<0.5 in Figure 16).

[36] This is also the case for the Nb-Zr-Y system of *Meschede* [1986] (Figures 17 and 18). The high misclassification rate of both the Ti-Zr-Y and Nb-Zr-Y diagrams is largely caused by the large spread of IAB compositions, which is likely caused by the complexity of magma generation underneath island arcs, where mixing of multiple melt sources often occurs. The Th-Ta-Hf system of *Wood* [1980], however, achieves a much better separation between the three tectonic affinities (Figures 19 and 20). The decision boundaries of the QDA (Figure 20) are much more complicated than those of the LDA (Figure 19), without substantially improving the overall misclassification risk. Therefore adding the extra parameters (covariances) was probably not worthwhile (see section 7).

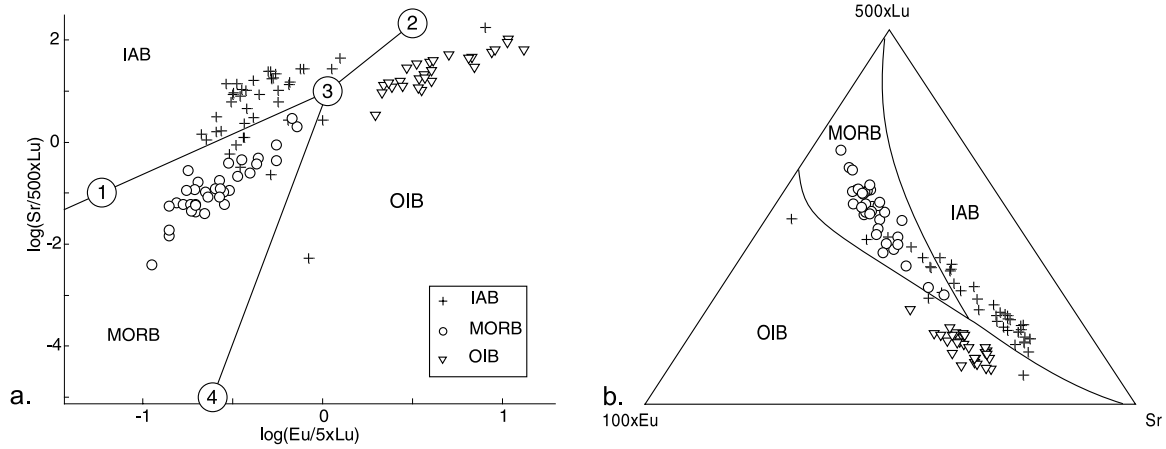

Figure 43. The test data (103/182 used) plotted on the Eu-Lu-Sr LDA diagram: (a and b) as in Figure 42.

