ANDRÉ LUÍS ALVES COSTA

ANÁLISE DE CORRESPONDÊNCIA SIMPLES COM NOVOS ESCORES E O USO DA ANÁLISE DE CORRESPONDÊNCIA MÚLTIPLA EM DADOS COMPOSICIONAIS DE GRANULOMETRIA DE GRÃOS DE CAFÉ

LAVRAS - MG

2016

ANDRÉ LUÍS ALVES COSTA

ANÁLISE DE CORRESPONDÊNCIA SIMPLES COM NOVOS ESCORES E O USO DA ANÁLISE DE CORRESPONDÊNCIA MÚLTIPLA EM DADOS COMPOSICIONAIS DE GRANULOMETRIA DE GRÃOS DE CAFÉ

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Orientador
Dr. Marcelo Ângelo Cirillo
Coorientadora
Dra. Carla Regina Guimarães Brighenti

LAVRAS - MG
2016

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Costa, André Luís Alves.
Análise de correspondência simples com novos escores e o uso da análise de correspondência múltipla em dados composicionais de granulometria de grãos de café / André Luís Alves Costa. – Lavras : UFLA, 2016.
94 p. : il.

Dissertação (mestrado acadêmico)–Universidade Federal de Lavras, 2016.
Orientador(a): Marcelo Ângelo Cirillo.
Bibliografia.

1. binomial correlacionada. 2. hipótese de independência. 3. dados composicionais. 4. transformações logarítmicas. I. Universidade Federal de Lavras. II. Título.

ANDRÉ LUÍS ALVES COSTA

ANÁLISE DE CORRESPONDÊNCIA SIMPLES COM NOVOS ESCORES E O USO DA ANÁLISE DE CORRESPONDÊNCIA MÚLTIPLA EM DADOS COMPOSICIONAIS DE GRANULOMETRIA DE GRÃOS DE CAFÉ

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 26 de agosto de 2016.

Dra.Gislene Araújo Pereira	UNIFAL
Dra.Izabela Regina Cardoso de Oliveira	UFLA
Dr. Augusto Ramalho de Morais	UFLA

Orientador
Dr. Marcelo Ângelo Cirillo
Coorientadora
Dra. Carla Regina Guimarães Brighenti

LAVRAS - MG

2016

A Deus acima de tudo, pois sem ele nada é possível.
Aos meus pais Renato e Erenice que de um modo especial contribuíram para a realização desse trabalho.
DEDICO.

AGRADECIMENTOS

Inicio meus agradecimentos por DEUS, que todos os dias da minha vida me deu forças para nunca desistir, e por Ele ter colocado pessoas tão especiais a meu lado, sem as quais certamente não teria dado conta!

A meus pais, Renato e Erenice, meu infinito agradecimento. Sempre acreditaram em minha capacidade e sempre estiveram presentes na minha vida. Isso só me fortaleceu e me fez tentar, não ser o melhor, mas a fazer o melhor de mim. Obrigado pelo amor incondicional! E por sempre aceitarem e apoiarem minhas escolhas. AMO VOCÊS!

A meu irmão Adriano que a seu modo, sempre se orgulhou de mim e confiou em meu trabalho. Obrigado pela confiança! AMO VOCÊ!

A minha irmã Renata, pelo amor, carinho, torcida, e por me mostrar que a vida independente das provações que ela oferece, é muito especial. Estou sempre ao seu lado para o que precisar. AMO VOCÊ!

A toda minha família pelo carinho, conselhos, confiança, e por toda ajuda nos momentos de dificuldade.

Ao Rafael por todo carinho de sempre, lealdade e amizade!

Aos Professores Cirillo e Carla, sempre disponíveis e dispostos a ajudar, querendo que eu aproveitasse cada segundo dentro do mestrado para absorver algum tipo de conhecimento. Vocês não foram somente orientador e coorientadora, mas em alguns momentos conselheiros e amigos. Vocês são referências profissionais e pessoais para meu crescimento. Tenho uma admiração muito grande por vocês. MUITO OBRIGADO!

Agradeço a todos os meus amigos de Corinto, em especial Taci, Nayara, Luís Otávio, Ricardo, Romanine, Marcela, Lívia, Joice, Wellington, Ely, Tom, Felipe, Marcelinho e Lindamara. Amizade de mais de 20 anos. Como eu sempre digo: Nada como amigos a quem você pode chamar de irmãos. Muito obrigado por tudo vocês são minha segunda família. Aos novos amigos que a vida vai nos proporcionando Carola, Delaine, Sapinho, Hugo, Aline, espero que estejam sempre presentes em minha vida! Adoro vocês.

A Kmofinha Érica, pela amizade, conselhos e lealdade de sempre. Você é uma irmã que a vida me deu, quero você sempre presente em todos os momentos da minha vida! Amo você.

A Isis e Telma, pelas conversas, risadas e carinho. Amo vocês Kmofinhas!

Agradecimento especial a Jack, por toda ajuda, conversas, estudos e carinho comigo. Que sua vida seja repleta de realizações! Sucesso no seu doutorado.

Agradecimento especial ao Sergio e Taís, por todo companheirismo nesses dois anos de Mestrado. Muito sucesso para vocês.

A todos os meus amigos e professores da Estatística UFOP.

A todos os colegas e professores do Dex UFLA, que de alguma forma contribuíram para esse trabalho.

Enfim, como já dizia Anitelle: "Sonho parece verdade quando a gente esquece de acordar". Hoje vivo uma realidade que parece um sonho, mas foi preciso muito esforço, determinação, paciência, perseverança, para chegar até aqui e, tenho certeza, nada disso eu conseguiria sozinho.

"Mudam-se os tempos, mudam-se as vontades,
Muda-se o ser, muda-se a confiança,
Todo o mundo é composto de mudança,
Tomando sempre novas qualidades."
(Luís de Camões)

RESUMO GERAL

Na composição deste trabalho estão presentes três tópicos. O primeiro tópico contém a fundamentação teórica do presente estudo, sobre os métodos da análise de correspondência empregado para o desenvolvimento desse trabalho. O segundo tópico contém um artigo científico, onde é proposta uma nova abordagem com incorporação de resíduos, para o cálculo das coordenadas da análise de correspondência simples, mediante a tabelas de contingência em que categorias apresentam diferentes níveis de correlação, utilizando simulação Monte Carlo na geração de frequências provenientes da distribuição binomial correlacionada $BC(n, \pi, \rho)$. Concluiu-se nesse primeiro artigo que em todos os cenários avaliados a abordagem é promissora, no sentido que os objetos foram melhor discriminados em relação a abordagem convencional. O terceiro e último tópico contém o segundo artigo científico, que aborda a aplicação de análise de correspondência múltipla a dados composicionais, para um estudo comparativo do efeito de transformações logarítmicas realizadas nos dados originais, sobre a granulometria de grãos de café. Concluiu-se que a utilização das transformações logarítmicas é adequada para a análise de dados composicionais utilizando análise de correspondência múltipla. Além disso, dentre as transformações utilizadas no presente trabalho, a transformação logarítmica isométrica foi a que discriminou mais amostras de café em relação as categorias dos componentes.

Palavras-chave: binomial correlacionada, hipótese de independência, transformações logarítmicas, dados composicionais.

GENERAL ABSTRACT

Three topics are presented in the composition of this work. The first one contains the theoretical basis of this study on the methods of correspondence analysis used for its development. The second topic contains a scientific article where a new approach on residuals incorporation is proposed to calculate the coordinates of the simple correspondence analysis by contingency tables in which categories have different levels of correlation, using Monte Carlo simulation in generation of frequencies from the correlated binomial distribution $BC(n, \pi, \rho)$. The first article led to the conclusion that in all scenarios this approach is promising in the sense that the subjects were better discriminated when compared to the conventional approach. The second scientific article, which discusses the application of multiple correlation analysis of compositional data for a comparative study of logarithmic transformation effects performed on the original data to a study on the granulometry of the coffee beans, is presented in the third and last point. The use of logarithmic transformation was found suitable for compositional data analysis using multiple correspondence analysis. Among the transformations used in this study, the isometric logarithmic was the one able to discriminate most coffee samples in relation to the categories of the components.

Keywords: Related binomial, independence hypothesis, logarithmic transformations, compositional data.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO GERAL

A análise de correspondência (AC) é um técnica multivariada aplicada às tabelas de contingência de duas ou mais entradas, usada para estudar as relações entre os níveis das variáveis categorizadas, e medir o grau de associação entre elas por meio de um gráfico denominado mapa perceptual.

A análise de correspondência pode ser segmentada em dois formatos, análise de correspondência simples (ACS) análise de correspondência múltipla (ACM). A análise de correspondência simples é aplicável principalmente na análise de dados apresentados na forma de tabelas de dupla entrada, levando a um mapa que facilita a visualização da associação entre duas variáveis categóricas.

Em síntese, a análise de correspondência múltipla é simplesmente a generalização da análise de correspondências simples para tabelas com dimensão igual ou superior a três, ou seja, consiste na aplicação do algoritmo da ACS a matrizes de dados categorizados com mais de duas variáveis. Ambas as análises têm a mesma interpretação nos resultados, porém ocorrem diferenciações nos cálculos algébricos envolvidos na obtenção dos escores.

Em se tratando da natureza dos dados, segundo Aitchison (1986), os dados composicionais são um conjunto de vetores denominados composições que representam frações de algum todo e satisfazem a restrição de que a soma dos componentes é uma constante. Contudo, dada essa restrição, o uso de técnicas multivariadas usuais torna-se inviável, sendo necessária a utilização de transformações logarítmicas.

Existem três principais transformações logarítmicas utilizadas em análise de dados composicionais. Aitchison (1986) menciona duas, a transformação logarítmica aditiva (alr) e a transformação logarítmica centrada (clr). A terceira proposta por Pawlowsky et al. (2010) é a transformação logarítmica isométrica (ilr).

Convém ressaltar que, embora a técnica da análise de correspondência seja uma técnica consagrada em inúmeras situações, nota-se ainda uma abertura para novas metodologias que agreguem informações ao cálculo dos escores. A título de exemplificação Beh (2012) propõe a utilização da ACS usando resíduos

ajustados. Baxter, Cool e Heyworth (1990) fizeram um estudo sobre a similaridade entre a análise de componentes principais e análise de correspondência em dados composicionais.

Diante do exposto, os objetivos a serem contemplados nesse trabalho são divididos em dois artigos. Desta forma, enuncia-se que o artigo 1, em síntese, tem por objetivo propor novos escores em situações que violam suposição de independência entre os níveis das variáveis categóricas. O artigo 2 consiste em uma aplicação da análise de correspondência a dados composicionais, com ênfase na granulometria de grãos de café, em comparação a diferentes transformações.

2 REFERENCIAL TEÓRICO

2.1 Tabelas de contingência

Utiliza-se uma tabela de contingência para organizar as informações e analisar o relacionamento entre duas ou mais variáveis, obtendo-se a frequência de uma variável em função das categorias de outra variável. Seguindo o *layout* descrito na Tabela 1 n_{ij} representa as frequências obtidas pela interação entre cada nível da variável A, indicadas no sentido linhas, com os níveis da variável B, indicadas no sentido coluna.

Tabela 1 Estrutura de uma tabela de contingência com I linhas e J colunas.

A	B						
	1	2	$\cdots$	j	$\cdots$	J	Total
1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	n_{i+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	n_{I+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+j}	...	n_{+J}	$n_{++} = n$

Assim, dadas I linhas e J colunas, resultantes da classificação das n_{++} (total geral) observações de uma amostra ou de uma população, temos a seguinte notação:

- n_{ij} representa a frequência observada das categorias de linhas e colunas denotadas, respectivamente, pela *i*-ésima categoria da variável A e *j*-ésima categoria da variável B;
- n_{i+} representa os totais marginais das linhas da categoria A;
- n_{+j} representa os totais marginais das colunas da categoria B;
- $n_{++} = n$ representa o total geral;

- $n_{i+} = \sum_{j=1}^{J} n_{ij}$, $n_{+j} = \sum_{i=1}^{I} n_{ij}$, $n_{++} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$.

2.2 Análise de correspondência simples

A análise de correspondência simples é uma técnica estatística multivariada usada para estudar a relação entre variáveis categorizadas, visando medir o grau de associação entre elas e obter uma representação gráfica multidimensional da dependência entre as linhas e/ou colunas de uma tabela de contingência.

De acordo com Greenacre (2010) o objetivo desta técnica é mostrar geometricamente as variáveis, suas categorias e os objetos observados na base de dados em um espaço de baixa dimensão, de modo que a proximidade no espaço indique associação entre as linhas e colunas.

Entre os inúmeros resultados que esta técnica proporciona, destacam-se os mapas perceptuais. Neles, as relações entre as variáveis categóricas são visualizadas por meio de distâncias, de tal forma que cada objeto é identificado por uma posição espacial, refletindo a relativa similaridade ou preferência em relação a outros objetos, segundo as dimensões nas quais o mapa foi construído (SANTOS, 2012).

A obtenção dos escores, bem como as medidas que expressam a qualidade da representação de cada objeto, das distâncias, das contribuições e correlações a cada eixo, inicia-se com a geração de uma matriz de frequências relativas, conhecida com matriz de correspondência **P**. Esta matriz $\mathbf{P} = p_{ij} = \frac{n_{ij}}{n}$ não altera as relações proporcionais entre as linhas e as colunas (NAITO, 2007).

Na matriz de correspondências, cada frequência observada n_{ij} (Tabela 1) é transformada em uma proporção conforme ilustrado na Tabela 2, onde cada p_{ij} representa uma proporção da frequência n_{ij} em relação ao total n. Logo, os totais de cada linha e cada coluna são dados por $p_{i+} = \sum_{j=1}^{J} p_{ij}$ e $p_{+j} = \sum_{i=1}^{I} p_{ij}$. A formalização das proporções marginais pode ser representada por vetores, sendo então denominados vetores de massas das linhas $\tilde{\boldsymbol{r}}$ e colunas $\tilde{\boldsymbol{c}}$, em que:

$$\tilde{\boldsymbol{r}} = \boldsymbol{P}\tilde{\boldsymbol{j}} = (p_{1+},p_{2+},\ldots,p_{I+})' = (\tfrac{n_{1+}}{n},\tfrac{n_{2+}}{n},\ldots,\tfrac{n_{I+}}{n})$$

$$\tilde{\boldsymbol{c}} = \tilde{\boldsymbol{j}}' \boldsymbol{P} = (p_{+1}, p_{+2}, \ldots, p_{+J}) = (\tfrac{n_{+1}}{n}, \tfrac{n_{+2}}{n}, \ldots, \tfrac{n_{+J}}{n})$$

em que $\tilde{\boldsymbol{j}}$ é um vetor $1 \times p$ de $1's$.

Tabela 2 Matriz de Correspondência

A	B						
	1	2	$\cdots$	j	$\cdots$	J	Total
1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	p_{1J}	p_{1+}
2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$	p_{2J}	p_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	p_{iJ}	p_{i+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	p_{I1}	p_{I2}	$\cdots$	p_{Ij}	$\cdots$	p_{IJ}	p_{I+}
Total	p_{+1}	p_{+2}	$\cdots$	p_{+j}	...	p_{+J}	$p_{++} = 1$

A interpretação dada a esses vetores é verificada como as frequências marginais da tabela de contingência interpretadas como pesos para um perfil, de modo que para cada linha i associa-se um vetor de probabilidades condicionais, assim como para cada coluna j. Esses vetores são denominados de perfil de linha e perfil de coluna respectivamente (ZIEGEL, 1993). Dessa forma, o i-ésimo perfil de linha, $\tilde{\boldsymbol{r}}_{\boldsymbol{i}}^{\boldsymbol{t}}$, $i = 1,2,\ldots,I$, é obtido por:

$$\tilde{\boldsymbol{r}}_{\boldsymbol{i}}^{\boldsymbol{t}} = (\tfrac{p_{i1}}{p_{i+}}, \tfrac{p_{i2}}{p_{i+}}, \ldots, \tfrac{p_{iJ}}{p_{i+}}) = (\tfrac{n_{i1}}{n_{i+}}, \tfrac{n_{i2}}{n_{i+}}, \ldots, \tfrac{n_{iJ}}{n_{i+}}).$$

Similarmente, o j-ésimo perfil de coluna $\tilde{\boldsymbol{c}}_{\boldsymbol{j}}$, $j = 1,2,\ldots,J$ é representado por:

$$\tilde{\boldsymbol{c}}_{\boldsymbol{j}} = (\tfrac{p_{1j}}{p_{+j}}, \tfrac{p_{2j}}{p_{+j}}, \ldots, \tfrac{p_{Ij}}{p_{+j}}) = (\tfrac{n_{1j}}{n_{+j}}, \tfrac{n_{2j}}{n_{+j}}, \ldots, \tfrac{n_{Ij}}{n_{+j}}).$$

Cada elemento do vetor $\tilde{\boldsymbol{r}}_{\boldsymbol{i}}^{\boldsymbol{t}}$ representa a probabilidade de ocorrer o evento j condicional ao evento i. Interpretação análoga pode ser feita para $\tilde{\boldsymbol{c}}_{\boldsymbol{j}}$. Esses conceitos de perfis de linha e coluna são importantes quando se deseja comparar as linhas entre si ou ainda as colunas entre si. Neste contexto, faz-se necessário transformar a matriz $\boldsymbol{P}$ objetivando-se eliminar a influência das suas respectivas marginais (NAITO, 2007). Para que se possa fazer uma análise, sem que a magnitude (frequências absolutas) das categorias influencie nas comparações, deve-se

fazer uma tabela com as frequências condicionais, isto é, comparando as frequências relativas com as respectivas frequências marginais.

Como a ideia básica da ACS é a representação gráfica das relações (correspondências) entre as categorias das variáveis envolvidas, faz-se necessário representar os perfis de linha e coluna, ou seja, as categorias de linha e coluna em um mesmo espaço multidimensional, já que as componentes de cada vetor seguem uma distribuição multinomial condicional ao total da linha ou coluna e sua soma é 1, o que gera uma dependência linear entre tais componentes.

Assim, o vetor $\tilde{\boldsymbol{r}}_{\boldsymbol{i}}^{\boldsymbol{t}}$ representa uma realização da distribuição multinomial, condicionada à *i*-ésima categoria da variável A dada por:

$$P(\tilde{r}_i^t) = \frac{n!}{r_1^t!,\ldots,r_I^t!}\theta_1^{r_1^t},\ldots,\theta_I^{r_I^t}, \quad r_i^t = 1,\ldots,I, \;\; \sum_{i=1}^{I} r_i^t = n$$

Analogamente, cada vetor $\tilde{\boldsymbol{c}}_{\boldsymbol{j}}$ representa uma realização da distribuição multinomial, condicionada à *j*-ésima categoria da variável B.

Assim sendo, cada um desses pontos vetoriais (perfil de linha ou coluna) podem ser representados em uma dimensão menor do que originalmente foram projetados, ou seja, dimensão $J - 1$para representar os perfis de linha e dimensão $I - 1$ para representar os perfis de coluna (SILVA, 2012).

A cada elemento é associada uma massa que é a frequência marginal correspondente, equivalente ao centroide da nuvem dos elementos de A ou B respectivamente.

O centroide ou perfil médio de linha pode ser determinado pelo próprio vetor de massa de coluna e, de forma similar, o perfil médio de coluna da nuvem pode ser determinado pelo próprio vetor de massa de linha. Portanto, é natural compreender que a massa afeta o centroide, assim, pontos com massa grande, deslocam o centroide na sua direção, originando mapas perceptuais assimétricos.

O centroide da nuvem representa o ponto que é esperado se as variáveis da linha e da coluna da tabela de contingência forem independentes. Portanto, medir a distância que cada ponto (perfil de linha ou coluna) de seu respectivo centroide significa quantificar a dispersão (SILVA, 2012).

Uma menor ou maior distância entre os pontos de uma nuvem pode indicar, respectivamente, relacionamento ou não entre as categorias das variáveis. A distância entre dois perfis de linha e dois perfis de colunas é definida por meio da

métrica do qui-quadrado, isto é, recorrendo à distância euclidiana ponderada pelo inverso da frequência relativa correspondente a cada termo, dado respectivamente por:

$$d^2(i,i^{'}) = \sum_{j=1}^{J} \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i^{'}j}}{p_{i^{'}+}} \right)^2$$

e

$$d^2(j,j^{'}) = \sum_{i=1}^{I} \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij^{'}}}{p_{+j^{'}}} \right)^2$$

Assim, $d^2(i,i')$ e $d^2(j,j')$ referem-se à distância euclidiana ponderada entre dois perfis de linha e dois perfis de coluna respectivamente.

A distância euclidiana é uma medida absoluta da dispersão de uma coordenada de um perfil em relação à respectiva coordenada do centroide, sem levar em consideração a dispersão relativa que essa distância representa (SILVA, 2012). A métrica entre os pontos da nuvem e o seu respectivo centro de gravidade é dada pela distância euclidiana ponderada, também denominada distância de qui-quadrado.

Uma outra justificativa para o uso da distância de qui-quadrado como métrica entre as distâncias dos pontos é que a mesma satisfaz a propriedade da equivalência distribucional (GREENACRE, 1992). Este princípio permite que a inércia da nuvem de perfis linha seja a mesma da nuvem de perfis colunas, onde a agregação de dois perfis de linha não altera a distância entre duas quaisquer categorias de B e a agregação de dois perfis de coluna não altera a distância entre duas quaisquer categorias de A. Um exemplo sobre o princípio da equivalência distribucional encontra-se no Anexo A . Logo, se duas linhas de uma matriz são proporcionais, podemos dizer que elas tem os mesmos perfis pelo princípio da equivalência distribucional demonstrado por Escofier (1978).

Como a diferença dos perfis linha ou colunas aos seus respectivos centroides é contemplada por meio da distância χ^2, considerando a matriz de correspondência $\boldsymbol{P}$, reproduz a matriz $\boldsymbol{W}$, definida por:

$$\boldsymbol{W} = \left[\frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}} \right],$$

em que $p_{i+} = \sum_{J}^{j=1} p_{ij}$ e $p_{+j} = \sum_{I}^{i=1} p_{ij}$.

Posteriormente, aplica-se a decomposição dos valores singulares (DVS), para obtenção das coordenadas (escores) de linhas ou colunas, de tal forma que $\boldsymbol{W}'\boldsymbol{W}$ corresponde à matriz de covariância associada ao perfil linha e $\boldsymbol{W}\boldsymbol{W}'$ representa a matriz de covariância associada ao perfil coluna.

O objetivo final da análise de correspondência é reduzir a dimensionalidade da nuvem de dados com o mínimo possível de perda de informação, ou seja, retendo o máximo possível da inércia (variabilidade).

A redução da dimensionalidade na análise de correspondência consiste em obter um sistema de coordenadas para representação dos objetos em uma projeção plana, para isso a inércia total da distribuição dos objetos no espaço multiplano poderá ser decomposta como inércias parciais para cada dimensão obtida (BLASUS et al., 2009). Dessa forma, tem-se o conhecimento da quantidade da inércia total que está sendo explicada por uma dada dimensão.

A inércia total é o percentual da variância que é explicado pela aplicação da análise de correspondência e corresponde à soma ponderada das distâncias dos pontos do conjunto a seu centroide, dada por:

$$IT = \sum_{i=1}^{I} \lambda_i^2,$$

em que λ_i são os autovalores não nulos da diagonal da matriz $\mathbf{\Lambda}$, $i = 1,2,\ldots,I$.

