DOI: 10.5433/1679-0359.2020v41n2p479

Proposal of a metric selection index for correspondence analysis: an application in the sensory evaluation of coffee blends

Proposta de um índice de seleção de métrica para análise de correspondência: uma aplicação na avaliação sensorial de blends de cafés

Adilson Silva da Costa[1]; Mariana Resende[2]; Eduardo Yoshio Nakano[3]; Marcelo Angelo Cirillo[4*]; Flávio Meira Borém[4]; Diego Egídio Ribeiro[5]

Highlights:
Different metrics were used in data analysis of coffee blends.
Improvement of correspondence analysis and use thereof in sensory analysis are proposed.
A simulation was employed for the choice of the metrics and its application in the selection of blends.

Abstract

Correspondence analysis is a multivariate dimensionality-reduction technique applied to data structured into contingency tables. The main outcome of this approach is the generation of perceptual maps aimed at the study of similarity between categorical levels. In most cases, interpretations of these similarities present subjectivity when different metrics are considered; e.g., Hellinger distance and Chi-square. Thus, in an attempt to minimize this subjectivity, the present study proposes an index that quantifies the shortest distance between those levels. A simulation study was undertaken in which the generated maps were discussed in relation to real data involving the similarity of blends formed by coffees of different species, with sensory evaluations considering the *flavor* and *acidity* attributes. In conclusion, the proposed index named metric selection index (MSI) made it possible to include a statistic that justifies the most suitable metric for correspondence analysis, thus preventing subjectivity in interpretations of similarities between blend types and grade classes. In the simulation studies, with the metric proposed by Hellinger distance, MSI showed stabler results regarding total inertia distribution on the first two axes.

Key words: *C. arabica*. *C. canephora.* Hellinger. Simulation.

Resumo

A análise de correspondência é uma técnica multivariada de redução de dimensionalidade aplicada a dados estruturados em tabelas de contingência. Como principal resultado, mapas perceptuais são gerados com o propósito de estudar a similaridade entre os níveis categóricos. Na maioria das vezes, as

[1] Discente de Mestrado, Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, Universidade Federal de Lavras, UFLA, Lavras, MG, Brasil. E-mail: adilsonsilvsac@hotmail.com

[2] Discente de Doutorado, Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, UFLA, Lavras, MG, Brasil. E-mail: mresende31@gmail.com

[3] Prof., Departamento de Estatística, Universidade de Brasília, UNB, Brasília, Brasil. E-mail: eynakano@gmail.com

[4] Profs., Departamento de Estatística, UFLA, Lavras, MG, Brasil. E-mail: macufla@gmail.com; flavioborem@ufla.br

[5] Pesquisador, Departamento de Engenharia, UFLA, Lavras, MG, Brasil. E-mail: diegoerib@gmail.com

[*] Author for correspondence

Received: May 06, 2019 - Approved: Oct. 22, 2019

interpretações dessas similaridades apresentam certa subjetividade, ao considerar diferentes métricas, como por exemplo, a distância de Hellinger e Qui-quadrado. Assim, com o intuito de minimizar essa subjetividade, esse trabalho teve como objetivo propor um índice que quantifique a menor distância entre esses níveis. Foi realizado um estudo de simulação, discutindo-se os mapas gerados em relação a dados reais envolvendo a similaridade de blends formados por cafés de diferentes espécies com avaliações sensoriais considerando os atributos sabor e acidez. Concluiu-se que a proposta do índice, denominado índice de seleção de métrica (ISM), permitiu agregar uma estatística que justifique a métrica mais adequada na análise de correspondência, evitando a subjetividade nas interpretações das similaridades entre os tipos de blends e classe de notas. Em relação aos estudos de simulação a métrica proposta pela distância de Hellinger, o ISM apresentou resultados mais estáveis em relação à distribuição da inércia total nos dois primeiros eixos.

Palavras-chave: *C. arábica. C. canéfora*. Hellinger. Simulação.

Introduction

One of the main activities related to commercial and industrial production of coffee and its derivatives is the formulation of blends with roasted and ground coffee of different species and quality, involving *Coffea arabica* and *Coffea canephora* beans. Each species has its peculiarities in terms of chemical and physiological composition of raw beans, which influence various sensory attributes. Additionally, external factors such as processing type and geographic denomination also interfere with those characteristics (Ramos, Ribeiro, Cirillo, & Borém, 2016; Ribeiro et al., 2016).

