Mathematical Geology, Vol. 38, No. 8, November 2006 (© 2007)
DOI: 10.1007/s11004-006-9059-3

Application of the Multivariate Runs Test to Compositional Data[1]

W. E. Sharp[2]

KEY WORDS: minimum spanning tree, ternary diagrams, taxi-cab metric.

INTRODUCTION

The graph median has recently been proposed (Sharp, 2006) as an alternative measure of central tendency in those instances when the arithmetic average is not a part of the data set. This property was illustrated with ternary diagrams by using a series of five three-component compositional data sets all of which had the same mean, but greatly differing patterns in their compositions. For these sets, a comparison of their means and, except for the case in which all compositions (hongite2) fell on the same line, a comparison of their variance would be of no value. Contemplation of this fact, suggested the need for a simple method that determines the extent to which two data sets with several components are coincident.

Suppose one visualizes each component of a composition as being plotted along a separate coordinate axis, then an individual compositon can be represented as a point in a space of three or higher dimension and the data set as a whole when plotted would form a cloud of points. If the two data sets are statistically the same, the two clouds of points should be of similar size, shape and internal point distribution. A test for this can be conceived as follows. Merge the two clouds of points into one and then connect each pair of points in the combined cloud with the one set of lines such that the total length of the lines is a minimum. In graph theory, the points are termed nodes, the lines as edges, the set of connected lines as a tree and the set of lines having a minimum total length as a minimum spanning tree (e.g. Reingold, Nievergelt, and Deo, 1977, chap. 8). If the two data

[1]Received 21 April 2006; accepted 23 June 2006; Published online: 17 February 2007.
[2]Department of Geological Sciences, University of South Carolina, Columbia, South Carolina 29208 USA; e-mail: sharp@sc.edu

sets are coincident, then the points comprising the minimum spanning tree (MST) should alternate between the two data sets in a random fashion as one traverses through the tree. As such a random alternation between two data sets would be analogous to the flipping of a coin it should be possible to count the number of times, a change between sets occurs. Whether or not the observed number of changes (runs) is statistically random can be tested using the well known runs test (Wald and Wolfowitz, 1940 or see Conover, 1971; Davis, 1973, p. 184–191). To see if this approach might have been used previously, a literature search was made and showed that the runs test has already been extended to the multivariate case through the use of the minimum spanning tree (Friedman and Radfsky, 1979). Although clearly established, the multivariate runs test (MVRT) seems to have been used rarely and it appears not to have been applied to compositional data.

MULTIVARIATE RUNS TEST

To perform a multivariate runs tests on two data sets, they must be merged into a single combined data set in such a way as to be able to keep tabs as to which samples belong to each individual data set. A simple way to do this is to append one data set to the other and then to number the observations in the combined data set sequentially. An observation number greater than the number of observations in the first data set must then belong to the second data set (see Table 1). As the data are compositional, the distances between all compositonal observations are calculated by using a composition (taxi-cab) metric (Miller, 2003; Shurtz, 2003, p. 930). From that result, the minimal spanning tree over the combined data set can then be determined by the method of Prim (1957; also see Reingold, Nievergelt, and Deo, 1977, p. 323–325).

Table 1. Two Contrived Suites of Samples R and S Composed of A, B and C

OBS	SMP	A	B	C
1	R1	44.00	49.00	7.00
2	R2	44.53	18.82	36.65
3	R3	48.58	26.98	24.44
4	R4	1.54	42.66	55.80
5	R5	55.01	42.55	2.44
6	R6	31.19	45.58	23.22
7	R7	37.03	46.59	16.38
8	S1	13.37	33.15	53.48
9	S2	35.73	25.49	38.79
10	S3	59.58	13.59	26.83
11	S4	12.37	39.75	47.89

