Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 701–707.

DEALING WITH ROUNDED ZEROS IN COMPOSITIONAL DATA UNDER DIRICHLET MODELS

Rafiq H. Hijazi
Department of Statistics
United Arab Emirates University
P. O. Box 17555, Al-Ain, UAE
E-mail: rhijazi@uaeu.ac.ae

ABSTRACT

One of the obstacles facing the application of the Dirichlet modeling of compositional data is the occurrence of zero. In the Dirichlet model, the presence of zeros makes the probability density function vanish. Zeros in compositional data are classified into "rounded" zeros and "essential" or true zeros. The rounded zero corresponds to a small proportion or below detection limit value while the essential zero is an indication of the complete absence of the component in the composition. Several parametric and non-parametric imputation techniques have been proposed to replace rounded zeros and model the essential zeros under logratio model. In this paper, a new method based on Beta regression is proposed for replacing rounded zeros in compositional data. The performance of the proposed method is analyzed using Monte Carlo simulation and an illustrative example using real data is given.

1. INTRODUCTION

Compositional data are non-negative proportions with unit-sum. This type of data arises whenever objects are classified into disjoint categories and the resulting relative frequencies are recorded, or partition a whole measurement into percentage contributions from its various parts. The sample space of compositional data is the simplex S^D defined as

$$S^D = \left\{ (x_1, \ldots, x_D) : x_j > 0 \text{ for } j = 1, 2, \ldots, D \text{ and } \sum_{j=1}^{D} x_j = 1 \right\}$$

Compositional data occur in nearly all disciplines, but recognition and modeling of their basic structure have gotten particular attention in geology, chemistry, political science, business and economics. For example, economists might be interested in how the composition of household income spent on food, housing, clothes, entertainment and services. Due to the unit-sum constraint and its consequences, traditional regression models are not suitable for modeling compositional data. Aitchison (1986) suggested an analysis based on the logratios of the compositional data. Campbell and Mosimann (1987) developed an alternative approach by extending the Dirichlet distribution to a class of Dirichlet Covariate Models. Hijazi and Jernigan (2009) developed maximum likelihood inference in Dirichlet regression models. Hijazi (2006, 2008) investigated the diagnostics checking and the residuals analysis in Dirichlet regression.

In compositional data analysis, the presence of zero components represents one of the main obstacles facing the application of both logratio analysis and Dirichlet regression. In a logratio

analysis, we cannot take the logarithm of zero when applying the additive logistic transformation. In the Dirichlet model, the presence of zeros makes the probability density function vanish.

In this paper, we propose a new technique, based on Beta regression, for replacing the zeros under Dirichlet model. Section 2 gives an overview of zeros and zero replacement strategies in compositional data besides the new proposed replacement method. A Monte Carlo simulation study to compare the proposed method with the multiplicative replacement strategy is presented in Section 3. An application to illustrate the use of the proposed technique is presented in Section 4. Finally, concluding remarks are given in Section 5.

2. ZERO REPLACEMENT IN COMPOSITIONAL DATA

2.1 Types of Zeros in Compositional Data

Aitchison (1986) classified the zeros in compositional data into "rounded" or trace elements zeros and "essential" or true zeros. The trace zero is an artifact of the measurement process, where observation is recorded as zero when it is below the detection limit (BDL). For example, in the porphyry deposits, assume that we record the amount on the different elements. If the scale used does not identify the presence of the element if it is less than 0.2%, then this component is recorded as zero. Thus the observed zero is a proxy for a very small number below 0.2%. On the other hand, often the observation is recorded as zero as an indication of the complete absence of the component in the composition. In the household budget, a household spending nothing on tobacco will have a zero for the share of the tobacco in the budget.

2.2 Common Zero Replacement Strategies

The treatment of the zero observations in compositional data should be done according to the cause of the zero (Aitchison 1986). Several attempts have been made to deal with the essential zeros using ranks (Bacon-Shone 1992) and conditional modeling (Aitchison and Kay 2003). In case of rounded zeros, Aitchison (1986) suggested the reduction of the number of components in the composition by amalgamation. That is, eliminating the components with zero observations by combining them with some other components. Such approach is not appropriate when the goal is modeling the original compositions or the model includes only three components. However, a more logical approach is to replace the rounded zeros by a small nonzero value that does not seriously distort the covariance structure of the data (Martìn-Fernández *et al.* 2003a). The first replacement method, the additive replacement, proposed by Aitchison (1986) is simply replacing the zeros by a small value δ and the normalizing the imputed compositions. Fry *et al.* (2000) showed that the additive replacement is not subcompositionally coherent and consequently, distorts the covariance structure of the data set.

