



PAULO CÉSAR OSSANI

**QUALIDADE DE CAFÉS ESPECIAIS E
NÃO ESPECIAIS POR MEIO DA
ANÁLISE DE MÚLTIPLOS FATORES
PARA TABELAS DE
CONTINGÊNCIAS**

**LAVRAS – MG
2015**

PAULO CESAR OSSANI

**QUALIDADE DE CAFÉS ESPECIAIS E NÃO ESPECIAIS
POR MEIO DA ANÁLISE DE MÚLTIPLOS FATORES PARA
TABELAS DE CONTINGÊNCIAS**

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 Análise Multivariada, para a obtenção do título de Mestre.

Orientador

Dr. Marcelo Ângelo Cirillo

Coorientador

Dr. Flávio Meira Borém

LAVRAS – MG

2015

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

Ossani, Paulo Cesar.

Qualidade de cafés especiais e não especiais por meio da
análise de múltiplos fatores para tabelas de contingências /

Paulo Cesar Ossani. – Lavras: UFLA, 2015.

107 p. : il.

Dissertação (mestrado acadêmico) - Universidade Federal de
Lavras, 2015.

Orientador(a): Marcelo Ângelo Cirillo.

Bibliografia.

1. Métodos MFA e MFACT. 2. Cafés. 3. Sensorial. 4. Per-
cepção. 5. Consumidores. I. Universidade Federal de Lavras. II.
Título.

PAULO CESAR OSSANI

**QUALIDADE DE CAFÉS ESPECIAIS E NÃO ESPECIAIS
POR MEIO DA ANÁLISE DE MÚLTIPLOS FATORES PARA
TABELAS DE CONTINGÊNCIAS**

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 Análise Multivariada, para a obtenção do título de Mestre.

APROVADO em 13 de fevereiro de 2015

Dr. Joel Augusto Muniz UFLA

Dr. Taciana Villela Savian USP

Orientador
Dr. Marcelo Ângelo Cirillo

**LAVRAS – MG
2015**

*À minha amada esposa, Brígida, pela paciência e carinho
dispensado a mim nas horas difíceis que passei estudando
para a apresentação deste trabalho.*

*Aos meus pais, Paulo e Sebastiana, que sempre me
apoiaram nas realizações dos meus sonhos.*

*E aos meus filhos, Gabriel e Rafael, que sempre afagam o
meu coração com muito amor e ternura.*

DEDICO

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, por estar sempre ao meu lado, me intuindo nas escolhas certas e guiando-me no caminho correto.

Aos meus pais, Paulo e Sebastiana, que sempre me apoiaram e incentivaram na jornada em busca de novas realizações.

À Universidade Federal de Lavras (UFLA), ao Departamento de Ciências Exatas e ao Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, pela oportunidade de cursar o mestrado.

Aos orientadores, professores Dr. Marcelo Ângelo Cirillo e Dr. Flávio Meira Borém, pela paciência ao orientar-me, e, acima de tudo, pelo conhecimento que ambos compartilharam comigo.

A todos meus amigos de Lavras, em especial aos do Departamento de Ciências Exatas que me auxiliaram durante o curso.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela bolsa concedida no mestrado, pois, sem a mesma, seria inviável a minha manutenção no decorrer do curso.

Enfim, a todos os meus amigos que fazem parte de minha carreira acadêmica.

"Aquele que não sabe, e não sabe, que não sabe, é um tolo. Aquele que não sabe, e sabe que não sabe, é um simples. Aquele que sabe, e não sabe que sabe, está dormindo. Aquele que sabe, e sabe que sabe, é um sábio".

Provérbio Árabe

RESUMO

Artigo 1 - A qualidade sensorial de cafés especiais, cujas características se relacionam com o meio geográfico, é apreciada pelo setor produtivo e o mercado, no qual, uma relação comercial é pautada na livre escolha, por parte do consumidor e na agregação de valores e diferenciação de preços, em função da qualidade do produto. Neste cenário, focar as tendências dos consumidores é primordial para que um café seja diferenciado em relação aos demais. Para isso, novas metodologias de análise devem ser exploradas, para que os resultados sejam dignos de contemplar inúmeros fatores inerentes às particularidades de cada consumidor e/ou produto. Com esse propósito, neste trabalho buscou-se propor o uso da técnica de múltiplos fatores, aplicada a tabelas de contingência (MFACT), em dados categorizados obtidos em um experimento sensorial realizado com diferentes grupos de consumidores, com a finalidade de identificar similaridades entre quatro cafés especiais. Concluiu-se que o uso dessa técnica é viável, por permitir discriminar os cafés especiais produzidos em diferentes ambientes (altitudes) e processamentos, considerando a heterogeneidade entre os consumidores envolvidos na análise sensorial.

Artigo 2 - A importância do café assume posição de destaque no cenário socioeconômico brasileiro, pois, além de ser relevante na geração empregos, e em toda economia, de modo direto ou indireto, também são importantes as relações sociais, sendo consumido por hábito e prazer, indicando que as motivações para o seu consumo estão associadas aos hábitos e aos costumes brasileiros. Assim, o estudo da percepção dos consumidores é uma ferramenta importante para a identificação das motivações nos processos de compra do café, nos diferentes segmentos desse mercado. Entretanto, ao considerar grupos de variáveis qualitativas, em uma abordagem, multivariada, as técnicas estatísticas usuais são limitadas ou inviáveis para descrever adequadamente o perfil sensorial dos consumidores. Tendo esse fato como premissa, este trabalho foi realizado com o objetivo de apresentar a técnica de múltiplos fatores (MFA), no estudo do perfil de consumidores de café, em uma pesquisa de opinião pública aplicada a um grupo de consumi-

dores localizados na cidade de Varginha, MG. Concluiu-se que o uso dessa técnica é viável, por permitir discriminar as percepções dos consumidores de cafés.

Palavras-chave: Métodos MFA e MFACT, Cafés, Sensorial, Percepção, Consumidores.

ABSTRACT

Paper 1 – The sensorial quality of specialty coffees, the characteristics of which are related to the geographic area is enjoyed by the productive sector and the market, in which, a commercial relationship is ruled on free choice, on the consumer's part and price differentiation owing to the product quality. In this scenario, to focus consumers' trends is primordial for a coffee to be distinct in relation to the others, for that purpose, new analysis methodologies should be exploited for the results to be worthy to behold a number of factors inherent the particularities of each consumer and/or product. With the purpose, the objective of that work is to propose the use of the multi-factor technique applied to contingency tables (MFACT), in data categorized obtained in a sensorial experiment conducted with different groups of consumers with the purpose of identifying similarities among four specialty coffees. It follows that use of that technique is viable for allowing discriminating the specialty coffees produced in different places (altitudes) and processings, considering heterogeneity among the consumers involved in the sensorial analysis.

Paper 2 – The importance of coffee takes over leading position in the Brazilian socioeconomic scenario, for in addition to be relevant in the generation of employment and in the economics of either direct or indirect way, the social relationships are also important, its being consumed by habit and pleasure, indicating that the motivations to its consumption is associated to the Brazilian habits and costumes. So, the study of the consumers' perception is an important tool to the identification of the motivations in the coffee purchase processes in the different segments of this market. However, in considering groups of qualitative variables, in multivariate approach, the usual statistic techniques are limited or unviable to describe adequately consumers' sensorial profile. Having this fact as a premise, that work intends to present a multi-factor technique (MFA), in the study of the coffee consumers' profile in a study of public pool applied to a group of consumers localized in the city of Varginha - MG. It follows that the use of this technique is viable, for allowing discriminating the coffee

consumers' perceptions.

Keywords: MFA and MFACT Methods, Coffees, Sensorial, Perception, Consumers.

LISTA DE FIGURAS

| | | |
|----|---|-----|
| 1 | Ilustração da rotação dos eixos das coordenadas, definidos por duas variáveis correlacionadas X_1 e X_2 , em que Y_1 e Y_2 correspondem aos componentes principais. | 19 |
| 2 | Tabela Global com os conjuntos de variáveis. | 24 |
| 3 | Tabelas de contingências justapostas | 32 |
| 4 | Questionário da pesquisa de análise sensorial de café. | 54 |
| 5 | Layout de tabelas de contingência justapostas. | 57 |
| 6 | Gráfico correspondente à análise global das faixas dos indivíduos. | 62 |
| 7 | Gráfico correspondente à análise global das faixas dos indivíduos com as análises dos grupos de indivíduos. | 63 |
| 8 | Gráfico correspondente à análise global das faixas dos indivíduos. | 65 |
| 9 | Gráfico correspondente à análise global das faixas dos indivíduos com as análises dos tipos de cafés. | 66 |
| 10 | Layout dos grupos de variáveis justapostas, representando um único grupo. | 76 |
| 11 | Gráfico da correspondente à análise global dos indivíduos. | 81 |
| 12 | Gráfico das análises dos grupos de indivíduos. | 81 |
| 13 | Gráfico da análise global dos homens e mulheres. | 92 |
| 14 | Gráfico dos Cafés com a Análise Global dos Homens e Mulheres. | 94 |
| 15 | Circulo das correlações entre os componentes principais (eixos) e as variáveis originais. | 96 |
| 16 | Gráfico das inércias dos cafés. | 97 |
| 17 | Gráfico da análise global dos homens e mulheres. | 103 |
| 18 | Gráfico dos grupos de variáveis com à análise global dos homens e mulheres. | 105 |
| 19 | Gráfico das inércias dos grupos de variáveis. | 107 |

LISTA DE TABELAS

| | | |
|----|---|----|
| 1 | Coefficiente relativo à tabela t nas análises CA e ICA, aplicado na tabela X_G | 35 |
| 2 | Escala de qualidade dos cafés. | 46 |
| 3 | Escala de classificação de cafés especiais. | 47 |
| 4 | Descrição dos cafés especiais avaliados na análise sensorial. | 55 |
| 5 | Layout da tabela justaposta para análise de cada grupo de indivíduos em relação às avaliações sensoriais para os cafés especiais. | 56 |
| 6 | Layout da tabela justaposta para análise dos cafés especiais, em relação às avaliações sensoriais para os cafés especiais. | 56 |
| 7 | Coefficiente relativo à tabela t nas análises CA e ICA, aplicado na tabela P | 58 |
| 8 | Proporção da variação amostral explicada pelos três primeiros componentes. | 61 |
| 9 | Valores referentes às inércias computadas nos três primeiros eixos para cada grupo de consumidor. | 64 |
| 10 | Proporção da variação amostral explicada pelos três primeiros componentes. | 64 |
| 11 | Valores referentes às inércias computadas nos três primeiros eixos, para cada grupo de tipos cafés. | 65 |
| 12 | Layout do questionário aplicado ao estudo sensorial dos consumidores de café referentes à cidade de Varginha, MG | 75 |
| 13 | Layout da tabela justaposta para análise de cada grupo de variáveis. | 76 |
| 14 | Proporção da variação amostral, explicada pelos dois primeiros componentes. | 80 |
| 15 | Valores referentes às inércias computadas nos dois primeiros componentes, para cada grupo de variáveis. | 82 |

| | | |
|----|--|-----|
| 16 | Dados sobre o número de pessoas que consomem cafés. | 88 |
| 17 | Tabela de frequência relativa dos dados. | 89 |
| 18 | Proporção explicada pelas variâncias. | 91 |
| 19 | Correlações entre os componentes principais (Eixos) e as variáveis originais. | 95 |
| 20 | Inércia dos cafés. | 97 |
| 21 | Tabela de dados. | 98 |
| 22 | Matriz indicadora dos dados categóricos. | 99 |
| 23 | Proporção explicada pelas variâncias. | 103 |
| 24 | Correlações entre os componentes principais (eixos) e as variáveis originais. | 106 |
| 25 | Inércia dos grupos de variáveis. | 106 |

SUMÁRIO

| | | |
|-------|---|----|
| | PRIMEIRA PARTE | 17 |
| 1 | INTRODUÇÃO | 17 |
| 2 | REFERENCIAL TEORICO GERAL | 17 |
| 3 | Análise de componentes principais | 17 |
| 3.1 | Componentes principais | 18 |
| 4 | Método de análise de múltiplos fatores | 23 |
| 4.1 | Notação | 23 |
| 4.2 | Problema geral e o balanceamento dos grupos de variáveis | 24 |
| 4.3 | Representação das observações e variáveis | 26 |
| 4.4 | Montando fatores comuns | 27 |
| 4.5 | Procedimento para análise de múltiplos fatores | 27 |
| 4.6 | MFA como uma PCA | 28 |
| 4.6.1 | Os escores dos fatores globais | 29 |
| 4.6.2 | Análise parcial. | 29 |
| 4.6.3 | Correlações entre os componentes principais e as variáveis originais | 29 |
| 4.6.4 | Inércia parcial entre os grupos de variáveis | 30 |
| 5 | Método de análise de múltiplos fatores para tabelas de con- tingências - MFACT | 31 |
| 5.1 | Notação | 31 |
| 5.1.1 | Tabela de contingências X | 31 |
| 5.1.2 | Tabelas de contingências justapostas | 32 |
| 5.2 | Análise de correspondência interna. | 32 |
| 5.2.1 | Equivalência entre a CA e uma PCA | 33 |
| 5.2.2 | Princípio de análise de correspondência interna. | 33 |
| 5.2.3 | Decomposição de valor singular generalizada - GSVD | 34 |
| 5.3 | Análise de múltiplos fatores para tabela de contingências | 37 |
| 5.3.1 | Combinando ICA e MFA | 37 |
| 5.3.2 | Os dois passos da metodologia do MFACT | 38 |
| 5.3.3 | MFACT como uma PCA ponderada da tabela global | 38 |

| | | |
|-------|--|----|
| 5.4 | Propriedades do MFACT | 39 |
| 5.4.1 | As distâncias entre as linhas e entre as colunas | 39 |
| 6 | Cafés especiais. | 44 |
| 7 | Classificação e análise sensorial de café | 45 |
| | REFERÊNCIAS | 48 |
| | SEGUNDA PARTE - Artigos. | 52 |
| | ARTIGO 1 - Qualidade de cafés especiais: uma avaliação sensorial feita com consumidores utilizando a técnica MFACT | 52 |
| 1 | INTRODUÇÃO | 52 |
| 2 | MATERIAL E MÉTODOS. | 54 |
| 2.1 | Planejamento e execução do experimento | 54 |
| 2.2 | Preparo das amostras e realização do experimento sensorial. | 55 |
| 2.3 | Notação matricial para implementação e entendimento do método MFACT | 56 |
| 2.4 | Análise de correspondência para tabelas justapostas. | 57 |
| 3 | RESULTADOS E DISCUSSÃO | 61 |
| 3.1 | Análise dos perfis dos consumidores, considerando o agrupamento das notas em função dos grupos de indivíduos | 61 |
| 3.2 | Análise dos perfis dos consumidores considerando o agrupamento das notas, em função dos cafés especiais. | 64 |
| 4 | CONCLUSÕES | 68 |
| | REFERÊNCIAS | 69 |
| | ARTIGO 2 - Análise de múltiplos fatores em dados qualitativos da percepção do consumidor de café | 72 |
| 1 | INTRODUÇÃO | 72 |
| 2 | MATERIAL E MÉTODOS. | 75 |
| 2.1 | Preparo e realização do experimento. | 75 |
| 2.2 | Procedimento matricial do método análise de múltiplos fatores | 77 |
| 3 | RESULTADOS E DISCUSSÃO | 80 |
| 4 | CONCLUSÕES | 83 |

| | |
|-----------------------|----|
| REFERÊNCIAS | 84 |
| APÊNDICE | 87 |

PRIMEIRA PARTE

1 INTRODUÇÃO

Neste trabalho, trata-se de duas técnicas multivariadas, a primeira chamada de análise de múltiplos fatores (MFA ou multiple factor analysis), a segunda de análise de múltiplos fatores para tabelas de contingências (MFACT ou multiple factor analysis for contingency tables). Ambas as técnicas lidam com comparações de grupos de variáveis, sendo que a primeira trabalha com dados quantitativos e categóricos e a segunda, com tabelas de frequências.

A priori, essas técnicas, dentro das pesquisas feitas para a realização deste trabalho, ainda não são utilizadas no Brasil. Portanto, este trabalho irá contribuir metodologicamente para a implantação dessas técnicas no país. Sua utilização possibilitará aos pesquisadores brasileiros uma nova ferramenta estatística, em que se busca a comparação de grupos de variáveis.

Na primeira parte deste trabalho, apresenta-se todo o aspecto teórico de ambas as técnicas utilizadas na segunda parte.

Também no intuito de facilitar o trabalho, para quem for utilizar as técnicas em questão, foi desenvolvido um código computacional para o software R, por meio do pacote chamado MVar.pt versão 1.0 (escrito em língua portuguesa). Como este software é de domínio público, todos poderão ter acesso aos códigos fontes.

2 REFERENCIAL TEORICO GERAL

3 Análise de componentes principais

Análise de componentes principais (PCA, do inglês principal components analysis) é uma técnica multivariada, introduzida por Karl Person, em 1901, e está fundamentada no artigo de Hotelling (1933). Têm como objetivo principal explicar a estrutura de variância e covariância de um vetor aleatório, composto de p -variáveis aleatórias, por meio da construção de combinações lineares das variáveis originais. Estas combinações lineares são

chamadas de componentes principais e não são correlacionadas entre si. Se houver p -variáveis originais, é possível obter-se p componentes principais. No entanto, em geral, deseja-se obter a redução do número de variáveis a serem avaliadas e a interpretação das combinações lineares construídas, ou seja, a informação contida nas p -variáveis originais é substituída pela informação contida em k ($k < p$) componentes principais não correlacionadas. Dessa forma, o sistema de variabilidade do vetor aleatório composto das p -variáveis originais é aproximado pelo sistema de variabilidade do vetor aleatório que contém os k componentes principais. A qualidade da aproximação depende do número de componentes mantidos no sistema e pode ser medida pela avaliação da proporção de variância total explicada por eles (MINGOTI, 2005).

