Compositional amalgamations and balances: a critical approach

G. Mateu-Figueras[1], **and J. Daunis-i-Estadella**[1]

[1] Universitat de Girona, Girona, E; *gloria.mateu@udg.edu, josep.daunis@udg.edu*

Abstract

The amalgamation operation is frequently used to reduce the number of parts of compositional data but it is a non-linear operation in the simplex with the usual geometry, the Aitchison geometry. The concept of balances between groups, a particular coordinate system designed over binary partitions of the parts, could be an alternative to the amalgamation in some cases. In this work we discuss the proper application of both concepts using a real data set corresponding to behavioral measures of pregnant sows.

Key words: Balances, Amalgamations, logratios.

1 Introduction

In the 80's, Aitchison (1982, 1986) showed that the standard operations we use in real space make no sense from a compositional point of view, and introduced the perturbation, $\oplus$, the power transformation, $\odot$, as the internal and external operations, and the Aitchison distance, d_a. Later, Billheimer et al. (2001) and Pawlowsky-Glahn and Egozcue (2001) introduced independently an inner product, $\langle\rangle_a$, and showed that the simplex with the mentioned operations has an Euclidean vector space structure of dimension $D-1$. Thus, the general theory of linear algebra guarantees the existence of a (non unique) orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$, which leads to a unique expression of a composition $\mathbf{x}$ as a linear combination,

$$\mathbf{x} = (\langle \mathbf{x}, \mathbf{e}_1\rangle_a \odot \mathbf{e}_1) \oplus (\langle \mathbf{x}, \mathbf{e}_2\rangle_a \odot \mathbf{e}_2) \oplus \ldots \oplus (\langle \mathbf{x}, \mathbf{e}_{D-1}\rangle_a \odot \mathbf{e}_{D-1}).$$

In what follows, the vector of coordinates $(\langle \mathbf{x}, \mathbf{e}_1\rangle_a, \langle \mathbf{x}, \mathbf{e}_2\rangle_a, \ldots \langle \mathbf{x}, \mathbf{e}_{D-1}\rangle_a)$ is denoted as $\mathrm{h}(\mathbf{x})$. Note that $\mathrm{h}(\mathbf{x}^* \oplus (\alpha \odot \mathbf{x})) = \mathrm{h}(\mathbf{x}^*) + \alpha \cdot \mathrm{h}(\mathbf{x})$, $\langle \mathbf{x}, \mathbf{x}^*\rangle_a = \langle \mathrm{h}(\mathbf{x}), \mathrm{h}(\mathbf{x}^*)\rangle$ and $d_a(\mathbf{x}, \mathbf{x}^*) = d(\mathrm{h}(\mathbf{x}), \mathrm{h}(\mathbf{x}^*))$, where the lack of a subindex denotes the standard operations in $\mathcal{R}^{D-1}$. This means that standard real analysis can be applied to the coordinates. If we work with coordinates we preserve distances and our results will be coherent form a compositional point of view. Like in every inner product space, the orthonormal basis is not unique but the important point is that, once an orthonormal basis has been chosen, all standard statistical methods can be applied to the coordinates and transferred to the simplex preserving their properties. Nevertheless, the vector of coordinates is not easily interpretable.

The amalgamation operation was introduced by Aitchison(1982) as a fundamental operation on compositions. Certainly the amalgamation of some parts may have a clear sense and may be completely justified. In particular, the amalgamation is a commonly operation used to reduce dimension or to avoid zeros. But it can be show that it is a non-linear operation in the simplex and consequently does not preserve distances. More recently Egozcue and Pawlowsky(2005) advert that amalgamation of parts cannot be considered as a compatible reduction of dimension and introduces the balances as an alternative.

The objective of this paper is to review and clarify the role of amalgamations and balances for compositional data to contribute to the understanding of the existing methodology. A real data set based on behavioural measures of pregnant sows is used to motivate the work and to illustrate the differences between the two mentioned operations.

