Precision of Compositional Data in a Stratified Two-Stage Cluster Sample: Comparison of the Swiss Earnings Structure Survey 2002 and 2004

Monique Graf
Statistical Methods Unit, Swiss Federal Statistical Office

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 needs 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. For surveys 2002 and 2004, we first investigate the incidence of components and then follow Aitchison's approach for the analysis of compositional data, which is based on logratios of components. Different multivariate analyses of the compositional data are performed and compared between the years, whereby the wage components are broken down by economic activity classes. Then we propose a number of ways to assess precision.

Keywords: Complex survey; compositional data; linearization; confidence domain; coefficient of variation.

1 Introduction

In a design-based framework, the variances cannot be given a subject matter interpretation, because they are influenced by the sampling design. It is exactly the same for correlations between variables. They are nevertheless of importance for the assessment of precision and of change in multivariate data.

A widely used way to assess precision is to release the coefficient of variation (CV). Being dimensionless, it enables easy comparisons of precision among variables with different orders of magnitude. However in the case of multivariate data which are correlated by nature, like parts of a whole, CV's are not enough to assess precision. We seek a generalization of the CV along the lines of multivariate statistics to be applied in the particular context of compositional data. This global CV will thus be related to the matrix norm of the covariance matrix of estimates.

This study takes the principles and methods of compositional data analysis, initiated by John Aitchison more than 20 years ago and applies them within the framework of a complex survey. Despite the fact that these principles and methods now enjoy considerable support from theorists, they have not yet been extensively applied to survey data. Applications of compositional analysis to public statistics have been published by Silva and Smith (2001), Brunsdon and Smith (1998), Larrosa (2003) and Anyadike-Danes (2003), but do not consider the sampling variability.

The case study, taken from the Swiss Earnings Structure Survey (SESS), will provide insight into the precision and variability of wage components using the framework of compositional data analysis as developed by Aitchison (1986). The principle is to compute the logarithm of ratios of the components. The total variance is proportional to the sum of the variances of all possible logratios of components. From the total variance, an average logratio variance can be obtained. It will be shown that the linearized form of this average variance can be interpreted as an average squared CV of all possible ratios of components and thus provides a summary measure of the wage compositional vector. Graf (2005) gives a first analysis of the 2002 data. A detailed comparison of the 2002 and 2004 wage components can be found in Graf (2006).

2 Compositional vectors

Compositional data are observations expressed as parts, thus have a unit sum constraint. A good mathematical summary of the principal notions is provided by Aitchison (2001), a less formal introduction in Aitchison (1997) and a thorough presentation of the classical theory in Aitchison (1986). Basic notions are recalled in Sections 2.1 and 2.2.

2.1 Geometric properties

A unit-sum compositional vector of length D, $(p_1, p_2, ...p_D)$ has strictly positive components that sum up to 1. The set of these vectors is the simplex S^D. A vector $\mathbf{w}$ with positive coordinates is made compositional by the closure operation, which means dividing each coordinate by their sum: $\mathbf{p} = \text{clo}(\mathbf{w}) = \mathbf{w}/\sum w_i$.

The unit sum constraint implies that there is necessarily a negative correlation between the components. This shows that the crude correlations are not directly interpretable. To release this constraint, Aitchison proposes different alternative transformations that generalize the logistic transformation $\ln(p/(1-p))$ for a 2-part compositional vector. Here for practical reasons related to the form of the published tables for the present survey, we use the *additive logratio transform*

$$\text{alr}(\mathbf{p}) = (\ln(p_1/p_D), \ln(p_2/p_D), ..., \ln(p_d/p_D)) \qquad (1)$$

where $d = D - 1$. Applying this transformation, the resulting vector is no longer constrained and correlations between components can be interpreted. The logistic transformation is exactly the alr for $D = 2$.

The vector space structure in the alr transform induces a vector space on the simplex. There are two basic operations in a vector space: the addition of two vectors and the multiplication of a vector by a scalar. In the case of compositions, these classical operations have different names. Perturbation corresponds to the vector addition in the alr representation:

$$\mathbf{a} \oplus \mathbf{p} = \mathrm{clo}(a_1 p_1, a_2 p_2, ..., a_D p_D) \tag{2}$$

Powering corresponds to the scalar multiplication in both representations:

$$b \odot \mathbf{p} = \mathrm{clo}(p_1^b, p_2^b, ...p_D^b) \tag{3}$$

Graphical effects of perturbation can be found in von Eymatten et al (2002).