In blends formulated with a higher proportion of *C. arabica* beans, the beverage is more aromatic and acidic. It has a distinguished taste compared to the beverage formulated with a higher *C. canephora* concentration, which is characterized by a bitter taste and denser consistency (Santos, Rosires, Freitas, & Corrêa, 2013).

In statistical analysis, considering the categorized data represented in a contingency table structured by different grade classes in relation to a given sensory attribute as a function of the number of components that form a blend, the blend can be classified as pure, binary or ternary. In this scenario, correspondence analysis can be employed as an alternative method for data analysis.

The technique is aimed at identifying associations between the categorical levels represented by the coordinates of profiles, which are obtained by an algebraic procedure that encompasses the distance, such as the Chi-square or Hellinger statistics. The coordinates are used to represent the profiles in a two-dimensional graph called perceptual map (Cuadras & Cuadras, 2006).

In general, to generate the coordinates for the construction of perceptual maps, statistical software programs utilize the Chi-square metric (Meyners, Castura, & Car, 2013). However, this metric is not suitable for levels with low marginal frequencies (Jaeger et al., 2014). Therefore, it is important to discuss which metric should be chosen for this technique. In this respect, the improvement and use of different metrics in various applications of correspondence analysis have been investigated.

Costa, Guimarães and Cirillo (2018) used the Chi-square metric in variables derived from sensory experiments and proposed an improvement through the incorporation of Pearson's residuals in the calculation of coordinates. Rao (1995) proposed the Hellinger distance to obtain coordinates of correspondence analysis, demonstrating that this distance meets the principle of distributional equivalence, with the possibility of two similar profiles being combined into a single profile whose weight equals the sum of individual weights, pertaining to the profiles considered in the combination. Thus, as an example for the profiles of rows r_i and $r_{i'}$, of weights a_{i+}/a_{++} and $a_{i'+}/a_{++}$, where $r_i \approx r_{i'}$, naturally, $a_{i+} \approx a_{i'+}$ and $a_{ij} \approx a_{i'j'}$. Accordingly, by summing the elements of the row profiles for each column, we have a new row profile with the elements $[r_{i1}+r_{i'1};...; r_{ij+}r_{i'j}]$.

Vidal, Tárrega, Antúnez, Ares and Jaeger (2015) examined the viability of using correspondence analysis on sensory data, comparing the Hellinger and Chi-square distances on data obtained using CATA. Considering the different characteristics of products identified in the CATA technique and categorized into frequency tables, the authors concluded that the metrics showed similar results. They describe, however, that because it generated more-unstable results, the Hellinger metric might not be applicable in studies involving non-trained consumers.

It should be emphasized that the interpretation of similarities between categorical levels of variables is subjective, since there is not a statistical criterion adapted to the technique of correspondence analysis that expresses the most recommendable metric as a coefficient or index.

On these bases, the present study was conducted to propose an index that justifies the choice of the metric to be used in correspondence analysis to be applied in the sensory analysis of blends of different coffee species, given the existence of heterogeneity in sensory attributes as well as external factors related to the development of experiments. Through this study, we expect to offer an instrument that provides an interpretation of similarities between blends and grade classes with greater accuracy and reliability.

Methodology

Experimental description

The blends are described in Table 1. Further details can be found in Paulino, Cirillo, Ribeiro, Matias and Borém (2019). The procedures were approved by the ethics committee of the Federal University of Lavras (CAAE: 14959413.1.0000.5148).

The formulation of blends involved three coffee types that were distinguished according to their quality. Roasted and ground beans of the species *Coffea arabica* L., variety Bourbon Amarelo, were used to represent the high-quality coffee, deemed “special”. This coffee was evaluated by trained tasters and following the protocol for sensory analysis of coffee as established by the U.S. Specialty Coffee Association (SCAA) (Lingle, 2011). The common coffee was represented a brand sold regionally, described as containing 100% of the species *C. arabica* L. Lastly, the coffee of inferior quality was represented by roasted and ground beans of the species *Coffea canephora* variety Robusta.

Except for the commercial coffee, the blend was prepared as proposed by the SCAA (Lingle, 2011). The roasting degree of the other coffees was determined visually by using Agtron disks (SCAA/AgtronRoast Color Classification System). Each beverage was prepared using two different concentrations: 7% (70 g of ground coffee to 1000 mL of water) and 10% (100 g of ground coffee to 1000 mL of water). Thus, four experiments were carried out at 24-h intervals, involving five tasters and following the blend formulations (Table 1).

