WP.41
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9–11 May 2011)

Topic (vii): New and emerging methods

A FORWARD SEARCH ALGORITHM FOR COMPOSITIONAL DATA

Key Invited Paper

Prepared by Filippo Palombi (ISTAT), Simona Toti (ISTAT),
Romina Filippini (ISTAT), Valeria Tomeo (ISTAT)

I. A SHORT REVIEW OF COMPOSITIONAL ANALYSIS

A. Aitchison Geometry on the Simplex

1. Compositional data analysis has been the subject of a number of papers, pioneered by J. Aitchison [1] over the past twenty years. As a methodology of statistical investigation, compositional analysis finds application in all cases where the main object of interest is an n-dimensional variate x, whose components are strictly positive continuous real variables, to be regarded as portions of a total amount κ, i.e. they belong to the n-dimensional simplex

$$\mathcal{S}^{(n)} = \left\{ x = (x_1, \ldots, x_n) : \quad 0 < x_k < \kappa\,, \quad \sum_{k=1}^{n} x_k = \kappa \right\}. \tag{1}$$

Elements of $\mathcal{S}^{(n)}$ are named compositions. Examples are easily found in the context of geochemical data analysis, balance sheets analysis, etc. Under such circumstances, it may be relevant to consider the x_k's only in terms of their relative importance, i.e. with no reference to their absolute size. If this perspective is adopted, comparing different samples of the variate becomes a matter of comparing one by one all the possible pairwise ratios of their components, since these are the only quantities that bring the relative information. Along this line, Aitchison has introduced an appropriate distance function,

$$d_{\mathrm{A}}(x, y) = \sqrt{\frac{1}{2n} \sum_{i,j=1}^{2n} \left[\ln\left(\frac{x_i}{x_j}\right) - \ln\left(\frac{y_i}{y_j}\right) \right]^2}, \tag{2}$$

where all these ratios enter equally weighted, smoothed and symmetrized by logarithms. The Aitchison metric looks different from the more popular Mahalanobis one, as it corresponds to evaluating the distance between observations according to a different criterion. The change of metric makes most classical statistical techniques useless and imposes the introduction of more sophisticated mathematical tools, in order to restore the conditions for a statistical investigation.

2. Usually, observations in a compositional data set cannot be directly compared, as their parts sum to different constants κ. This kind of situation occurs frequently in practice: think for instance of a chemical data set, where the observations consist of the amounts of certain chemical elements (expressed in a given unit) in each of the analyzed samples; obviously the observed values depend on the total weight of the sample. Nevertheless, all parts of each sample can be rescaled by a given factor, depending on the sample, chosen so that different observations fulfill the same compositional constraint. The rescaling is mathematically encoded via a closure transformation

$$\mathcal{C}(x) = \left(\frac{\kappa x_1}{\sum_{k=1}^{n} x_k}, \ldots, \frac{\kappa x_n}{\sum_{k=1}^{n} x_k} \right) , \tag{3}$$

which rescales x and leaves all the pairwise ratios of its components unchanged. It should be remarked that the Aitchison distance is closure–invariant, since it does not depend on ratios of components belonging to x and y respectively. Nevertheless, the closure transformation has a theoretical rôle in the construction of a theory of compositional analysis, since it turns an experimental multivariate data set into a set of elements of $\mathcal{S}^{(n)}$.

3. The Aitchison distance can be naturally obtained by providing the n–dimensional simplex with a linear vector space structure, as originally proposed by Aitchison (see ref. [1] for a detailed presentation of the idea). The sum of two elements of the simplex is here defined by

$$x \oplus y = \mathcal{C}(x_1 y_1, \ldots, x_n y_n) , \qquad x, y \in \mathcal{S}^{(n)} , \tag{4}$$

and the product of an element of the simplex by a real variable is defined by

$$\alpha \otimes x = \mathcal{C}(x_1^{\alpha}, \ldots, x_n^{\alpha}) , \qquad \alpha \in \mathbb{R} , x \in \mathcal{S}^{(n)} . \tag{5}$$

Note that the closure $\mathcal{C}$ is needed in both eqs. (4–5) in order to project the results of the operations onto $\mathcal{S}^{(n)}$. To complete the construction of a linear vector space, an inner product has to be assigned. This is given by

$$\langle x, y \rangle_{\mathrm{A}} = \frac{1}{2n} \sum_{i,j=1}^{n} \ln \left(\frac{x_i}{x_j} \right) \ln \left(\frac{y_i}{y_j} \right) . \tag{6}$$