2.2 Statistics on compositions

The centre of the distribution is given by the geometric - and not the arithmetic - mean of the compositions. Its theoretical counterpart (expressed with the help of the expectation of the alr coordinates) is:

$$\mathrm{cen}(\mathbf{p}) = \mathrm{clo}\exp\left(\mathrm{E}\left[\mathrm{alr}(\mathbf{p})\right]\right)$$

Consider the vector of ratios of the $d = D - 1$ first components to the last, that is

$$\mathbf{x} = (x_1, ..., x_d) = (p_1, ..., p_d)/p_D = \mathbf{p}_{-D}/p_D \tag{4}$$

Let us denote the $d \times d$ - covariance matrix of the logratios by:

$$\mathrm{Cov}(\mathrm{alr}(\mathbf{p})) = \Sigma = [\sigma_{ij}] \tag{5}$$

Under regularity conditions, $\mathbf{y} = (\ln x_1, \ln x_2, ..., \ln x_d)'$ is asymptotically normally distributed $N^d(\boldsymbol{\mu}, \Sigma)$ with Σ given by Equation (5).

Under the asymptotic distribution hypothesis and considering Σ as fixed,

- the confidence domain for $\mathbf{y}$, $\mathbf{D}_{1-\alpha}(\mathbf{y})$ 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

$$\mathbf{D'}_{1-\alpha}(\mathbf{y}) = \left\{\mathbf{y} \in \mathbb{R}^d \mid (\mathbf{y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu}) \le \chi^2_{d;1-\alpha}\right\}$$

- the corresponding domain for $\mathbf{p} = (p_1, ..., p_D)$ is a subset of the simplex S^D:

$$\mathbf{D}_{1-\alpha}(\mathbf{p}) = \{\ \mathbf{p} \in S^D \mid \left(\ln \frac{\mathbf{p}_{-D}}{p_D} - \boldsymbol{\mu}\right)' \Sigma^{-1} \left(\ln \frac{\mathbf{p}_{-D}}{p_D} - \boldsymbol{\mu}\right) \le \chi^2_{d;1-\alpha}\ \} \tag{6}$$

Aitchison defines (among other measures) the total variance for which different equivalent formulations exist (Aitchison 1986, Chapter 4), among them:

$$\begin{aligned} \mathrm{totvar}(\mathbf{p}) &= \frac{1}{D} \sum_{i<j} \mathrm{Var}\left(\ln \frac{p_i}{p_j}\right) & (7) \\ &= \mathrm{tr}(\Sigma) - \frac{1}{D} \mathbf{1}'_d \Sigma \mathbf{1}_d & (8) \end{aligned}$$

2.3 Proposition for a global measure of precision

Whereas a thorough discussion of precision must take into account the multivariate nature of compositions, it is also of importance to derive a simple summary measure in order to characterize the overall precision of a composition. The drawback of the total variance is its dependence on the dimension D. Taking the formulation in Equation (7), we define an average standard deviation for logratios by

$$\mathrm{stot}(\mathbf{p}) = \sqrt{\frac{\mathrm{totvar}(\mathbf{p})}{(D-1)/2}} = \sqrt{\frac{\sum\limits_{i<j} \mathrm{Var}\left(\ln \frac{p_i}{p_j}\right)}{D\,(D-1)\,/2}} \tag{9}$$

If the logratios variability is small, we can approximate the logratio variance by its first order linearized form.

$$\mathrm{Var}\left(\ln \frac{p_i}{p_D}\right) \approx \mathrm{CV}^2\left(\frac{p_i}{p_D}\right) = \frac{\mathrm{Var}\,(p_i/p_D)}{(p_i/p_D)^2} \doteq \frac{\widetilde{\sigma}_{ii}}{(p_i/p_D)^2} \tag{10}$$

