EXPLORATION OF GEOLOGICAL VARIABILITY AND POSSIBLE PROCESSES THROUGH THE USE OF COMPOSITIONAL DATA ANALYSIS: AN EXAMPLE USING SCOTTISH METAMORPHOSED

C. W. Thomas
British Geological Survey
West Main Road
Edinburgh EH9 3LA
SCOTLAND.
Email: cwt@bgs.ac.uk

J. Aitchison
Department of Statistics
University of Glasgow
Glasgow G12 8QQ
SCOTLAND
Email: john.aitchison@btinternet.com

ABSTRACT

Developments in the statistical analysis of compositional data over the last two decades have made possible a much deeper exploration of the nature of variability, and the possible processes associated with compositional data sets from many disciplines. In this paper we concentrate on geochemical data sets. First we explain how hypotheses of compositional variability may be formulated within the natural sample space, the unit simplex, including useful hypotheses of subcompositional discrimination and specific perturbational change. Then we develop through standard methodology, such as generalised likelihood ratio tests, statistical tools to allow the systematic investigation of a complete lattice of such hypotheses. Some of these tests are simple adaptations of existing multivariate tests but others require special construction. We comment on the use of graphical methods in compositional data analysis and on the ordination of specimens. The recent development of the concept of compositional processes is then explained together with the necessary tools for a staying-in-the-simplex approach, namely compositional singular value decompositions. All these statistical techniques are illustrated for a substantial compositional data set, consisting of 209 major-oxide and rare-element compositions of metamorphosed limestones from the Northeast and Central Highlands of Scotland. Finally we point out a number of unresolved problems in the statistical analysis of compositional processes.

1. Introduction

This is essentially a compositional case study. We use a suite of limestones and their geochemical compositions – major oxides and trace elements – to demonstrate that a whole variety of recent techniques of compositional data analysis can be brought into play to allow meaningful geological inferences. Our approach is applied mathematical. After a description of the provenance of the limestones we shall proceed to use appropriate modelling and compositional data analysis to answer a sequence of questions as posed by the geologist.

2. The limestones

Metamorphosed limestones form potentially useful markers within the much more abundant siliciclastic rocks in the Central and Northern Highlands of Scotland. However, limestones are relatively uncommon in the Central Highlands and resolution of their lithostratigraphical status has proved problematical by conventional

geological mapping. Limestones are more abundant in the Northern Highlands. Here the discriminatory power of their geochemistry, established in areas of good exposure and well constrained lithostratigraphy, can prove useful in aiding lithographical correlation areas of much poorer exposure.

In this study we confine attention to two limestone data sets consisting 49 Dufftoen limestones and 160 Inchrory limestones, each with 17-part geochemical compositions (Si, Al, Ti, Fe, Mg, Ca, Na, K, Mn, P, Loi,, Ba', Rb. Sr; V Y; Zr). The main problems considered in this study are to see how these may be considered as marker data sets, investigating in full the nature of their differences. Also, since Inchrory limestones are 'younger' than Dufftown limestones we consider how we may describe the process of change from one to the other.

For more geological details see Thomas and Aitchison (1998)

3. Are Duffinch and Inchrory limestones compositionally different?

It is tempting to investigate this question by considering the lattice of Aitchison (1986, Section 7.5 and Fig. 7.3) whereby we would test whether the Duffinch and Inchrory limestone compositions came from the same additive logistic normal distributions. The fact that the distributional form of the compositional variability cannot be assumed to be logistic normal and that although we may attempt to fit multivariate logistic skew normal distributions we would have to await a set of comprehensive multivariate skew normal tests before we could complete that route.

But a simple examination of the estimated centres of the distributions of the major oxides for Duffinch and Inchrory respectively::

0.0496 0.0061 0.0003 0.0029 0.0136 0.5023 0.0007 0.0013 0.0002 0.0002 0.4229

0.1276 0.0270 0.0011 0.0106 0.0069 0.4540 0.0025 0.0050 0.0003 0.0007 0.3644

will easily convince us that there are compositional differences between the two compositional suites. Subsequent analysis, not dependent on distributional form will confirm that there are substantial compositional differences between the two limestones.

4. What is the simplest way of discriminating between Duffinch and Inchrory limestonesr?

Aitchison (1986, Chapter 12) gave a number of practical situations where compositions play an explanatory or regressor role, where we may wish to see how a composition is changed by different treatments, where in experiments with mixtures we may attempt to determine the mixture which will provide the optimum response, and in classification or diagnostic problems where we may wish to use a composition as a convenient or efficient means of determining type or to find out if any subcomposition accounts for the substantive difference between the types.

