Mixing compositions and other scales

K. G. van den Boogaart[1], **and R. Tolosana-Delgado**[2]
[1] University of Greifswald, Germany and McGill University, Montreal Canada
boogaart@uni-greifswald.de [2] University of Göttingen, Germany

Abstract

Theory of compositional data analysis is often focused on the composition only. However in practical applications we often treat a composition together with covariables with some other scale. This contribution systematically gathers and develop statistical tools for this situation. For instance, for the graphical display of the dependence of a composition with a categorical variable, a colored set of ternary diagrams might be a good idea for a first look at the data, but it will fast hide important aspects if the composition has many parts, or it takes extreme values. On the other hand colored scatterplots of ilr components could not be very instructive for the analyst, if the conventional, black-box ilr is used.

Thinking on terms of the Euclidean structure of the simplex, we suggest to set up appropriate projections, which on one side show the compositional geometry and on the other side are still comprehensible by a non-expert analyst, readable for all locations and scales of the data. This is e.g. done by defining special balance displays with carefully-selected axes. Following this idea, we need to systematically ask how to display, explore, describe, and test the relation to complementary or explanatory data of categorical, real, ratio or again compositional scales.

This contribution shows that it is sufficient to use some basic concepts and very few advanced tools from multivariate statistics (principal covariances, multivariate linear models, trellis or parallel plots, etc.) to build appropriate procedures for all these combinations of scales. This has some fundamental implications in their software implementation, and how might they be taught to analysts not already experts in multivariate analysis.

Key words: alr, clr, MANOVA, multivariate modeling, regression, transformation

1 Introduction to scales

Compositions have been introduced as a new scale additional to the classical scales according to e.g. Stevens (1946): nominal, ordinal, interval and ratio (positive real). This list of different scales has often been criticized to be to limited. We will therefore discuss a more extensive list, selected according to our own understanding of important scales and the methods we would like to propose. To avoid misunderstandings we give a short definition.

- **Categorical**: each statistical individual is *classified* in exactly one of some a priori defined categories; e.g. species, facies, treatment. Ideally, the sampling procedure is defined to evenly cover all categories.
- **Nominal**: each statistical individual belongs to exactly one group. This is like categorical, but thee groups are not a priori defined, and typically *label* the individual, more than *classify* it; e.g. families, treating doctor, interviewer.
- **Dichotomous**: like categorical, but only with two categories; e.g. sex, before/after treatment, dead/alive, yes/no.
- **Sets**: like categorical, but each individual can belong to more than one category; e.g. cause of being included into the statistical population.
- **Ordinal**: Like categorical, but additionally there is a natural order defined on the categories; e.g.: strongly agree - agree - neutral - disagree - strongly disagree, coarse-medium-fine sand.
- **Interval (discrete)**: like the ordinal, but additionally the difference or "increment" between any two consecutive categories is considered constant in some sense.
- **Real / interval (continuous)**: the observations are represented by numbers and the natural geometry is represented by their arithmetic differences; e.g. gravity or electric potentials.
- **Ratio (i.e. positive real)**: the observations are represented by positive real numbers. The natural difference is represented by their (log)-ratios; e.g. amounts, distances, prices.
- **Portions**: the observations are a portion of a total, and they are represented by real numbers in the open interval (0,1) or percentages between 0 and 100% ; e.g. concentrations, volume percentages, mass percentages, proportion of women in CEO positions.
- **(Real) compositions** : the observation is represented by a vector of positive real numbers giving the relative amounts of different parts. Observations scaled by an arbitrary positive real represent the same composition; e.g. geochemical compositions, contribution of economic sectors to GDP.
- **(Count) compositions** The observation is represented by a vector of non-negative integer numbers giving the observed relative amount of different groups. The total number of observations may be considered irrelevant; e.g. species compositions, quartz-feldspats-feldspatoid abundances determined by point-counting, votes given to each party.

In this view a compositional information is seen as a "single variable" with a compositional scale. Although the information of this variable is internally represented by several numbers and thus inherently multivariate, it is also just a single element representing some non-separable information.

Discussing compositional data analysis from the perspective of looking at the compositional information *only* is in many aspects similar to univariate statistics. In this contribution we go towards a "bivariate" approach, offering a systematic set of proposals on how compositions can be represented and analyzed in datasets which also provide additional secondary information from another scale. We will see that in this context the many questions can be easily solved by applying classical multivariate techniques and the principle of working in coordinates.

Very much has been written in the past years about how to analyze a compositional dataset, and we will not repeat it here. The reader is referred to Aitchison and Egozcue (2005) for a review and to Pawlowsky-Glahn (2003) for a notation introduction. In contrast, few has been done on such a joint statistical analysis of compositions with other scales. The exceptions are: discrimination of several categories using compositions as explanatory variables (e.g. Barceló-Vidal, 1996; Thomas and Aitchison, 1998; Tolosana-Delgado et al., 2004), regression or ANOVA with a compositional response (e.g. Daunis-i Estadella et al., 2002; Thomas and Aitchison, 2005; Tolosana-Delgado and von Eynatten, pted), and canonical correlation and similar interactions between two compositions (e.g. Aitchison, 1997; Puig et al., 2008).

In this *bivariate* (in the sense of two-scale+multivariate) approach we must solve several basic problems,

- **display** the dependence between both variables graphically,
- **model** the dependence or the conditional distribution,
- **describe** the dependence graphically and numerically,
- **quantify** the dependence numerically,
- **test** for the dependence itself, and test the hypothesis of the model,

for two situations, namely that the composition is the response or that it is the explanatory variable. Each of these 2×5 problems will be addressed for the 11 scales, grouped in three kinds: univariate discrete scales, univariate continuous scales, and inherently multivariate scales. Two data sets will be used to illustrate the several methods and techniques presented. One is the petrographic composition on quartz (poly- and mono-crystalline) vs. feldspar vs. rock fragments, of some modern sand samples from a Precambrian metasedimentary bedrock, divided in 2 slightly different subareas: North vs. South, being the southern slightly more mature, i.e. quartz rich (Blue Ridge Mountains, NC, USA Grantham and Velbel, 1988). These compositions are complemented with information on the relief ratio and the annual average precipitation of the watershed where the samples were obtained. The second illustration data set is a joint normal simulation of a composition and a continuous interval variable, with some mutual dependence.

2 Compositions and discrete scales

2.1 Principles for constructing graphics with discrete covariables

To **display** the dependence of a compositional variable on a discrete variable (nominal, categorical, dichotomous, ordinal, or discrete interval), we can use different extension principles.

1. **Symbols/colours in plots**
   Use a compositional graphic (e.g. ternary diagrams, scatterplot matrices of log-ratios, clr, ilr or balances, biplots) which represents every statistical individual, and mark each individual by a different symbol and/or color. A legend representing the coding for the different categories is mandatory. Figure 1 is an example.

   The advantages are the easy interpretation, the possibility of a direct comparison, the small space needed for the graphic, and the possiblity to recognize the univariate compositional patterns at the same time. Disadvantages are the risk of overplotting between different categories and the loss of the intuitive perception of the groups if too many are involved or the group patterns are not clear enough.

2. **Trellis plots**
   Use any graphic (e.g. one of the above mentioned, balance boxplots, density estimates)

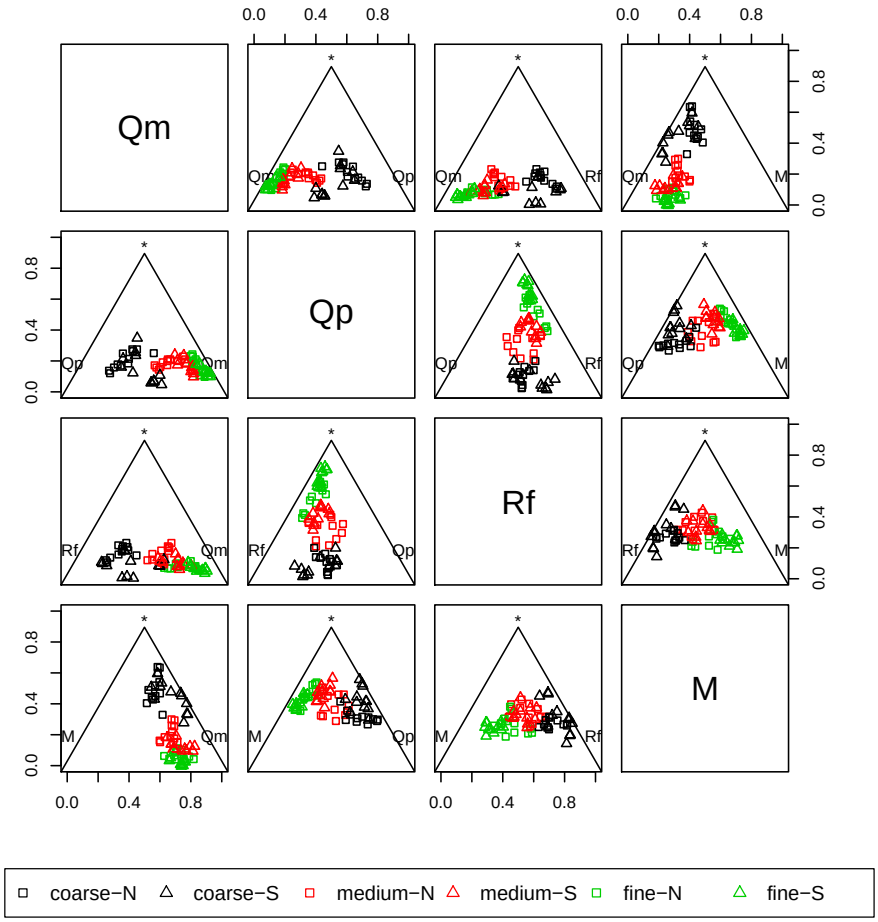

Figure 1: Sand data set, displaying the relation of a composition with a 3-category factor (using color) and a dichotomous variable (using symbols). This illustrates the symbol strategy. In each of these ternary diagrams, two parts are determined by row and column, and the third is the geometric average of the remaining parts.

displaying a compositional dataset and display it once for every category, each one only with the data from the given category. The different views should be labeled with the displayed category, as in Figure 2.