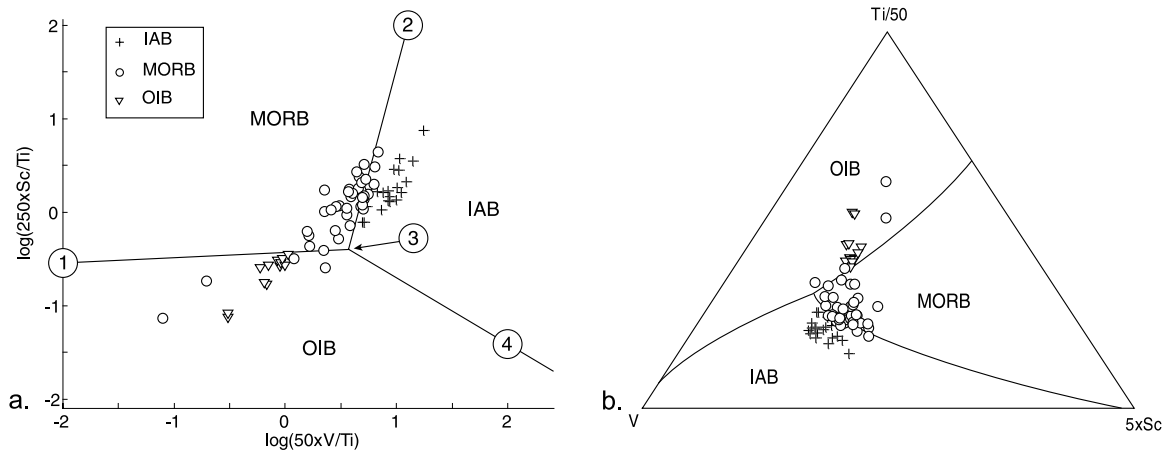

Figure 44. The test data (72/182 used) plotted on the Ti-V-Sc LDA diagram: (a and b) as in Figure 42.

5.3. Multielement Discriminant Function Analysis

[37] As illustrated by Figure 2, LDA offers the possibility of projecting a data set onto a subspace of lower dimensionality. As explained in section 2 this procedure is related to, but quite different from PCA. Therefore it is somewhat puzzling why *Butler and Woronow* [1986] performed a PCA on a data set of Zr, Ti, Y and Sr analyses of oceanic basalts. These authors were the first to note the significance of the constant sum constraint to the problem of tectonic discrimination, but they stopped short of doing a full discriminant analysis. Figure 21 does exactly that. The two linear discriminant functions (ld1 and ld2) are

$$\begin{aligned} \mathrm{ld1} &= -0.016\log(\mathrm{Zr/Ti}) - 2.961\log(\mathrm{Y/Ti}) + 1.500\log(\mathrm{Sr/Ti}) \\ \mathrm{ld2} &= -1.474\log(\mathrm{Zr/Ti}) + 2.143\log(\mathrm{Y/Ti}) + 1.840\log(\mathrm{Sr/Ti}) \end{aligned} \quad (7)$$

Note that the training data cluster quite well, that the clusters are of approximately equal size, and that they are well separated, resulting in a misclassification rate of only 8%.

[38] *Butler and Woronow* [1986] were the first ones to note the potential importance of data-closure in the context of tectonic discrimination of oceanic basalts. However, as said before, they did not use the log-ratio transformation to improve discriminant analysis, but performed a PCA instead, the implications of which are unclear. On the other hand, *Pearce* [1976] did perform a traditional multielement discriminant analysis, but since his paper predated the work of *Aitchison* [1982, 1986], he was unaware of the effects of closure. Figure 22 shows the results of a reanalysis of the major element abundances (except FeO) used by *Pearce* [1976]. The two linear discriminant functions are

$$\begin{aligned} \mathrm{ld1} &= 0.555\log(\mathrm{TiO_2/SiO_2}) + 3.822\log(\mathrm{Al_2O_3/SiO_2}) \\ &\quad + 0.522\log(\mathrm{CaO/SiO_2}) + 1.293\log(\mathrm{MgO/SiO_2}) \\ &\quad - 0.531\log(\mathrm{MnO/SiO_2}) - 0.145\log(\mathrm{K_2O/SiO_2}) \\ &\quad - 0.399\log(\mathrm{Na_2O/SiO_2}) \\ \mathrm{ld2} &= 3.796\log(\mathrm{TiO_2/SiO_2}) + 0.008\log(\mathrm{Al_2O_3/SiO_2}) \\ &\quad - 2.868\log(\mathrm{CaO/SiO_2}) + 0.313\log(\mathrm{MgO/SiO_2}) \\ &\quad + 0.650\log(\mathrm{MnO/SiO_2}) + 1.421\log(\mathrm{K_2O/SiO_2}) \\ &\quad - 3.017\log(\mathrm{Na_2O/SiO_2}) \end{aligned} \quad (8)$$

This discriminant analysis performs about as well as the Ti-Zr-Y-Sr diagram of Figure 21, although it uses many more elements. The benefits of multielement LDA are clearly a decrease in misclassification rate. This comes at the expense of interpretability, because the linear discriminant functions (ld1 and ld2) have no easily interpretable meaning, in contrast with their binary and ternary counterparts.

