Is compositional data analysis a way to see beyond the illusion?

Computers & Geosciences 50 (2013) 165–173

Contents lists available at SciVerse ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Is compositional data analysis a way to see beyond the illusion?

Antonella Buccianti

University of Florence (I), Department of Earth Sciences, Via G. La Pira 4, 50121 Florence (I), Italy

ARTICLE INFO

Article history:
Received 30 April 2012
Received in revised form
13 June 2012
Accepted 14 June 2012
Available online 28 June 2012

Keywords:
Compositional data analysis
Geochemistry
Worldwide lithology
River chemistry
Volcanic gases
Isometric log-ratio transformation

ABSTRACT

Notwithstanding the numerous contributions that have been published on theoretical and practical aspects of the management of compositional (constrained) data during the last thirty years, in geochemistry most of the scientific papers in international journals continue to ignore their peculiar features. In order to understand the reasons of the undervaluation of the effects of an incorrect choice of the sample space and, consequently, an incorrect application of the distance concept, case studies of comparison between methodologies will be presented and discussed. The aim is to evaluate the differences in interpretation of geochemical processes affecting rocks, water and gaseous samples, when the two different approaches, classical and compositional, are adopted. If we compare the results of case studies following the two paths it is possible to evaluate which type of error (and consequences) will affect our evaluations in geochemistry.

The presence of expected differences between the two approaches indicates that compositional data analysis can be a way to see beyond the illusion due to the constrained space. However, the possibility that the difference is tenuous in some situations, not revealed a priori, may be at the origin of the unconscious choice of the classical approach. Is this condition which some researchers call "common sense" frequently encountered in geochemistry? The paper is aimed to try to answer the proposed question, and to understand the difficulty of diffusion of compositional data analysis even if now simple tools of investigation, for different degrees of knowledge, are available.

1. Introduction

The development of graphical and numerical tools to perform compositional data analysis (CoDA) represents a benchmark problem in geological sciences and, in particular, in computational geochemistry. This discussion has long animated researchers working in different fields and, notwithstanding that appropriate tools are now available to correctly investigate the features of compositional data, most of the published papers in the international literature avoid facing this attractive question. Moreover, people that are interested in managing compositional data in a consistent[1] mathematical and statistical framework, often have to convince referees of the appropriateness of their methodology. Answers such as "nature is stronger than closure" or "a good geologist is able to recognise forced relationships from the natural one" may be typical. Why is it so difficult to convince the scientific community about the adoption of the CoDA approach? Are the expected differences between results obtained by using classical and compositional approaches able to convince us that a way to see beyond the illusion due to the constrained space is to take into account its geometry? Or the differences are usually so tenuous, that the unconscious choice of the classical approach is recognised as "common sense"? The paper is aimed to try to answer the proposed question and to understand the difficulty of diffusion of compositional data analysis even if now simple tools are available.

The first benchmark of the compositional data problem can be considered the presence of the *spurious correlations* recognised by Pearson (1897) affecting all data that measure parts of the same whole, such as percentages, proportions, ppm and so on. His work represented the first approach able to recognise that if X, Y and Z are uncorrelated, then X/Z and Y/Z ratios, will not be uncorrelated. Chayes (1960) found a mathematical demonstration of Pearson's work and showed that some of the correlations between the components of the composition must be negative due to the sum constraint, thus affecting interpretation of natural processes, biased by this effect.

Even if natural data are often non-negative, ranging in a sample space with a restriction to R_+^D (they are only positive, and variables move only on the positive parts of real space), compositional data have a further restriction since they have been

E-mail address: antonella.buccianti@unifi.it

[1] In the sense of coherent with the principles of compositional data analysis (Egozcue and Pawlowsky-Glahn, 2011) that is: (a) The relative character of the information carried by the data is taken into account, (b) The models that are compatible with the sample space and constraints do not need to be taken additionally into account.

0098-3004/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cageo.2012.06.012

scaled by the total of the components of the composition. This important operation of standardization is fundamental in interpreting and comparing results obtained by experimental measures in geochemistry, since data are related to the same weight (solid matrices) or volume (solutions and gaseous mixtures). The consequence is that compositional data with D components not only pertain to the positive part of R^D but occupy a restricted part of its axes, in general from 0 to the constant defined *a priori*. In mathematical terms compositional data are represented as pertaining to a sample space called the simplex S^D:

$$S^D = \left\{ \boldsymbol{x} = (x_1, x_2, x_D) : x_i > 0 \ (i = 1,2,D), \sum_{i=1}^{D} x_i = \kappa \right\} \quad (1)$$

where κ is a given positive constant, defined *a priori* and depending on how the parts are measured. The key question here is not to discuss the nature of compositional data, since they have to be defined exactly in this way if we want to make comparisons among cases, starting from the same baseline. The key question is whether standard statistical analysis which assumes that the sample space is R^D with D dimensions, where all the values from $\pm\infty$ are generable and have a probability to be found in a sampling process, is appropriate to represent the investigated phenomena. In other words, in an olivine, a silicate mineral represented by the formula (Mg, Fe)SiO_2, it is known that Mg and Fe substitute each other and that a rigorous stoichiometric law governs the process characterised by the competition of two ions for the same crystallographic site. It is also clear in this framework that when Mg decreases, Fe tends to increase (common sense). However, the classical approach, that simply represents this phenomenon in a binary diagram, where abundance of Fe and Mg are analysed in respect to each other, or versus Si considered a common base, does not represent a coherent geometry on which to base statistics, both descriptive and inferential, and to propose models able to indicate how natural phenomena work. In fact, considering compositional data as real data, the hypothesis that it is possible to obtain negative contents of Fe and Mg in a sampling process is considered as feasible. The formulation of this hypothesis is frequently performed, even if unconsciously, when a correlation coefficient is determined, or some modelling of the linear pattern of the data on binary diagrams is proposed. The probability of a negative concentration may be low, but it is not possible to know its value in advance and this also affects the determination of simple central tendency statistics (mean) and variability measures (variance). Consequently, as reported in Aitchison et al. (2000), it should be obvious that with compositional data only the statements about the ratios of the components are meaningful, since their use respects the fundamental principle of *scale invariance*. This item for a long time was indirectly recognised in geochemistry as is testified by the common use of ratio diagrams. However, often these diagrams were realised considering ratios with the same denominator, for example X/Z and Y/Z ratios that, as reported by Pearson (1897) were affected by spurious correlations.

In the statistical literature there is a long history of the search for a solution to the statistical analysis of compositional data (Aitchison and Egozcue, 2005). The main contribution to a solution is attributable to John Aitchison in the early 1980s (Aitchison, 1982) when he introduced the log-ratio approach using the intuitive concept of difference associated with the features of data. For example, the log-ratio approach was proposed to capture the difference between 5% and 10% and that between 45% and 50%, difference equal to 5 in both cases in the Euclidean real space. Following this approach, compositions are transformed to move them into real space using a log-ratio transformation, analysed by classical statistical methods, and results reported back to the simplex, by using the correspondent inverse transformation. A further key step was the recognition of the Euclidean space nature of the simplex (Pawlowsky-Glahn and Egozcue, 2001). In this framework compositions can be represented by their coordinates in the simplex with a suitable orthonormal basis, leading to the *ilr* (isometric log-ratio) transformation (Egozcue et al., 2003). Its use allow us to avoid the arbitrariness of denominator choice related to the *alr* (additive log-ratio) transformation and to the singularity of the *clr* (centered log-ratio) transformation, the two transformations originally proposed by Aitchison (1982).

