Assessing the Precision of Compositional Data in a Stratified Double Stage Cluster Sample: Application to the Swiss Earnings Structure Survey

Monique Graf
Statistical Methods Unit, Swiss Federal Statistical Office, CH
monique.graf@bfs.admin.ch

Abstract

Precision of released figures is not only an important quality feature of official statistics, it is also essential for a good understanding of the data. In this paper we show a case study of how precision could be conveyed if the multivariate nature of data has to be taken into account. In the official release of the Swiss earnings structure survey, the total salary is broken down into several wage components. We follow Aitchison's approach for the analysis of compositional data, which is based on logratios of components. We first present different multivariate analyses of the compositional data whereby the wage components are broken down by economic activity classes. Then we propose a number of ways to assess precision.

Key words: Compositional data; complex survey; linearization; confidence domain; precision; coefficient of variation.

1 Introduction

There are several aspects of quality of surveys in public statistics (see e.g. [Eurostat, 2003] for the definition advocated by Eurostat). One aspect of quality is accuracy, the first feature of accuracy being precision in relation with sampling variability. Among the many ways used to assess precision, the most recommended indicator is the coefficient of variation (CV) [Eurostat, 2003]. Being dimensionless, it permits easy comparisons of precision among variables with different orders of magnitude. However in the case of multivariate data that are correlated by nature, like parts of a total, CV's are not enough to assess precision. We seek a generalization of the CV along the lines of multivariate statistics. This global CV will thus be related to the matrix norm of the covariance matrix of estimates. In the present case study, one aspect of the Swiss earnings structure survey is studied, namely the estimation of population wage components. We use the framework of compositional data analysis as developed by J. Aitchison [Aitchison, 1986]. The principle is to compute the logarithm of ratios of the components. The total variance is the sum of the variances of all possible logratios of components. If we divide the total variance by the number of ratios, we obtain an average variance. It will be shown that the linearized form of this average variance can be interpreted as an average squared CV. Other applications of compositional analysis to public statistics can be found in [Brundson and Smith, 1998], [Silva and Smith, 2001], [Anyadike-Danes, 2003], [Larrosa, 2003].

1.1 Basic notions on compositional vectors

Basic notions and notations on compositional data are recalled here. Compositional data are observations expressed as parts, thus having a unit sum constraint. A good mathematical summary of the principal notions can be found in [Aitchison, 2001], a less formal introduction in [Aitchison, 1997] and a thorough presentation of the theory in [Aitchison, 1986]. A compositional vector of length D, $(p_1, p_2, ...p_D)$ has strictly positive components with sum equal to 1:

$$p_1 + p_2 + ... + p_D = 1 \tag{1}$$

The set of these vectors is the simplex S^D. Equation (1) implies that

$$\mathbf{V}(p_1 + p_2 + ... + p_D) = 0 \Rightarrow \sum_{i \neq j} \mathbf{Cov}(p_i, p_j) = -\mathbf{V}(p_j) \quad \forall j$$

so there is necessarily a negative correlation between the components. This shows that the correlations are not directly interpretable. To release this constraint, Aitchison proposes that we consider the vector of ratios of the $d = D - 1$ first components to the last, that is

$$x = (x_1, ..., x_d) = (p_1, ..., p_d)/p_D = p_{-D}/p_D \tag{2}$$

and then to take the logarithm $y = \ln x$. Applying this transformation , the resulting vector is no longer constrained and correlations between y_i and y_j can be interpreted.

1.1.1 Center

The center of the distribution for x is given by the geometric - and not the arithmetic - mean of the compositions. Its theoretical counterpart is:

$$\xi = \exp(\mathbf{E}(\ln x)) \tag{3}$$

It can be transformed back to give the center for p:

$$\begin{aligned} \mathbf{cen}(p_i) &= \frac{\xi_i}{1 + \sum_{j=1}^{d} \xi_j} \quad i = 1, ..., d \\ \mathbf{cen}(p_D) &= \frac{1}{1 + \sum_{j=1}^{d} \xi_j} \end{aligned}$$

1.1.2 Dispersion

There are different equivalent and linearly related dispersion matrices [Aitchison, 1986, § 4.8].

1. The $d \times d$ - covariance matrix of the logratios:

$$\mathbf{\Sigma} = [\sigma_{ij}] = \mathbf{Cov}(\ln x_i, \ln x_j) = \mathbf{Cov}\left(\ln \frac{p_i}{p_D}, \ln \frac{p_j}{p_D}\right) \tag{4}$$

   The only drawback in this setting is that the last component p_D is treated differently.

2. All components of p are handled symmetrically, if they are divided by the geometric average $g(p) = \left(\prod_1^D p_i\right)^{1/D}$. The $D \times D$ - centered covariance matrix is given by:

$$\mathbf{\Gamma} = [\gamma_{ij}] = \mathbf{Cov}\left(\ln \frac{p_i}{g(p)}, \ln \frac{p_j}{g(p)}\right) \tag{5}$$

   which is singular, because $\sum \ln \frac{p_i}{g(p)} = 0$.

3. The last possibility is to use the $D \times D$ -variation matrix, with all elements being variances:

$$\mathbf{T} = [\tau_{ij}] = \mathbf{V}\left(\ln \frac{p_i}{p_j}\right) \tag{6}$$

   This matrix has a zero principal diagonal and only one positive eigenvalue, corresponding to eigenvector 1_D.

1.1.3 Asymptotic distribution

Under regularity conditions, $y = (\ln x_1, \ln x_2, ..., \ln x_d)'$ is asymptotically normally distributed $N^d(\mu, \mathbf{\Sigma})$ (with $\mu = \ln \xi$ and $\mathbf{\Sigma}$, given by Equation (4). The derived distribution for $p = (p_1, p_2, ...p_D)$ on the simplex S^D is called the additive logistic normal distribution and denoted by $L^d(\mu, \mathbf{\Sigma})$.

1.1.4 Confidence domains

Under the asymptotic distribution hypothesis,

1. for y the confidence domain $D_{1-\alpha}(y)$ is limited by a d dimensional ellipsoid. Let $\chi^2_{d;1-\alpha}$ be the $(1-\alpha)$ quantile of the chi-square distribution with d degrees of freedom. Then
$$D_{1-\alpha}(y) = \left\{ y \in R^d \mid (y-\mu)' \mathbf{\Sigma}^{-1} (y-\mu) \leq \chi^2_{d;1-\alpha} \right\}$$
2. for x in Equation (2) , the equivalent domain is
$$D_{1-\alpha}(x) = \left\{ x \in R^d_+ \mid (\ln x - \mu)' \mathbf{\Sigma}^{-1} (\ln x - \mu) \leq \chi^2_{d;1-\alpha} \right\} \tag{7}$$
3. for $p = (p_1, ..., p_D)$, the domain is a subset of the simplex S^D:
$$D_{1-\alpha}(p) = \left\{ p \in S^D \mid \left(\ln \frac{p_{-D}}{p_D} - \mu \right)' \mathbf{\Sigma}^{-1} \left(\ln \frac{p_{-D}}{p_D} - \mu \right) \leq \chi^2_{d;1-\alpha} \right\} \tag{8}$$

1.1.5 Total variance

Whereas a thorough study of the precision of a composition implies computing a d-dimensional confidence domain, but we also need a simple global characterization of precision. This is generally given by a matrix norm of the covariance matrix. Aitchison defines (among other measures) the total variance for which different equivalent formulations exist [Aitchison, 1986, Chapter 4]:

$$\begin{aligned} \mathbf{totvar}(p) &= \mathbf{tr}(\mathbf{\Gamma}) = \sum_{i=1}^{D} \mathbf{V} \left(\ln \frac{p_i}{g(p)} \right) && (9) \\ &= \frac{1}{D} \sum_{i<j} \mathbf{V} \left(\ln \frac{p_i}{p_j} \right) = \frac{1}{2D} \sum_{i,j=1}^{D} \tau_{ij} && (10) \end{aligned}$$

or

$$\begin{aligned} \mathbf{totvar}(p) &= \frac{1}{2D} \sum_{i,j=1}^{D} \tau_{ij} = \frac{1}{2D} 2 \left(D\mathbf{tr}(\mathbf{\Sigma}) - 1_d' \mathbf{\Sigma} 1_d \right) \\ &= \mathbf{tr}(\mathbf{\Sigma}) - \frac{1}{D} 1_d' \mathbf{\Sigma} 1_d && (11) \end{aligned}$$

2 The Swiss earnings structure survey

The Swiss earnings structure survey (SESS) is a biennial written survey sent out to businesses. The survey is constructed on a stratified double stage cluster sampling scheme [Graf, 2004]. The 2002 sample is rather large: 1/3 of all businesses in Switzerland are involved. This means that finite population corrections (fpc) are indispensable for realistic estimates of the precision of the population values. The extrapolation weights and the finite population correction take non response into account (which we suppose is ignorable within the stratum). The variance estimation method applied here is the classical linearization of the estimators, see e.g. [Särndal and others, 1992]. Other aspects of precision computed for the 2000 survey were studied in (Graf 2002a, 2002b), see also [Eurostat, 2002]. A general report on the 2002 survey can be found in (SESS 2003, 2004).

2.1 Design

The sampling frame is the business register (BR) in its latest state at the time of sampling. The stratification was originally designed as a combination of 41 activity classes, 3 business size classes and 13 regional subdivisions. In the SESS, the activity classes are the NOGA at 2 digits level with some grouping in order to avoid the appearance of very small strata (see Table A1)[1]. Class 0 represents the total of all activities considered.