6. An Exhaustive Exploration of Binary and Ternary Discriminant Analyses

[39] Some of the popular discrimination diagrams discussed in section 5 use a choice of elements that is based on petrological reasons [e.g., *Shervais*, 1982]. However, more often the reasons are entirely statistical, i.e., those features are used that result in a "good" classification. If a database of N elements is used, there are $\binom{N}{2} = \mathrm{N(N-1)/2}$ possible

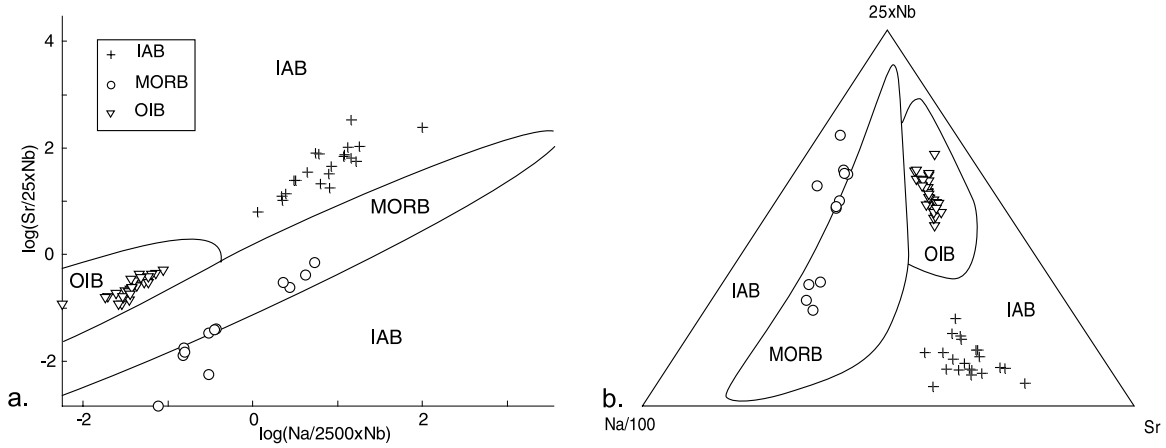

Figure 45. The test data (61/182 used) plotted on the Na-Nb-Sr QDA diagram with (a) the decision boundaries in log-ratio space and (b) mapped back to Δ_2.

binary diagrams and $\binom{N}{3} = N(N-1)(N-2)/6$ possible ternary diagrams. For the database of 11 major oxides, this corresponds to 55 binary and 165 ternary diagrams, whereas the database of 45 elements yields 990 binary and 14,190 ternary diagrams. To efficiently summarize the results of these thousands of discrimination diagrams, a matrix visualization was used.

6.1. Binary Discrimination Diagrams

[40] Figure 23 shows an example of such a visualization for all bivariate LDAs using the major oxides. Of the 756 training data, not all had been analyzed for all major elements. The upper right triangular part of the matrices in this figure show the number of analyses for which both elements were measured. Using the same color-code but a different scale, the lower left triangular parts of the matrices show the resubstitution errors of the 55 possible bivariate LDAs. For example, the lower left triangular matrices of Figure 23 show that only 13.5% of IABs, 15.2% of MORBs and 7.4% of OIBs were misclassified by an LDA using TiO_2 and K_2O. The overall resubstitution error is 12%. The upper right triangular parts of the same figure show that 229 out of 256 IABs, 230 out of 241 MORBs and 203 out of 259 OIBs were used for the construction of the LDA, accounting for a total of 662 out of 756 training data. Figure 24 shows the same thing for QDA.

[41] Figure 25 visualizes the results of all possible bivariate LDAs for the complete data set of 45 elements. On the whole, Ti jumps out as the apparently best overall discriminator. One might think that the Tm-Sc diagram performs very well, considering that the overall error (shown in the upper right triangle of the lower right matrix of Figure 25) is only 7.7%. 12% of the IABs, 8.8% of the OIBs and only 2.4% of the MORBs in the training data were misclassified. However, the upper right triangular matrices of the same figure show that only 101 of 756 training data were used for the classification. Only 25/256 of the IABs, 42/241 of the MORBs and 34/259 of the OIBs were analyzed for both Tm and Sc, thereby greatly reducing the reliability of the classification. Figure 26 shows the results of all possible bivariate QDAs for the database of 45 elements. The strikingly different colors of the lower triangular matrices on this figure illustrate the difficulties in classifying IABs. Both MORBs and IABs are relatively easy to separate, but the geochemical variability of IABs is much larger, for reasons discussed before.

