Mathematical Geology, Vol. 37, No. 7, October 2005 (© 2005)
DOI: 10.1007/s11004-005-7374-8

Some Basic Concepts of Compositional Geometry[1]

R. Tolosana-Delgado,[2] N. Otero,[3] and V. Pawlowsky-Glahn[2]

KEY WORDS: closed data, compositions, geochemistry, spurious correlation, ternary diagram.

The object of this short note is to give a synthetic overview of basic concepts of compositional geometry, which are necessary to follow the different analysis presented in subsequent papers in this issue of *Mathematical Geology*. To put them in a proper context, and to understand how to compute and operate with them, it is convenient to introduce them with simple examples. Let us take the major ion content of a groundwater sample and assume one observation has, among others, a content of $\text{Na}^+ = 26.17\,\text{ppm}$, $\text{Ca}^{2+} = 75.83\,\text{ppm}$, and $\text{Mg}^{2+} = 9.22\,\text{ppm}$. To represent these values in a *ternary diagram* (or *three-part simplex* in a general terminology), we have to normalize or *constrain* them:

$$\begin{aligned}[\text{Na}^+, \text{Ca}^{2+}, \text{Mg}^{2+}] &= \frac{100}{26.17 + 75.83 + 9.22}[26.17, 75.83, 9.22]\\ &= [23.5, 68.2, 8.3]\%.\end{aligned}$$

We call this the *closure operation*, $\mathcal{C}[\cdot]$, and write

$$[\text{Na}^+, \text{Ca}^{2+}, \text{Mg}^{2+}] = \mathcal{C}[26.17, 75.83, 9.22] = [23.5, 68.2, 8.3]\%. \qquad (1)$$

After closure, the three parts sum to a constant κ, in this case $\kappa = 100$. This approach takes $[26.17, 75.83, 9.22]$ ppm and $[23.5, 68.2, 8.3]\%$ as equivalent numerical representations of the three-part composition $[\text{Na}^+, \text{Ca}^{2+}, \text{Mg}^{2+}]$, because the *relative* content of the ions in the sample is the same, independent of the units used. This is important, as our approach is based on the assumption that we are

[1] Received 7 February 2005; accepted 28 February 2005.

[2] Departament d'Informàtica i Matemàtica Aplicada, Universitat de Girona, Girona, Spain; e-mail: raimon.tolosana@udg.es.

[3] Departament de Cristal·lografia, Mineralogia i Dipòsits Minerals, Facultat de Geologia Universitat de Barcelona, Barcelona, Spain; e-mail: notero@ub.edu.

0882-8121/05/1000-0673/1 © 2005 International Association for Mathematical Geology

only interested in the relative weight of each component, or that this is the only information we have. $[\mathrm{Na}^+, \mathrm{Ca}^{2+}, \mathrm{Mg}^{2+}]$ can be viewed as a *composition* of three parts, or as a *subcomposition* (or closed subvector) of the whole composition of major ion content of the groundwater sample concerned.

Consider now the bulk major oxide composition of a sediment sample $[\mathrm{Al_2O_3}, \mathrm{CaO} + \mathrm{Na_2O}, \mathrm{K_2O}] = [63.8, 22.7, 13.5]\%$. Assume that the material is exposed to weathering, i.e. altered and removed, and that, after a certain period of time, only half of the $\mathrm{Al_2O_3}$ content is left over. Similarly, only a proportion of 0.17 of $\mathrm{CaO} + \mathrm{Na_2O}$ and 0.33 of $\mathrm{K_2O}$ remain at the site. After the process has taken place, the composition of the sample is

$$\begin{aligned}[\mathrm{Al_2O_3}, \mathrm{CaO} + \mathrm{Na_2O}, \mathrm{K_2O}]_p &= \mathcal{C}\,[63.8 \times 0.5, 22.7 \times 0.17, 13.5 \times 0.33]\\ &= \mathcal{C}\,[31.9, 3.86, 4.45]\\ &= [79.3, 9.6, 11.1]\%.\end{aligned}$$

This operation models the change in a composition. This fundamental operation is known as *perturbation* (Aitchison, 2003) and is denoted with $\oplus$. We write

$$\begin{aligned}[\mathrm{Al_2O_3}, \mathrm{CaO} + \mathrm{Na_2O}, \mathrm{K_2O}]_p &= [63.8, 22.7, 13.5] \oplus [0.5, 0.17, 0.33]\\ &= [79.3, 9.6, 11.1]\%. \end{aligned} \tag{2}$$

In this example, we have used two compositions expressed in different units, percentages and proportions, but the result is independent from it, because it is scale independent. Perturbation is a usual operation in geology: parts are first separately scaled and then closed. It is very useful to better visualize data sets in ternary diagrams (von Eynatten, Pawlowsky-Glahn, and Egozcue, 2002; Pawlowsky-Glahn and Buccianti, 2002).