Showing that the operations introduced in eqs. (4–6) fulfill the standard algebraic properties of vector sum, scalar multiplication and inner product is straightforward, though it goes beyond the aims of the present report. We conclude this section by reminding the reader that once a vector space is provided with an inner product, a metric structure can be introduced as well. The norm induced by the inner product is given by $||x||_{\mathrm{A}} = \sqrt{\langle x, x \rangle_{\mathrm{A}}}$, and the distance induced by the norm is accordingly defined by

$$d_{\mathrm{A}}(x, y) = ||x \ominus y||_{\mathrm{A}} , \tag{7}$$

where $x \ominus y \equiv x \oplus [(-1) \otimes y]$. Eqs. (2) and (7) are equivalent.

B. Isometric Logratio Transformation

4. The vector space structure introduced in the previous section is built on top of objects (i.e. experimental observations) represented by an n–tuple, whose components fulfill a compositional constraint. This representation is definitely legitimate, though mathematically unappealing. The presence of a scalar constraint entails that the faithful dimension of our vector space is $n-1$ instead of n. In order to get rid of the redundancy induced by the constraint and to accordingly provide a minimal description of the n–dimensional simplex as a $(n-1)$–dimensional linear vector space, the simplest solution is to perform an appropriate projection of the original n–tuple onto a $(n-1)$–tuple with unconstrained components. Several ways of doing have been explored extensively in the literature (see ref. [2] for a review). Here we are interested in one particular mapping, known as the *Isometric Logratio Transformation* (ILR), firstly introduced in [3]. The feature that makes the ILR a remarkable transformation is that it is an isometry, i.e. it establishes a relation between different vector spaces that preserves the distance. The ILR is defined as follows. As a first step a set of generators of the simplex is introduced,

$$w_k = \mathcal{C}(\underbrace{1,\ldots,1}_{k-1 \text{ times}}, \mathrm{e}, \underbrace{1,\ldots,1}_{n-k \text{ times}}) , \qquad k = 1,\ldots,n , \tag{8}$$

where the Euler constant e is placed in the k^{th} position. As usual, the closure is needed to guarantee that $w_k \in \mathcal{S}^{(n)}$. Obviously, each element of the simplex can be represented as a linear combination of the w_k's with an appropriate choice of coefficients. Nevertheless, the n generators do not constitute a basis of the simplex, since their number exceeds the dimension of $\mathcal{S}^{(n)}$ by one. A basis can be obtained from the generators in a straightforward way: one of the generators is dropped, conventionally w_n, and then a Gram-Schmidt orthonormalization procedure (employing the inner product of eq. (6)) is applied to the remaining $n-1$ generators. The $n-1$ vectors $\{e_k\}_{k=1,\ldots,n-1}$ thus obtained constitute a basis of $\mathcal{S}^{(n)}$. Obviously, this basis is not unique and is related to any other basis of via a linear transformation of maximal rank. In what follows, we consider one particular choice of the vector basis, i.e.

$$e_k = \mathcal{C}\left(\underbrace{\exp\sqrt{\frac{1}{k(k+1)}}, \ldots, \exp\sqrt{\frac{1}{k(k+1)}}}_{k \text{ times}}, \exp\left\{-\sqrt{\frac{1}{k(k+1)}}\right\}, \underbrace{1,\ldots,1}_{n-k-1 \text{ times}} \right) , \tag{9}$$

with $k = 1,\ldots,n-1$. The *Isometric Logratio Transformation* is the mapping

$$x \in \mathcal{S}^{(n)} \;\rightarrow\; \mathrm{ilr}(x) = (\langle x, e_1\rangle_{\mathrm{A}}, \ldots, \langle x, e_{n-1}\rangle_{\mathrm{A}}) . \tag{10}$$

As previously mentioned, the ILR is an isometry. It maps the Aitchison distance in $\mathcal{S}^{(n)}$ onto the Euclidean distance in the transformed space, i.e.

$$d_{\mathrm{A}}(x,y) = d_{\mathrm{E}}\left(\mathrm{ilr}(x), \mathrm{ilr}(y)\right) , \qquad x, y \in \mathcal{S}^{(n)} , \tag{11}$$

where

$$d_{\mathrm{E}}(u,v) = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2} , \qquad u, v \in \mathbb{R}^{n-1} . \tag{12}$$

Eqs. (11-12) will be used in sect. III.

