BIOSCIENCE JOURNAL

GEE-LOGIT MODEL CORRECTED BIPLOTS WITH HARVEST EFFECTS ON COFFEE BEANS GRADING

Haiany Aparecida FERREIRA[1], Érica Resende de OLIVEIRA[2], Carla Regina Guimarães BRIGHENTI[3], Marcelo Ângelo CIRILLO[4]

[1] Postgraduate program in Statistics and Agricultural Experimentation, Federal University of Lavras, Lavras, Minas Gerais, Brazil.
[2] Food Engineering Department, Agronomy School, Federal University of Goiás, Goiânia, Goiás, Brazil.
[3] Animal Science Departament, Federal University of São João del-Rei, São João del-Rei, Minas Gerais, Brazil.
[4] Statistical Department, Federal University of Lavras, Lavras, Minas Gerais, Brazil.

Corresponding author:
Haiany Aparecida Ferreira
Email: haianyferreira@yahoo.com.br

How to cite: FERREIRA, H.A., et al. Gee-logit model corrected biplots with harvest effects on coffee beans grading. *Bioscience Journal*. 2021, **37**, e37044. https://doi.org/10.14393/BJ-v37n0a2021-53679

Abstract

In a granulometric analysis of coffee beans with different categories of defects, the data can be organized in contingency tables, and when considering the discrimination by harvest, they may have a structure that suggest a more complex model, by means of the counting of defective coffee beans compared to different crops interacting with the classification of defects and percentages of sieve grains, which characterizes a block design with multivariate responses. However, due to the techniques based on the analysis of variance, considering the uniform correlation structure for all plots, it becomes feasible to propose a model that allows contemplating different structures between the plots, associating the effects of the crops to the defects in the granulometric procedure applied to the coffee beans. Thus, the hypothesis of incorporating the effects of crops associated with defects arises using the biplot multivariate technique. This work aims to propose the use of corrected biplots by predictions obtained trhough the fit to the Generalized Linear Model in the coffee grain size classification, broken down by components of the effect of the harvests. In conclusion, the use of GEE models with the corrected biplot technique by the predictions is feasible for application to be applied to the granulometric analysis of defective coffee beans, presenting discrimination regarding the effects of harvests.

Keywords: DVS. Generalized Estimating Equations. Uniform Correlation Structure.

1. Introduction

Brazil currently accounts for approximately 35% of all the world's coffee production (Moura et al. 2015), reflecting in a high consumption and exportation rate. Given this, the classification of this product in Brazil has been increasingly encouraged, to obtain a higher quality drink. The granulometric classification is essential to discriminate special and traditional coffees, given the quality of the beans in relation to the absence and moderate or excessive presence of defective beans.

Basically, there are two classifications, given the defects found in coffee batches, which can be intrinsic, when attributed to the imperfections of the bean itself; or extrinsic, when caused by the presence of impurities. Extrinsic defects are caused by odd fractions present in the processed coffee, such as: coconut, sticks, marinheiro, bark and stones. Intrinsic defects result from beans that were badly granulated, or beans that are broken, dry, black, green, or burnt. Also, intrinsic defects may result from the beans own genetics and physiology or due to flaws in agricultural or industrial procedures.

According to Esquivel and Jiménez (2012), defective beans represent about 15 to 20% of coffee production. Among these defects, black, green and burnt grains can be highlighted as a single category, which are considered the worst defects, as they directly affect the quality and type of coffee.

According to the Classification Manual of the Coffee Trade Center of the state of Minas Gerais (CCCMG 2018), the number of defective beans evaluated following the table of Official Brazilian Classification (COB) is counted from each sample and will determine the type of coffee. For this, the determination of the number of defective grains is carried out using samples of 300 g each and these are converted into defects according to an equivalence table. This defect count has its result compared to a seven-level scale, in which each level corresponds to a type of defect that expresses an order of classification from less defects to more defects.

Regarding the classification of grains in sieves, the granulometry process is referred to the shape of the grains. More details on the characterization of these sieves are available at Soares et al. (2019).