6.2. Ternary Discrimination Diagrams

[42] As calculated in the previous section, there are 990 ways to choose three out of 11 major oxides, and 14,190 ways to choose three out of 45 major, minor and trace elements. Although all these possibilities were explored in the framework of this research, it is not practical to visually show all the results in this paper, even using the highly compact matrix visualization. Therefore only an (important) subset is shown of all ternary diagrams using Ti. As discussed before, many of the most effective bivariate discriminant analyses use Ti. In addition to being an excellent discriminator, Ti is also highly immobile, in contrast with for example Sr,

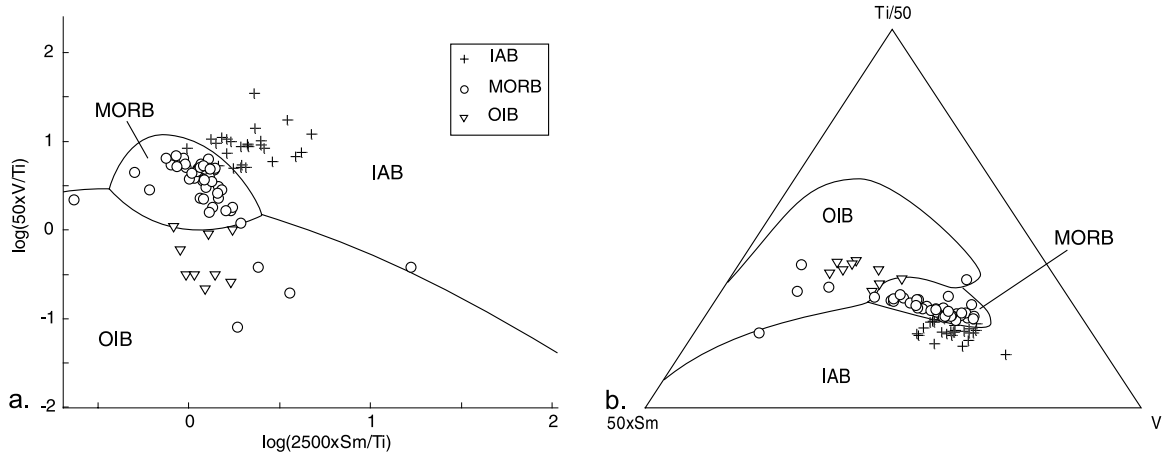

Figure 46. The test data (85/182 used) plotted on the Ti-V-Sm QDA diagram: (a and b) as in Figure 45.

which is another powerful discriminator. For these reasons, only the results of ternary LDAs and QDAs using Ti are shown in Figures 27, 28, 29, and 30.

[43] The resubstitution errors of all 14,190 ternary LDAs (i.e., not only those using Ti) were ranked to find the best combinations of elements. Table 1 shows the 100 best LDAs. Only those diagrams for which at least 100 IABs, 100 MORBs and 100 OIBs of the training data had been analyzed for all three elements were used. 2,333 out of 14,190 possible combinations fulfilled this requirement. The best ternary LDA uses the Si-Ti-Sr system. It has an overall resubstitution error of 6.2%, (2.7% for IABs, 2.8% for MORBs and 2.7% for OIBs), using nearly all the training data (221/256 IABs, 211/241 MORBs and 192/259 OIBs). Figure 31 shows the Si-Ti-Sr LDA in detail. Another powerful ternary diagram using minor and trace elements is the Eu-Lu-Sr system, which ranks third among all the ternary LDAs of Table 1. This diagram is shown in Figure 32. Many if not most of the best performing ternary LDAs use Sr as one of the elements. However, as discussed before, Sr is quite mobile during processes of alteration and metamorphism, potentially affecting the reliability of the discrimination diagrams using it. The Ti-V-Sc diagram, ranking 28th in Table 1, suffers much less from this problem and still has an overall misclassification rate of only 10.4% while using 374 out of 756 training data. Figure 33 shows the Ti-V-Sc diagram in detail. Table 2 lists the best performing (lowest resubstitution error) ternary LDAs, using the following 25 incompatible elements: Ti, La, Ce, Pr, Nd, Sm, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Sc, V, Cr, Y, Zr, Nb, Hf, Ta, Pb, Th, and U.