Table 1
Proportions of the blends of special Arabica (SAC), commercial C. arabica (CCA) and C. canephora (CC) coffees

Sample	SAC	CAC	CC
1 (P)	1.00	0.00	0.00
2 (B)	0.67	0.33	0.00
3 (T)	0.34	0.33	0.33
4 (B)	0.50	0.50	0.00
5 (B)	0.50	0.00	0.50
6 (B)	0.34	0.66	0.00
7 (B)	0.34	0.00	0.66
8 (P)	0.00	1.00	0.00
9 (P)	0.00	0.00	1.00

Classification of blends based on the coffee types employed: (P) a 100% proportion beverage, which may be Special Arabica, Canephora or commercial Arabica; (B) blends formed by two coffees; (T) blends formed by three coffees.

A non-structured nine-point hedonic scale was used to assign the grades referring to the *flavor* and *acidity* attributes.

Application of correspondence analysis with the Chi-Square and Hellinger metrics

The obtained data were used for the discretization of sensory evaluations, considering the classifications of blends [Pure (P), Binary (B) and Ternary (T)] and the classification of grades into four classes. The average of the grades assigned by the five evaluators was considered by combining the responses of both concentrations so as to determine a grading profile that would characterize the composition of the blends regardless of their concentration. The frequencies were obtained as described in Table 2.

Table 2
Layout of the contingency table structured by the frequency of grades, a_{ij} (i=1,..., 3 and j=1,...,4) of the blend types in relation to the grade classes (C_1,...,C_4)

	Sensory evaluation grade			
Blend	$C_1 < 3$	$3 \leq C_2 < 5$	$5 \leq C_3 < 7$	$C_4 \geq 7$
Pure (P)	a_{11}	a_{12}	a_{13}	a_{14}
Binary (B)	a_{21}	a_{22}	a_{23}	a_{24}
Ternary (T)	a_{31}	a_{32}	a_{33}	a_{34}

The metrics were specified according to the Chi-Square and Hellinger distances, considering the expression (1) proposed by Naito (2007):

$$\mathbf{M} = D_l^{\frac{1}{2}}\left(\mathrm{I}\text{-}\beta\right)\mathbf{1}l^t\left\{\frac{1}{\alpha\left(1+\beta\right)}\left(D_l^{-1}QD_c^{-1}\right)^{\alpha} - \mathbf{11}\beta\right\}\left(\left(\mathrm{I}\text{-}(1\text{-}\beta)\mathbf{1}c^t\right)^t D_c^{\frac{1}{2}}\right., \text{ where} \qquad (1)$$

D_l refers to the root square of the diagonal matrix of the marginal proportions of the categorical levels for the *blend* variable; D_c corresponds to the marginal proportions of the variables representing the grade classes; Q is the matrix of proportion corrected by the Chi-square distance usually employed in conventional correspondence analysis; lastly, I is the identity matrix.

The metric given by the Hellinger distance was specified by defining the constants α=0.5 and β=1. In the case of the metric defined by the Chi-Square distance, the defined values were given by α=1 and β=0.

For each metric, the estimate of coordinates was initially determined by applying the singular value decomposition theorem, which resulted in the normalized eigenvectors represented by U and V. Thus, the standardized coordinates were obtained by expressions (2) and (3):

$$L = D_r^{-\frac{1}{2}} U \qquad (2)$$

$$C = D_c^{-\frac{1}{2}} V \text{, where} \qquad (3)$$

D_r represents the diagonal matrix of the proportions of variables described in the row direction, that is, blends (4); and D_c is the diagonal matrix of proportions of variables described in the column direction (5), defined by the grade classes:

$$D_r = \begin{pmatrix} p_{1+} & 0 & \cdots & 0 \\ 0 & p_{2+} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_{I+} \end{pmatrix} \qquad (4)$$

$$D_C = \begin{pmatrix} p_{+1} & 0 & \cdots & 0 \\ 0 & p_{+2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_{+\mathrm{J}} \end{pmatrix} \qquad (5)$$

Based on these results, new principal coordinates referring to the blends (Y) and the columns (Z) for the grade classes, respectively, were given in (6) and (7):

$$Y = D_r^{-\frac{1}{2}} U^t C \qquad (6)$$

$$Z = D_c^{-\frac{1}{2}} V^t L \qquad (7)$$

Construction of an index to select the metric to be used in correspondence analysis

Once the Y(6) and Z(7) coordinates in the Chi-square and Hellinger distance metrics were obtained, the construction of the proposed index made it possible to determine which metric would be more suitable to explain the similarity between the categorical levels of blend types and the grade classes (Table 2). Aiming at a better understanding, the coding described in Table 3 was adopted for the described application.