A inércia total é uma medida utilizada para indicar a variabilidade dos dados no espaço determinado pelas dimensões. Ela é decomposta nos componentes de linha e coluna ao longo das dimensões principais. A análise destes componentes de inércia tem grande importância na interpretação na AC, pois fornece um diagnóstico que permite ao pesquisador identificar quais pontos tem maior contribuição para as dimensões principais (NASCIMENTO, 2011). A i-ésima coordenada principal tem a sua proporção de explicação em relação a inércia total (IT) dada por:

$$IT = \frac{\lambda_i^2}{\sum_{i=1}^{I} \lambda_i^2}.$$

A decomposição da inércia total inicia-se com a determinação dos autovalores $\boldsymbol{W}'\boldsymbol{W}$ e $\boldsymbol{W}\boldsymbol{W}'$ e seus respectivos autovetores $\boldsymbol{U}$ e $\boldsymbol{V}$. A matriz de $\boldsymbol{U}$ e

a matriz de $\boldsymbol{V}$ são ortogonais e contém os autovalores das matrizes de covariâncias $\boldsymbol{W^tW}$ e $\boldsymbol{WW^t}$, respectivamente. Com essas especificações, a normalização usada para distribuir a inércia pelos escores das linhas e das colunas é dada, respectivamente, por :

$$L = D_r^{-\frac{1}{2}}U$$
$$C = D_c^{-\frac{1}{2}}V$$

com as matrizes diagonais dos vetores massa de linha e de coluna $\boldsymbol{D_r}$ e $\boldsymbol{D_c}$, definidos por:

$$D_r = \begin{bmatrix} p_{1+} & 0 & \cdots & 0 \\ 0 & p_{2+} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & p_{I+} \end{bmatrix} D_c = \begin{bmatrix} p_{+1} & 0 & \cdots & 0 \\ 0 & p_{+2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & p_{+J} \end{bmatrix}$$

Com base nesses resultados, as coordenadas das linhas são dadas por:

$$Y = D_r^{-1}W^tC$$

e, analogamente, as coordenadas das colunas são dadas por:

$$Z = D_c^{-1}W^tL.$$

2.3 Análise de correspondência múltipla

A análise de correspondência múltipla pode ser considerada uma extensão da ACS quando estão envolvidas mais de duas variáveis, cuja dimensionalidade está ligada às categorias de cada variável. A aplicação é feita em tabelas multidimensionais, onde as linhas representam os objetos observados e as colunas as diferentes categorias de diferentes variáveis. A representação gráfica pode ser feita para os indivíduos, para as variáveis, para as categorias e para as categorias e indivíduos (ROUX; ROUANET, 2004).

A técnica permite a visualização de relações que normalmente não seriam reveladas por comparação de variáveis par a par. A ACM tem a capacidade de incorporar e ordenar um grande número de indicadores categóricos, permitindo a redução da dimensionalidade das entradas (TRAMARIN et al., 1997).

Um dos objetivos da ACM é classificar os indivíduos em diferentes tipos a partir do conceito de semelhança. Assim, indivíduos são considerados similares quando apresentam características em comum, ou seja, possuem uma representação próxima no espaço euclidiano gerado pela ACM (GREENACRE, 1984).

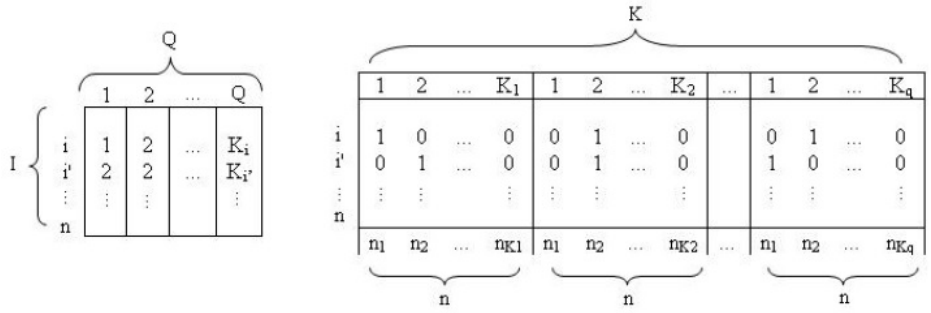
Essencialmente, enquanto na ACS a tabela de contingência é usada, na ACM a matriz de dados pode ser representada por duas maneiras: matriz indicadora Z ou matriz de Burt, $B = Z^{'}Z$. A matriz de Burt consiste em transformar a matriz retangular em uma matriz quadrada simétrica composta por tabelas de contingência bidimensionais realizando todos os cruzamentos possíveis entre as variáveis envolvidas.

Segundo Naito (2007) as duas maneiras são equivalentes. Os gráficos resultantes via matriz indicadora e via matriz de Burt são análogos, divergindo apenas na escala das variáveis.

A matriz indicadora tem em suas linhas os indivíduos (objetos) e nas colunas as categorias das variáveis. Cada linha contém todos os códigos correspondentes às categorias atribuídas a um indivíduo para cada uma das variáveis observadas. As categorias devem ser mutuamente exclusivas e exaustivas, isto é, cada indivíduo deve possuir uma e somente uma categoria para cada variável.

Os elementos são codificados como variáveis *dummy* na matriz indicadora, ou seja, 1 para a categoria escolhida como resposta de uma variável e 0 para as demais categorias da mesma variável, conforme ilustrado pela Figura 1. Nessa figura I representa o grupo de n indivíduos, Q o grupo de variáveis, K_q o número de categorias da questão q, K o número de total de categorias e K_i representa o grupo das Q categorias escolhidas pelo indivíduo i.

Figura 1 Representação de uma matriz indicadora



Fonte: (ROUX; ROUANET, 2010)

A Tabela 3 representa um exemplo de uma matriz indicadora, onde são considerados 6 indivíduos e três variáveis (V_1, V_2 e V_3), portanto $Q = 3$, sendo três categorias na primeira variável ($V_1 = \{v_1,v_2,v_3\}$), portanto $K_{V_1} = 3$, três na segunda variável ($V_2 = \{v_4, v_5, v_6\}$), logo ($K_{V_2} = 3$), e duas categorias na terceira variável ($V_3 = \{v_7, v_8\}$), portanto $K_{V_3} = 2$, temos então um total de $K = 8$ categorias. Utilizando a matriz indicadora (Tabela 3), pode-se ter a resposta de todos as observações, por exemplo, a observação 1 tem as seguintes respostas: $V_1 = v_3$; $V_2 = v_4$; $V3 = v_7$, e assim sucessivamente podemos obter as respostas de todas as observações.

Tabela 3 Matriz Indicadora Genérica

Observações	Variáveis e suas categorias								Total de Linha
	V_1			V_2			V_3		
	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	
1	0	0	1	1	0	0	1	0	3
2	1	0	0	0	1	0	1	0	3
3	1	0	0	0	1	0	1	0	3
4	0	1	0	0	0	1	0	1	3
5	0	0	1	0	1	0	0	1	3
6	0	0	1	0	0	1	1	0	3
Total	2	1	3	1	3	2	4	2	18

Seja $i = 1,2,\cdots,n$ e $k = 1,2,...,K$ a linha marginal da matriz indicadora Z, é a soma das observações na k-ésima coluna, que representa o número de

indivíduos que possuem a categoria k:

$$z_{+k} = \sum_{i=1}^{n} z_{ik}$$

A coluna marginal de Z é a soma das observações da i-ésima linha:

$$z_{i+} = \sum_{k=1}^{K} z_{ik}$$

que representa o número de variáveis, sendo $z_{i+} = Q \; \forall \; i = 1,2,\cdots,n$.

O total geral que representa o número de indivíduos (objetos) multiplicado pelo número de variáveis é dado por:

$$z_{++} = \sum_{i=1}^{n} z_{i+} = \sum_{k=1}^{K} z_{+k} = nQ.$$

O vetor de frequências relativas marginais é denominado como o vetor contendo todas as frequências totais, podendo ser de linha e de coluna. Esses vetores são conhecidos como massas e servem para normalizar as contribuições das linhas ou colunas, respectivamente, em função da distância euclidiana ponderada (BLASUS et al., 2009).

A massa de um elemento $i \in I$ é o quociente do total da *i*-ésima linha pelo total geral:

$$r_i = \frac{z_{i+}}{z++} = \frac{Q}{nQ} = \frac{1}{n}$$

e depende da quantidade de indivíduos, sendo a mesma para todo $i = (1,2,\cdots,n)$. A massa de um elemento $k \in K$ é o quociente do total da *k*-ésima coluna pelo total geral:

$$c_k = \frac{z_{+k}}{z++} = \frac{n_k}{nQ}.$$

que corresponde ao quociente do número de indivíduos que escolheram a k-ésima categoria pelo número total de indivíduos multiplicado pelo número de variáveis.

Um perfil do vetor de linha i que tem k elementos e um perfil do vetor de categoria k que tem i elementos, são dados respectivamente por:

$$r_k^i = \frac{z_{ik}}{z_{i+}} = \frac{z_{ik}}{Q}$$

$$c_I^k = \frac{z_{ik}}{z_{+k}} = \frac{z_{ik}}{n_k}$$

Segundo Roux e Rouanet (2010) os perfis de colunas podem ser considerados como uma primeira quantificação das categorias das variáveis qualitativas uma vez que, os valores são pesos relativos de cada categoria dentro da respectiva variável. Assim o perfil marginal linha corresponde à massas dos elementos $k \in K$, e o perfil marginal da coluna corresponde às massas dos elementos $i \in I$.

Assim como na análise de correspondência simples, a ACM é baseada na observação de uma "nuvem"de pontos, que é definida como um conjunto finito de pontos em um espaço geométrico (ROUX; ROUANET, 2010). O procedimento da ACM gera duas nuvens de pontos. Uma referente às linhas, nuvem dos I pontos dos indivíduos, e a outra referente às colunas, nuvem dos K pontos das categorias.

A nuvem de pontos dos indivíduos é formada a partir do conjunto dos perfis de cada linha, cada um associado à sua massa. Da mesma forma, a nuvem de pontos das categorias é formada a partir do conjunto de perfis de cada coluna, cada um associado à sua massa (SOUZA, 2004). A nuvem de pontos é projetada em planos que possuam a capacidade de preservar, ao máximo, a distância entre eles, mantendo assim a maior parte da informação original.

O ponto médio ou centroide G da nuvens de pontos, consiste no somatório das distâncias entre um ponto P qualquer e todos os pontos da nuvem dividido pelo número total de pontos. Ou seja G pode ser definido como as médias das coordenadas dos pontos, sendo que o ponto médio não depende da escolha do ponto P.

A distância entre dois indivíduos $d_q(i,i')$ somente pode ser calculada quando ambos escolhem diferentes categorias de uma mesma variável, pois quando a mesma categoria é escolhida os dois pontos coincidem na nuvem de pontos, sendo a distância nula. Segundo Roux e Rouanet (2010) a distância entre dois indivíduos é encontrada pelas diferentes escolhas de categoria para cada variável q denotada por $d_q(i,i')$, se os dois indivíduos escolherem a mesma categoria tem-se que $d_q(i,i') = 0$.

A distância ao quadrado entre dois indivíduos que escolheram categorias diferentes, ou seja, i escolheu a categoria k e o i' escolheu a categoria k', é dada

por:

$$d_q^2(i,i') = \frac{1}{f_k} + \frac{1}{f_{k'}},$$

sendo f_k a frequência relativa de indivíduos na categoria k, ou seja, $f_k = \frac{n_k}{n}$, sendo n_k o número de observações na categoria k e n o número total de observações.

Sendo Q o número de variáveis, o quadrado médio da distância total entre i e i' é definido por:

$$d^2(i,i') = \frac{1}{Q} \sum_{q=1}^{Q} d_q^2(i,i').$$

Segundo Roux e Rouanet (2010), quanto menor a frequência das diferentes categorias, maior é a distância entre indivíduos. Sendo assim, um ponto qualquer M^i ficará longe do centro, localizando-se na periferia da nuvem de pontos.

A linhas marginais, ou seja, o peso de um ponto M^k referente a uma categoria k é n_k, e a soma dos pesos para cada categoria de uma dada variável é n e para todas, o total de categorias é nQ.

A distância ao quadrado entre duas categorias é dada pela fórmula:

$$d^2(M^k,M^{k'}) = \frac{n_k + n_{k'} - 2n_{kk'}}{n_k n_{k'}/n},$$

sendo:

- M^k com $k = 1,2,\ldots,K$ os pontos da nuvem de categorias;
- n_k o peso do ponto M_k (número de indivíduos que escolheram a categoria k);
- $n_{k'}$ o peso do ponto $M_{k'}$ (número de indivíduos que escolheram a categoria k');
- $n_{kk'}$ o número de indivíduo que escolheram ambas categorias k e k'

Tem-se que $n_{kk'} = 0$ quando k e k' são duas diferentes categorias de uma mesma variável. Segundo Roux e Rouanet (2010) a distância entre M^k e $M^{k'}$ será menor quando o quanto mais categorias k e k' forem escolhidas pelo

mesmo indivíduo, e, quanto menor a frequência da categoria k, mais o ponto M^k se distancia do centro da nuvem.

Uma medida de dispersão da nuvem de pontos em relação a um ponto qualquer é a variância (inércia) da nuvem de pontos, que é calculada pela média dos quadrados das distâncias entre os M pontos de uma nuvem e um ponto P qualquer, menos o quadrado da distância entre o ponto médio G e o ponto P.

De acordo com Roux e Rouanet (2010) a nuvem de categorias tem a mesma inércia da nuvem de indivíduos. A frequência de indivíduos em uma certa categoria influência em sua contribuição, ou seja, categorias menos frequentes contribuem mais para a inércia global, tornando assim recomendável, o agrupamento de categorias com frequência abaixo de 5% e quanto mais categorias uma variável possui, mais esta contribui para a variância da nuvem.

A nuvem de pontos é projetada em planos que possuam a capacidade de preservar, ao máximo a distância entre eles, mantendo assim a maior parte da informação original. A projeção de uma nuvem corresponde à projeção ortogonal de seus pontos, portanto, a inércia total de uma nuvem ortogonalmente projetada é sempre menor ou igual á inércia total da nuvem inicial (ROUX; ROUANET, 2010).

A nuvem de pontos é projetada sobre os chamados eixos principais, que corresponde a sua projeção em eixos arbitrários, através das distâncias entre os pontos. A inércia de cada eixo é restituída pelos autovalores λ determinados pela fatoração da matriz indicadora através do método de decomposição por valores singulares (DVS). A soma dos autovalores λ é igual a inércia da nuvem de pontos.

As coordenadas principais dos pontos definem a nuvem referida aos seus eixos principais. De acordo com Prado (2012) a coordenada principal de um ponto M^i da nuvem de indivíduos em relação ao eixo principal é denotada por y_l^i. Para a nuvem de categorias, as coordenadas são definidas da mesma forma, sendo a coordenada principal de M^k definida como y_l^k.

Para cada eixo principal, tem-se que a média das coordenadas é nula e a variância é igual ao autovalor. Assim, tem-se que:

$$\sum \tfrac{1}{n} y_l^i = 0; \sum \tfrac{1}{n} (y_l^i)^2 = \lambda_l \text{ e } \sum p_k y_l^k = 0; \sum p_k (y_l^k)^2 = \lambda_l.$$

Como os eixos principais, as coordenadas são determinadas pela (DVS)

da matriz indicadora.

A interpretação gráfica dos resultados da ACM pode ser confusa, dependendo do número de variáveis estudadas. Algumas estatísticas, denominadas de contribuições, colaboram para a interpretação dos eixos obtidos a partir da aplicação da ACM (PRADO, 2012).

A contribuição de um ponto a um determinado eixo consiste na importância desse ponto ao eixo, isto é, o quanto da inércia do eixo é devido ao ponto. Essa contribuição é calculada por meio da multiplicação do peso p e sua coordenada nesse eixo dividido pela inércia do eixo. Sendo y a coordenada relativa ao eixo de inércia λ, a contribuição do ponto a um eixo, segundo Roux e Rouanet (2010), é dada por:

$$Ctr = \frac{(py^2)}{\lambda}.$$

Note que essa contribuição indica a quantidade da inércia do eixo explicada por um ponto, logo, define-se a contribuição do ponto referente a nuvem de indivíduos M^i e a contribuição do ponto das nuvens de categorias (M^k), respectivamente por :

$$Ctr_i = \frac{\frac{1}{n}(y^i)^2}{\lambda}$$

e

$$Ctr_k = \frac{\frac{f_k}{Q}(y^k)^2}{\lambda}.$$

A visualização das duas nuvens (indivíduos e categorias) de pontos é trabalhada separadamente. Para visualização das nuvens em um único espaço, utilizam-se as fórmulas de transição que tem como finalidade ligar as coordenadas principais referentes a nuvem de indivíduos y^i e as coordenadas principais referentes a nuvem de categorias y^k, o que permitiria analisar relações diferentes entre a nuvem de indivíduos e a nuvem de categorias. São elas:

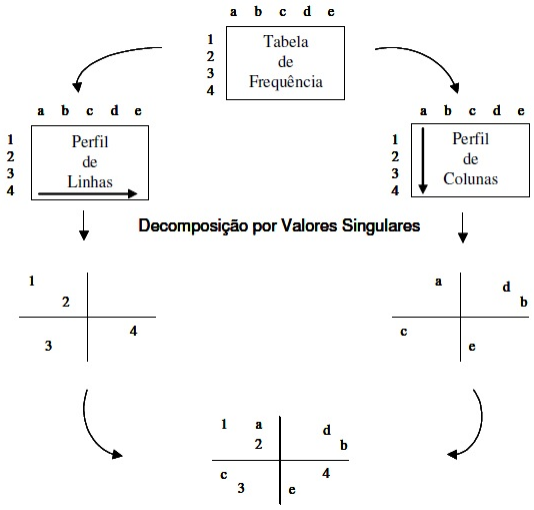
$$y^i = \frac{1}{\sqrt{\lambda}} \sum_{k \in K_i} \frac{y^k}{Q}$$

e

$$y^k = \frac{1}{\sqrt{\lambda}} \sum_{i \in I_k} \frac{y^i}{n_k}.$$

A primeira fórmula de transição permite calcular as coordenadas principais de um indivíduo através das categorias escolhidas por ele, e, com a segunda fórmula de transição, pode-se encontrar as coordenadas principais de uma categoria escolhida por um grupo de indivíduos. Na figura 2 tem-se de forma sucinta o resumo da análise de correspondência simples e múltipla.

Figura 2 Resumo da Análise de Correspondência Simples e Múltipla



Fonte: Nascimento (2011)

3 DADOS COMPOSICIONAIS

3.1 Dados composicionais: um caso particular para experimentos de mistura

Segundo Aitchison (1986) dados composicionais consistem em um conjunto de vetores, denominados composições, cujos elementos ou componentes $x_1, x_2, \cdots, x_D$ são positivos e definidos no intervalo $(0,1)$. Eles representam fra-

ções de um todo e satisfazem a restrição de que a soma dos componentes é igual a um. Assim, tem-se que:

$$x_1 \geq 0, x_2 \geq 0, \ldots, x_D \geq 0$$

e

$$x_1 + x_2 + x_3 + \ldots + x_D = 1.$$

Dados dessa forma apresentam obrigatoriamente certa correlação, já que o aumento em importância de determinado componente implica necessariamente a diminuição dos demais.

Uma característica importante neste tipo de dados é que eles carregam informação relativa e não absoluta sobre os valores das componentes, então a análise por valores absolutos seria inadequada para avaliação de dados composicionais (AITCHISON, 1986).

Uma tabela com dados composicionais gera uma tabela de múltipla entrada, em que as variáveis na coluna correspondem a cada um dos componentes x_1,$x_2, \ldots,x_D$ e o número de categorias de cada uma dessas é dado pela contagem de valores diferentes em cada coluna. Assim, diferentemente do que ocorre na análise de correspondência múltipla, para dados composicionais tem-se o número total de variáveis, e não o número de categorias associado a cada variável. A Tabela 4, exemplifica uma tabela de múltipla entrada com dados composicionais obtidos de seis amostras em que foram avaliados as valores de quatro componentes (X_1, X_2, X_3, X_4) que constituem cada amostra.

Tabela 4 Tabela de dados composicionais obtidos de seis amostras sendo avaliados quatro componentes.

	Componente			
Amostra	X_1	X_2	X_3	X_4
1	0,8	0,05	0,05	0.1
2	0,1	0,75	0,05	0,1
3	0,02	0,03	0,05	0,9
4	0,1	0,75	0,05	0,1
5	0,25	0,25	0,25	0,25
6	0,05	0,05	0,8	0,1

De acordo com Reyment e Savazzi (1999) e Labus (2005) as características principais de um conjunto de dados composicionais são:

- podem ser representados na forma de uma matriz;
- cada uma das linhas da matriz soma 1, no caso de proporções; 100% no caso de percentagens; ou alguma outra constante, de acordo com a forma particular de representação dos dados adotada pelo pesquisador como, por exemplo, unidades ppm (partes por milhão, equivalente a 1 miligrama por litro), ppb (partes por bilhão, equivalente a 1 micrograma por litro) ou outra unidade de concentração;
- cada coluna da matriz representa uma componente (parte);
- os coeficientes de correlação amostrais mudam se um dos componentes é excluído da matriz de dados, e a soma relativa 1 ou 100% é realizada novamente nas linhas. O mesmo ocorre se um novo componente é adicionado.

Essa última característica significa que alterar um ou mais componentes do conjunto de dados pode ter um efeito numericamente significativo nas correlações entre os restantes.

3.2 Espaço amostral dos dados composicionais

O espaço amostral restrito no qual são definidos os dados composicionais é conhecido como $D-$simplex, de dimensão igual ao número de componentes, dado pela equação (PAWLOWSKY et al., 2010):

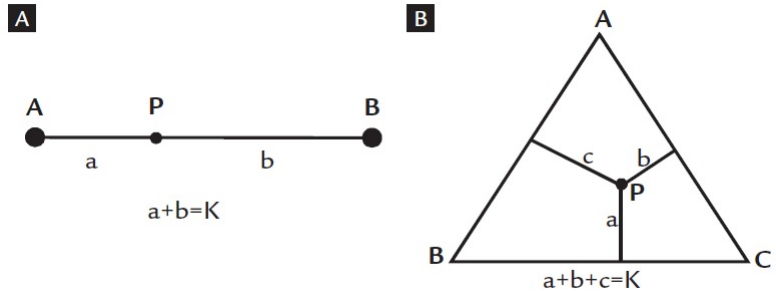
$$S^D = \{x = [x_1; \ldots; x_D \mid x_i \geq 0 \; e \; \sum_{j=1}^{D} x_j = K]\},$$

com $K = 1{,}100{,}10^6{,}10^9$(proporção,$\%$,ppm,etc.).

O D-simplex de uma composição com dois componentes ($2-$simplex), A e B, com soma constante K, é um segmento de reta, no qual a soma constante K é válida em cada ponto P da mesma. No caso de uma composição composta de 3 componentes, A, B e C, também com soma constante K, o 3-simplex é um

triângulo equilátero em que qualquer ponto P satisfaz a soma constante K, (Figura 3).