The number of runs is then found by removing all lines between points that belong to different sample sets and then counting the number of disjoint trees that are formed (Friedman and Rafsky, 1979, p. 700). Operationally, the number of observed runs is obtained simply by counting the number of lines connecting points which belong to different sample sets (Friedman and Rafsky, 1979, p. 709). The expected number of runs and the variance are given by the usual formulas based on the number of observations in the two samples (Wald and Wolfowitz, 1940, p. 151):

$$U = ((2n_1n_2)/(n_1 + n_2)) + 1$$

$$s^2 = (2n_1n_2)(2n_1n_2 - n_1 - n_2)/(n_1 + n_2)^2(n_1 + n_2 - 1)$$

The test statistic (Friedman and Rafsky, 1979, p. 698) is then given by:

$$Z = (U_\mathrm{o} - U)/s$$

where Z is normally distributed, $N(0, 1)$, U_o is the observed number of runs, U is the expected number of runs, s^2 is the expected variance in the runs, n_1 and n_2 are the number of samples in suites 1 and 2. These formulas apply provided both n_1 and n_2 are greater than 20. When they are less than 20, tables for testing are available in books on non-parametric statistics (e.g. Gibbons and Chakraborti, 2003, p. 568).

EXAMPLE

To illustrate these concepts, two suites of samples with three components were contrived (Table 1). The first with seven samples and the other with four and these were combined into a single set of 11 samples. The compositional distance between each observation can be calculated from:

$$d_{ij} = |A_i - A_j| + |B_i - B_j| + |C_i - C_j|$$

and the MST obtained by Prim's method. These results are summarized in Table 2 and a plot of the MST shown in a ternary diagram (Fig. 1) with lines indicating the taxi-cab paths between individual compositons in which the compositions of suite R are shown with dots and those of suite S as plus signs. This example is simple enough that inspection shows a total of three paths (dashed lines) connect compositions of the differing suites. The number of runs is also easily determined in Table 2 by counting the number of point pairs that join different suites.

Because the number of compositions and the number of runs are so small, a test of significance is best obtained by consulting tables (e.g. Gibbons and

Table 2. Links in the Minimum Spanning Tree for Combined Suites R and S

SMP	OBS	OBS	SMP	d_{ij}
S1	8	11	S4	13.19
R7	7	6	R6	13.69
R2	2	9	S2	17.61
R1	1	7	R7	18.76
S4	11	4	R4	21.65
R1	1	5	R5	22.02
R3	3	2	R2	24.42
R3	3	10	S3	26.78
R6	6	3	R3	37.21
S2	9	8	S1	44.71

Chakraborti, 2003, p. 568). For a count of three runs, the tables give a left-tail probability of 0.033 which leaves in doubt whether or not the two suites are statistically similar or different.

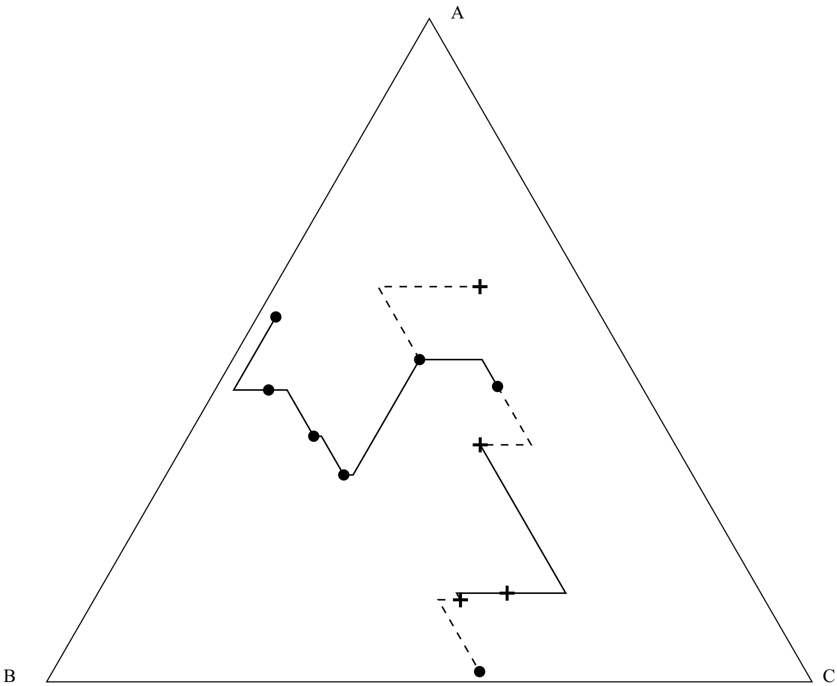

Figure 1. Ternary diagram showing the ABC compositions of suites R (•) and S (+) and the taxi-cab paths for the minimum spanning tree for the combined set.