Table 3
Coding of the categorical levels of blend types and grade classes

I	Blend
1	Pure (P)
2	Binary (B)
3	Ternary (T)
J	**Grade class**
1	$C_1 < 3$
2	$3 \leq C_2 < 5$
3	$5 \leq C_3 < 7$
4	$C_4 \geq 7$

Given the coding (Table 3), the Euclidean distances were obtained as follows (8):

$$d_{ij} = \sqrt{\left|Y_{1i} - Z_{1j}\right|^2 + \left|Y_{2i} - Z_{2j}\right|^2}\text{ , i=1, 2 and 3; j=1, 2, 3 and 4, where} \qquad (8)$$

Y_{1i} and Y_{2i} refer to the i-th coordinate of the categorical levels of the *blend* variable on axes 1 (abscissa) and 2 (ordinate). Analogously, Z_{1j} and Z_{2j} indicate the coordinates of the categorical levels of the *grade class* variable in relation to the same axes. Thus, considering the categorical variable defined by the blends as a reference (i=1, 2 and 3), the shortest distance calculated between the categorical levels of the *grade class* variable (j=1, 2, 3 and 4) was obtained for each blend, according to the expressions (9).

$$d^*_1=\min\{d_{11},d_{12},d_{13},d_{14}\};\ d^*_2=\min\{d_{21},d_{22},d_{23},d_{24}\};\ d^*_3=\min\{d_{31},d_{32},d_{33},d_{34}\}. \qquad (9)$$

Thus, the metric selection index (MSI), which considers the marginal proportions for each blend, referenced by p_{+i} (i=1,2,3), is given in (10).

$$MSI = \left(d_1^* \times p_{+1}\right) + \left(d_2^* \times p_{+2}\right) + \left(d_3^* \times p_{+3}\right) \qquad (10)$$

Monte Carlo simulation for a comparison of the Chi-square and Hellinger metrics in the explanation of inertia

The reduction of space dimensionality into a new plane resulted in a loss in sampling variation. To quantify this loss, the total inertia (11) concentrated in the sum of eigenvalues λ_i (i=1,...,I) was computed. The interpretation is a result of the weighted sum of the distances of points, which are identified by the coordinates of blends and grade classes in relation to the centroid.

Considering the structure of the contingency table (Table 2) defined by I=3 rows and J=4 columns, the highest number of axes to be considered in the construction of the maps is given by k=min (I,J). Thus, three eigenvalues should be estimated. However, because the categorical variables are represented in a two-dimensional plane, the proportion of total inertia (TI) explained on the first two axes is justified as described in expression (12):

$$TI=\sum_{i=1}^{I}\lambda_i \qquad (11)$$

$$I_1=\frac{\lambda_1}{TI}\ ;\ I_2=\frac{\lambda_2}{TI} \qquad (12)$$

Given the above, 1000 Monte Carlo steps were performed, simulating contingency tables with a structure similar to the real data (Table 2). The observed frequencies were defined by the random variable Y_{ij}, following the model given by the correlated binomial distribution. This probabilistic model is formed by the mixture of binomial distribution with the modified Bernoulli equation, which assumes a value of 0 or n, with probability ρ (Cirillo & Ramos, 2014). Therefore, as a scenario for the evaluation of the metrics, n_j was fixed at 50 and the (π_j) proportions of the correlated binomial model were considered 0.20, 0.50 and 0.90. For each proportion, the values for the degree of correlation (ρ) were specified as 0.20, 0.50 and 0.80.

Results and Discussion

Studies of simulation of the use of chi-square and hellinger metrics in correspondence analysis to explain inertia

The results obtained in the simulation study allowed for a comparative analysis of the performance of the Chi-square and Hellinger metrics in explaining the total inertia, to be distributed between the first two axes. The information of inertia on each axis corresponds to the estimates of eigenvalues (λ_i, i=1,2).

The proportion of sampling variation explained on the first axis (e_1), denoted by p_1, made it possible to determine whether one metric tends to concentrate most part of the sampling variation. Results are described in Table 4.