3.1 Componentes principais

Os componentes principais são definidos como combinações lineares de p variáveis correlacionadas, sendo que cada combinação linear é não correlacionada com a outra. As combinações lineares são em número igual ao de variáveis originais presentes no estudo. A ideia geométrica por trás dessa técnica é realizar uma rotação rígida no sistema de eixos coordenados, fazendo com que os novos eixos resultantes sejam posicionados no sentido de maior variabilidade. Em geral, a aplicação da técnica surtirá o efeito desejado quando há muitas variáveis ($p > 2$) que sejam intrínseca e fortemente relacionadas. Assim, pode-se reduzir as p dimensões originais em um sistema de $k < p$ coordenadas, sem que se tenha perdido muita informação contida no sistema original pela redução de dimensões. (FERREIRA, 2011)

Algebricamente, os componentes principais representam combinações lineares de p variáveis aleatórias X_1, X_2, \dots, X_p . Geometricamente, essas combinações lineares representam a seleção de novos eixos coordenados, que são obtidos por rotações do sistema de eixos original, representados por X_1, X_2, \dots, X_p . Os novos eixos representam as direções de máxima variabilidade. Para ilustrar a situação particular, com $p = 2$ variáveis aleatórias (X_1 e X_2) correlacionadas. Realizando uma rotação dos eixos coordena-

dos definidos pelas duas variáveis no sentido de maior variabilidade, têm-se as combinações lineares $Y_1 = a_1X_1 + a_2X_2$ e $Y_2 = b_1X_1 + b_2X_2$, sendo os mesmos componentes principais, Figura 1.

A obtenção dos componentes principais envolve a decomposição da matriz de covariância do vetor aleatório de interesse. Então seja $X = (X_1, X_2, \dots, X_p)'$ um vetor aleatório com médias $\mu = (\mu_1, \mu_2, \dots, \mu_p)'$ e matriz de covariância $\Sigma_{p \times p}$. Sejam $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ os autovalores da matriz $\Sigma_{p \times p}$, com os respectivos autovetores normalizados e_1, e_2, \dots, e_p , isto é, os autovetores e_i satisfazem às seguintes condições:

1. $e_i' e_j = 0$, para todo $i \neq j$ (pois são ortogonais);
2. $e_i' e_i = 1$, para todo $i = 1, 2, \dots, p$ (pois são ortogonais);
3. $\Sigma_{p \times p} e_i = \lambda_i e_i$, para todo $i = 1, 2, \dots, p$;

sendo o autovetor e_i denotado por $e_i = (e_{i1}, e_{i2}, \dots, e_{ip})'$.

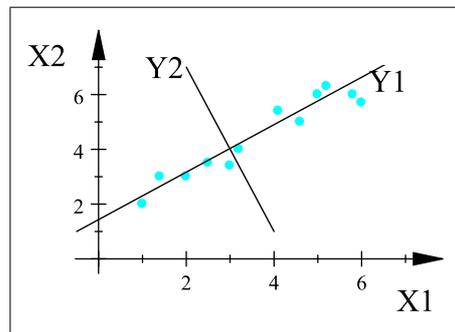


Figura 1 Ilustração da rotação dos eixos das coordenadas, definidos por duas variáveis correlacionadas X_1 e X_2 , em que Y_1 e Y_2 correspondem aos componentes principais.

Assumindo nenhuma distribuição para o vetor X , então, segue a definição.

Definição 1 *Define-se o i -ésimo componente principal Y_i como:*

$$Y_i = e_i' X = e_{i1}X_1 + e_{i2}X_2 + \dots + e_{ip}X_p,$$

em que o vetor desconhecido e_i estabelece a i -ésima combinação linear, para $i = 1, 2, \dots, p$.

A esperança, a variância e a covariância da componente Y_i são, respectivamente;

$$E[Y_i] = e_i' \mu = e_{i1} \mu_1 + e_{i2} \mu_2 + \dots + e_{ip} \mu_p, \quad i = 1, 2, \dots, p$$

$$\begin{aligned} \text{Var}[Y_i] &= \text{Var}[e_i' X] = e_i' \text{Var}[X] e_i = e_i' \sum_{p \times p} e_i \\ &= e_i' \lambda_i e_i = \lambda_i e_i' e_i = \lambda_i, \quad i = 1, 2, \dots, p \end{aligned}$$

$$\text{Cov}[Y_i, Y_k] = \text{Cov}[e_i' X, e_k' X] = e_i' \sum_{p \times p} e_k, \quad i, k = 1, 2, \dots, p.$$

Lembrando que e_i' e e_k são ortogonais, segue, ainda, que a covariância

$$\text{Cov}[Y_i, Y_k] = 0, i \neq k.$$

Usando a decomposição espectral na matriz Σ , tem-se que

$$\Sigma = P \Delta P',$$

em que P é a matriz de autovetores e Δ é a matriz de autovalores, então pode-se observar que

$$\text{tr}(\Sigma) = \text{tr}(P \Delta P') = \text{tr}(\Delta P' P) = \text{tr}(\Delta I) = \text{tr}(\Delta) = \sum_{i=1}^p \lambda_i.$$

Mas, $\text{tr}(\Sigma)$ é, obviamente, dada pela soma dos elementos da diagonal, ou seja

$$\text{tr}(\Sigma) = \sum_{i=1}^p \sigma_{ii},$$

em que se conclui que,

$$\sum_{i=1}^p \sigma_{ii} = \sum_{i=1}^p \lambda_i,$$

significa que a variabilidade total contida nas variáveis originais é igual à variabilidade total contida nos componentes principais. (FERREIRA, 2011)

Cada autovalor λ_i representa a variância de um componente principal Y_i . Como os autovalores estão ordenados em ordem decrescente, a primeira componente é a de maior variabilidade e a p -ésima é a de menor variabilidade (JOHNSON; WICHERN, 2007; MINGOTI, 2005).

Definição 2 *A proporção da variância total explicada pelo k -ésimo componente principal, mensurado de forma individual, expresso em porcentagem, é dada por*

$$P_k = \frac{\lambda_k}{\sum_{i=1}^p \sigma_{ii}} \times 100. \quad (1)$$

Se temos $k < p$ componentes principais na tentativa de gerar um modelo parcimonioso para a matriz de covariância populacional, deve-se basear em algum critério que informe quanto da variabilidade total foi explicada por ele. Para apresentar um critério como esse, considera-se, inicialmente, que o vetor de componentes principais $Y = [Y_1, Y_2, \dots, Y_p]'$ pode ser obtido de forma vetorial, dada por

$$Y = P'X.$$

Como a matriz P é ortonormal, sua inversa é igual à sua transposta, logo, o vetor X pode ser recuperado pela transformação não singular (FERREIRA, 2011)

$$X = PY.$$

Assim, segue que a covariância do vetor Y de componentes principais é dada por

$$\begin{aligned} Cov[Y] &= Cov(P'X) = P'Cov(X)P \\ &= P'\Sigma P = P'P\Delta P'P \\ &= \Delta. \end{aligned}$$

Para avaliar o grau de associação existente entre eles e as variáveis originais, deve-se obter a covariância de cada componente principal com as

variáveis originais, o que é dado por

$$Cov[Y, X] = Cov(P'X, X) = P'Cov(X) = P'\Sigma = \Delta P'. \quad (2)$$

Pode-se verificar que a covariância entre a k-ésima variável original e o i-ésimo componente principal é dada pelo produto entre o i-ésimo autovalor de Σ e o k-ésimo componente do i-ésimo autovetor, ou seja, por

$$Cov[Y_i, X_k] = \lambda_i e_{ik}.$$

Finalmente, podem-se obter as correlações entre os componentes principais e as variáveis originais da seguinte forma

$$\rho_{Y_i, X_k} = \Delta^{-\frac{1}{2}} \Delta P' V^{-\frac{1}{2}} = \Delta^{\frac{1}{2}} P' V^{-\frac{1}{2}}, \quad (3)$$

sendo $V = diag(\sigma_{ii})$. A i-ésima linha e a k-ésima coluna dessa matriz de correlações contém o coeficiente de correlação entre Y_i e X_k , dado, de forma escalar por

$$\rho_{Y_i, X_k} = \frac{\sqrt{\lambda_i} e_{ik}}{\sqrt{\sigma_{kk}}}.$$

4 Método de análise de múltiplos fatores

Análise de múltiplos fatores (MFA, do inglês *multiple factor analysis*) é uma técnica multivariada que estuda vários grupos de variáveis (colunas), que podem ser quantitativas ou categóricas, definidas no mesmo conjunto de observações (linhas). Esta técnica foi proposta pelos estatísticos franceses Escofier e Pagès (1982, 1990).

O cerne da técnica é uma análise de fator aplicada ao conjunto de variáveis nas quais cada grupo de variáveis é balanceado, conduzindo a uma representação das observações e variáveis, como em qualquer análise de fator. Devido ao balanceamento, esta análise de fator pode ser interpretada como uma análise canônica (ESCOFIER; PAGÈS, 1994, 2008).

Na Análise de Múltiplos Fatores o número de variáveis em cada grupo pode diferir e a natureza das variáveis (quantitativa ou categórica) pode variar de um grupo para o outro, mas as variáveis devem ser da mesma natureza no grupo dado (ABDI; VALENTIN, 2007).

Este método pode ser muito útil para a análise de estudos em que se podem identificar vários grupos de variáveis, ou para os estudos em que as mesmas perguntas são feitas em intervalos de tempo diversos (ABDI; WILLIAMS; VALENTIN, 2013).

A vantagem proporcionada por esse método consiste na visualização, num espaço de duas ou três dimensões, dos grupos de variáveis (sendo cada grupo representado por um ponto), das variáveis, dos eixos principais, e ainda das observações.

4.1 Notação

As I observações são descritas por J conjuntos de variáveis (Figura 2). A letra j refere-se a um conjunto, a letra k refere-se a uma coluna e K_j é o número de colunas no conjunto j . $K = \sum_{j \in J} K_j$ é o número de colunas em todos os conjuntos.

Uma tabela $I \times K_j$ está associada com cada conjunto j . As tabelas J juntas compõem uma tabela múltipla ou global $I \times K$. Para um conjunto j quantitativo, K_j é tanto o número de colunas e o número de variáveis.

Agora, para um conjunto categórico j , K_j é tanto o número de colunas e o número de categorias. Este tipo de conjunto é representado por uma tabela de variáveis indicadoras na qual a coluna k está associada com a categoria k (BÉCUE-BERTAUT; PAGÈS, 2008).

No cruzamento da linha i e coluna k (que pertencem à tabela j), temos:

- se j é um conjunto quantitativo, o valor x_{ikj} da variável k para a observação i ;
- se j é um conjunto categórico, $x_{ikj} = 1$ se i pertence à categoria k e 0, caso contrário;

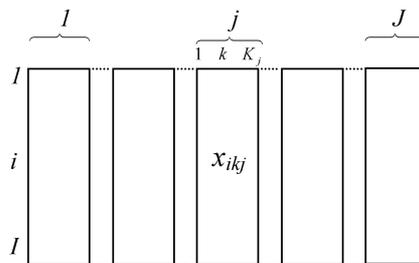


Figura 2 Tabela Global com os conjuntos de variáveis.

4.2 Problema geral e o balanceamento dos grupos de variáveis

O objetivo do MFA é integrar grupos diferentes de variáveis que descrevam as mesmas observações. Então este objetivo depara-se com os seguintes problemas:

- a) a topologia das observações descritas por todo conjunto de variáveis,
- b) avaliação das relações entre as variáveis,
- c) comparação dos grupos de variáveis.

Os itens (a) e (b) são clássicos na análise de fator (análise de componentes principal - PCA; análise de correspondência múltipla - MCA). O

item (c) sobrepõe objetivos posteriores. Assim, balancear os grupos de variáveis se faz necessário para que a influência de cada grupo seja comparável em uma análise global. Concretamente, de acordo com o ponto de vista de análise de fator, quer-se evitar a possibilidade que um único grupo tenha uma influência dominante no primeiro fator (nada pode ser requerido para fatores adicionais porque um grupo de multidimensional sempre influenciara mais fatores que um unidimensional) (ESCOFIER; PAGÈS, 1994, 2008).

Então, para cada um dos grupos de variáveis é feita uma análise PCA (para dados quantitativos) ou MCA (para dados categóricos), de acordo com o tipo das variáveis da tabela, ficando registrados os autovalores de cada análise de modo a balancear os vários grupos de variáveis. Este processo é preciso porque a análise direta obtida concatenando todas as variáveis seria dominada pelo grupo com a estrutura mais forte. Um problema semelhante pode acontecer em um PCA não normalizado: sem normalização, a estrutura é dominada pelas variáveis com a discrepância maior. Para PCA, a solução é normalizar (isto é, usar z-escore) cada variável, dividindo por seu desvio padrão. A solução proposta no MFA é semelhante: para comparar grupos de variáveis, cada grupo é normalizado dividindo-se todos seus elementos por uma quantidade chamada de primeiro valor singular, que é equivalente ao desvio padrão. Este passo é implementado executando-se um PCA ou MCA em cada grupo de variáveis, como já dito anteriormente. O primeiro valor singular é a raiz quadrada do primeiro autovalor do PCA ou MCA. Depois da normalização, as tabelas de dados são concatenadas em uma tabela de dados que é submetida a um PCA (ABDI; VALENTIN, 2007; ABDI; WILLIAMS, 2010; ABDI; WILLIAMS; VALENTIN, 2013).

Segundo Escofier e Pagès (1994), dados perdidos são permitidos no caso de variáveis categóricas, assim as observações em que não tem nenhuma categoria k , terá para 0 (zero) para todas as variáveis indicadoras associadas à categoria k .

Antes de balancear os grupos, denotam-se:

1. a_k é o peso da variável k . Para variáveis quantitativas, $a_k = 1$ para todo o k . Para variáveis categóricas, a_k é a proporção das observações

que não possuem categoria k .

2. λ_{j1} é o primeiro autovalor estimado da análise de fator aplicada ao grupo j (PCA no caso de grupo quantitativo e MCA, no caso de grupo categórico).

De acordo com Escofier e Pagès (1994, 2008), o balanceamento dos grupos consiste em atribuir o peso $a_k/\sqrt{\lambda_{j1}}$ para cada variável do grupo j . Desse modo, o primeiro autovalor estimado da análise de fator aplicada ao único grupo j torna-se 1. Portanto, os grupos são equilibrados no seguinte: em qualquer direção, a inércia máxima do subgrupo associado a um grupo é 1. Assim, em uma análise de fator global, é impossível para um único grupo causar o primeiro fator.

Claro que contribuições do grupo para análise global não são semelhantes; um grupo unidimensional não pode mostrar uma influência importante em mais de um fator e um grupo multidimensional influenciará vários fatores.

Este balanceamento é uma característica específica do MFA.

4.3 Representação das observações e variáveis

Estas representações correspondem ao clássico objetivo da análise de fator que é dizer:

- topologia das observações;
- topologia das variáveis;
- ligar os dois tipos de topologia (ESCOFIER; PAGÈS, 1994, 2008).

MFA é uma análise de fator aplicada ao conjunto de todos os grupos de variáveis.

O procedimento do método é equivalente ao PCA, no caso do grupo de variáveis quantitativas, ou MCA, no caso de grupo de variáveis categóricas.

O uso do peso $1/\sqrt{\lambda_{j1}}$ equilibra a inércia entre os diferentes grupos assim também equilibra suas influências. Levando em conta a equivalência de PCA-MCA mencionada anteriormente, este balanceamento permite aplicações nas quais alguns grupos são quantitativos e outros de variáveis categóricas (ESCOFIER; PAGÈS, 1994, 2008).

4.4 Montando fatores comuns

MFA pode ser considerado como uma análise canônica desde que o seguinte o princípio seja seguido: para medir a relação entre a variável balanceada z e o grupo j , MFA usa $L(z, j)$, ou seja, a inércia projetada do grupo j de variáveis ao longo da direção definida por z .

Esta medida tem algumas propriedades interessantes, como:

1. levando-se em conta o balanceamento das variáveis: $0 \leq L(z, j) < 1$; $L(z, j) = 0$, se z for ortogonal ao subespaço medido pelo grupo j de variáveis; $L(z, j) = 1$, se z for o primeiro componente principal do grupo j . Se todas as variáveis do grupo j forem ortogonais entre si, $L(z, j)$ será o coeficiente de correlação múltiplo;
2. em MFA, a variável z , relaciona-se com todos os grupos, satisfaz o critério: $\sum_j L(z, j)$ máximo. Este critério conduz ao primeiro componente principal de X ;
3. devido ao balanceamento das variáveis, a inércia das variáveis do mesmo grupo pode se interpretada como uma medida de relação. Isto permite interpretar o mesmo critério de um ponto de vista do PCA ou de um ponto de vista de análise canônica (ESCOFIER; PAGÈS, 1994, 2008).

4.5 Procedimento para análise de múltiplos fatores

Primeiramente, centralizam-se os dados por coluna, isto é, subtrai-se de cada elemento da coluna a sua respectiva média. Com isso, a média

de cada coluna será zero, ou seja

$$C_{ik} = x_{ik} - \bar{x}_k.$$

Em seguida, normalizam-se as colunas, isto é, dividi-se cada elemento da coluna pela raiz quadrada da soma do quadrado da respectiva coluna, ou seja

$$x_{ik}^* = \frac{C_{ik}}{\sqrt{\sum_{i=1}^I (C_{ik})^2}}.$$

Desse modo cada coluna irá se comportar como um vetor de módulo 1. Agora, se os dados forem quantitativos, aplica-se um PCA (MCA se forem de variáveis categóricas), em cada grupo X_j de variáveis para encontrar os primeiros autovalores λ_{j1} , sendo $\varphi_k = a_k / \sqrt{\lambda_{j1}}$, e segue que

$$Z_j = X_j \times \varphi_k$$

A matriz global será uma matriz $I \times K$ denotada por

$$Z = \begin{bmatrix} Z_1 & \cdots & Z_J \end{bmatrix} \quad (4)$$

Agora, como os dados já estão balanceados, segue a análise global em Z .