In a case study, where a new methodological approach to study the relationship among defects considering information about the crops was proposed, Brighenti and Cirillo (2018) addressed the analysis of this database using a data structure organized in a double entry table relating the counts of the types of defects as a function of the sieve percentage, for each block effect represented by the harvests (S2014 and S2015). Based on the above, the combined biplots technique was applied, following the decomposing singular values methodology enhanced by Greenacre (2003), with the advantage of filtering eigenvalues corresponding to the construction of biplots which are specific to the additive effects of the blocks (S2014 + S2015) and the difference between them (S2014-S2015).

As a result of this filtering, the exploratory detection of the correlated variables through the graphs becomes clearer. However, there is a procedure related to the attribution of greater importance to the scores that prioritize the generation of column coordinates (percentage of sieves) that generate a maximum quality of representation for the columns and minimum for the lines (types of defects), or vice-versa, and in a more balanced way, the same importance is given to the coordinates of lines (types of defects) and columns (percentage of sieves).

The procedure to be used to build the graphs consists of setting constants for the scale parameter δ, respectively at values 0, 1, and 0.5. Such aspects were not addressed in the statistical analysis of the granulometric classification given in Brighenti and Cirillo (2018), as well as in the construction of biplots described by Greenacre (2003).

Another issue to be highlighted arises from the fact that the whole procedure described above does not involve inferential aspects such as the association with models in which the predictions obtained are contemplated by the correlation structure associated with the repeated measures either for the sieve percentage or types of defects. In this sense, this study provides several advantages, which are related to the model adjustment and its contribution to the interpretation of a Biplot, which heterogeneity effect results in more asymmetric biplots. In such context, the researcher is expected to have more confidence while interpreting biplots, avoiding subjectivity in his interpretation.

Given this, a possible organization of the obtained data is from the counting of defective beans organized in contingency tables, whose frequencies correspond to the count of defective coffee beans classified by the types of defects and percentage of sieves. It is relevant to state that, in this conventional approach it is not possible to discriminate grains obtained in different harvests, and that the counts are not evaluated by a statistical model.

Given these arguments, this study aim was to propose the use of generalized estimating equations models (Weng and Wei 2021) together with the biplot technique (Greenacre 2003), incorporating the effects of crop components in an application related to granulometric analysis of coffee beans.

Thus, it is expected that the predictions obtained by the GEE model can introduce an inferential procedure to the biplot technique, adding predictable information regarding the estimation of eigenvectors and eigenvalues necessary to understand and interpret the defects of coffee beans in relation to different effects of harvests.

2. Material and Methods

The database for application of the proposed methodology was obtained from a study by Brighenti and Cirillo (2018), in which the processed samples of coffees from Catuaí cultivar were harvested by rural producers in the Southern region of Minas Gerais (Brazil). The proportion of defective beans or impurities contained in a 300g sample of processed coffee was obtained for each type of defect (Table 1), according to the COB table.

Table 1. Defect counts as percentage (p) of retained grains in the sieve regarding 2014 and 2015 harvests.

Harvest	Defect type	Percentage of retained grains (p)		
		(p < 20%)	(20% ≤ p ≤ 30%)	(p >30%)
S2014	d_1 (pest demaged)	187	223	151
	d_2 (black, green, and burnt)	744	916	759
	d_3 (broken)	588	498	465
	d_4 (shell)	127	252	156
	d_5 (impurities)	12	8	12
S2015	d_1 (pest demaged)	264	190	274
	d_2 (black, green, and burnt)	474	552	725
	d_3 (broken)	324	430	262
	d_4 (shell)	210	293	382
	d_5 (impurities)	12	0	53

A contingency table of three entries was organized considering five coded defects (d1 to d5) (Table 1), given the counts related to the types of defects and the percentage of flat grains (which have a flat side and a convex side) in the sample, which have higher commercial value. Following this data structuring, the additive effects (S2014 + S2015) and difference (S2014 - S2015) between the 2014 and 2015 harvests were incorporated as suggested by the methodology proposed in this study. The adjustment fit of the Generalized Linear Model using da GEE was made in function of the proportions obtained in relation to the marginal totals of the frequencies described in Table 1, assumed as binomial answers, represented by Y_{ijk}, the i-th defect counts (i = 1,..., 5) to the sieve percentage (j = 1,..3) observed in the k-th harvest (k = 1, 2). Therefore, each observed unit y_{ijk}, was represented by a vector, $Y_r=[Y_{r1},Y_{r2},..., Y_{rN}]$; r=1,...,N, where, N is the total number of parcels for each block, contextualized by the harvests.