The advantages are the clear separation of the groups giving a direct perception of group membership and avoiding overplotting between different groups. A comparison between groups is possible, but relies on the ability of transferring graphical positions from one view to the other. The comparison is thus not subconscious and only possible if the differences are sufficiently pronounced. A further disadvantage lies in the large space needed.

3. **Parallel plots**
   Use univariate projections of the data, e.g. balances, and display these by an univariate graphic such as a boxplot, a dotplot or a histogram. Display these graphics parallel to each other as in the trellis graphic (Fig. 3). If multiple projections (e.g. multiple balances) are used, one would display the different projections in different parts of the graphic, such that the same projection for different categories is displayed in the same coordinate system next to each other, sharing an axis such that a direct comparison is possible.

   This principle is quite similar to that of trellis and so are advantages and disadvantages.

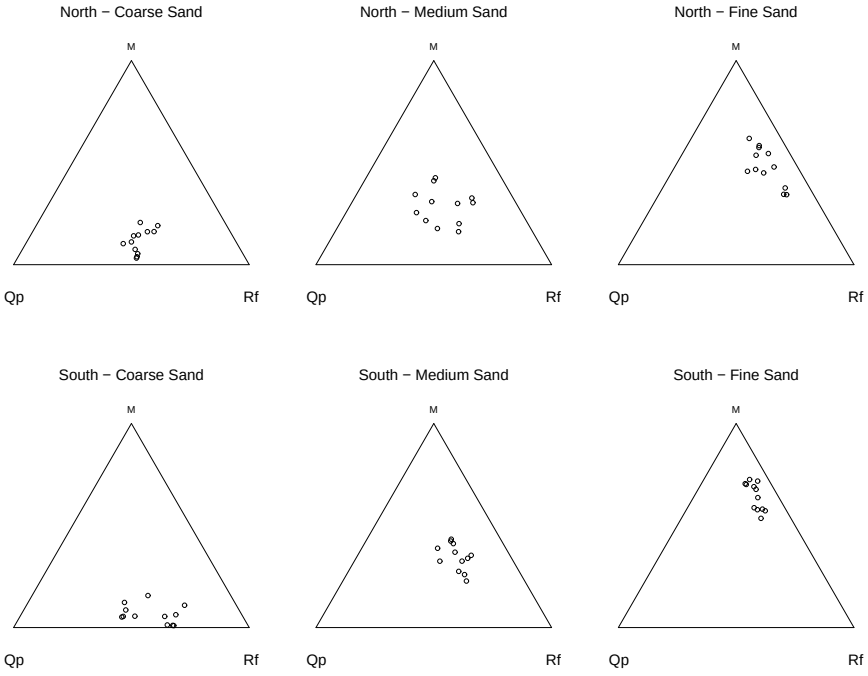

Figure 2: Subcomposition polycrystalline quartz-rock fragments-mica of the sand data set, displaying the relation of a composition with a 3-category factor and a dichotomous variable, using the trellis plot strategy.

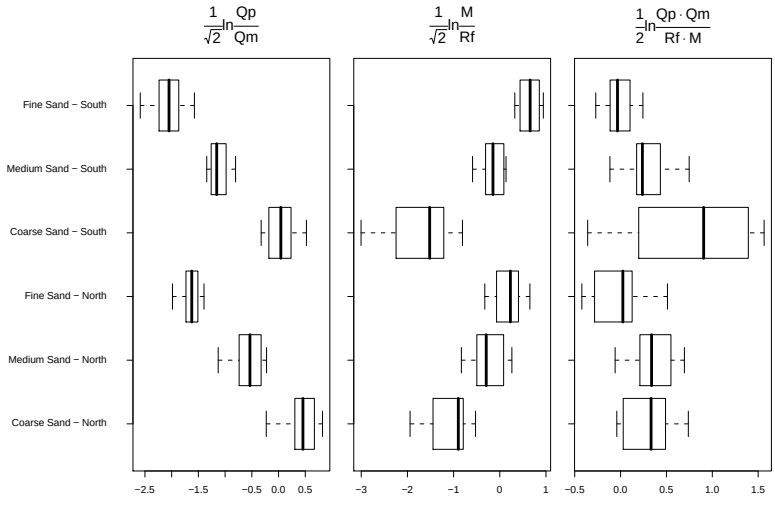

Figure 3: Coordinates of the sand data set, displaying the relation of a composition with a 3-category factor and a dichotomous variable, using the parallel plot strategy.

An advantage with respect to trellis is the simplification of the direct comparison, since the graphics to be compared are directly adjacent and use the same horizontal or vertical axis. The disadvantage lays in the restriction to univariate projections, which are more difficult to interpret by the inexperienced user and do not ease a mental reconstruction of the compositional nature of the data.

4. **The selection principle**
   If we use the overview with multiple categories in one of the above mentioned plot techniques we can replace the categorical variable by a dichotomous variable, which is coded as "is category A" and "is not category A". In this way we can compare each of the categories to the rest of the dataset one after the other. To avoid overplotting, we can plot first the non-selected observations in black, and then the selected ones using a highlighting color.

   The general advantages are clearly the direct comparability, avoiding of overplotting, an immediate quantitative comparison and the small space of graphics that needs to be displayed at the same time. The disadvantage is clearly that multiple different selections need to be used to get a complete picture. This method is thus more appropriate for an interactive data analysis rather than for printed results.

2.2 Modeling with discrete covariables

To **model** the conditional distribution of a composition $\mathbf{Y}$ given a discrete factor X with K levels, we can first reduce the problem to a multivariate regression problem by the principle of working in coordinates based on the ilr-transform. E.g. in the most simple case we get a multivariate analysis of variance model:

$$\text{ilr}(\mathbf{Y}) = \vec{a} + \vec{b}_X + \vec{\varepsilon},\ \vec{\varepsilon} \sim N(\vec{0}, \Sigma),\ \vec{a}, \vec{b}_1, \ldots, \vec{b}_K \in \mathbb{R}^{D-1}$$

which corresponds to model specify the conditional distribution as an additive logistic distribution (or normal distribution in the simplex) given by:

$$P(\mathbf{Y}|X = x) = ALN(\mathbf{a} \oplus \mathbf{b}_x, \mathbf{V}\Sigma\mathbf{V}^t)$$

where $\mathbf{a} = \text{ilr}^{-1}(\vec{a})$, $\mathbf{b}_i = \text{ilr}^{-1}(\vec{b}_i)$ and $\mathbf{V}$ is the ilr to clr transformation matrix. The notation $\oplus$ represents the perturbation operation, and $\ominus$ its inverse. the matrix defining in its columns the ilr-basis used. The estimated parameters thus have an interpretation as (ilr-transforms of) compositions and as compositional variances. They can thus be used to **describe** the dependency *numerically*, furthermore $\mathbf{a} \oplus \mathbf{b}_X$ and the corresponding ellipses defined by $\mathbf{V}\Sigma\mathbf{V}^t$ can be plotted into compositional diagrams to **display** that dependence *graphically*. Parameters $\mathbf{a}$ and $\mathbf{b}_i$ are themselves vectors, and can thus be represented accordingly, for instance as bar plots.

The multivariate analysis of variance model has associated **tests** to check for the presence of dependence. An R^2 like measure based on the trace of the variance matrices:

$$R^2_A = \frac{\sum_{i=0}^{n} \|\mathbf{a} \oplus \mathbf{b}_{x_i} \ominus \bar{\mathbf{Y}}\|^2_A}{\sum_{i=0}^{n} \|\mathbf{y}_i \ominus \bar{\mathbf{Y}}\|^2_A} \in [0, 1]$$

can be used to **quantify** the level of dependency in this model. Here $\|\cdot\|_A$ denotes the Aitchison norm, and $\bar{\mathbf{Y}}$ is the center (closed geometric mean) of the composition. The interpretation of this measure of dependence corresponds to the usual R^2.

The two general tests (that the conditional distribution is really additive log normal, and that the variance matrices are all equal) are equivalent to the corresponding test problem on the ilr-transformed data sets. Tests of additive logistic normality have been presented by Aitchison et al. (2004), and tests on the homogeneity of compositional variances can be traced back to (Aitchison, 1986, though for 2 groups and based on the alr transformation). Partial tests on the dependence

of a given subcomposition on a specific factor level are equivalent to the corresponding simple regression tests, briefly explained in section 3.2; more general independence structure tests (e.g., that two factor levels behave equally, or that a given subcomposition does not depend on any level of the factor) are quite tricky, and they will be presented in section 4, relative to multivariate regression.

2.3 Modeling the conditional distribution of the discrete variable

Modeling the conditional distribution of the discrete variable corresponds to the problem of discrimination analysis. If the conditional distribution given the category is additive logistic-normal the corresponding ilr-transform is multivariate normal. If the variances in all groups are equals the discrimination rule can thus be computed as a linear function, and if the variances in the groups are different one will use quadratic discrimination analysis, both based on the ilr-transformed compositions. A non-parametric way, especially for dichotomous variables is obviously also given by logistic regression or probit models with all elements of the ilr-transform as covariables. The (multinomial) logistic model will be reconsidered in section 4, as the vector of probabilities of belonging to each group forms a composition.

2.4 Compositions and categorical

In the case of categorical covariables the above mentioned models and methods can be applied without change.

2.5 Compositions and nominal

In case of a nominal covariable (in the above mentioned sense) the different levels of the factor have no qualitative difference. The multivariate analysis of variance model should thus be a random effects **model** rather than a fixed effects model. The proposed display of the empirical means does not provide a dataset with the correct variance, since every estimated mean has the variance of the mean and its own estimation variance. However we can estimate the mean and variance of the random effect, **display** its 2σ ellipsoid and **describe** this ellipsoid by its mean composition and its principal directions and principal variances (eigenvectors and eigenvalues respectively).