4.6 MFA como uma PCA

MFA consiste de um PCA na matriz global Z cujo termo geral é dado por (4). Do PCA, também chamado de análise global, os componentes principais são obtidos de modo usual por linhas e colunas de F (ABDI; VALENTIN, 2007; ABDI; WILLIAMS, 2010).

Usando a decomposição de valores singulares, sabe-se que $Z = \hat{U} \hat{\Lambda} \hat{V}'$, com $\hat{U}' \hat{U} = \hat{V}' \hat{V} = \mathbf{I}$, sendo \hat{U} e \hat{V} as matrizes de autovetores, $\hat{\Lambda} = \text{diag}(\sqrt{\lambda_i})$, e $\lambda_i > 0$ são os autovalores e \mathbf{I} é a matriz identidade, assim extraindo-se os autovalores e os autovetores da matriz Z obtém-se:

4.6.1 Os escores dos fatores globais

Os escores dos fatores globais são dados por

$$F = M^{\frac{1}{2}} \hat{U} \hat{\Lambda}, \quad (5)$$

sendo M a matriz diagonal das massas das observações, com $m_i = I$, sendo I o número de linhas da matriz Z . Em F , cada linha representa às observações, e cada coluna as variáveis.

4.6.2 Análise parcial

A análise global revela a estrutura comum das observações. Para ver como cada variável interpreta este espaço projeta-se o conjunto de dados de cada variável sobre a análise global. Isto é alcançado do seguinte modo: obtém-se a matriz de projeção, reescreve-se a equação (5), podendo os escores dos fatores globais ser computados como

$$F = M^{\frac{1}{2}} \hat{U} \hat{\Lambda} = (ZZ') \times \left(M^{\frac{1}{2}} \hat{U} \hat{\Lambda}^{-1} \right).$$

Isto mostra que

$$P = M^{\frac{1}{2}} \hat{U} \hat{\Lambda}^{-1}, \quad (6)$$

é a matriz de projeção que transforma a matriz ZZ' em escores dos fatores.

A matriz de projeção é usada para projetar os grupos sobre o espaço global, fazendo

$$F_j = J \times Z_j Z_j' P, \quad (7)$$

sendo J é o número de conjuntos de variáveis.

4.6.3 Correlações entre os componentes principais e as variáveis originais

Para obter as correlações entre os componentes principais e as variáveis originais, basta utilizar a equação (3).

4.6.4 Inércia parcial entre os grupos de variáveis

As relações entre os grupos de variáveis são analisadas calculando-se a inércia parcial de cada variável por cada dimensão da análise global. Isto é calculado, para cada variável, como a soma das projeções quadradas das variáveis do vetor singular \hat{V} de Z , multiplicado pelo autovalor correspondente. Porque os vetores singulares são normalizados, a soma das inércias parciais para todos os grupos para uma determinada dimensão é igual ao seu autovalor, ou seja

$$R_{j,i} = \lambda_p \times \sum_j^{J_k} \hat{V}_{j,i}^2, \quad (8)$$

em que j refere-se ao grupo, p é o componente principal (eixo) e λ_p é o autovalor da decomposição de valores singulares.

5 Método de análise de múltiplos fatores para tabelas de contingências - MFACT

Análise de múltiplos fatores para tabelas de contingências (MFACT, do inglês multiple factor analysis for contingency tables) é uma técnica multivariada que compara as estruturas de várias tabelas de contingências com a mesma linha e colunas diferentes (ou vice-versa), geralmente construídas de diferentes amostras ou populações. Uma apresentação teórica detalhada da metodologia pode ser encontrada em Bécue-Bertaut e Pagès (2004).

Por estrutura de uma tabela, quer-se dizer a relação entre linhas e colunas conforme expresso na análise de correspondência (AC) por meio de interfaces gráficas. As comparações têm de lidar com ambas as linhas e colunas, isto é, a estrutura induzida pelas linhas ao longo dos diferentes conjuntos de colunas e também a estrutura induzida sobre as linhas pelos diferentes conjuntos de colunas (BÉCUE-BERTAUT; PAGÈS, 2004; ESCOFIER; PAGÈS, 2008).

5.1 Notação

5.1.1 Tabela de contingências X

Considera-se uma tabela de contingências X de ordem $I \times J$. O termo geral p_{ij} é a frequência relativa, ou seja, a proporção com a qual cada linha i ($i = 1, \dots, I$) está associada com a coluna j ($j = 1, \dots, J$), de tal modo que $\sum_{ij} p_{ij} = 1$. Denotou-se a i -ésima linha marginal como $p_{i.} = \sum_j p_{ij}$ e a j -ésima coluna marginal por $p_{.j} = \sum_i p_{ij}$.

Ao inspecionar uma tabela de contingência, os dados são considerados por meio de proporções condicionais, geralmente chamadas de perfis. Assim, o i -ésimo perfil da linha é $\{p_{ij}/p_{i.}; j = 1, \dots, J\}$, e o j -ésimo perfil da coluna é $\{p_{ij}/p_{.j}; i = 1, \dots, I\}$. A média do perfil da linha é calculada pela ponderação das linhas usando a soma de sua margem $p_{i.}$. Analogamente, para a média do perfil da coluna usa-se $p_{.j}$ (BÉCUE-BERTAUT; PAGÈS, 2004; ESCOFIER; PAGÈS, 2008).

5.1.2 Tabelas de contingências justapostas

Dada T tabelas $X_1, \dots, X_t, \dots, X_T$, justapostas em uma única tabela conforme Figura 3, sendo as colunas não precisam ser as mesmas em todas as tabelas. A tabela X_G é chamada de tabela global, de dimensão $I \times J$, e a subtabela X_t de dimensão $I \times J_t$, de t -ésima subtabela. Denotou-se p_{ijt} a frequência relativa ou proporção, na tabela t ($t = 1, \dots, T$), com a qual a linha i (= observação i ; $i = 1, \dots, I$) está associada com a coluna j (= categoria j ; $j = 1, \dots, J_t$; $\bigcup_t J_t = J$), $\sum_{ijt} p_{ijt} = 1$.

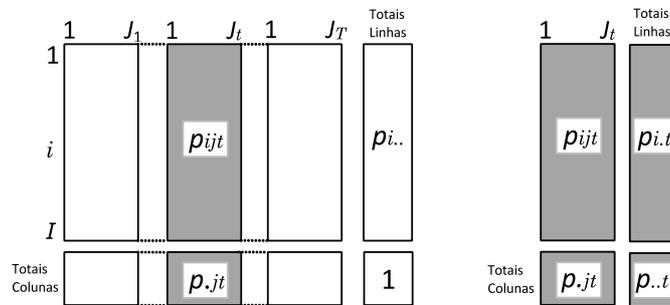


Figura 3 Tabelas de contingências justapostas

Denotou-se a linha marginal da tabela X_G como $p_{i..} = \sum_{jt} p_{ijt}$. A coluna marginal da tabela X_G é $p_{.jt} = \sum_i p_{ijt}$. A linha marginal da tabela t , como uma subtabela da tabela X_G , $p_{i.t} = \sum_j p_{ijt}$, e a soma dos termos da tabela t dentro da tabela X_G é $p_{..t} = \sum_{ij} p_{ijt}$ (BÉCUE-BERTAUT; PAGÈS, 2004; ESCOFIER; PAGÈS, 2008).

5.2 Análise de correspondência interna

Para comparar tabelas de contingências que têm margens diferentes, Benzécri (1983) e Escofier e Drouet (1983) apresentaram CA intratabelas (análise de correspondência intratabelas). Cazes e Moreau (1991, 2000), e mencionados por Greenacre e Blasius (2006), propuseram uma generalização desta metodologia que chamaram de CA interna (análise de correspondência interna - ICA), para lidar com tabelas com partições ou estruturas gráficas nas linhas e colunas.

5.2.1 Equivalência entre a CA e uma PCA

A análise de correspondência (CA) de uma simples tabela Y pode ser vista como uma análise de componentes principais (PCA) ponderada realizada na tabela, tendo o termo geral

$$\frac{p_{ij} - p_{i..}p_{.j}}{p_{i..}p_{.j}}, \quad (9)$$

usando $\{p_{i.}; i = 1, \dots, I\}$ como pesos das linhas (e como uma métrica no espaço das colunas) e $\{p_{.j}; j = 1, \dots, J\}$ como pesos das colunas (e como uma métrica no espaço das linhas). Assim, a CA da tabela T (tabelas justapostas) é equivalente a um PCA ponderado na tabela que tem o termo geral

$$\frac{p_{ijt} - p_{i..}p_{.jt}}{p_{i..}p_{.jt}}, \quad (10)$$

usando $\{p_{i.}; i = 1, \dots, I\}$ como pesos das linhas (e como uma métrica no espaço das colunas) e $\{p_{.jt}; j = 1, \dots, J_t; t = 1, \dots, T\}$ como pesos das colunas (e como uma métrica no espaço das linhas).

A inércia entre a distribuição observada $\{p_{ijt}\}$ e a distribuição que corresponde a $\{p_{i..}p_{.jt}\}$, é determinada pelo coeficiente ϕ , também conhecido como coeficiente de contingência da média quadrática de Person

$$\phi^2 = \sum_t \sum_j \sum_i \frac{(p_{ijt} - p_{i..}p_{.jt})^2}{p_{i..}p_{.jt}} = \frac{\chi^2}{n}, \quad (11)$$

sendo χ^2 é a estatística do qui-quadrado, e n o tamanho da amostra quando se trabalha com a tabela de frequência, e com a frequência relativa $n = 1$ (BÉCUE-BERTAUT; PAGÈS, 2004; GREENACRE; BLASIUS, 2006).

5.2.2 Princípio de análise de correspondência interna

Escofier (1984) propôs uma generalização de CA para qualquer modelo que tenham a mesmas margens como as tabelas de dados. ICA pode ser vista como uma generalização na qual o modelo de independência geral é substituído pela independência entre linhas e colunas dentro de cada tabela

t (modelo de independência de intratabelas).

A nuvem de pontos que corresponde à tabela de contingências X_t é denotada por N_{J_t} . Geometricamente, em ICA, cada N_{J_t} é transferida para que o centroide coincida com o centroide global.

Usando a equivalência entre a CA e o PCA de uma tabela convenientemente transformada (com pesos satisfatórios para linhas e colunas), os resultados de ICA também podem ser obtidos executando-se um PCA ponderado na tabela que tem o termo geral seguinte

$$\frac{p_{ijt} - \frac{p_{i.t} p_{.jt}}{p_{..t}}}{p_{i..} p_{.jt}} = \frac{1}{p_{i..}} \begin{bmatrix} p_{ijt} & - p_{i.t} \\ p_{.jt} & p_{.t} \end{bmatrix}, \quad (12)$$

usando $\{p_{i..}; i = 1, \dots, I\}$ como peso das linhas (e como uma métrica no espaço de colunas) e $\{p_{.jt}; j = 1, \dots, J_t; t = 1, \dots, T\}$ como pesos das colunas (e como uma métrica no espaço das linhas) (GREENACRE; BLASIUS, 2006).

Obsevação 1 *Existe uma diferença entre o N_{J_t} em sua CA separada e em ICA: a métrica do espaço de coluna difere e, assim, a forma da nuvem N_{J_t} é modificada. Na Tabela 1 recapitulam-se as diferenças entre os mesmos dados nas duas análises do ponto de vista do PCA. Tem-se, ainda, que quando as margens de linhas são proporcionais entre todas as tabelas, ICA é equivalente à CA aplicada às tabelas concatenadas (GREENACRE; BLASIUS, 2006).*

5.2.3 Decomposição de valor singular generalizada - GSVD

Dada a matriz A de ordem $n \times m$, a decomposição de valor singular generalizada envolve o uso de duas matrizes quadradas definidas positivas de ordem $n \times n$ e $m \times m$, respectivamente. Estas duas matrizes expressam restrições impostas respectivamente sobre as linhas e as colunas de A . Assim, se M é uma matriz de ordem $n \times n$ que expressa as restrições para as linhas de A , e W uma matriz de ordem $m \times m$ que expressa as restrições

Tabela 1 Coeficiente relativo à tabela t nas análises CA e ICA, aplicado na tabela X_G .

| | Termo Geral | Peso das Linhas (M) | Peso das Colunas (W) |
|---------------------------|---|---------------------------|---------------------------|
| CA aplicado em N_{J_t} | $\frac{p_{ijt} - \frac{p_{i.t}}{p_{..t}}p_{.jt}}{\frac{p_{i.t}}{p_{..t}}p_{.jt}}$ | $\frac{p_{i.t}}{p_{..t}}$ | $\frac{p_{.jt}}{p_{..t}}$ |
| ICA aplicado em N_{J_t} | $\frac{p_{ijt} - \frac{p_{i.t}}{p_{..t}}p_{.jt}}{p_{i..}p_{.jt}}$ | $p_{i..}$ | $p_{.jt}$ |

M = métrica no espaço das colunas e W = métrica no espaço das linhas.

para as colunas de A , logo, tem-se que

$$\tilde{A} = M^{\frac{1}{2}}AW^{\frac{1}{2}},$$

é a matriz com as respectivas restrições a linhas e colunas.

Também define-se a matriz A como

$$A = M^{-\frac{1}{2}}\tilde{A}W^{-\frac{1}{2}} \iff \tilde{A} = M^{\frac{1}{2}}AW^{\frac{1}{2}}.$$

Então, a matriz A pode ser decomposta em

$$A = \tilde{U}\tilde{\Lambda}\tilde{V}' \text{ com } \tilde{U}'M\tilde{U} = \tilde{V}'W\tilde{V} = \mathbf{I}. \quad (13)$$

Em outras palavras, na decomposição de valor singular generalizada (GSVD, do inglês generalized singular value decomposition), as matrizes dos autovetores singulares generalizados \tilde{U} e \tilde{V} são ortogonais sob as restrições impostas por M e W , ou seja, é igual à matriz identidade \mathbf{I} (ABDI, 2007).

Esta decomposição é obtida como resultado da decomposição dos valores singulares padrão (SVD, do inglês singular value decomposition) em \tilde{A} :

$$\tilde{A} = P\Lambda Q' \text{ com } P'P = Q'Q = \mathbf{I}_r,$$

sendo P ($n \times r$) e Q ($m \times r$) ortogonais por coluna, e r o posto da matriz

\tilde{A} .

As matrizes dos autovetores generalizados são obtidas por

$$\tilde{U} = M^{-\frac{1}{2}}P \text{ e } \tilde{V} = W^{-\frac{1}{2}}Q.$$

A matriz diagonal de valores singulares é simplesmente igual à matriz de valores singulares de $\tilde{\Lambda}$

$$\tilde{\Lambda} = \Lambda.$$

Ainda podemos verificar que

$$A = \tilde{U}\tilde{\Lambda}\tilde{V}',$$

e, por substituição, tem-se que

$$\begin{aligned} A &= M^{-\frac{1}{2}}\tilde{A}W^{-\frac{1}{2}} \\ &= M^{-\frac{1}{2}}P\Lambda Q'W^{-\frac{1}{2}} \\ &= \tilde{U}\tilde{\Lambda}\tilde{V}'. \end{aligned}$$

Assim, para mostrar a condição de (13), é suficiente mostrar que

$$\tilde{U}'M\tilde{U} = P'M^{-\frac{1}{2}}MM^{-\frac{1}{2}}P = P'P = \mathbf{I},$$

e, ainda, segue que

$$\tilde{V}'W\tilde{V} = Q'W^{-\frac{1}{2}}WW^{-\frac{1}{2}}Q = Q'Q = \mathbf{I}.$$

Bécue-Bertaut e Pagès (2004) e Escofier e Pagès (2008), citam que o termo geral da diagonal $\tilde{\Lambda} = \text{diag}(\sqrt{\lambda_i})_i$, sendo $\lambda_i > 0$, são os autovalores, e com estas notações, os escores das matrizes para as linhas e colunas são, respectivamente,

$$F = U\Lambda^{\frac{1}{2}}, \tag{14}$$

e

$$G = V\Lambda^{\frac{1}{2}}. \tag{15}$$

5.3 Análise de múltiplos fatores para tabela de contingências

5.3.1 Combinando ICA e MFA

Abdessemed e Escofier (1996) já propuseram uma extensão do MFA para comparar várias tabelas de contingências, mas o método é restringido ao caso no qual as margens de linhas das tabelas diferentes são idênticas. Esta extensão é generalizada para o caso de tabelas que têm margens de linhas diferentes. A ideia básica é combinar ICA (que lida com o problema das diferenças entre as linhas das margens) e MFA (que equilibra a influência das subtabelas na análise global). Este método é chamado MFA para tabelas de contingências (MFACT) e pode ser visto como um PCA.

- analisando a mesma tabela como ICA cujo termo geral é determinado em (12);
- preservando os pesos das linhas usado em ICA ($p_{i..}$);
- modificando pesos das colunas para equilibrar a influência das tabelas t .

Uma ponderação cumpre esta modificação dos pesos de coluna:

- de modo idêntico para todas as colunas de uma tabela X_t (para não modificar a estrutura de tabela X_t);
- para equilibrar a inércia máxima axial de cada nuvem N_{J_t} .

Esta ponderação consiste em multiplicar o peso ($p_{.jt}$) da coluna (j, t) em ICA por $1/\lambda_1^t$, sendo λ_1^t o primeiro autovalor do PCA separado da subtabela X_t cujo termo geral é determinado em (12), com os pesos usados em ICA ($p_{i..}$ para linhas i , $p_{.jt}$ para a coluna (j, t) ; conforme Tabela 1). Chama-se esta análise restringida de “pseudosseparada CA”, e a CA simples habitual da tabela X_t de “CA separada”.

