Frequency distributions and natural laws in Geochemistry

Frequency distributions and natural laws in geochemistry

A. BUCCIANTI[1], G. MATEU-FIGUERAS[2] & V. PAWLOWSKY-GLAHN[2]

[1]*Dipartimento di Scienze della Terra, Università di Firenze, Via G. La Pira 4, I-50125, Firenze, Italy (e-mail: antonella.buccianti@unifi.it)*

[2]*Departament Informàtica i Matemàtica Aplicada, Universitat de Girona, Campus Montilivi, P4, E-17071 Girona, Spain*

Abstract: The chemical composition of natural waters is derived from many different sources of solutes, including gases and aerosols from the atmosphere, weathering and erosion of rocks and soil, solution or precipitation reactions occurring below the land surface, and effects resulting from human activities. The chemical composition of the crustal rocks of the Earth, as well as the composition of the ocean and the atmosphere, are important in evaluating sources of solutes. Data used in the investigation of natural and non-natural contributions are obtained usually from chemical analysis of water samples, which may be statistically evaluated with the aim of summarizing the contained information. However, as these data are compositional and thus constrained to move in the simplex, application of usual statistical methodologies may lead to incorrect evaluations and/or interpretations. This paper focuses on how to draw information on natural processes by modelling univariate and multivariate frequency distributions using water data. The chemical composition of 977 samples collected in wells from Vulcano island (Italy) are used as a case study. The methodological approach can be transferred to the investigation of other geochemical or constrained data.

Modelling of geochemical processes affecting water chemistry by applying univariate and multivariate statistical methodologies may improve the interpretation of complex systems and, in the process of finding proper models, the sample space represents an important issue. Chemical constituents of water are reported usually in gravimetric units, such as milligrammes per litre (ppm) or milliequivalent per litre. In both cases, a sample space has been constrained with a metric different from the ordinary one, as measurements cannot fall outside the interval $(0, 10^6)$ or $(0, 10^9)$ of the real line, and the measure of difference between observations is relative. It is only reasonable to expect the statistical results to respect these facts and this is only possible if the model used does so. The problem is that most usual statistical models require, instead, the whole real line as sample space and an absolute measure of difference. Therefore, when constrained data with a relative scale are managed, the evaluation of the shape of a frequency distribution, as well as the application of statistical tests to evaluate the performance of a given model, is at the most a good approximation and, in any case, questionable.

Why is the knowledge about the distribution that can be used to model the data so important? There are at least two reasons. One is that distributions can frequently be associated with a generating process which might be helpful in understanding the studied phenomenon. This is well known in geochemistry. For example, concentrations of different elements in terrestrial waters displayed in cumulative frequency plots are used to evaluate the effects of natural phenomena (Davies & DeWiest 1996). For some species, such as Ca^{2+}, HCO_3^-, SiO_2, K^+ and F^-, a steep gradient was found, indicating that the solubility of a mineral places an upper limit on the maximum concentration of the species in natural waters. For most of the other species, the curve has a lower gradient, suggesting that the concentrations depend rather on their availability in rocks, on very slowly dissolving minerals, or are controlled by biological processes.

The other reason is that most statistical techniques assume some underlying distribution and it is important that it fits the data reasonably well for results to be valid. For example, standard correlation analysis, factor analysis, discriminant analysis, tests and computation of probability levels are usually based on the assumption of a normal (Gaussian) distribution.

In general, inspection about the best probability model to be used in data description is a topic widely neglected, although the problem has been discussed in several papers. Ahrens (1953, 1954*a*,*b* 1957) proposed the log-normal distribution for geochemical data as a universal law after analysing abundance of trace elements in crustal rocks, expressed as the frequency of appearance versus concentration. His idea encountered criticism (Aubrey 1954; 1956; Chayes 1954; Miller

From: BUCCIANTI, A., MATEU-FIGUERAS, G. & PAWLOWSKY-GLAHN, V. (eds) *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Geological Society, London, Special Publications, **264**, 175–189.
0305-8719/06/$15.00

& Goldberg 1955; Vistelius 1960), but in recent textbooks his assertion is commonly reported. Consequently, the log-transformation is used the most frequently when geochemical data are managed. However, besides the fact that a logarithmic transformation does not project data from a constrained sample space to the whole real line, log-transformed data do not often adjust sufficiently well to a normal distribution (McGrath & Loveland 1992), as they still present some skewness. This is generally attributed to imprecise measurements, to many potential sources of error involved in sampling, to sample preparation and analysis, to detection limit problems, to the presence of outliers, or to mixing of populations (Reimann & Filzmoser 1999). Another approach for interpreting empirical geochemical distributions has been given by Allègre & Lewin (1995), who considered that they are not of a unique type. They proposed a unified theory based on the action of differentiation-mixing operators able to give, respectively, fractal and Gaussian distributions. If their reasoning is inverted, when the distribution is Gaussian, mixing is the acting geochemical operator and, when the distribution is fractal or multifractal, differentiation is the dominant process. However, none of the cited works considers the sample space of geochemical data explicitly, nor do the authors justify the appropriateness of models with support for the whole real line (normal distribution), or the strictly positive real line (log-normal model), for data which are constrained to an interval. The conclusions on the shape of the frequency distribution are therefore in general inconsistent, at least from a formal point of view.

Recently, the skewness of log-transformed data has been tentatively modelled using the logskew-normal distribution (Mateu-Figueras 2003). This model is defined on the positive real line combining the logarithmic transformation and the univariate skew-normal model on the real line (Azzalini 1985). It is an extension of the log-normal distribution by adding a shape parameter. This model could be considered a good alternative, but it is defined on the whole positive real line, the same as the log-normal model. Thus, this model is also questionable for constrained data. But the multivariate extension by Mateu-Figueras *et al.* (2005) of the logistic normal distribution (Aitchison 1986), defined on the simplex, and thus on a constrained sample space, appears to be a reasonable alternative for compositional data. Following the (1986) approach of using log-ratios, the basic idea is to transform compositional data to the whole real space using an appropriate log-ratio transformation and to use the multivariate skew-normal distribution (Azzalini & Dalla Valle 1996; Azzalini & Capitanio 1999) for modelling the transformed sample. The advantages of this strategy are many: it takes into account the constraint character of the sample space, it is consistent with its algebraic–geometric structure, and it is flexible enough to account for some skewness in the transformed data.

The aim here is to draw information on the action of natural processes in a correct statistical framework so that inferential modelling can be developed. In this context, the potentiality of the logistic skew-normal distribution as a **natural law** for geochemical phenomena is evaluated. Strictly speaking, to attain this goal, first an appropriate basis of the simplex, expressed in terms of log-ratios, has to be found (Egozcue *et al.* 2003). From it, using basic linear algebra, any vector expressed in terms of log-ratios can be obtained. Therefore, in practice, log-ratios of interest can be defined using known geochemical properties, using appropriate coefficients which depend on the number of parts in the numerator and in the denominator (Egozcue & Pawlowsky-Glahn 2005), and this will be the approach used here.

Theoretical distributions and geochemical processes

Although natural processes and phenomena in geochemistry may combine many complex and poorly understood factors, their frequency distributions each appear to follow closely one of a few theoretical models. The theoretical frequency distribution provides a probability density function for predicting the probability of occurrence of certain events. For example, physical and geological observations that follow an apparently symmetrical frequency distribution have been compared to the normal distribution, also named Gaussian distribution after its originator Karl Friedrich Gauss (1777–1855). The normal distribution was introduced to model the pattern of non-systematic and additive errors of observation and measurement. It requires a physical property X ranging theoretically from $-\infty$ to $+\infty$, whose realizations show a strong tendency to cluster around their mean, and are equally likely to undershoot or overshoot the mean. The resulting frequency distribution is symmetrical, with long tails corresponding to rare events, far from the mean. The dispersion about the mean is defined by the variance. Normal processes are obtained under some conditions: the main one is the summation of many continuous random variables, and a mild condition is the independence of these random variables. The normal distribution is a mathematical model of dispersion commonly applied in geochemistry, although many variables do not range from $-\infty$ to $+\infty$, e.g. percentages, ppm or other similar units. Therefore, in

constrained systems, the use of the normal distribution is not appropriate. Moreover, using the normal distribution, predictions outside the sample space can be obtained, leading in general to negative values or, less frequently, to values larger than the maximum possible value (e.g. 100% for observations). This fact does not only invalidate the obviously wrong results, but puts a reasonable doubt on the correctness of the apparently good results, namely those which satisfy the constraints.