In this paper the comparison of the results obtained for some interesting compositional cases investigated by using the classical and the log-ratio approach allows us to verify how the illusion to see compositional data as real data may compromise our understanding of natural phenomena. To achieve this aim, the *ilr* transformation was used to represent a composition as a real vector. Even if the computation of *ilr* coordinates appears to be complex, there are different rules on how to generate them (Egozcue et al., 2003). The identification of balances, a particular form of *ilr* coordinates (Egozcue and Pawlowsky-Glahn, 2005) may simplify the adoption of this transformation. Balances, reflecting the relative variation of two groups of parts, represent a powerful tool for researchers to prove their geochemical hypothesis, translating ideas on natural phenomena in numbers moving in a coherent geometry. Balances in fact define coordinates of the samples within an orthogonal system of axes, i.e., they are usual random variables in real space.

2. Working on coordinates: the *ilr* transformation of compositional data and the balances approach

In statistics the real space R^k (k=number of dimensions) is assumed to be the natural sample space for a given set of observations. Standard statistics have been developed in R^k using its particular algebraic–geometric structure, which is commonly known as Euclidean geometry. Linear algebra allows us to translate standard statistics into any sample space, different from R^k, if it has an Euclidean vector space structure. Definition of basic operations in the simplex such as perturbation and powering, with the associated norm and distance, permits us to analyse data (Aitchison, 2001; Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001). In this framework, the procedure of the sequential binary partition to identify orthonormal coordinates, can be adopted (Egozcue and Pawlowsky-Glahn, 2005). In a first step the parts of the composition are divided into two groups: the parts of the first group are coded by $+1$ and the parts of the second group are coded by -1. In this way the first coordinate describing the balance between the $+1$ and -1 parts is obtained. In the second and following steps a previous group of parts is divided into new groups, similarly coded by $+1$ and -1 while the components that are not involved are coded with a zero. The number of steps required for all the groups to contain a single component is exactly $D-1$, dimensions of S^D. The whole procedure can be summarised in a table as reported in Egozcue and Pawlowsky-Glahn (2005). From a general point of view, in the kth step the balance z_k (Eq. (2)) between two groups is obtained so that the r_k ($+1$) parts are balanced with the s_k (-1) parts:

$$z_k = \sqrt{\frac{r_k s_k}{r_k + s_k}} \ln \frac{(x_{i1} x_{i2} x_{ir_k})^{1/r_k}}{(x_{j1} x_{j2} x_{js_k})^{1/s_k}}, \quad k = 1, D-1 \quad (2)$$

or:

$$z_k = \sqrt{\frac{rs}{r+s}} \ln \frac{g_m(\boldsymbol{x}_+)}{g_m(\boldsymbol{x}_-)}, \quad (3)$$

where $g_m(.)$ is the geometric mean of the parts reported in the argument in parenthesis (Eq. (3)). Balances may have a relatively easy interpretation as they are log-ratios of geometric means of groups of parts. Furthermore they can well describe processes of element partition, or groups of elements, when they present coherent (similar) behaviour in geochemistry (Buccianti, 2011a; Buccianti, 2011b). The identification of easily interpretable balances may also be useful to reveal natural laws governing proportions in several field of earth (Buccianti and Esposito, 2004; Nisi et al., 2008; Buccianti et al., 2009) and ecological sciences (Bertocchi et al., 2008; Tricarico et al., 2008). Although the sequential binary partition can be performed following rigorous mathematical steps (Egozcue et al., 2003, Hron et al., 2010), here in the case studies analysed, the choice of the balances starts from a geochemical hypothesis. Statistical analysis will be used to model the data variability. In this way a comparison with the usual plots used in geochemistry can be performed as well as a discussion about the results obtained.

3. Another look at the worldwide distribution of continental rock lithology

The weathering processes affecting silicate rocks, and the formation of carbonate rocks in the ocean, are the mechanisms of transfer of CO_2 from the atmosphere to the lithosphere, modifying the carbon cycle. This CO_2 uptake on the geological timescale is mainly controlled by the chemical properties of the rocks. The worldwide distribution of continental rock lithology and the implications for the atmospheric/soil CO_2 uptake processes have been investigated in Amiotte Suchet et al. (2003). On the other hand, investigation of sites locally characterised by high natural flux of CO_2 can be found in Tassi et al., 2009.

The flux of atmospheric/soil CO_2 consumed by rock weathering is generally considered mainly a function of runoff and of the rock type drained by surface water. If streams drain silicate rocks, the flux is considered equal to the whole alkalinity flux while if streams drain carbonate rocks the half of the alkalinity flux has to be considered. In river water bicarbonates (HCO_3^-) can be assumed to be equal to the alkalinity (Meybeck, 1987). In general, linear trends represent the relationships between the CO_2 flux (alkalinity) and runoff, for different type of rocks (Amiotte Suchet et al., 2003).

Few studies attempt to evaluate the abundance of various rock types on the continents and the estimates improved by the geological knowledge and the data that became available during the last century. The first data are attributable to Clarke (1924) who proposed that continental outcrops were composed of 75% of sedimentary rocks and of 25% of combined igneous and metamorphic rocks. Subsequent estimates established that the land surfaces were composed of 8% of extrusive crystalline rocks, 9% of intrusive crystalline rocks, 17% of metamorphic and Precambrian crystalline rocks and 66% of sedimentary rocks (Blatt and Jones, 1975). The results were then refined by Meybeck (1987), even if the first analysis of the spatial distribution of rocks is attributable to Bluth and Kump (1991), followed by the work of Gibbs and Kump (1994).

In the paper of Amiotte Suchet et al. (2003) the lithological composition of 39 large river basins, runoff and alkalinity (observed and calculated by modelling) have been analysed to quantify the flux of consumed CO_2 by weathering (10^3 mol/km^2/year). The categories, reflecting the chemical composition and the behaviour of rocks with regard to the chemical weathering, are given by sands and sandstones, shales, carbonate rocks, shield rocks (igneous and metamorphic rocks), and acid volcanic rocks (Amiotte Suchet and Probst, 1995; Amiotte Suchet et al., 2003).

In this paper another look at the analysis of the relationships between the lithological compositions of 39 major river basins of the world (in percent of the basin area) and the observed alkalinity (10^3 mol/km^2/year) is proposed. Data used in this analysis has been taken from Amiotte Suchet et al., 2003. The map of the most important major watersheds of the world, including the 39 basins here analysed, is reported in Fig. 1 (Revenga et al., 1998).

This approach, based on the use of the balances, or particular *ilr* transformation of lithological proportions, has the advantage of being easy, it can be based on geochemical hypothesis and allows us to perform a coherent classical statistical analysis.