C. Distributional transformations

5. In what follow, we shall describe an algorithm for the search of compositional outliers. The detection of one or more outliers in a given data set will be based on a statistical test, whose outcome rests in its turn on a distributional assumption on the data. Though in most applications the simplest statistical hypothesis would involve the normal distribution (or the multi-normal one, in case of multivariate observations), it can be easily realized that this is not the case: each component of a given compositional vector is a strictly positive number; as such, the tails of its marginal distribution can never extend to the negative semi-axis, as would be the case for a normal variable. In fact, Aitchison has shown in [4] that the most natural distributional assumption for compositional data is the log–normal one. From now on, we shall adopt this view. Accordingly, we shall assume that the observations of our data set $\mathcal{D} = \{x^{(k)} \in \mathbb{R}_+\}_{k=1,\ldots,N}$ tend to distribute according to an n–dimensional log–normal distribution $\Lambda_n(\xi, \Omega)$ with given log–scale vector ξ and scattering matrix Ω, i.e.

$$H_0 : \quad \{x^{(1)} \sim \Lambda_n(\xi, \Omega)\} \cap \{x^{(2)} \sim \Lambda_n(\xi, \Omega)\} \cap \ldots \cap \{x^{(N)} \sim \Lambda_n(\xi, \Omega)\} \,. \tag{13}$$

The question arises naturally as to how the ILR acts on $\Lambda_n(\xi, \Omega)$. In other words, if $x \sim \Lambda_n(\xi, \Omega)$, how does $\mathrm{ilr}\,(\mathcal{C}(x))$ distribute? This question has been partially addressed by Aitchison, who has shown in [4] the following statement:

Theorem. Iff $x \sim \Lambda_n(\xi, \Omega)$, then the closure $\mathcal{C}(x) \sim L_{n-1}(A\xi, A\Omega A^{\mathrm{T}})$, where the rectangular matrix A has dimension $(n-1) \times n$ and is given by $A = [\mathbb{I}_{n-1}, -e_{n-1}]$, $\mathbb{I}_n$ being the unit matrix of order n and e_n being an n–dimensional vector with unit components. □

Here we denote $L_{n-1}(\nu, \Gamma)$ the Logistic distribution with p.d.f. given by

$$\begin{aligned} f(u_1, \ldots, u_{n-1}) &= \frac{1}{(2\pi)^{n/2}|\Gamma|^{1/2}} \left[\prod_{j=1}^{n} u_j\right] \exp\left\{-\frac{1}{2}[\ln(u/u_n) - \nu]^{\mathrm{T}} \Gamma^{-1} [\ln(u/u_n) - \nu]\right\} , \\ u_n &= 1 - \sum_{j=1}^{n} u_j \,. \end{aligned} \tag{14}$$

By using this theorem, it is easy to show that $x \sim \Lambda_n(\xi, \Omega)$ entails $\mathrm{ilr}\,(\mathcal{C}(x)) \sim \mathcal{N}_{n-1}(\mu, \Sigma)$, i.e. $\mathrm{ilr}\,(\mathcal{C}(x))$ distributes normally with mean μ and covariance matrix Σ, easily related to ξ and Ω via known linear transformations. The converse is also true, i.e. $\mathrm{ilr}\,(\mathcal{C}(x)) \sim \mathcal{N}_{n-1}(\mu, \Sigma)$ entails $x \sim \Lambda_n(\xi, \Omega)$. In summary, the assumption of log–normality of the original data propagates to the closed data and the ILR–transformed ones according to the distributional chain

$$x \sim \Lambda_n \qquad \Leftrightarrow \qquad \mathcal{C}(x) \sim L_{n-1} \qquad \Leftrightarrow \qquad \mathrm{ilr}\,(\mathcal{C}(x)) \sim \mathcal{N}_{n-1} \,. \tag{15}$$