The fact that the normal distribution does not respect the constrained character of the sample space when dealing with compositional data is not the only reason for this model to be inappropriate. In fact, geochemical distributions cannot be compared with the normal distribution because multiplicative sources of variability are the best way to interpret their behaviour. Also, in some cases, the presence of a positively skewed distribution is observed, showing a long tail to the right, towards high values of measurement. It indicates that observations with large values are not unusual. Frequently, this positive skewness disappears after applying the logarithmic transformation. In this situation the distribution of the physical property X has been modelled frequently by the log-normal distribution, for which the sample space is the positive real line, $\Re^+$. Typical examples are concentrations of accessory minerals and other rare components, such as trace elements, which may show a positively skewed frequency distribution with a mode at a low concentration. Compared with the conditions necessary to obtain the normal distribution, a log-normal process is one in which the random variable of interest results from the product of many independent random variables multiplied together. This happens, for example, when the quantity present in each state is expressed as a random proportion of the quantity present in the immediately prior state. If each successive proportion is independent of the previous one, and if many states occur between the initial state and the final one, then the final result can be expressed as a product of random variables and the variable of interest will show some similarity with a log-normal distribution. A general mechanism that in nature can generate the right-skewed concentration distributions is explained under the 'Theory of Successive Random Dilutions'. It represents a special application of the 'Law of Proportionate Effects' originally proposed by Kapteyn (1903). It is considered to be especially appropriate for modelling substances released into environmental carrier media (air, water, soil) experiencing considerable physical movement and agitation. It considers a pollutant released at initially high concentrations into a carrier medium, which undergoes dilution in successive, independent stages. If a mixing stage occurs between each independent dilution stage, the final concentration will be the product of the initial concentration and a series of independent dilution factors. When the number of successive random dilutions becomes large, the distribution of the final concentration has been approximated by a log-normal (Wayne 1990), although the constraint due to the volume is not taken into account in this model.

Sometimes the normal model cannot fit the log-transformed data properly because of remaining skewness. To deal with it the skew-normal model can be used; it is given by a family of skewed distributions, including the normal one, defined on the whole real line, and characterized by an extra parameter which allows the density to have positive or negative skewness (Azzalini 1985). The notation $X \sim SN(\mu, \sigma, \lambda)$ is used. The parameter λ controls the shape of the distribution; when $\lambda = 0$, $SN(\mu, \sigma, 0)$ is equal to the $N(\mu, \sigma)$, while when $\lambda \to \pm\infty$ a half-normal density is the result (Azzalini 1985). Consequently, besides the mean and the variance, the skew-normal distribution depends on a skewness index γ that varies in the interval $(-0.995, +0.995)$ and when λ tends to $\pm\infty$, γ tends to ± 0.995. Starting from this model, Mateu-Figueras (2003) investigated the properties of the logskew-normal distribution for a positive random variable X, with support $\Re^+$; i.e. for $Y = \ln(X)$ following a $SN(\mu, \sigma, \lambda)$ distribution. The logskew-normal distribution is a generalization of the log-normal distribution and allows the density to incorporate positive or negative skewness. The use of the log-normal and the logskew-normal distribution as an approximate model for the concentration of the elements could be discussed. Nevertheless, in constrained systems the use of the log-normal or the logskew-normal distribution still has a major drawback, because they are defined on the whole positive real line.

To avoid this problem, the proposal here is to work with models defined exclusively on the support space of the geochemical data, with a relative measure of variability, using the log-ratio approach (Aitchison 1986) and an appropriate representation. First, some single log-ratios, constructed from two components properly chosen from a geochemical point of view, will be investigated using the univariate normal and the univariate skew-normal models. To model a compositional vector $\mathbf{X}$, multivariate models with sample space a simplex and with a relative measure of variability, have to be used. The logistic skew-normal distribution (Mateu-Figieras *et al.* 2005) on an appropriate basis is a reasonable choice. A D-part composition $\mathbf{X} = (X_1, X_2, \ldots, X_D)$ is said to follow a logistic skew-normal distribution when its transform by an appropriate log-ratio transformation,

$\mathbf{Y}$, follows a multivariate skew-normal distribution. It is denoted as $\mathbf{Y} \sim SN^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is a parameter that takes into account the presence of skewness (Azzalini & Dalla Valle 1996; Azzalini & Capitanio 1999). The logistic normal distribution (Aitchison 1982, 1986) is a particular case, obtained when $\mathbf{Y}$ follows a multivariate normal distribution or, equivalently, when $\boldsymbol{\lambda}$ is the null vector. Note that in the multivariate case bold notation will be used to indicate vectors and matrices.

Log-ratios and their univariate statistical modelling: results and discussion

Database and geochemistry of waters

For illustration, the log-ratio analysis is applied to water samples collected in the quiescent volcanic environment of Vulcano, an island belonging to the Aeolian Archipelago (Sicily, Southern Italy). The area, interpreted as a typical volcanic arc generated by subduction processes beneath the Tyrrhenian Sea, had the last eruption from 1888 to 1890. Since then, fumarolic activity of varying intensity has continued to the present day. Several years of geochemical investigations have produced some models to describe the evolution of the fluids (water and gases) related to the volcanic system in time (Martini 1980, 1989, 1996; Montalto 1996; Capasso *et al.* 1999, 2001; Di Liberto *et al.* 2002; Buccianti & Pawlowsky-Glahn 2005). These studies allowed the identification of at least two aquifers, a shallow one of meteoric origin and a deeper one influenced by thermal activity. From a hydrological point of view, the aquifers are probably not physically separated. A current hypothesis considers that the shallower less saline aquifer floats over the more saline one of marine origin and is affected by the interaction with volcanic fluids of deeper origin. The differences observed in the wells are then attributable to lateral permeability variations, local alteration processes, and/or to the presence of areas of preferential upflow of volcanic fluids. From a general point of view, the aggressive character of the water able to mobilize the elements in this environmental context is due to the input of carbon dioxide from the deep uprising gaseous flow into the aquifer. The presence of hydromagmatic deposits, which appear to have undergone early syn-depositional alteration processes, contributes elements from secondary minerals as calcium sulphate, calcium fluoride, sodium chloride and others. According to their solubility in aqueous solutions, chemical components can be leached away even if in the presence of a weak alteration of surface waters. Systematic studies have been in progress since 1977 by geochemists of the University of Florence. At present, 977 samples of groundwater, pertaining to 50 wells located in the northwest sector of the area surrounding the active crater, have been sampled and analysed at regular time intervals to look at the concentrations (ppm) of Ca^{2+}, Na^{+}, Mg^{2+}, K^{+}, HCO_3^{-}, SO_4^{2-} and Cl^{-}. Calcium is the most abundant of the alkaline-earth metals and is a major constituent of many common rock minerals. It can be derived from carbonate, gypsum, feldspar, pyroxene and amphibole; processes affecting its distribution are related to dissolution, while limits are due to the solubility of calcite. Consequently, the behaviour of calcium in natural aqueous systems is generally governed by the availability of the more soluble calcium-containing solids and by solution and gas phase equilibria that involve carbon dioxide species, as well as by the availability of sulphur in the form of sulphate. It also participates in cation-exchange equilibria of aluminosilicates and other mineral surfaces. The HCO_3^{-} species can be derived from carbonates and organic matter; processes affecting its concentration are in general due to soil-CO_2 pressure, equilibria involving carbon dioxide species (particularly in a volcanic environment) and weathering, while limits are posed by organic matter decomposition. The SO_4^{2-} species is contributed by the atmosphere, gypsum and sulphides, but in a volcanic area the input of gaseous components (affecting S cycle) from fumarolic activity is also important. Processes involved in its distribution are dissolution and oxidation, while concentration limits are attributable to removal by reduction. Sodium, the most abundant member of the alkali-metal group, when brought into solution, tends to remain in that status, once it has been liberated from silicate-mineral structures. It derives from feldspar, rock-salts, zeolite and the atmosphere by dissolution and cation exchange, particularly in coastal aquifers, with concentration limits due to silicate weathering. There are no important precipitation reactions that can involve sodium, as carbonate precipitation controls calcium concentration. Potassium is slightly less common than sodium in igneous rocks, but it is more abundant in all the sedimentary rocks. It is liberated with greater difficulty from silicate minerals (feldspar, mica) and it is affected by processes of dissolution, adsorption (it exhibits a strong tendency to be reincorporated into solid weathering products as clay minerals) and decomposition. Moreover, the element is involved in the biosphere processes, especially in vegetation and soil, and is essential for both plants and animals; limits on its concentration are due to solubility of clay minerals and vegetation uptake. Chloride, derived from rock-salts and atmosphere, is present in all natural waters, but mostly the