A escolha de $p_{i..}$ como peso para linhas i , parece natural. Em CA, o grupo i tem um peso alto quando representa muitas observações. No caso de tabelas com frequências mais altas que outras, estas podem dominar

esses pesos fortemente. Então, uma pergunta surge: o balanceamento das tabelas também deveria intervir definindo os pesos de linhas? Não há uma única resposta e os usuários diferentes podem ter opiniões diferentes e podem depender do conjunto de dados. Alternativamente, para equilibrar as influências das tabelas nos pesos das linhas, os dados podem ser transformados primeiro em frequência relativa ($n_{ijt}/n_{..t}$ em vez de n_{ijt}), antes da concatenação. A formulação de MFACT não depende desta transformação (GREENACRE; BLASIUS, 2006).

5.3.2 Os dois passos da metodologia do MFACT

Primeiro passo: análises da pseudosseparada

Análise de correspondência, usando pesos ($p_{i..}$) e ($p_{.jt}$), é aplicado a cada tabela X_t para calcular o primeiro autovalor λ_1^t . Estes pesos de linhas e colunas são distintos desses usados em CA separada que são, respectivamente, $p_{i..t}/p_{..t}$ e $p_{.jt}/p_{..t}$ (veja Tabela 1).

Segundo passo: análise global

Este passo estabelece resultados globais que são

- análogo a esses de CA aplicada nas tabelas concatenadas, aqui chamada de tabela Z .
- específico para tabelas múltiplas (principalmente, sobrepôs representações de estruturas de linha-grupo induzidas por cada tabela X_t , ou seja, representações de fatores derivaram de análises da pseudosseparada) (GREENACRE; BLASIUS, 2006).

5.3.3 MFACT como uma PCA ponderada da tabela global

MFA para tabelas de contingências consiste do PCA nas métricas M e W da tabela Z (tabelas concatenadas) cujo termo geral é dado por (12). Note que a tabela Z é dada pelas T subtabelas Z_t . É fácil verificar que a ponderação das linhas e a das colunas é centrada e a ponderação das colunas

de cada conjunto t também é centrada (BÉCUE-BERTAUT; PAGÈS, 2004; ESCOFIER; PAGÈS, 2008).

Do PCA, também chamado de análise global, os componentes principais são obtidos de modo usual por colunas e linhas conforme as equações (14) e (15), F_s e G_s respectivamente, sendo $s = 1, \dots, \min(I - 1, J - 1)$, sendo s o posto das matrizes de autovetores U_s e V_s , associado aos valores singulares λ_s (BÉCUE-BERTAUT; PAGÈS, 2004; ESCOFIER; PAGÈS, 2008).

Componentes canônicos - F_s podem ser escritos como

$$F_s = \frac{1}{\lambda_s} \sum_t Z_t W_t Z_t' M F_s.$$

No MFACT, os componentes canônicos, associados com os componentes principais globais F_s em cada grupo t , são definidos do seguinte modo

$$\begin{aligned} F_s &= \sum_t F_s', \\ F_s^t &= \frac{1}{\lambda_s} Z_t W_t Z_t' M F_s. \end{aligned} \quad (16)$$

Reconstrução de dados - A clássica fórmula do PCA conduz à seguinte expressão de decomposição

$$\begin{aligned} Z &= \sum_s F_s \lambda_s^{-\frac{1}{2}} G_s' = \sum_s U_s \sqrt{\lambda_s} V_s', \\ p_{ijt} &= \frac{p_{i.t}}{p_{.t}} \cdot p_{.jt} + p_{i..} \cdot p_{.jt} \left(\sum_s \frac{F_s(i) G_s(j, t)}{\sqrt{\lambda_s}} \right). \end{aligned} \quad (17)$$

5.4 Propriedades do MFACT

5.4.1 As distâncias entre as linhas e entre as colunas

As distâncias entre as linhas

O quadrado da distância entre as linhas i e l , calculado a partir das

coordenadas dadas em (12), utilizando o peso de $p_{.jt}/\lambda_1^t$ para a coluna j do conjunto de t , é

$$d^2(i, l) = \left[\sum_t \frac{1}{\lambda_1^t} \sum_{j \in J_t} \left(\frac{p_{ijt}}{p_{i..}} - \frac{p_{ljt}}{p_{l..}} \right)^2 \cdot \frac{1}{p_{.jt}} \right] - \left[\sum_t \frac{1}{\lambda_1^t \cdot p_{..t}} \left(\frac{p_{i.t}}{p_{i..}} - \frac{p_{l.t}}{p_{l..}} \right)^2 \right]. \quad (18)$$

Na expressão (18), desprezando a ponderação pelo inverso do primeiro autovalor:

- o primeiro termo corresponde à distância entre as linhas i e l na CA das tabelas justapostas;
- o segundo termo corresponde à distância entre as linhas i e l do CA da tabela que contém as somas por linha e por subtabela. O termo geral $i.t$ nesta tabela é a soma da linha i na tabela t . Vemos aqui como esta última tabela é neutralizada pela recentralização de cada subtabela em suas próprias margens (BÉCUE-BERTAUT; PAGÈS, 2004; ESCOFIER; PAGÈS, 2008).

Proximidades entre colunas

O quadrado da distância entre a coluna j (pertencente a tabela t) e a coluna k (pertencente a tabela de r), calculado a partir de coordenadas dadas em (12), é

$$d^2(j \in t, k \in r) = \sum_i \frac{1}{p_{i..}} \left[\left(\frac{p_{ijt}}{p_{.jt}} - \frac{p_{i.t}}{p_{..t}} \right) - \left(\frac{p_{ikr}}{p_{.kr}} - \frac{p_{i.r}}{p_{..r}} \right) \right]^2, \quad (19)$$

$$d^2(j \in t, k \in r) = \sum_i \frac{1}{p_{i..}} \left[\left(\frac{p_{ijt}}{p_{.jt}} - \frac{p_{ikr}}{p_{.kr}} \right) - \left(\frac{p_{i.t}}{p_{..t}} - \frac{p_{i.r}}{p_{..r}} \right) \right]^2. \quad (20)$$

As aproximações entre as colunas podem ser interpretadas como uma associação semelhante com linhas.

Caso 1: as colunas pertencem à mesma tabela ($t = r$).

A proximidade entre as duas colunas é interpretada em termos de semelhança entre os perfis, exatamente como na CA habitual.

Caso 2: as colunas pertencem a diferentes tabelas ($t \neq r$).

Os perfis das colunas são relativos aos perfis médios, como mostrado nas duas expressões do quadrado da distância. Expressão (19) mostra que o perfil de uma coluna intervém por seu desvio a partir do perfil significativo da tabela correspondente. A expressão (20) mostra como as diferenças entre os perfis de coluna são relativos a diferenças entre os perfis médios (BÉCUE-BERTAUT; PAGÈS, 2004; ESCOPIER; PAGÈS, 2008).

Representação global de linhas e colunas

O PCA realizado na seção (5.3.3) fornece uma representação das linhas e das colunas que deve ser analisadas em conjunto. Além disso, estas representações podem ser sobrepostas a um biplot, como na análise de correspondência. Esta é a consequência das fórmulas de transição que se referem às coordenadas de linhas e de colunas, tal como analisado a seguir.

Fórmulas de transição e regras de interpretação

Linhas entre colunas: a relação dando (ao longo do eixo s) a coordenada $F_s(i)$ da linha i a partir das coordenadas $\{G_s(j, t), j = 1, \dots, J_t; t = 1, \dots, T\}$ de colunas é

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_t \sum_{j \in J_t} \frac{p_{jt}}{\lambda_1^t} \left[\frac{1}{p_{i..}} \left[\frac{p_{ijt}}{p_{.jt}} - \frac{p_{i.t}}{p_{..t}} \right] \right] G_s(j, t). \quad (21)$$

O centroide do perfil coluna pertencente a J_t sendo 0, tem-se

$$\sum_{j \in J_t} p_{jt} G_s(j, t) = 0.$$

Usando esta propriedade, a fórmula de transição pode ser escrita da

seguinte forma

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_t \frac{1}{\lambda_1^t} \frac{p_{i.t}}{p_{i..}} \left[\sum_{j \in J_t} \frac{p_{ijt}}{p_{i.t}} G_s(j, t) \right]. \quad (22)$$

Exceto para uma constante, cada linha encontra-se no centro de gravidade das colunas associadas a esta linha.

Globalmente, uma linha é atraída pelas colunas com quais ela está associada.

As colunas entre linhas: a expressão (ao longo do eixo s) para a coordenada $G_s(j, t)$ da coluna j, t a partir das coordenadas $\{F_s(i), i = 1, \dots, I\}$ de linhas é

$$G_s(j, t) = \frac{1}{\sqrt{\lambda_s}} \left[\sum_i \left(\frac{p_{ijt}}{p_{.jt}} - \frac{p_{i.t}}{p_{..t}} \right) F_s(i) \right]. \quad (23)$$

Como o coeficiente de $F_s(i)$ pode ser negativo, as colunas não estão no centroide das linhas, exceto quando os pesos das linhas são os mesmos em todas as tabelas.

Este coeficiente mede a discrepância entre o perfil das colunas j , e a margem da coluna da tabela t . Uma coluna é atraída (ou repelida) pelas linhas que são mais (ou menos) associadas como se houvesse a independência ou não, entre as linhas e colunas na tabela t (BÉCUE-BERTAUT; PAGÈS, 2004; ESCOFIER; PAGÈS, 2008).

Representações das linhas por meio de componentes canônicos

A fim de comparar as estruturas das linhas induzidas pelas tabelas diferentes, é possível utilizar os componentes canônicos F_s^t , como dado por (16), para se obter uma representação sobrepostas dessas estruturas, chamadas estruturas parciais.

Na prática, os componentes canônicos das linhas, também chamados componentes parciais, podem ser calculados com muita facilidade. Portanto, as linhas das tabelas separadas t são completadas por zeros, de tal modo que elas têm o mesmo número de colunas que a tabela global, e projetadas como linhas suplementares, sobre os eixos globais. Desse modo, sobrepõe

as representações das linhas associadas com cada tabela t (chamadas linhas parciais) e as representações associadas com a tabela global X_G (chamadas linhas globais ou significativas). Em particular, estas representações parciais podem estar relacionadas com a representação da coluna por meio de uma fórmula de transição "restrita".

$$F_s^t(i) = \frac{1}{\sqrt{\lambda_s}} \cdot \frac{p_{i.t}}{p_{i..}} \left[\sum_{j \in J_t} \frac{p_{ijt}}{p_{i.t}} \cdot \frac{G_s(j, t)}{\lambda_1^t} \right]. \quad (24)$$

Esta expressão é derivada da equação (22) restringida às colunas da tabela t .

No gráfico sobrepondo representações parciais, estas representações parciais são dilatadas pelo coeficiente T (número de tabelas). Assim, um ponto (média) da linha está situado no centroide dos correspondentes pontos da linha parcial (BÉCUE-BERTAUT; PAGÈS, 2004; ESCOFIER; PAGÈS, 2008).

6 Cafés especiais

O café, uma das bebidas mais consumidas mundialmente, é cultivado em mais de 60 países (CHAGAS et al., 2009). De acordo com Arruda et al. (2009), o consumo do café indica que as motivações para o seu consumo no Brasil têm um significado social associado aos hábitos e costumes.

Segundo a Associação Brasileira da Indústria de Café - ABIC (2014), os brasileiros consomem 4,87 kg de café torrado por habitante/ano, perdendo apenas para o consumo de água. A ABIC estima que o consumo de café crescerá em 2015, alcançando 21 milhões de sacas.

Atualmente, o consumo de café tem sido incentivado por especialistas da área de saúde. Trabalhos como o de Alves, Casal e Oliveira (2009) mostram que o consumo do café está associado a efeitos benéficos no desenvolvimento de doenças como: diabetes tipo II, asma, cirrose alcoólica, determinados tipos de cancro, doença de Parkinson e Alzheimer

Neste cenário destacam-se os cafés especiais que, segundo a BSCA (Brazil Specialty Coffee Association) este segmento representa, cerca de 12% do mercado internacional da bebida (BRAZIL SPECIALTY COFFEE ASSOCIATION - BSCA, 2015).

Apesar de o café ser uma commodity, os cafés especiais têm vantagens na sua comercialização, como a fixação de contratos de longo prazo com preços mais estáveis em relação aos cafés não especiais (CHAGAS et al., 2009).

De acordo com a BSCA (2015), a definição de cafés especiais é ampla, pois os atributos da qualidade desses cafés devem levar em consideração desde a lavoura, a torra e o preparo da bebida. Assim, características físicas, como as variedades, origens, tamanho, cor, preocupações sócioambientais, além de condições de trabalho da mão de obra cafeeira, bem como os sistemas de produção, também devem ser consideradas nesta avaliação.

Os cafés especiais têm um valor de venda entre 30% e 40% maior em relação ao café cultivado de modo convencional, ultrapassando em alguns casos a barreira dos 100% (BSCA, 2015).

Os cafés especiais são divididos em cinco categorias principais que

são: qualidade superior (Gourmet), origem certificada (estate coffee), sombreado, orgânico e fair trade (LEÃO; PAULA, 2010), a saber:

Café gourmet: são cafés que apresentam características de melhor qualidade, com grãos de aspecto uniforme, quase isento de defeitos e uma bebida mole ou estritamente mole (BSCA, 2015; LEÃO; PAULA, 2010).

Café de origem certificada: São os cafés diferenciados pela região produtora, clima, solo, temperatura. Alguns dos atributos de qualidade do produto são inerentes à região onde a planta é cultivada (BSCA, 2015; LEÃO; PAULA, 2010).

Café sombreado: de acordo com Leão e Paula (2010) são cafés produzidos a partir de plantas cultivadas sob a cobertura de árvores, com isso preservando o meio ambiente. É uma tentativa de reproduzir o habitat natural da planta, podendo ser produzido sob diversos níveis de sombreamento.

Café orgânico: são cafés cultivados exclusivamente com fertilizantes orgânicos, e o controle de pragas e doenças, é feitos biologicamente, o que contribui para a manutenção dos solos e do meio ambiente (BSCA, 2015; LEÃO; PAULA, 2010).

Café fair trade: segundo Leão e Paula (2010) e Saes (2006), são cafés adquiridos por meio de cooperativas que recebem o certificado fair trade, também conhecido como comércio justo ou solidário, em que as preocupações são as condições sócioeconômicas dos produtores nos países em desenvolvimento, envolvendo o pagamento de salários justos aos trabalhadores, trabalho cooperativo, educação do consumidor, sustentabilidade ambiental, suporte técnico e financeiro e respeito à identidade cultural.

7 Classificação e análise sensorial de café

Na cadeia de processamento do café, a análise sensorial é um dos aspectos mais importantes, pois, além de garantir a homogeneidade nos lotes comercializados, também garante a qualidade dos mesmos.

Dentre as metodologias utilizadas na análise sensorial dos cafés especiais, a mais utilizada é o protocolo SCAA, refere-se a (Speciality Coffe

Association of America). Este protocolo determina quais os equipamentos necessários em uma análise sensorial, como preparar uma amostra, como avaliar a amostra, os procedimentos, os atributos a serem avaliados, e os resultados finais da avaliação. Este protocolo tem por objetivo permitir uma correta caracterização sensorial de um determinado tipo de café (SPECIALITY COFFEE ASSOCIATION OF AMERICA - SCAA, 2009).

As análises sensoriais têm com o propósito determinar diferentes características sensoriais entre diferentes amostras; descrever aromas e sabores e determinar uma preferência entre produtos. O protocolo possibilita a avaliação de onze atributos na degustação do café (SCAA, 2009), que são: fragrância/aroma, uniformidade, ausência de defeitos (xícara limpa), doçura, sabor, acidez, corpo, finalização, equilíbrio, defeitos e avaliação global.

Os resultados dessa avaliação são estabelecidos a partir da atribuição de notas entre 6 e 10 pontos, com intervalos de 0,25 (um quarto de ponto), conforme Tabela 2. As notas abaixo de 6 pontos são aplicáveis aos cafés comerciais (SCAA, 2009).

Tabela 2 Escala de qualidade dos cafés.

| Bom | Muito bom | Excelente | Excepcional |
|------|-----------|-----------|-------------|
| 6,00 | 7,00 | 8,00 | 9,00 |
| 6,25 | 7,25 | 8,25 | 9,25 |
| 6,50 | 7,50 | 8,50 | 9,50 |
| 6,75 | 7,75 | 8,75 | 9,75 |

Fonte: SCAA (2009).

O resultado final da análise sensorial é obtido calculando-se a soma dos 11 atributos pontuados, subtraindo-se o valor referente ao atributo defeitos. Esses resultados são expressos de acordo com a classificação da Tabela 3 (SCAA, 2009).

Tabela 3 Escala de classificação de cafés especiais.

| Pontuação | Descrição | Classificação |
|------------|-----------|---------------------|
| 90 - 100 | Exemplar | Cafés especiais |
| 85 - 89,99 | Excelente | Cafés especiais |
| 80 - 84,99 | Muito Bom | Cafés especiais |
| < 80 | Bom | Cafés não especiais |

Fonte: SCAA (2009).

