When zero doesn't mean it and other geomathematical mischief.

R.A. Valls Alvarez[1]

[1]Valls Geoconsultant., 1008-299 Glenlake Ave, Toronto, Ontario,M6P4A6, Canada
Corresponding author: vallsvg@aol.com

Abstract

There is almost not a case in exploration geology, where the studied data doesn't includes below detection limits and/or zero values, and since most of the geological data responds to lognormal distributions, these "zero data" represent a mathematical challenge for the interpretation.

We need to start by recognizing that there are zero values in geology. For example the amount of quartz in a foyaite (nepheline syenite) is zero, since quartz cannot co-exists with nepheline. Another common essential zero is a North azimuth, however we can always change that zero for the value of 360°. These are known as "Essential zeros", but what can we do with "Rounded zeros" that are the result of below the detection limit of the equipment?

Amalgamation, e.g. adding Na_2O and K_2O, as total alkalis is a solution, but sometimes we need to differentiate between a sodic and a potassic alteration. Pre-classification into groups requires a good knowledge of the distribution of the data and the geochemical characteristics of the groups which is not always available. Considering the zero values equal to the limit of detection of the used equipment will generate spurious distributions, especially in ternary diagrams. Same situation will occur if we replace the zero values by a small amount using non-parametric or parametric techniques (imputation).

The method that we are proposing takes into consideration the well known relationships between some elements. For example, in copper porphyry deposits, there is always a good direct correlation between the copper values and the molybdenum ones, but while copper will always be above the limit of detection, many of the molybdenum values will be "rounded zeros". So, we will take the lower quartile of the real molybdenum values and establish a regression equation with copper, and then we will estimate the "rounded" zero values of molybdenum by their corresponding copper values.

The method could be applied to any type of data, provided we establish first their correlation dependency.

One of the main advantages of this method is that we do not obtain a fixed value for the "rounded zeros", but one that depends on the value of the other variable.

Key words: compositional data analysis, treatment of zeros, essential zeros, rounded zeros, correlation dependency.

1. Are there any zeros in the house?

We need to start by recognizing that there are zero values in geology. For example the amount of quartz in a foyaite (nepheline syenite) is zero, since quartz cannot co-exists with nepheline (Trusova and Chernov, 1982). In binomial distributions, like for example the drilling of an ore body, you either will intersect the ore body (1) or not (0). Another common zero is a North azimuth, however we can always change that zero for the value of 360°. These are known as "Essential zeros" (Aitchison, 2003) or "Real zeros". They are not a problem for as long as their population does not respond to a log-normal distribution, since you can't take the logarithm of a zero. Then in geology, especially in geochemistry, we also have "Rounded zeros". In some cases, labs report bellow detection limit (b.d.l.) as zeros or non existent, while in most cases they just put the b.d.l. as the value for that parameter. These b.d.l. values are a similar problem to the "rounded zeros". Let us illustrate with the example proposed in Table 1.

Table 1. CLR transformed data for the used example. The original b.d.l. value was 0.5 (see appendix 1).

Au	Cu	Mo
0.23342	0.72138	0.0452
0.32663	0.61146	0.06191
0.12652	0.16198	0.71149
0.20133	0.28139	0.51728
0.20796	0.3302	0.46184
0.41506	0.51552	0.06942
0.20034	0.21824	0.58142
0.13951	0.18003	0.68046
0.14029	0.12599	0.73372
0.12876	0.17744	0.6938
0.13442	0.28513	0.58045
0.22589	0.15577	0.61834
0.18861	0.13306	0.67834
0.26028	0.28088	0.45884
0.19831	0.31928	0.48241
0.37797	0.58109	0.04094
0.47982	0.46952	0.05065
0.17888	0.314	0.50712
0.26791	0.17539	0.5567
0.64782	0.3135	0.03868

Au	Cu	Mo
0.22814	0.19547	0.57639
0.47232	0.48404	0.04363
0.21663	0.27648	0.50689
0.207	0.24005	0.55295
0.30235	0.24198	0.45567
0.20662	0.12838	0.665
0.25618	0.3309	0.41292
0.26629	0.29193	0.44178
0.50983	0.45371	0.03645
0.26377	0.25831	0.47792
0.50258	0.46207	0.03536
0.61358	0.3443	0.04212
0.34444	0.32052	0.33504
0.38402	0.1694	0.44658
0.24623	0.25965	0.49412
0.26424	0.17672	0.55904
0.42291	0.14872	0.42837
0.2371	0.18903	0.57387
0.27013	0.17273	0.55714
0.51564	0.45322	0.03113

A ternary diagram can show these results (Fig. 1).