concentration is low. Exceptions occur where the inflow of sea water can affect the chemical composition of groundwater, as in the case of Vulcano island. Chloride ions do not enter into oxidation or reduction reactions significantly, form no important solute complexes with other ions (unless the concentration is extremely high), do not form salts of low solubility and are not adsorbed significantly on mineral surfaces. Furthermore, they play a minor role in biogeochemical processes. On the whole, the circulation of chloride ions in the hydrological cycle is largely through physical processes, dominated by a conservative behaviour. The alkaline-earth metal Mg^{2+} shows only one oxidation state of significance in water chemistry and is a common element, essential in plant and animal nutrition. Processes that are important as sources in water are related to dissolution, while limits are posed by the solubility of clay minerals. Generally, it can derive from dolomite, serpentine, pyroxene, amphibole, olivine and mica, but a source of Mg^{2+} at Vulcano island can also be ascribed to the influence of deep-seated aquifers of marine-like composition, which occasionally inflow into the overlying water bodies. Even if magnesium has a behaviour similar to calcium (i.e. property of hardness), its ions are smaller and can be accommodated in the space at the centre of six octahedrically co-ordinated water molecules. This behaviour increases the tendency to precipitate crystalline compounds. Magnesium occurs in significant amounts in most limestones and the dissolution of this material can bring magnesium into solution. However, the process is not readily reversible, and the precipitate that forms from a solution may be nearly pure calcite. As a consequence, magnesium concentration would tend to increase along the flow path of a groundwater undergoing such processes, achieving a rather high Mg/Ca ratio.

Suitability of normal and skew-normal models

As the data are constrained with a relative scale, log-ratios are taken. A first phase involves the investigation of the shape of the frequency distributions of the log-ratios $(1/\sqrt{2})\ln(Ca^{2+}/HCO_3^-)$, $(1/\sqrt{2})\ln(Na^+/K^+)$, $(1/\sqrt{2})\ln(Na^+/Cl^-)$, $(1/\sqrt{2})\ln(Ca^{2+}/Mg^{2+})$, $(1/\sqrt{2})\ln(Mg^{2+}/SO_4^{2-})$ and $(1/\sqrt{2})\ln(Ca^{2+}/SO_4^{2-})$, considering cations and/or anions derived from similar sources as, for example, carbonates or sulphates, or showing a similar geochemical behaviour in natural processes. The coefficient $(1/\sqrt{2})$ is used to preserve the same scale as in the simplex, the sample space of the full composition. Note that the above log-ratios are not orthogonal in the geometry of the simplex, and therefore they are not appropriate to perform a standard multivariate analysis combining them (Egozcue & Pawlowsky-Glahn 2005).

To investigate the shape of the frequency distributions of mentioned log-ratios, whose terms have been chosen from a geochemical point of view (like anions and/or cations with the same charge, or species potentially derived from the same source), the normal and the skew-normal distribution are used. The parameters of the skew-normal distributions have been estimated using the maximum likelihood method, working with the routines of the Matlab version software library available at http://azzalini.stat.unipd.it/SN/index.html.

Note that an underlying hypothesis to this approach is random sampling from a single population. Nevertheless, a common practice in space–time monitoring of volcanic systems is to set up plans of investigation in which the same variable is measured on each experimental unit at a number of different occasions, and observations that are made at different times on the same experimental unit will frequently show some correlation. In general, observations made close together in time will be more highly correlated than observations taken far apart in time. Furthermore, the presence of outliers and/or groups of samples might be interpreted as a mixture of populations. All these aspects would require a complex model to study the frequency distributions of above mentioned log-ratios. The hypothesis of random sampling made in the present approach is based on the fact that the sampling time interval of available data is about six months, that the analysis of the behaviour in time of the log-ratios considered did not show any evidence of time-dependence, and that the presence of outliers and/or groups in the sample is attributable to the already mentioned influence of deep-seated aquifers of marine-like composition, which occasionally inflow into the overlying water bodies, to lateral permeability variations, and/or to local alteration processes.

To model the log-ratios $(1/\sqrt{2})\ln(Ca^{2+}/HCO_3^-)$, $(1/\sqrt{2})\ln(Na^+/K^+)$, $(1/\sqrt{2})\ln(Na^+/Cl^-)$, $(1/\sqrt{2})$, $\ln(Ca^{2+}/Mg^{2+})$, $(1/\sqrt{2})\ln(Mg^{2+}/SO_4^{2-})$, and $(1/\sqrt{2})\ln(Ca^{2+}/SO_4^{2-})$, a normal density has been used. This is equivalent to modelling the corresponding ratios using a log-normal model. The Kolmogorov–Smirnov goodness-of-fit test for normality only allows one to accept the Gaussian model for $(1/\sqrt{2})\ln(Ca^{2+}/HCO_3^-)$ ($\mu=-0.74$, $\sigma=0.72$) and $(1/\sqrt{2})\ln(Na^+/K^+)$ ($\mu=0.67$, $\sigma=0.31$), with a p-value >0.03. Their probability plots are reported in Figures 1 and 2.

If $(1/\sqrt{2})\ln(Ca^{2+}/HCO_3^-)$ and $(1/\sqrt{2})\ln(Na^+/K^+)$ are well represented by the normal model, then the ratios $(Ca^{2+}/HCO_3^-)^{1/\sqrt{2}}$ and $(Na^+/K^+)^{1/\sqrt{2}}$ follow the log-normal one. The

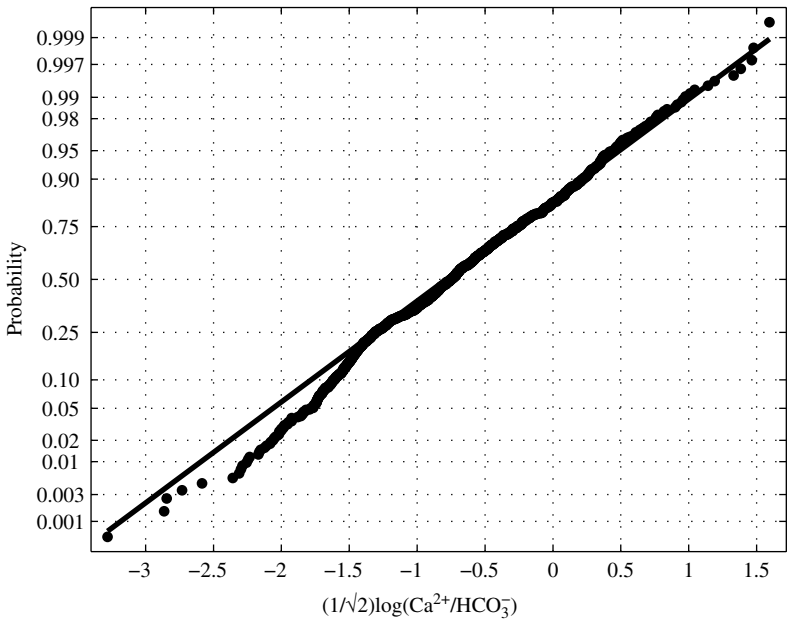

Fig. 1. Probability plot of $(1/\sqrt{2})\ln(Ca^{2+}/HCO_3^-)$. Continuous line represents a perfect normal distribution.

exponent affects only the scale of the ratios and is important to preserve consistency with the geometric properties of the full composition. Nevertheless, for interpretation, the scale can be changed, i.e. the exponent can be omitted, and the ratios Ca^{2+}/HCO_3^- and Na^+/K^+ will also follow a log-normal distribution.