Let Σ_{ij} be the 2×2-covariance matrix of $(\ln(p_i/p_D), \ln(p_j/p_D))'$ and $\widetilde{\Sigma}_{ij}$ be the corresponding matrix for $(p_i/p_D, p_j/p_D)'$. For other ratios of parts, the same approximation yields in matrix form:

$$\mathrm{Var}\left(\ln \frac{p_i}{p_j}\right) = \begin{pmatrix} 1 & -1 \end{pmatrix} \Sigma_{ij} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \cong \mathrm{CV}^2\left(\frac{p_i}{p_D}\right) \tag{11}$$

because by linearization the quadratic form in Equation (11) can be approximated by

$$\begin{pmatrix} (p_i/p_D)^{-1} & -(p_j/p_D)^{-1} \end{pmatrix} \widetilde{\Sigma}_{ij} \begin{pmatrix} (p_i/p_D)^{-1} \\ -(p_j/p_D)^{-1} \end{pmatrix}$$

Substituting approximations (10) and (11) into Equation (9), we can interpret stot($\mathbf{p}$) as an approximate L_2-average of the CV's of all possible ratios of components (i.e. the square root of the mean squared CV's). We call it *global CV*.

Practically, Σ is computed using the expression of the total variance in Equation (8) and the linearized form of Equation (5), that is

$$\Sigma \cong \left[\mathrm{diag}\left(\mathbf{p}_{-D}/p_D\right)\right]^{-1} \widetilde{\Sigma} \left[\mathrm{diag}\left(\mathbf{p}_{-D}/p_D\right)^{-1}\right] \tag{12}$$

Thus we only need to evaluate matrix $\widetilde{\Sigma}$, even if our global measure is interpreted using the form in Equation (7).

3 Compositional analysis of wage components

The Swiss earnings structure survey (SESS) is a biennial written survey sent out to businesses. The survey is constructed on a stratified two-stage sampling scheme (Graf 2004). The 2002 and 2004 samples are rather large: 1/3 of all businesses in Switzerland are involved. 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 et al (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 and 2004 surveys can be found in SESS(2002, 2003 and 2004).

The sampling frame is the business register 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. Class 0 represents the total of all activities considered. The survey design is a stratified two-stage 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 non response is assumed to be ignorable at the stratum level. 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. For a fuller description of the sampling design and extrapolation, see Graf (2004, 2006).

We concentrate on the decomposition of wage into five components (overtime earnings, hardship allowances, 13th or n-th salary, bonuses and non-standardized gross earnings), see SESS (2002, 2003). The “non standardized total monthly salary” is the sum of the five components. The defined components are summarized in Table 1. The wage percentage attributed to each component are computed relative to the fifth component, and not to their sum. They are reproduced here for the main economic activity groupings in 2002 and 2004 (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. Thus the published proportions are, with $D = 5$:

$$\mathbf{x} = \mathbf{p}_{-D}/p_D = (s_1, s_2, s_3, s_4)/s_5 \tag{13}$$

It is stressed that in this framework, the interest is not in the wage composition at the individual level, but in the global composition for 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 Incidence of components

Before the precision is computed, it is important to get a rough idea of the data. The only component that is always present is the gross earnings (component 5). The wage percentages in Table A1 have a different meaning, whether the corresponding mass is distributed over all or just a small number of ultimate units. In order to investigate this point, the average incidence of each components was computed (Table 2). It can be interpreted as a probability that a randomly chosen unit is getting a non-zero amount for that component. So an estimated proportion of 75-76% of wages encompasses a 13th salary (component 3). The probability of obtaining the components ”overtime earnings” and ”bonuses” defined in Table 1 is doubled, if the 13th salary is present. On the contrary, the probability of having the ”bonuses” share given the presence of ”13th salary” is slightly smaller.

Table 1: Wage mass attributed to the different components.

Component	Part p_i	Ratio	N.R.
Overtime earnings	p_1	x_1	0.3
Hardship allowances	p_2	x_2	0.7
13th month salary (/12)	p_3	x_3	6.3
Bonuses (/12)	p_4	x_4	3.4
Non-stand. gross earnings incl. social contributions	p_5	1	100
	$\sum p_i = 1$		

$x_i = p_i/p_5, i = 1, ..., 4$
N.R. = national ratio(%)

Table 2: Weighted proportion of wages with the considered component: global (first 2 lines) and given presence/absence of ”13th salary” (4 last lines). Numbering of components as in Table 1.