II. THE FORWARD SEARCH ALGORITHM (FSA)

A. Construction of the signal

6. In this report we consider the problem of detecting potential outliers within a compositional data set $\mathcal{D} = \{x^{(k)} \in \mathbb{R}_+\}_{k=1,\ldots,N}$. Among the presently known methodologies that allow to detect and eliminate outlying data, we focus on the one based on the *Forward Search Algorithm* (FSA), originally introduced in [5] and thoroughly discussed in [6] for the case of multivariate analysis. The original algorithm applies to the case of normally distributed data, i.e. the null hypothesis assumes that all the observations distribute simultaneously according to $x \sim \mathcal{N}_n(\mu_0, \Sigma_0)$, i.e.

$$H_0 : \quad \{x^{(1)} \sim \mathcal{N}_n(\mu_0, \Sigma_0)\} \cap \{x^{(2)} \sim \mathcal{N}_n(\mu_0, \Sigma_0)\} \cap \ldots \cap \{x^{(N)} \sim \mathcal{N}_n(\mu_0, \Sigma_0)\} \,. \tag{16}$$

The algorithm makes use of the Mahalanobis distance

$$d_{\mathrm{M}}(x, y|\Sigma) = \sqrt{(x-y)^{\mathrm{T}} \Sigma^{-1} (x-y)} \,, \qquad x, y \in \mathbb{R}^n \,, \tag{17}$$

and consists of a recursion of sorting steps. Initially, a subset $\mathcal{S}(m_0) \subset \mathcal{D}$ of m_0 observations is chosen (randomly or close to the bulk of the distribution) and is used in order to provide an estimate $\mu(m_0)$ of the true mean μ_0 and an estimate $\Sigma(m_0)$ of the true covariance matrix Σ_0. Then, the recursion starts and its m^{th} step ($m \geq m_0$) is described as follows:

- $\mathcal{D}$ is sorted according to the increasing values of the Mahalanobis distance $d_{\mathrm{M}}(x)$ of the single observation x from $\mu(m)$, given the sample estimate $\Sigma(m)$ of the covariance matrix,

$$d_{\mathrm{M}}(x) \equiv d_{\mathrm{M}}\left(x, \mu(m)|\Sigma(m)\right) \,; \tag{18}$$

- the square of the Mahalanobis distance of the $(m+1)^{\mathrm{th}}$ observation in $\mathcal{D}$ is stored together with the corresponding identification number and defines the m^{th} value of what is named the signal of the FSA;
- $\mathcal{S}(m)$ is replaced by a new subset $\mathcal{S}(m+1)$, made of the first $m+1$ observations of $\mathcal{D}$;
- a new estimate of μ_0 and Σ_0 is obtained from the observations in $\mathcal{S}(m+1)$. Since the observations in $\mathcal{D}$ are ordered, the estimate $\Sigma(m+1)$ of Σ_0 is biased and must be corrected in order to remove the effects of truncating $\mathcal{D}$ at the $(m+1)^{\mathrm{th}}$ observation. The correction is described in [7].

The sorting procedure repeats iteratively until the estimate of the distributional parameters involves all the inliers and the occurrence of the first outlier shows up as a break–point (or a jump) in the values taken by the signal as a function of m.

B. Testing the signal

7. In order to give a quantitative meaning to the break–point and to associate it with the occurrence of an outlier, a statistical test, resting on the normality hypothesis, is performed.

At each step of the recursion, the computation of the signal is accompanied by a measurement of its theoretically admissible lowest and highest values at – say – 99% confidence level under H_0. The envelope of the pointwise values of these lower and upper thresholds for $m_0 \leq m \leq N-1$ generates two curves that surround the signal until an outlier enters the estimate of the distributional parameters. The violation of the envelopes signals the income of an outlier within $\mathcal{S}(m)$. The computation of the thresholds is made possible by the fact that the squares of the Mahalanobis distance from the mean ideally distribute according to χ^2_n under H_0. When the dataset is sorted as prescribed by the FSA, the Mahalanobis distances from the mean constitute a set of n order statistics. The distribution of the signal, i.e. the $(m+1)^{\text{th}}$ order statistic, cannot be computed in closed form. Nevertheless, its percentiles (coinciding with the above mentioned thresholds at $\alpha = 0.01$ and $1-\alpha = 0.99$) can be obtained from a general result firstly derived in [8],

$$y_{m+1,N;\alpha} = (\chi^2_n)^{-1}\left(\frac{m+1}{m+1+(N-m)f_{2(N-m),2(m+1);1-\alpha}}\right), \qquad m_0 \leq m \leq N-1\,, \tag{19}$$

where the symbol $f_{a,b;\alpha}$ denotes the α–percentile of the Fisher distribution with parameters (a,b).