Martin-Fernandez *et al.* (2003a) proposed an alternative method using a multiplicative replacement which preserves the ratios of nonzero components. Let $x = (x_1,\ldots,x_D) \in S^D$ be a composition with rounded zeros. The multiplicative method replaces the composition x containing c zeros with a zero-free composition $r \in S^D$ according to the following replacement rule

$$r_j = \begin{cases} \delta & if \quad x_j = 0 \\ (1-c\delta)x_j & if \quad x_j > 0 \end{cases} \tag{1}$$

In addition, Martin-Fernandez *et al.* (2003a) emphasized that the best results are obtained when δ is close to 65% of the detection limit. However, since the multiplicative replacement imputes exactly the same value in all the zeros of the compositions, this replacement introduces artificial correlation between components which have zero values in the same composition.

Besides these nonparametric approaches, several parametric approaches based on applying a modified EM algorithm on the additive logratio transformation (Martin-Fernandez *et al.* (2003b), Palarea-Albaladejo *et al.* (2007) and Palarea-Albaladejo and Martìn-Fernández (2008)). However, none of these methods is applicable when the compositional data arise from Dirichlet model.

2.3 Beta Regression Based Strategy

As mentioned earlier, the existing parametric replacement strategies assume that the compositional data arise from the additive logistic normal distribution, the underlying distribution in logratio analysis (Aitchsion 1986). When the underlying distribution is Dirichlet, a natural imputation method should be based on Beta distribution as a marginal of Dirichlet distribution. Ferrari and Cribari-Neto (2004) have proposed a regression model when the response variable is beta distributed. The proposed model is based on the parameterization of the mean and dispersion parameters in the beta distribution as follows. If Y is beta distributed with parameters p and q, then the dispersion parameter $\phi=p+q$ and the mean parameter is then $\mu = \pi/\phi$. The density of Y using this parameterization is given by

$$f(y;\mu,\phi)= \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0<y<1 \tag{2}$$

Similar to OLS regression, the beta regression model is obtained by assuming that the mean μ can be written as a linear combination of some independent variables $x_1,\ldots,x_k$ using a link function $g(.)$ as

$$g(\mu)= \sum_{i=1}^{k} x_i\beta_i \tag{3}$$

where $\beta_1,\ldots,\beta_k$ is a vector of unknown parameters. Let $\mathbf{X}=(\mathbf{x_1},\ldots,\mathbf{x_D})$ be the compositional data with rounded zeros in component x_j. Our proposed approach works as follows:

1. Split **X** based on the existence of rounded zeros in x_j into a zero-free subdata $X_{(1)}$ and $X_{(2)}$, a subdata with all zeros in component x_j.
2. Apply beta regression on portion of $X_{(1)}$ where the values of the j^{th} component are close to the detection limit. The j^{th} component is the response variable and the rest of components as covariates.
3. Using the estimated regression parameters in (2), predict the imputed values of the rounded zeros in $X_{(2)}$.
4. Use the multiplicative replacement strategy using the imputed values to replace the rounded zeros.
5. Repeat this process sequentially on components with rounded zeros.

It is clear that this method takes into account the information included in the covariance structure and produces different imputed values for each composition. The method also assumes that the component with rounded zeros is correlated with the other components in the compositional data especially in $X_{(1)}$. It is noteworthy that external covariates related to the component with rounded zeros might be used in the regression model sole or jointly with the compositional components. In addition, this method would not replace the zeros by negative values but it might replace them by values over the detection limit.

3. SIMULATION-BASED RESULTS

Our interest is mainly focused on to what extent the zero replacement strategies affect to the estimation of the relationship between the components. Consider a 4-component random composition drawn from Dirichlet distribution with parameters 1, 4, 15 and 20 i.e., $\mathcal{D}(1,4,15,20)$. For our simulation purposes, 100 datasets **X** of size 200 are drawn from the above Dirichlet distribution. Next, small values under the detection limit in the first component of **X** are replaced by zero. A range of 10 detection limits is considered from 0.0025 to 0.025 with increments of 0.0025. Thus, 1000 datasets containing different number of rounded zeros are generated. To measure the distortion between the original data **X** and the imputed data $\mathbf{X}^*$, the mean squared distances (MSD) is used. The MSD is given by

$$MSD = \frac{\sum_{i=1}^{200} d_a^2(x_i, x_i^*)}{200} \tag{4}$$

where the Aitchison's distance (d_a) is defined as

$$d_a(x_i, x_i^*) = \sqrt{\sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{x_i^*}{x_j^*} \right)^2} \tag{5}$$

Figure (1) shows the mean MSD for the two replacement methods for the different detection limits. For small detection limits, the multiplicative replacement seems to perform better than our proposed method. However, as the detection limit increases and consequently the percentage of zeros, the beta regression based method outperforms the multiplicative replacement method. Same conclusion is drawn if the Euclidean distance is used instead of Aitchison distance in (4). To compare the effect of replacement method on the variability in the compositional data, the mean estimates of Dirichlet parameters are used. The variability in Dirichlet model is inversely proportional to the sum of its parameters. Figure (2) shows the mean estimate of the first parameters in the original data and the imputed data. Similar behavior is shown for the rest of the parameters. Compared to the proposed method, under the multiplicative replacement, the parameter is clearly overestimated and hence the total variability is underestimated. Such underestimation increases as the percentage of zeros increases. This is due to the replacement of all zeros with same value which is not the case in beta regression based method.