Moreover, assume that the process is cyclic, and that we are interested in knowing which is the composition after three cycles. The answer is rather simple: we repeat the perturbation operation as many times as necessary. It results in

$$\begin{aligned}&[\mathrm{Al_2O_3}, \mathrm{CaO} + \mathrm{Na_2O}, \mathrm{K_2O}]_{3p}\\ &\quad= [63.8, 22.7, 13.5] \oplus [0.5, 0.17, 0.33] \oplus [0.5, 0.17, 0.33] \oplus [0.5, 0.17, 0.33]\\ &\quad= [63.8, 22.7, 13.5] \oplus [0.5^3, 0.17^3, 0.33^3] = [93.04, 1.30, 5.66].\end{aligned}$$

As can be observed, the percentage of $\mathrm{Al_2O_3}$—the less affected major oxide—increases after three cycles, while the other two decrease. The operation of applying the same perturbation several times is known as *powering* or *power transformation*

(Aitchison, 2003). It is usually denoted either by $\odot$ or by $\otimes$,

$$[0.5, 0.17, 0.33] \oplus [0.5, 0.17, 0.33] \oplus [0.5, 0.17, 0.33] = 3 \odot [0.5, 0.17, 0.33], \tag{3}$$

leading to

$$\begin{aligned}[\mathrm{Al_2O_3}, \mathrm{CaO} + \mathrm{Na_2O}, \mathrm{K_2O}]_{3p} &= [63.8, 22.7, 13.5] \oplus (3 \odot [0.5, 0.17, 0.33]) \\ &= [93.04, 1.30, 5.66].\end{aligned}$$

Note that we can set a generic counter for the number of cycles,

$$[\mathrm{Al_2O_3}, \mathrm{CaO} + \mathrm{Na_2O}, \mathrm{K_2O}]_{3p} = [63.8, 22.7, 13.5] \oplus [\alpha \odot [0.5, 0.17, 0.33]]\,, \tag{4}$$

and then, we can answer questions like: How was the composition one cycle before we took the first measurement? How will it look like after three and a half cycles? The answer can be obtained by setting $\alpha = -1$, respectively $\alpha = 3.5$, and we readily observe that we have modelled a compositional process, as α can be any real number. For further examples see von Eynatten, Barceló-Vidal, and Pawlowsky-Glahn (2003), and Signorelli and others (1998).

Equation (4) represents a compositional process, but how shall we proceed when we do not know which one was the *perturbing* vector? Such a situation is encountered when we have an initial unaltered rock, $\mathbf{z}_0 = [\mathrm{Al_2O_3}, \mathrm{CaO} + \mathrm{Na_2O}, \mathrm{K_2O}]_0 = [0.603, 0.298, 0.099]$, and the final weathered result, $\mathbf{z}_\mathrm{f} = [\mathrm{Al_2O_3}, \mathrm{CaO} + \mathrm{Na_2O}, \mathrm{K_2O}]_\mathrm{f} = [0.955, 0.003, 0.041]$. The answer is again simple: take the vector of *compositional differences*,

$$\delta = \mathbf{z}_\mathrm{f} \ominus \mathbf{z}_0 = \mathbf{z}_\mathrm{f} \oplus (-1 \odot \mathbf{z}_0), \tag{5}$$

to obtain

$$\begin{aligned}\delta &= [0.955, 0.003, 0.041] \oplus (-1) \odot [0.603, 0.298, 0.099] \\ &= [0.955, 0.003, 0.041] \oplus [1/0.603, 1/0.298, 1/0.099] \\ &= \mathcal{C}\left[\frac{0.955}{0.603}, \frac{0.003}{0.298}, \frac{0.041}{0.099}\right] \\ &= \mathcal{C}\,[1.585, 0.011, 0.417] = [0.787, 0.005, 0.207],\end{aligned}$$

and then we can write

$$\mathbf{z}_\mathrm{f} = \mathbf{z}_0 \oplus (1 \odot \delta) = [0.603, 0.298, 0.099] \oplus [0.787, 0.005, 0.207].$$

Note that we handle the operations $\oplus$, $\ominus$ and $\odot$ in the simplex formally like we do with the standard vector operations $+$, $-$ and $\times$ in multidimensional real space, which makes it quite easy to work with them.