From a geochemical point of view, these ratios appear to be the product of many independent random processes, so that the quantity present in

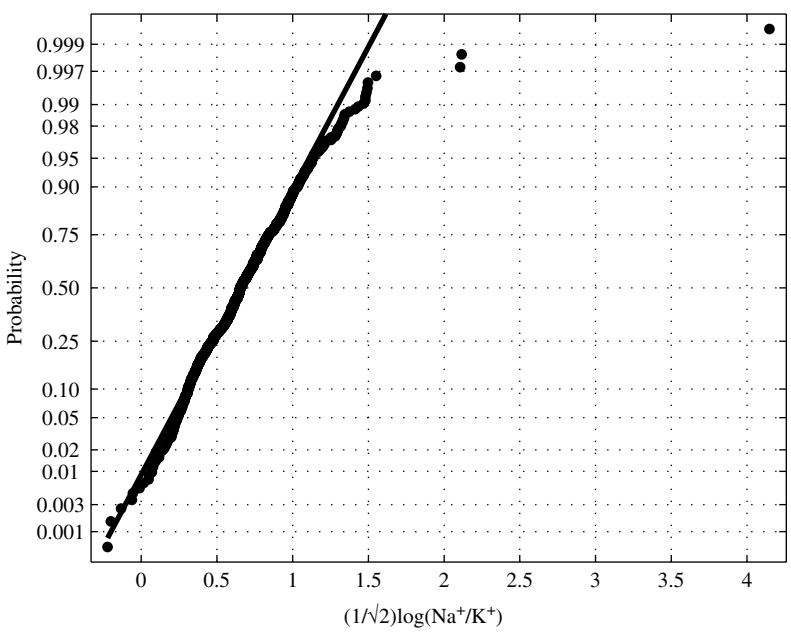

Fig. 2. Probability plot of $(1/\sqrt{2})\ln(Na^+/K^+)$. Continuous line represents a perfect normal distribution.

each state can be expressed as a random proportion of the quantity present in the immediately prior state. The chemical species involved in the ratios would have experienced considerable physical movement and agitation, as well as dilution in successive, independent stages. The resulting log-normal distributions would thus represent dominant and general phenomena affecting the investigated waters, where weathering of silicates and carbonates is important. The input of carbon dioxide from the deep uprising gaseous flow is, in fact, able to give an aggressive character to the water. It produces a weak rock weathering with consequent presence in water of Na^+, K^+, Ca^{2+} and HCO_3^-. Physical movement and agitation, as well as dilution in successive, independent stages can lead to the log-normal distribution.

The same simple modelling path cannot be accepted for the log-ratios $(1/\sqrt{2})\ln(Na^+/Cl^-)$, $(1/\sqrt{2})\ln(Ca^{2+}/Mg^{2+})$, $(1/\sqrt{2})\ln(Mg^{2+}/SO_4^{2-})$ and $(1/\sqrt{2})\ln(Ca^{2+}/SO_4^{2-})$ (all p-values of the Kolmogorov–Smirnov test are less than 0.01). Their probability plots, with the reference line of the Gaussian model, are reported in Figures 3, 4, 5 and 6. As the log-ratios display a moderate skewness, the performance of the skew-normal model or, equivalently, the logskew-normal model is explored for the corresponding ratios.

In Figures 7, 8, 9 and 10 the histograms and the estimated skew-normal curves obtained using the maximum likelihood estimation procedure (Azzalini 1985) are reported.

As can be seen, the skew-normal model appears to capture the skewness affecting the log-ratios; the log-likelihood function (a value similar to the sum of squared error in regression analysis) shows, in fact, a better value if compared with the normal model, for each log-ratio. Furthermore, the value of the likelihood ratio test statistic to compare both models (i.e. the null hypothesis that the shape parameter λ is zero) leads always to a p-value < 0.01. Thus, the conclusion is that the skew-normal model is significantly better than the normal one in all cases. As can be seen in Table 1, the shape parameter λ explains a skewness ranging from $\gamma = -0.48$ to 0.58.

Consequently, omitting the exponent for interpretational purposes, low values can be found with higher frequency for the ratios Na^+/Cl^- and Ca^{2+}/Mg^{2+}, and high values for Mg^{2+}/SO_4^{2-} and Ca^{2+}/SO_4^{2-}, when a comparison with the log-normal model is performed. However, the application of some goodness-of-fit tests for the skew-normal model (Kolmogorov–Smirnov, Kuiper, Anderson–Darling, Cramer–von Mises and Watson (Mateu-Figueras 2003)) indicates that only for the log-ratio $(1/\sqrt{2})\ln(Na^+/Cl^-)$ can the skew-normal model be considered statistically acceptable, taking a significance level of 0.01 (p-value > 0.01). In the other cases, the bimodality

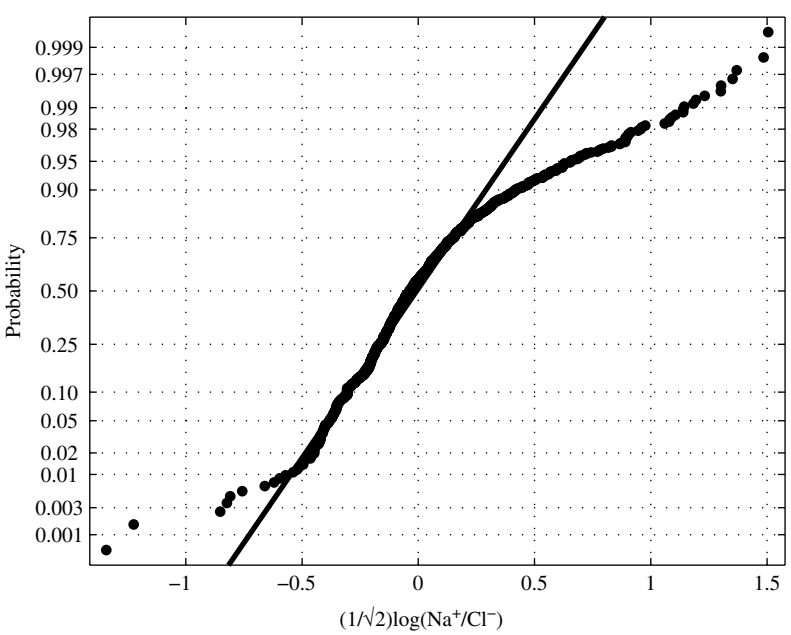

Fig. 3. Probability plot of $(1/\sqrt{2})\ln(Na^+/Cl^-)$. Continuous line represents a perfect normal distribution.

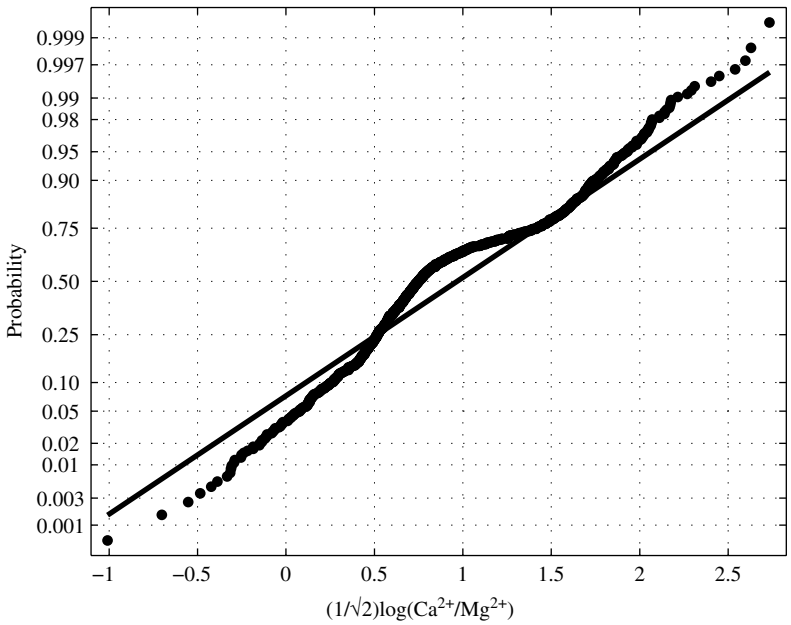

Fig. 4. Probability plot of $(1/\sqrt{2})\ln(Ca^{2+}/Mg^{2+})$. Continuous line represents a perfect normal distribution.

affecting the data indicates a more complex structure that cannot be explained considering only a moderate skewness.