2.6 Compositions and dichotomous

A dichotomous covariable is a special case of a categorical variable with only two categories. The selection principle does therefore not provide any additional flexibility compared to symbols. Standard logistic regression can be easily used to model the conditional probabilities of the dichotomous variable given the composition:

$$\operatorname{logit}(P(X=1|\mathbf{Y}=\mathbf{y})) = a + \sum_{j=1}^{D-1} b_j \operatorname{ilr}_j(\mathbf{Y}) = a + \langle \mathbf{b}, \mathbf{Y} \rangle_A, \tag{1}$$

where $\langle \cdot, \cdot \rangle_A$ is the Aitchison scalar product of two compositions, and

$$\xi = \operatorname{logit}(x) =: \log \frac{x}{1-x} \quad \text{and} \quad x = \operatorname{logit}^{-1}(\xi) =: \frac{\exp(\xi)}{1+\exp(\xi)},$$

is the logistic transformation (and its inverse). Note that the regression coefficients $\{b_j\}$ have be interpreted as the ilr transform of a composition $\mathbf{b}$, giving the direction of the simplex in which the conditional probability of the dichotomous variable actually changes. This direction can be displayed in a ternary diagram or in a scatterplot matrix of ilr-transformed observations.

2.7 Compositions and sets

In a given way, a set variable can be seen as a set of dichotomous variables. It is thus quite difficult to use the symbol technique to display the combination of possible memberships. However trellis and parallel plot techniques and the selection principle all allow a display of each possible membership. The only problem is that a case with an empty set would not be displayed at all in the trellis or parallel plot: we thus need to add the information "empty set" as an additional category.

The multivariate analysis of variance can be extended to a multiple multivariate analysis of variance, where each of the dichotomous variables is introduced as a regressor, eventually with interactions between them. For the modeling of the conditional distribution of the set, one could either model the dichotomous covariables independently, or use a log-linear model to address their joint distribution. In this last case, each coefficient (marginal effects and interactions alike) depends on the composition by a linear model. This advanced technique is not available in software now.

2.8 Compositions and ordinal

For all proposed **display** techniques there is a natural way of introducing an order. The symbol/color technique may admit some order on the colors or symbols mirroring that of the categories: e.g. use colors with increase hue, value or saturation, follow a standard scale like the rainbow colors, or use a symbol with increasing size, a set of stars or polygonal symbols with increasing number of rays/sides/parts, or a symbolic sequence such as numbers or letters. For the trellis and the parallel plot principles, the sequence of the ordinal factor should be honored in the sequence of the graphics. With the selection principle, the time sequence of the selection can be used to express the order of the factor. In a similar way the ordinal relation should be honored when displaying results: e.g., the bar plots of coefficients $\mathbf{b} = \mathrm{ilr}^{-1}(\vec{b}_i)$ for the different categories should be ordered.

For modeling the ordinal relation many constraints have been proposed in literature. This is nevertheless not a straightforward issue to adapt to the multivariate nature of compositions, and it is left for further research and discussion.

2.9 Compositions and discrete interval

The discrete interval scale is in many ways similar to an ordinal scale and can be used in the same way. However it can also be represented by integer numbers and in this way some of the techniques proposed in the next section might be applicable to them. In particular, the problem of describing the conditional distribution of a variable with discrete interval scale on a compositional variable may be seen from the light of generalized linear models. In this case, the response variable is considered to follow a certain parametric model of probability, and a convenient set of parameters are regressed against the ilr-transformed composition. Note that Eq. (1) already developed that idea for the case of a binomial variable, which log-odds parameter (the logistic transform of the probability parameter) was regressed against the composition.

3 Compositions and one dimensional numeric scales

3.1 Visualisation of dependent compositions

The central problem of visualising a dependent composition is given by the fact that the data set has more than two dimensions, as the composition itself already has two or more. The visualisation can follow one of these three stragies.

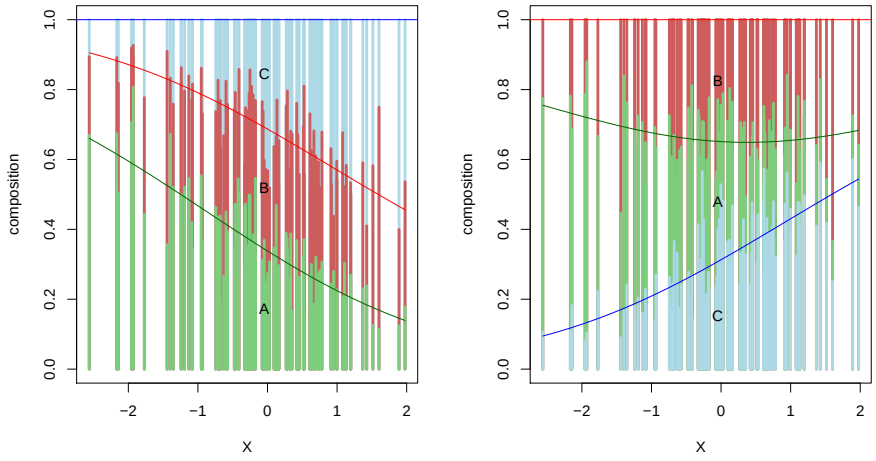

Figure 4: Two thin-bar representations of an explained composition and its dependence with respect to a continuous variable. The smooth lines correspond to the regression model. A simulated data set has been used, the same as in Figs. 5-8.

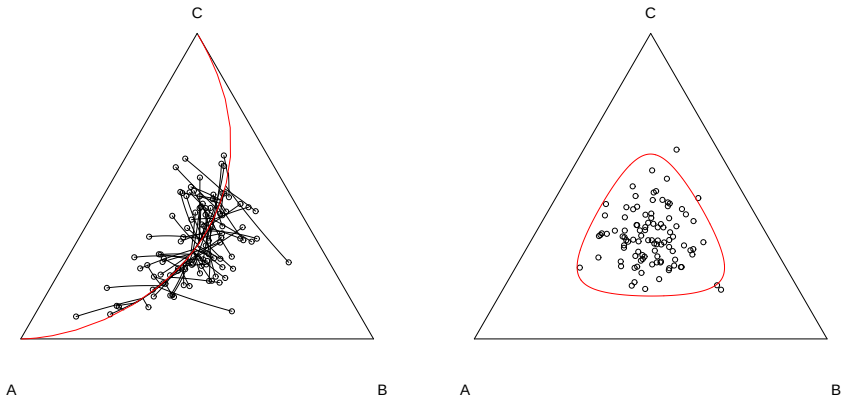

Figure 5: (Left) Displaying the dependence of a composition on a numerical covariable using a model; simulated data, the same as in Figs. 4-8. (Right) residuals of the regression, with a 95% normal confidence region based on its variance.

1. **Stacked plots:** one can express the whole composition in one single axis, with bars, lines and/or areas, in all cases better stacked.

   To actually represent the co-dependence with the numeric variable, the sample should be ordered with the later. Figure 4 shows an example with thin bars.

   As advantages, these plots cleanly show the relationship between the explanatory variable and the response composition, and they are quite familiar to spreadsheet users, as most of these software applications offer similar stacked bar and area plots. Additionally, bar plot representations can be ideal to show model parameters, as intercepts and slopes of regression models, or ANOVA-derived group means. Disadvantadges are several: it needs color or filling patterns to distinguish the parts; there is no easy way to correctly represent two individuals with exactly the same X value (critical when X is available only rounded); small parts are under-represented, and the compositional geometry is not well-represented; the general *appearance* strongly depends on the ordering of the parts (Fig. 4); and finally the fitted model may look weirdly non-linear, as what we are actually representing is the cummulative sum of some closed exponentials.

2. **Overlay plots:** one can overlay two compositional plots together, one showing the data and the other the model.

   In this case we need to draw a line following the predicted values, and to connect each data point with its prediction. The expected location serves as proxy for the explanatory variable, which is thus not directly visible in the plot. This principle is exemplified in Figure 5.

   The advantage of this type of plots lies in the direct link to the marginal plots of the composition. The figure more or less explains the marginal distribution by the dependence. Disadvantages are the high amount of inked space needed, since every statistical individual is represented by two locations linked by a connecting line, leading to a bad readability of the graphic in case of many observations. Finally, the explanatory variable itself is missing, present only through the prediction line, which might hinder the understanding of the dependence "shape".

3. **Projected compositions:** one can reduce the dimension of the composition to a one dimensional object, and represent several of these projections in parallel plots (Figs. 6-7).

   The advantage of this approach lies in the direct numerical display of the dependence. The disadvantages are those of parallel plots, mainly the difficulty to mentally reconstruct the composition as an object.

This mentioned reduction of dimensionality can be done, e.g. by a projection such as

- a subcomposition (Fig. 6),
- a balance or a coordinate (Fig. 7),
- a principal direction.

The advantage of using subcompositions lies in the easy interpretation. It is easy to locate a point in other graphics, e.g. ternary diagrams, based on subcompositions. However in these graphics the classical embedding geometry is used and thus linear dependences in the Aitchison simplex might appear as nonlinear relations (a problem shared with the stacked bar or area plot representation). The advantage of using balances, coordinates or principal values lies in the undistorted display of the Aitchison geometry. Indeed one can easily draw regression lines in these graphics to represent an appropriate linear model. However here it is quite difficult for the user to relate the displayed value to a concrete composition.

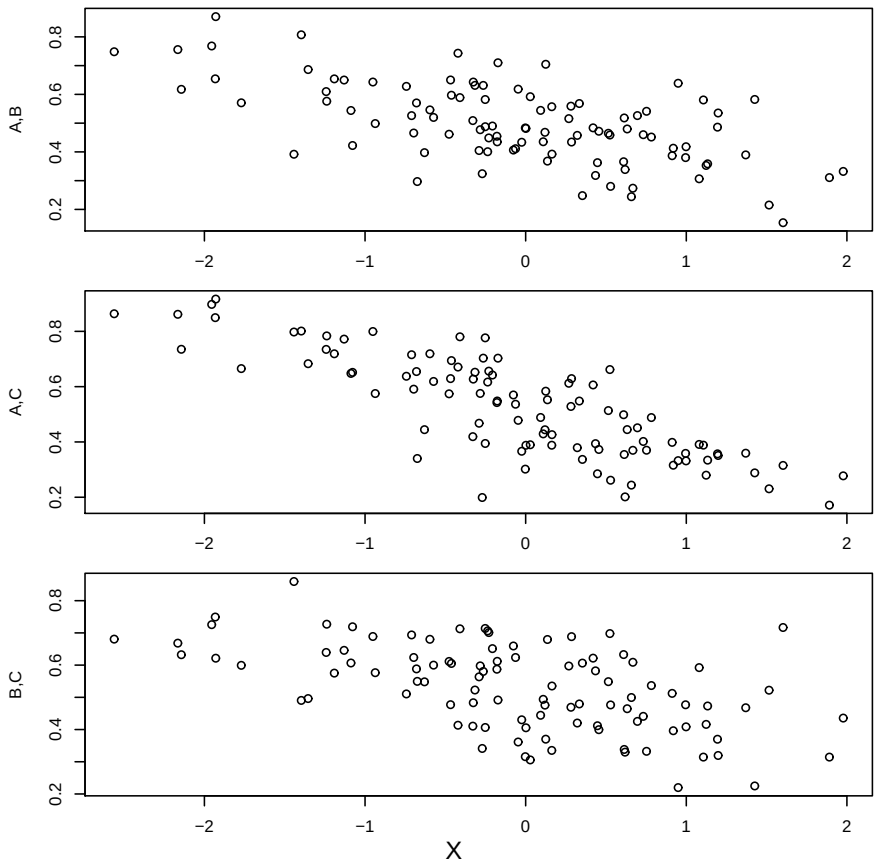

Figure 6: Displaying the dependence of a composition on a numerical covariable based on one dimensional subcompositions (simulated data, same as in Figs. 4-8).