With the π_j parameter values fixed, the explained percentage of sampling variation on the first axis (p_1) was higher under the high-correlation condition (ρ=0.8), for both metrics. However, when the Hellinger metric was used, this percentage was lower than that observed with the Chi-square metric. This finding demonstrates the advantage of using the Hellinger metric, as total inertia was more evenly distributed between the first two axes, which are those typically used to construct two-dimensional perceptual maps.

In this respect, for all evaluated situations, the results obtained with the Hellinger distance were more promising than those obtained with the Chi-square metric.

Table 4
Estimates of eigenvalues (e_i, i=1, 2 and 3) and sampling variation percentage explained on the 1st axis of the metrics defined by the Chi-square and Hellinger distances used in correspondence analysis

		Chi-square			Hellinger		
(π_j)	ρ	λ_1	λ_2	p_1	λ_1	λ_2	p_1
	0.2	0.32	0.08	78.65	0.26	0.10	71.84
0.2	0.5	0.98	0.16	85.34	0.35	0.14	70.89
	0.8	3.40	0.19	94.52	0.48	0.13	78.06
	0.2	0.50	0.24	67.37	0.19	0.12	62.18
0.5	0.5	0.81	0.41	66.26	0.29	0.16	64.64
	0.8	1.19	0.54	68.77	0.39	0.18	67.93
	0.2	0.41	0.15	72.49	0.29	0.07	80.01
0.9	0.5	0.66	0.27	71.01	0.14	0.09	60.08
	0.8	0.86	0.30	73.97	0.17	0.09	63.83

Application of correspondence analysis using the Chi-square and Hellinger metrics in the study of associations of blend types, in sensory evaluations

The frequencies described in Table 5 correspond to the total grades assigned by the two tasters for the pure (P) coffees, the binary blends (B; two coffee types) and the ternary mixtures (T; three coffee types).

Table 5
Frequencies for the sensory attributes and marginal proportion of the blends

			Flavor			
Blend	$C_1 < 3$	$3 \leq C_2 < 5$	$5 \leq C_3 < 7$	$C_4 \geq 7$	Total	Proportion (Row)
Pure (P)	1	6	2	3	12	0.33
Binary (B)	0	7	13	0	20	0.55
Ternary (T)	1	1	2	0	4	0.11
Total	2	14	17	3	36	1.00
			Acidity			
Pure (P)	1	8	1	2	12	0.33
Binary (B)	2	11	7	0	20	0.55
Ternary (T)	0	2	2	0	4	0.11
Total	3	21	10	2	36	1.00

Based on the frequencies observed for each sensory attribute (Table 5), sensory analysis was carried out considering the Chi-square and Hellinger metrics. Results pertaining to variability, explained by the first two axes (Table 6), are represented by the cumulative percentage of sampling variation.

Table 6
Sampling variation percentage (cp, %) and eigenvalues (λ_i) obtained in correspondence analysis, using the Chi-square and Hellinger distances to be explained by the first two main axes

	Chi-square				Hellinger			
	Flavor		Acidity		Flavor		Acidity	
Axis	λ_i	cp (%)	λ_i	cp (%)	λ_i	cp (%)	λ_i	cp (%)
1	0.68	53.23	0.53	61.86	0.38	55.25	0.34	60.61
2	0.52	93.52	0.26	92.86	0.23	89.61	0.15	84.76
3	0.08	-	0.06	-	0.07	-	0.07	-

For both attributes, the sampling variation - explained by the total inertia computed in the cumulative percentage explained on the first two axes - was adequate, with indices higher than 80%. This result agrees with those reported by Infantosi, Costa and Almeida (2014), confirming that the study of similarity between the categorical levels of the *blend* and *grade class* variables can be represented in a two-dimensional perceptual map.

In the comparison between metrics, for this application, the Hellinger distance contributed less to explaining the sampling percentage on the first axis compared to the Chi-square metric. This result is in line with the preliminary observations made in the Monte Carlo simulation. As such, it justifies obtaining the necessary coordinates for the reproduction of maps considering both approaches, described in Table 7.