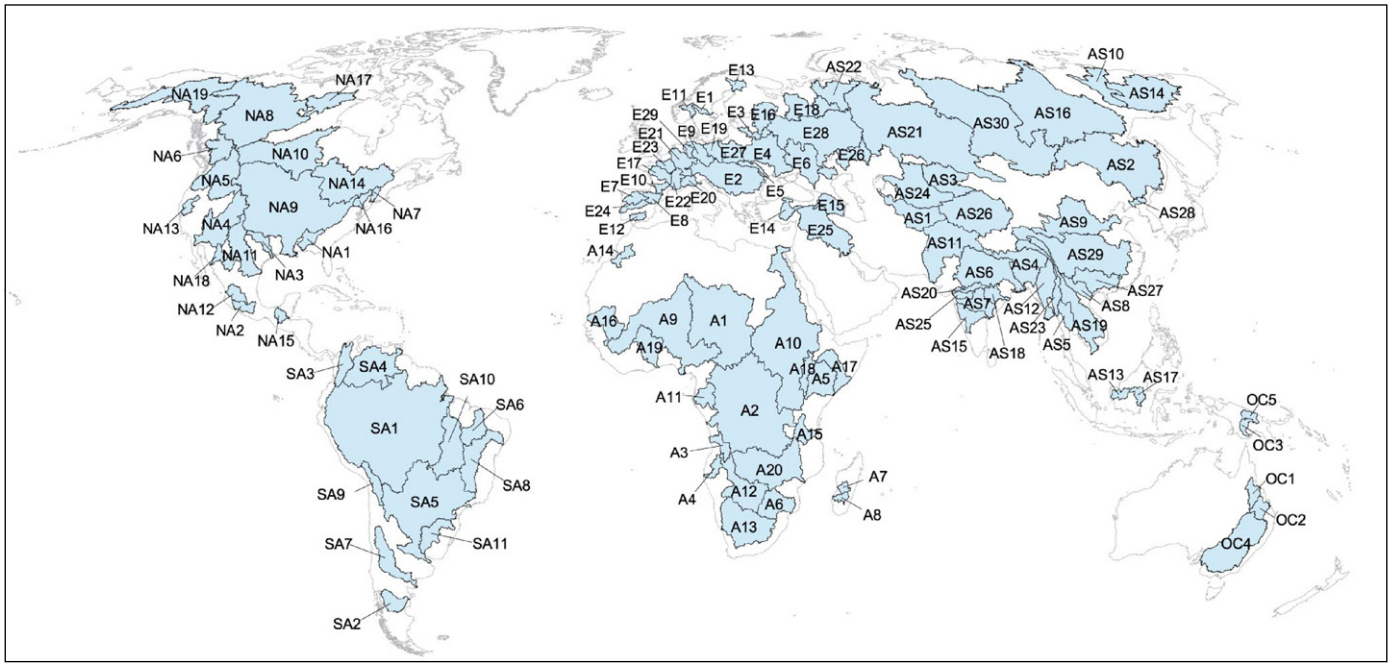
The six lithological categories can be reduced to three by considering their response to weathering processes as obtained in Amiotte Suchet et al. (2003): (a) easily weathered rocks (EWR, carbonates, high CO_2 consumption rates, >200 10^3 mol/km^2/year), (b) intermediate rocks (*IR*, basalts and shales, moderate CO_2 consumption rates, 50–100 10^3 mol/km^2/year) and (c) weathering resistant rocks (*WRR*, shield rocks, sands and sandstones, low CO_2 consumption rates, <50 10^3 mol/km^2/year). The balances (*ilr* coordinates) to be used in our considerations, by taking into account the sequence *EWR* (carbonates)→*IR* (basalts and shales)→*WRR* (shield rocks, sands and sandstones) can be:

$$ilr\_1 = \sqrt{\frac{2}{3}}\ln\frac{EWR}{\sqrt{IR \times WRR}}, \quad ilr\_2 = \frac{1}{\sqrt{2}}\ln\frac{IR}{WRR} \tag{4}$$

With the previous coordinates simple statistical analysis can be performed to analyse the relationships among the components of the composition in their sample space without distortion. They can also be used to evaluate the presence of relationships with other variables as for example observed alkalinity (Amiotte Suchet et al., 2003). In this framework, the first analysis concerns the probability plots reported in Fig. 2. The aim is to verify if the coordinates follow the Gaussian model, if there is a sort of equilibrium around a barycentre, or if the presence of anomalous values perturbs the distribution of the data. As we can see, distribution for *ilr*_1 is perturbed by a set of data characterised by low values, correspondent to the scarce presence of EWR (Eq. 4) while *ilr*_2 shows a more equilibrated distribution.

Graphical tools can be associated with hypothesis tests on the normality. Results of the tests for *ilr*_1 are controversial, as expected considering Fig. 2, while for *ilr*_2 the normality hypothesis can be accepted ($p \gg 0.05$). If the set of data characterised by low *ilr*_1 values are excluded, normality can be accepted with reasonable p values ($p \gg 0.05$).

Results indicate that on a global scale an equilibrium in the proportion of the three different typologies of rocks in the major river drainage basins can be found if some extreme conditions are not considered. In this case, the mean value of *ilr*_1 coordinate, approximately equal to -0.68, allows us to say that on the whole the proportion of easily weathered rocks is lower than the geometric mean which expresses the relative balance between the intermediate and resistant rocks. The *ilr*_1 value equal to -0.68 corresponds to a relationship similar to $EWR/(IR \times WRR)^{1/2}$ equal to 0.43 (or approximately 1:2). Analogous analysis can be performed for the *ilr*_2 coordinate, with mean value equal to -0.55 describing the balance between intermediate and weathering resistant rocks for which IR/WRR=0.58 (or approximately 1:2). The investigation of the frequency distribution of the *ilr* coordinates associated with the use of statistical tests, has allowed us to find a first approximation of the barycentre among the proportions of the different lithologies. This datum was critically proposed by Amiotte Suchet et al. (2003) considering raw percentages, due to the high variability from one drainage basin to another. However, the choice of tools consistent with the compositional nature of the data has increased our capacity to



Africa

A01 Lake Chad
A02 Congo
A03 Cuanza
A04 Cunene
A05 Jubba
A06 Limpopo
A07 Mangoky
A08 Mania
A09 Niger
A10 Nile
A11 Ogooue
A12 Okavango
A13 Orange
A14 Oued Draa
A15 Rufiji
A16 Senegal
A17 Shaballe
A18 Turkana
A19 Volta
A20 Zambezi

Europe

E01 Dalalven
E02 Danube
E03 Daugava
E04 Dnieper
E05 Dniester (Nistru)
E06 Don
E07 Duero
E08 Ebro
E09 Elbe
E10 Garonne
E11 Glomma-Laagen
E12 Guadalquivir
E13 Kemijoki
E14 Kizilirmak
E15 Kura-Araks
E16 Lake Ladoga
E17 Loire
E18 North Dvina
E19 Oder
E20 Po
E21 Rhine & Maas
E22 Rhone
E23 Seine
E24 Tagus
E25 Tigris & Euphrates
E26 Ural
E27 Vistula
E28 Volga
E29 Weser

Asia

AS01 Amu Darya
AS02 Amur
AS03 Lake Balkhash
AS04 Brahmaputra
AS05 Chao Phraya
AS06 Ganges
AS07 Godavari
AS08 Hong (Red River)
AS09 Huang He (Yellow River)
AS10 Indigirka
AS11 Indus
AS12 Irrawaddy
AS13 Kapuas
AS14 Kolyma
AS15 Krishna
AS16 Lena
AS17 Mahakam
AS18 Mahanadi
AS19 Mekong
AS20 Narmada
AS21 Ob
AS22 Pechora
AS23 Salween
AS24 Syr Darya
AS25 Tapti
AS26 Tarim
AS27 Xun Jiang
AS28 Yalu Jiang
AS29 Yangtze
AS30 Yenisey