Binary logistic discrimination. For two types, such as the Duffinch and Inchrory limestones here we may use binary logistic regression with logconstrasts of the compositional components as the regressor; for a *D*-part composition x a logcontrast is defined as

$$lc(a, x) = a \log x_1 + \ldots + a_D \log x_D \quad (a_1 + \ldots + a_D = 0).$$

For two types ($t = 0$, $t = 1$) the binary logistic model is thus defined by

$$pr(t = 0 \mid x) = 1 - pr(t = 1 \mid x) = \frac{\exp\{a_0 + lc(a, x)\}}{1 + \exp\{a_0 + lc(a, x)\}} .$$

Maximum likelihood estimation of the parameter a is straightforward. The beauty of this model is that the adequacy of a subcomposition say (1, . . . , *C*) can readily be tested since this hypothesis can be expressed as $a_{C+1} = \ldots = a_D = 0$. Thus the whole lattice of subcompositional hypotheses can be investigated and any adequate subcomposition identified. Examples of this procedure can be found for hongite-kongite discrimination and Permian and post-Permian: discrimination in Aitchison (1986, Sections 12.6, 12.7).

Figure 1 shows a portion of the substantial lattice of subcompositional hypotheses with some of the more interesting subcompositions highlighted. The maximum model at the top of the lattice retains the full 17-part composition as the explanatory variable. At the foot of the lattice is the hypothesis that compositional information is useless. At the next bottom level we have all the two-part subcompositions; at the next level all the three-part subcompositions, and so on. In such lattice testing we use generalised likelihood ratio tests always testing the hypothesis within the maximum model, starting with the simplest hypothesis, moving up the lattice to the next level only if we can reject all the hypotheses at the lower level and so on until we reach a hypothesis that we cannot reject. In the event that there are several non-rejectable hypotheses at a particular level that we either ‘accept’ the one with the smallest likelihood ratio or move to the next level and ‘accept’ the hypothesis with the smallest likelihood ratio. It should be remembered that in the absence of any loss structure in a multiple hypothesis situation all testing procedures are essentially ad hoc. We can only report that we have found that this form of lattice testing usually leads to sensible inferences.

The dramatic result for limestones was reported in Thomas and Aitchison (1998), where out of the 17-part geochemical composition a 3-part subcomposition, namely (Fe, Mg, Ca), is found to be an adequate discriminator. The separation of the duffinch and inchrory limestones on the basis of the (Fe, Mg, Ca) subcomposition was illustrated in a logratio scattergram, namely a (log(Fe/Ca), log(Mg/Ca)) plot. We show in Figure 2a the ternary diagram of the (Fe, Mg, Ca) subcompositions. Because of the smallness of Fe and Mg relative to Ca the subcompositional points crowd into the Ca corner of the ternary diagram providing little illumination of the differences in the duffinch and inchrory subcompositions. We can, however, use the device of a centering perturbation *von Eynatten, Pawlowsky-Glahn and Mateu Figueras, 2002) to provide a clearer picture of the separation. This is shown in Figure 2b, together

with the perturbed dividing logcontrast line which separates Duffinch from Inchrory..We record that the estimated logcontrast discriminator is

$$6.98 + 4.84 \log \text{Fe} - 3.05 \log \text{Mg} - 1.79 \log \text{Ca}.$$

It is worth commenting here that prior to this analysis, reliance on spider diagrams had suggested strongly that discrimation between the two limestones lay in the siliciclastic component, the (Si, Ti, Al, K, Rb, V) subcomposition, not in the carbonate component, the subcomposition (Fe, Mg, Ca). In the lattice of Figure 1 we show the results for the siliciclastic subcomposition. It is quite clear that the hypothesis that this had a differentiating effect has to be strongly rejected.

Another worthwhile comment here is about the use of perturbation, the operation of change in compositions. The use above in the ternary diagram is an excellent device for the provision of simpler visual interpretation of ternary scattergrams. The reader will have noticed that we have used the 11 major oxides (proportion by weight) and 6 trace elements (parts per million) as if these constituted a single composition, despite the fact that they are measured in different units. The reason for this is that it requires only a simple constant perturbation to attain common units. And such perturbations have useful invariance properties. In the present context it is simple to demonstrate that such a perturbation would affect only the value of a_0 in the binary logistic model and this has no relevance within the lattice testing procedure. See Aitchison (2003) for further comment.