[44] Table 3 shows the 100 best performing ternary QDAs. The Na-Nb-Sr system performs the best, with an overall resubstitution error of only 5%. As shown in Figure 34, this diagram misclassifies only 22 out of 425 training samples. However, Na is a very mobile element and not much faith can be had in a classification that uses it for basalt samples that are not perfectly fresh. The Ti-V-Sm diagram (Figure 35) is the best performing QDA using only relatively immobile elements. It is ranked 33rd in Table 3. Notice that both for LDA and QDA, the best performing ternary discrimination diagrams using immobile elements contain both Ti and V, apparently confirming the effectiveness of the approach used by *Shervais* [1982]. The latter author selected Ti and V for mostly petrological reasons, while the present paper arrived at the same elements using an entirely statistical method. The compatibility of both approaches lends more credibility to the results. Table 4 lists the best performing QDAs using ternary combinations of the 25 incompatible elements listed in the previous paragraph for which at least 100 training samples of each tectonic affinity were represented.

7. Testing the Results

[45] Some of the discrimination diagrams of the previous section were extremely good at classifying the training data. However, as briefly mentioned in section 5, the resubstitution error is not the best way to assess performance on future data. Furthermore, QDA nearly always performed better than LDA, because the former involves more parameters than the latter. As the number of parameters in a model increases, its ability to

Table 6. Anchor Points for Selected Linear Discriminant Analyses

Node	1	2	3	4
ld1 (equation (7))	−12	−12.23	−18	−8
ld2 (equation (7))	4	−1.37	−6.6	−6.45
ld1 (equation (8))	5.02	12.17	15.9	11.85
ld2 (equation (8))	−6.28	−12.23	−10.93	−16
log(Ti/(10^6-Ti-V))	−4.65	−6	−4.11	−5.22
log(V/(10^6-Ti-V))	7.36	10.5	7.36	10.5
log(Zr/(10^6-Ti-Zr))	−13	−7	−13	−7
log(Ti/(10^6-Ti-Zr))	−4.28	−4.45	−5.36	−4.72
log(100xZr/Ti)	−2.5	2	−2.5	2
log(300xY/Ti)	−0.48	0.53	−0.97	0.17
log(Zr/(8xNb))	−2	2	0.41	3
log(Y/(2xNb))	−1.49	2.92	0.19	1.81
log(3xTh/Hf)	−1.49	1.43	−0.07	2.81
log(3xTa/Hf)	−2.48	−1.5	−0.86	1
log(Si/(25xTi))	−1.26	−0.2	−0.05	2
log(40xSr/Ti)	−2.15	2.98	0.39	−0.77
log(Eu/(5xLu))	−1.23	0.5	0.03	−0.61
log(Sr/(500xLu))	−1	2.33	1	−5
log(50xV/Ti)	−2	1.1	0.57	2
log(250xSc/Ti)	−0.54	2	−0.39	−1.41

resolve even the smallest subtleties in the training data improves. In a regression context, this would correspond to adding terms to a polynomial interpolator (Figure 36). For a very large number of parameters (equaling or exceeding the number of data points), the curve will eventually pass through all the points and the "error" (e.g., squared distance) will become zero. In other words, the high-order polynomial model has zero bias. However, unbiased models rarely are the best predictive models, because they suffer from high variance. High-order polynomial models built on different sets of training data are likely to look significantly different because of irreproducible random variations in the sampling or measuring process. On the other hand, a one-parameter linear model will have low variance, but can be very biased (e.g., when the true model is really polynomial). This phenomenon is called the bias-variance tradeoff, and exists for all data mining methods.