Figura 3 Simplex com 2 e 3 componentes



Fonte: Boezio et al. (2011)

A operação que define o fechamento de uma composição em uma constante K para cada amostra (linha) é dada pela seguinte equação (PAWLOWSKY et al., 2010):

$$C(x) = \left[\frac{K \times x_1}{\sum_{i=1}^{D} x_i}, \frac{K \times x_2}{\sum_{i=1}^{D} x_i}, \frac{K \times x_3}{\sum_{i=1}^{D} x_i}, \cdots, \frac{K \times x_D}{\sum_{i=1}^{D} x_i}\right],$$

em que:

- $C(x) =$ operação de fechamento;
- $K =$ a constante de fechamento (geralmente 100%);
- x_i é o valor da i-ésima amostra.

No entanto a restrição de soma constante imposta pelos dados composicionais impõe limitações para aplicação direta da ACM.

Segundo Pawlowsky-Glahn e Olea (2004) os dados composicionais apresentam um efeito de correlação espúria devido à restrição de soma constante, o que significa que a aplicação dos métodos estatísticos padrões pode levar a resultados inconsistentes. De acordo com Labus (2005), outro problema é o fato de que os

componentes não são independentes. Segundo Pawlowsky-Glahn e Olea (2004) isso implica em singularidade da matriz de covariâncias de uma composição.

Baseado no fato de que os dados composicionais carregam informações relativas e não absolutas do valor das componentes, Aitchison (1986) propõe que as variações relativas entre componentes sejam medidas em escala logarítmica, por meio do uso de razões logarítmicas. Posteriormente, após a realização de alguma transformação (seção 4.3) os dados estão prontos para serem submetidos a uma análise estatística.

3.3 Transformações logarítmicas

Boezio (2010) menciona que o princípio das transformações de razões logarítmicas está baseado no fato de que existe uma correspondência um a um entre os vetores composicionais e os vetores das razões logarítmicas associadas.

A utilização de razões logarítmicas, no caso de valores nulos, Pawlowsky-Glahn e Buccianti (2011) recomendam que esses valores sejam substituídos pelo limite de detecção de cada componente analisado ou por um valor muito pequeno, tal como $1e^{-5}$.

Aitchison (1986) definiu duas das principais transformações logarítmicas utilizadas em dados composicionais, conhecidas como Transformações Logarítmicas Aditivas (alr) e Transformações Logarítmicas centradas (clr). Uma terceira transformação bastante citada na literatura sobre dados composicionais foi proposta por Egozcue et al. (2003), conhecida como transformações logarítmicas isométricas (ilr).

Nos três diferentes tipos de transformações logarítmicas, a composição de cada amostra é transformada em um vetor. Nas transformações alr e ilr, a amostra resulta com uma componente a menos, enquanto que na transformação clr o número de componentes é preservado.

3.3.1 Transformações logarítmicas aditivas *(alr)*

Definindo x como uma composição de D-variáveis no simplex S^D, então:

$$alr(\mathbf{x}) = (y_1, y_2, \ldots, y_{D-1}) = \left[\ln \frac{x_1}{x_D}; \ln \frac{x_2}{x_D}; \ldots; \ln \frac{X_{D-1}}{X_D}\right],$$

ou seja, é aplicado o logaritmo natural na divisão de cada componente pela última componente X_D. Logo, para a primeira componente, a transformação alr é definida como $alr(x_1) = \ln \frac{x_1}{x_D}$. Uma vez que a transformação alr divide os valores de dados por uma variável de referência, a escolha dessa variável de referência vai influenciar resultados, assim esta transformação é bastante subjetiva (FILZMOSER; HRON; REIMANN, 2009).

Para voltar ao espaço simplex saindo do espaço real, aplica-se o processo inverso dado pela seguinte equação:

$$x = C(exp(alr_1(x), alr_2(x), \ldots, alr_{D-1}(x),0))$$

onde C é a operação de fechamento definida na Seção 4.2.

A vantagem nesta transformação é que o espaço simplex de uma composição de D partes é reduzido para um espaço real de $D - 1$ partes. Entretanto, aponta-se como desvantagem a essa transformação, justamente por não preservar as distâncias no espaço real. Um exemplo da transformação dos dados originais é dado a seguir:

Tabela 5 Dados resultantes após transformação logarítmica aditiva (alr).

	Componentes Originais				Componentes Transformados		
Amostra	X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3
1	0,8	0,05	0,05	0,1	2,08	-0,69	-0,69
2	0,1	0,75	0,05	0,1	0,00	2,01	-0,69
3	0,02	0,03	0,05	0,9	-3,81	-3,40	-2,89
4	0,1	0,75	0,05	0,1	0,00	2,01	-0,69
5	0,25	0,25	0,25	0,25	0,00	0,00	0,00
6	0,05	0,05	0,80	0,10	-0,69	-0,69	2,08

3.3.2 Transformações logarítmicas centradas *(clr)*

A transformação clr é uma transformação de S^D para $\mathbb{R}^D$ e o resultado de uma observação $x \in S^D$ são os dados transformados $y \in \mathbb{R}^D$. Assim considerando $g(x)$ como a média geométrica dada por:

$$g(x) = \sqrt[D]{x_1 \times x_2 \times \cdots \times x_D}$$

os dados transformados (y) são obtidos por:

$$clr(\mathbf{x}) = (y_1, y_2, \ldots, y_D) = \left[\ln \frac{x_1}{g(x)}, \ldots, \ln \frac{x_D}{g(x)}\right]$$

Diferentemente das alr, as transformações clr preservam as distâncias no espaço real, uma vez que

$$\langle x,y \rangle = \langle clr(x), clr(y) \rangle,$$

$$\| x \| = \| clr(x) \|,$$

$$d(x,y) = d(clr(x), clr(y)),$$

em que $\langle x,y \rangle$ representa o produto interno, $\| x \|$ e $d(x,y)$ representam a norma e a distância, respectivamente.

Para voltar ao espaço simplex a partir do espaço real, o processo inverso é dado pela seguinte equação:

$$x = C(\exp(clr_1(x), clr_2(x), \ldots, clr_D(x))).$$

Pawlowsky et al. (2010) ressaltam que a principal desvantagem desta transformação é que os dados resultantes são colineares porque $\sum_{i=1}^{D} y_i = 0$. Um exemplo da transformação dos dados originais é apresentado na Tabela 6.

Tabela 6 Dados resultantes após transformação logarítmica centrada (clr).

	Componentes Originais				Componentes Transformados			
Amostra	X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3	Y_4
1	0,8	0,05	0,05	0,1	1,91	-0,87	-0,87	-0,17
2	0,1	0,75	0,05	0,1	-0,33	1,68	-1,02	-0,33
3	0,02	0,03	0,05	0,9	-1,28	-0,88	-0,37	2,52
4	0,1	0,75	0,05	0,1	-0,33	1,68	-1,02	-0,33
5	0,25	0,25	0,25	0,25	0,00	0,00	0,00	0,00
6	0,05	0,05	0,8	0,1	-0,87	-0,87	1,91	-0,17

3.3.3 Transformações logarítmicas isómetricas *(ilr)*

De acordo com Rubio (2014), essa transformação supre as desvantagens observadas nas transformações alr e clr, no entanto, baseia-se na escolha de uma base ortonormal de S^D, desta forma considerando as composições $e_1, e_2, \ldots, e_{D-1}$ tal que $\langle e_i, e_j \rangle = 0$ para $i \neq j$ e $\parallel e_i \parallel = 1$. Para uma base fixa, as coordenadas de uma composição são obtidas usando a função:

$$ilr(\mathbf{x}) = (y_1, y_2, \ldots, y_{D-1}) = (\langle x, e_1 \rangle, \langle x, e_2 \rangle_a, \ldots, \langle x, e_{D-1} \rangle),$$

$$\langle x,y \rangle = \langle ilr(x), ilr(y) \rangle,$$

$$\parallel x \parallel = \parallel ilr(x) \parallel,$$

$$d(x,y) = d(ilr(x), ilr(y)),$$

em que $\langle x,y \rangle$ representa o produto interno, $\parallel x \parallel$ e $d(x,y)$ representam a norma e a distância, respectivamente.

Nota-se que as equações acima são análogas às propriedades dadas nas equações da transformação clr. A única diferença é que o produto interno, a norma e a distância entre os vetores das coordenadas ilr correspondem à dimensão $D-1$ do espaço real (RUBIO, 2014). Um exemplo da transformação dos dados originais é apresentado na Tabela 7.

Tabela 7 Dados resultantes após transformação logarítmica isométrica (ilr).

	Componentes Originais				Componentes Transformados		
Amostra	X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3
1	0,8	0,05	0,05	0,1	-1,96	-1,13	-0,20
2	0,1	0,75	0,05	0,1	1,42	-1,39	-0,38
3	0,02	0,03	0,05	0,9	0,29	0,58	2,92
4	0,1	0,75	0,05	0,1	1,42	-1,39	-0,38
5	0,25	0,25	0,25	0,25	0,00	0,00	0,00
6	0,05	0,05	0,80	0,10	-0,00	2,26	-0,20

3.4 Análise de correspondência aplicada a dados composicionais

A análise de correspondência múltipla é normalmente utilizada para dados originalmente categóricos, no entanto Aitchison (1986) propôs a expansão dessa técnica para dados composicionais. Para melhores esclarecimentos, prosseguimos com um exemplo fictício ilustrando o cálculo das transformações logarítmicas aditiva, centrada e isométrica, a partir dos dados originais da Tabela 4.

Após cada uma das transformações utilizadas, para cada componente são avaliados os diferentes valores obtidos, visando organizá-los em categorias, originando desta forma uma tabela com dados de múltipla entrada, na qual pode ser utilizada a análise de correspondência múltipla.

Tabela 8 Tabela de múltipla entrada dos componentes pós categorização e transformação alr.

	Componentes com os dados transformados											
	Y_1				Y_2				Y_3			
Amostra	-3,81	-0,69	0	2,08	-3,40	-0,69	0	2,01	-2,89	-0,69	0	2,08
1	0	0	0	1	0	1	0	0	0	1	0	0
2	0	0	1	0	0	0	0	1	0	1	0	0
3	1	0	0	0	1	0	0	0	1	0	0	0
4	0	0	1	0	0	0	0	1	0	1	0	0
5	0	0	1	0	0	0	1	0	0	0	1	0
6	0	1	0	0	0	1	0	0	0	0	0	1

Seguindo a álgebra da análise de correspondência múltipla, a contribuição das amostras em cada um dos eixos, segundo a transformação alr é descrita na Tabela 9.

Utilizando a ACM para os dados com transformação alr, em uma escala de porcentagem, os resultados descritos na Tabela 9 evidenciam que a amostra 3 é a que possui maior destaque com contribuição de $83{,}33\%$ ao eixo 1. As amostras 6 e 5 estão relacionadas com o segundo eixo, sendo a contribuição das mesmas para o segundo eixo $56{,}25\%$ e 25%, respectivamente. O terceiro e quarto eixo destacam principalmente as amostras 5 e 1, com contribuição de 50% e $61{,}25\%$ respectivamente.

Tabela 9 Contribuição das amostras em cada um dos eixos, usando a transformação alr.

	Eixos			
Amostras	1	2	3	4
1	0,0333	0,0625	0,1250	**0,6125**
2	0,0333	0,0625	0,1250	0,1125
3	**0,8333**	0,0000	0,0000	0,0000
4	0,0333	0,0625	0,1250	0,1125
5	0,0333	0,2500	**0,500**0	0,0500
6	0,0333	**0,5625**	0,1250	0,1125

Utilizando a transformação clr, a Tabela 10 de múltipla entrada dos componentes obtidos pós categorização é obtida.

Tabela 10 Tabela de múltipla entrada dos componentes pós categorização e transformação clr.

	Componentes com os dados transformados																	
	Y_1					Y_2				Y_3					Y_4			
Amostra	-1,28	-0,87	-0,33	0	1,91	-0,88	-0,87	0	1,68	-1,02	-0,87	-0,37	0	1,91	-0,33	-0,17	0	2,52
1	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	1	0	0
2	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0
3	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
4	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0
5	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0
6	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0

Utilizando a ACM para os dados com transformação clr, (Tabela 11) percebe-se que a amostra 5 é a que possui maior destaque com contribuição $68{,}25\%$ ao eixo 1. As amostras 2 e 4 estão relacionadas com o segundo eixo, sendo a contribuição das mesmas para o segundo eixo $17{,}48\%$. O terceiro e quarto eixos destacam principalmente as amostras 3, 1 e 5, com contribuição de $66{,}67\%$, 50% e 50% respectivamente.

Tabela 11 Contribuição das amostras em cada um dos eixos, usando a transformação clr.

	Eixos			
Amostras	1	2	3	4
1	0,0001	0,1666	0,1667	**0,5000**
2	0,1586	**0,1748**	0,0000	0,0000
3	0,0001	0,1666	**0,6667**	0,0000
4	0,1586	**0,1748**	0,0000	0,0000
5	**0,6825**	0,1508	0,0000	0,0000
6	0,0001	0,1666	0,1667	**0,5000**

Por fim, considerando a transformação (ilr), a tabela de múltipla entrada dos componentes pós-categorização (Tabela 12) é apresentada a seguir.

Tabela 12 Tabela de múltipla entrada dos componentes pós categorização e transformação ilr.

	Componente com os dados transformados												
	Y_1				Y_2					Y_3			
Amostra	-1,96	0	0,29	1,42	-1,39	-1,13	0	0,58	2,26	-0,38	-0,2	0	2,92
1	1	0	0	0	0	1	0	0	0	0	1	0	0
2	0	0	0	1	1	0	0	0	0	1	0	0	0
3	0	0	1	0	0	0	0	1	0	0	0	0	1
4	0	0	0	1	1	0	0	0	0	1	0	0	0
5	0	1	0	0	0	0	1	0	0	0	0	1	0
6	0	1	0	0	0	0	0	0	1	0	1	0	0

Para os dados com transformação ilr, (Tabela 13) percebe-se que a amostra 3 possui maior contribuição em relação aos dois primeiros eixos com $19{,}04\%$ e $64{,}30\%$ de explicação, respectivamente. As amostras 1, 5 e 6 possuem maior contribuição para o segundo e terceiro eixo com 50% de explicação em relação as amostras 1 e 5 e contribuição de $66{,}67\%$ da amostra 6 para o eixo 4.

Na Tabela 14 pode-se observar as categorias dos componentes formadas pelas transformações alr, clr e ilr, sendo que as categorias que possuem maior inércia são as categorias das amostras que possuem maior contribuição em relação aos eixos. A interpretação gráfica para este exemplo foi restrita aos eixos 1 e 2 uma vez que a porcentagem de inércia restituída nesses eixos são as de maior proporção.

Tabela 13 Contribuição das amostras em cada um dos eixos, usando a transformação ilr.

	Eixos			
Amostras	1	2	3	4
1	0,1665	0,0002	**0,5000**	0,1667
2	0,1551	0,1782	0,0000	0,0000
3	**0,1904**	**0,6430**	0,0000	0,0000
4	0,1551	0,1782	0,0000	0,0000
5	0,1665	0,0002	**0,5000**	0,1667
6	0,1665	0,0002	0,0000	**0,6667**

Tem-se que pela transformação alr as categorias que se destacam em relação a contribuição dos eixos são (Y_1:-3,81/ Y_2:-3,4 / Y_3:2,89), sendo o grupo de categorias de maior inércia. Para a transformação clr as categorias que possuem maior contribuição aos eixos são (Y_1:0/Y_2:0/Y_3:0/Y_4:0), sendo o grupo que possui maior inércia. No caso da transformação ilr as categorias que se destacam em relação aos eixos são (Y_1:-0,29/Y_2:0,58/Y_3:2,92) sendo o grupo de categorias com maior inércia.

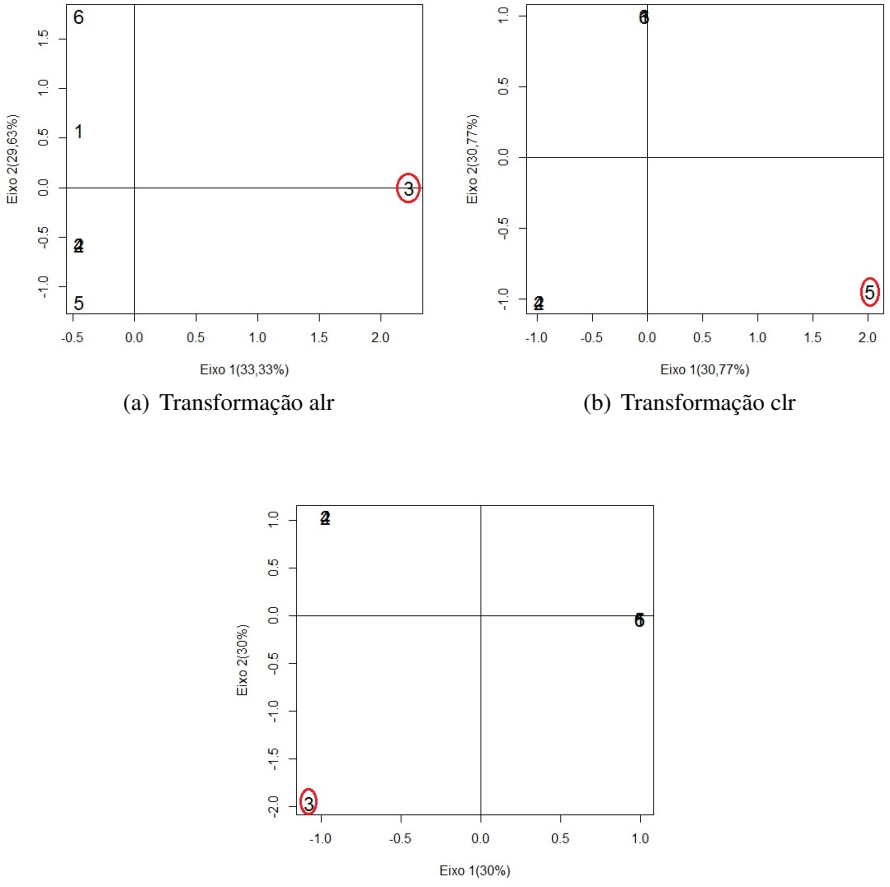
Tabela 14 Categorias e inércias referentes as três transformações

	Transformação alr		Transformação clr		Transformação ilr	
Grupo	Categorias	Inércia	Categorias	Inércia	Categorias	Inércia
1	Y_1:-3,81	0,278	Y_1:0/Y:0	0,208	Y_1:0,29	0,278
	Y_2:-3,4	0,278	Y_3:0	0,208	Y_2:0,58	0,278
	Y_3:-2,89	0,278	Y_4:0	0,208	Y_3:2,92	0,278
2	Y_1:-0,69	0,222	Y_1:-1,3/Y_1:-0,3/Y_3:-0,4	0,167	Y_1:-1,96/Y_1:1,42	0,222
	Y_3:2,08	0,222	Y_2:1,7/Y_3:-1/Y_4:2,5	0,167	Y_2:-1,39/Y_2:-1,13	0,222
3	Y_2:0/Y_3:0	0,204	Y_4:-0,2	0,146	Y_1:0/Y_3:-0,2	0,194
4	Y_2:-0,69	0,185	Y_1:-0,9/Y_1:1,9/Y_3:1,9	0,135	Y_2:0	0,1666
5	Y_1:2,08	0,148	Y_2:-0,9	0,125	-	-
	Y_2:2,01	0,148	-	-	-	-
6	Y_1:0	0,142	-	-	-	-
7	Y_3:-0,69	0,124	-	-	-	-

Na Figura 4 têm-se os mapas perceptuais das amostras referentes às três transformações alr, clr e ilr, dos eixos 1 e 2 que são os eixos que possuem os maiores percentuais de explicação obtidos em cada transformação.

Observa-se que a amostra 3 destaca-se em relação aos eixos 1 e 2, quando utilizada a transformação alr (Figura 4 a), sendo a amostra de maior contribuição para os eixos. No caso da transformação clr observa-se que a amostra 5 se destaca em relação as demais amostras, sendo a que possui maior contribuição em relação aos eixos (Figura 4 b). Em relação à transformação ilr observa-se que a amostra 3 é a amostra que se destaca em relação aos dois primeiros eixos (Figura 4 c).

Figura 4 Mapas perceptuais das amostras em relação aos eixos de maior contribuição nas três tranformações.



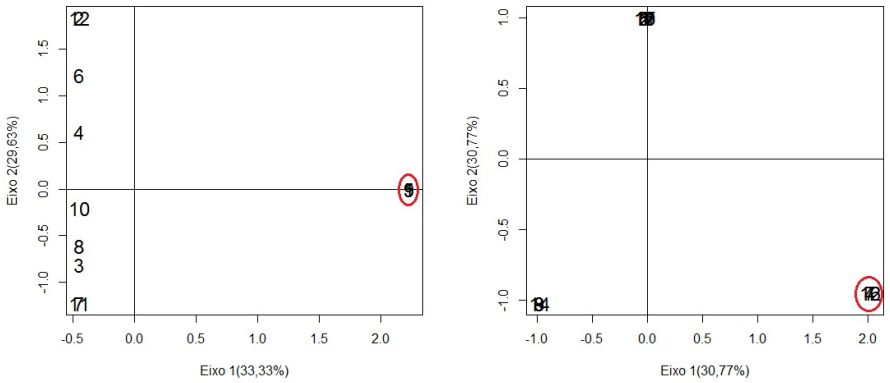
(a) Transformação alr

(b) Transformação clr

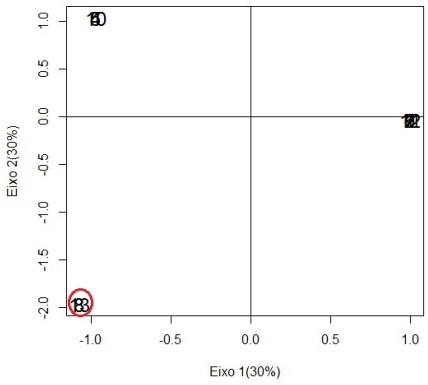
(c) Transformação ilr

Na Figura 5 têm-se os mapas perceptuais das categorias das amostras geradas pelas três transformações, onde as categorias em destaque são as categorias das amostras que se destacaram na Figura 4, sendo essas as categorias dos grupos de maior inércia da Tabela 14.

Figura 5 Mapas perceptuais das categorias em relação aos eixos de maior contribuição nas três transformações.



(a) Transformação alr (b) Transformação clr



(c) Transformação ilr

Referências

AITCHISON, J. **The Statistical Analysis of Compositional Data**. London, UK, UK: Chapman & Hall, Ltd., 1986.

BAXTER, M.; COOL, H.; HEYWORTH, M. Principal component and correspondence analysis of compositional data: some similarities. **Journal of Applied Statistics**, Taylor & Francis, v. 17, n. 2, p. 229–235, 1990.

BEH, E. J. Simple correspondence analysis using adjusted residuals. **Journal of Statistical Planning and Inference**, v. 142, n. 4, p. 965 – 973, 2012.

BLASUS, J. et al. Special issue on correspondence analysis and related methods. **Computational Statistics e Data Analysis, New York**, v. 53, n. 8, p. 3103–3106, 2009.

BOEZIO, M. et al. Cokrigagem de razões logarítmicas aditivas alr na estimativa de teores em depósitos de ferro. **Rem Revista Escola de Minas**, v. 49, p. 401–408, 2011.