	Component			
year	3	1	2	4
2002	0.753	0.031	0.106	0.258
2004	0.760	0.041	0.113	0.255
	given 3			
2002	0	0.018	0.063	0.295
2002	1	0.035	0.120	0.246
2004	0	0.016	0.051	0.273
2004	1	0.049	0.132	0.249

3.2 Multivariate analyzes of the estimated components

A multidimensional scaling on logratio estimates $\ln(\mathbf{p}_{-D}/p_D)$ with $D = 5$ (alr scale, see Section 1) was performed with the S-plus procedure CMDSCALE on the 2002 data, using as distance between 2 economic activities the Euclidian distance between the corresponding alr transformed vectors, see Equation (1). (The same result would be obtained by principal component analysis).

For each year the left pane in Figure 1 represents the projection onto the first two principal axes[1] computed

[1] The usual terminology is “principal components”; the expres-

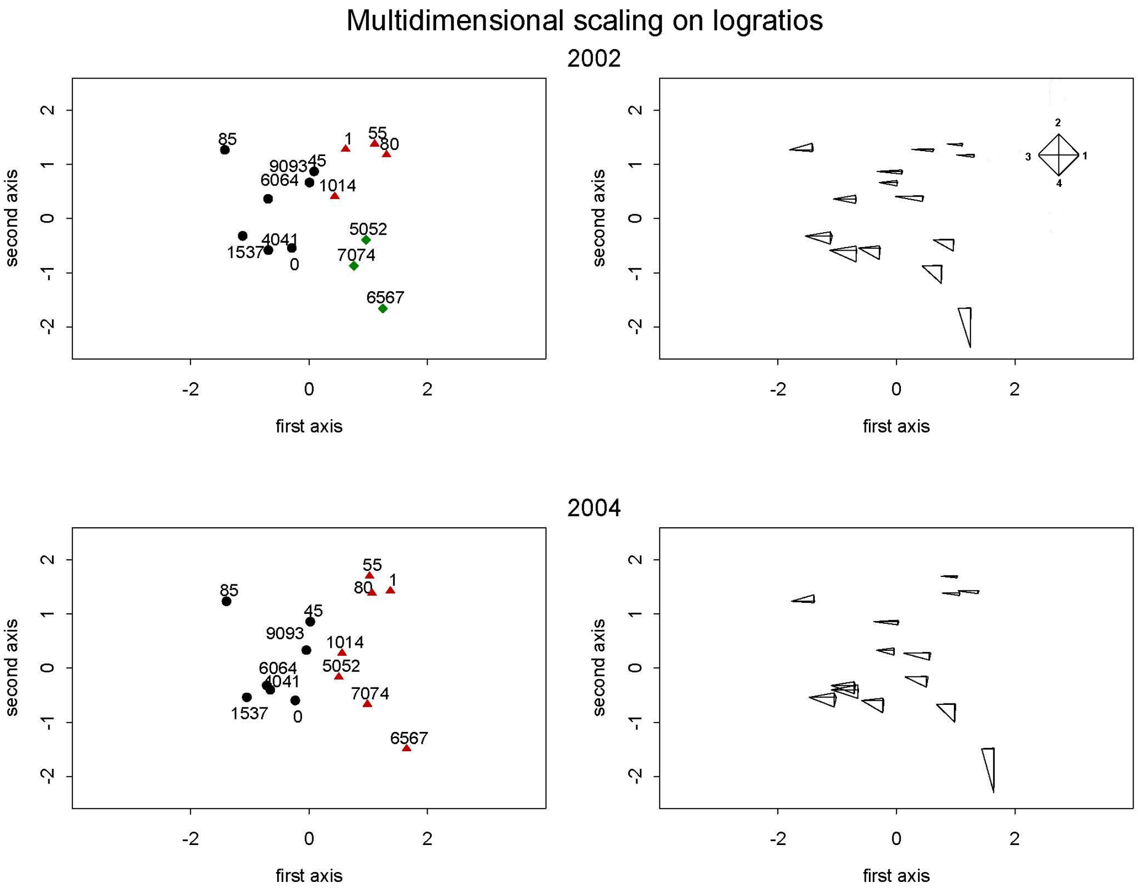

Figure 1: Multivariate representations of the 5-part composition for activity aggregates (2002: top, 2004: bottom). Left panes: Multidimensional scaling on the alr representation. Circles, triangles and diamonds give the group membership computed by PAM; numbers denote the NACE aggregates. Right panes: corresponding star plot.

for the 2 years separately. These planes explain 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. The activity aggregates are coded by their NACE2 code (see Table A1), whereas in the right panes, they are represented by a star plot for which the half diagonals of the quadrilateral are proportional to the components of $\mathrm{alr}(\mathbf{p}) = \ln(\mathbf{p}_{-D}/p_D)$.