Perturbation and powering are mathematically very interesting, as they define on the simplex a *vector space structure*, with perturbation as a commutative or Abelian group operation and powering as the external product. In this context, Equation (4) is the equation of a line, which justifies to call it a *compositional linear process* or a *compositional linear trend.*

Adding to perturbation and powering an inner product, we will have a *real Euclidean space structure* for the simplex (Billheimer, Guttorp, and Fagan, 2001; Pawlowsky-Glahn and Egozcue, 2001, 2002; Aitchison and others, 2002). This might sound rather complicated, but its meaning is simple. An inner product induces a measure of distance and a norm, and then we are able to do *geometry* in the simplex: lines, angles, circles, ellipses, and any geometric element we can think about will be at our disposal. Although mathematically equivalent, to avoid confusion with the usual Euclidean geometry in real space, we talk about the *Aitchison geometry* of the simplex. Inner products are associated with angles, and are needed to compute orthogonal projections. We use them implicitly when we determine e.g. principal components, which are known as *log-contrasts* (Aitchison, 2003) in a compositional framework. An account of the relationship between perturbation, log-contrasts and Singular Value Decomposition can be found in Aitchison and others (2002).

Formally, we are working with a set of points in the strictly positive orthant of D-dimensional real space, which we call the *D-part simplex*,

$$\mathcal{S}^D = \left\{ \mathbf{x} = [x_1, x_2, \ldots, x_D]; x_i > 0; \sum_{i=1}^{D} x_i = \kappa \right\} ,$$

which has two operations defined, *perturbation*,

$$\mathbf{x} \oplus \mathbf{y} = [x_1, x_2, \ldots, x_D] \oplus [y_1, y_2, \ldots, y_D] = \mathcal{C}\,[x_1 y_1, x_2 y_2, \ldots, x_D y_D] ,$$

and *powering*,

$$\alpha \odot [x_1, x_2, \ldots, x_D] = \mathcal{C}\left[x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha\right] ,$$

as well as an *inner product*,

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \sum_{i=1}^{D} \ln \frac{x_i}{g(\mathbf{x})} \ln \frac{y_i}{g(\mathbf{y})}, \tag{6}$$

where $g(\mathbf{x}) = (\prod_{j=1}^{D} x_j)^{1/D} = \sqrt[D]{\prod_{j=1}^{D} x_j}$ stands for the geometric mean.

For illustration on how to compute the inner product (6), consider two three-part compositions

$$\mathbf{x}_1 = [0.787, 0.005, 0.207] \quad \text{and} \quad \mathbf{x}_2 = [0.5, 0.17, 0.33]\,,$$

both closed to 1. First we need to compute the geometric means,

$$g(\mathbf{x}_1) = \sqrt[3]{0.787 \times 0.005 \times 0.207} = 0.095$$

and

$$g(\mathbf{x}_2) = \sqrt[3]{0.5 \times 0.17 \times 0.33} = 0.304\,,$$

then the log-quotients

$$\left[\ln \frac{0.787}{0.095}, \ln \frac{0.005}{0.095}, \ln \frac{0.207}{0.095}\right] = [2.111, -2.888, 0.777]$$

and

$$\left[\ln \frac{0.5}{0.304}, \ln \frac{0.17}{0.304}, \ln \frac{0.33}{0.304}\right] = [0.498, -0.581, 0.083]\,,$$

and finally the inner product

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = 2.111 \times 0.498 + (-2.888) \times (-0.581) + 0.777 \times 0.083 = 2.794.$$

If we consider $\mathbf{x}_1$ and $\mathbf{x}_2$ as two vectors in the three-part simplex $\mathcal{S}^3$, this means that they are not orthogonal, as in that case their inner product should be zero.