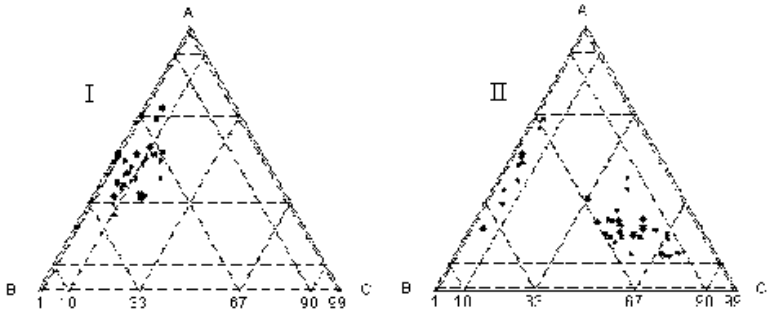

Figure 1. Ternary diagram of the studied data. To the left the CLR transformed data (I), at the right (II) the same data, but centered.

It is clear that even centering does not solve the "problem" of the b.l.d. data, which remain grouped along the AB axis.

2. Zero, zero… What shall I do with you?

Geologists, even those that are not knowledgeable of compositional data analysis, have being dealing with these problems for quite some time (Kashdan *et al*., 1979). One of the most frequently use technique is amalgamation (Aitchison, 1986). Amalgamation, e.g. adding Na_2O and K_2O, as total alkalis is a solution, but sometimes we need to differentiate between a sodic and a potassic alteration, and therefore amalgamation is not an option.

Pre-classification into groups is another solution, but it requires a good knowledge of the distribution of the data and the geochemical characteristics of the groups which is not always available.

Considering the zero values equal to the limit of detection of the used equipment, or substituting it by some other constant (e.g. half the limit of detection) will generate spurious distributions, especially in ternary diagrams as we show in Fig. 1.

Same situation will occur if we replace the zero values by a small amount (Bacon-Shone, 2003) using non-parametric or parametric techniques (imputation). Even if we add the same small value to all of the analyzed parameters, we will get the same spurious distribution.

3. How do I deal with spurious distributions?

The method that I am proposing takes into consideration the existence of well known relationships between some elements. For example, in copper porphide deposits, there is always a clear dependency between the copper values and the molybdenum ones, but while copper will always be above the limit of detection, many of the molybdenum values will include b.d.l. values ("rounded zeros").

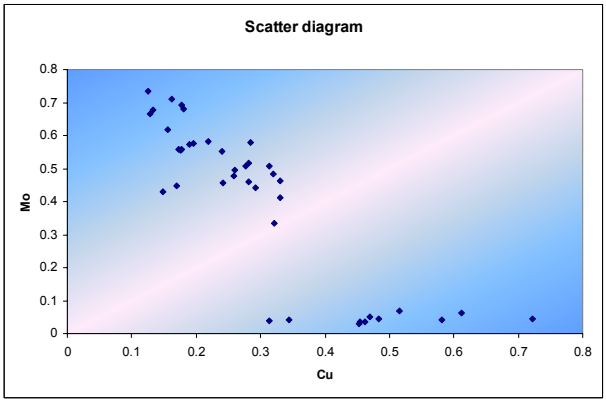

Figure 2. As in all Cu porphyde deposits, there is a strong correlation between Cu and Mo values. We can use such correlation to estimate the values b.d.l. for Mo.

In this case, I will take the lower quartile of the real molybdenum values (Table 2) and establish a regression equation with copper, and then we will estimate the "rounded" zero values of molybdenum by their corresponding copper values (Table 3).

Table 2. Values of the lower quartile of real data for Mo from this study.

Au	Cu	Mo
0.41506	0.51552	0.06942
0.34444	0.32052	0.33504
0.25618	0.3309	0.41292
0.42291	0.14872	0.42837
0.26629	0.29193	0.44178
0.38402	0.1694	0.44658
0.30235	0.24198	0.45567
0.26028	0.28088	0.45884
0.20796	0.3302	0.46184
0.26377	0.25831	0.47792

Table 3. Results of the regression analysis for the lower quartile of real molibdenum data.

Regression Statistics	
Multiple R	0.792759158
R Square	0.628467082
Standard Error	0.079131742
Observations	10

Significance F
0.006231332

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	0.084737701	0.084737701	13.53241
Residual	8	0.050094661	0.006261833	
Total	9	0.134832361		

	Coefficients	*Std. Error*	*t Stat*	*P-value*
Intercept	0.674594109	0.079027652	8.536178017	2.73E-05
X Variable 1	-0.954714886	0.259529113	-3.678642738	0.006231

So, according to Table 3, we could use regression Equation (1) in order to estimate the b.d.l. values for Mo.

(1) Mo = 0.675 – 0.955*Cu

Figure 3 shows that the obtained results are close to the predicted line.