A partition was also performed (using the 4 new coordinates) by the S-plus procedure PAM (partition around medoids) and an optimal number of 3 groups was obtained in 2002, whereas two of the former groups merged in 2004. The groups are visible on the left panes (circles, triangles and diamonds). For an interpretation of the groups, see Graf (2006).

The use of the alr representation for the multidimensional scaling is justified by the form of the published table (Table A1: relative values of the other components to the gross salary). The gross salary indeed plays a special role, because it is the only component that is always present (see preceding paragraph).

3.3 Precision

The variance-covariance matrix of the estimates is based on the sampling distribution of the wage components. The large sample size implies that finite population corrections (fpc) are indispensable for realistic estimates of the precision of the population values. The variance estimation method applied here relies on the linearization of the estimators as given in Equation (12) and adapted to the case of a stratified SI-SI design. Within a stratum, we linearize the variance $\widetilde{\sigma}_{ii}$ of the ratio $s_i/s_5, i = 1, ..., 4$, in Table 1, according to the formula in Särndal et al (1992) p. 180, and sum over the strata defining an economic activity aggregate. We neglect the variability in the extrapolation weights: a refinement would be to use Särndal and Lundström's approach (2005), with InfoU at PSU level (the number of businesses in the stratum is known from the register) and InfoS at the SSU level (we only use

sion "principal axes" is being used instead, in order to avoid confusion with the salary components.

the information on the total number of wages as given in the questionnaire). Because this auxiliary information is minimal and in view of their comment on top of p. 138, it won't probably change the results much.

Once the approximate covariance matrices of the logratios of the wage components to the gross salary are obtained, we are in position to assess the accuracy of the population composition estimates.

Recall that the coefficient of variation of a ratio is an estimate of the average logratio standard deviation and return to the summary measure defined in Section 2.3. In this application, the global CV is always between the extremes of the 4 corresponding univariate CV's (Table A1). It can be seen that the global CV has a tendency to be larger for a smaller geometric mean of the 5 parts.

3.4 Simplex view of confidence domains

In order to get a feeling of what a global CV means, let us look at the full compositional vector for which the 5 components in Table 1 are divided by the total salary. To visualize the 95% confidence domains given in Equation (6), let us split the 5-part composition into two 3-part compositions: an amalgamation $(p_1 + p_2 + p_4, p_3, p_5)$ where the 3 generally smallest parts 1, 2 and 4 are added, and a subcomposition $\mathrm{clo}(p_1, p_2, p_4)$[2]. Both are unit-sum compositional vectors. Because our approximation is linear, it is easy to deduce the corresponding covariance matrices from $\widetilde{\Sigma}$. It is interesting to note that the breaking down of the 5-dimensional composition into the above amalgamation and subcomposition is sufficient for the recovery of the original compositions, but not for the full original covariance matrix. For instance the elements corresponding to parts 1 and 3 cannot be recovered from the covariance matrices of the provided amalgamation and subcomposition.

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. Figure 2 shows the 95% confidence domains for the main economic activity aggregates in 2002 and 2004.

The amalgamations (left panes) are very precisely estimated, the worst is for "banking, insurance" 65-67 for which the uncertainty is essentially in the demarcation between the amalgamation of part 5 "gross earnings" and part 3 "13th salary" and the sum of the others (1+2+4). Apart for this group, (1+2+4) is never larger than 6%.

The right panes show how the small amount of (1+2+4)is distributed among the 3 components. In general, the subcompositions have a minimum of 50% share in "bonuses" (4) and very little parts of "overtime earnings" (1) and "hardship allowances" (2), except grouping 85 "health and social work" which shows narrowly 20% of bonuses (4) but more than 80% of hardship allowances (2). The subcompositions are fairly precisely estimated, as can be seen by the small 95% confidence domains.