REFERÊNCIAS

- ABDESSEMED, L.; ESCOFIER, B. Analyse factorielle multiple de tableaux de frequencys: comparaison avec l'analyse canonique des correspondences. **Journal de la Société de Statistique de Paris**, Paris, v. 137, n. 2, p. 3-18, 1996.
- ABDI, H. Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition (GSVD). In: SALKIND, N. J. (Ed.). **Encyclopedia of measurement and statistics**. Thousand Oaks: Sage, 2007. p. 907-912.
- ABDI, H.; VALENTIN, D. Multiple factor analysis (MFA). In: SALKIND, N. J. (Ed.). **Encyclopedia of measurement and statistics**. Thousand Oaks: Sage, 2007. p. 657-663.
- ABDI, H.; WILLIAMS, L. Principal component analysis. **WIRES Computational Statistics**, New York, v. 2, n. 4, p. 433-459, July/Aug. 2010.
- ABDI, H.; WILLIAMS, L.; VALENTIN, D. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. **WIRES Computational Statistics**, New York, v. 5, n. 2, p. 149-179, Feb. 2013.
- ALVES, R. C.; CASAL, S.; OLIVEIRA, B. Benefícios do café na saúde: mito ou realidade? **Química Nova**, São Paulo, v. 32, n. 8, p. 2169-2180, set. 2009.
- ARRUDA, A. C. et al. Justificativas e motivações do consumo e não consumo de café. **Ciência e Tecnologia de Alimentos**, Campinas, v. 29, n. 4, p. 754-763, out./dez. 2009.
- ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA DE CAFÉ. **Pesquisa nacional FIESP/IBOPE sobre o perfil do consumo de alimentos no Brasil**. Disponível em: <http://www.abic.com.br/media/EST_PESQFoodTrendsI.pdf>. Acesso em: 10 dez. 2014.
- BÉCUE-BERTAUT, M.; PAGÈS, J. Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data.

Computational Statistics & Data Analysis, New York, v. 52, n. 6, p. 3255-3268, Feb. 2008.

BÉCUE-BERTAUT, M.; PAGÈS, J. A principal axes method for comparing contingency tables: MFACT. **Computational Statistics & Data Analysis**, New York, v. 45, n. 3, p. 481-503, Feb. 2004.

BENZECRI, J. Analyse de l'inertie intraclasse par l'analyse d'un tableau de contingence: intra-classinertia analysis through the analysis of a contingency table. **Les Cahiers de l'Analyse des Données**, Paris, v. 8, n. 3, p. 351-358, 1983.

BRAZIL SPECIALTY COFFEE ASSOCIATION. **O que são cafés especiais**. Disponível em: <<http://bsca.com.br>>. Acesso em: 19 fev. 2015.

CAZES, P.; MOREAU, J. Analyse des correspondances d'un tableau de contingence dont les lignes et les colonnes sont munies d'une structure de graphes bistochastique. In: MOREAU, J.; DOUDIN, P. A.; CAZES, P. (Ed.). **In l'analyse des correspondances et les techniques connexes**: approches nouvelles pour l'analyse statistique des données. Berlin: Springer, 2000. p. 87-103.

CAZES, P.; MOREAU, J. Analysis of a contingency table in which the rows and the columns have a graph structure. In: DIDAY, E.; LECHEVALLIER, Y. (Ed.). **In symbolic-numeric data analysis and learning**. New York: Nova Science, 1991. p. 271-280.

CHAGAS, I. S. P. et al. Avaliação do mercado de cafés especiais. In: CONGRESSO SOCIEDADE BRASILEIRA DE ECONOMIA, ADMINISTRAÇÃO E SOCIOLOGIA RURAL, 47., 2009, Porto Alegre. **Anais...** Porto Alegre: UFRGS, 2009. 1 CD-ROM.

ESCOFIER, B. Analyse factorielle en référence à un modèle: application à l'analyse d'un tableau d'échanges. **Revue de Statistique Appliquée**, Paris, v. 32, n. 4, p. 25-36, 1984.

ESCOFIER, B.; DROUET, D. Analyse des différences entre plusieurs tableaux de fréquence. **Les Cahiers de l'Analyse des Données**, Paris, v. 8, n. 4, p. 491-499, 1983.

ESCOFIER, B.; PAGÈS, J. **Analyse factorielles simples et multiples**. Paris: Dunod, 1990. 267 p.

ESCOFIER, B.; PAGÈS, J. **Analyses factorielles simples et multiples**: objectifs, méthodes et interprétation. 4th ed. Paris: Dunod, 2008. 318 p.

ESCOFIER, B.; PAGÈS, J. **Comparaison de groupes de variables définies sur le même ensemble d'individus**: un exemple d'applications. Le Chesnay: Institut National de Recherche en Informatique et en Automatique, 1982. 121 p.

ESCOFIER, B.; PAGÈS, J. Multiple factor analysis (AFUMULT package). **Computational Statistics & Data Analysis**, New York, v. 18, n. 1, p. 121-140, Aug. 1994.

FERREIRA, D. F. **Estatística multivariada**. 2. ed. rev. e ampl. Lavras: UFLA, 2011. 675 p.

GREENACRE, M.; BLASIUS, J. **Multiple correspondence analysis and related methods**. New York: Taylor and Francis, 2006. 607 p.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, Arlington, v. 24, p. 417-441, Sept. 1933.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 6th ed. New Jersey: Prentice Hall, 2007. 794 p.

LEÃO, E. A.; PAULA, N. M. A produção de cafés especiais no brasil e a emergência de novos padrões de competitividade. In: ENCONTRO REGIONAL DE ECONOMIA, 13., 2010, Porto Alegre. **Anais...** Porto Alegre: ANPEC, 2010. 1 CD-ROM.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada**: uma abordagem aplicada. Belo Horizonte: UFMG, 2005. 297 p.

PAGÈS, J. Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. **Revue de Statistique Appliquée**, Paris, v. 50, n. 4, p. 5-37, 2002.

PAGÈS, J. Multiple factor analysis: main features and application to sensory data. **Revista Colombiana de Estadística**, Bogotá, v. 27, n. 1, p. 1-26, 2004.

SPECIALITY COFFE ASSOCIATION OF AMERICA. **SCAA protocols**: cupping specialty coffee. Long Beach, 2009. 7 p.

SEGUNDA PARTE - ARTIGOS

ARTIGO 1 - Qualidade de cafés especiais: uma avaliação sensorial feita com consumidores utilizando a técnica MFACT

1 INTRODUÇÃO

De modo geral, ao considerar os testes de aceitação ou preferência, na análise dos resultados de um experimento sensorial, o principal foco é avaliar a intenção do provador em discernir a qualidade sensorial de um produto em relação aos demais. A questão é que, além do problema de natureza estatística a ser exemplificado pelo erro de medida, no preenchimento de uma ficha sensorial ou na análise de dados categorizados, resultantes da distribuição de classes de frequência das notas sensoriais, alguns fatores externos pertinentes à formação do painel sensorial poderão não ser contemplados.

Diante desse fato, há evidências para acreditar que, as análises estatísticas poderão gerar resultados e conclusões incoerentes. Portanto, é notório um amplo campo de pesquisa, no que tange à proposta de uso de novas abordagens estatísticas que agreguem resultados mais precisos na análise, na aceitação e na discriminação da qualidade sensorial de um produto. Com esse propósito, a análise de dados categorizados multivariados, representados em tabelas de contingência, merece uma atenção especial, pois além de resultar em uma perda de informação, os fatores, em geral desconhecidos, porém, relacionados à qualidade sensorial, devem ser identificados.

Dentre as inúmeras técnicas propostas para analisar dados categorizados, a análise de múltiplos fatores aplicada a tabelas de contingência, conhecida por MFACT (BÉCUE-BERTAUT; PAGÈS, 2004; ESCOFIER; PAGÈS, 2008), se caracteriza por permitir analisar grupos de variáveis com tamanhos diferentes e de natureza distintas.

Esta técnica analisa simultaneamente várias tabelas de contingência. Entretanto, comparando-a com outras técnicas que permitem analisar simultaneamente diversas tabelas, Zárraga e Goitisoló (2009) ressaltam que, ao

analisar tabelas com diferentes estruturas, definidas pelo número de linhas e colunas, nas quais representam as relações entre as variáveis, certamente os resultados fatoriais são afetados. Diante desses problemas, nota-se que a técnica MFACT é promissora em relação à análise de procrustes generalizada (Sugnet et al., 2006), por considerar pesos atribuídos a cada tabela, na estimação dos fatores.

Em se tratando da aplicabilidade da técnica MFACT em experimentos sensoriais, uma característica a ser ressaltada na viabilidade do uso dessa técnica pode ser justificada por Calle (1956), Chagas, Carvalho e Costa (1996), ao mencionarem que a subjetividade da prova de xícara é limitada pela aptidão do provador, podendo ser deformada e que não é possível de ser medida. Nesse contexto, ao discretizar as notas relacionadas a um atributo sensorial, dada a variabilidade dos dados e o método utilizado para determinar o número de classes de uma tabela de contingência, diferentes estruturas poderão ser geradas. Ao considerar, outros atributos, múltiplas tabelas de contingência são geradas, porém, analisadas individualmente. A vantagem em utilizar a técnica MFACT é justamente analisar simultaneamente inúmeras tabelas.

Com relação a fatores externos, exemplificados pelos perfis dos consumidores, Chagas, Carvalho e Costa (1996) citam um experimento sensorial, relacionado à prova de xícara de café, em que as amostras foram provadas por diferentes degustadores. Decorrente da análise estatística, visando avaliar a validade da “prova de xícara”, os degustadores capacitados evidenciaram que a prova de xícara é perfeitamente válida, dentro de certos limites especificados pelos autores (FAIRBANKS BARBOSA et al., 1962).

Assim, em virtude do que foi mencionado, este trabalho foi realizado com o objetivo de propor uma análise estatística utilizando a técnica MFACT em um experimento sensorial formado a partir de quatro grupos de consumidores, com o propósito de avaliar a qualidade sensorial de quatro tipos de cafés especiais, produzidos com diferentes processamentos e altitudes.

2.2 Preparo das amostras e realização do experimento sensorial

Em concordância com o parecer consubstanciado do Conselho de Ética e Pesquisa, registrado no CAAE: 14959413.1.0000.5148, o preparo das amostras de café 100% arábica foi feito retirando-se todos os grãos defeituosos e torrados respeitando-se o prazo máximo de 24 horas para a degustação.

O ponto de torra foi determinado visualmente, utilizando-se o sistema de classificação de cor por meio de discos padronizados (SCAA/Agtron Roast Color Classification System). Em relação ao preparo da bebida, manteve-se a concentração de 7% m/v, utilizando água filtrada pronta para consumo, livre de qualquer contaminante e sem adição de açúcar. Com estas especificações, quatro tipos de cafés especiais foram codificados nas amostras por A, B, C e D, de acordo com a Tabela 4.

Tabela 4 Descrição dos cafés especiais avaliados na análise sensorial.

| Tipo | Genótipo | Altitude | Processamento |
|------|-----------------|------------------|-------------------|
| A | Bourbon Amarelo | acima de 1.200 m | Natural |
| B | Acaiá | abaixo de 1.100m | Cereja descascado |
| C | Acaiá | abaixo de 1.100m | Natural |
| D | Bourbon Amarelo | acima de 1.200m | Cereja descascado |

Para cada tipo de café, as características sensoriais foram avaliadas no teste de aceitação em relação às suas características sensoriais: acidez, corpo, doçura e nota geral em quatro sessões, nas quais os consumidores eram voluntários com conhecimento básico em relação à análise sensorial de cafés.

Visando melhores esclarecimentos, a estruturação da tabela justaposta considerando as notas sensoriais obtidas nos grupos sensoriais para os atributos nota, corpo, acidez e doçura é apresentada no layout na Tabela 5.

O grupo (G=1 e 2) foi formado por consumidores que receberam um treinamento em relação às avaliações sensoriais. Os integrantes dos demais

Tabela 5 Layout da tabela justaposta para análise de cada grupo de indivíduos em relação às avaliações sensoriais para os cafés especiais.

| | Grupo 1 | | | | ... | Grupo 4 | | | |
|-----------|-----------|-------|-------|-------|-----|-----------|-------|-------|-------|
| Notas | N_T | C_T | A_T | D_T | ... | N_T | C_T | A_T | D_T |
| [0; 2.5) | f_{iTG} | | | | ... | f_{iTG} | | | |
| [2.5; 5) | | | | | | | | | |
| [5; 7.5) | | | | | | | | | |
| [7.5; 10] | | | | | | | | | |

N = nota geral, C = corpo, A = acidez e D = doçura.

G = 1, 2, 3, 4 (grupos de indivíduos) e T = A, B, C, D (grupos dos cafés).

f_{iTG} = frequência da categoria i no grupo T e tabela G.

grupos (G=3 e 4) não receberam nenhum treinamento, porém, se caracterizam por serem profissionais técnicos ou pesquisadores da área relacionada à pesquisa cafeeira.

Em se tratando do perfil dos consumidores em relação a alguma preferência de cafés especiais, a estruturação da tabela justaposta é apresentada na Tabela 6.

Tabela 6 Layout da tabela justaposta para análise dos cafés especiais, em relação às avaliações sensoriais para os cafés especiais.

| | Café A | | | | ... | Café D | | | |
|-----------|-----------|-------|-------|-------|-----|-----------|-------|-------|-------|
| Notas | N_G | C_G | A_G | D_G | ... | N_G | C_G | A_G | D_G |
| [0; 2.5) | f_{iGT} | | | | ... | f_{iGT} | | | |
| [2.5; 5) | | | | | | | | | |
| [5; 7.5) | | | | | | | | | |
| [7.5; 10] | | | | | | | | | |

N = nota geral, C = corpo, A = acidez e D = doçura.

G = 1, 2, 3, 4 (grupos de indivíduos) e T = A, B, C, D (grupos dos cafés).

f_{iGT} = frequência da categoria i no grupo G e tabela T.

2.3 Notação matricial para implementação e entendimento do método MFACT

Segundo Bécue-Bertaut e Pagès (2008), a estrutura de uma tabela de contingência é representada pela relação entre linhas e colunas, con-

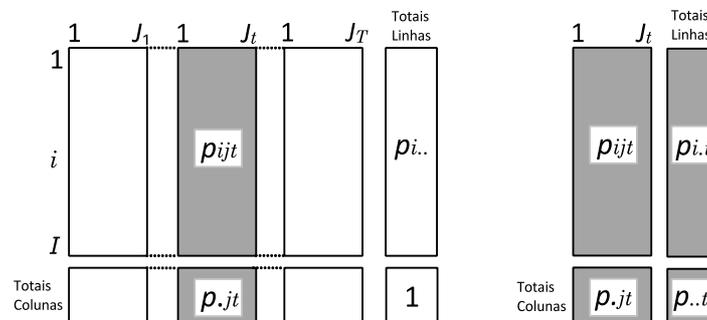
forme, usualmente, é feito na análise de correspondência (GREENACRE; BLASIUS, 2006; BÉCUE-BERTAUT; PAGÈS, 2004; ESCOFIER; PAGÈS, 2008).

Com este propósito, consideramos X como uma tabela de contingência de ordem $I \times J$. O termo geral p_{ij} representa a proporção relativa, com a qual a i -ésima linha i ($i = 1, \dots, I$) está associada com a j -ésima coluna j ($j = 1, \dots, J$), de tal modo que $\sum_{ij} p_{ij} = 1$. Assumindo, a i -ésima linha marginal representada por $p_{i.} = \sum_j p_{ij}$, sendo D_I uma matriz diagonal formada por $p_{i.}$. De forma análoga, temos a j -ésima coluna marginal definida por $p_{.j} = \sum_i p_{ij}$.

2.4 Análise de correspondência para tabelas justapostas

Consideremos várias tabelas de contingência, representadas por $X_1, \dots, X_t, \dots, X_T$, porém, justapostas em uma única tabela, conforme ilustrado na Figura 5.

Figura 5 Layout de tabelas de contingência justapostas.



Dado que cada tabela indica uma sub-tabela X_t , de dimensão $I \times J_t$, denotamos p_{ijt} como a frequência relativa, na tabela t ($t = 1, \dots, T$), com a qual a i -ésima linha i ($i = 1, \dots, I$) está associada com a j -ésima coluna j ($j = 1, \dots, J$), restrito a $\sum_{ijt} p_{ijt} = 1$.

Ressaltamos, por meio da Figura 5, que o número de colunas não é o mesmo e todas as sub-tabelas e denominando a tabela “global” por X_G com

dimensão $I \times J$, torna-se possível, representar a linha marginal de X_G pela proporção $p_{i..} = \sum_{jt} p_{ij t}$. Analogamente, a proporção da coluna marginal é dada por $p_{.jt} = \sum_i p_{ij t}$.

A implementação do método MFACT foi realizada considerando inicialmente a fundamentação teórica da análise de correspondência interna, cujo acrônimo é denominado por ICA, (GREENACRE; BLASIUS, 2006; BÉCUE-BERTAUT; PAGÈS, 2004; ESCOFIER; PAGÈS, 2008), sendo esta uma generalização da análise de correspondência simples, a qual permite comparar várias tabelas de frequência com estruturas e marginais diferentes. Contudo, as marginais referentes às variáveis categóricas descritas na linha são proporcionais entre todas as tabelas. Dessa forma, nota-se que ICA é equivalente à CA aplicada às tabelas concatenadas. (GREENACRE; BLASIUS, 2006; BÉCUE-BERTAUT; PAGÈS, 2004; ESCOFIER; PAGÈS, 2008)

Seguindo esta metodologia, o método MFACT foi implementado em dois procedimentos, descritos nos passos 1 e 2, descritos a seguir.

Passo 1: Implementou-se a análise de correspondência, considerando os pesos $p_{i..}$ e $p_{.jt}$, em cada tabela, representada por P_t (Tabela 7).

Tabela 7 Coeficiente relativo à tabela t nas análises CA e ICA, aplicado na tabela P .

| | Termo Geral | Peso das Linhas (M) | Peso das Colunas (W) |
|---------------------------|--|---------------------------|---------------------------|
| CA aplicado em N_{J_t} | $\frac{p_{ij t} - \frac{p_{i..} p_{.jt}}{p_{..t}}}{\frac{p_{i..} p_{.jt}}{p_{..t}}}$ | $\frac{p_{i..}}{p_{..t}}$ | $\frac{p_{.jt}}{p_{..t}}$ |
| ICA aplicado em N_{J_t} | $\frac{p_{ij t} - \frac{p_{i..} p_{.jt}}{p_{i..} p_{..t}}}{\frac{p_{i..} p_{.jt}}{p_{i..} p_{..t}}}$ | $p_{i..}$ | $p_{.jt}$ |

M = métrica no espaço das colunas e W = métrica no espaço das linhas. N_{J_t} corresponde à subnuvem do perfil da coluna que corresponde à tabela de contingência P . Em ICA, cada N_{J_t} é transferida para que o centroide coincida com o centroide global.