The inner product (6) induces a distance in the simplex, which is useful to evaluate the difference or distance between two composition. Recall that distances are also essential to understand variability within a data set, or to use techniques like cluster analysis. The *Aitchison distance*

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{D} \left(\ln \frac{x_i}{g(\mathbf{x})} - \ln \frac{y_i}{g(\mathbf{y})} \right)^2}\,. \tag{7}$$

has nice properties: it takes into account the relative nature of the information we are using, and it is compatible with the basic operations of perturbation and powering, something that is not accomplished by many other distances (Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn, 1998). To compute the distance between $\mathbf{z}_0$ and $\mathbf{z}_\mathrm{f}$ considered in Equation (5) we compute first the geometric means

$$g(\mathbf{z}_0) = \sqrt[3]{0.603 \times 0.298 \times 0.099} = 0.261\,,$$
$$g(\mathbf{z}_\mathrm{f}) = \sqrt[3]{0.955 \times 0.003 \times 0.041} = 0.050\,,$$

substitute in Equation (7),

$$\begin{aligned} d^2(\mathbf{z}_\mathrm{f}, \mathbf{z}_0) &= \left(\ln\frac{0603}{0.261} - \ln\frac{0.955}{0.050}\right)^2 + \left(\ln\frac{0.298}{0.261} - \ln\frac{0.003}{0.050}\right)^2 \\ &\quad + \left(\ln\frac{0.099}{0.261} - \ln\frac{0.041}{0.050}\right)^2 \\ &= (0.836 - 2.947)^2 + (0.132 - 2.756)^2 + (-0.968 + 0.191)^2 \\ &= 13.402, \end{aligned}$$

to obtain, after taking the square root,

$$d(\mathbf{z}_\mathrm{f}, \mathbf{z}_0) = 3.661$$

which, in this case, can be understood as a measure of the degree of alteration.

But not only geometric elements can be used with an Euclidean space structure: using the idea of coordinates with respect to an orthonormal basis, which exist for any Euclidean space, we have also integrals and derivatives in the simplex and thus, the whole battery of statistical methods, developed within the usual framework of multidimensional real space, can be transferred to the simplex (Pawlowsky-Glahn, 2003). We simply apply them to the coordinates. Working with coefficients with respect to an orthonormal basis is also more comfortable, as we can operate directly with the standard operations. Egozcue and others (2003) give such a basis. The coefficients for a D-part composition $\mathbf{x}$ are $D - 1$ and are given by

$$c_i = \left(\frac{1}{\sqrt{i(i+1)}} \ln \frac{x_1 \cdots x_i}{x_{i+1}^i}\right), \quad i = 1, 2, \ldots, D-1. \tag{8}$$

Thus, for the three-part composition $\{\mathrm{Al_2O_3}, \mathrm{CaO} + \mathrm{Na_2O}, \mathrm{K_2O}\} = [63.8, 22.7, 13.5]$, these coefficients are

$$c_1 = \frac{1}{\sqrt{2}} \ln \frac{\mathrm{Al_2O_3}}{\mathrm{CaO} + \mathrm{Na_2O}} = \frac{1}{\sqrt{2}} \ln \frac{63.8}{22.7} = 0.731$$

and

$$c_2 = \frac{1}{\sqrt{6}} \ln \frac{\mathrm{Al_2O_3(CaO + Na_2O)}}{\mathrm{K_2O^2}} = \frac{1}{\sqrt{6}} \ln \frac{63.8 \times 22.7}{13.5^2} = 0.846\,.$$

In the same basis, the compositional process (4), written in terms of coordinates, is given by two log-linear equations,

$$\begin{aligned} \frac{1}{\sqrt{2}} \ln \frac{\mathrm{Al_2O_3}}{\mathrm{CaO + Na_2O}} &= \frac{1}{\sqrt{2}} \ln \frac{63.8}{22.7} + \alpha \frac{1}{\sqrt{2}} \ln \frac{0.5}{0.17} \\ \frac{1}{\sqrt{6}} \ln \frac{\mathrm{Al_2O_3(CaO + Na_2O)}}{\mathrm{K_2O^2}} &= \frac{1}{\sqrt{6}} \ln \frac{63.8 \times 22.7}{13.5^2} + \alpha \frac{1}{\sqrt{6}} \ln \frac{0.5 \times 0.17}{0.33^2}\,, \end{aligned} \tag{9}$$

or, equivalently,

$$\begin{aligned} \ln(\mathrm{Al_2O_3}) - \ln(\mathrm{CaO + Na_2O}) &= \sqrt{2}(0.731 + \alpha\, 0.763) \\ \ln(\mathrm{Al_2O_3}) + \ln(\mathrm{CaO + Na_2O}) - 2\ln(\mathrm{K_2O}) &= \sqrt{6}(0.846 + \alpha(-0, 101)). \end{aligned}$$

Note that the representation in coordinates appears as a transformation of a three-part simplex into a two-dimensional real space. To be precise, this is a morphism which assigns coordinates to each composition. Its main property is that it translates inner products and distances in the Aitchison geometry into inner products and distances in the ordinary real space of coordinates.