Table 7
Coordinates of the categorical levels of the *blend* and *grade class* variables for the *flavor* and *acidity* attributes considering the Chi-square and Hellinger metrics

Chi-square distance				
	Flavor		Acidity	
Blend	Coord. (Axis 1)	Coord. (Axis 2)	Coord. (Axis 1)	Coord. (Axis 2)
Pure (P)	0.80	-0.06	0.22	0.23
Binary (B)	0.53	-0.40	0.59	0.11
Ternary (T)	-0.97	-1.26	0.78	0.63
Grade	Flavor		Acidity	
$C_1 < 3$	-2.27	-0.47	0.13	-0.67
$3 \leq C_2 < 5$	0.24	-0.47	0.50	0.16
$5 \leq C_3 < 7$	0.58	-0.01	-0.14	-0.09
$C_4 \geq 7$	-0.07	1.44	-1.50	0.52
Hellinger distance				
	Flavor		Acidity	
Blend	Coord. (Axis 1)	Coord. (Axis 2)	Coord. (Axis 1)	Coord. (Axis 2)
Pure (P)	0.15	-0.08	0.30	0.12
Binary (B)	0.23	-0.04	0.39	-0.91
Ternary (T)	0.23	0.37	0.19	0.33

continue

continuation

Grade	Flavor		Acidity	
$C_1 < 3$	0.41	0.30	-0.04	0.08
$3 \leq C_2 < 5$	-0.19	0.13	-0.43	0.03
$5 \leq C_3 < 7$	-0.05	-0.12	0.05	0.25
$C_4 \geq 7$	0.46	-0.01	0.43	0.22

The perceptual maps for each attribute are illustrated in Figures 1(A) to 1(D); the coding of points is described in Table 7.

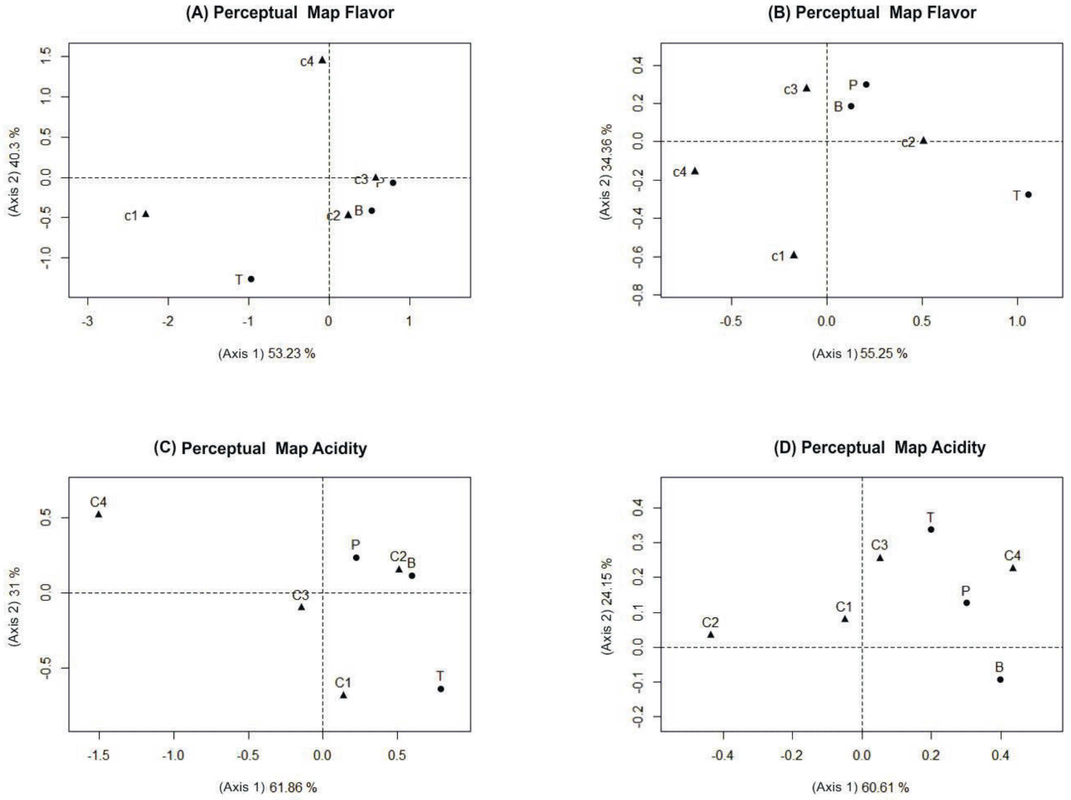

Figure 1. Perceptual map considering the evaluations of the attributes: flavor with the metric given by the Chi-square distance (A) and Hellinger (B); *acidity* with the metric given by the Chi-square distance (C) and Hellinger (D).