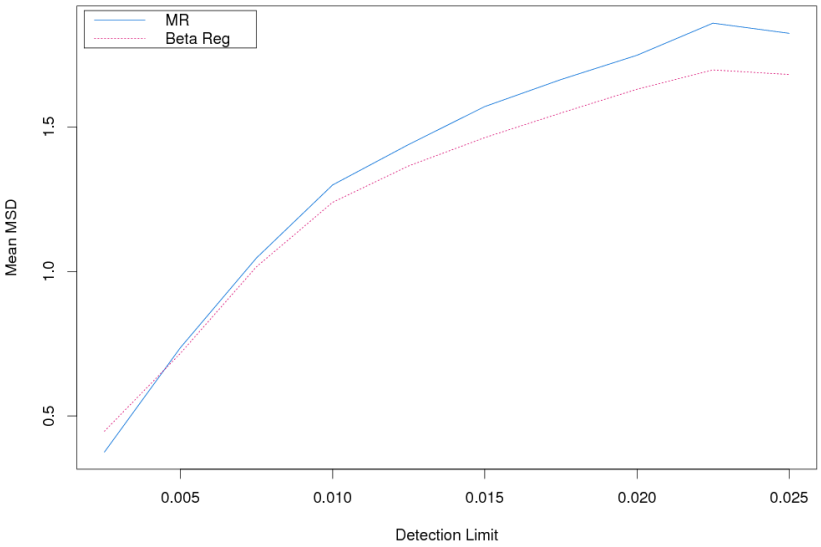

Figure 1: Replacement methods distortion

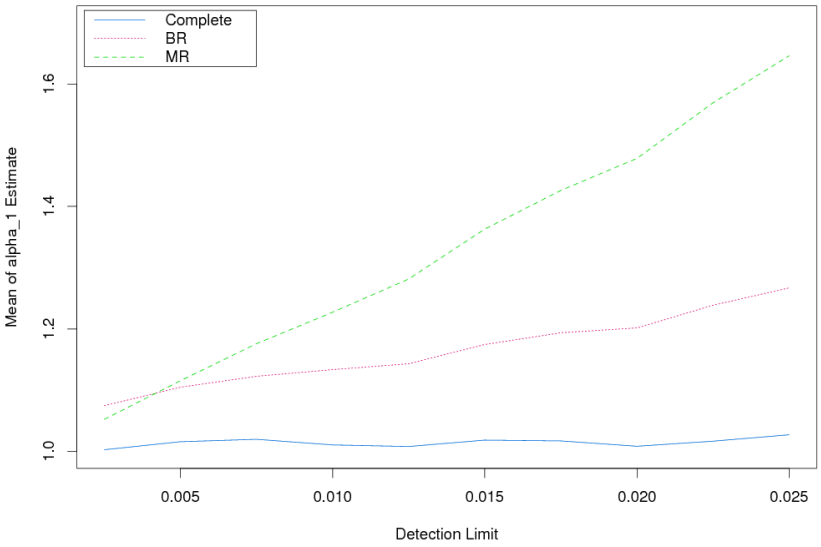

Figure 2: Parameter estimation under Replacement methods

4. APPLICATION

Consider the data collected from a deep-sea core measuring 478 cm in length from the Mediterranean Sea floor (Davis 2002). The core was split and grain-size analysis was made of 51 intervals. This paper focuses on the new proposed replacement method, we will consider that the instrument used does not detect the presence of any element if the percentage is less that 2.5%,

i.e. the detection limit is 2.5%. This will result in 9 compositions with sand component recorded as rounded zero as shown in Figure (3a). The Aitchison distance between the original data and the imputed datasets using the beta regression based method and multiplicative methods are 0.0042 and 0.0094, respectively. This indicates that the new replacement method yielded an imputed data which is closer to the original than the one produced by the multiplicative method. The compositions with rounded zeros and the corresponding imputed compositions are shown in Figure (3b).