BOEZIO, M. N. M. Estudo das metodologias alternativas da geoestatística multivariada aplicadas a estimativa de teores de depósitos de ferro. 2010.

EGOZCUE, J. et al. Isometric logratio transformations for compositional data analysis. **Mathematical Geology**, Kluwer Academic Publishers-Plenum Publishers, v. 35, n. 3, p. 279–300, 2003.

ESCOFIER, B. Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. **Revue de Statistique Appliquée**, v. 26, n. 4, p. 29–37, 1978.

FILZMOSER, P.; HRON, K.; REIMANN, C. Principal component analysis for compositional data with outliers. **Environmetrics**, Wiley Online Library, v. 20, n. 6, p. 621–632, 2009.

GREENACRE, M. Correspondence analysis in medical research. **Statistical methods in medical research**, SAGE Publications, v. 1, n. 1, p. 97–117, 1992.

GREENACRE, M. J. **Theory and Applications of Correspondence Analysis**. 2a ediÃ§Ã£o. ed. Orlando: Academic Press, 1984.

GREENACRE, M. J. Correspondence analysis. **Wiley Interdisciplinary Reviews:Computational Statistics**, John Wiley and Sons, Inc., v. 2, n. 5, p. 613–619, 2010.

LABUS, M. Compositional data analysis as a tool for interpretation of rock porosity parameters. **Geological Quarterly**, v. 49, 2005.

NAITO, S. **Análise de correspondências generalizada.** Universidade de Lisboa, Lisboa.: [s.n.], 2007.

NASCIMENTO, A. do. **Avaliação de Farmácias Hospitalares Brasileiras Utilizando Análise de Correspondência Múltipla**. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2011.

PAWLOWSKY, G. V. et al. **Lecture Notes on Compositional Data Analysis.** Technical University of Catalonia, Spain: [s.n.], 2010.

PAWLOWSKY-GLAHN, V.; BUCCIANTI, A. **Compositional data analysis: Theory and applications**. [S.l.]: John Wiley & Sons, 2011.

PAWLOWSKY-GLAHN, V.; OLEA, R. A. **. Geostatistical Analysis of Compositional Data.** New York: Oxford University Press, Inc: [s.n.], 2004.

PRADO, M. V. B. **Métodos de análise de correspondência múltipla: Estudo de caso aplicado á avaliação da qualidade do café.** Minas Gerais, MG, Brasil: [s.n.], 2012.

REYMENT, R. A.; SAVAZZI, E. **Aspects of Multivariate Statistical Analysis in Geology.** Elsevier: [s.n.], 1999.

ROUX, B. L.; ROUANET, H. **Geometric data analysis: from correspondence analysis to structured data analysis**. [S.l.]: Springer Science & Business Media, 2004.

ROUX, B. L.; ROUANET, H. **Multiple correspondence analysis**. Londres: Sage, 2010.

RUBIO, R. J. H. **CODA: Uma alternativa para estimativas multivariadas que envolvem balanços de massa granulométrico e das espécies químicas**. Porto Alegre, RS, Brasil: [s.n.], 2014.

SANTOS, E. N. F. **Álgebra intervalar na análise de correspondência: um estudo de caso em testes de aceitação sensorial com erros de medida**. Tese (Doutorado) — Universidade Federal de Lavras, 2012.

SILVA, Y. V. da. Dissertação de mestrado em Estatística Aplicada e Biometria, **Análise de Correspondência: uma abordagem geométrica**. Viçosa - MG: [s.n.], 2012.

SOUZA, A. C. **AnÁlise de Correspondência aplicada à ECINF: a diversidade do setor informal urbano no Brasil.** Rio de Janeiro - Rj: [s.n.], 2004.

TRAMARIN, A. et al. The influence of socioeconomic status on health service utilisation by patients with aids in north italy. **Social Science & Medicine**, Elsevier, v. 45, n. 6, p. 859–866, 1997.

ZIEGEL, E. R. Correspondence analysis handbook. **Technometrics**, Taylor & Francis, v. 35, n. 1, p. 103–103, 1993.

UNIVERSIDADE FEDERAL DE LAVRAS

Artigo 1

Uma nova abordagem da análise de correspondência simples com ênfase na violação da hipótese de independência dos níveis das variáveis categóricas

Versão preliminar de artigo - Sujeito a alterações pelo corpo editorial da revista

LAVRAS - MG

2016

RESUMO

A principal hipótese da análise de correspondência é dada pela independência entre os níveis das variáveis categóricas. Decorrente a violação dessa hipótese, esse trabalho tem por objetivo aprimorar a técnica da análise de correspondência, fornecendo uma nova abordagem para o cálculo das coordenadas através da incorporação de resíduos, mediante tabelas em que categorias apresentam diferentes níveis de correlação. Com esse propósito, utilizou-se a simulação Monte Carlo na geração de frequências provenientes da distribuição binomial correlacionada $BC(n, \pi, \rho)$. Concluiu-se que em todos os cenários avaliados a abordagem é promissora, no sentido que os objetos foram melhor discriminados em relação a abordagem convencional. Ainda, o procedimento proposto para obtenção das coordenadas é plausível de ser utilizado em dados reais conforme ilustra o exemplo de aplicação.

Palavras-chave: binomial correlacionada, resíduos padronizados, resíduos ajustados, coeficiente de correlação cofenética.

1 INTRODUÇÃO

A análise de dados categóricos envolve uma séries de métodos estatísticos aplicados a dados discretos, sejam eles representados por variáveis qualitativas ou quantitativas discretizadas. Nesse contexto, delineamentos amostrais associados aos modelos probabilísticos são propostos com o intuito de evidenciar informações relevantes por parâmetros, preservando as escalas de mensuração.

Em se tratando da organização dos dados, em síntese, resume-se na formação de tabelas de contingência, nas quais as principais características da análise estatística caracterizam-se por avaliar a independência entre as variáveis, estudar as distribuições condicionais, entender a associação entre as categorias, bem como uma representação gráfica compreensível a interpretação dos resultados. Diante do exposto, surge a técnica de análise de correspondência como uma alternativa importante por contemplar as características mencionadas.

Segundo Guedes et al. (1999) entende-se como análise de correspondência uma técnica exploratória da análise multivariada que permite obter uma representação gráfica multidimensional da dependência entre as linhas e/ou colunas de uma tabela de contingência de duas entradas, onde as linhas e as colunas representam categorias ou modalidades de variáveis categóricas. Análise de correspondência é um método de análise fatorial para variáveis categóricas. A representação gráfica é obtida pela distribuição dos escores das categorias de linhas e colunas e marcando estas categorias como pontos, onde os escores são utilizados como as coordenadas destes pontos.

A literatura contempla vários trabalhos que descrevem os aspectos teóricos e aplicações sobre a teoria de análise de correspondência, tais como: Greenacre (1984), Greenacre (1992), Greenacre (2007), Benzécri (1992), Blasus et al. (2009), Beh (2004) e Beh (2012) .

A viabilidade em se aplicar a análise de correspondência em uma tabela de contingência inicia-se com a aplicação de um teste Qui-Quadrado que avaliará a independência entre as variáveis categóricas. Convém ressaltar que a estatística deste teste na análise de correspondência é de suma importância para validar a aplicação, uma vez que o mesmo torna-se um indicativo da decomposição da inér-

cia restituída pelos componentes que identificam um espaço de dimensão reduzida em relação ao número de variáveis, de tal forma que a dispersão dos pontos possa ser representada da melhor forma possível, no sentido de proporcionar mapas perceptuais simétricos. Todavia, sabe-se que a estatística Qui-Quadrado é sensível a outliers que, no caso da análise de correspondência, podem ser identificados por pontos cuja massa de uma variável é bem superior a outra.

Em função desse problema, surge a necessidade de agregar novas informações à estatística Qui-Quadrado, na qual a decomposição de valores singulares é empregada. Veloso e Cirillo (2016) em um estudo de simulação propuseram um teste de significância para evidenciar componentes principais que melhor discriminam os outliers. Para isso recomendaram que as amostras devem ser corrigidas pelas distâncias Qui-Quadrado de Pearson e Yates.

Uma outra alternativa consiste em incorporar os resíduos de Pearson e suas diferentes versões descritas por Lee e Yick (1999). Desenvolvendo um procedimento que envolve os resíduos de Pearson utilizando a decomposição dos valores singulares em relação ao método de normalização, Beh (2012) propõe-se um procedimento para identificar as células que se desviam da hipótese de independência. Contudo, os pressupostos relativos a esse resíduo não são sempre satisfeitos e por isso tais resultados podem levar a conclusões questionáveis, uma vez que esses resíduos não apresentam variância unitária, é questionável utilizá-los para identificar as células que não são consistentes com a hipótese mencionada.

Decorrente ao fato de que a principal suposição da análise de correspondência é verificada na suposição de independência entre as variáveis categóricas, representadas nas linhas e colunas, mantem-se o foco na incorporação do resíduo padronizado proposto por Haberman (1973) e do resíduo ajustado proposto por Barnett e Lewis (1994) no cálculo dos escores da análise de correspondências simples, considerando estruturas de tabelas de contingência com diferentes graus de correlação entre os níveis das variáveis categóricas. Dessa forma, a análise de correspondência simples é aprimorada com a obtenção de novas coordenadas envolvendo esses resíduos.

Em virtude do que foi mencionado, este trabalho teve por objetivo avaliar e aprimorar a técnica da análise de correspondência, fornecendo uma nova abor-

dagem para o cálculo das coordenadas com incorporação de resíduos, mediante tabelas de contingência em que categorias apresentam diferentes níveis de correlação.

2 Metodologia

2.1 Geração das tabelas de contingência com amostras correlacionadas

A geração das tabelas de contingência representadas pela matriz $X_{l\times c}$, onde, fixada a *i*-ésima linha ($i = 1,\ldots,l$ e $l = 5$ ou 10), as frequências observadas y_{ij} ($j = 1,\ldots,C$ e $C = 5$ ou 10) foram simuladas seguindo a distribuição binomial correlacionada, ou seja, $y_{ij} \sim BC(n,\pi,\rho)$, definida por:

$$P(Y_j \mid n_j,\pi_j,\rho) = \left(\begin{array}{c} n_j \\ y_j \end{array} \right) \pi_j^{y_j}(1-\pi_j)^{n_j-y_j}(1-\rho)I_{A_1}(y_j)$$

$$+\pi^{y_j/n_j}(1-\pi_j)^{n_j-y_j/n_j}\rho I_{A_2}(y_j)\,,$$

com $A_1 = 0,1,\ldots,n_j$, $A_2 = 0, n_j$, $y_j = 0,\ldots,n_j$ e $0 \leq \rho \leq 1$.

Seguindo o procedimento dado por Cirillo e Ramos (2014), o vetor de variáveis aleatórias $\mathbf{Y} = (Y_1,Y_2,\ldots,Y_J)$, do qual cada componente representa o número de ocorrências na categoria j, para $j = 1,2,\ldots,C$, associado a um vetor $\pi_j = (\pi_1,\pi_2,\ldots,\pi_J)$, em que π_j corresponde a probabilidade de sucesso da binomial e ρ a taxa de mistura das distribuições binomial (n,π) com probabilidade $(1-\rho)$ é uma distribuição Bernoulli modificada representada por $BeM(\pi)$ assumindo 0 ou n valores com probabilidade ρ.

Assim, convém ressaltar que, fixado $\rho = 0$, o modelo $BC(n_j,\pi_j,\rho)$ é equivalente ao modelo binomial comum $B(n,\pi)$. Para $\rho \neq 0$ o modelo inclui variações extra-binomiais e $\rho \approx 1$ obtêm-se um excesso de n_j nas frequências simuladas. Neste caso, para evitar que as frequências matriciais sejam nulas, impossibilitando a realização das operações matriciais no cálculo dos escores, o valor nulo gerado foi substituído por 1. Dada essa adaptação a média mantém-se a mesma obtida por Tallis (1962), porem a variância torna-se aproximada, isto é:

$$E(Y_j) = n_j\pi_j$$

$$Var(Y_j) \cong \pi_j(1-\pi_j)\{n_j + \rho\, n_j(n_j-1)\}.$$

Seguindo essas especificações, os valores paramétricos utilizados na geração das tabelas de contingência encontram-se resumidos abaixo (Tabela 1).

Tabela 1 Valores paramétricos utilizados como cenários para gerar as tabelas de contingência para aplicação da análise de correspondência.

Dimensão	Proporção da Binomial	Grau de Correlação (ρ)
5×5 e 10×10	0,2	0,2
		0,5
		0,8
	0,5	0,2
		0,5
		0,8
	0,9	0,2
		0,5
		0,8

2.2 Obtenção dos escores convencionais e modificados da análise de correspondência

Para cada tabela de contingência gerada (Seção 2.1), os resíduos r_{ij} e $\tilde{r}_{ij}$, propostos, respectivamente, por Barnett e Lewis (1994) e Haberman (1973) foram computados por:

$$R_{BL} = \begin{bmatrix} r_{11} & \dots & r_{1c} \\ \vdots & r_{ij} & \vdots \\ r_{l1} & \dots & r_{lc} \end{bmatrix} \; sendo\, r_{ij} = \frac{y_{ij} - e_{ij}}{\sqrt{e_{ij}}},\ (i = 1,...,n; j = 1,...,n),$$

$e_{ij} = n \times p_{ij}p_{ij}$ com $p_{ij} = E[\frac{x_{ij}}{n}]$ estimada pela equação $\frac{x_{i+}x_{+j}}{n}$;

$$R_H = \begin{bmatrix} \tilde{r}_{11} & \dots & \tilde{r}_{1c} \\ \vdots & \tilde{r}_{ij} & \vdots \\ \tilde{r}_{l1} & \dots & \tilde{r}_{lc} \end{bmatrix} \; sendo\, \tilde{r}_{ij} = \frac{r_{ij}}{\sqrt{(1-\frac{y_{i+}}{n})(1-\frac{y_{+j}}{n})}}.$$

Seguindo essas especificações, os escores da análise de correspondência

na forma convencional C_1 e C_2 e modificada com a incorporação dos resíduos $C_1^{R_{BL}}$ e $C_2^{R_{BL}}$ e $C_1^{R_H}$ e $C_2^{R_H}$ foram obtidos, respectivamente, para os perfis de "linha" e "coluna", representados por:

$$C_1 = \begin{bmatrix} \frac{1}{(\frac{n_{1+}}{n})} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{(\frac{n_{p+}}{n})} \end{bmatrix} \begin{bmatrix} u_{11}\sqrt{\frac{n_{1+}}{n}} & \cdots & u_{1k}\sqrt{\frac{n_{1+}}{n}} \\ \vdots & \vdots & \vdots \\ u_{p1}\sqrt{\frac{n_{p+}}{n}} & \cdots & u_{pk}\sqrt{\frac{n_{p+}}{n}} \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix}$$

$$C_2 = \begin{bmatrix} \frac{1}{(\frac{n_{+1}}{n})} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{(\frac{n_{+q}}{n})} \end{bmatrix} \begin{bmatrix} v_{11}\sqrt{\frac{n_{+1}}{n}} & \cdots & v_{1k}\sqrt{\frac{n_{+1}}{n}} \\ \vdots & \vdots & \vdots \\ v_{q1}\sqrt{\frac{n_{+q}}{n}} & \cdots & v_{qk}\sqrt{\frac{n_{+q}}{n}} \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix}$$

$$C_1^{(R_{BL})} = \begin{bmatrix} r_{11} & \cdots & r_{1c} \\ \vdots & r_{ij} & \vdots \\ r_{l1} & \cdots & r_{lc} \end{bmatrix} \begin{bmatrix} u_{11}(\frac{n_{1+}}{n})^{-\frac{1}{2}} & \cdots & u_{1k}(\frac{n_{1+}}{n})^{-\frac{1}{2}} \\ \vdots & \vdots & \vdots \\ u_{p1}(\frac{n_{p+}}{n})^{-\frac{1}{2}} & \cdots & u_{pk}(\frac{n_{p+}}{n})^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix}$$

$$C_2^{(R_{BL})} = \begin{bmatrix} r_{11} & \cdots & r_{1c} \\ \vdots & r_{ij} & \vdots \\ r_{l1} & \cdots & r_{lc} \end{bmatrix} \begin{bmatrix} v_{11}(\frac{n_{+1}}{n})^{-\frac{1}{2}} & \cdots & v_{1k}(\frac{n_{+1}}{n})^{-\frac{1}{2}} \\ \vdots & \vdots & \vdots \\ v_{q1}(\frac{n_{+q}}{n})^{-\frac{1}{2}} & \cdots & v_{qk}(\frac{n_{+q}}{n})^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix}$$

$$C_1^{(R_H)} = \begin{bmatrix} \tilde{r}_{11} & \cdots & \tilde{r}_{1c} \\ \vdots & \tilde{r}_{ij} & \vdots \\ \tilde{r}_{l1} & \cdots & \tilde{r}_{lc} \end{bmatrix} \begin{bmatrix} u_{11}(\frac{n_{1+}}{n})^{-\frac{1}{2}} & \cdots & u_{1k}(\frac{n_{1+}}{n})^{-\frac{1}{2}} \\ \vdots & \vdots & \vdots \\ u_{p1}(\frac{n_{p+}}{n})^{-\frac{1}{2}} & \cdots & u_{pk}(\frac{n_{p+}}{n})^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix}$$

$$C_2^{(R_H)} = \begin{bmatrix} \tilde{r}_{11} & \cdots & \tilde{r}_{1c} \\ \vdots & \tilde{r}_{ij} & \vdots \\ \tilde{r}_{l1} & \cdots & \tilde{r}_{lc} \end{bmatrix} \begin{bmatrix} v_{11}(\frac{n_{+1}}{n})^{-\frac{1}{2}} & \cdots & v_{1k}(\frac{n_{+1}}{n})^{-\frac{1}{2}} \\ \vdots & \vdots & \vdots \\ v_{q1}(\frac{n_{+q}}{n})^{-\frac{1}{2}} & \cdots & v_{qk}(\frac{n_{+q}}{n})^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix}$$

sendo $(\lambda_1, \cdots, \lambda_K)$ autovalores, $(u_{11}, \cdots, u_{pk})$ e $(v_{11}, \cdots, v_{qk})$ autovetores calculados a partir da decomposição de valores singulares.

A análise da proximidade das coordenadas modificadas com a incorporação dos resíduos em relação às coordenadas convencionais foi realizada pelo coeficiente de correlação cofenético. A matriz fenética foi determinada por $\boldsymbol{S}$, matriz de dissimilaridade, onde cada coordenada foi dada pela distancia euclidiana entre os dados e a matriz cofenética $\boldsymbol{C}$ das distâncias cofenéticas para um agrupamento hierárquico gerado pelo método de ligação média, sendo esse utilizado com o propósito de evitar encadeamentos. Assim o coeficiente de correlação cofenético é dado por:

$$r_{cof} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-\bar{c})(s_{ij}-\bar{s})}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-\bar{c})^2}\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(s_{ij}-\bar{s})^2}},$$

em que c_{ij} é o valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética e s_{ij} valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz de similaridade em que $\overline{c} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}$ e $\overline{s} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s_{ij}$.

Para a obtenção dos resultados, foram computadas as médias do coeficiente de correlação cofenético, a inércia total e os escores considerando 2.000 realizações Monte Carlo, utilizando o *script* que se encontra no Anexo B.

3 Resultados e Discussão

3.1 Análise dos Escores com a inclusão dos Resíduos

Os resultados apresentados na Tabela 2 evidenciam que a incorporação dos resíduos no cálculo dos escores proporcionou uma melhoria para tabelas de dimensão 5×5 no sentido que as coordenadas apresentaram maior correlação aos dados agrupados de forma hierárquica considerando o método de ligação média. Observou-se um resultado mais promissor quando as amostras binomiais foram geradas considerando alto grau de correlação $(\rho = 0{,}8)$ para tabelas de dimensões maiores (10×10).

A correlação da matriz de dissimilaridade obtida pela distância euclidiana aplicada nos dados e da matriz obtida pelo método de ligação média foi maior na abordagem convencional ao considerar as amostras binomiais geradas $\pi = 0{,}2$ em todos os graus de correlação.

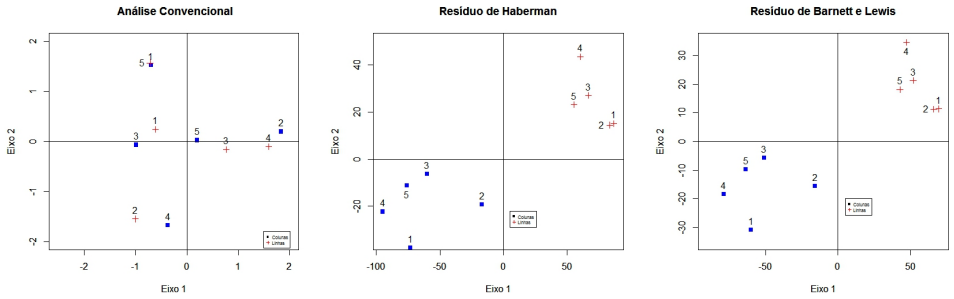
Tabela 2 Melhores desempenhos do coeficiente de correlação cofenética entre as matrizes de distância euclidiana e a matriz cofenética pelo método ligação média *(average)*.

Dimensão	Proporção da Binomial	Grau de Correlação (ρ)	Convencional	Barnett e Lewis	Haberman
5×5	0,2	0,2	0,2485	**0,6607**	0,6488
	0,5	0,8	0,1383	0,8215	**0,8246**
	0,9	0,5	0,5621	0,7916	**0,7977**
10×10	0,2	0,2	**0,6199**	0,1587	0,1554
	0,5	0,8	**0,5144**	0,4372	0,4388
	0,9	0,5	**0,4703**	0,3516	0,3513

Em relação aos mapas perceptuais, considerando as tabelas de dimensão 5×5 (Figura 1) e 10×10 (Figura 2) ilustra-se um caso em que o coeficiente de correlação cofenético (Tabela 2) apresentou uma estimativa superior ao utilizar os resíduos.

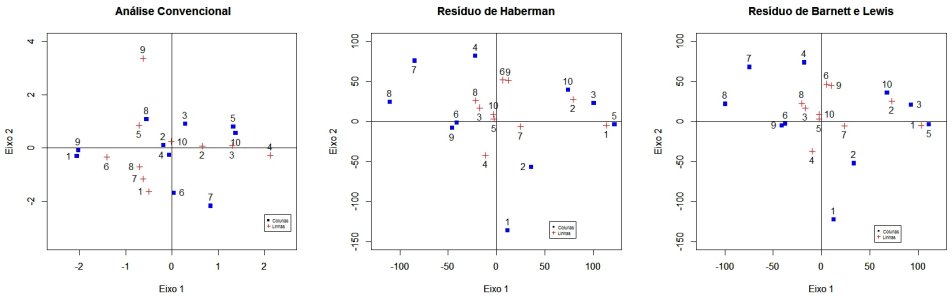
Figura 1 Mapas Perceptuais gerados a partir das coordenadas da análise de correspondência convencional e com a incorporação dos resíduos em tabelas de dimensão 5×5

.