North & Central America

NA01 Alabama & Tombigbee
NA02 Balsas
NA03 Brazos
NA04 Colorado
NA05 Columbia
NA06 Fraser
NA07 Hudson
NA08 Mackenzie
NA09 Mississippi
NA10 Nelson
NA11 Rio Grande
NA12 Rio Grande de Santiago
NA13 Sacramento
NA14 Saint Lawrence
NA15 San Pedro & Usumacinta
NA16 Susquehanna
NA17 Thelon
NA18 Yaqui
NA19 Yukon

South America

SA01 Amazon
SA02 Chubut
SA03 Magdalena
SA04 Orinoco
SA05 Parana
SA06 Parnaiba
SA07 Rio Colorado
SA08 Sao Francisco
SA09 Lake Titicaca & Salar de Uyuni
SA10 Tocantins
SA11 Uruguay

Oceania

OC01 Burdekin-Belyando
OC02 Dawson
OC03 Fly
OC04 Murray-Darling
OC05 Sepik

Fig. 1. Location of 106 major watersheds of the world. Omitted regions, shown in white, are primarily smaller coastal drainage basins or regions with no permanent rivers (Revenga et al., 1998).

synthesize the information contained in the analysed system. If the analysed data set is exhaustive, to explain the results on a global scale we can say that for one part of *EWR*, two parts with different proportions in IR and WRR are needed while for one part of *IR* two parts of *WRR* are required.

The obtained *ilr* coordinates, representing a mathematical tool where the relative proportions of the different rock categories are represented, can be used for further modelling since they are real variables to be used in various statistical analysis. In our case the relationships between the *ilr* coordinates and observed alkalinity have been investigated. The aim was to verify if they are able to explain the proportional contribution of lithology to the CO_2 uptake. Data on the observed alkalinity are from Amiotte Suchet et al., 2003, and they were transformed by using natural logarithm to take into account both the nature of the sample space (R_+) and the relative character of the information.

The diagrams of the *ilr* coordinates and the log-transformed observed alkalinity are reported in Figs. 3 and 4. Rivers have been

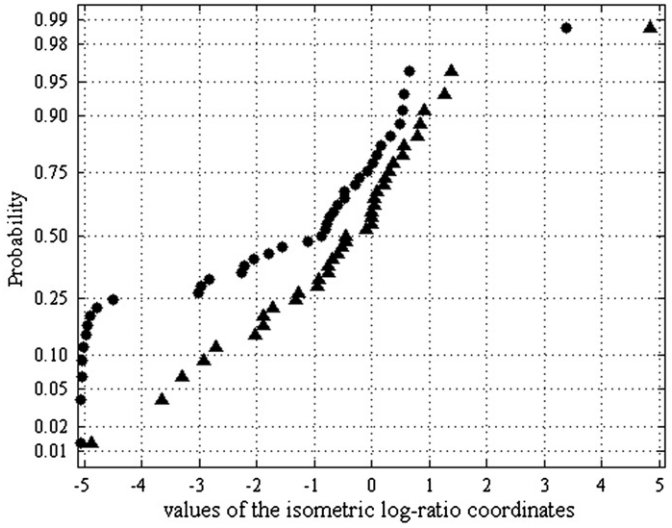


Fig. 2. Normal probability plots of *ilr*_1 (black circles) and *ilr*_2 coordinates (black triangles).

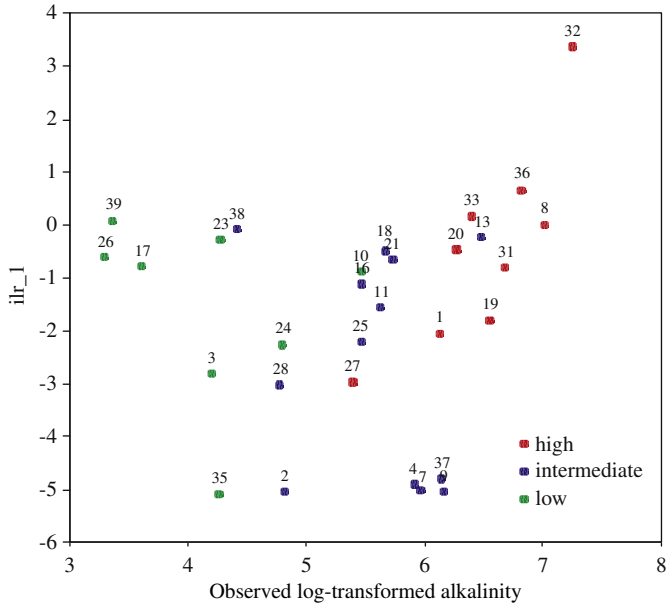


Fig. 3. Relationship between the *ilr*_1 coordinate and observed log-transformed alkalinity (10^3 mol/km^2/year); data have been discriminated by taking into account the CO_2 consumption rates (high $>200\times10^3$ mol/km^2/year; moderate $50–100\times10^3$ mol/km^2/year; low $<50\times10^3$ mol/km^2/year; Amiotte Suchet et al., 2003); labels: 1=Amazon; 2=Amour; 3=Colorado; 4=Columbia; 5=Danube; 6=Don; 7=Fraser; 8=Ganges–Brahmaputra; 9=Godavari; 10=Huangho; 11=Yenisei; 12=Indigirka; 13=Indus; 14=Irrawadi; 15=Kolyma; 16=Lena; 17=Limpopo; 18=Mackenzie; 19=Magdalena; 20=Mekong; 21=Mississippi; 22=Murray; 23=Niger; 24=Nile; 25=Ob; 26=Orange 27=Orinoco; 28=Parana; 29=Sao Francisco; 30=Senegal; 31=Severnaia Dvina; 32=Si Kiang; 33=St. Lawrence; 34=Tigris–Euphrates; 35=Yana; 36=Yangtze-Kiang; 37=Yukon; 38=Zaire; 39=Zambesi.

discriminated by considering the CO_2 consumption rates (high $>200\times10^3$ mol/km^2/year; moderate $50–100\times10^3$ mol/km^2/year; low $<50\times10^3$ mol/km^2/year) as reported in Amiotte Suchet et al., 2003. In the case of *ilr*_1 coordinate with the exception of two groups of data (17=Limpopo, 23=Niger, 26=Orange, 38=Zaire, 39=Zambesi, a group of African rivers in both tropical wet dry and humid tropical conditions, or 2=Amour, 4=Columbia, 7=Fraser, 9=Godavari, 35=Yana, 37=Yukon, mainly related to periglacial and dry continental conditions, with the exception of Godavari whose climatic setting

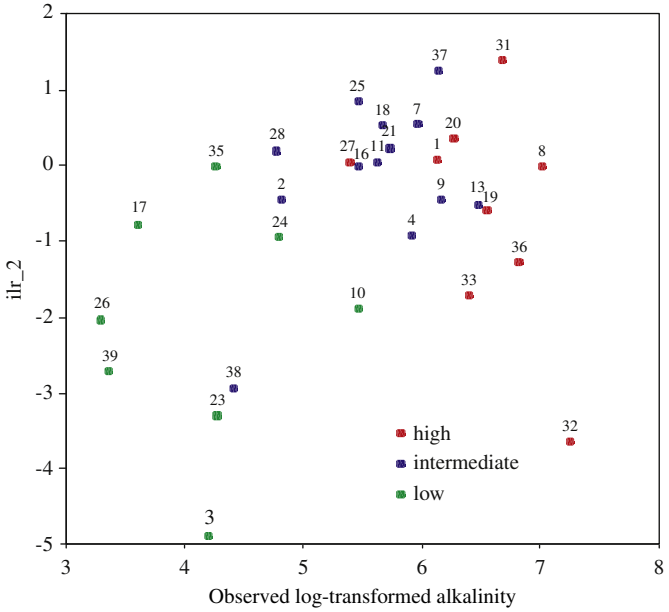


Fig. 4. Relationship between *ilr*_2 coordinate and observed log-transformed alkalinity (10^3 mol/km^2/year); data have been discriminated by taking into account the CO_2 consumption rates (high $>200\times10^3$ mol/km^2/year; moderate $50–100\times10^3$ mol/km^2/year; low $<50\times10^3$ mol/km^2/year; Amiotte Suchet et al., 2003). Labels as in Fig. 2.