Passo 2: Combinou-se a ICA com o método MFA, de modo a ponderar a influência da estrutura das subtabelas na análise global. Esta ponderação consistiu em multiplicar o peso ($p_{.jt}$) da coluna (j, t) em ICA por $1/\lambda_1^t$, sendo λ_1^t o primeiro autovalor do PCA separado da sub-tabela P_t , com os pesos determinados na ICA, isto é $p_{i..}$ para linhas i , $p_{.jt}$ para a coluna (j, t); (Tabela 7).

A combinação destes dois métodos é denominada por (MFACT), em que as distâncias entre as linhas i e l , fixada a j -ésima coluna do conjunto t , considerando o peso $\frac{p_{.jt}}{\lambda_1^t}$, é definida em (25).

$$d^2(i, l) = \left[\sum_t \frac{1}{\lambda_1^t} \sum_{j \in J_t} \left(\frac{p_{ijt}}{p_{i..}} - \frac{p_{ljt}}{p_{l..}} \right)^2 \cdot \frac{1}{p_{.jt}} \right] - \left[\sum_t \frac{1}{\lambda_1^t \cdot p_{..t}} \left(\frac{p_{i.t}}{p_{i..}} - \frac{p_{l.t}}{p_{l..}} \right)^2 \right]. \quad (25)$$

Ressalta-se que em (25), desprezando-se a ponderação pelo inverso do primeiro autovalor, temos que o primeiro termo corresponde à distância (entre os perfis i e l) na CA das tabelas justapostas. O segundo termo corresponde à distância (entre os perfis i e l) do CA da tabela que contém as somas por linha e por sub-tabela. O termo geral $i.t$ nesta tabela é a soma da linha i na tabela t . Em relação à distância entre as colunas, o quadrado da distância entre a coluna j pertencente à tabela t e a coluna k , pertencente a tabela de r , é definido em (26) e (27).

$$d^2(j \in t, k \in r) = \sum_i \frac{1}{p_{i..}} \left[\left(\frac{p_{ijt}}{p_{.jt}} - \frac{p_{i.t}}{p_{..t}} \right) - \left(\frac{p_{ikr}}{p_{.kr}} - \frac{p_{i.r}}{p_{..r}} \right) \right]^2, \quad (26)$$

$$d^2(j \in t, k \in r) = \sum_i \frac{1}{p_{i..}} \left[\left(\frac{p_{ijt}}{p_{.jt}} - \frac{p_{ikr}}{p_{.kr}} \right) - \left(\frac{p_{i.t}}{p_{..t}} - \frac{p_{i.r}}{p_{..r}} \right) \right]^2. \quad (27)$$

As aproximações entre as colunas podem ser interpretadas em duas situações:

1^a - Se as colunas pertencem à mesma tabela ($t = r$), a proximidade entre as duas colunas é interpretada em termos de semelhança entre os perfis, exatamente como na CA usual.

2^a - Se as colunas pertencem a diferentes tabelas ($t \neq r$), os perfis das colunas são relativos ao quadrado da distância, conforme apresentado em (26), evidenciando que o perfil de uma coluna intervém por seu desvio a partir do perfil significativo da tabela correspondente. Em (27), as diferenças entre os perfis de coluna são relativos às diferenças entre os perfis médios (BÉCUE-BERTAUT; PAGÈS, 2004; ESCOFIER; PAGÈS, 2008).

Determinadas as distâncias, as representações das linhas e colunas são analisadas conjuntamente, dispostas em um mapa perceptual, sobrepostas a um biplot. Dessa forma, seguem as interpretações de forma semelhante à análise de correspondência. Por fim, seguindo a metodologia apresentada, utilizou-se o software R para a implementação das rotinas computacionais que permitiram a obtenção dos resultados, descritos a seguir.

3 RESULTADOS E DISCUSSÃO

3.1 Análise dos perfis dos consumidores, considerando o agrupamento das notas em função dos grupos de indivíduos

Em concordância com a descrição da metodologia apresentada na seção anterior, seguindo a estruturação, conforme o layout da Tabela 5, os resultados descritos na Tabela 8 evidenciam que o estudo das associações, entre as variáveis categóricas representadas em cada subtabela, analisadas simultaneamente, poderá ser explicado em um gráfico bidimensional, uma vez que a proporção da variação explicada nos dois primeiros componentes principais é igual a 75,81%.

Tabela 8 Proporção da variação amostral explicada pelos três primeiros componentes.

| Componentes | λ_i | % da variação amostral | % da variação amostral acumulada |
|-------------|-------------|------------------------|----------------------------------|
| 1 | 3,003 | 46,74 | 46,74 |
| 2 | 1,869 | 29,08 | 75,81 |
| 3 | 1,554 | 24,19 | 100,00 |

Em continuidade a esse resultado, os perfis ilustrados na Figura 6 correspondem às notas dadas pelos indivíduos, classificadas em categorias. Nesse contexto, a interpretação dessa figura em relação aos grupos de indivíduos é compreendida como cada grupo de indivíduos está associado às notas classificadas nas categorias. Assim, torna-se possível visualizar que os grupos de consumidores diferiram as notas nas categorias $[5,0; 7,5)$ e $[7,5; 10]$, justamente por apresentar em suas respectivas localizações em quadrantes diferentes, em relação às demais categorias.

Com o propósito de buscar uma associação de cada grupo na classificação dessas notas, os resultados ilustrados na Figura 7 na qual, as projeções das categoriais de notas dos indivíduos, tendo por base cada categoria resultante da análise global, correspondem ao centroide das posições em relação aos grupos de consumidores avaliados. Dessa forma, as “linhas” que unem os

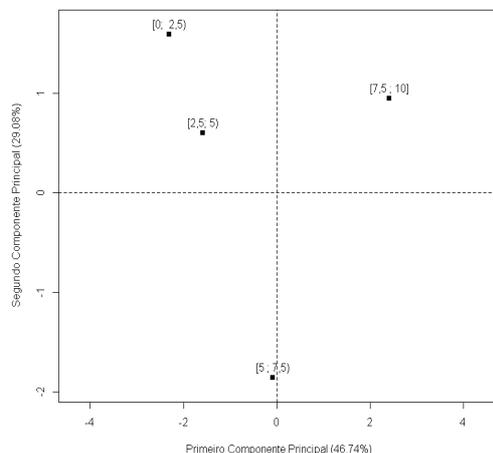


Figura 6 Gráfico correspondente à análise global das faixas dos indivíduos.

grupos com a posição global das categorias de notas dos indivíduos corroboraram em afirmar que o efeito do treinamento realizado com os consumidores classificados nos Grupo 1 resultou em uma avaliação sensorial com respostas homogêneas em relação aos demais grupos. Tal efeito foi perceptível para as categorias de notas [5,0; 7,5] e [7,5; 10].

Ressalta-se que esta última categoria de notas se enquadra nas especificações da Associação Americana de Cafés Especiais (SCAA), que classifica os cafés especiais com apresentação de nota final acima ou igual a 80 pontos (Alves, 2011; Lingle, 2011).

No caso dos outros grupos, nota-se, nas projeções, um distanciamento não homogêneo em relação ao centroide, o que impede de fazer mesmas afirmações relacionadas ao efeito do treinamento.

Uma vez que parte do pressuposto de que, em uma análise sensorial, provadores treinados com a habilidade de discriminar pequenas diferenças entre as amostras, dada a homogeneidade entre os resultados, proporcionados pelas avaliações dos treinadores, os resultados obtidos por meio de uma metodologia estatística são precisos e coerentes.

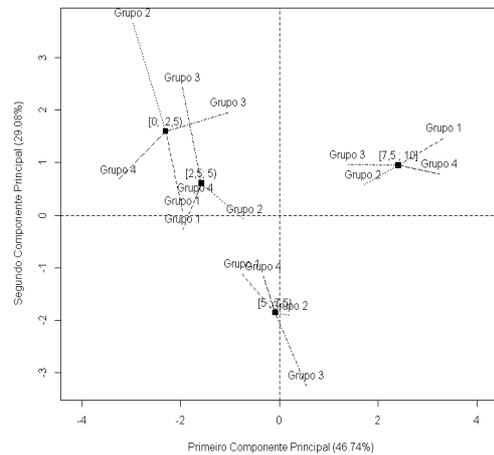


Figura 7 Gráfico correspondente à análise global das faixas dos indivíduos com as análises dos grupos de indivíduos.

Entretanto, ao considerar as notas provenientes de consumidores, naturalmente surge uma limitação nos estudos sensoriais, caracterizada por fatores diversos, como fadiga, indisposição para realizar todas as provas e diferença entre as habilidades dos consumidores. Enfim, todos estes fatores contribuem para que notas discrepantes possam surgir no decorrer da análise sensorial. Nesse sentido, o efeito do treinamento dado a um grupo de provadores não treinados para a realização desse experimento proporcionou um resultado favorável em relação a apresentar a respostas homogêneas com os demais grupos.

No que tange aos resultados referentes à inércia parcial estimada em cada grupo, sendo esta, interpretada como a quantidade de variação restituída em cada componente obtido na análise global, os resultados descritos na Tabela 9 corroboram as afirmações relacionadas à similaridade entre os Grupos 1 e 4.

Tabela 9 Valores referentes às inércias computadas nos três primeiros eixos para cada grupo de consumidor.

| Grupos de consumidores | | | | | |
|------------------------|--------|--------|--------|--------|---------------|
| Eixo | 1 | 2 | 3 | 4 | Inércia total |
| 1 | 0,9251 | 0,5943 | 0,4986 | 0,9852 | 3,0032 |
| 2 | 0,2734 | 0,5397 | 0,7693 | 0,2861 | 1,8685 |
| 3 | 0,1967 | 0,6720 | 0,4397 | 0,2459 | 1,5543 |

3.2 Análise dos perfis dos consumidores considerando o agrupamento das notas, em função dos cafés especiais

Seguindo a metodologia relacionada à técnica MFACT e à estruturação da tabela, conforme o layout representado na Tabela 6, os resultados descritos na Tabela 10 evidenciam que o estudo das associações entre as variáveis categóricas representadas em cada subtabela, analisadas simultaneamente, poderá ser explicado em um espaço bidimensional, uma vez que a proporção da variação explicada nos dois primeiros componentes é de 80,94% da variação amostral.

Tabela 10 Proporção da variação amostral explicada pelos três primeiros componentes.

| Componentes | λ_i | % da variação amostral | % da variação amostral acumulada |
|-------------|-------------|------------------------|----------------------------------|
| 1 | 2,989 | 47,31 | 47,31 |
| 2 | 2,125 | 33,63 | 80,94 |
| 3 | 1,204 | 19,06 | 100,00 |

Assim como na análise anterior, pelos resultados ilustrados na Figura 8, referentes às categorias de notas, percebe-se que, novamente, a categoria de notas [7,5; 10] é projetada em um quadrante separado em relação aos demais.

Em se tratando da distribuição das faixas de notas, dadas pelos consumidores em relação aos tipos de cafés avaliados, os resultados ilustrados na Figura 9, dadas as projeções das faixas dos indivíduos, na qual a posição

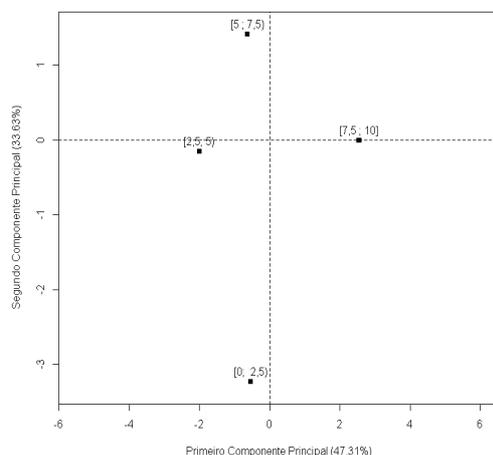


Figura 8 Gráfico correspondente à análise global das faixas dos indivíduos.

de cada faixa na análise global, corresponde ao centroide das posições em relação aos cafés especiais avaliados.

Logo, as linhas que unem os cafés com a posição global das faixas dos indivíduos corroboram a afirmação que há uma similaridade das notas em relação aos cafés A e D,. Ou seja, segundo a classificação dada na Tabela 4, ambos os cafés têm em comum, a altitude e o genótipo Bourbon Amarelo, diferenciando-se apenas no processamento, o que é corroborado pelos resultados descritos na Tabela 11.

Tabela 11 Valores referentes às inércias computadas nos três primeiros eixos, para cada grupo de tipos cafés..

| Eixo | Grupos de Tipos Especiais de Cafés | | | | Inércia total |
|------|------------------------------------|--------|--------|--------|---------------|
| | A | B | C | D | |
| 1 | 0,9923 | 0,6568 | 0,3521 | 0,9881 | 2,9893 |
| 2 | 0,3633 | 0,7379 | 0,8532 | 0,1709 | 2,1254 |
| 3 | 0,2963 | 0,5101 | 0,2707 | 0,1272 | 1,2043 |

Importante ressaltar que, embora os resultados sejam provenientes de uma análise exploratória, segundo a literatura, muitos trabalhos, como

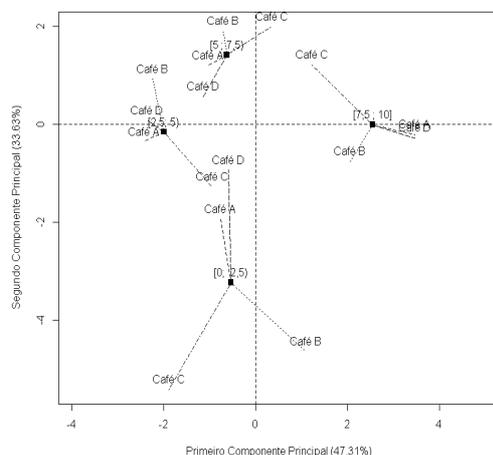


Figura 9 Gráfico correspondente à análise global das faixas dos indivíduos com as análises dos tipos de cafés.

os de Decazy, et al. (2003); Avelino et al. (2005); Bertrand et al. (2008); Barbosa et al. (2014); Barbosa et al. (2012) e Chagas et al. (2013) e Kitzberger et al. (2013) confirmam que a localização geográfica com altitudes mais elevadas é um dos principais fatores que influenciam a qualidade sensorial.

Em se tratando da inércia parcial de cada grupo, interpretada como a quantidade de variação restituída em cada componente inerente à análise global, os resultados descritos na Tabela 11 corroboram as afirmações relacionadas à similaridade dos cafés A e D, ou seja, verifica-se uma concordância entre a discriminação dos cafés A e D, sendo estes, diferenciados apenas no processamento, ou seja, o café A processado via seca (natural) e D, via úmida (cereja).

Em relação aos cafés B e C, segundo Abrahão et al. (2008), essa similaridade poderá ser explicada pela composição química dos grãos de café, uma vez que tal composição é influenciada por fatores genéticos, ambientais e condições de manejo pré e pós-colheita. Uma vez que, os cafés B e C são do mesmo genótipo, produzidos em um mesmo ambiente (altitude), há

evidências para afirmar que esses fatores contribuíram para a similaridade na análise dos resultados sensoriais para todos os grupos de consumidores.

Outra característica conveniente de ser ressaltada é que em comum os cafés B e C são caracterizados por frutos vermelhos e a produtividade desses cafés em regiões com altitude abaixo de 1.000m, em geral, resulta em bebidas com menor acidez, doçura e sabores menos complexos, predominando sabores de caramelo e corpo mediano a baixo. Com certa frequência, essas variedades podem também produzir cafés com leve adstringência. Esse perfil sensorial está mais próximo do padrão conhecido pela maioria dos consumidores brasileiros, excetuando-se o consumo de cafés defeituosos.

Em se tratando da similaridade entre os cafés A e D, justifica-se esse agrupamento inicialmente pelo fato de destes cafés serem de altíssima qualidade. Porém, era esperado que os consumidores apresentassem uma tendência em preferir o café do tipo D, uma vez que, este café é mais popular, por ser mais suave e adocicado, com fácil percepção na identificação de sabores agradáveis. Em se tratando do tipo de processamento, o café natural, por sua complexidade e presença de notas muito exóticas, dificulta essa percepção e preferência dos consumidores, pois se trata de uma bebida muito diferente, em relação aos cafés convencionais, os quais os consumidores estão acostumados a beber.

De modo geral, quanto às similaridades entre os cafés A, D e B, C vários componentes são atuantes, dignos de serem justificados em relação à qualidade sensorial. Os principais aspectos citados por Figueiredo et al. (2013) e Borém (2008), envolvem os sistemas de cultivo, no entanto, a característica final (fenótipo) depende da constituição genética e das condições ambientais (Leroy et al., 2006; Pereira et al., 2010).

4 CONCLUSÕES

1. Em concordância com os objetivos e a metodologia proposta, conclui-se que a técnica MFACT - Múltiplos fatores aplicada a tabela de contingência é viável para ser aplicada em análise de dados sensoriais categorizados na identificação de perfis sensoriais de grupos de consumidores com diferentes habilidades em relação à qualidade de cafés especiais.
2. Em se tratando da análise da qualidade sensorial dos cafés especiais, a técnica MFACT permitiu identificar a similaridade entre os cafés especiais produzidos em ambientes (altitudes) e processamentos diferentes, em todos os grupos de consumidores avaliados.