3.5 Multivariate test for change

Let us denote by $\ln(\mathbf{x}_A)$, $\ln(\mathbf{x}_B)$, $\mathit{\Sigma}_A$ and $\mathit{\Sigma}_B$ the estimated alr compositional vectors (in logarithmic scale) and their covariance matrices of a given activity grouping for 2 independent samples A and B respectively. By the independence between the samples, the covariance matrix of the difference between the compositional vectors (in logarithmic scale) is given by the sum $\mathit{\Sigma}_A + \mathit{\Sigma}_B$. Under the null hypothesis of no change, the Mahalanobis squared distance M in Equation (14) follows a chi-square distribution with 4 degrees of freedom:

$$M = (\ln(\mathbf{x}_B) - \ln(\mathbf{x}_A))^t (\mathit{\Sigma}_A + \mathit{\Sigma}_B)^{-1} (\ln(\mathbf{x}_B) - \ln(\mathbf{x}_A)) \tag{14}$$

If $M > \chi^2_{4;1-\alpha/2}$, the null hypothesis is rejected at the α-level.

The last two columns in Table A1 show the statistic M and the corresponding p-value, when the samples A and B correspond to the 2002 and 2004 data respectively. Choosing a 5% risk, one can see that the change in the compositional vectors is significant, except for two groupings, namely "Construction" and "Education".

4 Discussion

The whole study 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. 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 Equation (9), which is proportional to the square root of Aitchison's total variance divided by the degrees of freedom.

If the (univariate) CV is less than 0.3, the approximation is good; otherwise, the computed squared CV overestimates the logratio variance: assuming a lognormal distribution for a ratio of parts, if $\mathrm{CV} \doteq \sqrt{\exp(\sigma^2) - 1} = 0.50$, for example, the square root of the actual logratio variance would be around $\sigma = 0.47$. 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.

References

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. London and New York: Chapman and Hall, Monographs on Statistics and Probability.
Reedited by Caldwell: The Blackburn Press (2003).

Aitchison, J. (1997). The one-hour course in compositional data analysis or compositional data is simple. In Proceedings of the International Association of Mathematical Geology IAMG'97, Part I, pp.3-35, ed. Pawlowsky-Glahn,V.

Aitchison, J.(2001).Simplicial Inference. Comtemporary Mathematics, American Mathematical Society, 287.

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? In Proceedings of CODAWORK'03, ed. Thió-Henestrosa, S. and Martín-Fernández, JA.

[2] Aitchison (1986) gives a thorough description of amalgamation and subcomposition.