TESTING HONGITE VARIATIONS

To provide test compositions which were not "prohibitively large" but with "sufficient complexity to capture the typical difficulties" in the analysis of compositional data, Aitchison (1984, p. 533–536) contrived four sets of samples which he designated as hongite, kongite, boxite and coxite. Sharp (2006) then contrived a set of four variations on the ABC subcomposition of hongite (Aitchison, 1986, p. 9) to illustrate how distances based on the composition (taxi-cab) metric (Miller, 2002; Shurtz, 2003) could be used to obtain a consistent measure of central tendency, the graph median. As the point clouds of each variation of hongite when plotted in a ternary diagram were distinct (see Sharp, 2006), it seemed an interesting exercise to use MVRT to see how they would compare with one another statistically. The results of this exercise (Table 3) generally show that as expected each of the variations are statistically distinct.

The typical overlap among the hongite variations is illustrated by a plot of hongite with hongite4 (Fig. 2). While the two clouds of points have some overlap, there is a fair degree of separation between the two hongite variations. Now in the case of hongite and hongite5 (Fig. 3), the two suites can still be discerned because the method by which hongite5 was derived from hongite is known. However, the MVRT shows that the two suites (Table 3) cannot be distinguished statistically ($Z = -0.286$). Careful inspection of this ternary plot shows that the two suites are interwined so in the absence of a specfic model, they are indeed statistically indistinguishable.

Table 3. Results of the Multivariate Runs Test Comparing the Variations in the Hongite Subcompostions ABC

Suites	U_o	Z
H–H2	14	−3.429
H–H3	15	−3.144
H–H4	11	−4.287
H–H5	25	−0.286
H2–H3	11	−4.287
H2–H4	8	−5.144
H2–H5	11	−4.287
H3–H4	6	−5.715
H3–H5	8	−5.144
H4–H5	5	−6.001

Note. H: hongite, H2: hongite2, H3: hongite3, H4: hongite4, H5: hongite5.
$n_1 = n_2 = 25$, $U = 26$, $s = 3.50$, $Z(1\%) = -2.326$, $Z(5\%) = -1.645$.
U_o: observed runs, Z: expected normal score.

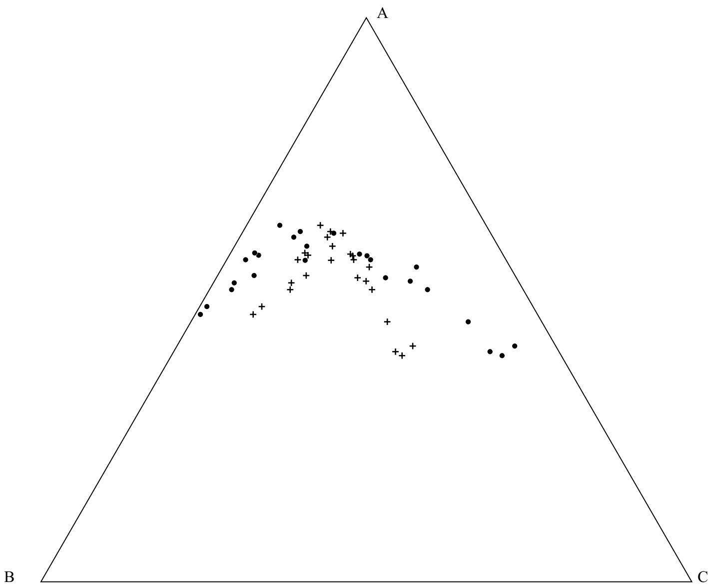

Figure 2. Ternary diagram showing the ABC subcompositions for both hongite (•) and hongite4 (+). Note that good separation between hongite and hongite4.