III. EXTENSION OF THE FSA TO THE COMPOSITIONAL CASE

8. In order to be applied to a compositional data set, the FSA has to be modified in two respects:

- the construction of the signal has to be performed with the Mahalanobis distance being replaced by the Aitchison one. There is a practical advantage in implementing this step. While the measurement of the Mahalanobis distance at the m^{th} step rests on the m^{th} estimate of the covariance matrix, this does not happen with the Aitchison distance. If the covariance matrix is estimated with a finite level of precision, the Mahalanobis distance is affected by a statistical uncertainty which is absent in the Aitchison distance. In other words, the fluctuations in the estimate of the covariance matrix turn the value of the Mahalanobis distance into a range of values.

- the statistical tests, i.e. the construction of the envelopes has to take into account the distributional changes in the null hypothesis. We have seen in sect. I that a convenient choice for the distributional hypothesis of a compositional multivariate is expressed in terms of the log–normal distribution. Accordingly, the statistical tests performed on the signal have to be modified so as to compare the value of the $(m+1)^{\text{th}}$ ordered Aitchison distance with the thresholds corresponding to – say – the $\alpha = 0.01$ and $1-\alpha = 0.99$ percentiles of its distribution.

In order to obtain the percentiles of the distribution of the $(m+1)^{\text{th}}$ ordered Aitchison distance, we use the isometric property of the ILR. Under the null hypothesis, the observations of $\mathcal{D}$ are log–normally distributed. This entails that the observations of $\text{ilr}\,(\mathcal{C}(\mathcal{D}))$ are normally distributed. Moreover, the Aitchison distance between two observations of $\mathcal{D}$ equal the Euclidean distance between the corresponding ILR–transformed data, as shown in eq. (11). It follows that the distribution of the Aitchison distance under the null hypothesis coincides with the distribution of

the Euclidean distance under the normality hypothesis in the Euclidean space obtained from the ILR transformation.

9. The distribution of the squared Euclidean distance under the hypothesis of normally distributed multivariate data has been studied in [9]. Here it is shown that the cumulative distribution function $H(d^2;\mu,\Sigma)$ of the squared Euclidean distance $d^2 = d^2_{\mathrm{E}}(x,\mu)$ of a normally distributed variable $x \sim \mathcal{N}_n(\mu,\Sigma)$ from the mean μ, can be expressed in terms of an infinite sum of distributions of χ^2–type. More precisely, if $F_n(d^2)$ denotes the cumulative distribution function of χ^2_n, a central chi-square with n degrees of freedom, then $H(d^2;\mu,\Sigma)$ is given by

$$H(d^2;\mu,\Sigma) = \sum_{j=0}^{\infty} c_j(\mu,\Sigma,p) F_{n+2j}(d^2/p)\,, \tag{20}$$

where the coefficients c_j depend on the distributional parameters and can be recursively obtained, and p is a scale factor that can be adjusted to improve the convergence properties of the series. A description of the recursion is beyond the aims of the present report and is detailed in [10], as well as some technical aspects of the numerical implementation of eq. (20). Here, we just observe that the infinite sum can be truncated with no uncontrolled systematic error and that on average a very good approximation is given by the first few terms of the expansion. This leads to a fast evaluation of $H(d^2;\mu,\Sigma)$, and consequently of the envelopes of the FSA.

10. It should also be noticed that $H(d^2;\mu,\Sigma)$ depends on the distributional parameters μ and Σ, i.e. it is not a pivotal distribution. Accordingly, the advantage of having a clean signal, i.e. a signal which does not depend on the estimate of the covariance matrix, is compensated by the disadvantage of having thresholds that depend on such estimates. Unfortunately, this is the only possible way of designing a FSA for compositional data.