The similarity between the categorical levels of blend types and grade classes is interpreted by the distance between their coordinates. However, the interpretation is subjective, since there is no limit value that allows the measurement of the magnitude of these distances. Thus, after the maps with both metrics are obtained, MSI can be used as a selection criterion. Because there is no interference with formulations of tests of hypotheses associated with this index, there is no evidence of the presence of type-I or II error. Therefore, an index must be determined that synthesizes the distances between these levels, with the selection of a metric founded upon a statistical criterion for greater reliability.

Table 8 contains the distances between the coordinates of each blend and the coordinates of grade classes in relation to axis 1, which presented the highest concentration of inertia compared to the other axes. Differences were denoted as dP, dB and dT, and the MSI index was estimated in accordance with expression (10).

In general, the maps are interpreted based on the distance between the points defined by the coordinates of the categorical levels for each variable (in this case, represented by the blends and grade classes), with shorter distances meaning greater evidence that a given blend is associated with a given grade class. Agreeing with the previously described results (Table 8), the MSI estimated for each attribute made it possible to determine the most suitable metric to explain the similarity for each attribute. In this respect, because it showed a lower value, the Hellinger metric is justified for the *flavor* attribute and the Chi-square metric is appropriate for the *acidity* attribute. The respective perceptual maps are represented in Figures 1(B) and 1(D).

As regards the *flavor* attribute (Figure 1(B)), for which the Hellinger metric was justified by the MSI (Table 8), the similarity between the pure (P) blends and the binary (B) mixtures is explained, as it maintained some uniformity in the blend characteristics; i.e., the flavor referring to the binary mixture that approaches those observed for pure coffees.

Table 8
Differences between the coordinates of the first axis of blends and grade classes represented by dP, dB and dT and the metric selection index (MSI) considering the Hellinger and Chi-square metrics for the flavor and acidity attributes

			Flavor					
Distance	Chi-square				Hellinger			
dP	3.10	0.68	0.21	1.74	0.46	0.41	0.21	0.31
dB	2.80	0.29	0.39	1.94	0.39	0.46	0.29	0.23
dT	1.52	1.45	1.99	2.85	0.18	0.49	0.57	0.44
MSI	0.39				0.21			
			Acidity					
dP	0.92	0.29	0.49	1.75	0.35	0.74	0.27	0.16
dB	0.91	0.09	0.76	2.13	1.09	1.26	1.22	1.14
dT	1.46	0.55	1.18	2.29	0.35	0.70	0.16	0.26
MSI	0.21				0.67			

Considering that the blends were dry- and wet-processed and adopting the results described by Lima et al. (2011) for blend composition as a reference (eight formulations of Arábica and Conilon coffees, similarly to the present formulations), the acceptance tests indicated no difference in the averages of sensory acceptance grades assigned by the consumers. Another relevant factor is the approximation to the C3 grade category, whose scoring characterizes the coffees as recommended for consumption.

The perceptual maps for the *acidity* attribute are illustrated in Figures 1(C) and 1(D). For this attribute, the metric showed to have a great influence in relation to centroid displacement, resulting in asymmetric maps. A notable result was observed in the formation of trends between the categorical levels. When the Hellinger metric was used (Figure 1(C)), we observed the existence of a linear relationship between the categorical levels of the *blend* variable and a polynomial trend for the levels of the *grade* variable.

Regarding the similarity between grade classes, with the categorical levels of blends, based on the Chi-square metric selection (Table 8), Galli and Barbas (2004) described that acidity is one of the main attributes in the evaluation of beverage quality. In the present study, because the blends contained different proportions of Arabica and Robusta coffees, the acidity corresponding to the two species is rather distinct.

It should be stressed that the proposed index, MSI, does not assume any probabilistic model for its construction or application. It is desirable for inferential purposes such as the construction of confidence intervals and hypothesis tests. Thus, for further studies, it would be interesting to relate the coordinates obtained with both metrics through a probabilistic distribution and/or models for categorized data with the prospect of including new covariates.

Conclusion

The proposal of the metric selection index (MSI) allowed the inclusion of a statistic that justifies the most suitable metric for correspondence analysis, preventing subjectivity in the interpretation of similarities between blend types and grade classes. In the simulation studies, the metric proposed by the Hellinger distance provided a more “balanced” result compared to total inertia distribution between the first two axes.

Acknowledgments

The authors thank the National Council for Scientific and Technological Development (CNPq) for the financial support though the research productivity project no. 304974/2015-3 and for the Master’s and Doctoral fellowship grants provided to the researchers involved in the project 140242/2019-8.