The maximum likelihood estimates of the original data and the imputed data are given in Table (1). It is clear that the multiplicative replacement method underestimated the model parameters compared to the original data but the beta regression based method overestimated such parameters. The sum of the estimates under the proposed method is slightly larger than the corresponding sum in the original data resulting in a slightly smaller estimate of the variability. However, the multiplicative replacement method yielded slightly larger estimate of the variability.

Table 1: Maximum likelihood estimates of original and imputed sediments data

	Clay	Silt	Sand	Sum
Original Data	10.599	28.702	2.672	41.973
Beta regression replacement	11.041	29.934	2.831	43.806
Multiplicative replacement	10.243	27.709	2.538	40.490

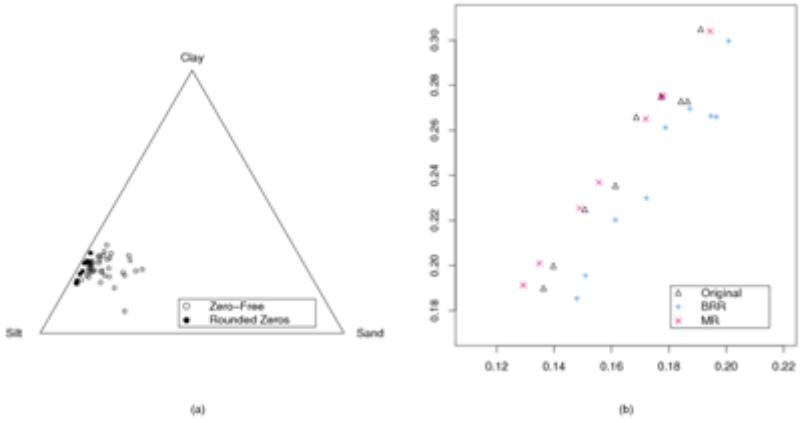

Figure 3: (a) The ternary diagram of the sediments data (b) The original compositions before rounding, imputed compositions with BRR (beta regression based replacement) and imputed compositions with MR (multiplicative replacement).

5. COMMENTS AND CONCLUSION

In this work we have proposed a new replacement method based on beta regression under Dirichlet model. The proposed method was compared with the multiplicative replacement method through simulation study and real data example implemented in S-Plus. The new method outperforms the multiplicative replacement method especially in datasets with large percentage

of zeros. This method gives positive imputed value but does not take into account the detection limit of the part. The method should be modified to overcome this deficiency. The proposed method is expected to be less efficient in the absence of correlation between the component with zeros and other components or external variables.

REFERENCES

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, New York.

Aitchison J., Kay J. W. (2003). Possible solutions of some essential zero problems in compositional data analysis. In: Thió-Henestrosa S., and Martìn-Fernández, J.A. (Eds.), Proceedings of CODAWORK'05, The 2nd Compositional Data Analysis Workshop, October 19-21, University of Girona, Girona (Spain).

Bacon-Shone, J. (1992).Ranking Methods for Compositional Data. *Applied Statistics*, 41, 533-537.

Campbell, G., and Mosimann, J. E. (1987). Multivariate methods for proportional shape. *ASA Proceedings of the Section on Statistical Graphics*, 10-17.

Davis, J. C. (2002). *Statistics and data analysis in geology*. John Wiley & Sons, New York.

Ferrari, S. L. P., Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics*, 31(7), 799–815.

Fry, J. M., Fry, T. R. L., and McLaren, K. R. (2000). Compositional data analysis and zeros in micro data: *Applied Economics*, 32(8), 953-959.

Hijazi, R. (2006). Residuals and Diagnostics in Dirichlet Regression. *ASA Proceedings of the Joint Statistical Meetings 2006, American Statistical Association*, 1190-1196.

Hijazi, R. (2008). Residuals Analysis of the Dirichlet Regression. *Egyptian Statistical Journal*, 52 (2), 109-120.

Hijazi, R., Jernigan, W. (2009). Modeling Compositional Data Using Dirichlet Regression. *Journal of Applied Probability and Statistics*, 4 (1), 77-91.

Martìn-Fernández, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V., (2003a). Dealing with zeros and missing values in compositional data sets. *Mathematical Geology*, 35, 253-278.

Martìn-Fernández, J.A., Palarea-Albaladejo J, Gómez-Garcìa, J. (2003b). Markov chain Monte Carlo method applied to rounding zeros of compositional data: first approach. In: Thió-Henestrosa S., and Martìn-Fernández, J.A. (Eds.), Proceedings of CODAWORK'05, The 2nd Compositional Data Analysis Workshop, October 19-21, University of Girona, Girona (Spain).

Palarea-Albaladejo, J., Martìn-Fernández, (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences*, 34, 902-917.

Palarea-Albaladejo, J., Martìn-Fernández, J.A., Gómez-Garcìa, J., (2007). A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology*, 39, 625-645.