The adoption of the skew-normal model for $(1/\sqrt{2})\ln(Na^{+}/Cl^{-})$ implies that Na^{+}/Cl^{-} (omitting again the exponent) follows the logskew-normal one, so that a sort of mechanism able to generate a further skewness (compared with the log-normal) is present. This result indicates that the samples have not experienced dilution as

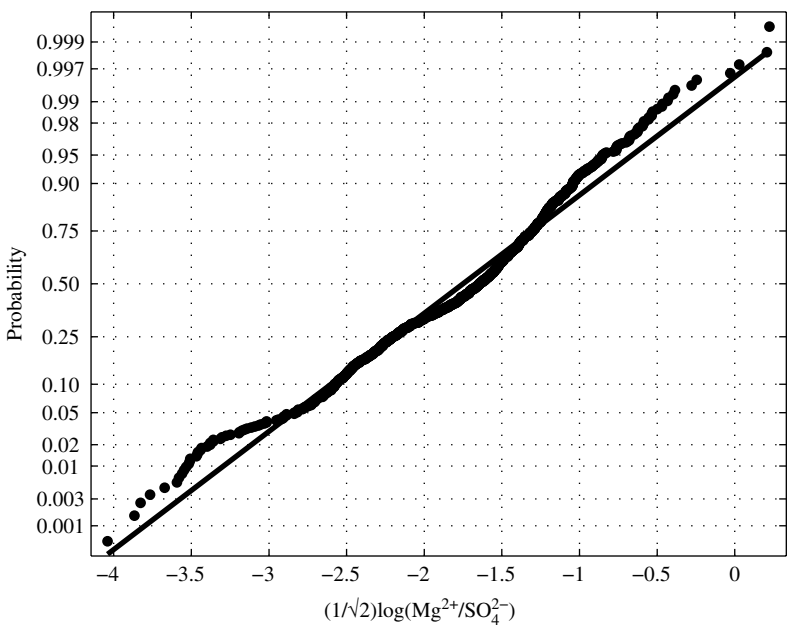

Fig. 5. Probability plot of $(1/\sqrt{2})\ln(Mg^{2+}/SO_4^{2-})$. Continuous line represents a perfect normal distribution.

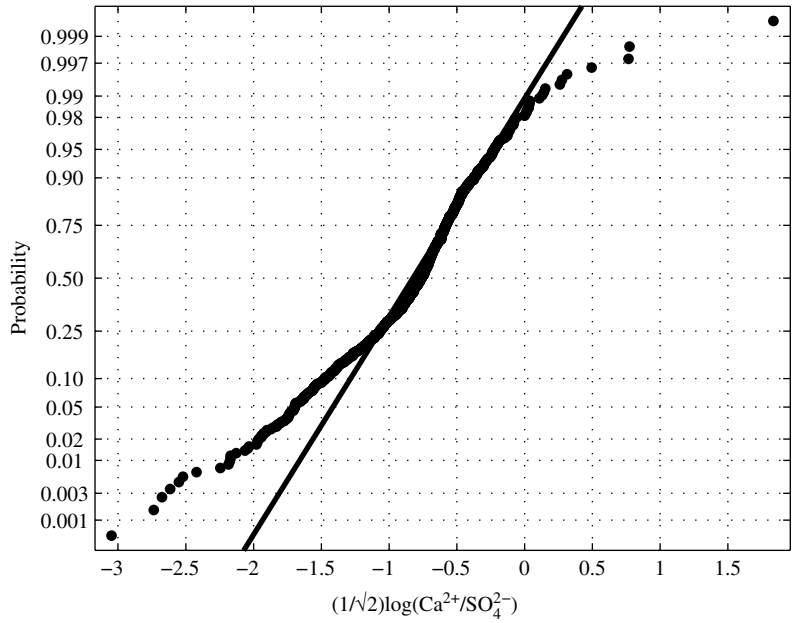

Fig. 6. Probability plot of $(1/\sqrt{2})\ln(Ca^{2+}/SO_4^{2-})$. Continuous line represents a perfect normal distribution.

the dominant process, but that the influence of marine water ($Na^+/Cl^- \sim 0.55$) is not able to generate clearly different groups of observations. In the case of the log-ratios $(1/\sqrt{2})\ln(Ca^{2+}/Mg^{2+})$, $(1/\sqrt{2})\ln(Mg^{2+}/SO_4^{2-})$ and $(1/\sqrt{2})\ln(Ca^{2+}/SO_4^{2-})$, the solution of minerals typical of weathered hydromagmatic deposits not homogeneously present in the area (i.e. sulphates) might be the mechanism able to

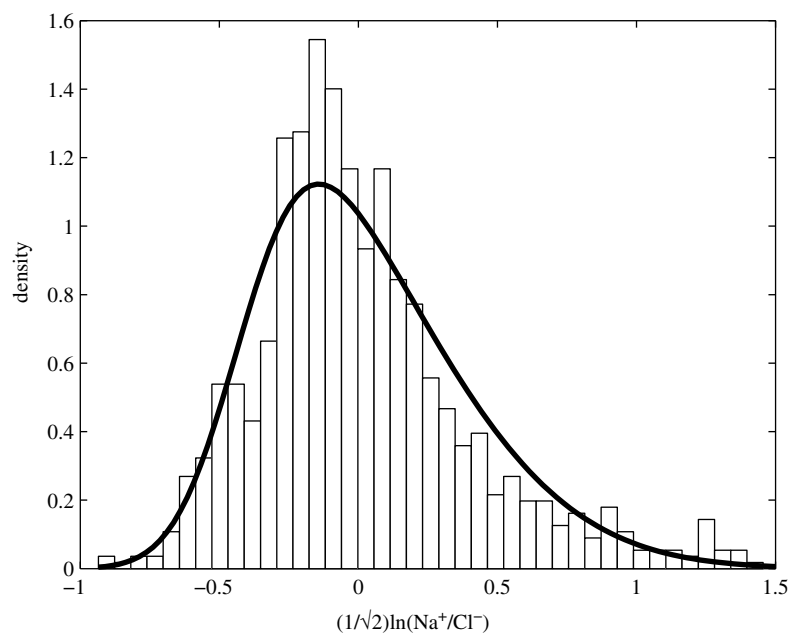

Fig. 7. Histogram of $(1/\sqrt{2})\ln(Na^+/Cl^-)$ values and fitted skew-normal densities.

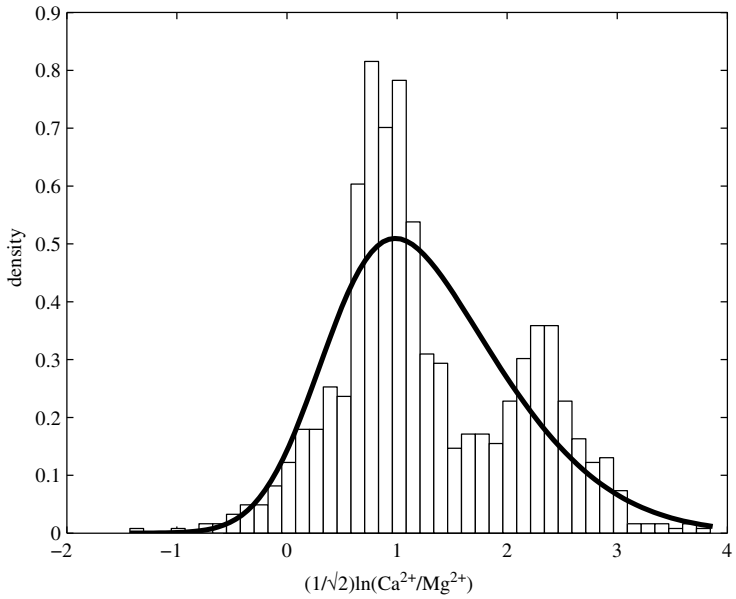

Fig. 8. Histogram of $(1/\sqrt{2})\ln(Ca^{2+}/Mg^{2+})$ values and fitted skew-normal densities.

increase the ratio of Ca^{2+}/SO_4^{2-} in part of the waters. Here, to take into account moderate skewness is insufficient to describe reality and the presence of groups of data appears to be the dominant feature.