[46] By assuming equal covariance between the different classes of the training data, LDA is a very crude approximation of the data space. Therefore it is likely to be quite biased in many cases. However, because of the bias-variance tradeoff, the variance of the LDAs described in previous sections is low. Therefore the resubstitution error might actually be a decent estimator of future performance. However, things are different for QDA because it estimates the covariance of each of the classes from the training data, thereby dramatically increasing the number of parameters in the model. Although this reduces the bias (i.e., a QDA describes the training data better than an LDA), it causes an increased variance. For example, some of the intricate structure of Figures 16 or 20 might not be very stable. Therefore the resubstitution error is not a good predictor of future performance. It must also not be used for comparing the performances of bivariate and ternary discrimination diagrams.

[47] The easiest way to obtain a more objective estimate of future performance is to use a second database of test data, which had not been used for the construction of the discrimination diagrams. Implementing this idea, a database of 182 test data was compiled from three locations:

[48] • 67 IABs from the Aleutian arc.

[49] • 55 MORBs from the Galapagos ridge.

[50] • 60 OIBs from the Pitcairn islands.

[51] All previously discussed discrimination diagrams are represented in the error analysis of Table 5. The left part of the table shows the resubstitution errors, while the right side shows the performance on the test-data. Figures 37–46 show the test data plotted on the binary and ternary discrimination diagrams. The new decision boundaries are shown in both log-ratio space and conventional compositional data space. As explained in section 2, the decision boundaries are linear for LDA in log-ratio space. To allow an easy reproduction of these decision boundaries, four "anchor points" are provided for each LDA in Figure 21, 22, 37–46 and Table 6. Figures 37–41 and Table 7 allow a direct comparison of the decision boundaries of *Shervais* [1982], *Pearce and Cann* [1973]), *Meschede* [1986], and *Wood* [1980] with the new decision boundaries constructed using LDA and QDA. Although it is hard to make a definite comparison due to the relatively small size of the effectively used test data set, the new decision boundaries seem to always perform at least as well as the old ones. Because the test data set is much smaller than the training data set, it is more likely affected by the missing-data problem. For example, the test data contained no MORBs that had been

Table 7. Comparison Between Old and New Decision Boundaries Using the Test Data

	Ti-Zr								
	Pearce and Cann [1973]			LDA			QDA		
True Affinity	IAB	MORB	Out of Bounds	IAB	MORB		IAB	MORB	
IAB	18	22	2	36	11		32	15	
MORB	0	5	6	0	11		3	8	
	Ti-V								
	Shervais [1982]				LDA			QDA	
True Affinity	IAB	MORB	OIB	Out of Bounds	IAB	MORB	OIB	IAB	MORB	OIB
IAB	17	10	0	0	19	8	0	19	8	0
MORB	0	47	2	3	2	39	12	4	37	12
OIB	0	2	28	6	0	0	36	0	0	36
	Ti-Zr-Y								
	Pearce and Cann [1973]			LDA			QDA		
True Affinity	IAB and MORB		OIB	Out of Bounds	IAB and MORB		OIB	IAB and MORB		OIB
IAB	36		3	6	42		3	42		3
MORB	8		0	3	11		0	11		0
OIB	3		24	2	5		24	5		24
	Zr-Y-Nb								
	Meschede [1986]			LDA			QDA		
True Affinity	IAB and MORB		OIB	Out of Bounds	IAB and MORB		OIB	IAB and MORB		OIB
IAB	22		0	1	21		2	23		0
MORB	8		0	0	8		0	8		0
OIB	3		23	1	4		23	4		23
	Th-Ta-Hf								
	Wood [1980]			LDA			QDA		
True Affinity	IAB		MORB and OIB	Out of Bounds	IAB		MORB and OIB	IAB		MORB and OIB
IAB	12		14	0	12		14	14		12
MORB	0		0	0	0		0	0		0
OIB	0		10	0	0		10	0		10