Considering this specification, the logit binding function (1) was made as a function of the linear predictor η_{ijk}, and the parameter estimates were obtained by the solution of the system (2) and the Gee-Logit model (Silva and Cirillo, 2018), was obtained (3).

$$\log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right)=\eta_{ijk} \quad (1) \qquad \sum_{i=1}^{N}\left(\frac{\partial\mu_i(\beta)}{\partial\beta}\right)R_{i(\alpha)}^{-1}\left(Y_i-\hat{\mu}_i(\beta)\right)=0 \quad (2) \qquad \mu_{ijk}=\frac{\exp\left(\eta_{ijk}\right)}{1+\exp\left(\eta_{ijk}\right)} \quad (3)$$

Where, $\hat{\mu}_i(\beta)$ is the vector of the adjusted proportions; $R_i(\alpha)$ is the uniform working correlation structure (Weng and Wei 2021) to the responses of the i-th defect and is given by (4).

$$R_i=\begin{bmatrix}1 & \hat{\alpha} & \cdots & \hat{\alpha}\\ \hat{\alpha} & 1 & \cdots & \hat{\alpha}\\ \vdots & \cdots & \ddots & \vdots\\ \hat{\alpha} & \hat{\alpha} & \cdots & 1\end{bmatrix}, \quad (4)$$

The estimate of the association parameter (α) and dispersion (ϕ) is give, respectively, by (5).

$$\hat{\alpha} = \frac{\phi \sum_{r=1}^{N} \sum_{i \geq r} \hat{e}_{r_i} \hat{e}_{r_{i'}}}{\left\{ \sum_{r=1}^{N} \frac{1}{2} n_r \left(n_r - 1 \right) - d \right\}} \quad ; \quad \hat{\phi} = \left(\sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} e_{ij}^2 \right) \frac{1}{N} \qquad (5)$$

Where, d is the number of parameters and is the estimate of the dispersion parameter in function of the Pearsons's residue (e_{ij}). Indeed, we chose this matrix due to the nature of the problem, which is related to the granulometry.

The repeated measures within the block, S_k, (k = 1,2), for each type of defect (D), and sieve percentage (P) do not present a temporal order that justifies other structures such as AR (1) or M-dependent. Therefore, there was no need to test other correlation structures.

The results were illustrated using the biplots technique (Chambers, 2018), considering the predicted values obtained in (3) as the origin. In order to incorporate the effect of the harvest associated with the defects, the following biplots were considered: GH-Biplot (δ=0), RMP-Biplot (δ=1), and SQRT-Biplot (δ=1/2) (Mair, 2018).

Overall, the interpretation of a biplot is (Figure 1) in line with the identification of the sample and variable units.

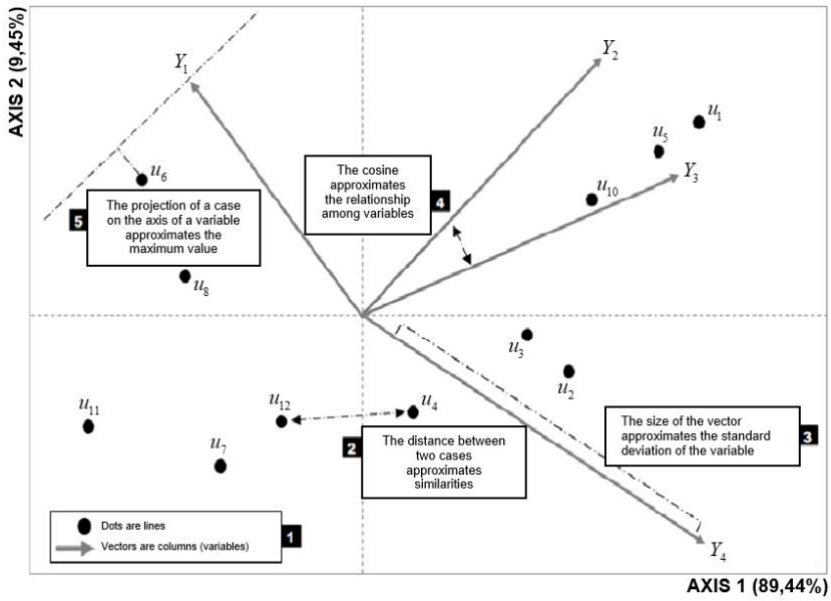

Figure 1. Biplot representation.