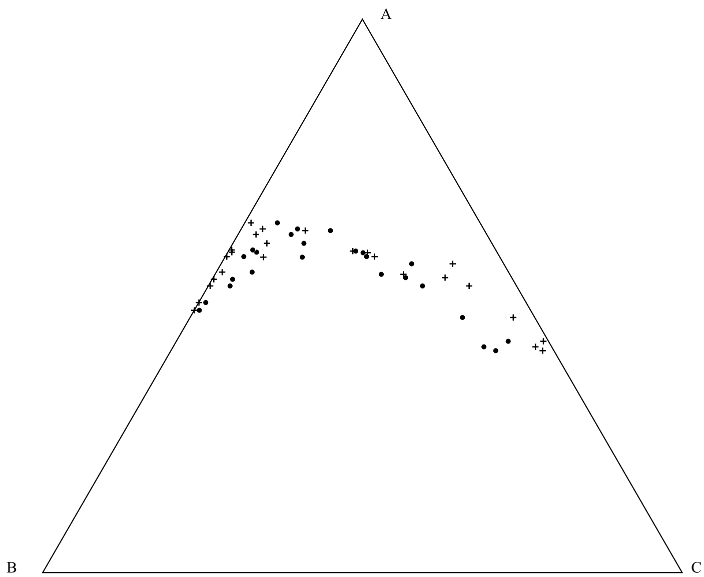

Figure 3. Ternary diagram showing the ABC subcompositions for both hongite (•) and hongite5 (+). Note how the compositions of hongite and hongite5 are intertwined.

TESTING HONGITE, KONGITE, BOXITE AND COXITE

In an early discussion of geochemical compositions, Aitchison (1984, p. 533–536) presented four suites of samples having five compositional components: A, B, C, D, E. As these compositions are sufficiently complex, simple inspection of the tabulated values will not reveal if they are similar nor can they easily be tested graphically for similarity with simple diagrams. This is just the circumstance for which the MVRT is most suitable. Each of the four suites was tested one against the other by MVRT. The results (Table 4) show, except for hongite and kongite, each of these suites is compositionally distinct.

Generally when comparing clouds of points with the MVRT if the compositions have little or no coincidence, the number of observed runs will be significantly less than the expected number of runs. This is the case normally encountered and the usual one here and also in the case of the variations on hongite (Table 3). If the observed number of runs is greater than the expected number, this would suggest the presence of some form of systematic alternation. Now in the case of hongite and kongite, the number of runs is slightly greater than the expected number. This would suggest that not only are hongite and kongite statistically the same, but that kongite was derived from hongite by some kind of offset.

Table 4. Results of the Multivariate Runs Test Comparing Four Suites of with the Five Components: A, B, C, D, E

Suites	U_o	Z
H–K	29	+0.857
H–B	6	−5.715
H–C	5	−6.001
K–B	11	−4.287
K–C	9	−4.858
B–C	1	−7.144

Note. H: hongite, K: kongite, B: boxite, C: coxite.
$n_1 = n_2 = 25$, $U = 26$, $s = 3.50$, $Z(1\%) = -2.326$, $Z(5\%) = -1.645$.
U_o: observed runs, Z: expected normal score.

TESTING SOME ALKALINE GLASSES

Thus far all of the examples used have been contrived. To illustrate the application of the method to some actual data, a set of alkaline glass analyses were selected (Baxter, Cool, and Jackson, 2005, p. 64–68). These glasses are colorless Romano-British vessel glasses collected from various archaeological sites in England. They consist of cast bowls, facet-cut beakers, wheel-cut beakers and cylindrical cups. These are interesting in that these manufactured glasses are very similar in composition and it would be nice to determine whether or not the compositions differ because of differences in their source, their date or their manufacturing process. Each of the four suites of glasses were tested against one another (Table 5). The key result is that the glass used to make the cast bowls and the wheel-cut beakers are statistically the same. In contrast, the glass used to make facet-cut beakers is statistically distinct from all of the other glasses. In this particular case, simple inspection of the glass analyses (Baxter, Cool, and Jackson, 2005) shows that the facet-cut beakers have elevated amounts of antimony and lead. Interestingly, even if one removes the lead and antimony analyses and computes the residual subcompositions, the facet-cut beakers are still statistically distinct indicating some other compositional effect is involved. In the case of the cylindrical cups, these are clearly different from the cast bowls and the wheel-cut beakers but simple inspection does not reveal the nature of that difference. While MVRT allows one to see if two compositional data sets are coincident, it does not indicate how they might differ.