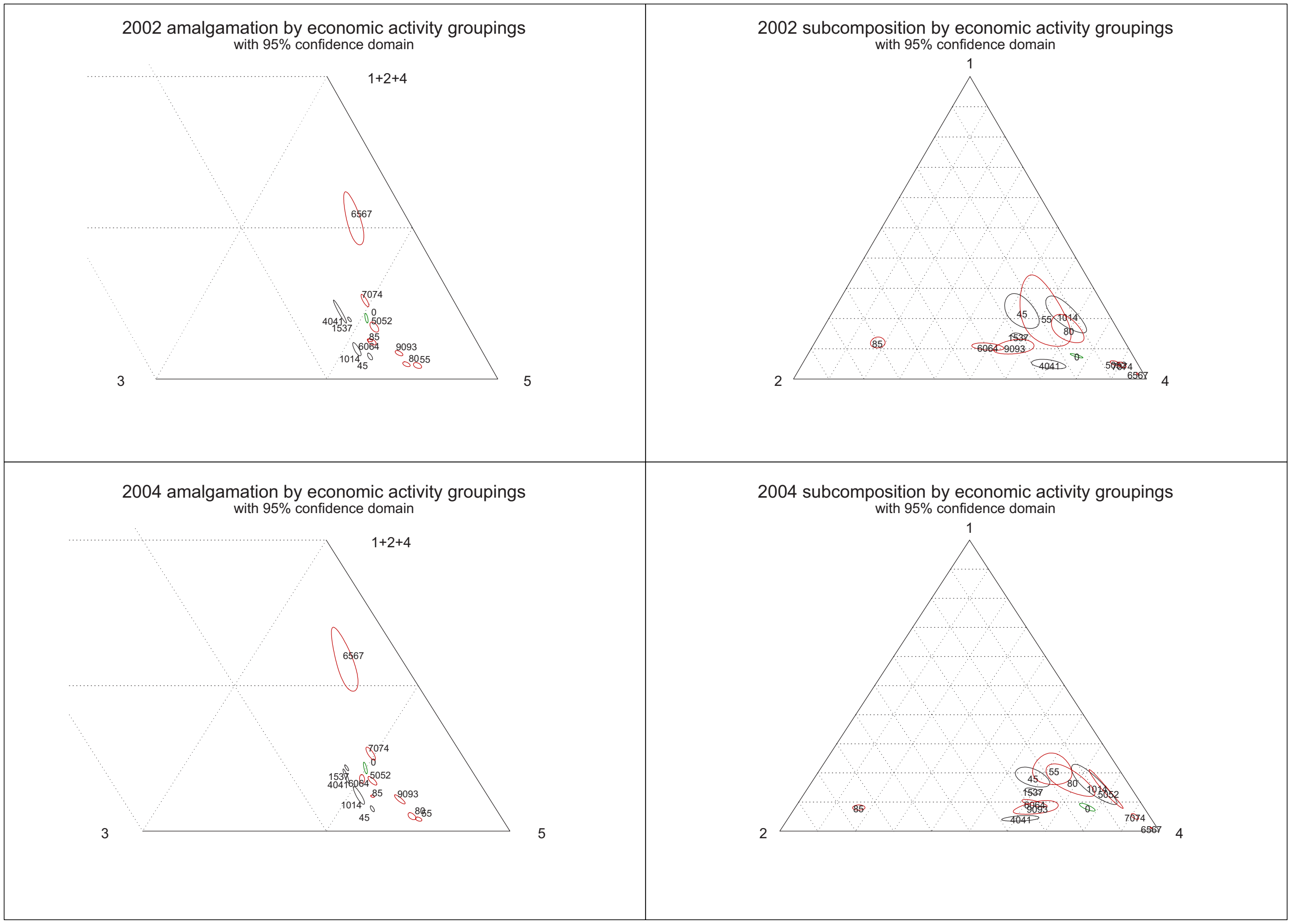

Figure 2: Representation within the simplex. Top: 2002; bottom: 2004 data. Left: amalgamations (zoom on the edge corresponding to part 5, gross earnings); right: subcompositions. Dotted lines are 10% apart.

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

Egozcue, J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003). Isometric Logratio Transformations for Compositional Data Analysis. Mathematical Geology, 35, 3, 279-300.

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

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

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

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. In Proceedings of the Joint Statistical Meeting 2002.

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

Graf, M. (2005). Assessing the Precision of Compositional Data in a Stratified Double Stage Cluster Sample: Application to the Swiss Earnings Structure Survey. In Proceedings of CODAWORK'05, ed. Thió-Henestrosa, S., Pawlowski-Glahn V. and Martín-Fernández, JA.

Graf, M. (2006). Swiss Earnings Structure Survey 2002-2004. Compositional data in a stratified two-stage sample: Analysis and precision assessment of wage components. Neuchâtel: Federal Statistical Office, Methodology report 338-0038.

Särndal, C.-E. & Lundström, S. (2005). Estimation in Surveys with Nonresponse. Chichester: Wiley Series in Survey Methodology.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer Series in Statistics.

SESS (2003). Enquête suisse sur la structure des salaires 2002 (ESS 2002). Neuchâtel: Office Fédéral de la Statistique, Actualités, 3, Vie active et rémunération du travail.

SESS (2004). L'enquête suisse sur la structure des salaires 2002. Résultats commentés et tableaux. Neuchâtel: Office Fédéral de la Statistique, Statistique de la Suisse.

SESS (2006). L'enquête suisse sur la structure des salaires 20024. Résultats nationaux. Neuchâtel: Office Fédéral de la Statistique, Statistique de la Suisse.

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

von Eynatten, H., Pawlovski-Glahn, V. & Egozcue, J.J. (2002). Understanding Perturbation on the Simplex: A Simple Method to Better Visualize and Interpret Compositional Data in Ternary Diagrams. Mathematical Geology 34, 3, 249-257.