The survey design is a stratified two stages cluster sampling, with a simple random (SI) sample of businesses in each stratum and a SI sample of salaries within each sampled business. The sampling fraction at both stages depends on the size class. The largest businesses form exhaustive strata. For medium and small businesses, a Neyman allocation based on the variance of the mean standardized gross earnings of the preceding survey is computed. The strata sizes are then modified so that a minimum of 10 units are sampled (if stratum size permits). Large businesses have to furnish 33% of the earnings paid out in October. Medium size businesses furnish 50% while the smallest businesses give them all. The desired sampling fraction and the expected non-response rate are used to determine the number of businesses to contact.

2.2 Calibration - robustification of weights

The non response is assumed to be ignorable at the stratum level and the Horvitz-Thompson weights at both stages are in principle used (the actual number of salaries paid by the business in October is asked in the questionnaire). Few expansion weights are large due to non-response. To robustify the procedure, these weights were trimmed, first at the cluster level, and then at the stratum level. The resulting weights are recalibrated using the CALMAR raking procedure[2], in such a way that the marginal total weights on the 3 stratum classifications remain constant. Thus the "unreliable" estimates are weighted down without changing the total. If the whole population is considered, this procedure changes the results very little. The sampling plan was designed for the main variable, namely the monthly standardized gross earnings. In this study, we are interested in the compositional analysis of the weighted total of monthly non standardized total salary.

3 Compositional analysis of wage components

In the SESS, 5 wage components are published (social security contributions, overtime earnings, hardship allowances, 13th or n-th salary, bonuses), see [SESS, 2004], [SESS, 2003]. They are reproduced here (Table A1, Appendix). In Table A1, wage mass is defined for an economic branch as the extrapolated sum of all sampled salaries, using the above calibrated weight. The "non standardized total monthly salary", MBLIU, is the sum of the 5 components and the rest (the "naked" salary), which forms a 6th component and is never published as such. The non-standardized gross monthly salary BLIMOK, includes the "naked" salary and social security contributions, but *excludes* components 2 to 5. The defined components are summarized in Table 1. The wage percentage attributed to each component in Table A1 are computed relative to BLIMOK, and not to MBLIU. Thus the published proportions are:

$$(s_1, s_2, s_3, s_4, s_5)\,/(s_1 + s_6) \tag{12}$$

We see that Table 12 contains two different subcompositions, the first is a 2-dimensional composition expressed as a part $s_1/(s_1 + s_6)$, and the second is 5-dimensional, expressed as ratios of components $(s_2, s_3, s_4, s_5)\,/(s_1 + s_6)$.

These two subcompositions will be analyzed separately. We stress that in this framework, the interest is not in the wage composition at the individual level, but in the global composition for

[1]The NOGA is the Swiss version of the Statistical Classification of Economic Activities in the European Community, Revision 1 (NACE Rev. 1). Both classifications are till the 4th level identical.

[2]SAS macro written at the French national statistical office INSEE.

Table 1: Wage mass attributed to the different components

Definition	Code	Total amount
social security contributions	SOZABG	s_1
overtime earnings	VERDUZ	s_2
hardship allowances	ZULAGEN	s_3
13th or n-th month salary (/12)	XIII12E = ROUND(XIIILOHN/12)	s_4
Special payements/12 bonuses	SOND12E = ROUND(SONDERZA/12)	s_5
"naked" salary	-	s_6
non-standardized gross earnings with social contributions	BLIMOK	$s_1 + s_6$
monthly non standardized total salary	MBLIU	$\sum_{i=1}^{6} s_i$

segments of the population. The advantage from a mathematical point of view is that no zero components are observed, while they exist at the individual level.

3.1 Variance estimation

The variance-covariance matrix of the estimates is based on the sampling distribution of the wage components. Computing the variance of these global compositions in a stratified double stage cluster sample is a complex task, because no closed formula for the variance is available. The large sample size implies that finite population corrections (fpc) are indispensable for realistic estimates of the precision of the population values. The extrapolation weights and the finite population correction take the non response into account (which we suppose ignorable within the stratum). The variance estimation method applied here relies on the linearization of the estimators and is equivalent to the recovery of the compositional variation array from the crude mean vector and covariance matrix, see [Aitchison, 1986, §4.4]. (An alternative would be to use resampling methods, but it would be extremely cumbersome in this large survey). In fact we simply use the first order approximation, which is a slight overestimation:

$$\mathbf{V}\left(\ln \widehat{X}\right) \cong \mathbf{E}\left(\ln \widehat{X} - \ln X\right)^2 \cong \mathbf{E}\left(\frac{\widehat{X} - X}{X}\right)^2 = \mathbf{CV}^2\left(\widehat{X}\right) \tag{13}$$

where **CV** is the coefficient of variation.

For a ratio:

$$\mathbf{V}\left(\ln \frac{\widehat{X}}{\widehat{Y}}\right) \cong \mathbf{E}\left(\frac{\widehat{X} - X}{X} - \frac{\widehat{Y} - Y}{Y}\right)^2 \tag{14}$$

$$= \left(\frac{X^2}{Y^2}\right)^{-1} \left\{\frac{1}{Y^2}\mathbf{E}\left(\widehat{X} - \frac{X}{Y}\widehat{Y}\right)^2\right\} \cong \mathbf{CV}^2\left(\frac{\widehat{X}}{\widehat{Y}}\right) \tag{15}$$

We recognize in the left expression in brackets in Equation (15) the formula for the linearization of the variance of a ratio. Practically a program for computing the linearized variance of a ratio will do the job.

In matrix form:

$$\mathbf{V}\left(\ln \frac{\widehat{X}}{\widehat{Y}}\right) \cong \begin{pmatrix} \frac{1}{X} & -\frac{1}{Y} \end{pmatrix} \Sigma_{\widehat{X},\widehat{Y}} \begin{pmatrix} \frac{1}{X} \\ -\frac{1}{Y} \end{pmatrix} \tag{16}$$

where $\Sigma_{\widehat{X},\widehat{Y}}$ is the covariance matrix of $\widehat{X}$ et $\widehat{Y}$. The covariance is found by a variance computation using

$$\mathbf{Cov}\left(\widehat{X},\widehat{Y}\right)=\frac{1}{2}\left[\mathbf{V}\left(\widehat{X}+\widehat{Y}\right)-\mathbf{V}\left(\widehat{X}\right)-\mathbf{V}\left(\widehat{Y}\right)\right] \tag{17}$$

Once the covariance matrices of the logratio of the wage components to the gross salary are obtained, we are in position to 1. assess the accuracy of the population composition estimates, and 2. test hypotheses on differences in composition between subpopulations. Graphical representations and interpretations of the results will be presented.

3.2 Proportion of the social contributions within the gross salary

With s_1 and s_6 as defined in Table 1, let

$$q=s_1/(s_1+s_6) \tag{18}$$

Our variable of interest is q, which represents the proportion of the social security contributions within the non-standardized gross monthly earnings BLIMOK. $1-q$ is the "naked salary" part.

In this case the compositional vector is of length $D=2$ and is denoted by $\widetilde{q}=(q,1-q)$. $\widetilde{q}$ can be replaced by the equivalent form of length $d=1$

$$x'=q/(1-q)$$

Having computed $\mathbf{CV}(q)$,the following 95% confidence intervals for q are obtained:

1. *Normal approximation CI for* q:

$$[bn_{l95},bn_{u95}]=q\left(1\pm 1.96\mathbf{CV}(q)\right) \tag{19}$$

2. *Log-normal approximation CI for* $\ln\left(\frac{q}{1-q}\right)$, using Equation (16):

$$[bl_{l95},bl_{u95}]=\ln\left(\frac{q}{1-q}\right)\pm 1.96\frac{\mathbf{CV}(q)}{(1-q)}$$

3. *CI for* q *deduced from 2. (logistic normal approximation)*

$$[b_{l95},b_{u95}]=\left[\frac{\exp\left(bl_{l95}\right)}{\exp\left(bl_{l95}\right)+1},\frac{\exp\left(bl_{u95}\right)}{\exp\left(bl_{u95}\right)+1}\right] \tag{20}$$

Results The intervals given by Equations (19) and (20) are very similar: the maximum difference is 0.01% (NOGA2=61, water transport, which is marginal in Switzerland). We prefer Equation (20) which has the advantage to guarantee that the bounds are in $[0,1]$.

The social security contributions by economic activity classes are presented in Figure 1. Because observations in different activity classes are independent, we can easily test 2 by 2 differences. If q_1 and q_2 are the proportions of the social security contributions in classes 1 and 2, then equality of proportions is rejected with risk α, if

$$\frac{\left|\ln\left(\frac{q_1}{1-q_1}\right)-\ln\left(\frac{q_2}{1-q_2}\right)\right|}{\sqrt{\left(\frac{\mathbf{CV}(q_1)}{(1-q_1)}\right)^2+\left(\frac{\mathbf{CV}(q_2)}{(1-q_2)}\right)^2}}>z_{1-\alpha/2}$$

where is the $(1-\alpha/2)$-quantile of the standard normal distribution.