3.2 Compositions as response variable

The **modeling** of a composition on a numeric covariable is straightforward, provided that it has (been scaled to) a real geometry. A simple linear regression of the ilr-coordinate vector on the explanatory variable will do the job:

$$\text{ilr}(\mathbf{Y}) = \vec{a} + X \cdot \vec{b} + \vec{\varepsilon},\ \vec{\varepsilon} \sim N(\vec{0}, \Sigma),\ \vec{a}, \vec{b} \in \mathbb{R}^{D-1}.$$

Again, as happened with ANOVA models of dependence on a categorical variable (section 2.2), the distribution of the explained composition is from the additive logistic normal family,

$$P(\mathbf{Y}|X = x) = ALN(\mathbf{a} \oplus X \odot \mathbf{b}, \mathbf{V}\Sigma\mathbf{V}^t),$$

where $\oplus$ and $\odot$ respectively denote perturbation and powering. The estimated parameters thus have again an interpretation as (ilr-transforms of) compositions and as compositional variances, and both can be used to numerically (in tables) or graphically (in bar plots) **describe** this dependence.

As with ANOVA, one can use multivariate ANOVA tests measure to *globally* **test** the independence and again a multivariate version of R^2 based on the trace of the variance matrices to **quantify** dependence:

$$R^2_A = \frac{\sum_{i=0}^{n} \|\mathbf{a} \oplus x_i \odot \mathbf{b} \ominus \bar{\mathbf{Y}}\|^2_A}{\sum_{i=0}^{n} \|\mathbf{y}_i \ominus \bar{\mathbf{Y}}\|^2_A} \in [0, 1], \tag{2}$$

with the usual interpretation of an R^2 coefficient (Daunis-i Estadella et al., 2002). The other global

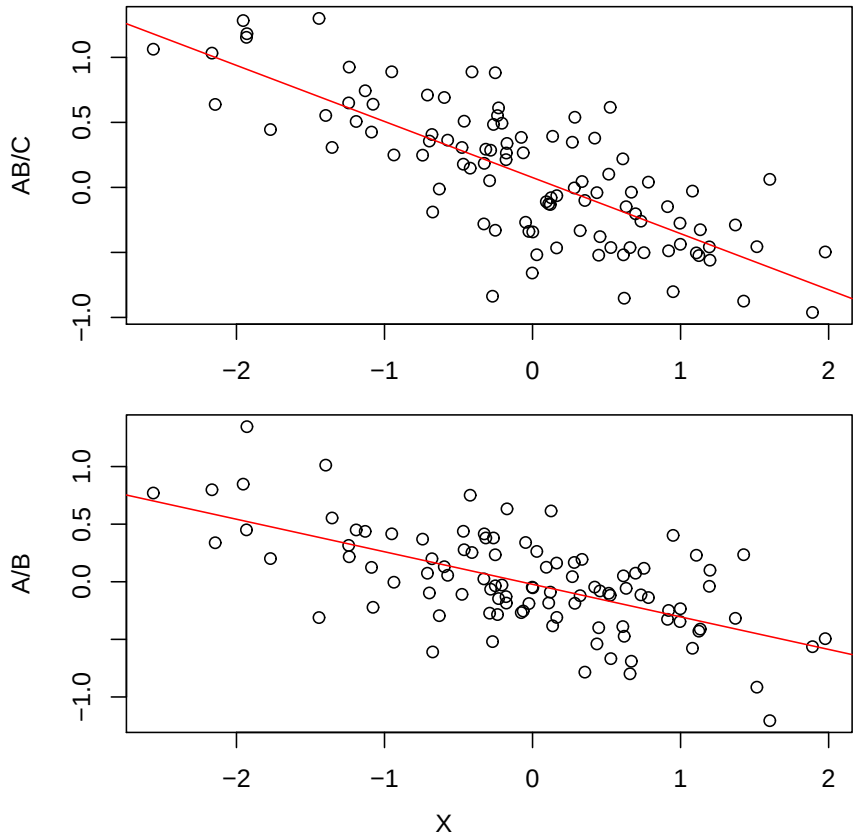

Figure 7: Displaying the dependence of a composition on a numerical covariable using a balance basis (simulated data, same as in Figs. 4-8).

test, additve logistic normality of the residuals, follows the same guidelines given in section 2.2, and presented in detail by Aitchison et al. (2004).

If one wants to test that a given part or subcomposition does not depend on the explanatory variable (a *partial independence test*), classical regression tests might serve if one carefully chooses the ilr basis following a partition scheme. We split the parts in two subcompositions, the dependent against the independent parts, build sub-bases for each group and build the balance between the two groups (Egozcue and Pawlowsky-Glahn, 2005). The target subcomposition will be totally independent of the explanatory variable if all the regression coefficients of its corresponding ilr-coordinates *and of the balance* can be simplified to zero. In the case that this last condition is not satisfied (the coefficient linked to the balance is not significantly different from zero), then the *size* of the target subcomposition depends on the explanatory variable, but the *relations* between its parts do not: this is a situation much likely to occur in practice than the total independence, though it has no sense with single-part "subcompositions".

3.3 Visualisation of dependence on a composition

The qualitative **display** of the covariation of a real variable on a composition (Fig. 8) can be achieved by using color: one can build (almost) a continuous color scale by attaching a different color to the minimum and the maximum observed values of the explained variable, and interpolating between them in one color space (e.g., RGB, CMYK, hue-value-saturation). Each point in the ternary diagram (or in a matrix of ilr-scatterplots) would be then given a color according to that

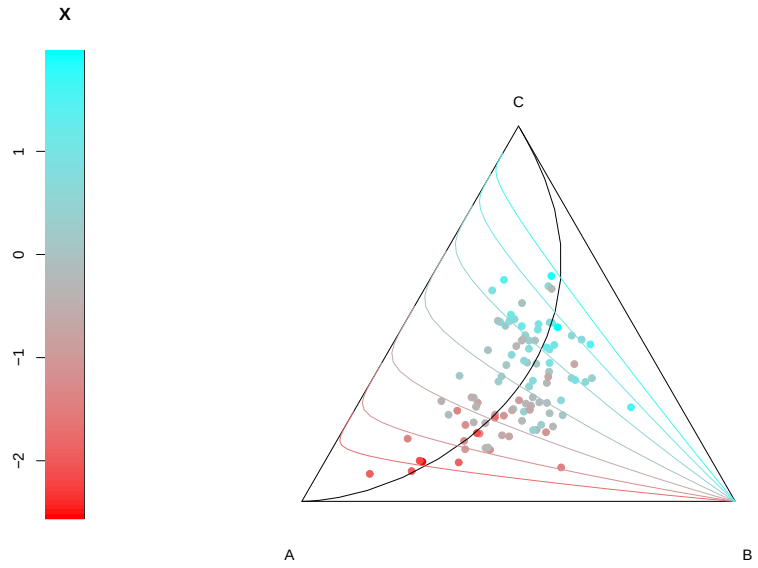

Figure 8: Composition in an explanatory role of a continuous interval variable: color is used to represent it, as a third dimension. Dots represent the data, and colored lines the hyperplane model fitted. Note that the black line gives the gradient of the plane (simulated data, same as in Figs. 4-7).

scale.

A simpler visualization is provided by the symbol principle, after splitting the range of the explained variable in some ordered categories, which effectively brings us to the case explained in section 2.8. Colored symbols are specially suited for this case, as a symbol encoding with increasing size or rays leads to visual underestimation of the importance of the lower parts of the scale. The key issue here is the definition of the classes: if the dependent variable is adequately scaled, equally-spaced intervals should be preferred, but equally-probable intervals (according to the empirical distribution of the explained variable) can be a good alternative if one is not sure of the scaling.

In any of the preceding cases, one should *never* forget to add the scale legend, relating color/symbols with values or intervals of the explained variable.

The parallel plot principle may also be useful: in this case, we would represent any of the one-dimensional projections of section 3.1 against the explained variable, though now the compositional variable would be represented in the abscissa axis. The advantages and problems of these several displays have been already explained in that section.

3.4 Compositions as explanatory variable

Standard linear regression can be easily used to **model** the relation of a continuous variable on the composition, by means of the ilr transformation,

$$Y = a + \sum_{j=1}^{D-1} b_j \mathrm{ilr}_j(\mathbf{X}) + \varepsilon = a + \langle \mathbf{b}, \mathbf{X} \rangle_A + \varepsilon. \tag{3}$$

Here the regression coefficient vector $\vec{b}_j = \mathrm{ilr}(\mathbf{b})$ gives the direction of the simplex in which the explained variable actually changes. The conditional distribution of the response variable given a value of the composition is

$$P(Y|\mathbf{X} = \mathbf{x}) = N(a + \langle \mathbf{b}, \mathbf{x} \rangle_A, \sigma_\varepsilon^2),$$

where σ_ε^2 is the residual variance. The global **quantification** of dependence and the **test** for normality of the residuals follow the standard procedures, as they depend on the residuals and the explained variable, which are treated in a conventional way. Partial **tests** of independence (e.g., that a subcomposition does not influence the explained variable) follow the same approach given in section 3.2: choose an ilr basis isolating the target subcomposition and test that its coefficients (now including the balancing element) are zero.