Appendix

Table A1 : Wage components in overall wage bill - private and public sector (Confederation) combined

Switzerland 2002 and 2004

NACE2	Economic activities	Overtime earnings		Hardship allowances		13th or nth month wage / salary		Special payments / bonuses		Global	Geometric	Mahal-anobis	p-
		estimate	CV	estimate	CV	estimate	CV	estimate	CV	CV	mean	distance	value
0	TOTAL	0.3	2.2	0.7	1.6	6.3	0.9	3.4	4.4	3.4	3.15	13.4	0.01
		0.4	5.4	0.7	1.4	6.3	1.0	3.7	5.1	3.5	3.29		
01	Horticulture	0.3	17.9	0.2	27.8	6.4	1.6	0.6	10.6	21.8	1.76	29.4	0.00
		0.4	13.2	0.1	9.4	5.9	1.7	0.8	12.7	9.2	1.57		
10-14	Mining and quarrying of stone	0.4	8.9	0.3	12.3	8.0	0.9	1.4	13.5	12.8	2.40	11.2	0.02
		0.4	20.0	0.2	15.3	7.7	1.0	2.0	13.0	11.9	2.44		
15-37	Manufacturing	0.6	2.6	1.3	2.2	7.5	0.5	2.5	2.9	2.9	3.86	20.9	0.00
		0.7	2.5	1.3	2.2	7.5	0.4	2.9	3.4	2.2	4.03		
40,41	Electricity, gas and water supply	0.3	14.4	1.3	4.0	7.9	0.7	4.0	10.0	10.8	3.38	10.1	0.04
		0.2	11.3	1.4	4.9	7.9	0.6	2.6	7.8	6.2	3.14		
45	Construction	0.4	12.8	0.4	7.6	7.3	0.8	0.9	8.9	10.7	2.24	2.7	0.61
		0.3	7.6	0.4	7.0	7.3	0.6	1.0	8.2	5.7	2.24		
50-52	Sale, repair	0.2	5.3	0.2	7.2	6.0	1.7	3.1	4.8	6.7	2.23	12.2	0.02
		0.5	22.6	0.3	9.6	6.3	1.3	3.0	3.9	10.3	2.74		
55	Hotels and restaurants	0.2	28.9	0.2	15.5	4.4	2.2	0.6	10.0	21.6	1.44	10.3	0.03
		0.2	14.5	0.1	14.0	4.8	1.5	0.5	7.7	9.4	1.39		
60-64	Transport, storage and communication	0.3	3.5	1.0	4.7	6.8	1.5	1.2	6.4	5.6	2.80	133.5	0.00
		0.4	3.3	1.1	3.2	6.9	1.2	2.5	5.8	3.5	3.36		
65-67	Banking; insurance	0.2	10.8	0.2	9.0	3.6	7.7	11.7	7.5	11.1	2.42	13.4	0.01
		0.1	6.0	0.2	8.9	3.7	9.2	13.3	8.6	7.6	2.30		
70-74	Real estate, computer, research & development	0.3	8.0	0.3	8.2	5.8	1.1	5.2	3.7	7.5	2.62	14.2	0.01
		0.3	7.5	0.2	7.2	5.5	1.4	5.4	4.0	5.1	2.59		
80	Education	0.2	12.1	0.1	10.4	5.2	1.7	0.7	9.0	11.7	1.46	0.9	0.92
		0.2	10.3	0.2	10.8	5.1	1.7	0.7	14.2	9.3	1.50		
85	Health and social work	0.3	6.5	2.0	1.9	6.8	1.1	0.5	5.4	5.7	2.70	55.6	0.00
		0.2	4.8	2.0	1.4	6.9	0.6	0.4	4.8	3.1	2.40		
90-93	Other community, social and personal service activities	0.2	9.0	0.4	9.5	5.1	1.7	1.1	5.7	9.3	2.13	33.4	0.00
		0.2	8.9	0.7	13.7	5.3	1.6	1.5	6.0	8.0	2.31		

For each economic activity , first line: 2002 ; second line: 2004 ; estimates and CV's are expressed in %.

CV: univariate coefficient of variation.

p-value: corresponding p-value under the null hypothesis of no change between 2002 and 2004.

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

Geometric mean: geometric mean of the 5 wage components ("overtime earnings" to "special payments", plus "non-standardized gross earnings with soc. contrib."), expressed as parts of the non-standa

Mahalanobis distance: squared distance in the alr scale between the 2002 and 2004 compositional vectors considering the covariance matrix for change.

Source: Swiss Federal Statistical Office, Swiss Earnings Structure Survey (SESS) 2002 a Original table: wage components only.