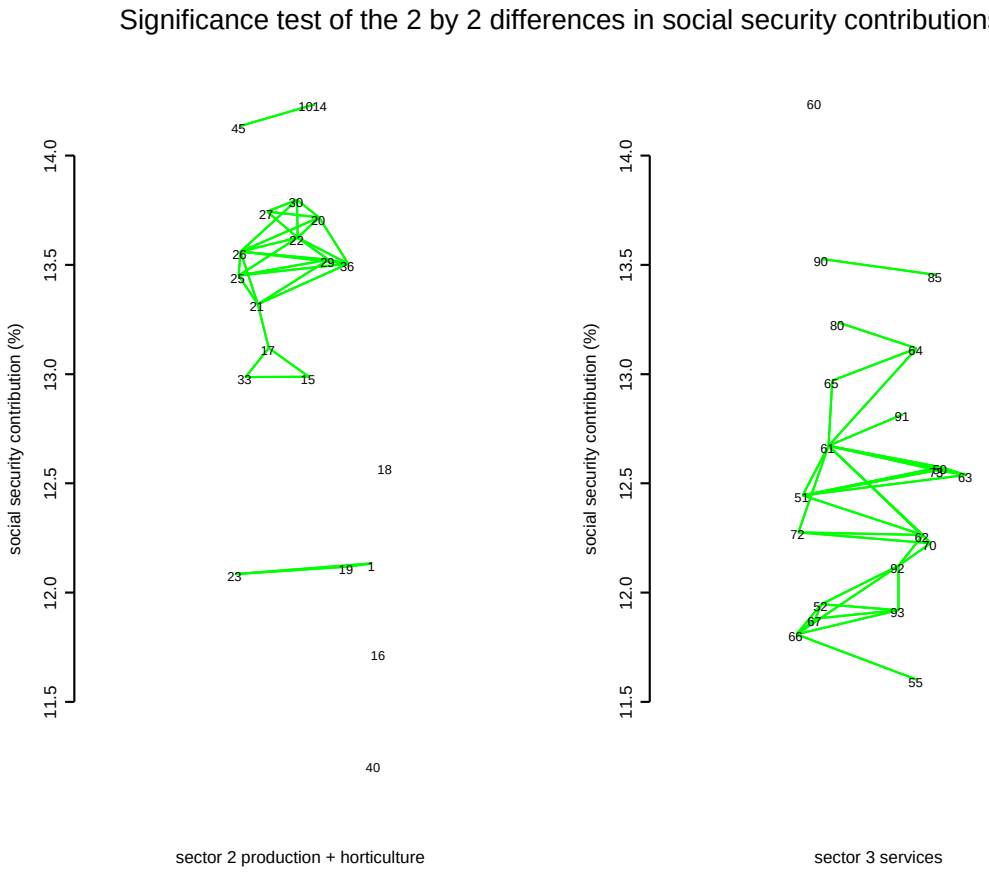

Figure 1: Non significant differences in proportions at the 5 % risk are represented by a connecting segment.

Results are given in Figure 1, separately for production and services. Non significant differences in proportions (with $\alpha = 5\%$) are connected by a segment. The horizontal axis has no meaning: a small random quantity is generated for the abscissa so that the connecting segments become distinguishable. It can be seen that in the secondary sector, a larger proportion of the non-standardized gross monthly salary BLIMOK is generally devoted to social security contributions than in the third. Activities 45 (construction), 10-14 (Mining and quarrying of stone) and 60 (land transport/pipelines) have the highest contributions. At the other extreme, 40 (electricity, gas and water supply) devotes the smallest part of the overall wage bill to social security. 61 (water transport) is the least precise, being connected to remote classes on the vertical scale. Table A2 (Appendix) shows the corresponding p-values.

3.3 Other components of the gross salary

From the six-parts compositional vector $(p_1, p_2, p_3, p_4, p_5, p_6)$, let us form a new composition by amalgamation of components 1 and 6:

$$p = (p_2, p_3, p_4, p_5, p_1 + p_6) = \frac{(s_2, s_3, s_4, s_5, s_1 + s_6)}{\sum_{i=1}^{6} s_i} \tag{21}$$

This vector can be written in the equivalent form

$$x = (x_1, x_2, x_3, x_4) = \left(\frac{s_2}{s_1 + s_6}, \frac{s_3}{s_1 + s_6}, \frac{s_4}{s_1 + s_6}, \frac{s_5}{s_1 + s_6} \right)$$

Interpreting $s_i, i = 1, ..., 6$ as in Table 1, we see that p represents a decomposition of the non standardized total monthly salary MBLIU into 5 components, and that x equals the ratios of the 4 first components VERDUZ ... SOND12E to the fifth (the non-standardized gross monthly salary BLIMOK). For each economic activity grouping, the last 4 columns of Table A1 are the components x_i, expressed in %.

3.3.1 Multivariate analyzes of the estimated components

Before we proceed to the computation of the precision, it is interesting to get a rough idea of the data. A multidimensional scaling on logratio estimates was performed, using as distance between 2 economic activities the Euclidian distance between the corresponding logratio vectors. (The same result would be obtained by principal component analysis). The 2 panes in Figure 2 represent the same projection onto the first two principal axes [3]. This plane explains 90% of the total variability. Thus the distance between 2 points in the plane can be interpreted as a measure of discrepancy between the corresponding compositional vectors. In the left pane activity classes are coded by their NOGA2 code, whereas in the right pane, they are represented by a star plot (the half diagonals of the quadrilateral are proportional to the $x_i, i = 1...4$). A partition was also performed (using the 4 new coordinates) by the PAM method (partition around medoids) and an optimal number of 3 groups was obtained. The groups are visible on the left pane (circles, triangles and diamonds). It is typical for the first group (diamonds) to report a large share of "special payments/bonuses" and a relatively small portion of "13th month salary". It is also typical for the share of "overtime pay" and "hardship allowances" to be practically nonexistent. All eight branches in this group fall in the tertiary sector.

The second group (triangles) counterbalances the first group to a certain extent. Apart from "13th month salary", the branches in this group report practically no wage components. This indicates that in these branches pay is limited to base monthly salary. The nine branches in this group are arranged according to economic sector and number of employees. That said, the vast majority of the branches in this group fall in the tertiary sector.

It is typical for the third group (circles) to report relatively small shares of wage components, with the exception of "13th month salary", especially when it comes to "overtime pay". In addition, there are three times more branches of trade in this group than in the first two groups. Most (i.e. sixteen of the twenty-four branches) fall in the secondary sector.

All things considered, it can be said that the shares of the four wage components vary considerably. The share of "13th month salary" is about the same in all branches; the share of "hardship allowances" and "overtime pay" are small to very small, which makes it difficult to assess them in the chart; in contrast, the share of "special payments/bonuses" varies considerably from branch to branch.

[3] The usual terminology is "principal components"; the expression "principal axes" is being used instead, in order to avoid confusion with the salary components.

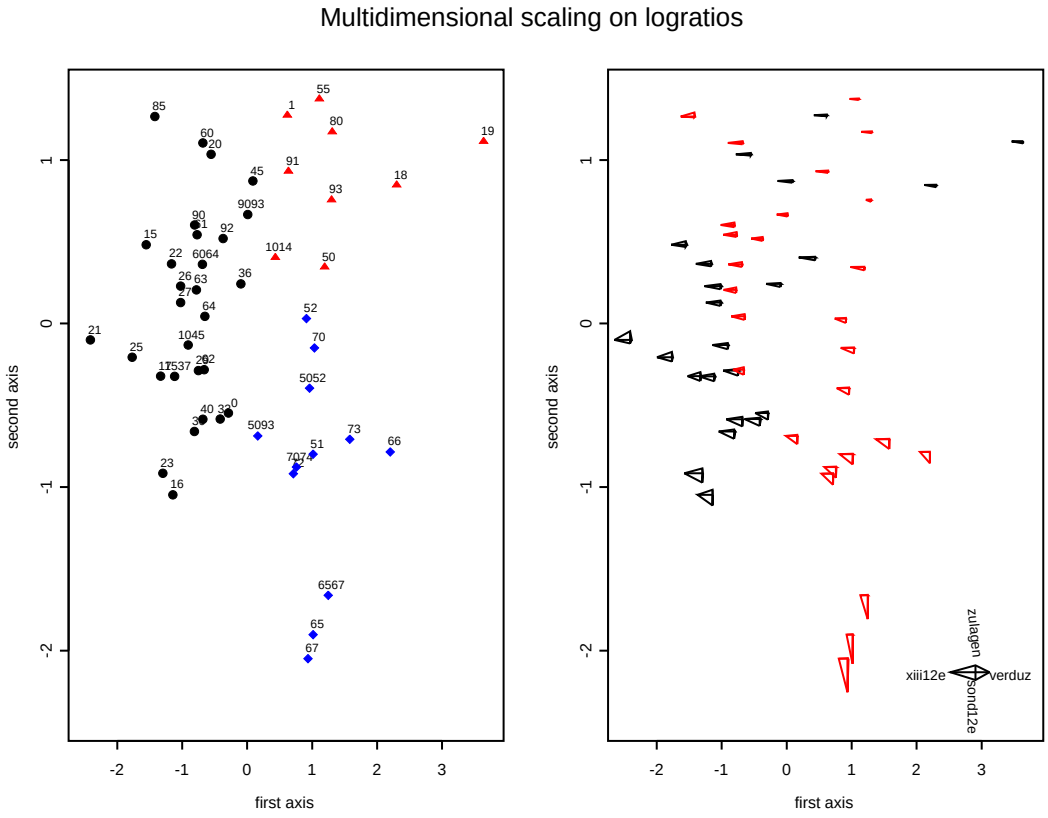

Figure 2: Multidimensional scaling on estimated logratios corresponding to the 5-dimensional composition for all activity classes. Circles, triangles and diamonds on the left pane give the group membership computed by PAM; numbers are the NOGA classes. The right pane shows star plots.