The paper is organized as follows. In Section 2 and 3 we first review the concepts of amalgamations and balances. In Section 4 we give a description of the data set and in Section 5 we perform the analysis based on amalgamations and balances. Lastly, Section 6 concludes with a final discussion.

2 Amalgamations

If the D parts of a composition are separated into $C \leq D$ mutually exclusive and exhaustive subsets and the components of each subset are added together, the resulting C-part composition is termed an amalgamation (Aitchison, 1982, 1986). For example, from the 6-part composition $(x_1, x_2, x_3, x_4, x_5, x_6)$ we can obtain the following 3-part amalgamation $(x_1 + x_2, x_3, x_4, +x_5 + x_6)$. An amalgamation can be regarded as a composition in a simplex with fewer parts, and thus belonging to a space of lower dimension. For this reason, amalgamations of parts has been extensively used to achieve reduced dimension.

Nevertheless, the amalgamation operation is a non-linear operation in the simplex with respect to the Aitchison geometry described above. It can be interpreted as a projection in a simplex of lower dimension. But the amalgamation operation does not preserve Aitchison distances under perturbation. The consequences might be important, suppose for example that we perform a cluster analysis to a compositional data set; the results might be completely different if we use the original parts or if we work with amalgamations. Moreover, when the analysis of the amalgamated parts is performed simultaneously with the analysis of the original and non-amalgamated parts, difficulties in interpretation and incompatibilities might arise (see Egozcue and Pawlowsky, 2005, for an illustrative example involving perturbations).

In some cases the nature of the sampling method or the particular characteristics of our data leads to amalgamate some components, specially if we have a large amount of zeros. Also, it may be of interest to amalgamate components of a composition and to work with the new amalgamated composition. One very classical case of amalgamation is hidden under the logit transformation. When we consider the $logit(p)$ in a sample of more than two parts, the logratio $\ln(p/(1-p))$ is considered and the term $1-p$ is the amalgamation of all parts except p.

3 Balances

The concept of balance between groups is introduced in Egozcue and Pawlowsky (2005) as a tool to design a particular orthonormal basis on the simplex in order to easily make the corresponding coordinates interpretable. The main idea is to provide a method to analyze grouped parts of a compositional vector thought the adequate coordinates in an orthonormal basis. The method is based on a sequential binary partition of a D-part composition into non-overlapping groups. At each step, a group of parts is partitioned into two non-overlapping groups.

In practice, there is no need to know the exact expression of this basis, as the coordinates can be computed using a one-to-one transformation, and for values of interest the inverse transformation can be used. For example, at i-th step two groups of parts are considered, denoted here as G_{i1} and G_{i2}, then the balance is

$$b_i = \sqrt{\frac{r_i \cdot s_i}{r_i + s_i}} \ln \frac{(\prod_{x_j \in G_{i1}} x_j)^{1/r_i}}{(\prod_{x_\ell \in G_{i2}} x_\ell)^{1/s_i}} .$$

with r_i and s_i representing the number of parts in G_{i1} and G_{i2} respectively. In other terms, the balance is defined as the natural logarithm of ratio between the geometric mean of parts in each group, normalized by a coefficient to guarantee unit length of the vectors of the basis. Therefore, balances are coordinates with respect to an orthonormal basis, denoted here as $\mathrm{h}(\mathbf{x}) = (b_1, b_2, \ldots, b_{D-1})$, and they behave like real random vectors, thus all standard methods can be applied.

Observe that using balances we could easily compare the relative behavior between two groups of variables and using the sequential binary partition we could design the adequate groups. Thus, Egozcue and Pawlowsky (2005) propose the balances as an alternative to the amalgamations because the whole composition is analyzed but also some lower-dimensional representations could be made. Using balances the analysis is compatible and coherent with the Aitchison geometry, in particular we have the invariance of distances under perturbation.

4 Description of the data

The largest amount of information about the welfare of the sows is obtained from the measures of behaviour, particularly measures of activity and stereotypies. Stereotypies are related to poor welfare because they are developed in situation of stress, frustration or lack of control. They reflect a past or present difficulty to cope with the environment. Therefore, the decrease in stereotypies level in group-housing systems could already be considered as a welfare improvement.