REFERÊNCIAS

- ABRAHÃO, S.A. et al. Compostos bioativos em café integral e descafeinado e qualidade sensorial da bebida. **Pesquisa Agropecuária Brasileira**. Brasília, v.43, n.12, p.1799-1804, 2008.
- ALVES, H.M.R. et al. Características ambientais e qualidade da bebidas dos cafés do estado de Minas Gerais. **Informe Agropecuário**. Belo Horizonte, v.32, n.261, p.18-29, 2011.
- AVELINO, J. et al. Effects of slope exposure, altitude and yield on coffee quality in two altitude terroirs of Costa Rica, Orosi and Santa María de Dota. **Journal of Science Food and Agriculture**, Sussex, v.85, n.11, p.1869-1876, 2005.
- BARBOSA, J.N. et al. Coffee Quality and its interactions with environmental factors in Minas Gerais, Brazil. **Journal of Agricultural Science**, Toronto, n.4, v.5,p.181-190, 2012.
- BARBOSA, J. N. et al. Discrimination of Production Environments of Specialty Coffees by Means of Stable Isotopes and Discriminant Model. **Journal of Agricultural Science**, v. 6, p. 55-64, 2014.
- BÉCUE-BERTAUT, M.;PAGÈS,J.; A principal axes method for comparing contingency tables: MFACT. **Computational Statistics & Data Analysis**, v. 45, p. 481-503, 2004.
- BÉCUE-BERTAUT, M.;PAGÈS,J.; Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. **Computational Statistics & Data Analysis**, v. 52, p. 3255-3268, 2008.
- BERTRAND, B. et al. Comparison of the effectiveness of fatty acids, chlorogenic acids, and elements for the chemometric discrimination of coffee (*Coffea Arabica* L.) varieties and growing origins. **Journal of Agricultural and Food Chemistry**. Washington, v.56, n.6, p.2273-2280, 2008.
- BORÉM, F.M. Processamento do café. In: BORÉM, F. M. (Ed.). **Pós-Colheita do Café**. Lavras: Editora UFLA, 2008.

CHAGAS, S.J.R.; CARVALHO, V.D.; COSTA, L. Caracterização química e qualitativa de cafés de alguns municípios de três regiões produtoras de Minas Gerais. **Pesquisa Agropecuária Brasileira**, v.31., n.8, p.555-561, 1996.

CHAGAS, E. N. et al. Selection of robust estimators used in analysis of sensory characteristics and identification of environments conducive to specialty coffee production. **Advanced Crop Science**, v. 3, p. 515-524, 2013.

CALLE, H. V. Bom ou mau café? **Boletim da Superintendência dos Serviços do Café**, São Paulo, v. 31, n. 354, p. 51-52, 1956.

DECAZY, F. et al. Quality of different Honduran coffees in relation to several environments. **Journal of Food Science**, Chicago, v.68, n.7, p.2356-2361, 2003.

ESCOFIER, B.;PAGÈS,J.; **Analyses factorielles simples et multiples: objectifs, méthodes et interprétation**. 4th. ed. Paris: Dunod, 2008. 318 p.

FAIRBANKS BARBOSA et al. **Estudos preliminares sobre a prova de xícara de café**. São Paulo: Secretaria da Agricultura, 1962. 38 p.

FIGUEIREDO, L. P. et al. The Potential for High Quality Bourbon Coffees From Different Environments. **Journal of Agricultural Science**, v. 5, p. 87-98, 2013.

GREENACRE, M.; BLASIUS, J.; **Multiple Correspondence Analysis and Related Methods**. 1th. ed. New York: Taylor and Francis Group, 2006. 607 p.

KITZBERGER, G.C.S.et al. Composição química de cafés árabica de cultivares tradicionais e modernas. **Pesquisa Agropecuária Brasileira. Brasília**, v.48, n.11, p.1498-1506, 2013.

LEROY, T. et al. Genetics of coffee quality. **Brazilian Journal of Plant Physiology**. Piracicaba, v.18, n.1, p.229-242, 2006.

LINGLE, T.R.; **The coffee cupper's handbook: systematic guide to the sensory evaluation of coffee's flavor**. 4.ed. Long Beach: Specialty Coffee Association of America, 66p. 2011.

OSSANI, P.C.; CIRILLO, M.A.; **MVar.pt**: análise Multivariada (Brazilian Portuguese). URL <http://CRAN.R-project.org/package=MVar.pt>. R package version 1.0. 2014.

PEREIRA, M.C.; CHALFOUN, S.M ; CARVALHO, G.R.; SAVIAN, T.V. Multivariate analysis of sensory characteristics of coffee grains (*Coffea arabica* L.) in the region of Upper Paranaíba. **Acta Scientiarum Agronomy**, v. 32, p. 635-641, 2010.

SPECIALITY COFFEE ASSOCIATION OF AMERICA. 2009. **SCAA Protocols - Cupping Specialty Coffee**. Long Beach: SCAA.2009. 7p.

SUGNET, G.; GOWER, J.C.; LE ROUX, N.J. A synthesis of canonical variate analysis, generalised canonical correlation and Procrustes analysis. **Computational Statistics & Data Analysis**, v.50 n.1, p.107-134, 2006.

ZÁRRAGA, A.; GOITISOLO, B. Simultaneous analysis and multiple factor analysis for contingency tables: Two methods for the joint study of contingency tables. **Computational Statistics & Data analysis**. v.53, n.8,p.3171-3182, 2009.

ARTIGO 2 - Identificação dos perfis dos consumidores de cafés através da análise de múltiplos fatores em dados qualitativos ordinais

1 INTRODUÇÃO

Segundo dados da Associação Brasileira da Indústria de Café - ABIC, os brasileiros consomem 4,87 kg de café torrado por habitante/ano (ABIC, 2014). Os efeitos benéficos, seja no âmbito social, do agronegócio ou da saúde, são destacados na mídia, visto como fator motivador para o consumo e a produção de cafés no Brasil. Decorrente desse fato, naturalmente sugere o foco no estudo da percepção dos consumidores, sendo essa uma importante ferramenta, no que tange à compreensão dos fatores envolvidos no processo de compra e na identificação dos diferentes segmentos do mercado de café. Segundo Noronha (2005), existem diferenças nas percepções individuais, resultantes de fatores genéticos, ou culturais, que influenciam os hábitos alimentares.

As percepções também podem ser influenciadas pelo processo cognitivo dos indivíduos, supostamente justificadas pelas diferenças de personalidades e interações do processamento em relação ao meio ambiente (COREN; WARD; ENNS, 1994).

Pesquisa realizada pela ABIC, em 2010, mostrou que o café é consumido por hábito, prazer, tradição, e que o sabor e o aroma são marcantes em sua qualidade.

Convém ressaltar que a importância do café, refletida na interação social, é mencionada por Arruda et al. (2009), ao relatar em que os consumidores associam o consumo do café ao hábito, prazer, família, amizade e ao trabalho, indicando que as motivações para o seu consumo têm um significado social associado aos hábitos e aos costumes brasileiros.

De acordo com Fehr et al (2012) o agronegócio café tem relevante importância social devido à sua capacidade de gerar empregos, além de atuar no desenvolvimento socioeconômico brasileiro. Os autores também ressaltam um aumento cada vez maior no consumo interno do café.

Alves, Casal e Oliveira (2009) apontam que os efeitos benéficos do consumo de café incluem a prevenção de doenças, tais como diabetes tipo II, asma, cirrose alcoólica, determinados tipos de cancro, doença de Parkinson e Alzheimer.

Evidentemente, o consumo de café dada a preferência de uma determinada marca, envolve a decisão do consumidor, relacionada a um processo cognitivo e afetivo (Orth, McGarry e Dodd, 2005). Para exemplificar melhor essa relação, observamos que, na maioria das situações, os consumidores, normalmente, escolhem características simples, nas quais baseiam suas percepções de marca, e adotam as marcas com personalidades que refletem a sua auto-imagem. Os consumidores avaliam melhor os produtos oriundos de seus países de origem e também os produtos e marcas proveniente de países mais desenvolvidos, em relação aos menos desenvolvidos (BOAS et al., 2009).

Saes e Spers (2006) notaram que os produtos diferenciados pela marca da torrefadora são mais atrativos quando comparados à região de origem ou ao processo de produção agrícola.

O consumidor de café no Brasil modificou seus hábitos de consumo e adquiriu novas percepções a respeito da bebida. Com isso, novas estratégias buscam a valorização do produto com atributos diferenciados, com aspectos tangíveis ou intangíveis, assim a concorrência não se dá apenas via preços, mas também por meio de produtos com características inovadoras (NICOLELI; MOLLER, 2006).

Segundo Spers, Saes e Souza (2004), o segmento de cafés diferenciados é o que mais cresce e que a compra do café está ligada aos atributos marca e sabor, que se relacionam às experiências passadas inerentes à memória sensorial que caracterizam cada consumidor, o que associa a fidelidade à marca ao sabor preferido. Portanto, há evidências de que os consumidores brasileiros de café estão preparados para cafés de qualidade, uma vez que existem diferenças entre as preferências dos consumidores, e que os segmentos de cafés deveriam ser atendidos por meio de estratégias de marketing que envolvam padrões de diferenciação, o que elevaria a qua-

lidade, de forma a agregar valor a satisfação do consumidor.

Arruda et al. (2009) afirmam que marca, qualidade e preço são os atributos mais observados na escolha do café, e a presença do selo de pureza, além de aumentar a credibilidade, valida a qualidade do produto.

Em relação às metodologias estatísticas empregadas nos setores de qualidade, nota-se uma carência quando os dados são de natureza qualitativa. Mormente, utilizam-se testes de hipóteses não paramétricos, sendo esses limitados a conclusões que tratam simultaneamente as relações entre as variáveis envolvidas na mensuração do processo de qualidade. Nesse contexto, em virtude do que foi mencionado, este trabalho tem por objetivo, propor uma análise estatística utilizando a técnica *multiple factor analysis* ou MFA, em uma pesquisa de percepção de consumidores de café, formada a partir de seis grupos de variáveis categóricas, relacionadas a uma pesquisa de opinião pública sobre a preferência e a percepção dos consumidores em relação à qualidade de café. Com esse propósito determinou-se as correlações entre esses grupos, permitindo identificar os atributos que levam à compra deste produto.

A análise de múltiplos fatores (MFA) se caracteriza por permitir analisar grupos de variáveis com tamanhos diferentes e de natureza distintas, que podem ser quantitativas ou categóricas, definidas no mesmo conjunto de observações (ESCOFIER; PAGÈS, 1982; ESCOFIER; PAGÈS, 1990, ESCOFIER; PAGÈS, 2008, ABDI; WILLIAMS; VALENTIN, 2013).

2 MATERIAL E MÉTODOS

2.1 Preparo e realização do experimento.

Tomaram-se por base os dados obtidos em uma pesquisa descritiva realizada na cidade de Varginha, MG, como o propósito de estudar a percepção sensorial do consumidor em relação à qualidade de café.

Diante do exposto, seguindo as especificações dadas por Saes e Spers (2006), consideraram-se as variáveis qualitativas em uma escala ordinal, de modo que as questões elaboradas foram estruturadas de forma a mensurar a importância atribuída pelo consumidor a diferentes marcas, conforme segue escrito na Tabela 12.

Tabela 12 Layout do questionário aplicado ao estudo sensorial dos consumidores de café referentes à cidade de Varginha, MG

| Natureza dos Grupos | Ref. | Questões | Descrição das Categorias |
|--|------|--|--|
| Escolaridade e Renda | A1 | Escolaridade | Analfabeto ou EF incompleto: 1; EF completo ou EM incompleto: 2; EM completo ou S incompleto: 3; S completo: 4; |
| | A2 | Renda Familiar | < R\$ 1.000,00: 1; [R\$ 1.000,00; R\$ 2.000,00[: 2; [R\$ 2.000,00; R\$ 4.000,00[: 3; [R\$ 4.000,00; R\$ 6.000,00[: 4; ≥ R\$ 6.000,00: 5; |
| Quantidade de xícaras de café consumidas | B1 | Quantidade de xícaras de café consumida em casa. | 1; 2; 3; 4; mais de 4; |
| | B2 | Quantidade de xícaras de café consumida no trabalho. | 1; 2; 3; 4; mais de 4; |
| Marca e Qualidade | C1 | Qual a importância você atribui a marca ao escolher o café? | Nenhuma: 1; Pouca: 2; Moderada: 3; Acentuada: 4; Muito Acentuada: 5; |
| | C2 | A marca do café interfere na qualidade do café. | 1; 2; 3; 4; 5; |
| Torrefadora e Produtor | D1 | A qualidade do café depende mais da firma torrefadora e não do produtor rural. | 1; 2; 3; 4; 5; |
| | D2 | O produtor de café interfere diretamente na qualidade do café. | 1; 2; 3; 4; 5; |
| Região e Produção | E1 | A região onde é produzido o café interfere na qualidade. | 1; 2; 3; 4; 5; |
| | E2 | Quando compro café, procuro informações sobre a região de produção. | 1; 2; 3; 4; 5; |
| Preferência e Preço | F1 | Prefiro cafés com a identificação da origem, mesmo que sejam mais caros. | 1; 2; 3; 4; 5; |
| | F2 | Prefiro cafés de qualidade superior, mesmo que sejam mais caros. | 1; 2; 3; 4; 5; |
| | F3 | Preço maior significa que a qualidade do café é superior. | 1; 2; 3; 4; mais de 4; |

Escolaridade - EF: Ensino Fundamental (1º grau); EM: Ensino Médio (2º grau); S: Superior
Fonte: Pereira (2014)

Com os dados obtidos, considerando uma amostra de $n = 50$ indivíduos, procedeu-se à construção das tabelas para cada grupo, sendo elas organizadas em uma estrutura conveniente de ser extraída os múltiplos fatores, denominada de tabela justapostas (Tabela 13).

Tabela 13 Layout da tabela justaposta para análise de cada grupo de variáveis.

| | Grupo 1 | | Grupo 2 | | Grupo 3 | | Grupo 4 | | Grupo 5 | | Grupo 6 | | |
|----------|-----------|----|-----------|----|-----------|----|-----------|----|-----------|----|-----------|----|----|
| Ind. | A1 | A2 | B1 | B2 | C1 | C2 | D1 | D2 | E1 | E2 | F1 | F2 | F3 |
| 1 | x_{ik1} | | x_{ik2} | | x_{ik3} | | x_{ik4} | | x_{ik5} | | x_{ik6} | | |
| 2 | | | | | | | | | | | | | |
| \vdots | | | | | | | | | | | | | |
| 50 | | | | | | | | | | | | | |

x_{ikG} = observação i na variável k do grupo G .

Grupo 1 - Escolaridade e renda

Grupo 2 - Quantidade de xícaras de café consumidas

Grupo 3 - Marca e qualidade

Grupo 4 - Torrefadora e produtor

Grupo 5 - Região e produção

Grupo 6 - Preferência e preço

Para cada grupo a notação é representada por $x_1, \dots, x_j, \dots, x_J$, que compreendem um único conjunto de dados, conforme ilustrado na Figura 10.

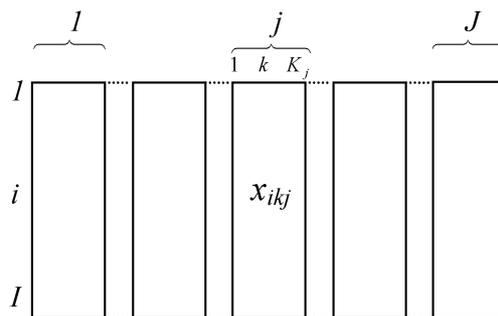


Figura 10 Layout dos grupos de variáveis justapostas, representando um único grupo.

As I observações foram representadas por J conjuntos de variáveis, o índice j referiu-se a um conjunto, o índice k referiu-se a uma coluna, e por fim K_j correspondeu ao número de colunas no conjunto j . De modo que $K = \sum_{j \in J} K_j$ indica número de colunas em todos os conjuntos.

Com essas especificações, uma tabela $I \times K_j$ foi associada com cada conjunto j . As tabelas J juntas formam uma tabela múltipla ou global $I \times K$. Para um j -ésimo conjunto, K_j representou o número de colunas e o número de variáveis, entretanto, considerando um conjunto categórico j , K_j indicou o número de colunas e o número de categorias. Uma tabela dessa forma para esse conjunto é formada, na qual o cruzamento da i -ésima linha com a k -ésima fixada na j -ésima tabela, implica nos valores x_{ikj} , sendo j um conjunto categórico, $x_{ikj} = 1$ se i pertence à categoria k e 0, caso contrário. Este tipo de conjunto é representado por uma tabela de variáveis indicadoras na qual a coluna k está associada com a categoria k (BÉCUE-BERTAUT; PAGÈS, 2008, PAGÈS, 2002, PAGÈS, 2008).

Contudo, as observações que não têm nenhuma categoria k , assumirão 0 para todas as variáveis indicadoras associadas à categoria k . (ESCOFIER; PAGÈS, 1994)

Seguindo essa notação, procedeu-se com a análise de múltiplos fatores, descritas a seguir.

2.2 Procedimento matricial do método análise de múltiplos fatores

Considerando os dados qualitativos centralizados por coluna representado por (28)

$$C_{ik} = x_{ik} - \bar{x}_k. \quad (28)$$

As colunas normalizadas conforme a expressão (29) obtida de (28);

$$x_{ik}^* = \frac{C_{ik}}{\sqrt{\sum_{i=1}^I (C_{ik})^2}}. \quad (29)$$

Para os dados aplicou-se uma Análise de Correspondência Múltipla, em cada grupo X_j de variáveis, formado por x_{ik}^* valores, com o propósito de determinar os primeiros autovalores λ_{j1} . Assim, tornou-se viável calcular $\varphi_k = a_k / \sqrt{\lambda_{j1}}$, sendo que a_k corresponde à proporção das observações que não possuem categoria k . Assim, segue que

$$Z_j = X_j \times \varphi_k \quad (30)$$

A matriz global será uma matriz $I \times K$ denotada por

$$Z = \begin{bmatrix} Z_1 & \cdots & Z_J \end{bmatrix} \quad (31)$$

Estimada a matriz global em Z , aplicou-se a decomposição de valores singulares (32).

$$Z = \hat{U} \hat{\Lambda} \hat{V}', \quad (32)$$

com $\hat{U}'\hat{U} = \hat{V}'\hat{V} = \mathbf{I}$, sendo \hat{U} e \hat{V} as matrizes de autovetores, $\hat{\Lambda} = \text{diag}(\sqrt{\lambda_i})$, e $\lambda_i > 0$ são os autovalores e \mathbf{I} é a matriz identidade. Logo os escores dos fatores globais foram estimados em (33).

$$F = M^{\frac{1}{2}} \hat{U} \hat{\Lambda}, \quad (33)$$

sendo M a matriz diagonal das massas das observações, com $m_i = I$, tendo I como o número de linhas da matriz Z . Em F , cada linha representa as observações e cada coluna, as variáveis.