3.3.2 Univariate statistics

Let $\xi_i = \mathbf{E}(x_i)$ and $\mu_i = \mathbf{E}(\ln x_i)$. Denote the coefficient of variation by $\mathbf{CV}$. By linearization and by Equation (13)

$$\mu_i \cong \ln \xi_i \tag{22}$$

$$\ln x_i - \mu_i \cong \frac{x_i - \xi_i}{\xi_i} \tag{23}$$

$$\mathbf{V}(\ln x_i) = \mathbf{E}(\ln x_i - \mu_i)^2 \cong \frac{\mathbf{V}(x_i)}{\xi_i^2} = \mathbf{CV}^2(x_i) \tag{24}$$

We get the following 95% CI:

1. *Normal approximation for* x_i:

$$\left[bn_{l95}^{(i)}, bn_{u95}^{(i)}\right] = x_i\left(1 \pm 1.96\mathbf{CV}(x_i)\right)$$

2. *Normal approximation for* $\ln(x_i)$:

$$\left[bl_{l95}^{(i)}, bl_{u95}^{(i)}\right] = \ln(x_i) \pm 1.96\mathbf{CV}(x_i)$$

3. *Deduced CI for* x_i (*lognormal approximation*):

$$\left[b_{l95}^{(i)}, b_{u95}^{(i)}\right] = \left[\exp\left(bl_{l95}^{(i)}\right), \exp\left(bl_{u95}^{(i)}\right)\right] = x_i \exp\left(\pm 1.96 \mathbf{CV}(x_i)\right) \tag{25}$$

The **CV**'s can be found in Table A1. If we postulate a lognormal distribution for x, the $\mathbf{CV}(x_i)$ is given by $\exp(\sigma_{ii}) - 1$ which slightly overestimates $\sigma_{ii} = \mathbf{V}(\ln x_i)$.

3.3.3 Covariances and correlations

The univariate confidence intervals are misleading because they ignore the dependencies between components. By the linear approximation in Equation 23, matrix $\mathbf{\Sigma} = [\sigma_{ij}]$ is given by

$$\sigma_{ij} = \mathbf{Cov}(\ln x_i, \ln x_j) \cong \frac{\mathbf{Cov}(x_i, x_j)}{\xi_i \xi_j} \tag{26}$$

The approximation of $\mathbf{\Sigma}$ given by Equation 26 can be seen as a multivariate form of the coefficient of variation. Moreover the correlations are given by correlations of x:

$$\rho_{ij} = \mathbf{Cor}(\ln x_i, \ln x_j) \cong \frac{\mathbf{Cov}(x_i, x_j)}{\xi_i \xi_j \mathbf{CV}(x_i)\mathbf{CV}(x_j)} = \frac{\mathbf{Cov}(x_i, x_j)}{\sqrt{\mathbf{V}(x_i)\mathbf{V}(x_j)}} = \mathbf{Cor}(x_i, x_j) \tag{27}$$

which is not surprising, because the approximation of $\ln x_i$ is linear in x_i.

These correlations (see Table A3, Appendix) should not be used for sociological interpretations, because the finite population correction implies that exhaustive strata (with full response) are excluded from the calculations and that other strata have different weights. Thus the correlations are only useful for evaluating the precision of the global ratios, and have no other interpretation.

3.3.4 Confidence domains

Let $\mathbf{R} = [\rho_{ij}]$ be the 4×4 correlation matrix with elements given by Equation 27, and let us approximate μ by $\ln \xi$ [Eq. 22].

The approximate confidence domain [Eq.7] at level $1 - \alpha$ is given by:

$$D_{1-\alpha}(x) = \left\{x \in R_+^4 \mid \mathbf{Q}(x) \leq \chi^2_{4;1-\alpha}\right\} \tag{28}$$

where

$$\mathbf{Q}(x) = \left(\begin{array}{cccc} \frac{\ln x_1 - \ln \xi_1}{\mathbf{CV}(x_1)} & \frac{\ln x_2 - \ln \xi_2}{\mathbf{CV}(x_2)} & \frac{\ln x_3 - \ln \xi_3}{\mathbf{CV}(x_3)} & \frac{\ln x_4 - \ln \xi_4}{\mathbf{CV}(x_4)} \end{array} \right) \mathbf{R}^{-1} \left(\begin{array}{c} \frac{\ln x_1 - \ln \xi_1}{\mathbf{CV}(x_1)} \\ \frac{\ln x_2 - \ln \xi_2}{\mathbf{CV}(x_2)} \\ \frac{\ln x_3 - \ln \xi_3}{\mathbf{CV}(x_3)} \\ \frac{\ln x_4 - \ln \xi_4}{\mathbf{CV}(x_4)} \end{array} \right) \tag{29}$$

In the coordinates $(\ln x_1, \ln x_2, \ln x_3, \ln x_4)$, this domain is a 4-dimensional ellipsoid. In the coordinates (x_1, x_2, x_3, x_4), the shape of the domain is similar to a drop. There is no direct relationship between the length of the one-dimensional confidence intervals and the limits of the corresponding 4-dimensional confidence domain.

3.3.5 Barycentric coordinates

3-part compositions can be seen as points within an equilateral triangle with height 1, in which each vertex represents 100% in the corresponding part. To visualize the 95% confidence domains above, let us split the 5-part composition into two 3-part compositions: an amalgamation

(VERDUZ+ZULAGEN+SOND12E, XIII13E, BLIMOK), and a sub-composition (ZULAGEN, SOND12E, VERDUZ) (see [Aitchison, 1986] for a thorough description of amalgamation and subcomposition). Both are unit-sum compositional vectors. Because our approximation is linear, it is easy to deduce the corresponding covariance matrices from $\mathbf{\Sigma}$.

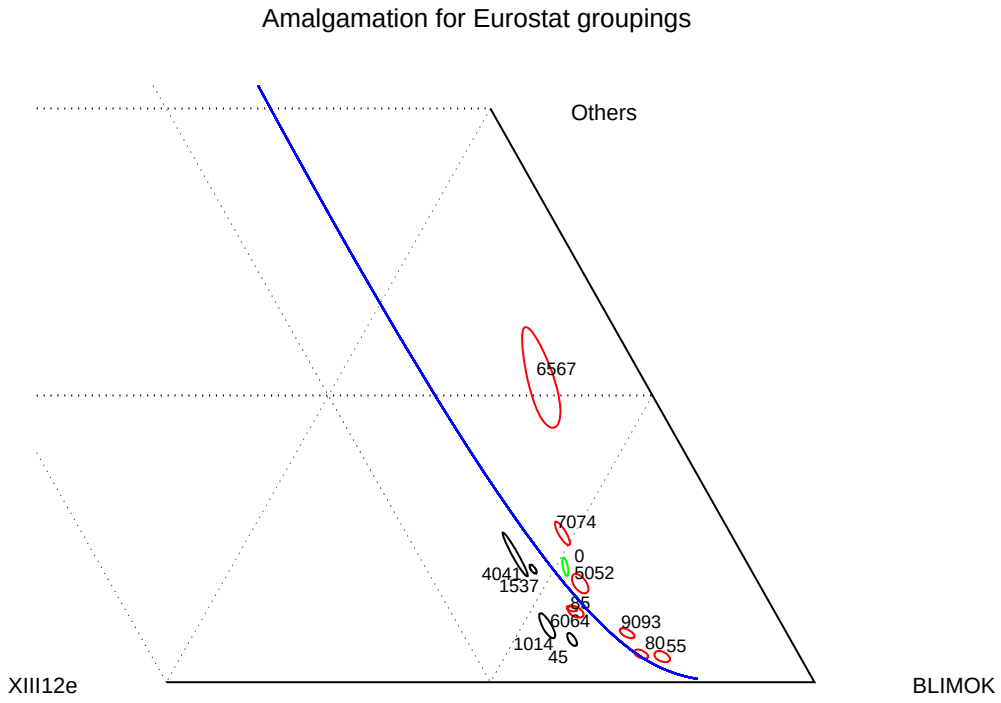

Figure 3: 95% confidence domains per activity classes for the amalgamation (ZULAGEN+SOND12E+VERDUZ, XIII13E, BLIMOK) and image of a robust regression line in the logratio 2-dimensional space. Dotted lines are spaced by 10%.

Figures 3 and 4 show the results for the economic activity groupings requested by Eurostat. We see that the amalgamations (Figure 3) are very precisely estimated, the worst is for banking, insurance 65-67 for which the uncertainty is essentially in the demarcation between the gross earnings (BLIMOK) and OTHERS. Apart for this group, OTHERS is never larger than 5%. The image of a robust regression line[4] of $\ln(x_3)$ onto $\ln(x_1 + x_2 + x_4)$ shows that groupings from Production are always below, indicating a larger share of 13th salary (XIII12E) in Production than in Services. Figure 4 shows how the small amount of OTHERS is distributed among the remaining components. In general, the subcompositions have a minimum of 50% share in bonuses (SOND12E) and very little parts of overtime earnings (VERDUZ) and hardship allowances (ZULAGEN), except grouping 85 "health and social work" which shows narrowly 20% of SOND12E but more than 80% of VERDUZ. The subcompositions are fairly precisely estimated. One exception is the grouping 55 "Hotels

[4] An ordinary least squares line would have been attracted by 65-67.

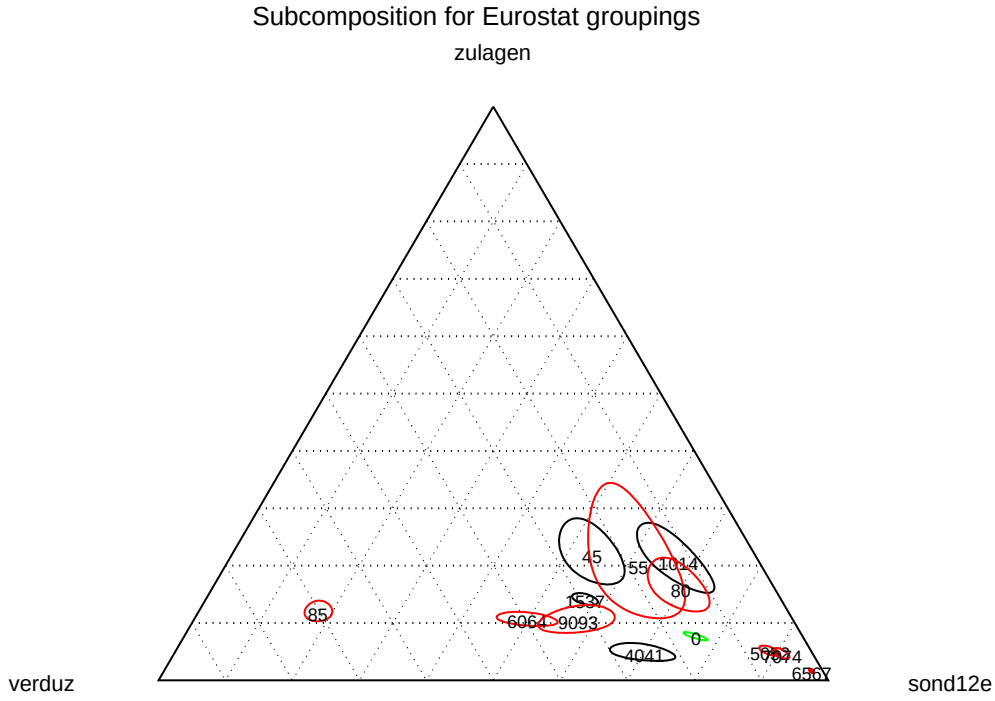

Figure 4: 95% confidence domains for a subcomposition.

and restaurants" which indicates a rather large uncertainty in the separation between SOND12E and ZULAGEN. Considered in perspective of Figure 3, where it is seen that the part OTHERS of the grouping 55 represents only 1%, the rather large uncertainty visible in Figure 4 looses its importance.