Summarizing, univariate frequency distributions in water geochemistry of Vulcano island can be characterized by three different models, log-normal, logskew-normal and multimodal. The goodness of fit of the skew-normal (log-skew) model compared with the normal (log-normal) one depends on the persistence and continuity of natural processes affecting a part of the population so that a moderate skewness is

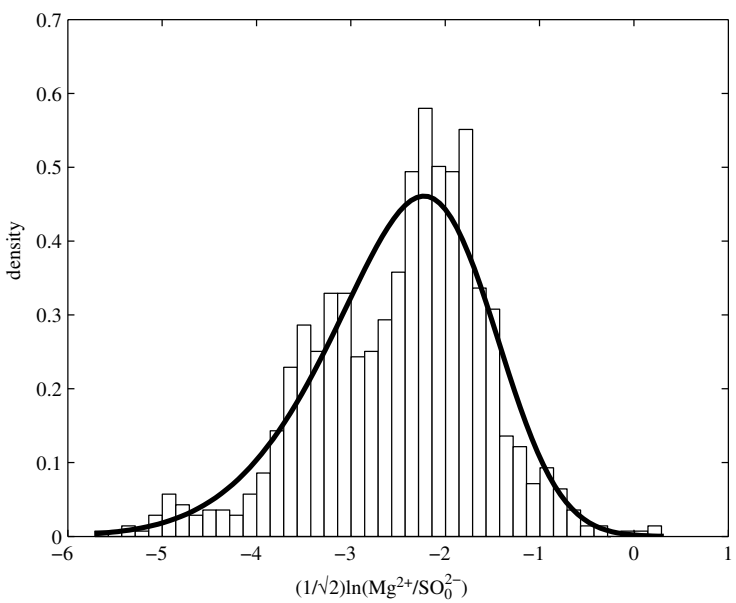

Fig. 9. Histogram of $(1/\sqrt{2})\ln(Mg^{2+}/SO_4^{2-})$ values and fitted skew-normal densities.

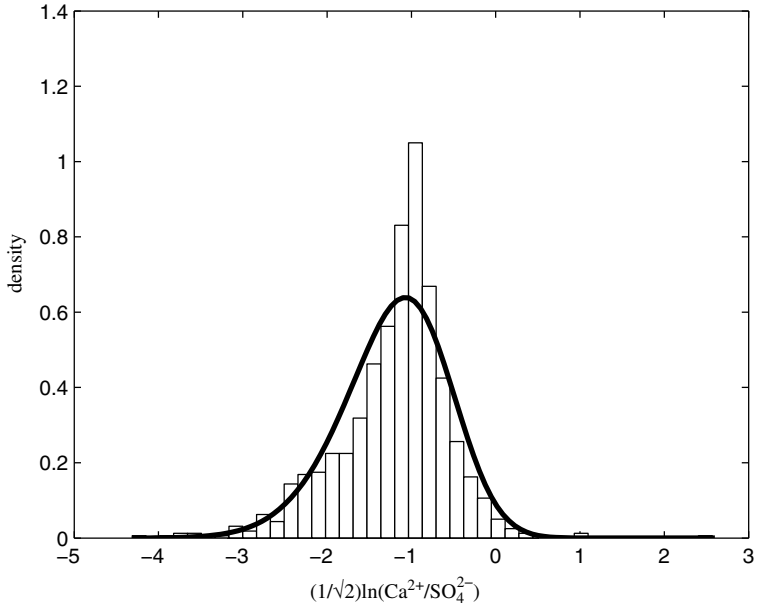

Fig. 10. Histogram of $(1/\sqrt{2})\ln(Ca^{2+}/SO_4^{2-})$ values and fitted skew-normal densities.

generated. However, when the geochemical phenomena affect with recurrence only a part of the population, groups tend to develop and a more complicated analysis is needed. The geochemical–statistical approach presented can be used to verify how successive independent stages of dilution are able to describe the behaviour of the data and how other processes are able to sign their presence and persistence in time and/or space until generation of groups of cases occur.

Log-ratios and their multivariate statistical modelling in water chemistry: results and discussion

In the previous part, log-ratios were analysed using a geochemical–statistical procedure to investigate their behaviour. The terms of the log-ratios were chosen following geochemical criteria.

Table 1. *Value of the estimated shape parameter, $\hat{\lambda}$, and the corresponding estimated skewness index, $\hat{\gamma}$, of each log-ratio*

Log-ratio	$\hat{\lambda}$	$\hat{\gamma}$
$(1/\sqrt{2})\ln(Na^+/Cl^-)$	2.97	0.55
$(1/\sqrt{2})\ln(Ca^{2+}/Mg^{2+})$	2.52	0.58
$(1/\sqrt{2})\ln(Mg^{2+}/SO_4^{2-})$	−2.11	−0.48
$(1/\sqrt{2})\ln(Ca^{2+}/SO_4^{2-})$	−1.74	−0.38

However, water chemistry can be considered also as a whole, and to do so the shape of the frequency distribution of the composition $\mathbf{X} = (Na^+, K^+, Ca^{2+}, Mg^{2+}, HCO_3^-, SO_4^{2-}, Cl^-)$ should be considered. In this case, one can verify if $\mathbf{X}$ follows a logistic normal distribution, or a logistic skew-normal one. The importance of processes able to introduce skewness in single log-ratios is considered when all the members of the composition are analysed together. The adequacy of the models was investigated applying the isometric log-ratio transformation, $\mathrm{ilr}(\mathbf{X}) = (\mathrm{ilr}_1, \mathrm{ilr}_2, \mathrm{ilr}_3, \mathrm{ilr}_4, \mathrm{ilr}_5, \mathrm{ilr}_6)$, given by Egozcue *et al.* (2003)

$$
\begin{aligned}
\mathrm{ilr}_1 &= \frac{1}{\sqrt{2}}\ln\left(\frac{Na^+}{K^+}\right),\\
\mathrm{ilr}_2 &= \frac{1}{\sqrt{6}}\ln\left(\frac{Na^+K^+}{(SO_4^{2-})^2}\right),\\
\mathrm{ilr}_3 &= \frac{1}{\sqrt{12}}\ln\left(\frac{Na^+K^+SO_4^{2-}}{(Cl^-)^3}\right),\\
\mathrm{ilr}_4 &= \frac{1}{\sqrt{20}}\ln\left(\frac{Na^+K^+SO_4^{2-}Cl^-}{(HCO_3^-)^4}\right),\\
\mathrm{ilr}_5 &= \frac{1}{\sqrt{30}}\ln\left(\frac{Na^+K^+SO_4^{2-}Cl^-HCO_3^-}{(Ca^{2+})^5}\right),\\
\mathrm{ilr}_6 &= \frac{1}{\sqrt{42}}\ln\left(\frac{Na^+K^+SO_4^{2-}Cl^-HCO_3^-Ca^{2+}}{(Mg^{2+})^6}\right).
\end{aligned}
\tag{1}
$$

Then, a multivariate normal model and a multivariate skew-normal model are fitted to the transformed samples. The maximum likelihood method is used to obtain estimates of the parameters. The two fitted models are compared by applying the likelihood ratio test (i.e. the null hypothesis that all the components of the shape parameter $\boldsymbol{\lambda}$ are 0). As the p-value is <0.01 (the value of the associated statistic is approximately 233), the conclusion is that the multivariate skew-normal model is significantly better than the multivariate normal one. Thus, the composition of Vulcano water $\mathbf{X}$ is described better by the logistic skew-normal model, compared with the logistic normal one. But it is necessary to validate the logistic skew-normal model for the composition $\mathbf{X}$ or, equivalently, the skew-normal model for the ilr-transformed vector. The univariate skew-normality of each component is a necessary but not sufficient condition for the skew-normality of the whole vector. Application of goodness-of-fit tests on the univariate distributions (Watson, Anderson–Darling, Cramer–von Mises, Kolmogorov–Smirnov), as reported in Mateu-Figueras (2003), indicates that, except for the first log-ratio, the other single log-ratios cannot be statistically well described by a univariate skew-normal model (p-value < 0.01). As can be seen in Table 2, the log-ratios show a λ value ranging from -2.96 to 3.39, corresponding to skewness indices ranging from $\gamma = -0.66$ to 0.72; However, investigations on the univariate frequency distributions allow verification of a strong presence of multimodality, as well as of outliers (Figs 11 and 12), as expected from the univariate analysis presented in the previous section.