Once we have fitted a linear model, thus defined a regression plane, we may **display** it on our ternary diagrams or ilr- scatterplot matrices by plotting the gradient vector, the direction of compositional change which effectively modifies the explained variable. Alternatively, iso-level curves (actually, Aitchison lines orthogonal to the gradient) may also represent the fitted surface (Fig. 8). If the surface is *not* linear, the iso-level curve representation is the most common choice, though one can also represent it as a field, by plotting a descriptive set of local gradients.

Finally, comparison of fitted and observed values of the dependent variable can be provided by a scatter plot of these two variables (the common choice). In this case, a good fit would correspond to a narrow spread around the identity line. The disadvantage of this representation is that we do not display the model itself, i.e. the relation between the composition and the explained variable. As an alternative, one can use the overlay technique of section 3.1, by plotting together, e.g. the surface as iso-level curves and the data as circles, all filled with color following the same scale (Fig. 8).

3.5 Compositions and continuous interval

In the case of continuous interval covariables, the above mentioned models and methods (and every standard statistical method as well) can be applied without change.

3.6 Compositions and ratio

The ratio scale is univocally related to a continuous interval scale by the logarithmic transformation. Thus, all mentioned models and methods can be directly applied to the log-transformed variable.

In the case of modeling the dependence of a ratio variable on a composition, Equation (3) may be seen as giving the mean of $\log(Y)$, or as giving the location parameter of a lognormal model for Y (in the scope of generalized linear models).

3.7 Compositions and portions

The portion scale is univocally related to a continuous interval scale by the logistic transformation. All mentioned models and methods can therefore be directly applied to the logit-transformed variable. The logistic transformation was already used in Equation (1), to transform a probability (with portion scale) into log-odds (a variable with continuous interval scale). As with the preceding case, when modeling the dependence of a portion variable on a composition, Equation (3) may either give the mean of $\text{logit}(Y)$, or the location parameter of a logistic normal model for Y.

4 Compositions and other multidimensional scales

4.1 Examples

There are many applications in which one explores the relationship between two random vectors, each with its scale, for instance:

- two different compositional data sets, as a composition used to discriminate several groups via multinomial logistic regression, or a subcomposition used to predict another one;

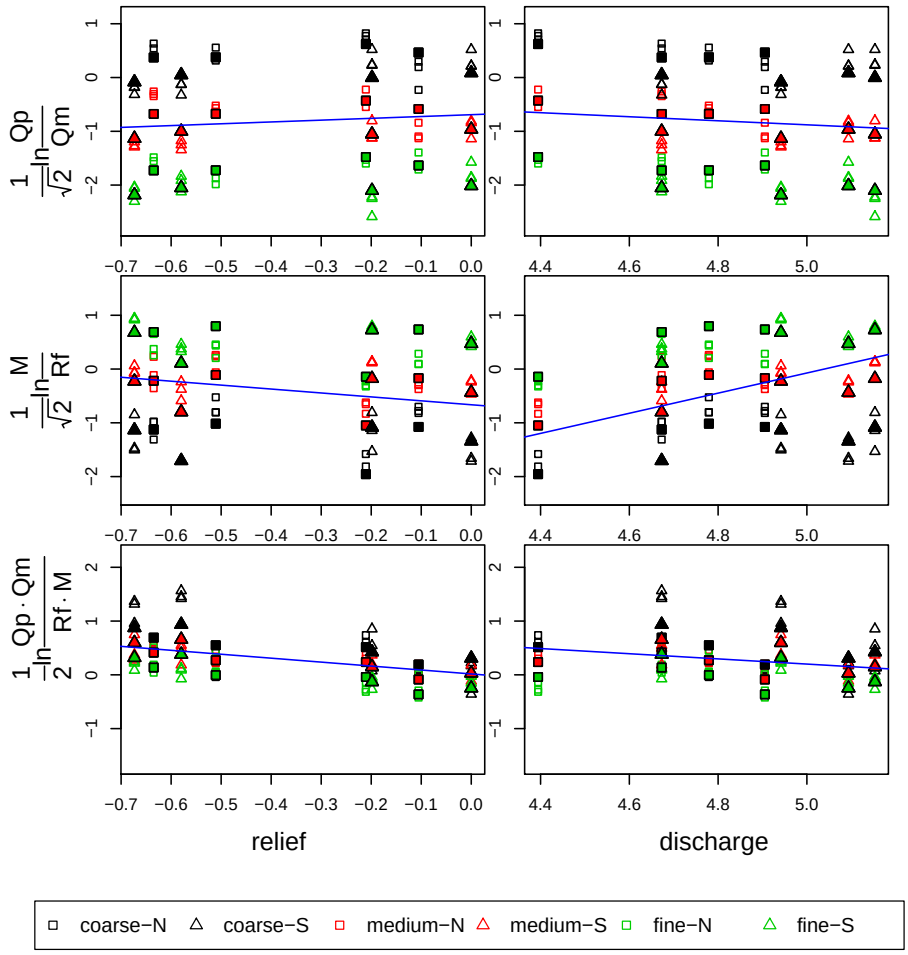

Figure 9: Representation of the linear model of dependence of the sand composition on log-relief, log-discharge, grain size in the ϕ scale (all 3 of class continuous interval) and position/geology (as dichotomous). Note that from here on, grain size is taken not as an ordered categorical variable, but as a continuous variable, with values ϕ =1, 2 or 3 (corresponding to coarse, medium and fine sand). Empty symbols represent the data set, and filled symbols the predictions provided by the model. The lines show the model line, passing through the average (X,Y) point of the plot, and with slope the coefficient in the linear model of the explanatory variable represented in axis X. Note that the Y axes have the same span, as the compositional coordinates are isometric and can thus be mutually compared.

- a compositional data set and a vector with multivariate ratio scale, for instance the major oxide composition and the trace element composition (although this is a simplification of the compositional nature of the whole set of variables);
- a compositional data set and a set of different, real (continuous interval) variables; e.g. with isotopic delta variables (Puig et al., 2008).

The last case is in fact the general one, as the idea is always to express the two data sets in some orthonormal bases (each with respect to its own Euclidean structure), and study the relationship among the coordinates of the two objects. In the following developments, we will consider that the composition is the response of the model, and that its explanation is provided by the real vectors (responding to a continuous interval scale, either being themselves also coordinates of objects with its Euclidean structure, or values embedded in the classical geometry). Inverting the roles (predicting trace elements from a composition, for instance), does not entail extra difficulty and will not be treated here.

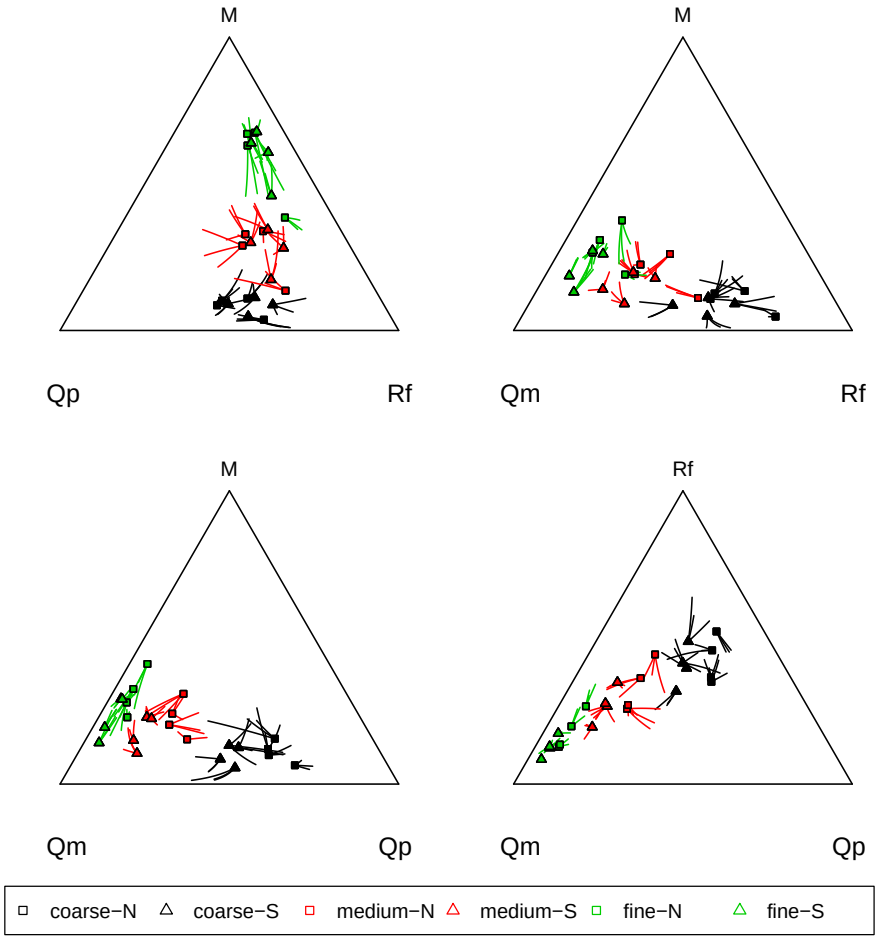

Figure 10: Three-part subcompositions of the sand data linear model, overlaying the predictions (symbols) and prediction errors (segments), thus the segment ends show the observations. The linear model takes as explanatory variables log-relief, log-discharge, grain size in the ϕ scale (all 3 of class continuous interval) and position/geology (as dichotomous).