3.4 Global CV

The total variance per activity class is computed by [Eq. (11)]. Taking the first formulation in [Eq. (10)], we define an average standard deviation for logratios by

$$\mathbf{stot}(p) = \sqrt{\frac{\mathbf{vartot}(p)}{(D-1)/2}} = \sqrt{\frac{\sum_{i<j} \mathbf{V}\left(\ln \frac{p_i}{p_j}\right)}{D\,(D-1)\,/2}} \tag{30}$$

In its linearized form, $\mathbf{V}\left(\ln \frac{p_i}{p_j}\right) \approx \mathbf{CV}^2\left(\frac{p_i}{p_j}\right)$. We can interpret [Eq. (30)] as a L_2-average of the CV's of all possible ratios of components (i.e. the square root of the mean square CV's). Practically, it is computed using the linearized form of [Eq.(4)] and the second interpretation of the total variance [Eq.(10)]. It is given in the last column of Table A1 under the heading "global CV". The global CV is a good candidate for an overall assessment of precision and has a theoretical counterpart in the theory of compositions. In this application, the global CV is always between the extremes of the 4 corresponding CV's. The barycentric representation of the 10 economic activity classes with the largest average standard deviation (Fig. 5 and 6) show a differing pattern

of variability between classes. For class 67 the observed uncertainty is essentially in the sharing out between gross salary BLIMOK and "others", but category "others" in this case is practically only Special payments SOND12E. For class 18, which has the second largest average standard deviation for logratios, the amalgamation is very precisely estimated. The largest variability is in the subcomposition where we see that the variability lies in the relative parts of SOND12E and ZULAGEN. It is interesting to note that the breaking down of the 5-dimensional composition [Eq.(21)] into the above amalgamation and subcomposition is sufficient for the recovery of the original compositions, but not for the complete original covariance matrix.

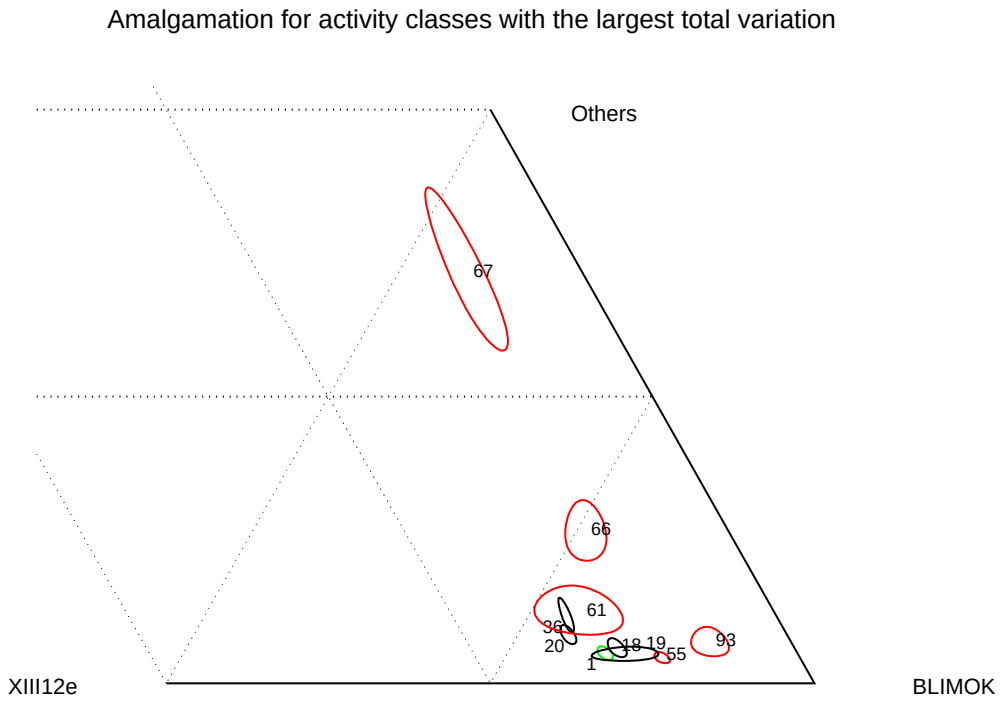

Figure 5: Representation of the amalgamation (ZULAGEN+SOND12E+VERDUZ, XIII13E, BLIMOK) of the activity classes with the 10 largest total variation. (Dotted lines: 10% apart).

3.5 Discussion

The lack of precision in Table A1 is linked with very small proportions, i.e. with a large discrepancy between components. If the discrepancy in components is large, the geometric mean $g(p)$ will be small. For a given dimension D, $\max(g(p)) = 1/D$ is attained for the uniform composition. A plot of $g(p)$ in function of **stot**(p) (Fig. 7) shows a clear relationship between discrepancy and variability. Points are coded by the PAM groupings (Fig. 2). This gives a further interpretation of the groups: dots group has the least discrepant and least variable compositions; diamonds group has more discrepancy but is still precisely estimated; while for the triangles group, the compositions

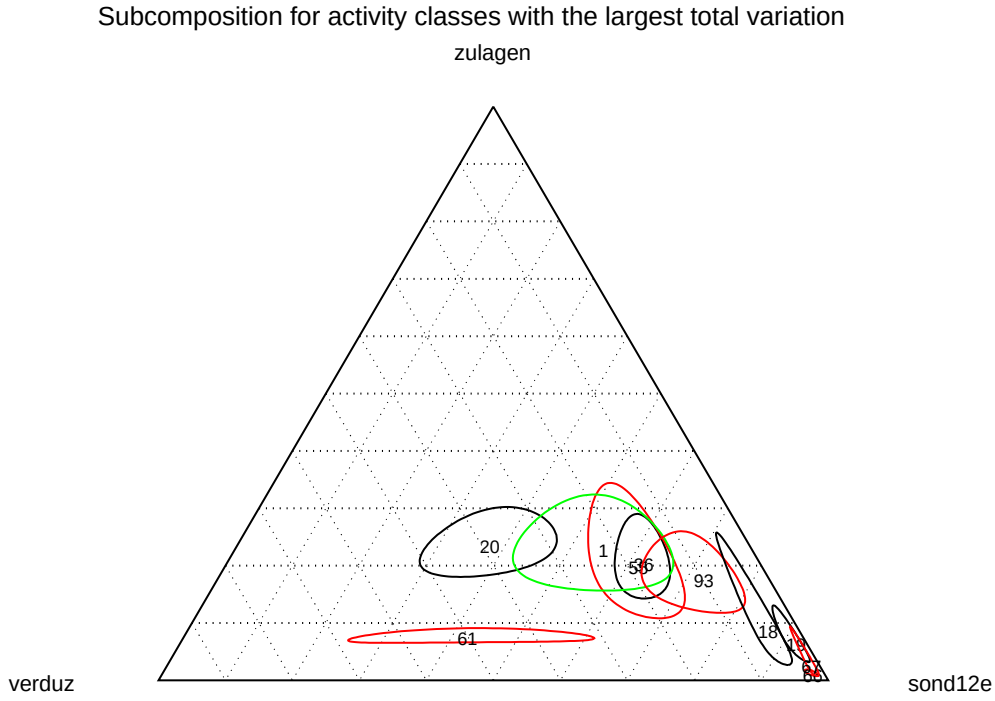

Figure 6: 95% confidence domains for the subcomposition of activity classes with the largest total variation.

generally have at least one very small component and also the largest average standard deviation. We see that the top half of the dots group is exclusively formed by activities from the production sector.

Using the independence between the estimates of compositions for two different activity classes, we can compute the covariance matrix of the difference of compositional logratios. Under the asymptotic distribution, 2 by 2 tests of differences between the 5-dimensional compositions [Eq. (21)] of economic activity classes within sectors have been processed at the 5% risk, and show that the null hypothesis of no difference is generally rejected. A small group from 60-64 (transport, storage and communication) are mutually not different in wage compositions, namely 61 (water transport), 63 (supporting and auxiliary transport activities) and 64 (post and telecommunications). Only one other non-significant difference is found between 61 (water transport) and 90 (water processing and other disposal). We conclude that the SESS has a good discriminating power for the wage compositional data.