(a) $(\rho = 0{,}5, \pi = 0{,}9)$ (b) $(\rho = 0{,}5, \pi = 0{,}9)$ (c) $(\rho = 0{,}5, \pi = 0{,}9)$

Figura 2 Mapas Perceptuais gerados a partir das coordenadas da análise de correspondência convencional, e com a incorporação dos resíduos, em tabelas de dimensão 10×10.



(a) $(\rho = 0{,}8, \pi = 0{,}5)$ (b) $(\rho = 0{,}8, \pi = 0{,}5)$ (c) $(\rho = 0{,}8, \pi = 0{,}5)$

Os resultados ilustrados na Figura 1 indicam que a incorporação dos resíduos do cálculo dos escores proporcionou uma melhor discriminação dos objetos. Aumentando a dimensão das tabelas de contingência (Figura 2) o mesmo efeito foi observado, entretanto, os mapas apresentaram uma tendência assimétrica em relação ao centroide.

Em ambas as situações, os resultados corroboraram com as recomendações feitas por Beh (2012), as quais se baseiam na violação da hipótese de independência e que a variabilidade dos pontos é afetada, sendo refletida na assimetria dos quadrantes. Para os demais casos, as coordenadas utilizadas para gerar os mapas encontram-se no Anexo C.

Em relação à porcentagem da inércia restituída nos eixos, dado os casos ilustrados nas Figuras 1 e 2 por meio dos resultados ilustrados na Tabela 3, nota-se que, para os dois primeiros eixos, os resultados foram adequados por apresentar elevadas proporções. É possível observar que quando aumenta a dimensão da tabela de contingência a proporção de explicação fica reduzida. Para tabelas de dimensão 5×5 onde se obtiveram melhores resultados com a incorporação dos resíduos, a proporção de explicação é maior em relação as tabelas de dimensão 10×10 onde a análise convencional obteve melhores resultados.

Tabela 3 Decomposição da inércia

	Parâmetros de melhor desempenho					
	$(\pi = 0{,}9; \rho = 0{,}5)$ (5×5)			$(\pi = 0{,}5; \rho = 0{,}8)$ (10×10)		
Inércia	Inércia	Proporção	Prop.Acum(%)	Inércia	Proporção	Prop.Acum(%)
$Eixo\ 1$	0,0006	0,5081	0,5081	0,2489	0,3483	0,3483
$Eixo\ 2$	0,004	0,3594	0,8675	0,1495	0,2092	0,5575
$Eixo\ 3$	0,0001	0,1143	0,9818	0,1215	0,1700	0,7274
$Inercia\ Total$	0,0012	–	–	0,7148	–	–

3.2 Exemplo de Aplicação

A classificação do café por defeitos e tipos é feita na forma de contagem de grãos defeituosos ou das impurezas contidas numa amostra de $300g$ de grãos beneficiados. Esta classificação obedece à tabela de Classificação Oficial Brasileira (COB), de acordo com a qual cada tipo corresponde a um maior ou menor número de defeitos encontrados em uma amostra de café. Na classificação de cafés são considerados defeitos os grãos imperfeitos e as impurezas (defeitos intrínsecos e extrínsecos).

A comercialização do café é feita segundo diversas classificações em vigor. As mais importantes são as classificações por peneira, por tipo, por bebida entre outras. Para esse exemplo a base de dados contém informações sobre os defeitos intrínsecos que são aqueles contidos no grão de café, causados pela utilização incorreta dos processos agrícolas e industriais e modificações de origem fisiológica ou genética, como por exemplo grão preto, grão ardido, grão brocado, grão verde, entre outros.

Percebe-se que muitos produtores de café não se preocupam com o percentual de defeito no momento do preparo do café para a comercialização, o que pode impactar nos preços praticados pelo mercado, sendo que o preço é um dos fatores decisivos na atividade cafeeira.

Um estudo sobre proporção de grãos defeituosos em peneiras $17/18$ é importante, pois os grãos de café que são classificados nessas peneiras. Por serem grãos graúdos e de maior valor comercial, são os de maior interesse para o exportadores de café, sendo que quanto menor a proporção de grãos café defeituosos menor será a influência no preço de venda por parte do produtor. A Tabela 4 con-

tém informações sobre a proporção de grãos de café defeituosos em uma amostra com $300g$, classificados por algum tipo de defeito encontrado após catação em peneiras $(17/18)$.

Tabela 4 Contagem de grãos retidos em relação aos defeitos em peneiras (17/18)

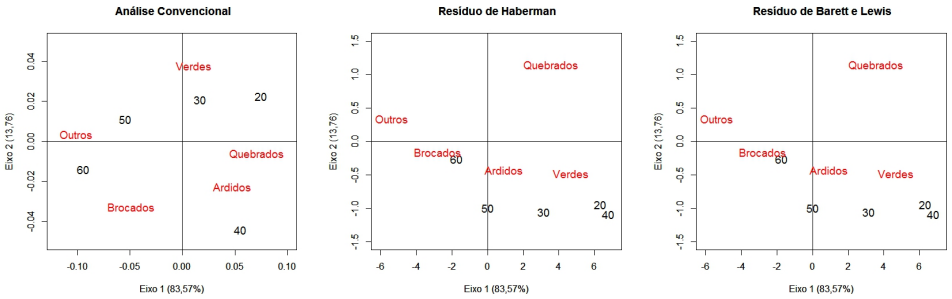
Proporção de defeitos na amostra(%)	**Tipo de defeito do grão**				
	Verdes	**Ardidos**	**Quebrado**	**Brocados**	**Outros**
20	18	21	26	21	18
30	11	11	8	8	2
40	25	24	23	18	7
50	7	10	2	11	11
60	3	12	7	22	18

Preliminarmente, aplicou-se o teste qui-quadrado, no qual a hipótese de independência foi rejeitada (valor-p < 0,05). Assim, justificamos um exemplo que viole a suposição básica da análise de correspondência, sendo, portanto, uma situação que sugere que a incorporação dos resíduos possam apresentar alguma melhoria na identificação das associações.

A interpretação gráfica dos mapas perceptuais Figura 3, ficará restrita aos dois primeiros eixos, uma vez que os eixos apresentam elevados valores de inércia. Em relação aos mapas perceptuais, percebe-se que os mapas são simétricos em relação ao centroide e que, quando utilizado os resíduos, tem-se melhor discriminação entre a proporção de grãos e seu defeitos, o que se percebe para os grãos Brocados, que está fortemente relacionado a 60% dos grãos e que os grãos quebrados não possuem nenhuma associação com as proporções estudadas, ou seja, são independentes das proporções.

O grão quebrado é um defeito que ocorre pela secagem excessiva do grão de café ou também pela secagem rápida e secadores mecânicos, o que o difere dos demais defeitos que são decorrentes geralmente de origem genética, fisiológica e pragas. O grão quebrado é o de menor gravidade em relação aos demais defeitos, não afetando tanto a qualidade do café.

Figura 3 Mapas Perceptuais gerados a partir das coordenadas da análise de correspondência convencional, e com a incorporação dos resíduos sobre a proporção de grãos de café defeituosos em peneiras $17/18$.



4 Conclusões

Para os cenários simulados, os resultados indicaram que é viável incorporar os resíduos na obtenção das coordenadas, incluindo as situações que a hipótese de independência entre as categorias de linhas e ou colunas foram violadas. O efeito desses resíduos evidenciou uma melhor discriminação dos objetos, com pequenas diferenças em relação a dimensão das tabelas.

O uso dos dois resíduos utilizados apresentou resultados similares, o que pode se observado pelos coeficientes de correlação cofenéticas e mapas perceptuais gerados.

A aplicação da análise de correspondência simples modificada com a incorporação dos resíduos a dados reais é viável, pois percebe-se, na amostra estudada, uma melhor discriminação entre a proporção de grãos e seus defeitos quando comparada com a análise de correspondência simples convencional.

Referências

BARNETT, V.; LEWIS, T. **Outliers in Statistical Data**. 3. ed. London: Academic Press, 1994.

BEH, E. J. Simple correspondence analysis: a bibliographic review. **International**

Statistical Review, Wiley Online Library, v. 72, n. 2, p. 257–284, 2004.

BEH, E. J. Simple correspondence analysis using adjusted residuals. **Journal of Statistical Planning and Inference**, v. 142, n. 4, p. 965 – 973, 2012.

BENZÉCRI, J.-P. **Correspondence Analysis Handbook**. 2. ed. London: Taylor & Francis, 1992.

BLASUS, J. et al. Special issue on correspondence analysis and related methods. **Computational Statistics e Data Analysis, New York**, v. 53, n. 8, p. 3103–3106, 2009.

CIRILLO, M. A.; RAMOS, P. de S. Goodness-of-fit tests for modified multinomial logit model. **Chilean Journal of Statistics**, v. 5, n. 1, p. 73–85, 2014.

GREENACRE, M. Correspondence analysis in medical research. **Statistical methods in medical research**, SAGE Publications, v. 1, n. 1, p. 97–117, 1992.

GREENACRE, M. **Correspondence analysis in practice**. 2. ed. Orlando: CRC press, 2007.

GREENACRE, M. J. **Theory and Applications of Correspondence Analysis**. 2th. ed. Orlando: Academic Press, 1984.

GUEDES, T. A. et al. Seleção de variáveis categóricas utilizando análise de correspondência e análise correspondência e análise procrustes procrustes procrustes. **Acta Scientiarum**, v. 21, n. 4, p. 861–868, 1999.

HABERMAN, S. J. The analysis of residuals in cross-classified tables. **Journal of Mathematical Modelling and Algorithms**, v. 29, n. 1, p. 205–220, 1973.

LEE, A. H.; YICK, J. S. Theory & methods: A perturbation approach to outlier detection in two-way contingency tables. **Australian & New Zealand Journal of**

Statistics, Blackwell Publishers Ltd, v. 41, n. 3, p. 305–314, 1999.

TALLIS, G. M. The use of a generalized multinomial distribution in the estimation of correlation in discrete data. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 24, n. 2, p. 530–534, 1962.

VELOSO, M. V. de S.; CIRILLO, M. A. Principal components in the discrimination of outliers: A study in simulation sample data corrected by pearson's and yatesÂ´s chi-square distance. **Acta Scientiarum**, v. 38, n. 2, p. 193–200, 2016.

UNIVERSIDADE FEDERAL DE LAVRAS

Artigo 2

Classificação granulométrica de cafés: uma proposta de avaliação utilizando a análise de correspondência aplicada a dados composicionais

Versão preliminar de artigo - Sujeito a alterações pelo corpo editorial da revista

LAVRAS - MG

2016

RESUMO

Ao verificar a classificação de cafés originada por tamanho de grãos em peneiras, obtém-se uma tabela com dados composicionais. Dados composicionais consistem em um conjunto de vetores denominados composições, cujos elementos ou componentes são positivos e definidos no intervalo $(0,1)$ cuja soma vale 1. A restrição de soma constante dos dados composicionais impõe limitações para aplicação de técnicas estatísticas multivariadas usuais, sendo necessária a utilização de transformações logarítmicas. Visto que, na análise de classificação de qualidade do café, é importante considerar a composição da amostra por peneiras para atribuição de preços de mercado, o presente estudo foi desenvolvido com o objetivo de aplicar a técnica multivariada de análise de correspondência múltipla a dados composicionais, para um estudo comparativo do efeito das transformações logarítmicas aditiva, centrada e isométrica na avaliações da granulometria de cafés. Concluiu-se que a utilização das transformações logarítmicas é adequada para análise de dados composicionais quando utilizada a análise de correspondência múltipla, pois retira as limitações impostas pelos dados composicionais. Dentre as transformações utilizadas no presente trabalho, a transformação logarítmica isométrica foi a que discriminou mais amostras de café em relação as categorias dos componentes.

Palavras-chave: análise multivariada, classificação de grãos, transformações logarítmicas.

1 Introdução

1.1 Aspectos gerais da classificação granulométrica de cafés

A qualidade dos grãos de café é bastante diversificada em função das diferentes condições de produção (ABRAHAO et al., 2009). No Brasil são vários os critérios utilizados na classificação dos grãos de café contemplando aspectos físicos, como o tamanho, a cor, o número de defeitos e as características sensoriais da bebida. É através dessa classificação que os valores comerciais são definidos, tanto no mercado interno quanto no externo.

Para que ocorra essa classificação do café, é de extrema importância a utilização de materiais e equipamentos apropriados, como: balança, amostra de café desejada, jogo de peneiras (utilizadas para determinar o tamanho e a forma das favas dentro da amostra analisada) entre outros equipamentos; e, para determinar a sua qualidade, deve-se analisar os diversos fatores que indicam o seu grau de aceitação pelo mercado consumidor dentro de um escala de comparação (LOURES; ALVES; JÚNIOR, 2007).

Classificações simples, como a medida do tamanho dos grãos, realizada através de peneiras, são capazes de indicar potencial produtivo dos cultivares e é um critério importante na comercialização do café. A importância se dá principalmente pelo rendimento e pela possibilidade de uniformizar os grãos para o processo de torração (MATIELLO et al., 2010).

A Instrução Normativa número 08 de 11 de junho de 2003 do Ministério da Agricultura (BRASIL, 2003) especifica as normas das características mínimas de qualidade para a classificação do Café Beneficiado Grão Cru de acordo com o tamanho dos grãos e da dimensão dos crivos circulares das peneiras. O café é primeiro categorizado como *Coffea arabica* e *Coffea canephora*, depois, o grão é classificado segundo o seu formato e a sua granulometria. As principais subcategorias de formato são: o chato, que é constituído de grãos com superfície dorsal convexa e a ventral plana ou ligeiramente côncava com a ranhura central no sentido longitudinal e o moca, constituído de grãos com formato ovoide, também com ranhura central no sentido longitudinal.

Em se tratando de dados composicionais, nota-se que a aplicação relacionada à classificação granulométrica é de suma importância para que novos resultados sejam associados à qualidade do café, com vistas à comercialização e internacionalização dos produtos derivados.

1.2 Definição de dados composicionais e relação com a classificação granulométrica

Dados composicionais consistem em um conjunto de vetores denominados composições, cujos elementos ou componentes, $x_1, x_2, \ldots, x_D$, são positivos e definidos no intervalo $(0,1)$.

Segundo Aitchison (1986), os dados composicionais representam frações de algum todo e satisfazem a restrição de que a soma dos componentes é uma constante. Assim tem-se que:

$$x_1 \geq 0, x_2 \geq 0, \ldots, x_D \geq 0$$

e

$$x_1 + x_2 + x_3 + \ldots + x_D = 1.$$

Essa restrição impõe limitações para a aplicação de técnicas estatísticas multivariadas usuais.

O espaço amostral (ou conjunto de valores possíveis) restrito no qual são definidos os dados composicionais é conhecido como $D-$simplex, de dimensão igual ao número de componentes, dado pela seguinte equação (PAWLOWSKY et al., 2010):

$$S^D = \{x = [x_1; \ldots; x_D \mid x_i \geq 0 \; e \; \sum_{j=1}^{D} x_j = K]\},$$

com $K = 1{,}100{,}10^6{,}10^9$(proporção,%,ppm,etc).

A operação que define o fechamento de uma composição em uma constante K para cada amostra (linha) é dada pela seguinte equação (PAWLOWSKY et al., 2010):

$$C(x) = \left[\frac{K \times x_1}{\sum_{i=1}^{D} x_i}, \frac{K \times x_2}{\sum_{i=1}^{D} x_i}, \frac{K \times x_3}{\sum_{i=1}^{D} x_i}, \cdots, \frac{K \times x_D}{\sum_{i=1}^{D} x_i}\right]$$

em que:

- $C(x) =$ operação de fechamento;
- $K =$ a constante de fechamento (geralmente 1);
- x_i é o valor do i-ésimo componente de uma amostra.

Uma tabela com dados composicionais gera uma tabela de múltipla entrada, em que as variáveis correspondem a cada um dos componentes x_1,$x_2, \ldots$,x_D e o número de categorias de cada uma dessas é dado pela contagem de valores diferentes em cada coluna. Assim, inicialmente, para dados composicionais, tem-se o número total de variáveis, mas não o número de categorias, correspondente a cada variável.

Os dados apresentados dessa forma mostram obrigatoriamente certa correlação, já que o aumento em importância de determinado componente implica necessariamente a diminuição dos demais. Segundo Aitchison (1986), uma característica importante neste tipo de dados é que eles carregaram informação relativa, o que acarreta dependência entre os componentes, ficando inviável a aplicação da análise de correspondência múltipla para os dados composicionais originais. Por essa razão, é proposto o uso de transformações nos dados originais.

Na classificação granulometrica de cafés, é obtida uma tabela das proporções de grãos de cada componente. Assim, por exemplo, a cada $300g$ de café para uma determinada amostra, teremos o percentual relativo de todos os componentes onde a soma final do percentual de cada componente é uma constante, o que configura dados composicionais.

A relação da análise de dados composicionais com a classificação granulométrica de cafés é verificada pelo fato de que essa classificação do café é uma fase muito importante no processo da produção e comercialização uma vez que o produtor possa conhecer o sistema, avaliar o seu produto e as prioridades de sua comercialização. Seguindo essa motivação, enunciamos o objetivo deste trabalho na proposta de aplicar a técnica análise de correspondência para dados composicionais provenientes de uma avaliação granulométrica de cafés *Coffea arabica* com o propósito de apresentar o uso dessa técnica como um alternativa promissora de ser aplicada a dados dessa natureza.

2 Metodologia

2.1 Descrição do conjunto de dados

O conjunto de dados utilizado nesse trabalho foi estruturado de acordo com informações de 40 amostras de café *Coffea arabica* oriunda da safra 2014/2015 produzido na região do Sul de Minas Gerais.

A classificação de grãos seguiu as normas específicas, sendo obtida a proporção de grãos de cada uma das amostras, conforme o tamanho determinado pelo tamanho dos crivos das peneiras.

Deve-se ressaltar que o jogo de peneiras é diferente conforme o formato do grão, ou seja, há um jogo específico para grãos chatos e outros para os grãos mocas, assim, entre as subcategorias chato e moca, há ainda subdivisão em relação à granulometria, denominando-se os grãos em graúdo, médio e miúdo segundo tamanho dos crivos das peneiras (Tabela 1).

Tabela 1 Classificação granulométrica dos grãos de café segundo tamanho dos crivos das peneiras

Formato	Classificação	Tamanho dos crivos
Chato	Graúdo	17,18,19 e 20
	Médio	15 e 16
	Miúdo	12,13 e 14
Moca	Graúdo	12 e 13
	Médio	10 e 11
	Miúdo (moquinha)	8 e 9

O restante dos grãos pode ser classificado como fundo de peneira (FDO) ou catação. O FDO é o percentual relativo ao vazamento da peneira, e a catação representa a proporção de grãos defeituosos, que é calculada para avaliar a qualidade do café e indicar os pontos onde o produtor pode amenizar e evitar possíveis defeitos. São considerados defeituosos os grãos com imperfeições causadas pelas utilização incorreta dos processos agrícolas e modificações de origem fisiológica ou genética, ou, ainda, as impurezas presentes na amostra, tais como grãos ou sementes de outra espécie.

As porcentagens de tamanho de grãos avaliadas foram agrupadas em seis

componentes: chato graúdo (CG), chato médio(CME), chato miúdo(CMI), moca (MK), catação (CAT) e fundo de peneira (FDO), conforme modelo da Tabela 2 obtendo-se uma tabela com dados composicionais.

Tabela 2 Dados Composicionais da Granulometria de Café conforme classificação por formato e tamanho

Amostra	**CG**	**CME**	**CMI**	**MK**	**FDO**	**CAT**
1	27,5	24,2	19,9	8,4	1,8	18,2
2	30,6	32,6	4,8	7,6	1,8	22,6
3	20,0	22,0	28,0	8,0	3,0	19,0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
40	0,5	18,5	51,0	8,0	14,0	8,0

2.2 Transformações utilizadas para realização da análise de dados composicionais e implementação computacional

As transformações logarítmicas utilizadas para a análise de dados composicionais foram as transformações logarítmica aditiva (alr), logarítmica centrada (clr) e logarítmica isométrica (ilr) .

Para transformação alr aplicou-se o logaritmo natural $\ln$ na divisão de cada componente pela última componente x_D. Definindo x como uma composição de D-partes no simplex S^D, então:

$$alr(\mathbf{x}) = (y_1, y_2, \ldots, y_{D-1}) = \left[\ln \frac{x_1}{x_D}; \ln \frac{x_2}{x_D}; \ldots; \ln \frac{x_{D-1}}{x_D}\right].$$

A escolha do componente x_D como variável razão é subjetiva, podendo ser qualquer um dos componentes contidos na amostra. Essa subjetividade interfere no cálculo da transformação logarítmica aditiva, levando a interpretações diferentes de acordo com a escolha do componente x_D.

Para voltar ao espaço simplex saindo do espaço real, aplica-se o processo inverso dado pela seguinte equação:

$$x = C(exp(alr_1(x), alr_2(x), \ldots, alr_{D-1}(x),0))$$

onde C é a operação de fechamento.

No caso da transformação clr o resultado de uma observação $x \in S^D$ são os dados transformados $y \in \mathbb{R}^D$ através da média geométrica $g(x)$ dada por:

$$g(x) = \sqrt[D]{x_1 \times x_2 \times \cdots \times x_D}\,,$$

Desta forma tem-se:

$$clr(\mathbf{x}) = (y_1, \ldots, y_D) = \left[\ln \tfrac{x_1}{g(x)}, \ldots, \ln \tfrac{x_D}{g(x)}\right].$$

Para voltar ao espaço simplex a partir do espaço real, o processo inverso é dado pela seguinte equação:

$$x = C(\exp(clr_1(x), clr_2(x), \ldots, clr_D(x))).$$

onde C é a operação de fechamento.

A transformação irl baseia-se na escolha de uma base ortonormal de S^D, que é uma série de composições $e_1, e_2, \ldots, e_{D-1}$ tal que o produto interno $\langle e_i, e_j \rangle = 0$ para $i \neq j$ e a norma de e_i seja $\parallel e_i \parallel = 1$ (RUBIO, 2014). Para uma base fixa, as coordenadas de uma composição são obtidas usando a função:

$$ilr(\mathbf{x}) = (y_1, y_2, \ldots, y_{D-1}) = (\langle x, e_1 \rangle, \langle x, e_2 \rangle, \ldots, \langle x, e_{D-1} \rangle)\,.$$

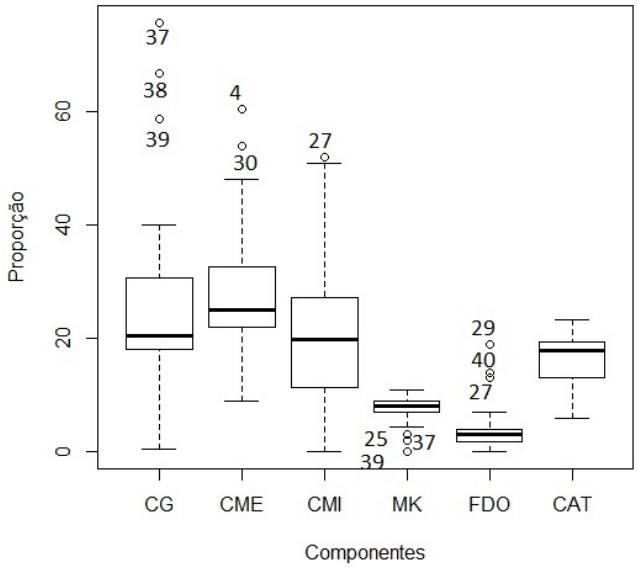
Uma vez organizada a tabela com os dados transformados, foi realizada a contagem de categorias de acordo com cada classificação, estruturando os resultados em uma tabela de múltiplas entradas para realização da análise de correspondência. As análises foram realizadas utilizando os pacotes do *software* R *"compositions"*, desenvolvido por Boogaart e Tolosana-Delgado (2008), para efetuar as referidas transformações e operações de fechamento, e o pacote *"ca"* para obtenção das inércias de cada categoria, coordenadas dos mapas perceptuais e contribuições das amostra para cada transformação utilizada.