4.2 The model and its typical display

To **model** the relationship of a compositional random vector $\mathbf{Y} \in S^D$ on a real vector $\vec{X} \in E$ (with E a real P-dimensional Euclidean space with the classical vector operations), the linear model is established by an affine linear application $\mathbf{Y} = \mathbf{a} \oplus \mathrm{B}(\vec{X}) \oplus \boldsymbol{\epsilon}$, with $\mathrm{B} : E \to S^D$ a linear form. After choosing orthonormal bases of S^D and E, we can compute the coordinates of all vectors involved: the transformations $\mathrm{ilr}(\mathbf{x})$ and $\mathrm{idt}(\vec{x})$ respectively give the coordinates of any generic vectors $\mathbf{x} \in S^D$ or $\vec{x} \in E$. But also the linear applications get coordinate expressions:

$$\mathrm{ilr}(\mathbf{Y}) = \mathrm{ilr}(\mathbf{a}) + \mathbf{B} \cdot \mathrm{idt}(\vec{X}) + \mathrm{ilr}(\boldsymbol{\epsilon}) = \mathrm{ilr}(\mathbf{a}) + \sum_{j=1}^{P} \mathrm{idt}_j(\vec{X}) \cdot \mathrm{ilr}(\mathbf{b}_j) + \mathrm{ilr}(\boldsymbol{\epsilon}), \tag{4}$$

with $\mathrm{ilr}(\mathbf{b}_j)$ the columns of $\mathbf{B}$ identifying (ilr-transformed) compositions. Multivariate regression of each $\mathrm{ilr}_i(\mathbf{Y})$ on the vector of all $\mathrm{idt}(\vec{X})$ is now easy to apply, and results can be assembled together back to vectors: e.g., joining all intercepts we get $\mathrm{ilr}(\mathbf{a})$.

The most immediate **display** of this model is by a matrix of scatter plots (Fig. 9), where the plot in the position (i, j) shows $\mathrm{ilr}_i(\mathbf{Y})$ as a function of $\mathrm{idt}_j(\vec{X})$. All plots in a row should share the same vertical scale, and those in a column the horizontal one (following the principle of parallel plots). The flaw of this representation becomes evident when we try to include the model in the form of a hyperplane: for a given value of $\mathrm{idt}_j(\vec{X})$ we may obtain a wide range of predictions of $\mathrm{ilr}_i(\mathbf{Y})$ (as it depends on other $\vec{X}$ coordinates), thus we cannot draw a line. If we want to add the predictions, then we may follow the overlay strategy (Fig. 10), with the problems already explained of "over-inking".

As an alternative, one can plot the observed values (of $\mathbf{Y}$, $\mathrm{ilr}_i(\mathbf{Y})$ or the set of all pair-wise log-ratios) against their predicted values: the identity line helps in visually assessing the fit of the model. This has the problem of not representing $\vec{X}$ itself, but through the linear function of $\mathbf{Y}$.

Bar plot representations of the model parameters, the compositions $\mathbf{b}_j$ from [Eq. (4)], may be a good complementary graphical **description**, as each displays the composition perturbing the prediction of $\mathbf{Y}$ when we change a unit a given coordinate of $\vec{X}$ (Fig. 12). This set of parameters can also be displayed as arrows in compositional scatter plots (ternary diagrams or ilr maps, Fig 11), in which case the relative lengths of the vectors on each represented subcomposition tells us the importance of each explanatory variable for that specific subspace. In this line, it is well-known that the role of the intercept $\mathbf{a}$ is to ensure that the image of the mean of $\vec{X}$ is the mean of $\mathbf{Y}$, and it represents the average value of the response composition when all explanatory variables are zero. We can display it as the point from which the arrows depart, though these arrows spreading from the mean of the plot may also be informative.

To **quantify** the global quality of the model, the R^2 measure given in [Eq. (2)] is equally valid for the case of multidimensional explanatory variables. The **test** on additive-logistic normality of the residuals given by Aitchison et al. (2004) is adequate as well. Partial tests of independence are nevertheless quite more complex. First, consider hypothesis of the form "*subcomposition S does not linearly depend on the coordinates K*". Using the generalized likelihood ratio tests or the union-intersection principle (Mardia et al., 1979), one can test simplifications of the form $\mathbf{A} \cdot \mathbf{B} \cdot \mathbf{C} = \mathbf{0}$, i.e. that a certain subspace of the origin space (the subspace spanned by the vectors of the basis of E described in the rows of $\mathbf{A}$ and linked to the coordinates in K) does not influence a given subspace of the image space (the subspace of S^D defined by the columns of $\mathbf{C}$ and linked to the subcomposition S, perhaps including the balance of S against the rest of parts of the composition). Aitchison (1986) presents in detail the case of $\mathbf{C} = \mathbf{I}$ the identity. Beyond the capabilities of this approach lie nevertheless those hypotheses of the form "*several subcompositions S_i are each independent of a set of explanatory variables K_i*", where the subcompositions $\mathrm{S}_1,\ldots,\mathrm{S}_p$ may have parts in common, and the sets $\mathrm{K}_1,\ldots,\mathrm{K}_p$ may have explanatory variables/coordinates in common. For some of these cases, Tolosana-Delgado and von Eynatten (2008) have developed an algorithm that searches a basis of S^D where such a complex hypothesis can be expressed as $D-1$ simultaneous

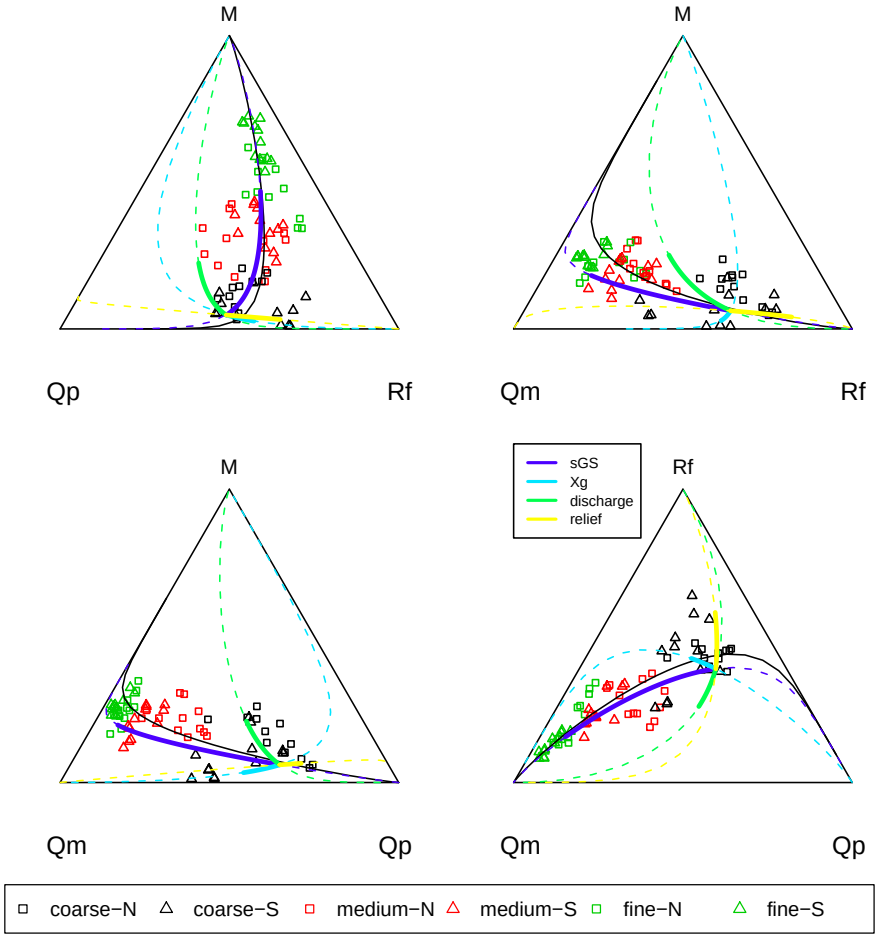

Figure 11: Sand data set, together with a vector representation of the fitted model: the coefficients of each explanatory variable are grouped together to define a compositional vector, and this is plotted as a thick segment; the associated (complete) compositional line is also displayed as a dashed curve. The principal component of the composition is also displayed (black line), showing that it is highly parallel to the grain size vector. Explanatory variables: sGS=grain size in ϕ scale (1=coarse, 2=medium, 3=fine); Xg=geology/position (0=north, 1=south).

conditions of the form "*the i-th ilr coordinate depends on the set of explanatory coordinates J_i*", to be independently fitted with univariate multiple regression, and checked together with generalized likelihood ratio tests on their ilr-back-transformed residuals. Unfortunately, not every hypothesis of this kind is compatible with an ilr basis. For these cases, the only alternative is the completely general Wald (1943) large-sample test, beyond the scope of this contribution.

4.3 Singular value decomposition of the model

It is not very well-known that the singular value decomposition (SVD) of B (or its coordinate expression $\mathbf{B}$) is a meaningful manipulation. For matrices, the SVD is typically expressed as

$$\mathbf{B} = \mathbf{U} \cdot \mathbf{D} \cdot \mathbf{V}^t,$$

but for linear applications (Eaton, 1983) it can be also written with the help of outer products (denoted by $\times$)[1] as

$$B = \bigoplus_{i=1}^{R} d_i \odot (\mathbf{u}_i \times \vec{v}_i).$$

The left and right singular vectors ($\mathbf{u}_i$ and $\vec{v}_i$, obtained by respectively applying $\mathrm{ilr}^{-1}(\cdot)$ and $\mathrm{idt}^{-1}(\cdot)$ to the columns of $\mathbf{U}$ and $\mathbf{V}$) are orthonormal sets, and each forms a basis of its space (or properly speaking of the image space and the domain space of $B(\cdot)$, respectively R-dimensional sub-spaces of S^D and E). This SVD of the linear application suggests many ways to display the obtained model, beyond the intuitive scatter plot of explanatory coordinates against response coordinates.