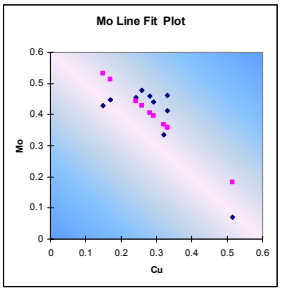

Figure 3. Molybdenum line fit plot showing a good correlation with the predicted values (in pink).

So, did we get ride of the spurious effect?

As Figure 4 clearly shows, only one value of Mo was really close to zero, while the rest has now a value that is a geological reflection of the geochemical characteristics of the data.

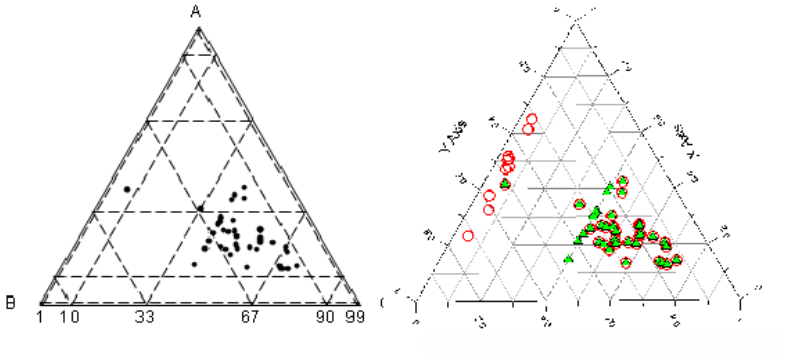

Figure 4. The ternary diagram of the left clearly shows that only one sample has really low value of Mo. The ternary diagram to the right compares the original data in red with the new estimated values of Mo in green.

4. Conclusions and recommendations

One of the main advantages of this method is that we do not obtain a fixed value for the "rounded zeros", but one that depends on the value of the other variable.

The proposed method depends on the geological characteristics of the data, and therefore is less bios or random than other methods. It also presents a viable alternative to amalgamation and an effective way to deal with "Essential zeros" in a population.

The sequence of the method is as follows:

1. We transform the data using CoDaPack or other similar software.
2. We select the lower quartile of real data for the element with the b.d.l. values.
3. Within this dataset, we test the relationship between the element with the b.d.l. values with one (or more) element without b.d.l. values. In most cases, these elements will correspond with well established geological relationship like between Pb and Zn on polymetallic deposits, or between Au and Pb in hydrothermal deposits, or between Cu and Mo in porphyritic deposits, as in the case I presented here.
4. We establish the regression equation.
5. We then substitute the b.d.l. values by those estimated with the obtained equation of regression.

The method could be applied to any type of data, provided we establish first their correlation dependency.

I would also like to recommend that this method will be included as an option for dealing with zeros in the next version of CoDaPack.

Appendix A. Original data

Au	Cu	Mo
30.34	93.90	0.50
31.00	58.11	0.50
35.13	43.69	0.50
54.25	83.52	0.50
55.66	54.54	0.50
63.61	65.27	0.50
82.17	73.23	0.50
83.52	76.90	0.50
85.59	48.09	0.50
97.32	85.66	0.50
98.40	47.68	0.50
34.42	54.73	6.50
85.83	79.98	7.10
32.81	45.92	7.17
49.67	53.67	7.45
95.67	33.69	8.25
68.20	54.66	8.75
35.87	39.13	8.86
89.56	39.56	8.86
71.16	92.04	9.76

Au	Cu	Mo
56.36	36.95	9.97
50.71	81.76	10.50
45.15	31.18	10.52
78.65	86.34	11.10
82.55	80.95	12.73
59.41	50.97	12.77
66.49	84.97	13.24
55.77	98.03	13.45
46.91	33.14	14.36
36.92	47.71	15.33
32.23	41.32	15.43
68.11	79.09	15.48
42.28	89.81	15.54
91.04	96.13	15.55
92.81	62.16	16.71
37.92	34.11	16.88
97.10	62.18	17.04
68.31	42.50	18.71
41.98	57.93	19.25
96.95	77.40	19.97

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman & Hall.

Aitchison, J. and J.W. Kay. (2003). Possible solutions of some essential zero problems in compositional data analysis. *Proceedings of Compositional Data Analysis Workshop, 2003*.

Bacon-Hone, J. (2005). Modelling structural zeros in compositional data. *Proceedings of Compositional Data Analysis Workshop, 2003*.

Kashan, A.B., Guskov, O.I. and A.A. Shimonsky. (1979). *Mathematical modeling in prospection (original in Russian)*. Moscow: Nedra.

Trusova, I. F and V. I. Chernov. (1982). *Petrography of magmatic and metamorphic rocks (original in Russian)*. Moscow: Nedra.