The data used in this study are obtained from the comparison among two different commercial housing and feeding systems (trickle feeding and electronic sow feeder) and conventional stalls for pregnant sows. One hundred and eighty pregnant sows were selected on a commercial farm and used in three different replicas (60 sows per replica). In each replica, 20 sows were housed in conventional stalls (Stall), 20 sows were observed using the trickle feeding system (Trick) and 20 sows more using the electronic sow feeder system (Fitmix). Sows were observed for 11 non-consecutive days for 4 h a day. General activity and stereotypies were measured by scan-sampling observation (10-min intervals) in all the systems. The final data set contain information about 177 individuals as 3 of the 180 initially selected sows cannot conclude the study.

Our data are a 5-part vectors containing the observed frequencies of 4 oronasofacial behaviours: interacion with the equipment (E), floor manipulation (T), drinking (D), sham-chewing (S) and one residual part (H). This data are previously studied in Chapinal (2006) and Daunis-i-Estadella et al. (2006a).

The distribution of the observed behaviours may be thought of as coming from a multinomial distribution, with unknown parameters. In a first step, the probability of each behaviour is estimated. In order to correctly estimate the probabilities, the presence of zeros has no sense. Observe that a zero would mean that the behaviour is not possible and we know that the behaviour is possible. Thus a zero is only related to a small time periods of recording observations and consequently a correction has to be implemented (see Daunis-i-Estadella et al., 2006a). In those situations the Jeffrey's estimation (Jeffreys, 1961) is used and obtained by adding 1/2 to each component and applying the closure operation to the resulting vector. Nevertheless, some other studies proposes as an alternative a bayesian estimation with uniform prior distribution (Daunis-i-Estadella et al., 2008). Using this correction our final composition is obtained by adding 1 to each observed frequency and applying the closure operation. Finally, for each sow, we have the estimated composition

$$(e, t, d, s, h) = \mathcal{C}(E + 1, T + 1, D + 1, S + 1, H + 1).$$

5 Balances vs amalgamation

In this section we focus our attention to a specific problem that we met when we start the analysis of this data set. Remember that the objective of this paper is to clarify the role of amalgamations and balances for compositional data, thus a complete study of this data set is not provided here. The reader interested in exploratory compositional data applied to this data set could see Daunis-i-Estadella et al. (2006a) or a work on more general exploratory compositional data tools Daunis-i-Estadella et al. (2006b).

Specialists consider that floor manipulation (t part) is highly associated to the interaction with the equipment (e part). Consequently, if t component has no difference for housing systems they suggest that it may be regrouped with e component.

To analyse if t component discriminates, two different methodologies could be applied. The first one is based on balances and the second one is based on amalgamations. Using the results of these two different studies we discuss and compare the role of amalgamations and balances.

Using balances we can easily compare the t component with the remaining components. At first step our composition is partitioned into two groups $\{t\}$ and $\{e, d, s, h\}$. Thus the first balance is

$$\frac{2}{\sqrt{5}} \ln \frac{t}{(e \cdot d \cdot s \cdot h)^{1/4}} .$$

The other balances depend on the groups formed in the following steps but they are not used here. As balances are coordinates with respect to an orthonormal basis, the standard real methodology can be applied and the analysis of variance (ANOVA) can be used for testing the equality of the means using the housing system as the factor variable. The value of the F statistic with 2 and 174 degrees of freedom is 6.40 and the corresponding p-value is 0.002. Thus our conclusion is that there are significant differences among the 3 means, consequently there is significant evidence for a housing effect in the t component. The assumptions of homogeneity of variances and the normality of observations are checked using the Levene's test (p-value=0.914) and the Anderson-Darling test of normality (p-value=0.716). As a conclusion, we decide not to amalgamate t component with e component and to carry on the exploratory analysis with the 5-part composition (e, t, d, s, h). Thus, it is possible to establish a conclusion on welfare terms related to the different proportion of time or frequencies spent in floor manipulation.