simultaneously analyzed for Th, Ta and Hf. For all the discrimination diagrams of Table 5, QDA performs better than LDA on the training data. On the other hand, LDA often performs better than QDA on the test data because of its lower variance. For example, LDA misclassified 17 out of 85 test samples using Ti, Zr and Y, whereas QDA misclassified 38 using the same three elements (Table 5). However, in most cases the difference is not so dramatic.

8. Conclusions

[52] This paper revisited the observation by *Butler and Woronow* [1986] that traditional statistical analyses of geochemical data is flawed because it ignores the effects of data-closure. Since the work of *Aitchison* [1982, 1986], it is possible to account and correct for the constant-sum constraint by transforming the data to log-ratio space. *Butler and Woronow* [1986] then went on to do a principal component analysis. The present paper instead uses the log-ratio method for the related, albeit different technique of discriminant analysis.

[53] First, a number of popular discrimination diagrams were revisited. Many of these historically important diagrams were not derived from a real discriminant analysis sensu *Fisher* [1936], but were instead obtained by drawing decision boundaries by eye. A positive side-effect of this is that the resulting diagrams are much less affected by the constant-sum constraint discussed before. A negative consequence remains, however, that all statistical rigor is lost. Nevertheless, it is not the intention of this paper to discredit the discrimination diagrams of *Pearce and Cann* [1973], *Wood* [1980], *Shervais* [1982], *Meschede* [1986], and others. Rather, the paper merely explains how to perform discriminant analysis of geochemical data in a statistically more rigorous manner.

[54] After revisiting these historically important discrimination diagrams, an exhaustive exploration was done of all possible linear and quadratic discriminant analyses using a data set of 756 IABs, MORBs and OIBs. The best overall performance was given by the Si-Ti-Sr (LDA) and Na-Nb-Sr (QDA) systems. The best LDA and QDA using only immobile elements are the Ti-V-Sc and Ti-V-Sm systems, respectively. One of the features of the old discrimination diagrams was a field of "not classifiable" compositions. If an unknown sample plotted outside the predefined fields tectonic affinity fields, it would be labeled as "other." The revisited discriminant analyses discussed above do not have this feature. On the one hand, it might be considered a positive thing that the method no longer "breaks down" when encountering "difficult" samples. On the other hand, one might wonder what would happen if we were to plot a rock of very different affinity on the discrimination diagrams. To mitigate this "garbage in, garbage out" effect, we might want to opt for a hybrid solution, and only accept results for data that plot inside the old (hand-drawn) affinity fields, or within the clouds of training data shown on all discrimination diagrams in this paper (Figures 11–22 and 31–35).

[55] Historically, discrimination diagrams and discriminant analysis have been the method of choice for geochemists to statistically classify rocks of different environments. However, discriminant analysis is not the only "data mining" method that can be used for this purpose. For examples, *Vermeesch* [2006] introduces classification trees as a potentially very useful tool for tectonic classification. Some of the advantages of classification trees over discriminant analysis are that the former (1) do not make any distributional assumptions, (2) can handle an unlimited number of geochemical species, isotopic ratios or other features, while still being easily interpretable as a two-dimensional graph, and (3) can still be used if some of these features are not available. Two trees were constructed using the same training data as in the present paper: one tree using 51 elements and isotopic ratios and one using only 23 High Field Strength (HFS) elements and isotopic ratios. Both trees were evaluated with the same test data used on the discrimination diagrams. The full tree misclassifies 23 and the HFS tree 41 out of the 182 test data. Presently, the Si-Ti-Sr and Eu-Lu-Sr LDAs, and the Na-Nb-Sr and Ti-V-Sm QDAs introduced in this paper still outperform the trees of *Vermeesch* [2006]. However, this is likely to change for trees created from a larger training set. Whereas discriminant analysis does not gain much from using exceedingly large training sets, classification trees continue to improve with growing sets of training data. Furthermore, the classification trees succeeded in classifying all 182 test data, even for samples missing several geochemical features. None of the discrimination diagrams achieved this. Therefore it is probably a good idea to use a combination of both methods.