is tropical wet dry) characterised, respectively, by scarce shales contribution (high values of *ilr*_1 and alkalinity $<100\times10^3$ mol/km^2/year) or scarce carbonate contribution (low *ilr*_1 values and alkalinity ranging between 0 and 500×10^3 mol/km^2/year), a significant positive relationship is detected (parametric correlation=0.77, non parametric correlation=0.83, $p<0.005$, with little changes, 0.75 and 0.79, respectively, excluding the case 32 relative to Si Kiang, possibly acting as leverage point). On the whole *ilr*_1 shows correlation with observed alkalinity and the basins characterised by different fluxes of consumed CO_2 are sufficiently discriminated. Since changes in the climatic settings are also represented (increases of *ilr*_1 values correspond to a passage into humid mid and tropical wet dry latitudes), the coordinate appears to represent a good index to consider weathering processes and changes in water/rock ratios jointly.

If the relationships among the proportion of the three categories of rocks and observed alkalinity values are investigated with classical tools (binary diagrams and determination of parametric and non parametric correlation coefficients), significant values ($p<0.05$) are obtained only for easily weathered rocks (positive correlation) and weathering resistant rocks (negative correlation), here obscuring the role played by intermediate lithologies such as basalts and shales. Are these results obtained in the classical framework reasonable? Do they satisfy geological common sense? Even if the answer may be yes, the role of intermediate lithologies could be completely crushed when the simplex geometry is treated as a real space. However, the contribution of shales and basalts to the carbon cycle and, consequently, to the global climate on Earth was discussed in Dessert et al. (2001) for the Deccan Traps and the approach here proposed is an attempt to emphasize their worldwide role.

The analysis of the relationship of the *ilr*_2 coordinate, dependent on the proportion between intermediate and weathering resistant rocks, and the log-transformed observed alkalinity (Fig. 4) allows us to investigate this aspect in more detail.

As we can see, data in Fig. 4 apparently follow parallel linear patterns. A first group of rivers is given by Amazon (1), Amour (2), Fraser (7), Yenisei (11), Lena (16), Limpopo (17), Mackenzie (18),

Mekong (20), Mississippi (21), Nile (24), Ob (25), Orange (26), Orinoco (27), Parana (28), Severnaia Dvina (31), Yana (35), Yukon (37), Zambesi (39), and shows a good correlation with log-transformed alkalinity (parametric and non parametric coefficient >0.8, $p<0.001$). Rivers are mainly located in periglacial and humid mid latitude in the northern hemisphere and in humid tropical or tropical wet dry conditions, these mainly in Africa and South America. Their basins are characterised by a higher presence of rocks with intermediate lithologies whose weight increases with alkalinity.

A second group of rivers given by Columbia (4), Ganges–Brahmaputra (8), Godavari (9), Huangho (10), Indus (13), Magdalena (19), Niger (23) and Zaire (38) is also characterised by a good correlation (parametric and non parametric coefficient >0.85, $p<0.002$). Most of these rivers are located in Asia and Oceania and with the exception of Columbia (North America) are mainly located below the Tropic of Cancer (humid tropical and tropical wet dry conditions). Their basins are characterised by a higher presence of weathering resistant rocks in comparison to the intermediate ones.

In this framework the rivers Colorado (3), Si Kiang (32), St. Lawrence (33) and Yangtze-Kiang (36) present moderate and important shifts with respect to the previous patterns due to the scarce presence of intermediate lithologies.

Results indicate that the *ilr*_2 coordinate due to its correlation with alkalinity is able to give indirect information about the CO_2 uptake and that the role of intermediate rocks is not negligible.

Considering runoff values, when data are analysed using raw percentages, significant correlation (positive) is obtained only for easily weathered rocks, while for *ilr* coordinates no significant correlations are revealed. It is known that local effects resulting from the distribution of rainfall in both space and time are superimposed on the continental-scale factors influencing the amount (variability) of river runoff (Berner and Berner, 1996). Geographic heterogeneity can increase the runoff by as much as 20% and *ilr* coordinates appear to better capture this phenomenon when compared with raw percentages. In this case too, the use of raw percentages, even if it leads to reasonable and geologically interpretable results, mortifies data variability forcing a relationship with runoff that could not be present.

4. Graphical representation of water chemistry and distortion related to the use of compositional data

Two types of diagrams are commonly used to represent water chemistry, Stiff and Piper plots. Stiff diagrams show the concentrations (in milliequivalent) of the major ions (both cations and anions) as a shape that gives the relative abundance of the various species and the total abundance (Stiff, 1951).

In the Stiff diagram the shape of the field is considered a representation of the relative proportions of the various ions, and its size shows the total ionic concentration. Stiff diagrams are considered particularly useful when plotted on a map because they give a graphical representation of regional variations in water chemistry. An example of the use of this type of representation is reported in Fig. 5 with reference to the Rio Grande and Columbia rivers and Central Pennsylvania ground waters (carbonate dominated). In the diagram *x*-axis is expressed in meq/L while an arbitrary scale (equal steps) is chosen for *y*-axis. As we can see, the Rio Grande has a much higher concentration of all ionic species than the Columbia river and the bicarbonate ion, relative to the sulphate ion, is much less important in the Rio Grande with respect to the Columbia river. By considering the groundwater, the diagram reveals that the dominant cation is Ca^{2+} and the dominant anion is HCO_3^-, as is usual for carbonatic acquifers.

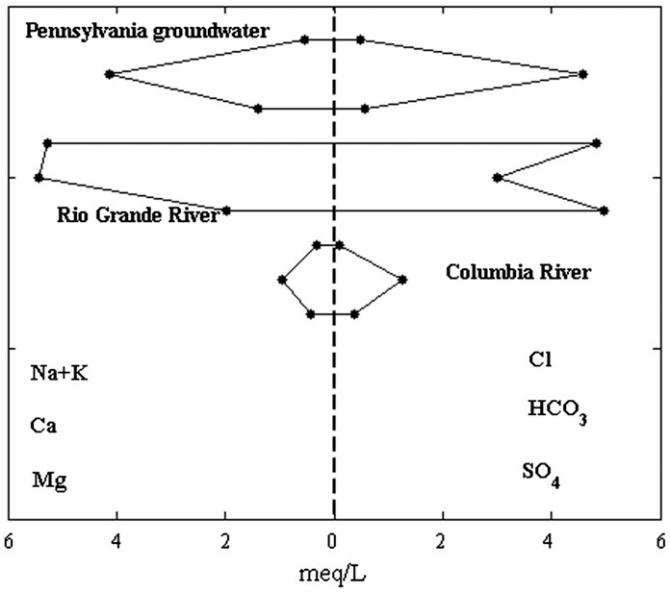


Fig. 5. Stiff diagrams for graphical representation of water chemistry (data from Nelson 2004). Values on horizontal axis are expressed in meq/L while an arbitrary scale (equal steps) is chosen for vertical axis.