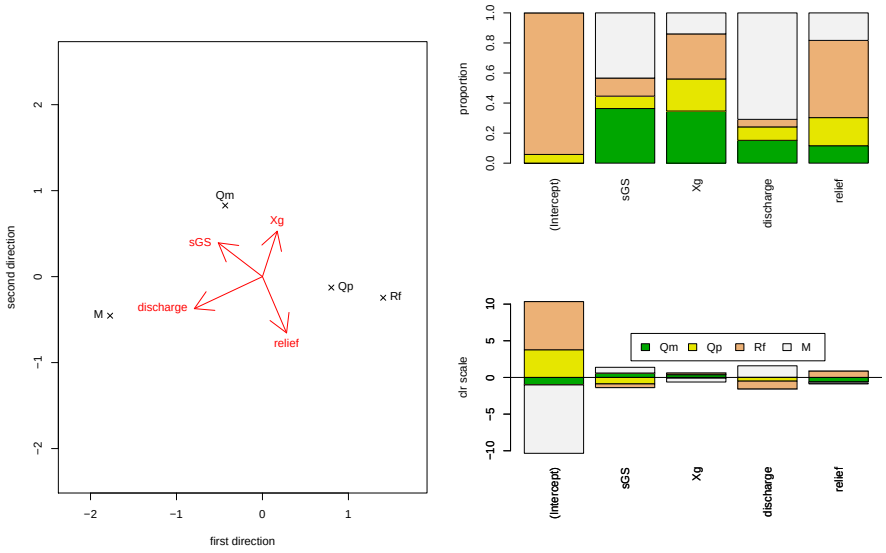

Figure 12: Three representations of the coefficients of the linear model. (Left) Biplot of the singular value decomposition of the linear application. (Right) Bar plots of each column of the linear application, as a composition (top) or as clr coefficients (bottom). The variables Xg and sGS are respectively the geology/position dichotomous factor (thus, the perturbation to pass from north to south) and grain size in ϕ scale (where the same increment of one unit leads from coarse to medium and from medium to fine grained sand). The intercept represents an ideal composition with all explanatory variables to zero (i.e., a very-coarse-grained sand from a northern watershed with 1mm yearly precipitation and relief ratios of 1:1 or 45^o).

For instance, we can represent $\mathbf{B}$ in a biplot (Fig. 12), exactly as it is done with a data set (Gower and Hand, 1996). Given the characteristics of compositions, in this case we would recommend to plot the variables from E as rays and the compositional variables as points, and recall the interpreter to look at the links between compositional variables. Note that as in any biplot, a decision must be made on how to "split" the singular values. In this case, the answer should come from the goal of the display. If we want to see how much changes the image composition when changing one unit of the explanatory variables (a typical choice), then we should represent $(\mathbf{U} \cdot \mathbf{D})$ and $\mathbf{V}$. But for instance in multinomial logistic regression, it might be interesting to show how strong must be the change in the explanatory variables (the composition) to modify one unit the log-odds (=log-ratios of probabilities) in favour of a given group, which would be attained by taking $\mathbf{U}$ and $(\mathbf{V} \cdot \mathbf{D})$.

By extension of the concepts explained up to here, each of these vector sets can be also represented in its own space: the compositions $(d_i \odot \mathbf{u}_i)$ may be displayed in bar plots, or as (curved) arrows in ternary diagrams, while the vectors from the explanatory space $\vec{v}_i$ may also be represented in scatter

[1]Recall that an outer product is a bi-linear operation taking a couple of vectors from two spaces, $\mathbf{u}_i \in S^D$ and $\vec{v}_i \in E$, and building as image a "vector" $\mathrm{T} = \mathbf{u}_i \times \vec{v}_i$ from $L(E, S^D)$, the space of linear applications from E to S^D. Recall also that $L(E, S^D)$ is a vector space with the operations inherited from S^D.

plots as arrows. In such representation it is important to link somehow the pairs $\{\vec{v}_i;\ (d_i \cdot \mathbf{u}_i)\}$, for instance using the same colour for both arrows. Finally, as each of these vectors represent bases, we can compute the coordinates of the observations on these bases and plot the results in a $R \times R$ matrix of scatterplots (Fig. 13): here, the model *can* be added as a line in the diagonal plots, where it is equal to the singular values d_i.

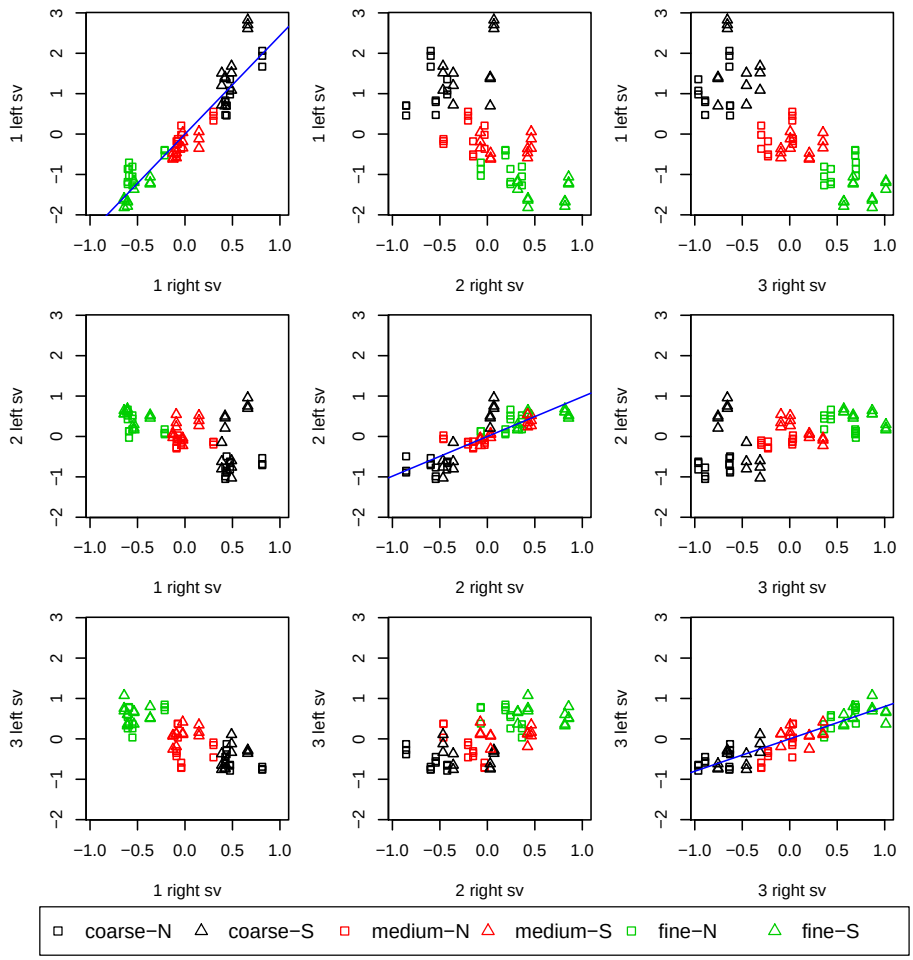

Figure 13: Sand data linear model, represented in the bases formed by its own SVD left and right singular vectors (abbr. "sv"). The slopes of the regression lines on the diagonal are equal to the singular values. Note that all horizontal and vertical axes are the same, as these coordinates are mutually comparable.

5 Compositions and count compositions

5.1 The basics proplem of count compositions

In our understanding a count composition is typically a binomial or Poisson observation with relative portions comming from a random real composition. Thus, two different process interfere:

- the randomness of the underlying real composition e.g. represended by a Mahalanobis distance in Aitchison geometry,

- the randomness of the random counts, expressed by the chord distance (e.g. Simpson, 2007),

$$d_c((x_i),(y_i)) = \sqrt{\sum_i \left(\sqrt{\frac{x_i}{\sum_j x_j}} - \sqrt{\frac{y_i}{\sum_j y_j}}\right)^2} \tag{5}$$

or by a χ^2 distance

$$d_{\chi^2}((x_i),(y_i)) = \sqrt{\sum_i \frac{(x_i - e_{xi})^2}{e_{xi}} + \sum_i \frac{(y_i - e_{yi})^2}{e_{yi}}} \tag{6}$$

with $e_{xi} = (x_i + y_i)\frac{\sum_j x_j}{\sum_j x_j + y_j}$.

However these different geometries are mutually incompatible. For instance, it is quite probable to observe a zero count for a component with a small portion. However from the point of view of real compositions that 0 would put the composition at the infinity circle. Moreover, the scale of the observations depends on their locations, seen in the Euclidean sense as well as in the Aitchison geometry. And it is well known that the χ^2-distance itself does not work correctly with count values near zero, since it is valid only asymptotically. On the other hand, it is dependent on the total number of observation: should a composition not be independent of its total "size"? Yes, but the χ-square distance measures a probabilistic difference, which actually depends mainly on the sample size. In other words, a correct interpretation of a count composition asks us to know the its reliability, which depends on its total number sum. Thus there is no unique geometry or an isometric transformation catching both the compositional and the statistical aspect of the data.

5.2 One dimensional informations about count compositions

For a joint **display** of a count composition, we need to reduce it to a one-dimensional information (section 3.1), e.g. by a balances, summarizing groups of parts or taking a subcomposition. However since the underlying composition is badly defined in Aitchison geometry, when inferred based on count data, we propose to use simple subcompositional marginals like:

$$s_{A,B}(x) := \frac{\sum_{i\in A} x_i}{\sum_{i\in B} x_i} \tag{7}$$

where A and B are disjoint subsets of the components of the composition x. Examples are $s_{\{i\},\{j\}}$ representing the ratio of x_i and x_j or $s_{\{1\},\{1,2,3,4,\ldots,D\}}$ giving the portion of x_1. These bivariate subcompositions can be easily interpreted. Some might argue that this type of marginalization is not invariant with respect to a change of units. However count compositions have the same unit in every observation, as the counts are natural numbers and these natural numbers yield again comparable units when normalized: probabilities. Thus the visualization may be meaningful even if it is not invariant under perturbation.

Furthermore, depending on the total number of counts, these ratios and portions [Eq. (7)] will have different variability in log scale. And if zeroes occur, this variance will even be infinite. Thus the analysis should not be invariant under the total number of individuals in the composition, since then we would lose that information on the precision of the composition.

Our proposal is to display the precision of the projection information [Eq. (7)] by confidence intervals. However confidence intervals for geometric means and thus for balances based on Poisson distributed observations are quite difficult to compute, and large in Aitchison geometry if a zero count is observed. We thus propose to use (arithmetic) balances like given in Equation (7). For these ratios the confidence limits for the probability:

$$p_{A|A\cap B} = P(x \in A | x \in A \cup B)$$

can be computed based on its distribution, the well-known binomial model.