As can be seen in (9), these equations involve several log-ratios, that come from the coordinate expressions. These kind of log-linear equations appear in many instances in compositional data analysis, specially when we use techniques like Principal Component Analysis or Factor Analysis. Examples can be found in the subsequent articles in this issue.

To conclude this short note, recall that it is well known, from early work by Pearson (1897) and Chayes (1960), that compositional data present what they called the *spurious correlation effect*. Actually, it means that standard statistical methods applied to raw compositional data might lead to inconsistent results and that, when results seem to be reasonable, we cannot be sure if they are really the best we can get out of our data. A way out to this problem is given by the above log-ratio approach, initiated by John Aitchison back in the 1980s.

REFERENCES

Aitchison, J., 2003, The statistical analysis of compositional data: (Reprint) Blackburn Press, Caldwell, NJ, 416 p.

Aitchison, J., Barceló-Vidal, C., Egozcue, J. J., and Pawlowsky-Glahn, V., 2002, A concise guide for the algebraic–geometric structure of the simplex, the sample space for compositional data analysis, *in* Bayer, U., Burger, H., and Skala, W., eds., Proceedings of IAMG'02—The Eigth Annual Conference of the International Association for Mathematical Geology, Terra Nostro, no. 3, p. 387–392.

Billheimer, D., Guttorp, P., and Fagan, W., 2001, Statistical interpretation of species composition: J. Am. Stat. Assoc., v. 96, no. 456, p. 1205–1214.

Chayes, F., 1960, On correlation between variables of constant sum: J. Geophys. Res., v. 65, no. 12, p. 4185–4193.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C., 2003, Isometric logratio transformations for compositional data analysis: Math. Geol., v. 35, no. 3, p. 279–300.

Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 1998, A critical approach to non-parametric classification of compositional data, *in* Rizzi, A., Vichi, M., and Bock, H.-H., eds., Advances in data science and classification; Springer-Verlag, Berlin, p. 49–56.

Pawlowsky-Glahn, V., 2003, Statistical modelling on coordinates, *in* Thió-Henestrosa, S. and Martín-Fernández, J. A., eds., Compositional Data Analysis Workshop—CoDaWork'03, Proceedings, Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/.

Pawlowsky-Glahn, V., and Buccianti, A., 2002, Visualization and modeling of subpopulations of compositional data: Statistical methods illustrated by means of geochemical data from fumarolic fluids: Int. J. Earth Sci. (Geologische Rundschau), v. 91, no. 2, p. 357–368.

Pawlowsky-Glahn, V., and Egozcue, J. J., 2001, Geometric approach to statistical analysis on the simplex: Stochastic Environ. Res. Risk Assess. (SERRA), v. 15, no. 5, p. 384–398.

Pawlowsky-Glahn, V., and Egozcue, J. J., 2002, BLU estimators and compositional data: Math. Geol., v. 34, no. 3, p. 259–274.

Pearson, K., 1897, Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs: Proc. R. Soc. Lond., v. LX, p. 489–502.

Signorelli, S., A. Buccianti, M. Martini, and G. Piccardi, 1998, Arsenic in fumarolic gases of Vulcano (Aeolian Islands, Italy) from 1978 to 1993: Geochemical evidence from multivariate analysis: Geochem. J., v. 32, p. 367–382.

von Eynatten, H., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 2003, Modelling compositional change: The example of chemical weathering of granitoid rocks: Math. Geol., v. 35, no. 3, p. 231–251.

von Eynatten, H., Pawlowsky-Glahn, V., and Egozcue, J. J., 2002, Understanding perturbation on the simplex: A simple method to better visualise and interpret compositional data in ternary diagrams: Math. Geol., v. 34, no. 3, p. 249–257.