Table 2. *Value of the estimated shape parameter, $\hat{\lambda}$, and the corresponding estimated skewness index, $\hat{\gamma}$, of each log-ratio*

Component	$\hat{\lambda}$	$\hat{\gamma}$
ilr_1	1.94	0.44
ilr_2	−2.96	−0.66
ilr_3	1.72	0.37
ilr_4	−1.17	−0.19
ilr_5	−0.58	−0.04
ilr_6	3.39	0.72

The multivariate skew-normal model can also be validated considering the Mahalanobis distances $d_i = (\mathbf{y_i} - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y_i} - \hat{\boldsymbol{\mu}})$, where $\mathbf{y_i}$ represent each

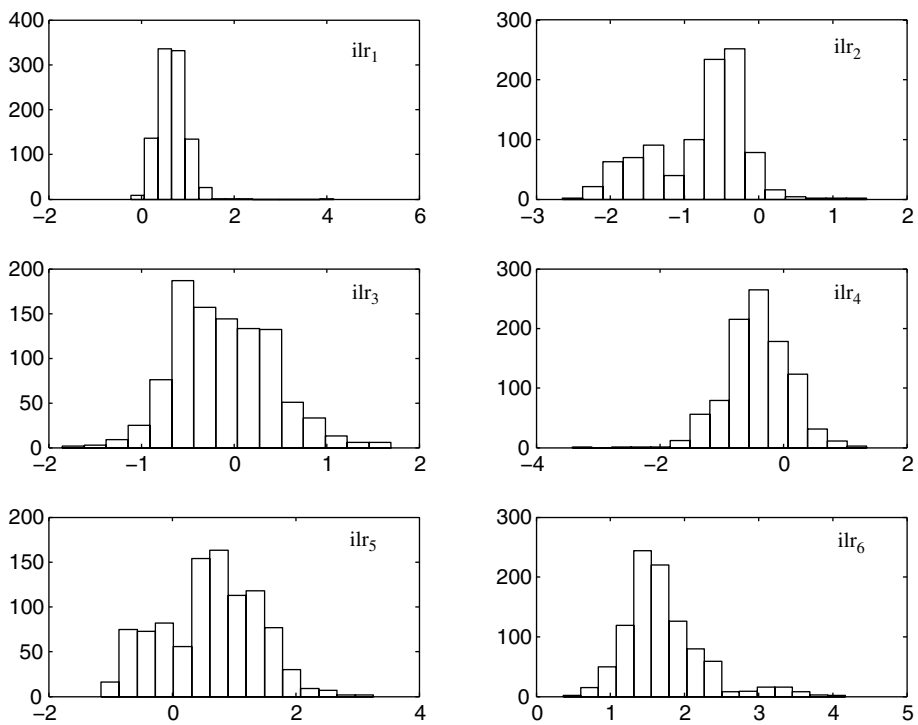

Fig. 11. Histograms of the investigated log-ratios.

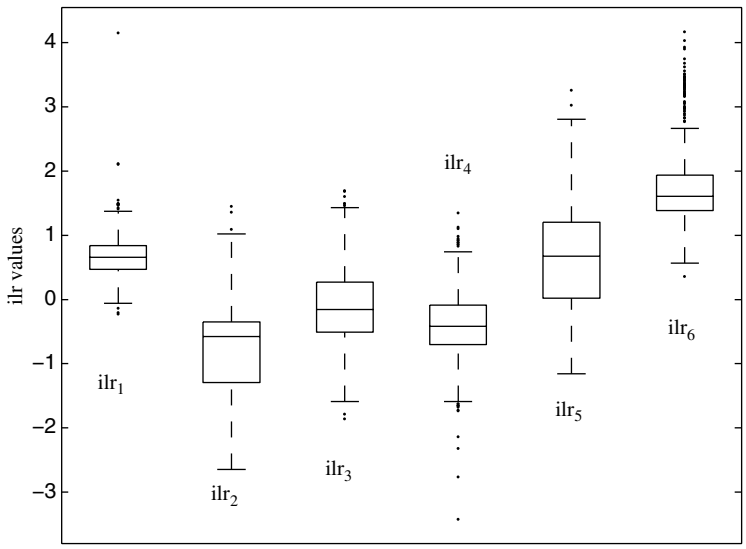

Fig. 12. Box-plots of the investigated log-ratios.

observation of the vector $\mathbf{Y}$ ($i = 1, \ldots, 977$) and $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the estimates of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of a skew-normal model (Azzalini & Capitanio 1999). If the log-ratio vector $\mathbf{Y}$ has a multivariate skew-normal distribution, then the Mahalanobis distances are sampled from a χ^2_5 distribution. However, application of graphical tests (P–P plot and Q–Q plot of the Mahalanobis distance versus χ^2_5 values shown in Fig. 13) and numerical tests (based on the Anderson–Darling, Cramer–von Mises and Kolmogorov–Smirnov statistics) to validate the χ^2_5 distribution of values d_i allow one to conclude that the multivariate skew-normal model is not yet able to describe statistically the composition in an exhaustive way and a more complex investigation about the presence of groups and anomalous values is needed. The water composition of some wells at Vulcano island appears thus to be affected by persistence phenomena (i.e. influence of uprising gas and/or presence of secondary minerals) able to generate isolated groups of observations with homogeneous behaviour.

Conclusions

In any aquifer chemical relationships between the different species are affected by the development of acid–base and redox reactions, solution–precipitation processes and adsorption phenomena. Which process dominates at any time depends on the mineralogy of the aquifer, the hydrogeological environment and the history of the groundwater movement (i.e. residence time). In the investigated volcanic environment the mobilization of chemical species as the result of fumarolic activity and of the alteration of volcanic products directly affects the unconfined aquifer feeding the wells of Vulcano. Several investigations in this area indicate that the weak acidity of the circulating solutions, able to leach chemical species, is attributable to the volcanic CO_2 provided at a discontinuous rate to the shallow-water body. Consequently, the mobilization depends on the CO_2 input, on the rate of neutralization by rock weathering and on the intensity of rainfall acting as a diluting factor. In this general framework, if solutions circulate in volcanic products involved in syn-depositional changes, significant quantities of calcium sulphate, sodium chloride, fluorine and other trace elements can be provided to the groundwater and no natural mechanism can remove them to a significant extent. The only limiting factor is the saturation of the solution with respect to the mineral. Further occasional contributions have to be ascribed to marine-like solutions that inflow into the overlying water bodies. In this situation, investigation of the shape of the frequency distribution of single log-ratios for the Vulcano waters (properly chosen from a geochemical point of view) has been used to verify: (1) whether one general process, dominated by dilution, is able to describe the behaviour of the data; or (2) whether further processes are overlapping dilution, so that a moderate negative or positive skewness is present; or (3) whether these processes are important enough to generate a