The whole paper is based on the interplay between Aitchison's theory of compositional data and the first order approximation of the logratio covariance matrix, interpreted as a multivariate coefficient of variation. If the (univariate) CV is less than 10%, the approximation is good; otherwise, the computed CV overestimates the logratio variance: for a CV=50%, the actual variance would be around 40%. The global CV can be viewed as the square root of the average squared CV for all possible ratios of components. It is also the linearized form of [Eq. (30)], the square root of

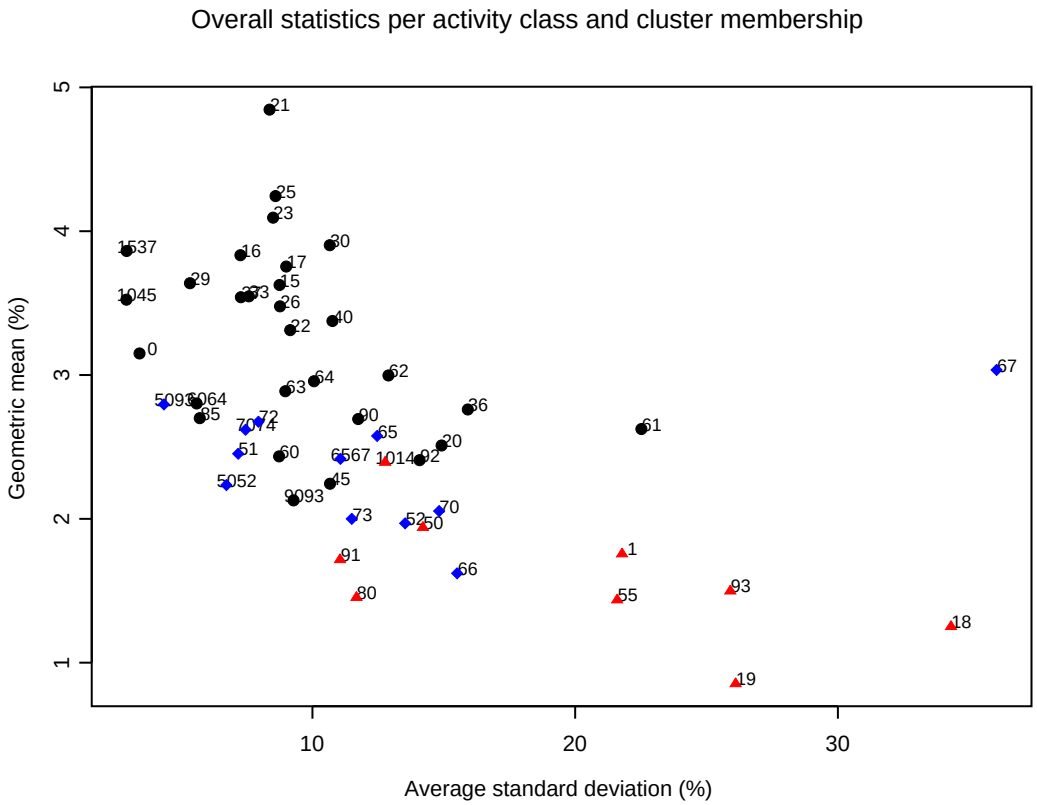

Figure 7: Geometric mean and square root of mean total variation. Clustering is based on a partition computed from the multidimensional scaling components (see Fig. 2).

Aitchison's total variance divided by the degrees of freedom. If the variability of the global estimates of the components is small enough for the linear approximation to be valid, the proposed approach shows a way for generalizing to multivariate compositional data Eurostat's recommendations for communicating precision by CV's. Should the variability be too large, we would suggest that CV's be replaced by the variance of logratios, along the lines given for the analysis of compositional data.

Acknowledgements

The author thanks Sara Keel, from the section "Wages and Working Conditions", for many interesting discussions and her input in the comments for several graphics.

References

[Aitchison, 1986] Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman and Hall, Monographs on Statistics and Probability.

[Aitchison, 1997] Aitchison, J. (1997). The one-hour course in compositional data analysis or compositional data is simple. *Pawlowsky-Glahn, V., Cimne, eds., Proceedings of Int. Assoc. of Mathematical Geology IAMG'97*, Part **I**, pp.3-35.

[Aitchison, 2001] Aitchison, J. (2001). Simplicial Inference. *Comtemporary Mathematics* **287**, AMS.

[Anyadike-Danes, 2003] Anyadike-Danes, M. (2003). The allometry of non-employment. What can compositional data analysis tell us about labour market performance across the UK's regions?, *Thió-Henestrosa, S. and Martín-Fernández, JA (Eds.), Proceedings of CODAWORK'03.*

[Brundson and Smith, 1998] Brunsdon, Teresa M., and Smith, T. M. F. (1998). The time series analysis of compositional data, *Journal of Official Statistics*, **14**, 237-253.

[Larrosa, 2003] Larrosa, J.M. (2003) A compositional statistical analysis of capital stock. *Thió-Henestrosa, S. and Martín-Fernández, JA (Eds.), Proceedings of CODAWORK'03.*

[Eurostat, 2002] Eurostat (2002). *Variance estimation methods in the European Union.* Monographs in official statistics. ISSN 17-25 15-67.

[Eurostat, 2003] Eurostat (2003). *Standard Quality Indicators, Producer-Oriented*, Doc. Eurostat/A4/Quality/03/General/Standard Indicators, Eurostat Working Group "Quality in Statistics" 2003.

[Graf, 2002a] Graf, M. (2002a). Enquête suisse sur la structure des salaires 2000. Plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé. *Rapport de méthode 338-0010*, Office fédéral de la statistique.

[Graf, 2002b] Graf, M. (2002b). Assessing the Accuracy of the Median in a Stratified Double Stage Cluster Sample by means of a Nonparametric Confidence Interval: Application to the Swiss Earnings Structure Survey. *Proceedings of the Joint Statistical Meeting 2002.*

[Graf, 2004] Graf, M. (2004). Enquête suisse sur la structure des salaires 2002. Plan d'échantillonnage et extrapolation pour le secteur privé. *Rapport de méthode 338-0025*, Office fédéral de la statistique, Neuchâtel.

[Särndal and others, 1992] Särndal, C.-E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer Series in Statistics.

[SESS, 2003] SESS (2003). Enquête suisse sur la structure des salaires 2002 (ESS 2002). *Actualités OFS,* **3** Vie active et rémunération du travail.

[SESS, 2004] SESS (2004). L'enquête suisse sur la structure des salaires 2002. Résultats commentés et tableaux. *Statistique de la Suisse*, OFS.

[Silva and Smith, 2001] Silva, D. B. N., and Smith, T. M. F. (2001), Modelling compositional time series from repeated surveys, *Survey Methodology*, **27** (2), 205-215.