IV. A CASE STUDY: THE 2010 ITALIAN AGRICULTURAL CENSUS

11. As an example, we show an application of our compositional FSA to some preliminary data of the sixth Italian agricultural census, held in October 2010. Agricultural holdings are stratified by type of crop within each region of Italy. We focus on the province of Alessandria and the local holdings growing three cereals: *soft and spelt wheat*, *barley* and *corn*. Our data set is made of $N = 148$ trivariate observations, normalized so that $\kappa = 1$. Inverse fractional parts are shown against each other on a log–log scale in scatter-plots of Figs. 1–2. Logarithmic scales are needed in order to emphasize data with extremely low values of one or more parts. Out of the $N = 148$ observations, the compositional FSA detects 7 outliers, shown as green circled points. Last plot of Fig. 2 (forward plot) shows the signal *vs.* m and the $\alpha = 0.005$ and $1-\alpha = 0.995$ over–imposed envelopes at $N^* = 142$ (see ref. [6] for an explanation of the meaning of the over–imposed envelopes). Though not far from the bulk of the distribution, the units marked as outliers are statistically incompatible with the distribution of the remaining data.

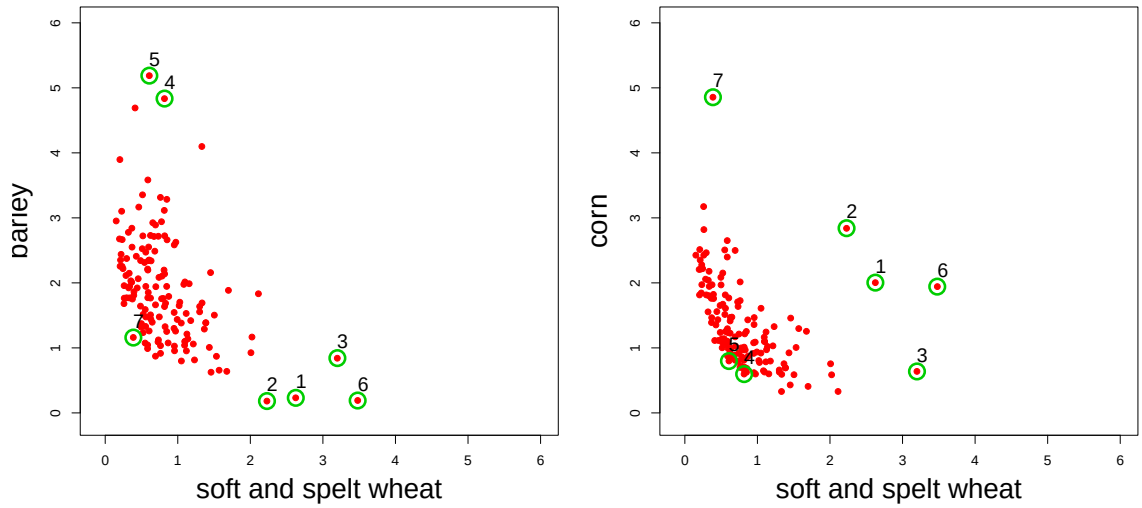

Figure 1: Agricultural surfaces of Italian holdings growing soft and spelt wheat, barley and corn in the province of Alessandria (source: ISTAT, Italian Agricultural Census 2010). Left/Right plots: inverse fractional surface parts (left: barley *vs.* soft and spelt wheat; right: corn *vs.* soft and spelt wheat) on a log–log scale; green circles denote data that have been marked as outliers by the compositional forward search algorithm.