Table 5. Results of the Multivariate Runs Test Comparing Four Suites of Alkaline Glasses with 12 Oxide Components: Al, Fe, Mg, Ca, Na, K, Ti, P, Mn, Sb, Pb, Si

Suites	n_1	n_2	U	s	U_o	Z
cast–wheel	34	51	41.8	4.40	36	− 1.319
cast–cups	34	97	51.4	4.37	32	− 4.426
cast–facet	34	63	45.2	4.45	22	− 5.198
wheel–cups	51	97	67.9	5.47	40	− 5.089
wheel–facet	51	63	57.4	5.26	25	− 6.159
cups–facet	97	63	77.4	6.02	17	−10.034

Note. $Z(1\%) = -2.326$, $Z(5\%) = -1.645$.
n_1 and n_2: number of observations, U: expected runs, s: std. dev. in U, U_o: observed runs, Z: expected normal score.

SUMMARY

As an integral part of routine geological and geochemical studies of rocks, large numbers of whole rock chemical analyses are performed each year in order to characterise these rocks or suites of rocks. The results for major-element analyses are commonly interpreted by the use of variation (subcomposition) diagrams (Rollinson, 1993, p. 66–83) and various discrimination diagrams (Rollinson, 1993, p. 49–52). To these procedures should be added the multivariate runs test.

The multivariate runs test is a simple method that allows one to test whether or not two suites of compositionally complex data, such as rock analyes, are statistically coincident. The method requires determining the taxi-cab distance between all pairs of analyses and then using the method of Prim to find the minimum spanning tree through the combined set of observations. The observed number of runs is easily obtained by counting just those lines in the MST which join points between two different sample suites. The expected number of runs, its variance and the normal score are readily obtained from the number of samples in each suite using formulas found in many books on non-parametric statistics.

Unfortunately at the present time, anyone who sets out to apply MVRT to compositional data will find that while algorithms for determining the MST are common enough, they are usually not incorporated into the standard statistical packages and if they are, may not use the taxi-cab metric.

While this procedure provides a simple method of testing the coincidence of two complex compositional data sets, it gives no information as to how two compositional data sets may differ. Inspection of the table of analyses may suggest that the difference may lie in certain specific components. If that is the case, a test of the renormalized subcomposition with these components removed may prove useful. One might also try a dimension reducing method such as principal component analysis or a special projection of the MST such as suggested by Friedman and Rafsky (1981).

REFERENCES

Aitchison, J., 1984, The statistical analysis of geochemical compositions: Math. Geol., v. 16, no. 6, p. 531–564.

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

Baxter, M. J., Cool, H. E. M., and Jackson, C. M., 2005, Further studies in the compositional variability of colorless Romano-British vessel glasses: Archaeometry, v. 47, no. 1, p. 47–68.

Conover, W. J., 1971, Practical nonparametric statistics: John Wiley, New York, 462 p.

Davis, J. C., 1973, Statistics and data analysis in geology: John Wiley, New York, 550 p.

Friedman, J. H., and Rafsky, L. C., 1979, Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample testing: Ann. Stat., v. 7, no. 4, p. 697–717.

Friedman, J. H., and Rafsky, L. C., 1981, Graphics for the multivariate two-sample problem: J. Am. Stat. Assoc., v. 76, no. 374, p. 277–287.

Gibbons, J. D., and Chakraborti, S., 2003, Nonparametric statistical inference, 4th ed.: Marcel Dekker, New York, 645 p.

Miller, W. E., 2003, Revisting the geometry of a ternary diagram with the half-taxi metric: Math. Geol., v. 34, no. 3, p. 275–290.

Prim, R. C., 1957, Shortest connection networks and some generalizations: Bell Syst. Tech. J., v. 36, p. 1389–1401.

Reingold, E. M., Nievergelt, J., and Deo, N., 1977, Combinatorial algorithms: Theory and practice: Prentice-Hall, Englewood Cliffs, 433 p.

Rollinson, H. R., 1993, Using geochemical data: Longman, Edinburgh Gate, 352 p.

Sharp, W. E., 2006, The graph median—A stable alternative measure of central tendency for compositional data sets: Math. Geol., v. 38, no. 2, p. 221–229.

Shurtz, R. F., 2003, Compositional geometry and mass conservation: Math. Geol., v. 35, no. 8, p. 927–937.

Wald, A., and Wolfowitz, J., 1940, On a test whether two samples are from the same population: Ann. Math. Stat, v. 11, p. 147–162.