As illustration, figure 14 shows a simulated dataset, of a real composition of 3 components A, B, C and a count composition of corresponding components a, b, c. The count composition is simulated as a conditional Poisson distribution which expected total and relative proportions are given by the real composition. Thus we expect a clear dependency between both compositions. For the real composition, we use the two (geometric) balances AB against C and A against B. Each of these two balances are backtransformed to a two part composition, by the corresponding two-part ilr-transform. In this way we get for each of the two dimensions of the "Aitchison-composition" problem, a one-dimensional projected composition. In matrix notation, that is:

$$g_{A,B}(x) = \mathcal{C}\left(\exp\left(\ln \mathbf{x} \cdot \vec{v}_{A,B}\right) \cdot \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}\right),$$

where $\vec{v}_{E,F}$ is the vector balancing the geometric mean of those parts in group $E = \{A, B\}$ against the geometric mean of those in group $F = \{C\}$, e.g. in this case the vector $\vec{v}_{E,F} = [1, 1, -2]/\sqrt{6}$ (Egozcue and Pawlowsky-Glahn, 2005). Similarly the count composition is represented by the corresponding portions [Eq. (7)], each represented by a short horizontal line displaying the estimated probability given by the observed portion, and a corresponding (conservative) 95%-confidence limit, represented by a gray vertical lines. Each of the two balances is plotted against each of the portions. The two plots on the main diagonal show clearly the proportionality of the corresponding balances and ratios. The off-diagonal plots show no clear dependence. Finally, even the strong outlier that apparently lays completely out of the main trend may be interpreted, as its confidence limits displays that it is the result of unprecise estimations due to small total counts in the count composition.

5.3 Modeling of dependency to count compositions

To **model** the dependence of a count composition $\mathbf{Y}$ on a real composition $\mathbf{X}$ a generalized linear model with a conditional multinomial distribution like:

$$\mathbf{Y}_i \sim Mu(n_i, \mathrm{ilr}^{-1}[\vec{a} + \mathbf{A}\mathrm{ilr}(\mathbf{X}_i)])$$

where n_i is the total number of composition Y at observation i, and $\mathrm{ilr}^{-1}(\vec{a} + \mathbf{A}\mathrm{ilr}(X_i))$ is an arbitrary affine linear function of the real composition with some matrix $\mathbf{A}$. Eventually one may consider overdispersion parameters. A **quantification** of the dependency might be possible through information criteria such as AIC. However we consider this an open problem.

6 Conclusions

This contribution presents a structured overview of the possibilities for the joint analysis of compositions with data from other scales, gathering existing methods and developing some new ones. The proposed methods are mainly straightforward and consequent applications of known principles of statistical graphical and multivariate modeling, together with the principle of working on coordinates. It is still not so easy to apply these techniques for the non-expert: none of the methods handle all the possible questions at the same time. Different models and methods have to be selected to display, highlight, model or quantify the different aspects of the data and its structure.

Note that the multivariate character of the compositions can not perfectly be represented in any graphics: each of them lose some information, do not properly display the distance, are counterintuitive to read or quite complex to build. The interpreter of the results thus needs to know to which sort of graphic he is going to look at and to understand the mathematical background of the projection methods used. This either means that he has to select the graphic himself or that a detailed description has to be passed with each result. It is nevertheless quite good recalling that this utterly applies to all plots in use: a lay can seldom read a ternary diagram, and the only

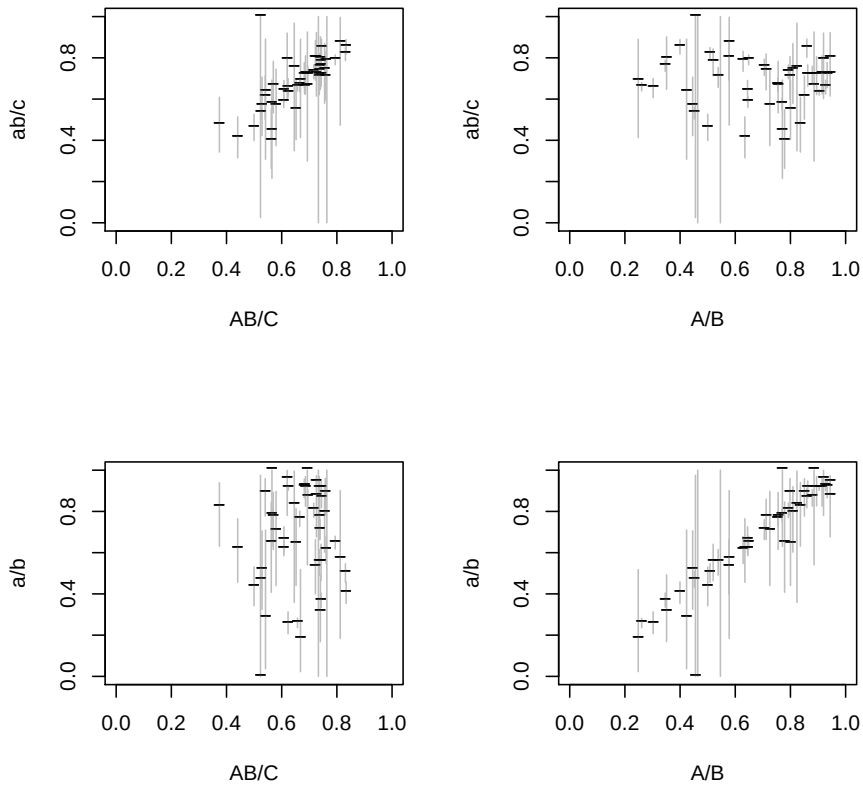

Figure 14: Displaying a dependent count composition. The lowercase letters a,b,c correspond to the components in the count composition. Note that the symbolic ratios represented in this vertical axis are arithmetic, e.g. ab/c means (a+b)/c. The uppercase letters A,B,C correspond to related real composition, giving the expectation of the count composition. In this horizontal axes, the symbolic ratios represent the two subcompositions used to build that particular two-part projected composition. In this case, AB/C means the portion of $\exp(\ln(A \cdot B)/\sqrt{12})$ vs. $\exp(\ln(C)/\sqrt{3})$.

reason most people can read a simple map is because we are taught to. Interpretability is not an intrinsic property of the representation, but the result of the analyst training.

Different projections emphasize different aspects of the dataset. Thus the projections should be selected according to the question to be answered or the aspect to be highlighted. It is thus pointless to define single canonical graphics a priori.

Finally, this contribution is by no means a full stop on the subject. Each of the proposed methods might deserve a more comprehensive analysis, and there is plenty of room for more methods to be developed according to more specialized problems of dependence analysis involving compositions and other scales.

Acknowledgments

The authors acknowledge funding from the Department of Universities, Research and Information Society (DURSI, grant 2005 BP-A 10116) of the *Generalitat de Catalunya*.

REFERENCES

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Aitchison, J. (1997). The one-hour course in compositional data analysis or compositional data analysis is simple. In V. Pawlowsky-Glahn (Ed.), *Proceedings of IAMG'97 — The third annual conference of the International Association for Mathematical Geology*, Volume I, II and addendum, pp. 3–35. International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), 1100 p.

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

Aitchison, J., G. Mateu-Figueras, and K. W. Ng (2004). Characterisation of distributional forms for compositional data and associated distributional tests. *Mathematical Geology 35*(6), 667–680.

Barceló-Vidal, C. (1996). *Mixturas de Datos Composicionales*. Ph. D. thesis, Universitat Politècnica de Catalunya, Barcelona (E). 261 p.

Daunis-i Estadella, J., J. J. Egozcue, and V. Pawlowsky-Glahn (2002). Least squares regression in the simplex. In U. Bayer, H. Burger, and W. Skala (Eds.), *Proceedings of IAMG'02 — The eigth annual conference of the International Association for Mathematical Geology*, Volume I and II, pp. 411–416. Selbstverlag der Alfred-Wegener-Stiftung, Berlin, 1106 p.

Eaton, M. L. (1983). *Multivariate Statistics. A Vector Space Approach*. John Wiley & Sons.

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

Gower, J. C. and D. J. Hand (1996). *Biplots*. London (UK): Chapman and Hall Ltd. 277 p.

Grantham, J. H. and M. A. Velbel (1988). The influence of climate and topography on rock-fragment abundance in modern fluvial sands of the southern blue ridge mountains, north carolina. *Journal Sedimentary Research 58*, 219–227.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate Analysis*. Academic Press, London (GB). 518 p.

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

Puig, R., N. Otero, R. Tolosana-Delgado, C. Torrentó, A. Menció, A. Folch, A. Soler, J. Back, and J. Mas-Pla (2008). Multi-isotopic and compositional exploration of factors controlling nitrate pollution. In J. Martín-Fernández and J. Daunis-i Estadella (Eds.), *Compositional Data Analysis Workshop – CoDaWork'08, Proceedings*. Universitat de Girona, http://ima.udg.es/Activitats/CoDaWork08/.

Simpson, G. L. (2007). Analogoue methods in palaeoecology: Using the analogue package. *Journal of Statistical Software 22*(2).

Stevens, S. (1946). On the theory of scales of measurement. *Science 103*, 677–680.

Thomas, C. W. and J. Aitchison (1998). The use of logratios in subcompositional analysis and geochemical discrimination of metamorphosed limestones from the northeast and central Scottish Highlands. In A. Buccianti, G. Nardi, and R. Potenza (Eds.), *Proceedings of IAMG'98 — The fourth annual conference of the International Association for Mathematical Geology*, Volume I and II, pp. 549–554. De Frede Editore, Napoli (I), 969 p.

Thomas, C. W. and J. Aitchison (2005). Compositional Data Analysis of Geological Variability and Process: a Case Study. *Mathematical Geology 37*(7), 753–772.

Tolosana-Delgado, R., N. Gorelikova, and V. Pawlowsky-Glahn (2004). Statistical classification of Tin deposits from Cassiterite compositions: application to Russian Far East Tin region. Volume 6, pp. 233–236. SGE.

Tolosana-Delgado, R. and H. von Eynatten (2008). Simplifying compositional multiple regression: application to grain size controls on sediment geochemistry. *Computers and Geosciences, submitted*.

Tolosana-Delgado, R. and H. von Eynatten (accepted). Grain-size control on petrographic composition of sediments: compositional regression and rounded zeroes. *Mathematical Geology*.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society 54*(3), 426–482.