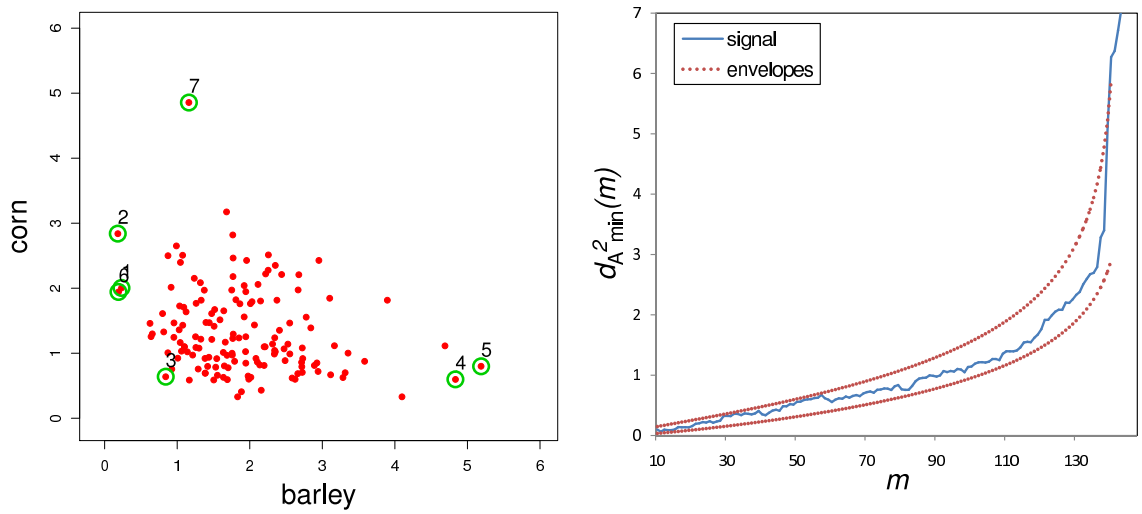

Figure 2: Agricultural surfaces of Italian holdings growing soft and spelt wheat, barley and corn in the province of Alessandria (source: ISTAT, Italian Agricultural Census 2010). Left plot: inverse fractional surface parts (corn *vs.* barley); green circles denote data that have been marked as outliers by the compositional forward search algorithm. Right plot: forward plot, the $(m+1)^{\text{th}}$ ordered Aitchison squared distance is plotted versus m; a true signal break–point takes place at $N^* = 142$ ($N = 148$), thus signalling the presence of 7 outliers.

V. CONCLUSIONS

12. In this report we have presented an extension of the *Forward Search Algorithm* for multi-

variate outlier detection to the case where data components satisfy a compositional constraint, i.e. they are represented by positive numbers with fixed sum (in a mathematical sense, compositional data live on a multidimensional simplex). The original *Forward Search Algorithm*, based on the Mahalanobis metric distance, is reformulated in terms of the Aitchison's one, which is better suited to the analysis of compositional data. The new formulation involves both the construction of a signal, represented by the minimum squared distance from the mean of an iteratively growing data subset, and the test statistics that are employed in order to detect signal breakpoints produced by outliers. The main ingredients of our construction are: (i) an Euclideanization of compositional data in terms of the *Isometric Logratio Transform* originally introduced by Aitchison; (ii) a numerical reconstruction of the distribution of Euclidean squared distances in terms of a chi-square series, according to an established result by Ruben (1962) (see [9]). As a case study, we probed the new algorithm with data coming from the latest Italian agricultural census (2010).

VI. ACKNOWLEDGEMENTS

13. The authors would like to thank Andrea Cerioli, Domenico Perrotta, Marco Riani and Francesca Torti for enlightening discussions on the FSA and a warm hospitality at the University of Parma, where this work has been firstly conceived.

References

[1] J. Aitchison, *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., 1986, London (UK).

[2] J. Aitchison and J. J. Egozcue, *Compositional Data Analysis: Where Are We and Where Should We Be Heading?*. Mathematical Geology, Vol. 37, No. 7(October 2005).

[3] J.J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras and C. Barcelò-Vidal, *Isometric Logratio Transformations for Compositional Data Analysis*. Mathematical Geology, Vol. 35, No. 3, April 2003, pp. 279-300.

[4] J. Aitchison, S. M. Shen, *Logistic-Normal Distributions: Some Properties and Uses*. Biometrika, Vol. 67, No. 2 (Aug., 1980), pp. 261-272

[5] A. C. Atkinson, *Fast very robust methods for the detection of multiple outliers*. Journal of the American Statistical Association , 89 (1994), pp. 1329-1339.

[6] M. Riani, A. Atkinson and A. Cerioli, *Finding an unknown number of multivariate outliers*, Journal of the Royal Statistical Society B (2009) 71, Part 2, pp. 447466

[7] G.M. Tallis, *Elliptical and Radial Truncation in Normal Populations*. The Annals of Mathematical Statistics, Vol. 34, No. 3 (Sep. 1963), pp. 940-944

[8] W.C. Günther, *An easy method for obtaining percentage points of order statistics*. Technometrics, Vol. 19, No. 3 (Aug. 1977), pp. 319-321

[9] H. Ruben, *Probability Content of Regions Under Spherical Normal Distributions, IV: The Distribution of Homogeneous and Non-Homogeneous Quadratic Functions of Normal Variables.* Annals of Mathematical Statistics, Vol. 33, No. 2 (1962), pp. 542-570

[10] F. Palombi, S. Toti, R. Filippini, A. Reale, *A forward search algorithm for compositional analysis.* In preparation.