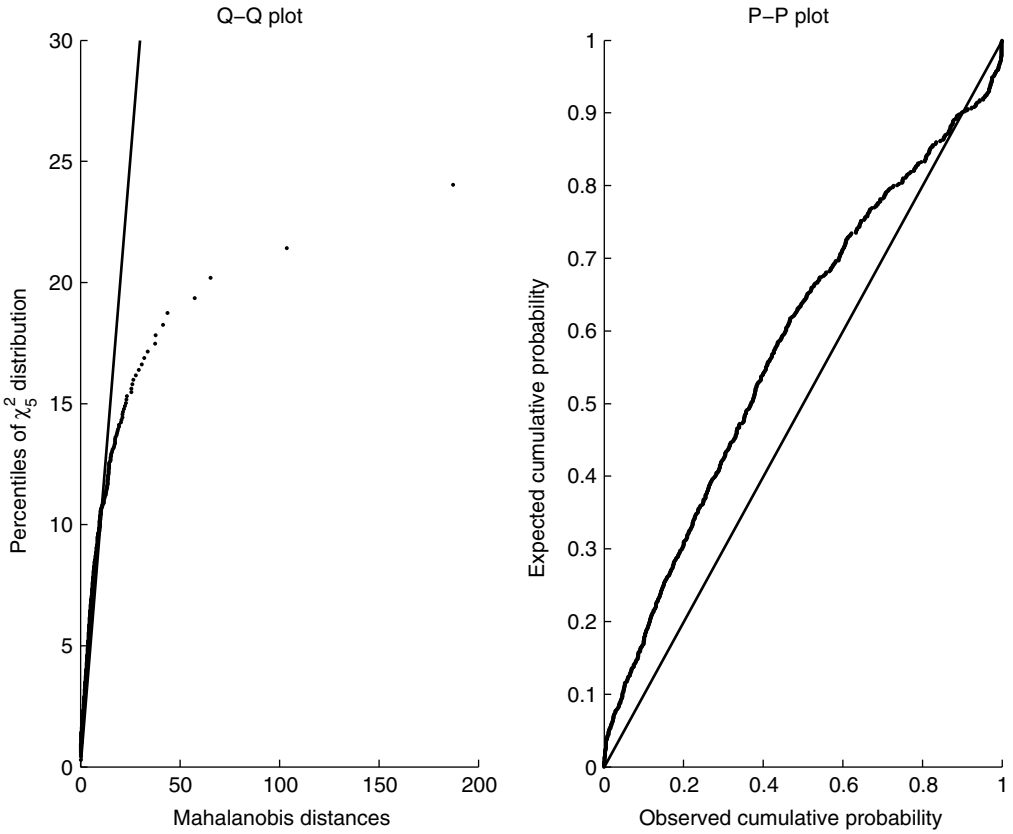

Fig. 13. P–P Plot and Q–Q plot for multivariate skew-normal distribution.

multimodal complex distribution. The analysis can be extended to the multivariate case, choosing an appropriate transformation. It has indicated that the multivariate distribution is described better by the skew-normal model compared with the normal one. Thus, the processes affecting the composition would tend to follow a log-skew normal law and dilution is not the only present and dominant phenomenon. However, the multivariate skew-normal model is not yet completely adequate to describe the simultaneous relationships among the variables. The presence of bimodality in some marginals, as well as of anomalous values, may be the underlying reason.

Summarizing the results, in water chemistry univariate frequency distributions of log-ratios can show three fundamental features, log-normal, logskew-normal, and overlapping processes leading to bimodality. In this context, the skew-normal distribution family appears to have an important intermediate role in a better description of natural phenomena where dilution is not the only phenomenon, but the persistence and continuity of other processes has not yet clearly generated different groups of observations. When marginal distributions of all the three previous types are considered together in a multivariate framework, the reciprocal relationships are complex. In such a case, the multivariate skew-normal model may be better than the normal one, but it is not yet adequate to describe the whole composition. Other statistical procedures are required to identify samples pertaining to different potential groups.

This research has been financially supported by Italian MIUR (Ministero dell'Istruzione, dell'Università e della Ricerca Scientifica e Tecnologica), PRIN 2004, through the GEOBASI project (prot. 2004048813-002) and by the Dirección General de Enseñanza Superior (DGES) of the Spanish Ministry for Education and Culture through the project BFM2003-05640.

References

AHRENS, L. H. 1953. A fundamental law of geochemistry. *Nature*, **172**, 1148.

AHRENS, L. H. 1954*a*. The lognormal distribution of the elements a fundamental law of geochemistry and its subsidiary. *Geochimica et Cosmochimica Acta*, **6**, 49–74.

AHRENS, L. H. 1954*b*. The lognormal distribution of the elements. ii. *Geochimica et Cosmochimica Acta*, **6**, 121–132.

AHRENS, L. H. 1957. Lognormal type distribution. iii. *Geochimica et Cosmochimica Acta*, **11**, 205–213.

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

AITCHISON, J. 1986. *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.

ALLÈGRE, C. J. & LEWIN, E. 1995. Scaling laws and geochemical distributions. *Earth and Planetary Science Letters*, **132**, 1–13.

AUBREY, K. V. 1954. Frequency distribution of the concentrations of the elements in rocks. *Nature*, **174**, 141–142.

AUBREY, K. V. 1956. Frequency distributions of elements in igneous rocks. *Geochimica et Cosmochimica Acta*, **9**, 83–90.

AZZALINI, A. 1985. A class of distribution which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171–178.

AZZALINI, A. & CAPITANIO, A. 1999. Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society, Series B*, **61** (3), 579–602.

AZZALINI, A. & DALLA VALLE, A. 1996. The mutlivariate skew-normal distribution. *Biometrika*, **83** (4), 715–726.

BUCCIANTI, A. & PAWLOWSKY-GLAHN, V. 2005. New perspectives on water chemistry and compositional data analysis. *Mathematical Geology*, **37** (7), 703–727.

CAPASSO, G., FAVARA, R., FRACOFONTE, S. & INGUAGGIATO, S. 1999. Chemical and isotopic variations in fumarolic discharge and thermal waters at Vulcano island Aeolian islands, Italy during 1996: evidence of resumed volcanic activity. *Journal of Volcanology and Geothermal Research*, **88**, 167–175.

CAPASSO, G., D'ALESSANDRO, W., FAVARA, R., INGUAGGIATO, S. & PARELLO, F. 2001. Interaction between the deep fluids and the shallow groundwaters on Vulcano island (Italy). *Journal of Volcanology and Geothermal Research*, **108**, 187–198.

CHAYES, F. 1954. The lognormal distribution of elements: a discussion. *Geochimica et Cosmochimica Acta*, **6**, 119–121.

DAVIES, S. N. & DEWIEST, R. C. M. 1996. *Hydrogeology*. Wiley and Sons, New York.

DI LIBERTO, V., NUCCIO, P. M. & PAONITA, A. 2002. Genesis of chlorine and sulphur in fumarolic emissions at Vulcano island (Italy): assessment of ph and redox conditions in the hydrothermal system. *Journal of Volcanology and Geothermal Research*, **116**, 137–150.

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-VIDÁL, C. 2003. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, **35** (3), 279–300.

KAPTEYN, J. C. 1903. *Skew Frequency Curves in Biology and Statistics*. Astronomical Laboratory, Groningen, Noordhoff.

MARTINI, M. 1980. Geochemical survey on the phreatic waters of vulcano (Aeolian Islands, Italy) *Bulletin of Volcanology*, **43** (1), 265–274.

MARTINI, M. 1989. The forecasting significance of chemical indicators in areas of quiescent volcanism: examples from bulcano and phlegrean fields (Italy). *In*: LATTER, J. H. (ed.) *Volcanic Hazard, IAV-CEI Proceedings in Volcanology 1*. Springer-Verlag, Berlin Heidelberg, Germany, 372–383.

MARTINI, M. 1996. Chemical characters of the gaseous phase in different stages of volcanism: precursors and volcanic activity. *In*: SCARPA, R. & TILLING, R. I. (eds) *Montoring and Mitigation of Volcanic Hazard*. Springer-Verlag, Berlin Heidelberg, Germany, 200–219.

MATEU-FIGUERAS, G. 2003. *Models de distribució sobre el símplex*. PhD thesis, Universitat Politècnica de Catalunya, Barcelona, Spain.

MATEU-FIGUERAS, G., PAWLOWSKY-GLAHN, V. & BARCELÓ-VIDAL, C. 2005. The additive logistic skew-normal distribution on the simplex. *Stochastic Environmental Reserach and Risk Assessment* (*SERRA*), **19** (3), 205–214.

MCGRATH, S. P. & LOVELAND, P. J. 1992. *The Soil Geochemical Atlas of England and Wales*. Blackie Academic, London.

MILLER, R. L. & GOLDBERG, E. D. 1955. The normal distribution in geochemistry. *Geochimica et Cosmochimica Acta*, **8**, 53–62.

MONTALTO, A. 1996. Signs of potential renewal of eruptive activity at La Fossa (Vulcano, Aeolian Islands). *Bulletin of Volcanology*, **57**, 483–492.

REIMANN, C. & FILZMOSER, P. 1999. Normal and lognormal data distribution in geochemistry: death of a myth. consequences for the statistical tretment of geochemical and environmental data. *Environmental Geology*, **39** (9), 1001–1014.

VISTELIUS, A. B. 1960. The shew frequency distributions and the fundamental law of the geochemical processes. *Journal of Geology*, **68**, 1–22.

WAYNE, R. O. 1990. A physical explanation of the lognormality of pollutant concentrations. *Journal of Air & Waste Management Associaties*, **40** (10), 1378–1383.