Acknowledgments

[56] Many thanks to Cameron Snow for proofreading the first draft of this paper. Careful reviews by Nick Arndt, Geoff Fitton, and particularly John Rudge are gratefully acknowledged.

References

Aitchison, J. (1982), The statistical analysis of compositional data, *J. R. Stat. Soc.*, *44*, 139–177.

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, 416 pp., CRC Press, Boca Raton, Fla.

Butler, R., and A. Woronow (1986), Discrimination among tectonic settings using trace element abundances of basalts, *J. Geophys. Res.*, *91*, 10,289–10,300.

Chayes, F. (1949), On ratio correlation in petrography, *J. Geol.*, *57*(3), 239–254.

Chayes, F. (1960), On correlation between variables of constant sum, *J. Geophys. Res.*, *65*, 4185–4193.

Chayes, F. (1971), *Ratio Correlation: A Manual for Students of Petrology and Geochemistry*, 99 pp., Chicago Univ. Press, Chicago, Ill.

Fisher, R. A. (1936), The use of multiple measurements in taxonomic problems, *Ann. Eugenics*, *7*, 179–188.

Lehnert, K., Y. Su, C. H. Langmuir, B. Sarbas, and U. Nohl (2000), A global geochemical database structure for rocks, *Geochem. Geophys. Geosyst.*, *1*(5), doi:10.1029/1999GC000026.

Meschede, M. (1986), A method of discriminating between different types of mid-ocean ridge basalts and continental tholeiites with the Nb-Zr-Y diagram, *Chem. Geol.*, *56*, 207–218.

Pearce, J. A. (1976), Statistical analysis of major element patterns in basalts, *J. Petrol.*, *17*(1), 15–43.

Pearce, J. A. (1982), Trace element characteristics of lavas from destructive plate boundaries, in *Andesites*, edited by R. S. Thorpe, pp. 525–548, John Wiley, Hoboken, N. J.

Pearce, J. A., and J. R. Cann (1971), Ophiolite origin investigated by discriminant analysis using Ti, Zr and Y, *Earth Planet. Sci. Lett.*, *12*(3), 339–349.

Pearce, J. A., and J. R. Cann (1973), Tectonic setting of basic volcanic rocks determined using trace element analyses, *Earth Planet. Sci. Lett.*, *19*(2), 290–300.

Pearce, J. A., and G. H. Gale (1977), Identification of ore-deposition environment from trace element geochemistry of associated igneous host rocks, *Spec. Publ. Geol. Soc. London*, *7*, 14–24.

Pearce, J. A., and M. J. Norry (1979), Petrogenetic implications of Ti, Zr, Y and Nb variations in volcanic rocks, *Contrib. Mineral. Petrol.*, *69*, 33–47.

Pearson, K. (1897), On a form of spurious correlation which may arise when indices are used in the measurement of organs, *Proc. R. Soc. London*, *60*, 489–502.

Shervais, J. W. (1982), Ti-V plots and the petrogenesis of modern ophiolitic lavas, *Earth Planet. Sci. Lett.*, *59*, 101–118.

Vermeesch, P. (2006), Tectonic discrimination of basalts with classification trees, *Geochim. Cosmochim. Acta*, *70*(7), 1839–1848.

Weltje, G. J. (2002), Quantitative analysis of detrital modes: Statistically rigorous confidence regions in ternary diagrams and their use in sedimentary petrology, *Earth Sci. Rev.*, *57*(3–4), 211–253.

Wood, D. A. (1980), The application of a Th-Hf-Ta diagram to problems of tectonomagmatic classification and to establishing the nature of crustal contamination of basaltic lavas of the British Tertiary volcanic province, *Earth Planet. Sci. Lett.*, *50*(1), 11–30.