Another often used approach to analyse if t component has or not differences is to study the logratio between t and $1 - t$. Observe that $1 - t$ is obtained as the amalgamation $e + d + s + h$, i.e. we work with composition $(t, e + d + s + h)$. As the dimension is reduced, we could easily compare t component with the rest. Now, following the standard compositional data analysis methodology, we work with the logratio

$$\frac{1}{\sqrt{2}} \ln \frac{t}{(e + d + s + h)},$$

that is, the coordinates of composition $(t, e + d + s + h)$ with respect to the orthonormal basis stated in Egozcue et al. (2003). As in the previous case, the analysis of variance (ANOVA) can be applied for testing the equality of the means using the housing system as the factor variable. In this case, the analysis is equivalent to the classical analysis of the $\text{logit}(t) = \ln(t/1 - t)$ except for the constant $1/\sqrt{2}$. This constant only guarantees the unit length of the vector of the basis (note that we have a two part composition and the simplex has dimension 1) but the ANOVA results are not affected. The value of the F statistic with 2 and 174 degrees of freedom is 0.280 and the corresponding p-value is 0.753. Thus our conclusion is that there is no significant difference among the 3 means, that is there is no significant evidence for a housing effect in the t component. The assumptions of homogeneity of variances and the normality of observations are checked using the Levene's test (p-value=0.124) and the Anderson-Darling test of normality (p-value=0.110). At this point we decide to amalgamate t component with e component and we follow our study with the 4-part composition $(e + t, d, s, h)$. Note that the previous amalgamation $e + d + s + h$ will not longer be considered.

6 Discussion

Two completely opposite conclusions are obtained using balances or amalgamations. In both cases the standard compositional methodology is used, as we work with logratios or coordinates with respect to an orthonormal basis. Therefore, which is the most suitable analysis in this case?

Remember that our dilemma here is to choose between the 5-part composition (e, t, d, s, h) or the 4-part composition $(e + t, d, s, h)$. It is important to note that using amalgamations, we first work with the 2-part composition $(t, e + d + s + h)$ but a conclusion in terms of the original and not amalgamated parts is finally obtained. This is not the case using balances. In those situations we have to be extremely careful and to avoid the amalgamations because we know that it is a non linear operation and it don't conserve the scale.

Only for illustrative purposes, the centered data set is now considered. The centering transformation was introduced by Martín-Fernández et al. (1999) as a perturbation that serves to move our data set into the center of the simplex. It is equivalent to translate the corresponding orthonormal coordinates to the origin of coordinates in the real space. In our case, the perturbation to the original parts is first applied and the same analysis using balances, amalgamations and the ANOVA methodology is repeated. As a perturbation is a translation on the simplex, the results of our analysis must be the same. Nevertheless, using amalgamations we obtain 7.31 as the F statistic with a p-value equal to 0.001. Now, our conclusion using amalgamations is that there are significant differences among the 3 means, consequently there is significant evidence for a housing effect in the t component. Using balances we obtain exactly the same results as before, i.e. the value of the F statistic is 6.40 and the corresponding p-value is 0.002.

We have showed that the amalgamation operation doesn't conserve distances. Thus, if an analysis with some amalgamated parts is combined with an analysis involving non amalgamated parts, incompatibilities and problems could arise. Thus, in this case, an analysis using balances is the most appropriate technique.

However, balances are not a perfect alternative to amalgamations. One clear example is a compositional data set with zeros. The Aitchison theory in general and balances in particular excludes dealing with zeros because logratios among components are used. Martín-Fernández et al. (2003) or more recently Palarea-Albaladejo et al. (2007) propose some replacement methods. But, if the compositional data set have a large amount of zeros, the amalgamation could be the solution. In fact, Martín-Fernández et al. (1997) try to perform a classification using a 8-part compositional data set from the Darss Sill area with a large amount of zeros. As a first step an amalgamation is proposed to reduce the number of zeros. In this particular case, the zeros are concentrated in a few components. This is interpreted as a sign of overdimension concerning the number of components thus the amalgamation is justified.