References

Cirillo, M. A., & Ramos, P. S. (2014). Goodness-of-fit tests for modified mutinomial logit model. *Chilean Journal of Statistics*, *5*(1), 73-85.

Costa, A. L. A., Guimarães, C. R., & Cirillo, M. A. (2018). A new approach to simple correspondence analysis with emphasis on the violation of the independence assumption of the levels of categorical variables. *Acta Scientiarum. Technology, 40*(1), 1-8. doi: 10.4025/actascitechnol.v40i1.34953

Cuadras, C. M., & Cuadras, D. (2006). A parametric approach to correspondence analysis. *Linear Algebra and Its Applications*, *417*(1), 64-74. doi: 10.1016/j.laa.2005.10.029

Galli, V., & Barbas, C. (2004). Capillary electrophoresis for the analysis of short-chain organic acids in coffee. *Journal of Chromatography, 1032*(1-2), 299-304. doi: 10.1016/j.chroma.2003.09.028

Infantosi, A. F. C., Costa, J. C. G. D., & Almeida, R. M. V. R. (2014). Análise de correspondência: bases teóricas na interpretação de dados categóricos em ciências da saúde. *Caderno de Saúde Pública*, *30*(3), 473-486. doi: 10.1590/0102-311X00128513

Jaeger, S. R., Cadena, R. S., Torres-Moreno, M., Antúnez, L., Giménez, A., Hunter, D. C.,... Ares, G. (2014). Comparison of check-all-that-apply and forced-choice yes/no question format for sensory characterization. *Food Quality and Preference*, *35*(1), 32-40. doi: 10.1016/j.foodqual.2014.02.004

Lima, T. Fº., Della Lucia, S. M., Saraiva, S. H., Carneiro, J. C. S., & Roberto, C. D. (2011). Perfil sensorial e aceitabilidade de bebidas de café tipo expresso preparadas a partir de blends de café arábica e conillon. *Enciclopédia Biosfera*, *7*(12), 1-17. doi: 10.1590/0034-737X201562040001

Lingle, R. T. (2011). *The coffee cupper´s handbook*: systematic guide to the sensory evaluation of coffee´s flavor. 4nd ed. Long Beach: S.C.A.A. SCAAed.

Meyners, M., Castura, J. C., & Carr, B. T. (2013). Existing and new approaches for the analysis of CATA data. *Food Quality and Preference*, *30*(2), 309-319. doi: 10.1016/j.foodqual.2013.06.010.

Naito, S. D. N. P. (2007). *Análise de correspondência generalizada*. Dissertação de mestrado, Universidade de Lisboa, Lisboa, Portugal.

Paulino, A. L. B., Cirillo, M. A., Ribeiro, D. E., Matias, G. C., & Borém, F. M. (2019). A mixed model applied to joint analysis in experiments with coffee blends using the least squares method. *Revista Ciência Agronômica*, *50*(3), 1-8. doi: 10.5935/1806-6690.20190041

Ramos, M. F., Ribeiro, D. E., Cirillo, M. A., & Borém, F. M. (2016). Discrimination of the sensory quality of the Coffea arabica L. (cv. Yellow Bourbon) produced in different altitudes using decision trees obtained by the CHAID method. *Journal of the Science of Food and Agriculture*, *96*(10), 3543-3551. doi: 10.1002/jsfa.7539

Rao, C. R. (1995). A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Questiió*, *19*(1), 23-63.

Ribeiro, D. E., Borém, F. M., Cirillo, M. A., Prado, M. V. B., Ferraz, V. P., Alves, H. M. R., & Taveira, J. H. S. (2016). Interaction of genotype, environment and processing in the chemical composition expression and sensorial quality of Arabica coffee. *African Journal of Agricultural Research*, *11*(27), 2412-2422. doi: 10.5897/AJAR2016.10832

Santos, E. S. M., Rosires, D., Freitas, D. G. C., & Corrêa, F. M. (2013). Efeito de grãos conilon no perfil sensorial e aceitação de bebidas de café. *Semina: Ciências Agrárias*, *34*(5), 2297-2306. doi: 10.5433/1679-0359.2013v34n5p2297

Vidal, L., Tárrega, A., Antúnez, L., Ares, G., & Jaeger, S. R. (2015). Comparison of Correspondence Analysis based on Hellinger and chi-square distances to obtain sensory spaces from check-all-that-apply (CATA) questions. *Food Quality and Preference*, *43*(1), 106-112. doi: 10.1016/j.foodqual.2015.03.003