Figure 1 shows the correlation between the variables, which can be verified through the angles formed between the vectors and is usually represented by the cosine of the angles. Thus, the correlation will be positive when the angle formed is acute, negative when the angle is obtuse and without correlation when the angle is straight. In Figure 1, dots represent the sampling units, vectors represent the variables, and the projection of a unit on the axis of a variable approximates the maximum value (Torres-Salinas et al. 2013).

The coordinates necessary for the construction of the biplots were computed by applying the singular value decomposition, divided into two blocks (Greenacre 2003), symbolized by the particle size evaluations of the samples collected in the 2014 and 2015 harvests. Considering this partition, the coordinates that characterize the additive effect and difference between harvests obtained for the construction of the biplot were given by S2014+S2015 and S2014-S2015 (R Core Team 2016). In this context, different biplot graphs were generated with different δ scale parameters. Thus, the graphs were characterized by the Column Metric Preserving (CMP or GH) Biplot (δ = 0), which prioritizes the column coordinates in the obtaining of

the coordinates. Similarly, the Row Metric Preserving (RMP or JK) Biplot (δ = 1) prioritizes line coordinates. Finally, the Square Root (SQRT) Biplot (δ = 1/2) gives equal importance to row and column coordinates.

3. Results

Having as a reference the first level of each factor – harvest year, defects, and sieve percentage - the estimates of the parameters (4) that make up the linear predictor is given below considering the uniform correlation structure.

$$\begin{aligned}\log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) = & -0.0196\times S_2 - 0.2199\times d_2 + 0.0187\times d_3 - 0.4495\times d_4 - 0.2633\times d_5 \\ & -0.1157\times p_2 - 0.3779\times p_3 + 0.2950\times d_2p_2 + 0.1283\times d_3p_2 \\ & +0.6998\times d_4p_2 - 0.8597\times d_5p_2 + 0.3445\times d_2p_3 - 0.1956\times d_3p_3 \\ & +0.5741\times d_4p_3 + 1.0537\times d_5p_3 + 0.0003\times S_2p_2 + 0.0020\times S_2p_3\end{aligned} \quad (6)$$

The estimation of the uniform working correlation structure R (α) is given in (5), however, it should be noted that although the correlation is weak, suggesting parameter estimates could be given by a simpler model considering the independent correlation structure, it is recommended to maintain the approach of estimating generalized equations by interpreting the results. In the case of uniform correlation, it is understood that the order of observations within the harvest does not matter.

$$R(\alpha) = \begin{bmatrix} 1 & -0.0301 & -0.0301 & -0.0301 \\ -0.0301 & 1 & -0.0301 & -0.0301 \\ -0.0301 & -0.0301 & 1 & -0.0301 \\ -0.0301 & -0.0301 & -0.0301 & 1 \end{bmatrix}. \quad (7)$$

As mentioned, based on the values predicted by the adjusted GEE model, the association of the evaluations given in the defect sieve percentages, regarding the defects, determined by the GH biplot, the RMP biplot and the SQRT biplot obtained in the components S2014+S2015 and S2014-S2015 are discussed below. It was observed that regardless of the biplot type considered in this study, related to the harvest additive effect (S2014+S2015) (Figure 2 (A), (B), and (C)), it is clear that defect d2 (Table 1) is characterized in the classification obtained at 20% ≤ p ≤ 30%, indicating, from a practical point of view, a higher incidence of black, green, and burnt grains.