Table 1
Sequential binary partitions (SBP) to compute balances between anions and cations related to the chemical composition of natural waters.

ilr	Na^+	K^+	Cl^-	Ca^{2+}	HCO_3^-	Mg^{2+}	SO_4^{2-}
*ilr*_1	+1	+1	+1	−1	−1	−1	−1
*ilr*_2	+1	+1	−1	0	0	0	0
*ilr*_3	+1	−1	0	0	0	0	0
*ilr*_4	0	0	0	+1	+1	−1	−1
*ilr*_5	0	0	0	+1	−1	0	0
*ilr*_6	0	0	0	0	0	+1	−1

An alternative to this representation of the chemistry of water can be obtained using *ilr* coordinates. The sequential binary partition (SBP) required to identify the *ilr* coordinates necessary to develop a diagram similar to that of Stiff is reported in Table 1. Among the different transformations that can be obtained with seven variables, three represent the balances to be used. They are the coordinates *ilr*_2, describing the balance between Na and K versus Cl, *ilr*_5, investigating the balance between Ca and HCO_3 and, finally, *ilr*_6, where Mg and SO_4 are compared with each other. If the values of the three coordinates for each sample are joined and reported in a binary plot, different triangles are obtained. In this framework, an equilateral triangle indicates an equilibrium among the different coordinates and reveals the relationships inside the whole chemical composition of water. Stretched triangles, on the other hand, point out the presence of dominant components. Results are reported in Fig. 6 for the same samples of Fig. 5. In the diagram *x*-axis is expressed in real numbers while an arbitrary scale (equal steps) is chosen for *y*-axis as in Fig. 5. The advantage of this representation is that it is not geometrically biased and the distances between points for each composition are in an Euclidean geometry (Buccianti and Magli, 2011). Moreover, ratios are easily interpreted in terms of stoichiometry of the mineral phases from which chemical species can be derived, one of the motives for which the measure unit of the equivalents is used in Stiff diagrams. When the *ilr* coordinates are equal to zero, $Na/Cl=Cl/K$, $Ca/HCO_3=1$ and $Mg/SO_4=1$.

As we can see, the Rio Grande river presents higher negative values for *ilr*_2 coordinate, indicating that $Na/Cl<Cl/K$ while in

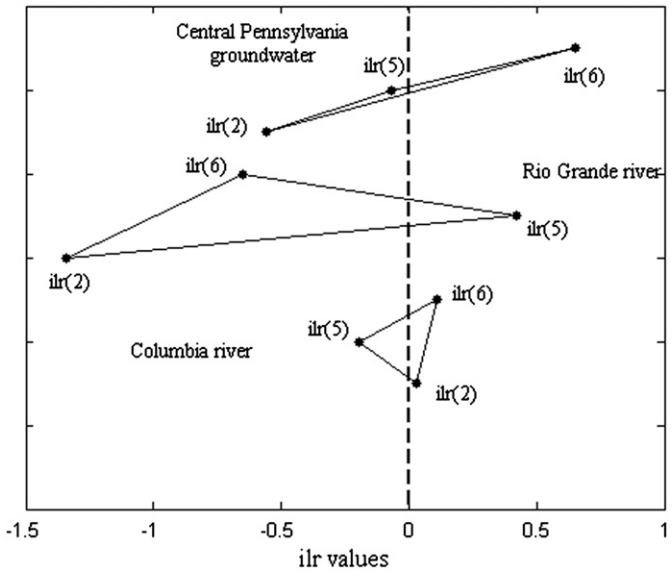


Fig. 6. An alternative diagram for graphical representation of water chemistry by using *ilr* coordinates (data from Nelson 2004). Values on horizontal axis are real numbers while an arbitrary scale (equal steps) is chosen for vertical axis as in Fig. 5.

Fig. 5 the obtained conclusion is that $Na+K \approx Cl$. The Central Pennsylvania groundwater shows a similar condition even if the coordinate values are lower in absolute value with regard to the Rio Grande River. A positive value, but near to zero, is registered for the Columbia river, indicating that $Na/Cl \approx Cl/K$ instead of $Na+K > Cl$.

It is evident in this type of representation that the stretched triangles of Fig. 6 reveal the presence of differences inside the chemical components characterising the coordinates and in particular among Na, K and Cl for *ilr*_2. This situation was not revealed in the Stiff diagram where, for example, Na+K is compared with Cl instead of the geometric mean of $Na \times K$ with respect to Cl. Consequently the shape of the Stiff diagram has to be evaluated with caution to extract information about the relative proportions of the chemical components, particularly if addition of values are reported on the axes. Combination of the parts of the composition by addition increases the illusion, since more distortion is added to the graphical representation of constrained data in real space (Egozcue and Pawlowsky-Glahn, 2005). Implications may be important: Na, K and Cl are components highly affected by pollution and their relative changes should be investigated in a coherent way.

The proposed *ilr* diagrams for investigating graphically the chemistry of water can be used to compare single compositions as shown above. However, since coordinates are real variables, the diagram can be constructed with mean or median values calculated for samples pertaining to different groups in order to compare them. Standard deviations or MAD (median absolute deviation) can also be reported to evaluate overlapping fields and to test hypotheses about differences. Further developments on this item are in progress.

5. Volcanic gas composition and geochemical processes: A statistical evaluation of the homogeneity of a fumarolic field

The composition of gases from active or quiescent volcanic and hydrothermal areas depends on deep processes, such as vapour–melt separation during the generation and rise of the magma and shallow processes active within the volcanic apparatus. Generally, the variability is widely attributable to shallow processes such as re-equilibration in response to cooling and dilution by meteoric water and interaction with fluids of associated hydrothermal systems. The chemical composition of volcanic and hydrothermal gases is expressed by concentrations, for example micromoles/mole. Since the relative presence of chemical species is attributable to partition phenomena between different phases, governed by chemical reactions, the investigation of the behaviour of single components is not fully informative. In this perspective the use of ratios as indices of relative changes is a common practice, as well as the adoption of diagrams where ratios present a common denominator, even if criticised as early as 1897 (Pearson, 1897).

The present case study aims to investigate the relationships among the chemical components constituting the gases emitted from different discharges located in several parts of the fumarolic field of Vulcano Island (Aeolian archipelago, Sicily, southern Italy). When the sampling of fumarolic gases is performed considering discharges located in different positions of the volcanic apparatus, it is important to evaluate whether significant differences can exist in their chemistry and if spatially the conditions are or are not homogeneous. Considering the wide variability of fractures in the system and the complexity of the water/rock interactions in this type of environment, differences are generally expected even for short distances (Signorelli et al., 1998).

Three groups of fumaroles were selected as representative of the spatial conditions and have been monitored from 2000 to 2009. Two of them are located in the outer part of the crater rim (labelled FNB and FZ), two on the rim itself (FNA and F5) and three in the inner part of the crater rim (F14, F27, F202). The chemistry of the gases includes the following components: H_2O, CO_2, SO_2, H_2S, S_2, HCl, HF, N_2, O_2, H_2, Ar, Ne, He, CO and CH_4 in μmol/mol. Previous investigations revealed that by considering outlet temperatures data appeared to be drawn from different populations with different mean values increasing from 100 °C (FZ) to 211 °C (F5), 247 °C (FNB), 285 °C (FNA), 306 °C (F14), 363 °C (F27) and 403 °C (F202) (Buccianti et al., 2006). Considering the three groups of fumaroles, temperature values range as follows: rim 180–337 °C (median = 268 °C), inner part 247–419 °C (median = 342), outer part 88.7–319 °C (median = 157).