But, in which other cases could the amalgamation operation be used? Basically it a question related to decide how many parts are reasonable to consider in an initial step. We can always apply the amalgamation operation if it has a clear sense and we are only interested in studying the relative variability of the parts of the new amalgamated composition. For example, let's suppose that our interest is to study the relative variability of the 2-part composition $(t, e + d + s + h)$. To avoid incompatibilities, we have to start with this 2-part composition and to perform our analysis with these only two parts. After having amalgamated, we will have no problems. We can, for example, center our data 2-part compositional data set and apply the ANOVA methodology for testing the equality of the means using the centered parts. In this case we obtain $F = 0.280$ and the corresponding p-value is 0.753, the same as we obtain with the amalgamated but not centered data.

In conclusion, the amalgamation operation could be considered in an initial step. Nevertheless, once the amalgamation is made, we have to work with the resulting and amalgamated parts because when the analysis of the amalgamated parts is performed before or simultaneously with the analysis of the original and non-amalgamated parts misinterpretations and incompatibilities could arise. Using balances the scale is conserved and the compositional coherence is always achieved.

Acknowledgements

This work has been supported by the Spanish Ministry of Education and Science under project 'Ingenio Mathematica (i-MATH)' No. CSD2006-00032 (Consolider – Ingenio 2010) and under project MTM2006-03040.

References

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

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

Billheimer, D., Guttorp, P. and Fagan, W. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association*, **96(456)**, 1205-1214.

Chapinal, N. (2006). *Effect of the housing and feeding system on the welfare and productivity of pregnant sows.* Ph-D Thesis. Dept. Ciència dels Animals i dels Aliments. UAB-Barcelona

Daunis-i-Estadella, J., Mateu-Figueras, G., Chapinal, N., Manteca, X. and Ruiz de la Torre, J. L.(2006a). Aplicación de técnicas composicionales al estudio del comportamiento de cerdas gestantes. In: *Libro de actas del XXIX Congreso de la SEIO.*

Daunis-i-Estadella, J., Barceló-Vidal, C. and Buccianti, A.(2006b). Exploratory compositional data analysis. In: *Compositional Data Analysis in the Geosciences: From Theory to Practice* (eds: A. Buccianti, G. Mateu-Figueras and V. Pawlowsky-Glahn), Geological Society, London, Special Publications, 264, 161–174.

Daunis-i-Estadella, J., Martín-Fernández, J.A. and Palarea-Albaladejo, J.(2008). Bayesian tools for zero counts in compositinal data In: *Proceedings of the 3r International Workshop on Compositional Data, CoDaWork'08.*

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

Egozcue, J. J. Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C.(2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, **35(3)**, 279-300.

Jeffreys, H. (1961). *Theory of Probability.* (3rd ed.) Oxford : Clarendon Press. 447 p.

Martín-Fernández, J.A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1997). Different Classifications of the Darss Sill Data Set Based on Mixture Models for Compositional Data. In: *Proceedings of the Third annual conference of the International Association for Mathematical Geology, IAMG'97* (ed: V. Pawlowsky-Glahn), 152-156, CIMNE, Barcelona(E).

Martín-Fernández, J.A., Bren, M., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1999). A measure of difference for compositional data based on measures of divergence. In: *Proceedings of IAMG'99, the fifth annual conference of the International Association for Mathematical Geology* (eds: S.J. Lippard, A. Næss and R. Sinding-Larsen), pp. 211–216. Tapir, Trondheim (N).

Martín-Fernández, J.A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (2003) Dealing with Zeros and Missing Values in Compositional Data Sets. *Mathematical Geology*, **35(3)**, 253-278.

Palarea-Albaladejo, J., Martín-Fernández, J. A., Gómez-García, J.A. (2007) Parametric Approach for Dealing with Compositional Rounded Zeros. *Mathematical Geology*, **39(7)**, 625-645.

Pawlowsky-Glahn, V. and Egozcue, J. J.(2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)*, **15(5)**, 384-398.