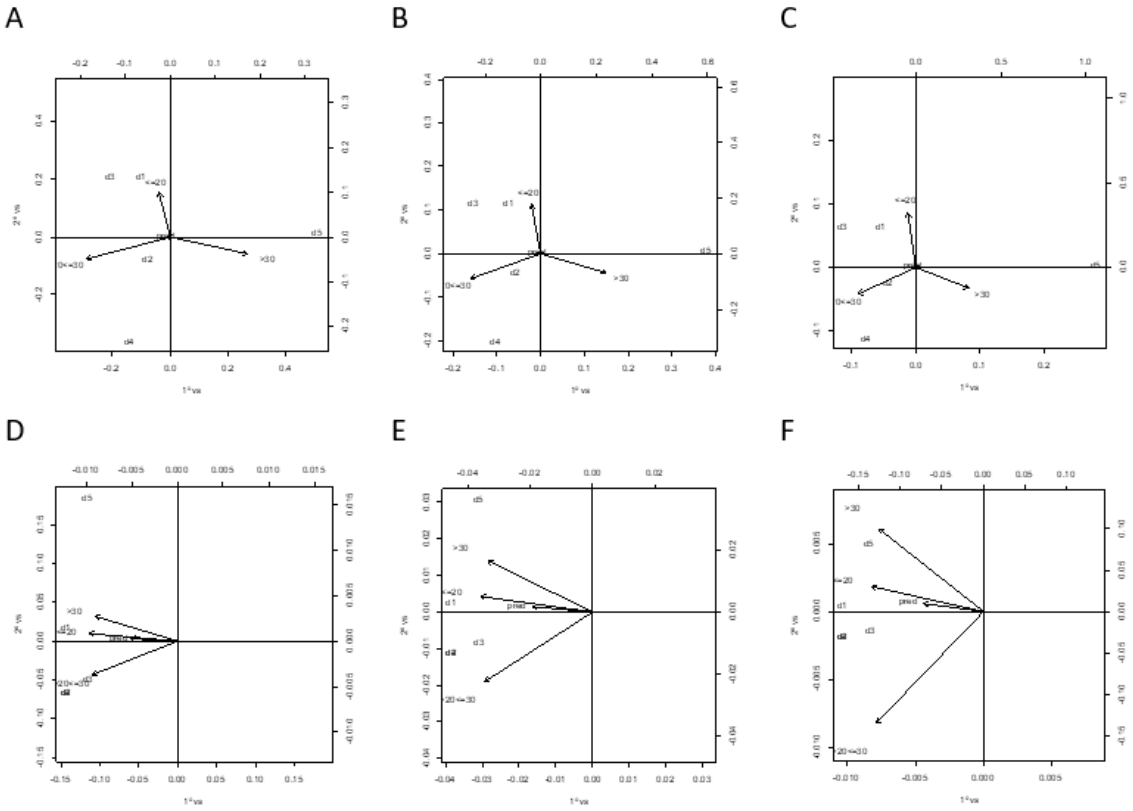

Figure 2. Biplots centered on predicted values of the GEE model with logit binding function for harvests additive effect (S2014 + S2015) and to difference effect (S2014 - S2015). A – RMP-biplot centered on predicted values of the GEE model with logit binding function for harvests additive effect; B – SQRT-biplot centered on predicted values of the GEE model with logit binding function for harvests additive effect; C – GH-biplot centered on predicted values of the GEE model with logit binding function for harvests additive effect; D – RMP-biplot biplot centered on predicted values of the GEE model with logit binding function for harvests difference effect; E – SQRT-biplot centered on predicted values of the GEE model with logit binding function for harvests difference effect; F – GH-biplot centered on predicted values of the GEE model with logit binding function for harvests difference effect.

4. Discussion

Regarding the effect of harvest differences, S2014-S2015 (Figure 2 (D), (E), and (F)), when considering the different types of biplots, respectively, it was noted that the variations in the results of the associations of defects in relation to sieve counts were distinct, which misrepresents the interpretation, by confusing associations with defects. In the case of GH-biplot, the results were not informative, so as to suggest new classifications.

Therefore, when considering the effects of association and difference between harvests, it was found that all biplots corrected by the predictions of the GEE models led to identify the grain defect (black, green, and burnt) associated with the percentage of sieve grains between 20 and 30%, which is consistent with existing literature (Brighenti and Cirillo 2018; Costa et al. 2018). In the case of the component S2014-S2015, the biplots presented different interpretations, so the results were not conclusive. Thus, it is noteworthy that for this purpose the biplots with effect of the sum of the harvests (S2014+S2015) are recommended for maintaining the same behavior.