Water is the most important constituent of volcanic gases and is involved in several key chemical reactions, as for example (1) $H_2+CO_2 \leftrightarrow CO+H_2O$, (2) $4H_2+2SO_2 \leftrightarrow S_2+2H_2O$ and (3) $CO_2+4H_2 \leftrightarrow CH_4+2H_2O$, all highly dependent on temperature and redox conditions of the system.

The previous reactions have represented the starting point to choose balances, or *ilr* transformed variables, able to describe the relationships between reactants (denominator of the balances) and products (numerator of the balances). Transformed values have then been used in a subsequent phase to test differences among the three groups of fumaroles and consequently to evaluate the homogeneity of the fumarolic field. Graphical results are reported in Fig. 7. As we can see, data for reaction (1) and (3) are mainly higher than zero, independently from their spatial position, indicating that for the whole fumarolic field, the barycentre of the systems $CO–H_2O$, reaction (1), and $CH_4–H_2O$, reaction (3) are systematically higher than that of $H_2–CO_2$. On the other hand, values for reaction (2) decrease in the sequence crater rim→ inner crater→ external crater, so that the barycentre of the system $S_2–H_2O$ may be higher or lower when compared with that of $H_2–SO_2$, depending on spatial position.

The application of normality tests indicates that balances follow the Gaussian model for each sampling site ($p > 0.05$), so that parametric tests for comparing mean values can be applied. The one way analysis of variance (ANOVA) allows us to verify that the balance related to reaction (1) $H_2+CO_2 \leftrightarrow CO+H_2O$, shows significant differences ($p < 0.01$) between the crater rim with

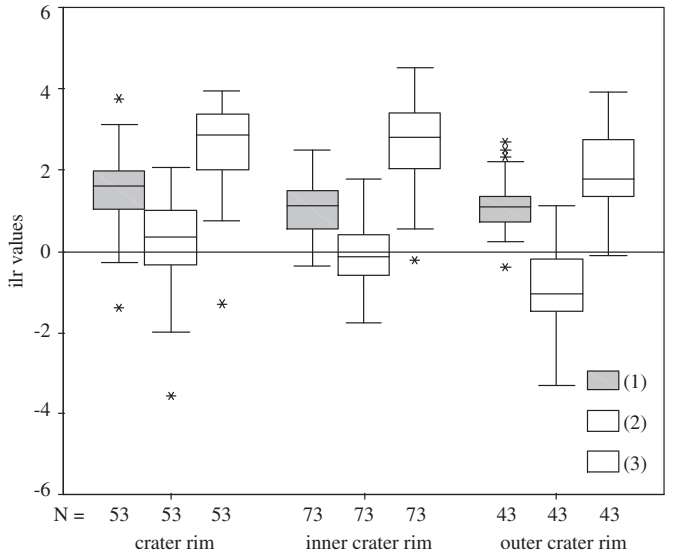


Fig. 7. Box-plots of the balances (*ilr* transformed variables) representing the reactions (1) $H_2+CO_2 \leftrightarrow CO+H_2O$, (2) $4H_2+2SO_2 \leftrightarrow S_2+2H_2O$ and (3) $CO_2+4H_2 \leftrightarrow CH_4+2H_2O$, for the three sites of sampling of the fumarolic field. Symbol * indicates anomalous values, *N* the number of sample for each site.

respect to outer and inner areas. Balances related to reactions (2) $4H_2+2SO_2 \leftrightarrow S_2+2H_2O$ and (3) $CO_2+4H_2 \leftrightarrow CH_4+2H_2O$, point out significant differences attributable to the outer part of the crater when compared with the other parts.

In the investigation of the chemical composition of volcanic gas discharges increased concentrations of CO_2 and SO_2 are often considered gas precursors of eruptions, while H_2O is attributable to magmatic or meteoric sources. However, an increase in SO_2 is not always observed prior to volcanic events due to the presence of groundwater or surface water that scrub magmatic gases during their path toward the surface. Geochemical modelling suggests that CO_2 is the main component to monitor if scrubbing phenomena affect volcanic discharges, since concentrations in SO_2 will be significantly influenced (Symonds, et al., 2001).

If the ANOVA is applied on raw concentrations, significant differences ($p < 0.05$) are obtained only when data collected from the outer part of the crater rim are compared with that pertaining to other sites and only for three variables, H_2, CO and S_2. In this contest, SO_2 abundance discriminates all the sites, while CO_2, H_2O and CH_4 are not relevant for this purpose. Results may be interpretable in the classical framework, considering the geochemical behaviour of the single species. For example, the discriminating role of SO_2 may refer to the indirect presence/absence of water and related scrubbing processes in all sites, while H_2, CO and S_2 may be related to high temperature conditions (H_2 and CO) and surficial conditions (S_2) affecting the sampling sites in a different way. However, it should be noticed that the balance built on the base of reaction (3), including CO_2, H_2O and CH_4, revealed, in the compositional approach, the presence of significant differences between the outer part of the crater and the other sites, giving us another look at the relative behaviour of the chemical species.

By taking into account the previous results, it is evident that the spatial position is important in determining the chemical composition of the collected samples. The result is expected and confirmed by using the different approaches proposed. However, it is clear that the balance approach is more powerful and allows us to understand better how natural phenomena are working. If we consider for example the balances related to reactions (1) $H_2+CO_2 \leftrightarrow CO+H_2O$ and (3) $CO_2+4H_2 \leftrightarrow CH_4+2H_2O$, we can see that higher values represent environmental conditions affected by water presence able to scrub chemical species; on the other hand, lower values may describe the minor contribution of water, as confirmed by the presence of CO_2 and H_2. This condition characterises all the sampling sites and differences among them may be related to the position of the barycentre between CO and H_2O or CH_4 and H_2O since the C-species are stable in different temperature ranges. In fact, the balance associated with reaction (3) shows higher values, but the median value of temperature is lower when compared with other sites, a condition favouring the presence of CH_4. To be noticed here is that CO_2, H_2O and CH_4 are not relevant to discriminate spatial positions in the classical approach. Finally, considering the balance associated to reaction (2) $4H_2+2SO_2 \leftrightarrow S_2+2H_2O$, due to the fact that its values cross the zero, it is possible to find compositions with a different compensation between the numerator and denominator of the balance. In particular, for most of the data of the crater rim the barycentre of the system S_2–H_2O is higher than that of H_2–SO_2, revealing a clear influence of scrubbing phenomena. Changes internal to the system S_2–H_2O can be easily related to shallow water/rock interactions for intense hydrothermal circulation. Data collected from the inner part of the crater rim present an equilibrium between these two different conditions, while for the outer rim area, presence of water, with consequent increased hydrothermal circulation, appears to be the dominant phenomena.