5. What is the nature of the change between duffinch and inchrory limestones?

Inchrory limestones are younger than Dufftown limestones and so it is reasonable to ask whether we can describe how a generic Inchrory limestone may be altered into a generic Dufftown limestone. We confine our discussion to the major oxide compositions. We recall that the operation of change for compositions is a perturbation Suppose that x and X denotes the compositions of generic Inchrory and Dufftown limestones and that X arises as a perturbation p of x, namely, $X = p \oplus x$. Then we know that in terms of centres of the distributions

$$cen(X) = cen(p) \oplus cen(x),$$

so that

$$cen(p) = cen(X) \ominus cen(x),$$

and is easily determined from the centres in Section 3 as

Si	Al	Ti	Fe	Mg	Ca	Na	K	Mn	P	Loi
0.0554	0.0321	0.0381	0.0394	0.2823	0.1578	0.0390	0.0366	0.1038	0.0499	0.1655

We have in our modelling allowed the perturbations to be variable and are using the centre of these to give an indication of the nature of the perturbation, This centre is, of

course, exactly what we obtain if we assumed that there was a constant perturbation at work.

First we note a trivial point, namely that the there is near equality pf the Ca and Loi components. This is simply recording the stability of the (Ca, Loi) subcomposition, in other words the fact that Loi will always match the Ca component.

Second, since we saw the importance of the (Fe, Mg, Ca) subcomposition in discriminating between the two liimestones we note the considerable differences in the perturbation components of these oxides, namely [0.0394 0.2823 0.1578], showing how in the change from Inchrory to Dufftown Fe has given way substantially to Mg and less so to Ca. The perturbation components of all the other oxides rxcept from Mn are reasonable constant indicating stability of the associated subcomposition. The question then arises: is there some geological explanation fo the relative increase in Mn?

6. Are there any helpful graphical representations of the variability in the limestones?

We have already seen the use of a logratio scattergram to demonstrate the separation between Duffinch and Inchrory limestones. There is hopefully no need to repeat the warnings of Chayes and others on the misuse of such other graphical tools such as Harker diagrams. We have also seen the use of a centering perturbation to improve the use of a ternary diagram to show the relevance of the carbonate (Fe, Mg, Ca) subcomposition as a discriminator between the Duffinch and Inchrory.limestones.

The main advance in other graphical descriptions of compositional variability has been the extension of the familiar and powerful unconstrained biplot technique to compositional data; see Aitchison (1990b, 1997, 2001) and Aitchison and Greenacre (2002) for details. We are currently exploring these and will present an example at CODAWORK03.

7. Is it possible to order specimens within the limestone suites?

If it is assumed that there is a process of change from Inchrory-ness to Dufftown-ness can we somehow order the limestones within this process there are simple compositional more or less equivalent ways of doing this. From the discrimination analysis we can compute for each limestone the probability of allocation to the Inchrory set and arrange these probabilities in decreasing order. Or we use the principal logcontrast component (Aitchison, 1986, Section 8.3) or equivalently by use of the first coefficients in a singular value decomposition (Aitchison, 2003). With such techniques it is also possible to interpret the possible position of other non-Inchrory, non Duffinch limestones with the Inchrory to Duffinch process. We do not report details here; for some earlier results, see Thomas and Aitchison (1998).

8. Can we identify possible compositional processes from the limestone data?

Aitchison and Thomas (1998) explained how modern compositional data analysis could be used to investigate identify and analyse possible compositional processes, using two simple examples: Arctic lake sediments to identify the process in relation to depth and simple olivine data to identify possible equilibrium relationships. The statistical technique used depended largely on compositional regression analysis and logcontrast principal component analysis. In a complementary paper Aitchison and Barceló-Vidal (2002) used compositional singular value decompositions to provide a staying-in-the-simplex approach where the compositions of the data set are expressed as power-perturbation combinations. Thus for an $N \times D$ compositional data matrix with *n*th row the composition x_n the singular value composition provides the expression

$$x_n = \hat{x} \oplus (u_{n1} p_1 \otimes b_1) \oplus \ldots \oplus (u_{nR} p_R \otimes b_R),$$

where $\hat{x}$ is the estimate of the centre of the data set, and p_i $(i = 1, \ldots, R)$ are positive 'singular values' in descending order of magnitude, the b_i $(i = 1, \ldots, R)$ are orthogonal compositions, R is a readily defined rank of the compositional data set and the u's are power components specific to each composition. . In practice R is commonly $D-1$, the full dimension of the simplex. In a way similar to that for data sets in R^D we may consider an approximation of order $r < R$ to the compositional data set given by

$$x_n^{(r)} = \hat{x} \oplus (u_{n1} p_1 \otimes b_1) \oplus \ldots \oplus (u_{nr} p_r \otimes b_r).$$

Such an approximation retains a proportion

$$(p_1^2 + \ldots + p_r^2) / (p_1^2 + \ldots + p_R^2)$$

of the total variability of the $N \times D$ compositional data matrix as measured by the trace of the estimated centered logratio covariance matrix or equivalently in terms of the total mutual squared distances as

$$\{N(N-1)\}^{-1} \sum_{m<n}^{D} \Delta_S^2 (x_m, x_n).$$

We have presented the process above as relative to the centre $\hat{x}$ but an alternative and possible more meaningful way is simply to omit this centering step.