3 Resultados e Discussão

Inicialmente, realizou-se uma análise descritiva dos dados, pela qual pode-se observar que os componentes apresentam níveis médios diferentes da proporção de grãos em cada amostra na (Figura 1).

Quanto à dispersão das proporções das classes componentes das amostras dos grãos, observa-se que para MK e FDO estas são menores, enquanto que CMI é a maior. Para CG observa-se grande assimetria, o que indica diferença acentuada entre as amostras.

Figura 1 Boxplot da proporção de grãos para cada componente



A proporção de alguns componentes se destaca em relação a algumas amostras, como no caso das amostras $37, 38$ e 39 em que a proporção de grãos chatos graúdos é mais elevada. As amostras 4, 30 e 27 possuem proporção de grãos chato médio e chato miúdo mais elevadas (respectivamente). As amostras 25,37 e 39 destacam-se por apresentarem níveis de grãos do tipo moca muito pequenos, e as amostras 27,29 e 40 apresentam níveis de grãos retidos no fundo de peneira mais elevados quando comparadas com as demais amostras.

Para cada uma das transformações utilizadas, os dados foram inicialmente organizados a partir da contagem referente a classificação das 40 amostras de café produzidas no sul de Minas Gerais, de acordo com as categorias geradas.

Devido à diferença na estrutura das transformações, o número de categorias geradas em cada uma delas é também diferente, sendo este maior para a transformação clr já que esta preserva o número de variáveis originais (Tabela 3). De acordo com a Tabela 3, nas três transformações obteve-se um alto numero de categorias por classificação, ou seja, a proporção de grãos de café nas 40 amostras é diversificada entre os diferentes formatos e tamanhos, indicando baixa uniformidade dos grãos, o que prejudica o valor comercial.

Tabela 3 Número de categorias por classificação de tamanho e formato dos grãos

Classificação	Transformações		
	alr	clr	ilr
Clato graúdo (CG)	17	20	16
Clato médio (CME)	14	15	19
Clato miúdo (CMI)	17	17	16
Moca (MK)	20	12	15
Catação (CAT)	-	15	-
Fundo de Peneira (FDO)	18	18	16
Total de Categorias	86	97	82

Nota-se que, como consequência do elevado número de categorias obtido, o percentual de explicação em cada um dos autovalores é baixo. Assim, torna-se necessário avaliar a inércia de cada categoria nos eixos para verificar a importância relativa em cada caso (Tabela 4).

Tabela 4 Primeiros autovalores e respectivos percentuais de explicação obtidos em cada transformação.

		Primeiros Eixos					
Transfomação		1	2	3	4	5	6
alr	Autovalor	1	0,968	0,965	0,945	0,919	0,892
	(%)	6,17	5,98	5,96	5,84	5,67	5,51
clr	Autovalor	1	1	0,885	0,879	0,857	0,821
	(%)	6,59	6,59	5,84	5,8	5,65	5,41
ilr	Autovalor	1	0,951	0,928	0,882	0,858	0,811
	(%)	6,49	6,18	6,03	5,73	5,57	5,27

Para a transformação alr foram obtidas 86 categorias entre as seis classi-

ficações padrão, as quais referem-se ao valor obtido do percentual de cada classe após a transformação, sendo obtidos 11 grupos de categoria em função da magnitude da inércia. A inércia obtida para cada uma das categorias indica o grau de importância em relação a dispersão das amostras (Anexo E). No caso da transformação alr a menor inércia foi igual a $0{,}061$ obtida para a categoria equivalente nesta transformação à porcentagem de grão chato graúdo em relação a catação em proporções iguais, ou seja, em ln(CG/CAT)=0. Assim, pode-se dizer que proporções iguais destas duas classes são as mais usuais entre as amostras. Já a maior inércia foi de $0{,}195$ para razões entre classes com as menores proporções de grão chato miúdo, moca e FDO, em relação a catação. Assim, é possível afirmar que, quanto maior a proporção de catação, maior é o índice de grãos de menor granulometria, ou seja, de menor qualidade.

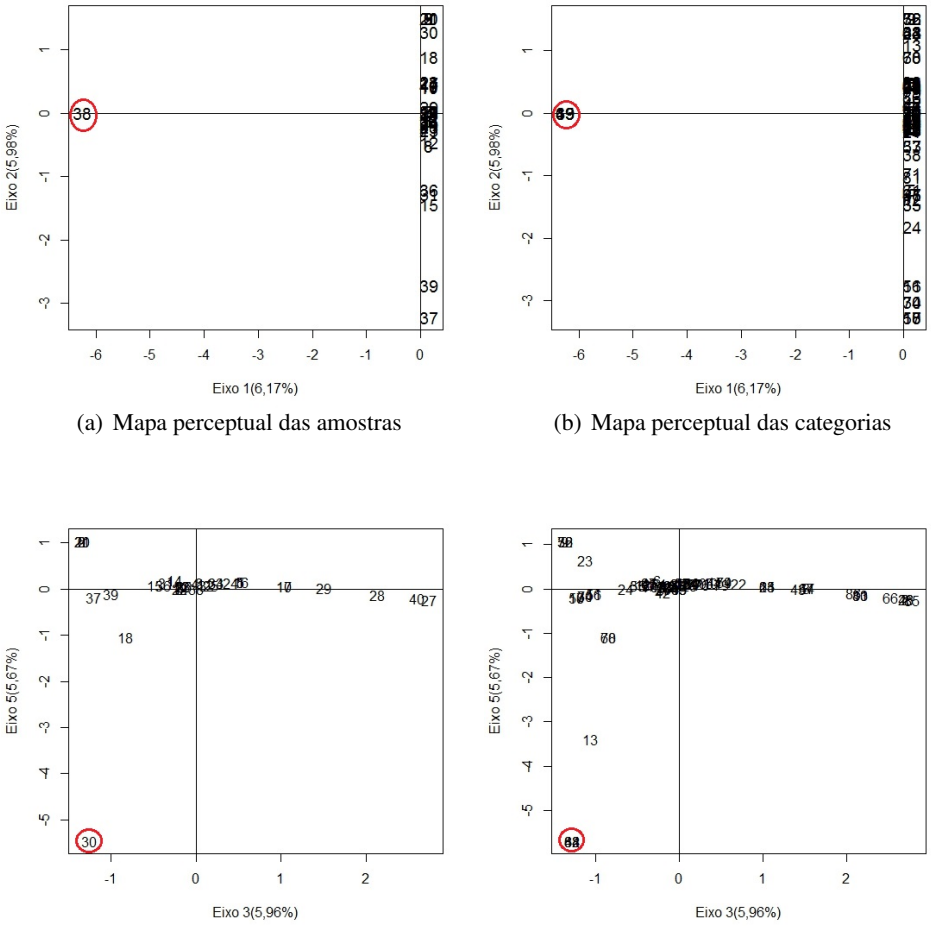
Na Figura 2 têm-se os mapas perceptuais das amostras de café e suas categorias para os eixos com maior contribuição da dispersão. Observa-se que a amostra 38 e suas categorias se destacam nos mapas perceptuais dos eixos 1 e 2 (Figura 2a e 2b). A amostra de café 38 destaca-se pela elevada proporção de grãos chatos graúdos, o que é visto como fator positivo, pois possuem maior valor comercial.

Quanto às contribuições das amostras de café para a variância dos eixos, observou-se que a amostra de café 38 está fortemente relacionada com o primeiro eixo, sendo sua contribuição de $97{,}50\%$. As amostras de café que mais contribuíram para a variância do segundo eixo foram as amostras 37 e 39, com contribuição de $26{,}79\%$ e $19{,}07\%$, respectivamente. As amostras que estão mais relacionadas aos eixos 3 e 4 são 27, 28, 40 e 37, sendo suas contribuições, respectivamente, de $19{,}38\%$, $11{,}68\%$, $17{,}41\%$ e $17{,}52\%$. Em relação ao eixo 5, a amostra 30 é a que está fortemente relacionada, com contribuição de $81{,}19\%$. A interpretação gráfica foi restrita aos eixos 1 e 2, e aos eixos 3 e 5, uma vez que a porcentagem da inércia restituída nos eixos são as de maior proporção (Anexo D).

Com relação aos mapas perceptuais dos eixos 3 e 5 da Figura 2 (c e d) observa-se que a amostra 30 foi a que mais se destacou, neste caso negativamente, pois há grande presença de grãos médios e mocas, sendo uma amostra que possui valor comercial inferior. Portanto as amostras de café 38 e 30 foram as que mais se

destacaram quando utilizada a transformação logarítmica aditiva, a primeira pela sua alta qualidade e a segunda de qualidade inferior.

Figura 2 Mapa perceptual das categorias e amostras em relação aos eixos com maiores contribuições, transformação alr.



(a) Mapa perceptual das amostras

(b) Mapa perceptual das categorias

(c) Mapa perceptual das amostras

(d) Mapa perceptual das categorias

Para a transformação clr foram obtidas 97 categorias entre as seis classificações padrão (Tabela 3). Percebe-se, para a transformação clr, o agrupamento das categorias em 9 grupos em função da magnitude da inércia obtida, mantendo-

se neste caso as categorias da componente CAT (Anexo E).

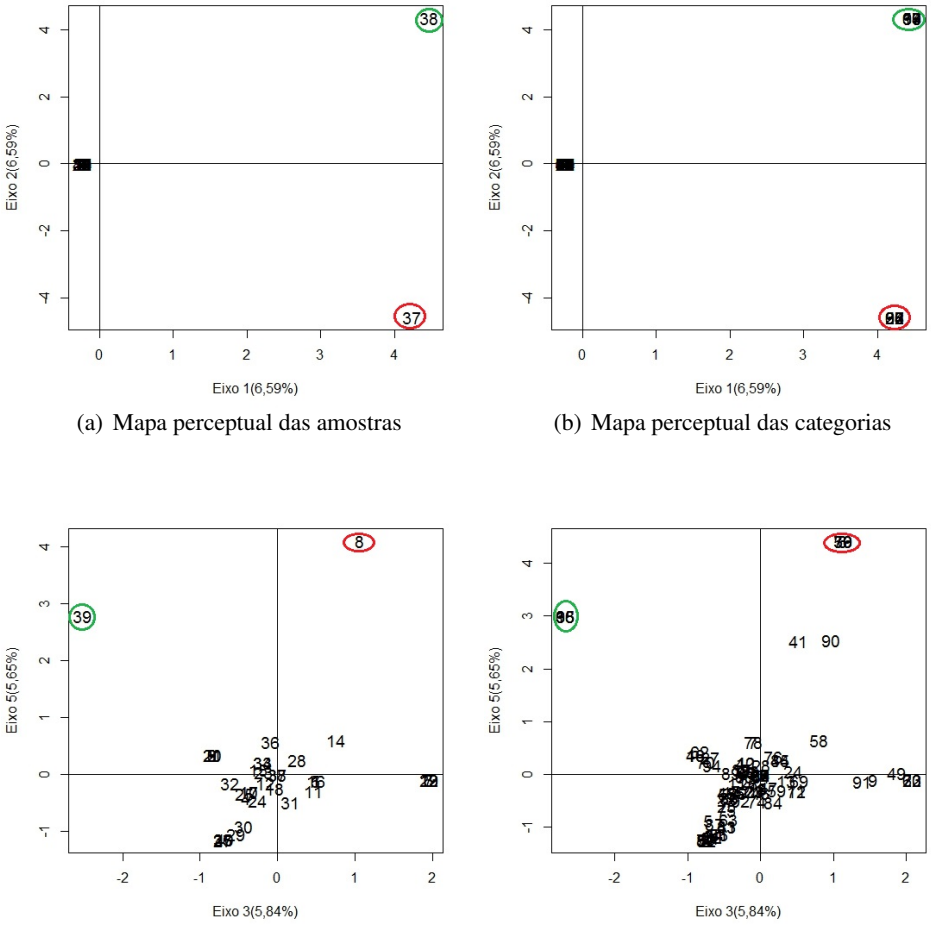
Na Figura 3 têm-se os mapas perceptuais das amostras de café, e suas categorias para os eixos que possuem maior contribuição. Quanto a contribuições das amostras de café, percebe-se que as amostras de café 37 e 38 estão fortemente relacionadas com os dois primeiros eixos, sendo suas contribuições de $44{,}93\%$ e $50{,}07\%$ para o primeiro eixo e $47{,}43\%$ para o segundo eixo, sendo que as demais possuem contribuição para o eixos quase nulas. Os eixos 3 e 5 tiveram varias amostras com proporção de explicação considerada significativa, variando de $10{,}81\%$ a $48{,}96\%$, sendo a amostra 8 de maior destaque em relação ao eixo 5 e a amostra 39 para o eixo 3 (Tabela 8 - Anexo D).

Observa-se que as amostras 37 e 38 e suas categorias se destacam nos mapas perceptuais dos eixos 1 e 2 (Figura 3a e 3b). As amostras 37 e 38 possuem alta proporção de grãos chatos graúdos, o que é bem visto pelo mercado, pois além de serem grãos graúdos possuem maior valor comercial.

Com relação aos mapas perceptuais dos eixos 3 e 5 da Figura 3 (c e d) observa-se que as amostras 8 e 39 foram as que mais se destacaram, pois são as que possuem maior contribuição aos eixos. Essas amostras se destacam pela alta proporção de grãos chatos médios, chatos miúdos e catação, sendo amostras inferiores em relação às amostras destacadas nos primeiros eixos. Portanto as amostras de café 8, 37, 38 e 39 são as que mais se destacaram quando utilizada a transformação logarítmica centrada, sendo que as amostras 37 e 38 são de alta qualidade e valor comercial e as amostras 8 e 39 de qualidade inferior.

Deve-se ressaltar que a porcentagem de explicação nos eixos nesta transformação, $13{,}18\%$, é maior que na transformação alr, $12{,}15\%$. Além disso, a transformação alr não preserva a distância no espaço real, sendo assim a transformação clr mostra-se mais plausível em relação a transformação alr para discriminar as amostras de café.

Figura 3 Mapa perceptual das categorias e amostras, transformação clr.



(a) Mapa perceptual das amostras

(b) Mapa perceptual das categorias

(c) Mapa perceptual das amostras

(d) Mapa perceptual das categorias

No caso da transformação ilr foram obtidas 82 categorias entre as seis classificações padrão. Para a transformação ilr foi possível agrupar as categorias em 11 grupos em função da magnitude da inércia obtida (Anexo E). Diferente das outras transformações, a ilr teve como categoria de maior inércia a MK:-2,5, ou seja, menor proporção de moca em relação a porcentagem de catação.

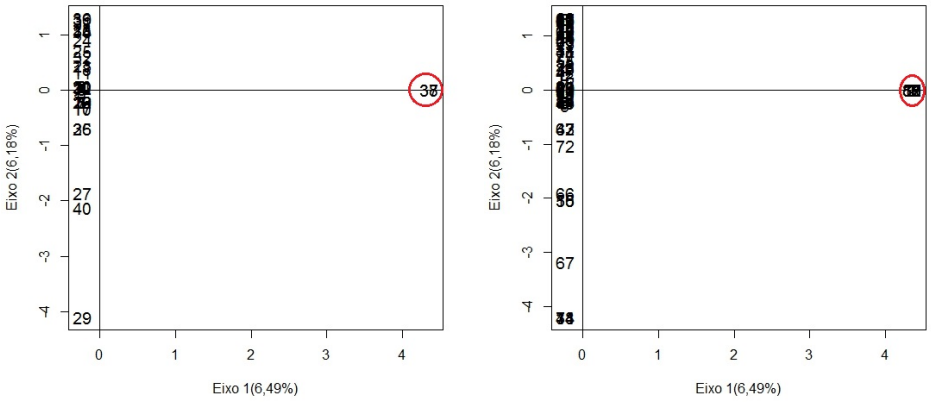
Na Figura 4 têm-se os mapas perceptuais das amostras de café e suas categorias para os eixos que possuem maior contribuição. Observa-se que as amostras 37 e 38 e suas categorias se destacam nos mapas perceptuais dos eixos 1 e 2 (Figura 4 a e 4 b).

Analisando a Tabela 9 que se encontra no anexo D, nota-se que as amostras de café 37 e 38 estão relacionadas com o primeiro eixo, sendo a contribuição das mesmas de $47{,}50\%$. As amostras 29 e 40 são responsáveis por explicar respectivamente, $44{,}31\%$ e $11{,}79\%$ da variância do segundo eixo. A amostra 29, também relacionada ao terceiro eixo, contribui com $23{,}02\%$ para a variância deste eixo. As amostras 4 e 24, estão relacionadas ao quarto e quinto eixo e contribuem com $13{,}23\%$ e $26{,}17\%$ para a variância do quarto eixo, e $27{,}52\%$ e $23{,}15\%$ da variância do quinto eixo, respectivamente.

Com relação aos mapas perceptuais dos eixos 3 e 5 (Figura 4 c e 4 d) observa-se que as amostras 4, 24, e 29 são as que possuem o maior percentual de contribuição e são as que mais se destacaram no mapa perceptual. Essa amostras se destacam pela presença de grãos chatos médios e miúdos. Então, são amostras relevantes ao comércio de café, mas com valor comercial inferior as amostras em que a proporção de grãos chato graúdo é maior. As amostras de café 37, 38, 4, 24 e 29 são as que mais se destacaram quando utilizada a transformação logarítmica isométrica.

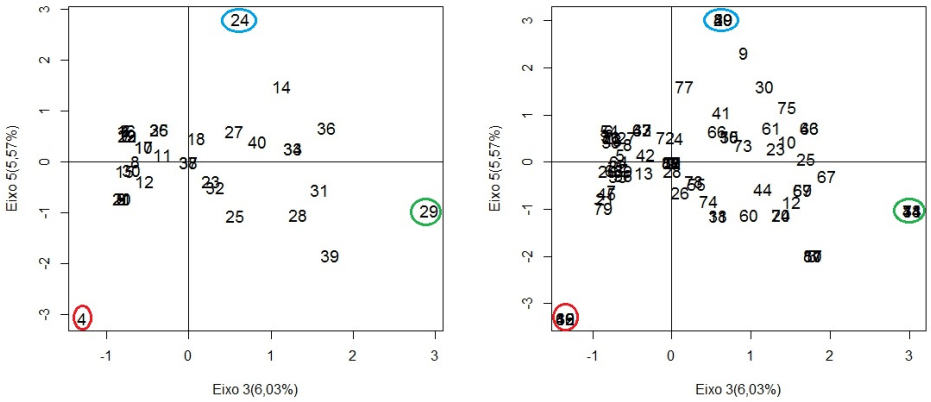
Deve-se ressaltar que a transformação clr preserva as distâncias no espaço real, mas tem a desvantagem dos dados resultantes serem colineares, diferente da transformação ilr que não apresenta as desvantagens das transformações alr e clr e que discriminou um número maior de amostras de café.

Figura 4 Mapa perceptual das categorias e amostras, transformação ilr.



(a) Mapa perceptual das amostras

(b) Mapa perceptual das categorias



(c) Mapa perceptual das amostras

(d) Mapa perceptual das categorias

4 Conclusão

Concluiu-se que a utilização das transformações logarítmicas é adequada para análise de dados composicionais em análise de correspondência múltipla, pois retira as limitações impostas pelos dados composicionais para aplicação de técnicas estatísticas multivariadas usais.

A análise de correspondência múltipla de dados de composição da granulometria de café mostrou-se uma técnica promissora para distinguir componentes que distinguem as amostras.

Entre as transformações utilizadas, a transformação logarítmica isométrica se mostrou mais plausível e foi aquela que discriminou mais amostras de café em relação às categorias dos componentes, indicando o que leva as amostras de café a ter um maior ou menor valor comercial.

Referências

ABRAHAO, A. A. et al. Classificação física e composição química do café submetido a diferentes tratamentos fungicidas. **Coffee Science**, v. 4, n. 2, p. 100–109, 2009.

AITCHISON, J. **The Statistical Analysis of Compositional Data**. London, UK, UK: Chapman & Hall, Ltd., 1986. ISBN 0-412-28060-4.

BOOGAART, K. G. van den; TOLOSANA-DELGADO, R. "compositions": A unified r package to analyze compositional data. **Computers & Geosciences**, v. 34, n. 4, p. 320 – 338, 2008.

BRASIL. Regras para análise de sementes. **Ministério da Agricultura, Pecuária e Abastecimento.**, Brasília, DF: Mapa/ACS, Brasil, p. 395, 2003.

LOURES, C. R.; ALVES, O. A. A. R.; JÚNIOR, R. A. **Classificação e Degustação do Café (Coffea arabica)**. [S.l.]: Senar, 2007.

MATIELLO, J. B. et al. **Cultura de café no Brasil: manual de recomendações.** 2. ed. Rio de Janeiro: Senar, 2010.

PAWLOWSKY, G. V. et al. **Lecture Notes on Compositional Data Analysis.** Technical University of Catalonia, Spain: [s.n.], 2010.

RUBIO, R. J. H. **CODA: Uma alternativa para estimativas multivariadas que envolvem balanços de massa granulométrico e das espécies químicas**. Porto Alegre, RS, Brasil: [s.n.], 2014.

SILVA, Y. V. da. Mestrado em Estatística Aplicada e Biometria, **Análise de Correspondência: uma abordagem geométrica**. Viçosa - MG: [s.n.], 2012.

Anexo A- Equivalência Distribucional

Exemplo extraído de Silva (2012), mostra a propriedade da equivalência distribucional. Considere uma matriz de correspondência H_{3x3} em que a linha 3 é k vezes a linha 2.