Appendix

Table A1 : Wage components in overall wage bill, in %

Private and public sector (Confederation) combined

Switzerland 2002

TA14	Economic activities	Social security contributions		Overtime earnings		Hardship allowances		13th or nth month wage/salary		Special payments bonuses		Global CV
		in %	CV (%)	in %	CV (%)	in %	CV (%)	in %	CV (%)	in %	CV (%)	in %
0	TOTAL	12.9	***0.1***	0.3	***2.2***	0.7	***1.6***	6.3	***0.9***	3.4	***4.4***	***3.4***
01	Horticulture	12.1	*0.6*	0.3	*17.9*	0.2	*27.8*	6.4	*1.6*	0.6	*10.6*	*21.8*
10-45	SECTOR 2 PRODUCTION	13.5	***0.2***	0.5	***2.9***	1.1	***2.2***	7.5	***0.4***	2.1	***2.8***	***2.9***
10-14	Mining and quarrying of stone	14.2	*0.5*	0.4	*8.9*	0.3	*12.3*	8.0	*0.9*	1.4	*13.5*	*12.8*
15-37	Manufacturing	13.3	***0.2***	0.6	***2.6***	1.3	***2.2***	7.5	***0.5***	2.5	***2.9***	***2.9***
15	Manufacture of food products and beverages	13.0	*0.3*	0.8	*10.3*	1.7	*6.0*	7.1	*0.9*	1.1	*7.3*	*8.7*
16	Manufacture of tobacco products	11.7	*0.1*	0.2	*11.0*	2.4	*1.7*	7.3	*0.2*	5.3	*2.9*	*7.3*
17	Manufacture of textiles	13.1	*0.5*	0.4	*9.5*	1.9	*5.1*	6.2	*3.1*	2.4	*8.5*	*9.0*
18	Manufact. of wearing apparel; dressing and dyeing of fur	12.6	*0.8*	0.1	*50.4*	0.1	*17.3*	5.9	*2.0*	1.1	*11.6*	*34.3*
19	Manufacture of leather and leather products	12.1	*1.0*	0.1	*29.3*	0.0	*26.9*	5.6	*8.6*	1.0	*10.7*	*26.1*
20	Manufacture of wood and wood products	13.7	*0.5*	0.4	*9.9*	0.7	*18.7*	7.4	*1.3*	0.7	*10.1*	*14.9*
21	Manufacture of pulp, paper and paper products	13.3	*0.8*	0.9	*10.5*	4.4	*6.0*	7.7	*1.0*	1.8	*5.2*	*8.4*
22	Publishing, printing, reproduction	13.5	*0.5*	0.5	*8.8*	1.4	*10.4*	7.5	*1.3*	1.3	*5.7*	*9.2*
23,24	Manufacture of coke, chemicals	12.1	*1.1*	0.2	*8.8*	2.4	*5.6*	8.9	*2.2*	4.6	*7.1*	*8.5*
25	Manufacture of rubber and plastic products	13.4	*0.6*	0.7	*8.2*	2.5	*5.3*	7.3	*1.2*	1.9	*9.2*	*8.6*
26	Manufacture of other non-metallic mineral products	13.6	*0.6*	0.8	*7.7*	1.0	*6.7*	7.8	*0.9*	1.5	*10.9*	*8.8*
27,28	Manufacture of basic metals	13.7	*0.3*	0.8	*5.8*	1.0	*5.4*	7.1	*0.9*	1.6	*7.8*	*7.3*
29,34,35	Manufact. of machinery & eq. N.E.C., -of motor vehicles	13.5	*0.3*	0.7	*4.2*	0.8	*5.8*	7.3	*1.5*	2.5	*4.1*	*5.3*
30-32	Manufact. of electrical equipment, precision machinery	13.8	*0.7*	0.6	*12.1*	1.0	*7.2*	7.6	*0.7*	3.6	*7.5*	*10.7*
33	Manufacture of medical and precision instruments	13.0	*0.6*	0.6	*5.2*	0.6	*7.8*	7.6	*0.9*	3.5	*7.6*	*7.6*
36,37	Manufacturing N.E.C.	13.5	*0.7*	0.4	*23.4*	0.4	*7.4*	7.1	*1.1*	1.6	*10.7*	*15.9*
40,41	Electricity, gas and water supply	11.1	*0.8*	0.3	*14.4*	1.3	*4.0*	7.9	*0.7*	4.0	*10.0*	*10.8*
45	Construction	14.1	*0.4*	0.4	*12.8*	0.4	*7.6*	7.3	*0.8*	0.9	*8.9*	*10.7*
50-93	SECTOR 3 SERVICES	12.6	***0.2***	0.2	***3.2***	0.5	***2.2***	5.6	***1.4***	4.0	***5.4***	***4.4***
50-52	Sale, repair	12.2	***0.4***	0.2	***5.3***	0.2	***7.2***	6.0	***1.7***	3.1	***4.8***	***6.7***
50	Sale, repair of motor vehicles	12.6	*0.5*	0.3	*13.7*	0.1	*17.1*	6.7	*1.2*	1.6	*6.2*	*14.2*
51	Brokerage, wholesale trade	12.5	*0.3*	0.2	*5.4*	0.2	*8.1*	6.4	*1.0*	4.5	*5.8*	*7.2*
52	Retail trade, repair of pers. & household goods	11.9	*0.7*	0.1	*10.5*	0.3	*11.9*	5.5	*4.2*	2.2	*11.5*	*13.5*
55	Hotels and restaurants	11.5	*0.7*	0.2	*28.9*	0.2	*15.5*	4.4	*2.2*	0.6	*10.0*	*21.6*
60-64	Transport, storage and communication	13.4	***0.3***	0.3	***3.5***	1.0	***4.7***	6.8	***1.5***	1.2	***6.4***	***5.6***
60	Land transport/pipelines	14.2	*0.2*	0.3	*6.3*	0.9	*7.7*	7.3	*0.6*	0.6	*9.2*	*8.7*
61	Water transport	12.7	*1.9*	0.2	*16.1*	1.3	*5.5*	6.5	*9.7*	1.1	*32.8*	*22.5*
62	Air transport	12.3	*0.7*	0.2	*16.7*	1.3	*8.6*	5.3	*1.8*	2.5	*8.5*	*12.9*
63	Supporting and auxiliary transport activities	12.6	*0.5*	0.3	*9.4*	1.1	*9.2*	6.1	*5.8*	1.6	*8.2*	*9.0*
64	Post and telecommunications	13.2	*0.4*	0.3	*3.2*	1.0	*8.4*	6.9	*1.8*	1.6	*11.7*	*10.1*
65-67	Banking; insurance	12.5	***0.7***	0.2	***10.8***	0.2	***9.0***	3.6	***7.7***	11.7	***7.5***	***11.1***
65	Banking	13.0	*0.5*	0.2	*9.1*	0.3	*8.9*	3.0	*11.5*	14.2	*8.1*	*12.5*
66	Insurance	11.8	*1.5*	0.1	*15.8*	0.1	*16.0*	4.8	*7.3*	5.6	*8.8*	*15.5*
67	Activities relating to banking/insurance	11.9	*0.9*	0.5	*49.9*	0.2	*24.6*	4.3	*6.9*	16.6	*9.8*	*36.0*
70-74	Real estate, computer, research & development	12.3	***0.3***	0.3	***8.0***	0.3	***8.2***	5.8	***1.1***	5.2	***3.7***	***7.5***
70,71	Real estate activities/renting of machinery & equipment	12.2	*0.7*	0.1	*11.7*	0.2	*18.6*	6.5	*3.1*	2.6	*6.4*	*14.8*
72,74	Computer and related activities; other business activities	12.3	*0.3*	0.3	*8.4*	0.3	*8.8*	5.7	*1.2*	5.4	*3.9*	*7.9*
73	Research and development	12.5	*0.5*	0.1	*12.3*	0.2	*8.3*	6.6	*3.3*	4.7	*9.8*	*11.5*
75	Public administration, national defence; social security	14.9	*1)*	0.2	*1)*	0.5	*1)*	7.8	*1)*	0.5	*1)*	
80	Education	13.2	*0.4*	0.2	*12.1*	0.1	*10.4*	5.2	*1.7*	0.7	*9.0*	*11.7*
85	Health and social work	13.4	*0.3*	0.3	*6.5*	2.0	*1.9*	6.8	*1.1*	0.5	*5.4*	*5.7*
90-93	Other community, social and personal service activities	12.4	***0.3***	0.2	***9.0***	0.4	***9.5***	5.1	***1.7***	1.1	***5.7***	***9.3***
90	Waste processing and other disposal	13.5	*0.9*	0.3	*7.5*	1.2	*10.9*	6.6	*2.3*	1.1	*13.1*	*11.7*
91	Activities of membership organizations n.e.c.	12.8	*0.3*	0.1	*11.3*	0.4	*11.3*	6.1	*1.4*	0.9	*7.5*	*11.1*
92	Recreational, cultural and sporting activities	12.1	*0.5*	0.3	*16.2*	0.5	*11.0*	5.3	*2.4*	1.3	*8.5*	*14.1*
93	Other service activities	11.9	*0.8*	0.3	*15.1*	0.1	*35.0*	2.6	*9.9*	1.0	*17.5*	*25.9*

Wage bill : Total of non-standardised gross monthly salary

Non-standardised gross salary : Gross salary in the month of October (incl. employee social insurance contributions, benefits in kind, regularly paid shares in bonuses, turnover or commission), but without any overtime pay, hardship allowances (for shift, night and Sunday work), 13th month salary and annual special payments.

CV: coefficient of variation; 1): not computed, because the results are not based on a random sample.

Global CV: Linearized form of the average standard deviation, that is square root of average total linearized variance of logratios of components (see text).

Source: Swiss Federal Statistical Office, Swiss Earnings Structure Survey (SESS) 2002

Original table: wage components only.

Table A2: p-values for the 2 by 2 equality tests for social security contributions

1) Sector 2 + horticulture

Noga2	40	16	23	19	1	18	33	15	17	21	25	36	29	26	22	20	27	30	45	1014
40	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
16	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
23	.	.	0	57	62	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
19	.	.	57	0	54	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1	.	.	62	54	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
18	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
33	.	.	.	.	.	.	0	50	90	.	.	.	.	.	.	.	.	.	.	.
15	.	.	.	.	.	.	50	0	94	.	.	.	.	.	.	.	.	.	.	.
17	.	.	.	.	.	.	90	94	0	95	.	.	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.	95	0	84	90	96	96	99	.	.	.	.	.
25	.	.	.	.	.	.	.	.	.	84	0	67	78	83	94	99	.	.	.	.
36	.	.	.	.	.	.	.	.	.	90	67	0	57	67	83	96	99	99	.	.
29	.	.	.	.	.	.	.	.	.	96	78	57	0	65	88	99	.	.	.	.
26	.	.	.	.	.	.	.	.	.	96	83	67	65	0	72	92	98	97	.	.
22	.	.	.	.	.	.	.	.	.	99	94	83	88	72	0	81	92	93	.	.
20	.	.	.	.	.	.	.	.	.	.	99	96	99	92	81	0	63	75	.	.
27	.	.	.	.	.	.	.	.	.	.	.	99	.	98	92	63	0	71	.	.
30	.	.	.	.	.	.	.	.	.	.	.	99	.	97	93	75	71	0	.	.
45	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	87
1014	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	87	0

2) Sector 3

Noga2	55	66	67	93	52	92	70	62	72	51	63	73	50	61	91	65	64	80	85	90	60
55	0	85	98	99	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
66	85	0	63	71	75	95	98	99	99	.	.	.	.	.	.	.	.	.	.	.	.
67	98	63	0	61	69	98	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
93	99	71	61	0	58	97	99	.	.	.	.	.	.	.	.	.	.	.	.	.	.
52	.	75	69	58	0	96	99	.	.	.	.	.	.	.	.	.	.	.	.	.	.
92	.	95	98	97	96	0	85	91	99	.	.	.	.	99	.	.	.	.	.	.	.
70	.	98	.	99	99	85	0	63	71	99	.	.	.	96	.	.	.	.	.	.	.
62	.	99	.	.	.	91	63	0	55	97	99	.	.	95	.	.	.	.	.	.	.
72	.	99	.	.	.	99	71	55	0	.	.	.	.	95	.	.	.	.	.	.	.
51	.	.	.	.	.	.	99	97	.	0	89	95	97	83	.	.	.	.	.	.	.
63	.	.	.	.	.	.	.	99	.	89	0	59	66	70	.	.	.	.	.	.	.
73	.	.	.	.	.	.	.	.	.	95	59	0	57	67	.	.	.	.	.	.	.
50	.	.	.	.	.	.	.	.	.	97	66	57	0	65	.	.	.	.	.	.	.
61	.	.	.	.	.	99	96	95	95	83	70	67	65	0	72	88	96	99	.	.	.
91	.	.	.	.	.	.	.	.	.	.	.	.	.	72	0	99	.	.	.	.	.
65	.	.	.	.	.	.	.	.	.	.	.	.	.	88	99	0	96	.	.	.	.
64	.	.	.	.	.	.	.	.	.	.	.	.	.	96	.	96	0	94	.	.	.
80	.	.	.	.	.	.	.	.	.	.	.	.	.	99	.	.	94	0	.	98	.
85	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	71	.
90	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	98	71	0	.
60	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0

Less than 97.5% p-values show non-significant differences.
Larger than 99.5% % p-values are replaced by dots
Noga2 are ordered by increasing social security contributions.