5. Conclusions

Therefore, through this work, when considering the application resulting from the sum component S2014+S2015, it was observed that all biplots corrected by the predictions of the GEE models led to the identification of the PVA grain defect associated with the percentage of sieves between 20% and 30 %

consistent with existing literature. In the case of the S2014-S2015 component, the biplots presented different interpretations, so no conclusive results were obtained.

Authors' Contributions: FERREIRA, H.A.: analysis and interpretation of data, and drafting the article; OLIVEIRA, E.R.: drafting the article; BRIGHENTI, C.R.G.: conception and design, and acquisition of data; CIRILLO, M.A.: conception and design, acquisition of data, analysis and interpretation of data. All authors have read and approved the final version of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Ethics Approval: Not applicable.

Acknowledgments: The authors would like to thank the funding for the realization of this study provided by the Brazilian agencies CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil), and FAPEMIG (Fundação de Amparo a Pesquisa do Estado de Minas Gerais - Brasil), Finance Code APQ -0242-16.

References

BRIGHENTI, C.R.G. and CIRILLO, M.A. Analysis of defects in coffee beans compared to biplots for simultaneous tables. *Revista Ciência Agronômica*. 2018, **49**(1), 62-69. https://doi.org/10.5935/1806-6690.20180007

CENTRO DE COMERCIO DE CAFÉ DO ESTADO DE MINAS GERAIS. *Manual de Classificação: Métodos de Classificação de Café utilizados pelo CCCMG*. 2018. Available from: http://cccmg.com.br/manual-de-classificacao/

COSTA, A.L.A., BRIGHENTI, C.R.G. and CIRILLO, M.A. A new approach to simple correspondence analysis with emphasis on the violation of the independence assumption of the levels of categorical variables. *Acta Scientiarum*. 2018, **40**, 1-7. https://doi.org/10.4025/actascitechnol.v40i1.34953

ESQUIVEL, P. and JIMÉNEZ, V.M. Functional properties of coffee and coffee by-products. *Food Research Internationa*l. 2012, **46**(2), 488–495. https://doi.org/10.1016/j.foodres.2011.05.028

CHAMBERS, J.M. *Graphical methods for data analysis*. Boca Raton: CRC Press, 2018.

GREENACRE, M. Singular value decomposition of matched matrices. *Journal of Applied Statistics*. 2003, **30**(10), 1101-1113. https://doi.org/10.1080/0266476032000107132

MAIR, P. *Modern psychometrics with R*. New York: Springer International Publishing, 2018.

MOURA, W.M., et al. Genetic diversity in arabica coffee grown in potassium-constrained environment. *Ciência e Agrotecnologia*. 2015, **39**(1), 23-31. https://doi.org/10.1590/S1413-70542015000100003

WANG, L. and WEI, M.A. Improved empirical likelihood inference and variable selection for generalized linear models with longitudinal nonignorable dropouts. *Annals of the Institute of Statistical Mathematics*. 2021, **73**(3), 623-647. https://doi.org/10.1007/s10463-020-00761-4.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2016. Available from: http://www.r-project.org.

SILVA, J.A. and CIRILLO, M.A. Selection criterion of work matrix as a function of limiting estimates of the covariance matrix of correlated data in GEE. *Biometrical Journal*. 2018, **60**(5), 979-990. https://doi.org/10.1002/bimj.201800035

SOARES, W.L., et al. Qualidade do café arábica por diferentes granulometrais. *Ciência Agrícola*. 2019, **17**(1), 31-35. https://doi.org/10.28998/rca.v17i1.6495

TORRES-SALINAS, D., et al. On the use of biplot analysis for multivariate bibliometric and scientific indicators. *Journal of the American Society for Information Science and Technology*. 2013, **64**(7), 1468-1479. https://doi.org/10.1002/asi.22837

Received: 9 April 2020 | **Accepted:** 11 September 2020 | **Published:** 20 August 2021