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ kh_{21} & kh_{22} & kh_{23} \end{bmatrix}$$

Matriz de perfil de linha :

$$R = \begin{bmatrix} \mathbf{r}_1^t \\ \mathbf{r}_2^t \\ \mathbf{r}_3^t \end{bmatrix} = \begin{bmatrix} \frac{h_{11}}{h_{1+}} & \frac{h_{12}}{h_{1+}} & \frac{h_{13}}{h_{1+}} \\ \frac{h_{21}}{h_{2+}} & \frac{h_{22}}{h_{2+}} & \frac{h_{23}}{h_{2+}} \\ \frac{kh_{21}}{kh_{2+}} & \frac{kh_{22}}{kh_{2+}} & \frac{kh_{23}}{kh_{2+}} \end{bmatrix}$$

Matriz de perfil de coluna:

$$C = \begin{bmatrix} \mathbf{c}_1^t \\ \mathbf{c}_2^t \\ \mathbf{c}_3^t \end{bmatrix} = \begin{bmatrix} \frac{h_{11}}{h_{+1}} & \frac{h_{21}}{h_{+1}} & \frac{kh_{21}}{h_{+1}} \\ \frac{h_{12}}{h_{+2}} & \frac{h_{22}}{h_{+2}} & \frac{kh_{22}}{h_{+2}} \\ \frac{h_{13}}{h_{+3}} & \frac{h_{23}}{h_{+3}} & \frac{kh_{23}}{h_{+3}} \end{bmatrix}$$

Que nos leva a verificar que o segundo e o terceiro perfis de linha são idênticos, sendo projetados no mesmo ponto da nuvem $N(I)$. Assim, pode-se somar a segunda e terceira linhas da matriz de correspondência H, gerando-se assim uma nova matriz H' e por consequência novos perfis de linha e coluna.

$$H' = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ (k+1)h_{21} & (k+1)h_{22} & (k+1)h_{23} \end{bmatrix}$$

novos perfis de linha

$$R' = \begin{bmatrix} r'^t_1 \\ r'^t_0 \end{bmatrix} = \begin{bmatrix} \frac{h_{11}}{h_{1+}} & \frac{h_{12}}{h_{1+}} & \frac{h_{13}}{h_{1+}} \\ \frac{(k+1)h_{21}}{(k+1)h_{2+}} & \frac{(k+1)h_{22}}{(k+1)h_{2+}} & \frac{(k+1)h_{23}}{(k+1)h_{2+}} \end{bmatrix}$$

Sendo assim, percebe-se que o vetor r'_0 é idêntico aos vetores de origem r_2 e r_3. Portanto, o importante resultado oriundo do princípio da equivalência distribucional é que depois de unificados os perfis semelhantes, sua projeção não se altera na nuvem de origem, como também não modifica as distâncias da outra nuvem, conforme exemplificado a seguir:

$$C' = \begin{bmatrix} c'^t_1 \\ c'^t_2 \\ c'^t_3 \end{bmatrix} = \begin{bmatrix} \frac{h_{11}}{h_{+1}} & \frac{(k+1)h_{21}}{h_{+1}} \\ \frac{h_{12}}{h_{+2}} & \frac{(k+1)h_{22}}{h_{+2}} \\ \frac{h_{13}}{h_{+3}} & \frac{(k+1)h_{23}}{h_{+3}} \end{bmatrix}$$

Calculando a distância entre o ponto correspondente ao vetor c^t_1 , (antes da união dos vetores semelhantes) e o seu respectivo centro de gravidade h_c, através da distância de qui-quadrado, tem-se:

$$d^2(c_1, h_c) = \frac{1}{h_{1+}} \left(\frac{h_{11}}{h_{+1}} - h_{1+} \right)^2 + \frac{1}{h_{2+}} \left(\frac{h_{21}}{h_{+1}} - h_{2+} \right)^2 + \frac{1}{kh_{2+}} \left(\frac{kh_{21}}{h_{+1}} - kh_{2+} \right)^2$$

$$d^2(c_1, h_c) = \frac{1}{h_{1+}} \left(\frac{h_{11}}{h_{+1}} - h_{1+} \right)^2 + \frac{1}{h_{2+}} \left(\frac{h_{21}}{h_{+1}} - h_{2+} \right)^2 + \frac{k^2}{kh_{2+}} \left(\frac{h_{21}}{h_{+1}} - h_{2+} \right)^2$$

$$d^2(c_1, h_c) = \frac{1}{h_{1+}} \left(\frac{h_{11}}{h_{+1}} - h_{1+} \right)^2 + \frac{k+1}{h_{2+}} \left(\frac{h_{21}}{h_{+1}} - h_{2+} \right)^2$$

$$d^2(c_1, h_c) = \frac{1}{h_{1+}} \left(\frac{h_{11}}{h_{+1}} - h_{1+} \right)^2 + \frac{(k+1)^2}{(k+1)h_{2+}} \left(\frac{h_{21}}{h_{+1}} - h_{2+} \right)^2$$

$$d^2(c_1, h_c) = \frac{1}{h_{1+}} \left(\frac{h_{11}}{h_{+1}} - h_{1+} \right)^2 + \frac{(1}{(k+1)h_{2+}} \left(\frac{(k+1)h_{21}}{h_{+1}} - (k+1)h_{2+} \right)^2$$

Calculando-se a distância entre c'_1 e o centro de gravidade h'_c:

$$d^2(c'_1, h'_c) = \frac{1}{h_{1+}} \left(\frac{h_{11}}{h_{+1}} - h_{1+} \right)^2 + \frac{(1}{(k+1)h_{2+}} \left(\frac{(k+1)h_{21}}{h_{+1}} - (k+1)h_{2+} \right)^2$$

Portanto, pode-se observar que $d^2(c_1, h_c)$ é igual $d^2(c'_1, h'_c)$ sendo assim conclui-se que ao utilizar a distância de qui-quadrado para medir a distância entre pontos de uma nuvem, conservam-se as distâncias entre os pontos, mesmo quando se unificam perfis.

Anexo B- Rotina Utiliza software R

```
# ## Gerar uma amostra da Binomial correlacionada ###
bc=function(N,P,RO)
{
x1=rbinom(1,N,P)
x2=rbinom(1,1,P)*N
if (x2==0) x2=1
u=rbinom(1,1,RO)
y=(1-u)*x1+u*x2
return (y)
}
infcor <- function (dados)
{
MP=as.matrix(dados*1/sum(dados))  # ## Prop ## #
plin=apply(MP,1,sum) #  total linha #
pcol=apply(MP,2,sum) #  total coluna  #
# ### marginais ### #
tcol=apply(dados,2,sum) #  total coluna #
tlin=apply(dados,1,sum) #  total linha  #
```

```
# ########## Perfil Coluna ############# #
Pcol=matrix(0,nrow(dados),ncol(dados))
for (j in 1:length(tcol))
{
auxcol=tcol[j]
for (i in 1:length(tlin))
{
Pcol[i,j]=dados[i,j]*(1/auxcol)}}
# ########### Perfil linha ############ #
Plin=matrix(0,nrow(dados),ncol(dados))
for (i in 1:length(tlin))
{
```

```
auxlin=tlin[i]
for (j in 1:length(tcol))
{
Plin[i,j]=dados[i,j]*(1/auxlin)}}
P=as.matrix(dados*1/sum(dados))  # ## Prop ## #
plin=apply(P,1,sum) #  total linha #
pcol=apply(P,2,sum) #  total coluna  #
```

```
# ##### Aplicando a correção qui-quadrado #### #
Q=matrix(0,nrow(dados),ncol(dados))
for (i in 1:length(plin))
{
for (j in 1:length(pcol))
{
Q[i,j]=(P[i,j]-(plin[i]*pcol[j]))/
sqrt(plin[i]*pcol[j])
 } }
# ######## Matrizes de covariancia ######### #
covc=t(Q)%*%Q            # cov.perfis coluna #
covl=Q%*%t(Q)            # cov.perfis linha #
return(list(covcol=covc,covlin=covl,
auxpcol=pcol,auxplin=plin,auxP=P))}
restab= function(dados,plin,pcol,P){
n=sum(dados)
eij=dados/n
res_BL=(dados-eij)/sqrt(eij)
res_H=(res_BL)/sqrt((1-plin)*(1-pcol))
xi_marg=apply(dados,1,sum) #  total linha #
xj_marg=apply(dados,2,sum) #  total coluna #
return(list(rBL=res_BL,rH=res_H))}
escores_conv <- function(covl,covc,pcol,plin,P,res)
```

```
# ## Coord dos perfis linha ao longo das colunas ### #
Avcol <- svd(covc)
U <- Avcol$v
A=diag(1/sqrt(pcol))%*%U  # padronizada #
F=diag(1/plin)%*%P%*%A
Anew=res%*%solve(diag(1/sqrt(pcol)))%*%
(diag(1/abs(pcol)))^0.5%*%U
# ## Coord do perfis coluna ao longo das medias
das linhas ### #
Avlin <- svd(covl)
V <- Avlin$u
B=diag(1/sqrt(plin))%*%V # padronizada #
G=diag(1/pcol)%*%t(P)%*%B
Bnew=res%*%solve(diag(1/sqrt(plin)))%*%
(diag(1/abs(plin)))^0.5%*%V
return(list(EA=A,EB=B,EA2=Anew,EB2=Bnew,EG=G,EF=F)}
```

Anexo C- Coordenadas

Tabela 5 Coordenadas utilizadas para gerar os mapas perceptuais das tabelas (5×5)

Parâmetros	Análise	Coord.1	Coord.2	Coord.3	Coord.4
$(\pi = 0{,}2; \rho = 0{,}2)$	Convencional	-0,7879	-1,1981	-0,9829	0,8799
		-1,8116	0,0459	0,2933	0,3223
		0,2139	0,9437	1,3213	-1,2554
		1,3335	-1,1602	-1,7956	-1,5969
		0,4058	1,0340	0,3820	0,8978
	Haberman	-18,9653	-6,9214	-19,2386	-4,4775
		-16,5696	-9,0200	-30,5574	-0,8205
		0,4449	3,4415	-35,1112	-12,5473
		-4,9382	-5,0532	-34,6888	-4,9709
		-9,2619	-4,8375	-32,8349	-13,2618
	Barnett e Lewis	-17,4561	-6,3208	-16,6720	-4,1512
		-16,5701	-7,5093	-24,3861	-1,3580
		-2,9232	2,3615	-27,9254	-9,4178
		-7,5548	-4,2930	-27,8916	-3,8085
		-10,3352	-3,8313	-26,1440	-10,1523
$(\pi = 0{,}5; \rho = 0{,}2)$	Convencional	-2,2371	-0,2066	-0,4177	0,4217
		0,2278	1,7717	-0,5038	0,6711
		0,5651	0,0704	-0,3628	-1,9345
		0,4933	-1,2346	-0,5353	0,7119
		0,4716	-0,3715	2,1888	0,1134
	Haberman	-9,4256	-4,4473	24,1288	-3,0829
		-34,7330	-0,2789	18,9255	37,4268
		-6,2632	-10,7454	9,5083	14,4605
		-37,0053	-5,1507	26,6567	-11,5101
		-8,1851	-4,5233	7,3248	12,2373
	Barnett e Lewis	-8,1725	-3,7932	19,8288	-3,0711
		-28,7696	-0,0886	15,2296	30,8849
		-7,1237	-9,0346	6,7979	11,8563
		-31,5563	-4,4455	20,5891	-9,0572
		-7,2133	-3,1264	5,8013	9,7821
$(\pi = 0{,}9; \rho = 0{,}2)$	Convencional	-1,8879	0,2541	-0,0419	0,0551
		0,4643	1,0162	-1,1526	0,0609
		0,9425	-0,4678	-1,0536	0,3499
		0,5155	0,9018	0,7833	-1,7637
		-0,0023	-1,7063	1,3815	1,2900
	Haberman	-47,8453	0,7783	57,8510	16,5086
		-20,9333	3,9139	41,9077	6,7942
		-64,4482	-11,8888	31,8137	-8,0544
		-13,5029	-5,6460	63,0205	12,2637
		-53,5279	-0,0146	30,0226	6,3448
	Barnett e Lewis	-39,2836	0,5240	46,4876	13,3708
		-17,9926	3,2052	32,9419	5,5636
		-53,6411	-9,7262	24,4250	-6,5553
		-11,9604	-4,6162	50,3276	9,8884
		-44,8948	-0,0785	23,0761	5,1036

Parâmetros	Análise	Coord.1	Coord.2	Coord.3	Coord.4
$(\pi = 0{,}2; \rho = 0{,}5)$	Convencional	-0,0988	-1,1934	-1,1780	-0,5001
		-0,8880	0,4944	2,0590	-2,1589
		1,2699	0,6532	0,4901	0,0832
		-1,5632	1,2514	0,1858	0,6157
		-0,1819	-1,4046	0,6685	1,7565
	Haberman	-13,3021	12,0217	2,4334	-11,2624
		20,3950	-10,9553	-24,8832	-0,0352
		10,3408	8,1234	-8,3153	-22,2471
		-9,5650	-5,2886	-28,6641	-10,8394
		-20,7007	-9,7308	-4,4755	-13,3131
	Barnett e Lewis	-9,5625	9,9340	2,0267	-7,7362
		14,6017	-8,9033	-19,2313	0,8867
		7,8516	6,5433	-6,5534	-17,3616
		-7,5657	-4,2153	-22,0427	-8,2795
		-13,9969	-8,4310	-3,6812	-11,7201
$(\pi = 0{,}5; \rho = 0{,}5)$	Convencional	-0,5559	-1,6630	-0,9463	-0,7130
		-2,2896	1,0050	1,3322	-0,6986
		0,8887	0,1365	-0,7074	1,6931
		0,6861	-0,0879	1,1781	0,9437
		0,3030	1,0923	-0,5415	-0,7081
	Haberman	18,5283	-25,3367	-16,8591	-21,3659
		-54,4731	14,0494	-7,5327	-8,0221
		5,3491	-3,1006	-13,1445	-7,1016
		9,6204	12,4416	-46,5967	-11,0401
		-35,0746	-19,9982	-17,8021	-0,1515
	Barnett e Lewis	17,1725	-20,7901	-11,6527	-16,0583
		-43,0124	11,0328	-6,2395	-5,4292
		3,3082	-2,8851	-10,8497	-4,8251
		7,9848	9,5548	-37,7255	-7,6345
		-28,3052	-16,8394	-13,7093	0,2473
$(\pi = 0{,}2; \rho = 0{,}8)$	Convencional	-1,7801	-0,9163	-1,5189	1,9195
		-0,7469	0,1828	-0,5352	-1,0810
		0,7676	0,6256	-1,0429	-0,2040
		0,4715	1,2108	1,3740	0,6452
		1,1696	-1,3534	-0,0467	-0,7483
	Haberman	-10,0704	9,1325	16,2542	-21,2592
		18,9959	-5,9637	-10,7066	12,7408
		-23,2093	-10,2464	-24,4817	-19,6012
		-6,6153	11,5198	8,3950	-27,0889
		-43,6607	-6,3848	-27,0469	-9,6000
	Barnett e Lewis	-6,3356	6,8958	13,2611	-13,9097
		17,7620	-3,7509	-4,5421	11,1530
		-14,9894	-7,6735	-14,8163	-12,8123
		-2,6523	8,3906	7,7295	-18,5765
		-29,2317	-3,6519	-17,7712	-6,3154

Parâmetros	Análise	Coord.1	Coord.2	Coord.3	Coord.4
		-1,4580	-0,8165	-1,7548	-3,4245
		-0,1051	0,2715	1,1660	-0,0942
	Convencional	-1,0235	0,0292	-1,4327	0,2461
		0,9330	1,2917	-0,0436	1,1171
		1,4153	-1,8338	0,4831	-1,0488
		-38,0865	-2,5728	50,1741	-55,3252
		64,5392	49,1138	45,2380	12,8431
$(\pi = 0{,}5; \rho = 0{,}8)$	Haberman	24,2811	-23,8728	3,8247	15,0729
		15,2483	-30,5240	22,5132	-17,6022
		-84,2481	-7,7556	-36,0024	-37,4907
		-27,8726	-0,3957	42,1441	-44,8139
		53,3672	40,5192	37,0725	12,4211
	Barnett e Lewis	21,8651	-19,2682	3,2469	13,3547
		14,5497	-25,0975	18,1844	-14,6900
		-65,4560	-7,8457	-28,2785	-30,0909
		-0,0243	0,8197	-0,0859	-0,5707
		-0,0228	0,7197	0,0169	-0,8072
	Convencional	1,6604	-1,4093	1,7084	1,3178
		0,1191	0,6564	0,0060	-0,8085
		-1,7133	-1,3115	-1,6743	1,4036
		-69,7414	4,8987	100,3598	12,8779
		-81,0177	-29,8803	63,4506	-15,1224
$(\pi = 0{,}9; \rho = 0{,}8)$	Haberman	-80,5427	5,0201	82,2458	-7,0351
		-92,5160	-24,9824	56,6923	29,1595
		-20,3550	24,4090	70,7618	18,1778
		-56,2689	3,1405	81,1728	10,1617
		-67,5858	-23,9907	49,0757	-11,9600
	Barnett e Lewis	-68,5583	2,9367	64,3011	-6,1616
		-77,2052	-21,2942	43,7886	23,3601
		-18,1589	18,6289	55,3713	14,0389

Tabela 6 Coordenadas utilizadas para gerar os mapas perceptuais das tabelas (10×10)

Parâmetros	Análise	Coord.1	Coord.2	Coord.3	Coord.4
		-0,1352	-0,1166	-1,5253	-0,1872
		-0,0961	0,0481	-0,3073	1,4251
		0,2886	0,6222	0,0960	0,3193
		-0,0313	0,1559	0,2956	-0,1656
	Convencional	-0,0229	-0,1560	-0,3963	0,7994
		1,9438	-1,7957	0,2736	0,1916
		0,5864	1,2055	1,6754	-1,5320
		-1,7182	-1,0340	0,8281	1,4844
		-0,4544	1,7469	-1,3183	-1,2804
		0,3240	0,6825	0,3753	-0,2966
		-24,5320	3,5937	-32,3461	5,8268
		-20,5552	-10,6729	-26,6888	10,7933
		-36,7826	21,1189	-32,2040	6,1615
		8,5552	16,8846	-39,5771	9,8004
$(\pi = 0{,}2; \rho = 0{,}2)$	Haberman	-11,0446	-24,6294	-20,1049	-11,3540
		2,2006	-16,5846	-19,1116	4,5656
		-39,7961	-4,1130	-10,8266	20,4151
		-6,4532	10,9316	-49,2734	-20,8707
		-29,9374	12,9525	-36,0055	19,1012
		2,6597	-8,3289	-21,9779	5,1937
		-22,9670	3,1762	-28,9443	5,3349
		-19,4910	-9,4843	-24,0970	9,7835
		-34,2141	18,7330	-28,8854	5,5575
		6,4215	14,9452	-35,8348	8,6231
	Barnett e Lewis	-11,2046	-21,7738	-18,2491	-9,9817
		0,9943	-14,2079	-17,4107	3,9668
		-36,7760	-4,0317	-10,0735	18,3286
		-6,9417	9,8458	-44,0485	-18,2994
		-27,9920	11,8481	-32,4341	17,1297
		0,9910	-7,5960	-19,6026	4,5529
		-0,2060	0,4217	-0,6730	-0,6018
		0,6994	0,7012	1,1149	0,2247
		0,3699	0,1392	-0,8432	-0,3997
		0,3227	1,5207	-0,6463	-0,0501
	Convencional	-0,7513	-0,1713	1,6994	0,1091
		0,3711	0,1044	-1,1538	2,2828
		0,4222	-1,5618	-0,7106	-1,1487
		0,1914	0,8396	1,3811	-0,6268
		-3,4853	-0,2646	-0,4036	-0,6734
		0,6740	-1,6760	1,1305	0,8499
		85,5537	21,3566	-6,1709	13,2805
		-68,1259	-19,3048	11,1310	-6,3631
		55,6038	16,3697	32,0919	-9,2920
		-63,9044	0,3281	27,7089	-25,3948
$(\pi = 0{,}5; \rho = 0{,}2)$	Haberman	30,7862	32,5740	16,7024	-9,6698
		-10,0782	-11,0799	-3,7858	-16,8482
		-65,3233	21,0704	15,5244	-38,0307
		23,8587	-48,6696	20,7287	24,1348
		-27,6194	22,0486	16,7087	-1,4591
		-99,0772	-18,9705	36,7972	5,8587
		77,3755	19,1184	-5,1346	12,1522
		-61,6907	-17,2485	9,7739	-5,5113
		50,0490	14,7948	28,9958	-8,5483
		-57,9559	0,0028	25,0185	-22,7498
	Barnett e Lewis	27,3062	29,5524	14,9754	-8,8597
		-9,3247	-10,0766	-3,5705	-15,1933
		-58,6396	19,0887	14,0362	-34,5736
		21,0872	-44,0379	18,7346	21,9250
		-25,3019	20,1716	14,9969	-0,8840
		-90,4107	-17,0409	32,9693	5,2182

Parâmetros	Análise	Coord.1	Coord.2	Coord.3	Coord.4
		-0,1362	-0,3042	-0,2095	1,4190
		-0,1865	1,6730	-0,5318	-0,1688
		-0,2547	1,2042	-0,1304	-0,6087
		-0,5515	-0,5827	-0,1268	-1,8787
	Convencional	-0,5895	-0,5316	-0,1485	1,1249
		-0,5369	1,2738	3,1115	0,010
		-0,2102	-1,5986	-0,4573	0,6569
		-0,1814	-0,6439	-0,4601	-1,3686
		3,0701	0,0263	-0,2310	0,4959
		-0,1978	-0,5736	-0,5263	0,3778
		-150,1881	-10,7258	63,7161	27,9676
		-113,7624	4,0678	101,0004	25,7982
		-30,2355	-22,4178	103,1199	6,7680
		-93,8810	-49,9240	99,9632	41,8975
$(\pi = 0{,}9; \rho = 0{,}2)$	Haberman	-111,9764	5,8404	91,0596	13,3815
		-25,1530	-35,4704	100,1831	17,3125
		-148,1924	13,9712	109,3007	-18,3867
		-55,2681	-32,8154	108,4768	7,6085
		-94,6022	-2,6522	97,3532	12,2362
		-46,6154	-12,2965	96,3532	12,2362
		-136,2952	-9,9239	57,3476	25,3481
		-103,4695	3,5098	90,7174	23,1845
		-27,9667	-20,4248	92,6621	6,0731
		-85,6167	-45,1219	89,7735	37,8937
	Barnett e Lewis	-101,9236	5,1502	81,6888	12,1475
		-23,1811	-32,1121	90,0193	15,5537
		-134,6346	12,5277	98,1450	-16,6380
		-50,5677	-29,5177	97,5375	6,8037
		-86,1445	-2,5722	87,4626	11,0697
		-42,5165	-11,2043	87,1522	34,2860
		-0,2283	-1,1785	-0,4778	0,5311
		0,0109	-0,7677	-0,8998	-0,6970
		-0,0044	0,4882	0,5358	0,4733
		1,9700	0,8633	0,5250	0,3681
	Convencional	-0,1600	-0,5545	0,0616	-0,4713
		-2,4084	2,4917	-1,9819	2,0303
		-0,7815	-0,4464	1,8449	0,9400
		1,0190	0,9361	0,1585	-0,7842
		-0,5992	-0,8905	-0,3792	-1,1510
		0,9102	-0,1650	-0,3625	-1,0419
		20,9526	60,0163	-9,8951	-7,5480
		-22,0443	12,4055	3,8527	7,4521
		27,9913	9,6185	15,1420	14,8631
		-20,3339	4,1103	-20,3545	12,1152
$(\pi = 0{,}2; \rho = 0{,}5)$	Haberman	11,3755	-18,4515	9,3753	-14,3200
		-69,9703	6,9772	-6,7812	18,3114
		-57,6642	-40,5932	-23,8223	14,1411
		4,9598	-15,0046	-13,6627	14,1525
		-11,9896	24,1702	-1,8685	14,5837
		16,6982	1,8512	4,2373	23,4491
		18,6592	52,3679	-8,8781	-6,9509
		-21,4473	11,1132	3,6074	6,7354
		23,872	9,1876	13,6501	12,4187
		-17,8076	3,2070	-17,7797	10,6547
	Barnett e Lewis	10,7259	-16,5297	8,3190	-12,9121
		-60,76595	5,6991	-6,4103	17,2945
		-50,8417	-34,9689	-21,4852	12,6161
		5,4421	-11,9246	-11,1302	13,1424
		-10,2869	21,4549	-1,6435	13,7735
		14,4008	3,4078	3,9500	20,6048