Table A3: Correlation matrices for composants 2 to 5 (total, horticulture and secondary sector)

1
0
1.00 0.14 0.00 0.02
0.14 1.00 0.18 -0.21
0.00 0.18 1.00 -0.74
0.02 -0.21 -0.74 1.00
2
1
1.00 0.16 0.01 0.01
0.16 1.00 0.28 -0.14
0.01 0.28 1.00 -0.18
0.01 -0.14 -0.18 1.00
3
1045
1.00 0.05 -0.04 -0.07
0.05 1.00 0.09 -0.05
-0.04 0.09 1.00 0.09
-0.07 -0.05 0.09 1.00
4
1014
1.00 0.08 -0.04 -0.08
0.08 1.00 -0.03 0.07
-0.04 -0.03 1.00 0.08
-0.08 0.07 0.08 1.00
5
1537
1.00 0.00 -0.10 -0.16
0.00 1.00 0.09 -0.16
-0.10 0.09 1.00 0.08
-0.16 -0.16 0.08 1.00
6
15
1.00 0.21 -0.06 -0.04
0.21 1.00 0.35 0.01
-0.06 0.35 1.00 -0.30
-0.04 0.01 -0.30 1.00
7
16
1.00 0.47 -0.04 -0.11
0.47 1.00 -0.36 -0.37
-0.04 -0.36 1.00 0.34
-0.11 -0.37 0.34 1.00
8
17
1.00 0.12 -0.04 -0.02
0.12 1.00 -0.09 0.11
-0.04 -0.09 1.00 -0.70
-0.02 0.11 -0.70 1.00
9
18
1.00 0.00 -0.11 0.00
0.00 1.00 0.35 0.35
-0.11 0.35 1.00 0.08
0.00 0.35 0.08 1.00
10
19
1.00 0.09 0.16 0.01
0.09 1.00 0.20 -0.04
0.16 0.20 1.00 -0.23
0.01 -0.04 -0.23 1.00
11
20
1.00 0.11 0.09 -0.08
0.11 1.00 0.15 -0.11
0.09 0.15 1.00 -0.36
-0.08 -0.11 -0.36 1.00
12
21
1.00 -0.05 0.11 0.10
-0.05 1.00 0.31 -0.20
0.11 0.31 1.00 -0.12
0.10 -0.20 -0.12 1.00
13
22
1.00 0.40 -0.04 -0.19
0.40 1.00 -0.09 -0.10
-0.04 -0.09 1.00 0.23
-0.19 -0.10 0.23 1.00
14
23
1.00 0.18 -0.17 -0.22
0.18 1.00 -0.09 -0.47
-0.17 -0.09 1.00 -0.37
-0.22 -0.47 -0.37 1.00
15
25
1.00 -0.17 -0.03 0.02
-0.17 1.00 0.03 0.00
-0.03 0.03 1.00 -0.03
0.02 0.00 -0.03 1.00
16
26
1.00 0.25 -0.08 0.24
0.25 1.00 -0.14 0.46
-0.08 -0.14 1.00 -0.15
0.24 0.46 -0.15 1.00
17
27
1.00 0.04 0.02 -0.19
0.04 1.00 0.05 -0.16
0.02 0.05 1.00 0.14
-0.19 -0.16 0.14 1.00
18
29
1.00 -0.09 -0.07 -0.05
-0.09 1.00 0.13 0.01
-0.07 0.13 1.00 0.11
-0.05 0.01 0.11 1.00
19
30
1.00 -0.10 0.30 -0.33
-0.10 1.00 -0.22 -0.40
0.30 -0.22 1.00 0.26
-0.33 -0.40 0.26 1.00
20
33
1.00 -0.07 0.09 -0.09
-0.07 1.00 0.37 0.07
0.09 0.37 1.00 0.36
-0.09 0.07 0.36 1.00
21
36
1.00 0.07 -0.35 0.71
0.07 1.00 0.19 -0.03
-0.35 0.19 1.00 -0.48
0.71 -0.03 -0.48 1.00
22
40
1.00 0.39 0.10 0.26
0.39 1.00 -0.05 0.07
0.10 -0.05 1.00 0.44
0.26 0.07 0.44 1.00
23
45
1.00 0.15 0.06 0.07
0.15 1.00 -0.08 0.11
0.06 -0.08 1.00 -0.02
0.07 0.11 -0.02 1.00

Explanations :
Matrices are represented on 3 columns.

By column :
1st line : sequential number

2nd line : NOGA2 class number

3rd à 6th lines : correlation matrices of (x_1, x_2, x_3, x_4).

Table A3 (continued): Correlation matrices for composants 2 to 5 (third sector)

24
5093
1.00 0.12 -0.16 0.20
0.12 1.00 0.06 -0.15
-0.16 0.06 1.00 -0.74
0.20 -0.15 -0.74 1.00
25
5052
1.00 0.08 -0.18 -0.21
0.08 1.00 -0.42 -0.16
-0.18 -0.42 1.00 0.02
-0.21 -0.16 0.02 1.00
26
50
1.00 0.10 -0.03 0.03
0.10 1.00 -0.20 0.12
-0.03 -0.20 1.00 -0.03
0.03 0.12 -0.03 1.00
27
51
1.00 0.11 -0.02 -0.09
0.11 1.00 0.13 -0.06
-0.02 0.13 1.00 0.12
-0.09 -0.06 0.12 1.00
28
52
1.00 0.09 -0.22 -0.45
0.09 1.00 -0.53 -0.32
-0.22 -0.53 1.00 0.24
-0.45 -0.32 0.24 1.00
29
55
1.00 0.04 -0.01 0.00
0.04 1.00 0.12 0.00
-0.01 0.12 1.00 0.04
0.00 0.00 0.04 1.00
30
6064
1.00 0.01 0.27 -0.06
0.01 1.00 0.20 -0.14
0.27 0.20 1.00 -0.03
-0.06 -0.14 -0.03 1.00
31
60
1.00 -0.09 -0.20 -0.02
-0.09 1.00 0.20 -0.08
-0.20 0.20 1.00 -0.02
-0.02 -0.08 -0.02 1.00
32
61
1.00 -0.27 -0.23 0.93
-0.27 1.00 0.48 -0.23
-0.23 0.48 1.00 -0.10
0.93 -0.23 -0.10 1.00

33
62
1.00 0.23 0.02 -0.10
0.23 1.00 -0.06 0.17
0.02 -0.06 1.00 -0.26
-0.10 0.17 -0.26 1.00
34
63
1.00 0.35 0.35 0.24
0.35 1.00 0.67 0.15
0.35 0.67 1.00 0.66
0.24 0.15 0.66 1.00
35
64
1.00 -0.26 0.52 -0.33
-0.26 1.00 -0.30 -0.33
0.52 -0.30 1.00 -0.46
-0.33 -0.33 -0.46 1.00
36
6567
1.00 0.53 -0.43 0.55
0.53 1.00 -0.77 0.70
-0.43 -0.77 1.00 -0.67
0.55 0.70 -0.67 1.00
37
65
1.00 0.65 -0.62 0.69
0.65 1.00 -0.86 0.64
-0.62 -0.86 1.00 -0.58
0.69 0.64 -0.58 1.00
38
66
1.00 0.31 0.47 -0.20
0.31 1.00 0.24 -0.19
0.47 0.24 1.00 -0.54
-0.20 -0.19 -0.54 1.00
39
67
1.00 -0.15 0.49 -0.19
-0.15 1.00 -0.18 0.61
0.49 -0.18 1.00 -0.31
-0.19 0.61 -0.31 1.00
40
7074
1.00 0.17 0.05 0.03
0.17 1.00 0.11 -0.08
0.05 0.11 1.00 0.08
0.03 -0.08 0.08 1.00
41
70
1.00 -0.13 -0.07 -0.09
-0.13 1.00 0.20 -0.10
-0.07 0.20 1.00 0.26
-0.09 -0.10 0.26 1.00

42
72
1.00 0.17 0.05 0.03
0.17 1.00 0.11 -0.08
0.05 0.11 1.00 0.07
0.03 -0.08 0.07 1.00
43
73
1.00 0.41 0.43 -0.22
0.41 1.00 -0.01 -0.59
0.43 -0.01 1.00 0.30
-0.22 -0.59 0.30 1.00
44
80
1.00 -0.05 0.11 0.04
-0.05 1.00 -0.11 -0.06
0.11 -0.11 1.00 0.19
0.04 -0.06 0.19 1.00
45
85
1.00 -0.01 0.12 -0.16
-0.01 1.00 0.37 -0.17
0.12 0.37 1.00 -0.54
-0.16 -0.17 -0.54 1.00
46
9093
1.00 -0.28 -0.19 0.15
-0.28 1.00 0.63 -0.11
-0.19 0.63 1.00 -0.12
0.15 -0.11 -0.12 1.00
47
90
1.00 -0.05 -0.26 0.02
-0.05 1.00 -0.18 0.30
-0.26 -0.18 1.00 -0.53
0.02 0.30 -0.53 1.00
48
91
1.00 0.24 -0.05 -0.04
0.24 1.00 0.25 -0.14
-0.05 0.25 1.00 0.06
-0.04 -0.14 0.06 1.00
49
92
1.00 -0.39 -0.26 0.2
-0.39 1.00 0.67 -0.3
-0.26 0.67 1.00 -0.2
0.20 -0.30 -0.20 1.0
50
93
1.00 0.26 -0.11 0.23
0.26 1.00 -0.07 0.37
-0.11 -0.07 1.00 -0.20
0.23 0.37 -0.20 1.00