6. Conclusions

Compositional data analysis (CoDA) has been a major issue of discussion for more than 100 years. It is difficult to realise that many statisticians and users of statistics in the geochemical community are unaware of the features of compositional data, and also that now some simple tools for their coherent analysis are available. Why is it not sufficient to propose coherent mathematical tools? In this paper the reasons at the base of this difficulty have been tentatively analysed considering classical and compositional statistical approaches for three different case studies involving several interesting items. Results indicate that often classical statistics applied on compositional data give us reasonable and interpretable results that unite common sense and acquired knowledge. However when results are compared with those obtained with the compositional approach it is easy to verify that the behaviour of some components of the composition may be obscured and their variability seriously mortified. This is an important item, since part of the information about compositional natural systems is completely lost and it is not possible to know the weight of this effect a priori. If the level of investigation of compositional data with classical tools moves from a descriptive phase to the application of statistical tests, erroneous evaluation of the formulated hypothesis may be the result, a serious point when environmental questions are the matter of debate. Thus, the answer to the question posed in the title of this paper is yes, compositional data approach is the way to go beyond the illusion to (1) have obtained an exhaustive understanding on how natural phenomena work and (2) have extracted all the information contained in data variability.

Acknowledgement

This work was supported by the University of Florence (I) (ex-60% contribution for 2010/2011), and by the Region of Tuscany (research and innovation for environmental and territorial projects, 2009–2011). I would like to thank the reviewers for their constructive and thorough criticism and Lisa Merli for the English language.

References

Aitchison, J., 1982. The statistical analysis of compositional data (with discussion). Journal of the Royal Statistical Society, Series B (Statistical Methodology) 44 (2), 139–177.

Aitchison, J., 2001. Simplicial inference. In: Viana, M.A., Richards, D.S. (Eds.), Algebraic Methods in Statistics and Probability, 287. American Mathematical Society, Providence, RI (USA), pp. 1–22.

Aitchison, J., Egozcue, J.J., 2005. Compositional data analysis: where are we and where should we be heading? Mathematical Geology 37 (7), 829–850.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A., Pawlowsky-Glahn, V., 2000. Logratio analysis and compositional distance. Mathematical Geology 32 (3), 271–275.

Amiotte Suchet, P., Probst, J.L., 1995. A global model for present day atmospheric/soil CO_2 consumption by chemical erosion of continental rocks (GEM–CO_2). Tellus Series B 47, 273–280.

Amiotte Suchet, P., Probst, J.L., Ludwig, W., 2003. Worldwide distribution of continental rock lithology: implications for the atmospheric/soil CO_2 uptake by continental weathering and alkalinity river transport to the oceans. Global Biogeochemical Cycles 17 (2), 7/1–7/13.

Berner, E., Berner, R., 1996. Global Environment: Water, Air and Geochemical Cycles. Prectice Hall, Inc, New Jersey 463 pp.

Bertocchi, S., Brusconi, S., Gherardi, F., Buccianti, A., Scalici, M., 2008. Morphometrical characterization of the *Austropotamobius pallipes* species complex. Journal of Natural History 42 (31,32), 2063–2077.

Billheimer, D., Guttorp, P., Fagan, W., 2001. Statistical interpretation of species composition. Journal of the American Statistical Association 96 (456), 1205–1214.

Blatt, H., Jones, R.L., 1975. Proportions of exposed igneous, metamorphic and sedimentary rocks. Geological Society American Bulletin 86, 1085–1088.

Bluth, G.J.S., Kump, R.L., 1991. Phanerozoic paleogeology. Americam Journal of Sciences 291, 281–308.

Buccianti, A., 2011a. Natural laws governing the distribution of the elements in geochemistry: the role of the log-ratio approach. In: Pawlowsky, Glahn, Buccianti (Eds.), Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Ltd., pp. 255–266.

Buccianti, A., 2011b. Isometric log-ratio co-ordinates and their simple use in water geochemistry. Boletin Geologico y Minero 122 (4), 453–458.

Buccianti, A., Esposito, P., 2004. Insights into late quaternary calcareous nannoplankton assemblages under the theory of statistical analysis for compositional data. Palaeogeography, Palaeoclimatology, Palaeoecology 202 (3–4), 209–227.

Buccianti, A., Magli, R., 2011. Metric concepts and implications in describing compositional changes for world river's water chemistry. Computers & Geosciences 37, 670–676.

Buccianti, A., Tassi, F., Vaselli, O., 2006. Compositional changes in a fumarolic field, Vulcano Island, Italy: a statistical case study. In: Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (Eds.), Compositional Data Analysis in the Geosciences: From Theory to Practice, 264. Geological Society, London, pp. 67–77, Special Publications.

Buccianti, A., Apollaro, C., Bloise, A., De Rosa, R., Falcone, G., Scarmiglia, F., Tallarico, A., Vecchio, G., 2009. Natural radioactivity levels (K, Th, U and Rn) in the Cecita Lake area (Sila massif, Calabria, Southern Italy): an attempt to discover correlations with soil features on a statistical base. Geoderma 152 (1–2), 145–156.

Chayes, F., 1960. On correlation between variables of constant sum. Journal of Geophysical Research 65 (12), 4185–4193.

Clarke, F., 1924. The data of geochemistry. U.S. Geological Survey Bulletin 770, 841.

Dessert, C., Dupré, L., Francois, L.M., Schott, J., Gaillardet, J., Chakrapani, G., Bajpai, S., 2001. Erosion of Deccan traps determined by river geochemistry: impact on the global climate and the $^{87}Sr/^{86}Sr$ ratio of seawater. Earth Planetary Science Letters 188, 459–474.

Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Mathematical Geology 37 (7), 795–828.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric loratio transformations for compositional data analysis. Mathematical Geology 35 (3), 279–300.

Gibbs, M.T., Kump, L.R., 1994. Global chemical erosion during the last glacial maximum and the present: sensitivity to changes in lithology and hydrogeology. Paleoceanography 9, 529–543.

Hron, K., Templ, M., Filzmoser, P., 2010. Imputation of missing values for compositional data using classical and robust methods. Computational Statistics and Data Analysis 54 (12), 3095–3107.

Meybeck, M., 1987. Global chemical weathering of surficial rocks estimated from river dissolved loads. American Journal of Sciences 287, 401–428.

Nelson, EbyG., 2004. Principles of Environmental Geochemistry. Brooks/Cole-Thomson Learning Inc, USA, isbn: 0-122-29061-5.

Nisi, B., Buccianti, A., Vaselli, O, Perini, G., Tassi, F., Minissale, A., Montegrossi, G., 2008. Hydrogeochemistry and strontium isotopes in the Arno river basin (Tuscany, Italy): constraints on natural control by statistical modelling. Journal of Hydrology 360, 166–183.

Pawlowsky-Glahn, V., Egozcue, J.J., 2001. Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment 15 (5), 384–398.

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. Proceedings of the Royal Society of London, LX, 489-502.

Revenga, C., Murray, S., Abramovitz, J., Hammond, A., 1998. Watersheds of the World: Ecological Value and Vulnerability. World Resources Institute, Washington, DC.

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

Stiff, H.A., 1951. The interpretation of chemical water analysis by means of patterns. Journal of Petroleum Technology 3/10, 15–17.

Symonds, R., Gerlach, T., Reed, T., 2001. Magmatic gas scrubbing: implications for volcano monitoring. Journal of Volcanology and Geothermal Research 108, 303–341.

Tassi, F., Vaselli, O., Cuccoli, F., Buccianti, A., Nisi, B., Lognoli, E., Montegrossi, G., 2009. A geochemical multi-methodological approach in hazard assessment of CO_2-rich gas emissions at Mt. Amiata volcano (Tuscany, Central Italy). Water, Air and Soil Pollution: Focus 9, 117–127.

Tricarico, E., Benvenuto, C., Buccianti, A., Gherardi, F., 2008. Morphological traits determine the winner of "symmetric" fights in hermit crabs. Journal of Experimental Marine Biology and Ecology 354, 150–159.