Parâmetros	Análise	Coord.1	Coord.2	Coord.3	Coord.4
		-0,2455	0,2464	-1,2800	-2,0077
		0,0239	-0,9151	0,6207	0,6827
		1,8245	-0,4458	0,2203	-0,4747
		-1,2128	0,3021	-0,2407	-0,5251
	Convencional	1,1417	2,0224	0,2818	1,4014
		0,9314	-1,4222	1,9948	-0,9219
		-0,5101	-0,9217	0,1229	0,3712
		-0,3716	1,5771	1,2995	-0,7817
		-1,8893	-0,4223	-2,0101	0,3592
		-0,1147	-0,0848	-0,06697	1,1476
		47,1558	-92,0237	18,0594	8,8834
		-9,3554	47,5226	-29,4364	-17,6130
		-16,6074	-28,9143	22,3753	-56,3938
		10,8065	76,4870	28,4474	-18,3830
$(\pi = 0{,}5; \rho = 0{,}5)$	Haberman	-97,6941	-32,8145	20,2425	7,8765
		81,5051	67,39	41,04	11,2703
		11,4531	-37,0289	31,0955	-22,6442
		-34,7378	3,5677	-37,6463	12,2620
		32,3377	-3,3239	77,6354	-34,08
		-30,3356	-44,3099	21,1102	2,8601
		43,5519	-83,0212	16,7253	7,5209
		-9,0733	43,1243	-26,7353	-15,2274
		-15,2086	-26,4654	20,2916	-51,0030
		9,6669	69,2835	25,6553	-17,1469
	Barnett e Lewis	-88,5339	-30,3346	18,4401	6,3109
		74,8051	60,2657	37,2392	9,8643
		11,3767	-33,1121	27,6506	-20,2465
		-31,2949	3,6309	-33,6751	11,1914
		30,2658	-1,6860	69,9186	-30,2215
		-27,3118	-38,9587	18,7222	2,9421
		-0,2274	-0,0019	-2,3053	-1,2337
		0,2856	0,0361	-0,2994	2,6662
		0,3140	0,0196	0,1448	0,2449
		0,3093	0,0394	-0,1465	-0,0493
	Convencional	0,3136	2,4488	1,4824	0,1322
		-1,8110	-1,4939	-0,1961	-0,0619
		2,0662	-1,5474	-0,0406	-0,2300
		-1,6209	0,2242	1,6705	-1,4053
		0,0366	0,2281	-0,3411	-0,1382
		0,2526	0,0157	0,0328	0,1545
		-178,7675	-13,2266	115,7443	6,1511
		-105,8232	-43,7580	99,3414	-7,8408
		-79,3090	-20,9866	109,9477	31,2597
		-32,7960	-88,1276	111,7987	7,6073
$(\pi = 0{,}9; \rho = 0{,}5)$	Haberman	-126,8830	-15,5789	96,0651	-17,6659
		-139,1557	-47,7286	151,3573	-5,4422
		-46,6310	17,9574	87,4402	50,1887
		-85,7226	-37,4424	108,4340	19,7219
		-39,4453	-23,6023	96,8036	49,9787
		-120,8128	15,6817	100,7046	45,2972
		-162,4228	-12,2933	104,0901	5,2090
		-96,4113	-39,5908	88,9106	-7,0914
		-72,9818	-18,8484	98,6822	27,9658
		-30,8042	-79,8031	100,2099	6,7041
	Barnett e Lewis	-115,6558	-14,4859	86,0800	-16,2024
		-127,0353	-43,1865	136,3350	-4,9290
		-43,5165	15,7706	78,3223	45,2223
		-78,5178	-34,0520	97,2911	17,7508
		-36,5519	-21,31222	86,6628	44,9554
		-110,3978	13,9419	90,3677	40,8377

Parâmetros	Análise	Coord.1	Coord.2	Coord.3	Coord.4
		-1,2255	0,8628	-0,8451	0,7901
		0,8579	-0,7904	0,4394	0,7163
		-0,4073	-1,7135	1,9145	0,1013
		0,7488	0,6550	0,5647	0,7040
	Convencional	-1,4340	0,3071	1,4627	0,4968
		-0,0381	-2,5199	-0,8738	1,0472
		0,4526	0,1172	-0,2544	-1,0619
		-0,3203	-2,2169	0,0285	-1,9973
		-1,7496	0,4716	-0,9978	0,3047
		1,0678	0,8868	-1,0485	0,0130
		15,1580	23,9893	16,7614	-6,4717
		57,9499	5,7955	-21,0656	-5,0691
		-43,3955	-7,3165	6,3009	7,4878
		-56,1363	-30,7995	-27,1527	23,3923
$(\pi = 0{,}2; \rho = 0{,}8)$	Haberman	-73,3263	18,7411	12,2168	-20,9783
		-19,5327	13,4631	28,6426	-37,0215
		28,7717	-30,8266	-15,7460	-15,3023
		-35,7045	16,6234	-21,1657	10,5219
		51,5809	-20,4697	-25,7526	-24,6012
		-32,6059	13,2195	34,4894	2,6508
		16,4291	21,8925	15,1149	-5,6739
		52,5831	4,4682	-18,6709	-3,9138
		-35,5596	-7,4101	6,1208	6,8215
		-46,6381	-28,2747	-24,5566	20,3796
	Barnett e Lewis	-64,7563	18,2368	11,3811	-17,9126
		-16,3123	11,8299	25,7331	-31,7792
		28,8219	-27,5292	-13,0617	-13,0964
		-28,7770	15,1334	-19,6992	10,4813
		45,7950	-17,7710	-23,7112	-22,1758
		-28,4075	11,4327	31,6575	3,0959
		-0,0358	0,2091	-0,1360	-0,2084
		1,4375	-1,4030	1,5571	1,2440
		1,5509	-1,2068	-0,1434	-0,3164
		0,2235	1,0754	0,0709	-1,0539
	Convencional	-0,1368	0,2282	-0,1360	-0,2084
		-0,3077	1,0142	-0,1062	-0,3087
		-0,1279	0,3338	-0,2965	-1,1831
		-0,1133	1,0885	1,4872	1,3134
		-0,2108	0,2646	-2,4972	1,9629
		-2,5036	-1,9009	-0,0252	-0,9163
		-197,3477	23,7762	110,1734	54,0385
		-150,9433	3,7032	122,0799	33,3665
		-48,9374	-46,4411	93,3257	6,6799
		-56,8945	-102,4305	71,4199	85,4476
$(\pi = 0{,}9; \rho = 0{,}8)$	Haberman	20,9921	-3,1747	98,9854	0,8249
		-48,9191	-40,5688	100,4645	20,2897
		-61,3714	-4,4649	93,9654	25,3294
		-99,9293	-79,7032	39,2078	41,8693
		-71,8732	-34,6175	74,9997	72,1436
		-101,7835	9,4058	113,2123	44,8848
		-179,3457	20,9503	99,3123	48,5497
		-137,4634	2,8250	109,6887	30,1019
		-45,5189	-41,6315	83,7691	5,9447
		-52,3039	-93,0152	63,5579	77,3856
	Barnett e Lewis	18,0063	-3,0639	88,6094	0,3841
		-45,0803	-37,1693	90,0504	18,3342
		-56,4038	-3,9207	84,4140	22,6050
		-91,9629	-72,6284	34,4769	37,4831
		-66,9441	-31,2548	67,2057	65,1635
		-93,5687	8,2182	101,7377	40,4756

Anexo D- Contribuição das amostras em relação aos eixos.

Tabela 7 Contribuição das amostras em cada um dos eixos, transformação alr

	Eixos					
Amostras	1	2	3	4	5	6
1	0,0006	0,0004	0,0071	0,0207	0,0006	0,0666
2	0,0006	0,0584	0,0457	0,0280	0,0293	0,0045
3	0,0006	0,0000	0,0009	0,0140	0,0000	0,0177
4	0,0006	0,0000	0,0000	0,0219	0,0003	0,0009
5	0,0006	0,0004	0,0071	0,0207	0,0006	0,0666
6	0,0006	0,0584	0,0457	0,0280	0,0293	0,0045
7	0,0006	0,0000	0,0009	0,0140	0,0000	0,0177
8	0,0006	0,0067	0,0001	0,0122	0,0005	0,0077
9	0,0006	0,0584	0,0457	0,0280	0,0293	0,0045
10	0,0006	0,0045	0,0280	0,0089	0,0001	**0,1450**
11	0,0006	0,0015	0,0003	0,0246	0,0004	0,0005
12	0,0006	0,0054	0,0007	0,0102	0,0001	0,0116
13	0,0006	0,0000	0,0009	0,0140	0,0000	0,0177
14	0,0006	0,0059	0,0016	0,0002	0,0009	0,0127
15	0,0006	0,0537	0,0060	0,0006	0,0001	0,0039
16	0,0006	0,0004	0,0071	0,0207	0,0006	0,0666
17	0,0006	0,0045	0,0280	0,0089	0,0001	**0,1450**
18	0,0006	0,0201	0,0175	0,0009	0,0298	0,0000
19	0,0006	0,0000	0,0009	0,0140	0,0000	0,0177
20	0,0006	0,0584	0,0457	0,0280	0,0293	0,0045
21	0,0006	0,0584	0,0457	0,0280	0,0293	0,0045
22	0,0006	0,0000	0,0009	0,0140	0,0000	0,0177
23	0,0006	0,0019	0,0007	0,0154	0,0001	0,0135
24	0,0006	0,0007	0,0044	0,0195	0,0005	0,0482
25	0,0006	0,0000	0,0008	0,0080	0,0001	0,0009
26	0,0006	0,0012	0,0005	0,0239	0,0001	0,0265
27	0,0006	0,0059	**0,1938**	0,0889	0,0016	0,0083
28	0,0006	0,0063	**0,1168**	0,0558	0,0004	0,0528
29	0,0006	0,0003	0,0582	0,0003	0,0000	0,1165
30	0,0006	0,0426	0,0403	0,0041	**0,8119**	0,0057
31	0,0006	0,0423	0,0033	0,0009	0,0005	0,0007
32	0,0006	0,0002	0,0004	0,0177	0,0001	0,0087
33	0,0006	0,0004	0,0014	0,0218	0,0005	0,0152
34	0,0006	0,0004	0,0014	0,0218	0,0005	0,0152
35	0,0006	0,0012	0,0005	0,0239	0,0001	0,0265
36	0,0006	0,0376	0,0038	0,0001	0,0001	0,0019
37	0,0006	**0,2679**	0,0372	**0,1752**	0,0010	0,0054
38	**0,9750**	0,0000	0,0000	0,0000	0,0000	0,0000
39	0,0006	**0,1907**	0,0256	0,0926	0,0003	0,0011
40	0,0006	0,0049	**0,1741**	0,0702	0,0012	0,0150

Tabela 8 Contribuição das amostras em cada um dos eixos, transformação clr

	Eixos					
Amostras	1	2	3	4	5	6
1	0,0013	0,0000	0,0071	0,0021	0,0003	0,0030
2	0,0013	0,0000	0,0196	0,0125	0,0034	0,0356
3	0,0013	0,0000	**0,1081**	0,0169	0,0002	0,0182
4	0,0013	0,0000	0,0046	0,0015	0,0049	0,0121
5	0,0013	0,0000	0,0071	0,0021	0,0003	0,0030
6	0,0013	0,0000	0,0196	0,0125	0,0034	0,0356
7	0,0013	0,0000	**0,1081**	0,0169	0,0002	0,0182
8	0,0013	0,0000	0,0322	**0,3202**	**0,4896**	0,0089
9	0,0013	0,0000	0,0196	0,0125	0,0034	0,0356
10	0,0013	0,0000	0,0040	0,0015	0,0029	0,0016
11	0,0013	0,0000	0,0060	0,0076	0,0026	0,0028
12	0,0013	0,0000	0,0007	0,0007	0,0007	0,0106
13	0,0013	0,0000	**0,1081**	0,0169	0,0002	0,0182
14	0,0013	0,0000	0,0162	0,0068	0,0105	0,0001
15	0,0013	0,0000	0,0017	0,0040	0,0002	0,0293
16	0,0013	0,0000	0,0071	0,0021	0,0003	0,0030
17	0,0013	0,0000	0,0040	0,0015	0,0029	0,0016
18	0,0013	0,0000	0,0000	0,0049	0,0019	0,0117
19	0,0013	0,0000	**0,1081**	0,0169	0,0002	0,0182
20	0,0013	0,0000	0,0196	0,0125	0,0034	0,0356
21	0,0013	0,0000	0,0196	0,0125	0,0034	0,0356
22	0,0013	0,0000	**0,1081**	0,0169	0,0002	0,0182
23	0,0013	0,0000	0,0009	0,0000	0,0000	0,0113
24	0,0013	0,0000	0,0018	0,0222	0,0062	0,0081
25	0,0013	0,0000	0,0051	0,0183	0,0034	0,0034
26	0,0013	0,0000	0,0138	0,0117	0,0380	0,0048
27	0,0013	0,0000	0,0138	0,0602	0,0391	0,0464
28	0,0013	0,0000	0,0018	0,0100	0,0017	0,0034
29	0,0013	0,0000	0,0083	0,0178	0,0317	0,0030
30	0,0013	0,0000	0,0055	0,0749	0,0242	0,0683
31	0,0013	0,0000	0,0008	0,0005	0,0069	0,0022
32	0,0013	0,0000	0,0107	0,0016	0,0007	0,0002
33	0,0013	0,0000	0,0010	0,0032	0,0012	0,0326
34	0,0013	0,0000	0,0010	0,0032	0,0012	0,0326
35	0,0013	0,0000	0,0138	0,0117	0,0380	0,0048
36	0,0013	0,0000	0,0003	0,0525	0,0093	0,0000
37	**0,4493**	**0,5257**	0,0000	0,0000	0,0000	0,0000
38	**0,5007**	**0,4743**	0,0000	0,0000	0,0000	0,0000
39	0,0013	0,0000	**0,1786**	**0,1572**	**0,2248**	**0,3853**
40	0,0013	0,0000	0,0130	0,0532	0,0381	0,0368

Tabela 9 Contribuição das amostras em cada um dos eixos, transformação ilr

	Eixos					
Amostras	1	2	3	4	5	6
1	0,0013	0,0011	0,0154	0,0141	0,0115	0,0078
2	0,0013	0,0001	0,0176	0,0300	0,0145	0,0002
3	0,0013	0,0012	0,0150	0,0195	0,0078	0,0075
4	0,0013	0,0000	0,0450	**0,1323**	**0,2752**	0,0864
5	0,0013	0,0011	0,0154	0,0141	0,0115	0,0078
6	0,0013	0,0001	0,0176	0,0300	0,0145	0,0002
7	0,0013	0,0012	0,0150	0,0195	0,0078	0,0075
8	0,0013	0,0000	0,0110	0,0087	0,0000	0,0254
9	0,0013	0,0001	0,0176	0,0300	0,0145	0,0002
10	0,0013	0,0030	0,0082	0,0008	0,0028	0,0016
11	0,0013	0,0027	0,0028	0,0476	0,0007	0,0041
12	0,0013	0,0000	0,0079	0,0013	0,0040	0,0039
13	0,0013	0,0012	0,0150	0,0195	0,0078	0,0075
14	0,0013	0,0287	0,0341	0,0411	0,0646	0,0028
15	0,0013	0,0005	0,0163	0,0000	0,0009	0,0004
16	0,0013	0,0011	0,0154	0,0141	0,0115	0,0078
17	0,0013	0,0030	0,0082	0,0008	0,0028	0,0016
18	0,0013	0,0039	0,0002	0,0044	0,0066	0,0003
19	0,0013	0,0012	0,0150	0,0195	0,0078	0,0075
20	0,0013	0,0001	0,0176	0,0300	0,0145	0,0002
21	0,0013	0,0001	0,0176	0,0300	0,0145	0,0002
22	0,0013	0,0012	0,0150	0,0195	0,0078	0,0075
23	0,0013	0,0052	0,0019	0,0150	0,0039	0,0043
24	0,0013	0,0214	0,0105	**0,2617**	**0,2315**	0,0098
25	0,0013	0,0134	0,0085	0,0231	0,0322	0,0311
26	0,0013	0,0128	0,0036	0,0199	0,0122	0,0413
27	0,0013	**0,0903**	0,0082	0,0040	0,0108	0,2717
28	0,0013	0,0324	0,0474	0,0249	0,0309	0,0037
29	0,0013	**0,4431**	**0,2302**	0,0324	0,0268	0,1913
30	0,0013	0,0000	0,0130	0,0010	0,0006	0,0181
31	0,0013	0,0406	0,0678	0,0110	0,0087	0,0012
32	0,0013	0,0093	0,0029	0,0040	0,0070	0,0120
33	0,0013	0,0303	0,0428	0,0001	0,0022	0,0044
34	0,0013	0,0303	0,0428	0,0001	0,0022	0,0044
35	0,0013	0,0128	0,0036	0,0199	0,0122	0,0413
36	0,0013	0,0448	0,0756	0,0026	0,0136	0,0149
37	**0,4750**	0,0000	0,0000	0,0000	0,0000	0,0000
38	**0,4750**	0,0000	0,0000	0,0000	0,0000	0,0000
39	0,0013	0,0435	0,0796	0,0526	0,0970	0,0127
40	0,0013	**0,1179**	0,0190	0,0010	0,0047	0,1494

Anexo E- Inércia e categorias da transformação alr, clr e ilr

Tabela 10 Inércias transformação alr

Ord. relevante	Inércia	Categoria
1	0,195	CG:1,2/CME:-0,4/CMI:-5,3/MK:-5,3/FDO:-5,3
2	0,176	CME:2,2/CMI:-5/CMI:0,7/MK:0,5/FDO:-5/FDO:-0,7
3	0,166	CG:-2,8/CG:0,3/CME:0,8/CMI:-1,5/CMI:0,6/CMI:1,9 MK:-1,1/MK:0/FDO:-2,5
4	0,156	CG:0,8/CG:0,9/CG:1/CG:1,6/CME:-0,5/CME:0,7 MK:-5/MK:-0,3
5	0,148	CG:-1,2/CG:0,7/CME:0,3/CME:0,4/CME:0,5/CME:1,6/MK:0,2/CME:1,7/CMI:0,1/ CMI:0,8/CMI:1,3/CMI:1,5/MK:-0,2/MK:-0,1/FDO:-0,9/FDO:-0,8/FDO-1
6	0,134	CG:-3,1/CG:0,4/CG:1,4/CME:0,2/CM:1,4/CMI:-1,6/CG:0,1 CMI:-0,2/MK:-2/MK:-1,2/MK:-0,8/FDO:-2,3/FDO:-1,3/FDO:0,6
7	0,121	CG:-0,2/CG:-0,1/CME:0,1/CME:1/CMI:-0,1/MK:-0,9 MK:0,6/MK:-0,4/FDO:-1,9/FDO:-1,8/FDO:-1,5
8	0,113	CG:0,2 /CMI:-0,5/ CMI:-0,3 / CMI:0,4 /MK:-1,4 /MK:-1/ MK:-0,7/FDO:-2,8/FDO:-2,1/FDO:-1,7/ FDO:0,5
9	0,084	CME:1,1 / Mk:-1,3 / FDO:-2,4
10	0,069	MK:-1,5
11	0,061	CG:0

Tabela 11 Inércias transformação clr

Ord. relevante	Inércia	Categoria
1	0,163	CG:3,7/CG:3,9/CME:1,8/CME:2,1/CMI:-2,8/CMI:-2,7/MK:-2,8/MK:-2,7/FDO:-2,8 FDO:-2,7/CAT:2,3/CAT:2,5
2	0,135	CG:0,1/CG:3,1/CME:1,4/CME:2,2/CMI:-3,3/MK:-1/FDO:-3,3/FDO:-0,7/CAT:1,7
3	0,121	CME:0,5/CMI:-0,1/MK:-0,5/FDO:-1,5/
4	0,117	CME:0,1/CME:1,6/CMI:0,1/CMI:0,7/CMI:1,8/MK:-0,1/CAT:-0,6/CAT:0,3
5	0,102	CG:-1,4/CG:0,4/CG:1,1/CME:0,6/CME:0,8/CME:1,1/CMI:-0,8/CMI:1,2/CMI:1,4 MK:-0,2/FDO:-1,3/FDO:-1/FDO:0,5/CAT:-0,2/CAT:-0,1/CAT:0,4/CAT:0,7
6	0,091	CG:-2,4/CG:0/CG:0,3/CG:0,5/CG:0,6/CME:1,2/CME:2/CMI:9/CMI:0,5/CMI:0,6 MK:-0,6/MK:-0,3/FDO:-2,1/FDO:-1,8/FDO:-1,6/FDO:-1,2/FDO:-0,6/CAT:0,1/CAT:0,2
7	0,086	CG:-2,8/CG:0,2/CG:0,7/CG:0,8/CG:1,5/CME:0,7/CME:1,3/CMI:-0,9CMI:0,2/CMI:0,3/CMI:0,4 MK:-0,8/MK:-1,3/MK:-0,4/FDO:-1,9/FDO:-1,7/FDO:-1,4/FDO:-0,8/FDO:0,4/CAT:-0,3/CAT:0,6
8	0,076	CG:-2,9/CG:0,9/CME:0,9/CAT:-0,4/CAT:0,5
9	0,064	CG:1/MK:-0,7

Tabela 12 Inércias transformação ilr

Ord. relevante	Inércia	Categoria
1	0,190	MK:-2,5
2	0,175	CG:-1,5/CG:-1,2/CG:1/CME:-4,6/CME:-4,5/CME:1,6 CMI:-3,3/CMI:-3,2/CMI:-0,4/FDO:-0,2/FDO:2,5/FDO:2,8
3	0,160	CG:-0,6/CG:3,2/CME:-4,8/CME:0,5/CMI:-0,2/CMI:-3,3/MK:-0,5/FDO:1,9
4	0,155	CG:0,3/CG:2,6/CME:0/CME:2,3/CMI:-1,2/CMI:0/MK:-1,5
5	0,141	CG:0,6/CME:-0,4/MK:-1,4/MK:0,5/FDO:0,5
6	0,132	CG:0/CG:0,4/CG:0,5/CME:-1,5/CME:0,9/CMI:-1,8 CMI:-0,8/CMI:-0,6/CMI:-0,5/MK:-1,3/MK:-1,1/MK:-0,8/FDO:0,3
7	0,120	CME:-0,7/CME:-0,6/CME:-0,5/CMI:-0,7/MK:-1,7/FDO:-0,5 FDO:-0,4/FDO:-0,1FDO:0,1/FDO:0,2/FDO:0,4/FDO:0,8/
8	0,110	CG:-0,1/CG:0,1/CG:0,9/CME:-0,3/CME:-0,2/CME:0,1/CME:0,2/CMI:-1,1/CMI:-0,9 MK:-2,1/MK:-1,8/MK:-1,6
9	0,101	CME:-0,1/MK:-1,9/MK:-0,7/MK:0,4
10	0,093	CMI:-1,3/CMI:-1/MK:-2,2/FDO:-0,7
11	0,081	CG:-0,3/CG:-0,2/CME:-1,9/FDO:0,6