In applying this to the possible limestone process we have chosen not to centre and applied the singular value decomposition to the combined inchrory duffinch data. Up to the fourth order the *b* compositions are

Si	Al	Ti	Fe	Mg	Ca	Na	K	Mn	P	Loi
0.1175	0.0955	0.0648	0.0857	0.0865	0.1426	0.0717	0.0781	0.0566	0.0618	0.1391

0.0675	0.0622	0.0788	0.0724	0.1254	0.1102	0.0560	0.0701	0.1460	0.0991	0.1123
0.0993	0.1111	0.1080	0.1028	0.0709	0.0807	0.0389	0.1190	0.0815	0.1076	0.0803
0.0674	0.0810	0.0899	0.0765	0.0422	0.1190	0.1011	0.0845	0.0899	0.1354	0.1132

with successive degrees of approximation 0.46, 0.69, 0.84, 0.92.

We note that in the first approximation the major relative increase is in Ca and to a lesser extent Si. At the second approximation, both Mg and Ca show a relative increase and the increase in Mn seen earlier is also evident. At the third approximation the main feature seems to be some relative depletion of Na, to a relative gain in K? And so on in a tentative interpretation.

There are substantial statistical deficiencies in this area of compositional data analysis. For example at the various stages of approximation the near equality of a subset of components suggests that the corresponding subcomposition is stable for that stage. But can we devise suitable statistical tests of such a subcompositional stability hypothesis, Also although the singular value decomposition is identifying orthogonal aspects of the process it may be that some other representation, say of the first two stages may provide a more enlightening view of the process. This would be a restructuring rather similar to the varimax technique in principal component analysis, but how should it be done compositionally? Again, suppose that the geologist suggest that there may be two basic processes at work, possibly independently, a carbonate and a siliciclastic process. A possible method of modelling might then be to set

$$x_n \hat{=} (u_{n1} p_1 \otimes b_1) \oplus (u_{n2} p_2 \otimes b_2) \oplus p,$$

where b_1, b_2 are D-part compositions somehow representing carbonate and siliciclastic subcompositions, but how, p_1, p_2 are parameters representing the extent of the carbonate and siliciclastic importance and requiring estimation, the u coordinates again requiring to be determined and the p a composition, rather like the error term in a regression analysis. It is obvious that there is much statistical research remaining to be done in this important area of compositional data analysis. We are currently working in this area and will certainly report any new work at CODAWORK03.

REFERENCES

Aitchison, J., 1986, The Statistical Analysis of Compositional Data: Chapman and Hall, London. Reprinted in 2003 with additional material by The Blackburn Press.

Aitchison, J., 1990, Relative variation diagrams for describing patterns of variability of compositional data: Math. Geology, v. 22, p. 487-512.

Aitchison, J., 1997, The one-hour course in compositional data analysis or compositional data analysis is easy, *in* Pawlowsky Glahn, V., ed., Proceedings of the Third Annual Conference of the International Association for Mathematical Geology: CIMNE, Barcelona, p. 3-35.

Aitchison, J., 2001, Simplicial inference, *in* Viana, M.A.G. and Richards, D.St.P., eds., Algebraic Methods in Statistics and Probability: Contemporary Mathematics Series 287, American Mathematical Society, Providence, Rhode Island, p. 1-22.

Aitchison, J., 2003. Compositional data analysis: where are we and where are we heading? : paper at CODAWORK03.

Aitchison, J. and Barceló-Vidal, C., 2002,. Compositional processes: a statistical search for understanding, *in* Proceedings of the Eighth Annual Conference of the International Association for Mathematical Geology, to appear.

Aitchison, J. and Greenacre, M., 2002, Biplots for compositional data: Appl. Statist., v. 51.. p. 375-382.

Aitchison, J. and Thomas, C. W. (1998) Differential perturbation processes: a tool for the study of compositional processes. In *Proceedings of the Fourth Annual Conference of the International Association for Mathematical Geology* (Buccianti A., Nardi, G. and Potenza, R., eds), pp. 499-504. Naplesi: De Frede.

Thomas, C. W. and Aitchison, J. (1998). The use of logratios in subcompositional analysis and geochemical discrimination of metamorphosed limestones from the northeast and central Scottish Highlands. In: A. Buccianti, G. Nardi and R. Potenza, Eds., *Proceedings of IAMG98, The Fourth Annual Conference of the International Association for Mathematical Geology*, De Frede, Naples, 549-554...

von Eynatten, H,. Pawlowsky-Glahn and MateuFigueras, G., 2002, Understanding perturbation on the simplex: a simple method to better visualise and interpret compositional data in ternary diagrams: Math. Geology, v. 34, p. 249-257.