O conjunto de dados Z_j , incorporado na matriz global Z , resulta na matriz de projeção (35) reescrevendo a equação F (33), conforme a expressão (34), sendo esta, utilizada na obtenção dos escores dos fatores globais.

$$F = M^{\frac{1}{2}} \hat{U} \hat{\Lambda} = (ZZ') \times \left(M^{\frac{1}{2}} \hat{U} \hat{\Lambda}^{-1} \right). \quad (34)$$

$$P = M^{\frac{1}{2}} \hat{U} \hat{\Lambda}^{-1}. \quad (35)$$

Uma vez que é a matriz de projeção P (35) que transforma a matriz ZZ' em escores dos fatores, a projeção dos grupos no espaço, representado pela matriz global Z , é realizada na obtenção dos escores em (36).

$$F_j = J \times Z_j Z_j' P, \quad (36)$$

As relações entre as variáveis e os escores globais foram analisadas calculando-se a inércia parcial de cada variável, por cada dimensão da matriz global Z . Dessa forma, para cada variável, como a soma das projeções quadradas das variáveis do vetor singular \hat{V} de Z multiplicado pelo autovalor correspondente (37) (ABDI; VALENTIN, 2007; ABDI; WILLIAMS, 2010, PAGÈS, 2002, PAGÈS, 2004).

$$R_{j,i} = \lambda_p \times \sum_j^{J_k} \hat{V}_{j,i}^2, \quad (37)$$

sendo que j refere-se ao grupo, p o componente principal, e λ_p é o autovalor da decomposição de valores singulares de Z .

Com base nas especificações matriciais, a obtenção dos resultados foi feita por meio do desenvolvimento de scripts no software R para uso público por meio do pacote MVar.pt versão 1.0.

3 RESULTADOS E DISCUSSÃO

Os resultados descritos na Tabela 14 evidenciaram que o estudo das associações entre as variáveis representadas em cada sub-tabela, analisadas simultaneamente, poderá ser explicado em um gráfico bidimensional, uma vez que, λ_i é a proporção da variação explicada nos dois primeiros componentes é igual a 17,14%. Entretanto, convém ressaltar que, embora esse percentual seja um número pouco expressivo, a qualidade de ajuste da técnica MFA prioriza a similaridade entre os grupos.

Tabela 14 Proporção da variação amostral, explicada pelos dois primeiros componentes.

| Componentes | λ_i | % da variação amostral | % da variação amostral acumulada |
|-------------|-------------|------------------------|----------------------------------|
| 1 | 2,6854 | 8,82 | 8,82 |
| 2 | 2,5327 | 8,32 | 17,14 |

As coordenadas obtidas que correspondem à posição global, isto é, considerando as respostas dos indivíduos a todos os sub-conjuntos de variáveis, estão representadas no mapa perceptual ilustrado na Figura 11, o qual permite interpretar, em termos práticos, que em relação às questões abordadas em diferentes sub-conjuntos, não é possível descrever um perfil que caracteriza os consumidores. Tal afirmação é perceptível pelo fato de a distribuição dos escores estar localizada próximo ao centroide, o que leva a crer que, dentre os fatores apontados por Malhotra (2001), que explicam o comportamento do consumidor, reportando os resultados obtidos, há evidências para afirmar que a distribuição dos escores em torno da média dar-se-á pelo fato de que localização, hábitos, mídia e reação às promoções e sensibilidade a preços não apresentaram um efeito que caracterize um perfil diferenciado em relação ao comportamento do consumidor, mediante a aquisição do produto.

Em se tratando da influência que cada grupo de variáveis exerceu em relação às respostas dos consumidores (Figura 11), os resultados ilustrados na Figura 12 indicam que as linhas que unem os grupos com os escores

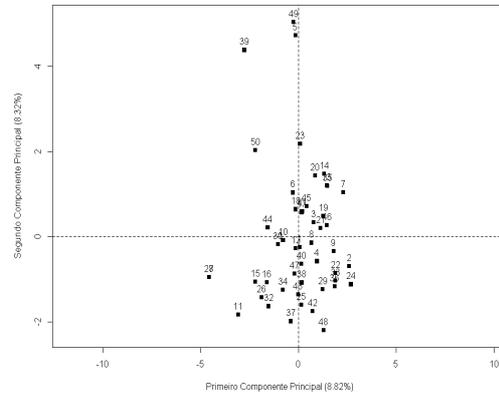


Figura 11 Gráfico da correspondente à análise global dos indivíduos.

provenientes da análise global foram inconclusivas ao destacar algum conjunto de variáveis relevantes a respostas dos consumidores, uma vez que estes se encontram aglomerados próximo ao centroide, não revelando nenhum grupo de consumidores distinto em relação à amostra.

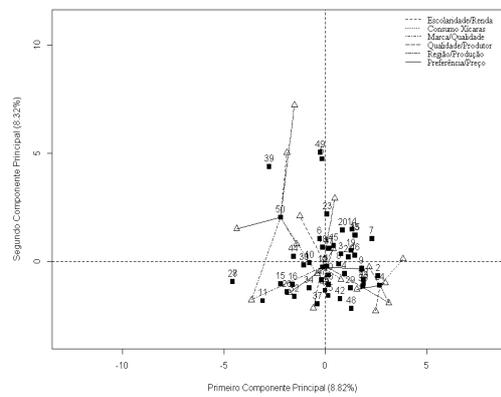


Figura 12 Gráfico das análises dos grupos de indivíduos.

Moori, Bido e Oliveira (2011), em um estudo para detectar as variáveis relevantes do consumidor do café solúvel por meio da modelagem

de equações estruturais, as variáveis que caracterizam a classe social, apresentam um pequeno efeito do consumo de café. Entretanto, a idade do respondente apresentou um efeito significativo. Contudo, esse estudo evidenciou que, no mercado interno brasileiro, o principal motivo que leva as pessoas a consumirem o café solúvel, está associado ao hábito. Dessa forma, justifica-se a baixa contribuição do grupo de variáveis utilizadas neste trabalho, em relação à determinação do perfil do consumidor da cidade de Varginha, MG.

Nota-se, pela magnitude dos coeficientes escritos na Tabela 15, que os grupos (Marca e qualidade) e (Preferência e preço) apresentaram maior contribuição para a formação dos componentes. Tal resultado permite afirmar que esses conjuntos de variáveis caracterizam os componentes que explicam a variabilidade total em duas dimensões.

Tabela 15 Valores referentes às inércias computadas nos dois primeiros componentes, para cada grupo de variáveis.

| Comp. | Grupo de variáveis | | | | | | Inércia total |
|-------|--------------------|--------|--------|--------|--------|--------|---------------|
| | 1 | 2 | 3 | 4 | 5 | 6 | |
| 1 | 0,2493 | 0,4917 | 0,6076 | 0,3681 | 0,3119 | 0,6569 | 2,6854 |
| 2 | 0,4560 | 0,4133 | 0,4972 | 0,1617 | 0,4367 | 0,5677 | 2,5327 |

Convém ressaltar que, embora essas dimensões tenham apresentado baixas porcentagens de explicação da variabilidade amostral, dado que os sub-conjuntos de variáveis (Marca e qualidade) e (Preferência e preço) foram mais marcantes, é sabido que esses conjuntos de variáveis são voláteis no tempo e dependentes da conjectura econômica e ou regional. Além desses fatores, as conclusões são limitadas ao tamanho amostral. Entretanto, por se tratar de uma pesquisa empírica, os resultados obtidos são apreciados, no contexto de direcionar estratégias metodológicas de marketing e segmentação de mercado.

4 CONCLUSÕES

A similaridade entre os grupos (Marca e qualidade) e (Preferência e preço) sugere que a importância atribuída à marca do café interfere na percepção de sua qualidade, e ainda, que a sua origem está associada a uma qualidade superior, com esses fatores interagindo na preferência por produtos de preços superiores, mostrando, com isso, a viabilidade da técnica análise de múltiplos fatores na identificação dos perfis de grupos de consumidores com diferentes características qualitativas.

REFERÊNCIAS

ABDI, H.; VALENTIN, D.; Multiple factor analysis (MFA). In: Salkind NJ, ed. **Encyclopedia of Measurement and Statistics**. Thousand Oaks, CA: Sage; 2007, 657-663.

ABDI, H.; WILLIAMS, L.; Principal Component Analysis. **WIREs Comp Stat** 2010 2 433-459 DOI: 10.1002/wics.101

ABDI, H.; WILLIAMS, L.; VALENTIN, D.; Multiple factor analysis: principal component analysis for multitable and multiblock data sets. **WIREs Comput Stat** 2013. doi: 10.1002/wics.1246

ALVES, R.C.; CASAL, S.; OLIVEIRA, B.. Benefícios do café na saúde: mito ou realidade?. **Química Nova**, São Paulo, v. 32, n. 8, 2009.

ARRUDA, A.C. et al.; Justificativas e motivações do consumo e não consumo de café. **Ciência e Tecnologia de Alimentos**, Campinas, v. 29, n.4, p. 754-763, out.-dez. 2009.

ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA DE CAFÉ - ABIC.
Tendências do Consumo de Café no Brasil em 2010. Tendências de consumo de café - VIII - 2010. Disponível em:
<http://www.abic.com.br/publique/media/EST_PESQTendenciasConsumo2010.pdf>. Acesso em: Dezembro 2014.

ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA DE CAFÉ - ABIC.
Estatísticas – indicadores da indústria de café no Brasil - 2013. Disponível em: <<http://www.abic.com.br/publique/cgi/cgilua.exe/sys/start.htm?sid=61#1910>>. Acesso em: Dezembro 2014.

ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA DE CAFÉ - ABIC.
Pesquisa Nacional Fiesp/IBOPE sobre o Perfil do Consumo de Alimentos no Brasil. Disponível em:
<http://www.abic.com.br/media/EST_PESQFoodTrends1.pdf>. Acesso em: Dezembro 2014.

BÉCUE-BERTAUT, M.; PAGÈS, J.; Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data.

Computational Statistics & Data Analysis, v. 52, p. 3255-3268, 2008.

BOAS, L.H. et al.; Marcas regionais de café em Uberlândia: a percepção de valor dos consumidores e a relação com o “ser regional”. in: 47º Congresso Sociedade Brasileira de Economia, Administração e Sociologia Rural (SOBER), 47., 2009, Porto Alegre. **Anais...Porto Alegre: UFRGS**, 2009.

COREN, S.; WARD, L. M.; ENNS, J. T. **Sensation and perception**. Orlando: Harcourt Brace College Publ., 1994. 747p.

ESCOFIER, B.; PAGÈS, J.; **Analyse Factorielles Simples et Multiples**. Paris: Dunod., 1990.

ESCOFIER, B.; PAGÈS, J.; **Analyses factorielles simples et multiples: objectifs, méthodes et interprétation**. 4th. ed. Paris: Dunod, 2008. 318 p.

ESCOFIER, B.; PAGÈS, J.; Comparaison de groupes de variables définies sur le même ensemble d'individus: un exemple d'applications. Working paper 165. **Institut National de Recherche en Informatique et en Automatique**, Le Chesnay, France. 1982

ESCOFIER, B.; PAGÈS, J.; Multiple Factor Analysis (AFUMULT package) **Computational Statistics & Data Analysis**, v. 18, p. 121-140, 1994.

FEHR, L.C.F. de A; DUARTE, S.L; TAVARES, M; REIS, E.A. dos., Análise temporal das variáveis de custos da cultura do café arábica nas principais regiões produtoras do Brasil. **Custos e Agronegócio Online**, v. 8, n. 1 – Jan/Mar - 2012.

MALHOTRA, N. K. **Pesquisa de Marketing: uma orientação aplicada**. 3. ed. Porto Alegre: Bookman, 2001. p. 720.

MOORI, R. G. ; BIDO, D. S. ; OLIVEIRA, L. H. de . **Variáveis Relevantes do Consumidor do Café Solúvel sob o Enfoque da Diferenciação**. **Organizações Rurais e Agroindustriais (UFLA)** , v. 13, p. 124-138, 2011.

NICOLELI, M.; MOLLER, H. D., Análise da competitividade dos custos do café orgânico sombreado irrigado. **Custos e Agronegócio Online**, v. 2, n. 1 - Jan/Jun - 2006.

NORONHA, R. L. F.; DELIZA, R.; SILVA, M. A. A. P.; A expectativa do consumidor e seus feitos na avaliação sensorial e aceitação de produtos alimentícios. **Alimentos e Nutrição**, Araraquara, v.16, n.3, p. 299-308, jul./set. 2005.

ORTH, ULRICH R.; MCGARRY, MARIANNE; DODD, WOLF TIM H. Dimensions of wine region equity and their impact on consumer preferences. **Journal of Product & Brand Management** 14/2 88-97, 2005.

OSSANI, P.C.; CIRILLO, M.A.; **MVar.pt**: análise Multivariada (Brazilian Portuguese). URL <http://CRAN.R-project.org/package=MVar.pt>. R package version 1.0. 2014.

PAGÈS, J. Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. **Revue de Statistique Appliquée**. 50 (4), 5-37, 2002.

PAGÈS, J.; Multiple Factor Analysis: main features and application to sensory data. **Revista Colombiana de Estadística**. vol. 27 n° 1, p. 1-26, 2004.

PEREIRA, G.A.; **Estimadores de Ridge Generalizados adaptados em modelos de equações estruturais: Estudo de simulação e aplicação no perfil de consumidores de café**. 2014. 80 p. Tese (Doutorado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2014.

SAES, M.S.M.; SPERS, E.E. Percepção do consumidor sobre os atributos de diferenciação no segmento rural: café no mercado interno. **Organizações Rurais & Agroindustriais**. Lavras, v.8, n.3, p.354-367, nov. 2006.

SPERS, E.E.; SAES, M.S.M.; SOUZA, M.C.M.. Análise das preferências do consumidor brasileiro de café: um estudo exploratório dos mercados de São Paulo e Belo Horizonte. **RAUSP - Revista de Administração da Universidade de São Paulo**, São Paulo, v. 39, n. 1, p. 53-61, 2004.

APÊNDICE

No intuito de deixar este trabalho o mais didático possível, apresentam-se, neste apêndice, exemplos da utilização do método MFA para dados quantitativos, categóricos, de frequência e dados mistos. Espera-se que os exemplos estejam bem compreensíveis para o leitor. Ainda, foram criados códigos computacionais para a execução dos cálculos, por meio de um pacote chamado de MVar.pt (o ".pt" significa que o pacote está em português), usado no software R versão 3.1.2 e RStudio versão 0.98.1091, de modo que todos possam ter acesso aos códigos fonte, o que facilitará a verificação dos cálculos apresentados nos artigos e exemplos dados. Assim, para ter acesso ao código fonte, basta instalar o pacote no software R, ou baixá-lo da internet, usando o endereço <http://CRAN.R-project.org/package=MVar.pt>. Foi tomado o cuidado de deixar os códigos bem legíveis e simples, com comentários onde fosse pertinente, para facilitar a compreensão do leitor que queira utilizá-los. Assim, um programador mais experiente perceberá que muitas linhas de comandos poderiam ser simplificadas, mas isto levaria a uma maior complexidade na compreensão do código, o que se evitou. Ao acessar este pacote, o leitor poderá verificar que as bases de dados utilizadas neste trabalho estão disponíveis e foram utilizadas nos exemplos. Neste pacote existem muitas funções úteis na análise multivariada e, certamente, outras funções importantes serão acrescentadas ao longo do tempo.

A) Exemplo aplicado do método MFACT

Para ilustrar o método MFACT, utilizou-se a Tabela 16, com dados simulados sobre o número de homens e mulheres que consomem café, nas cidades de Paris, Nova Iorque e Tóquio. O objetivo, foi obter a topologia do sexo e de idade das pessoas e saber se existem diferenças entre os grupos de café consumidos.

Embora o método MFACT não exija que se trabalhe com a frequência relativa, para fins didáticos, trabalhar-se-á com a mesma (Tabela 17).

Tabela 16 Dados sobre o número de pessoas que consomem cafés.

| Sexo e idade | Café Bourbon | | | Café Acaiá | | Café Catuaí | | |
|-----------------|--------------|----|--------|------------|--------|-------------|----|--------|
| | Paris | NI | Tóquio | NI | Tóquio | Paris | NI | Tóquio |
| Homens < 30 | 14 | 12 | 2 | 14 | 17 | 11 | 11 | 5 |
| Homens 30-50 | 7 | 13 | 1 | 12 | 4 | 11 | 17 | 8 |
| Homens > 50 | 1 | 5 | 0 | 4 | 2 | 32 | 24 | 17 |
| Mulheres 30 | 8 | 13 | 7 | 4 | 9 | 22 | 26 | 5 |
| Mulheres 30-50 | 3 | 14 | 1 | 1 | 8 | 32 | 25 | 22 |
| Mulheres > 50 | 0 | 7 | 0 | 3 | 2 | 47 | 20 | 38 |

Fonte: Dados extraídos de Greenacre e Blasius (2006, pg. 318) e adaptados a Cafés.

Agora, usando a equação (12), segue que a matriz da análise de correspondência interna (ICA) da tabela de frequência relativa de cada grupo de variáveis são dadas por:

$$\begin{aligned}
 X_1 &= \begin{bmatrix} 1.0762 & -0.4681 & -0.5052 \\ 0.1358 & 0.0667 & -0.7957 \\ \vdots & \vdots & \vdots \\ -0.3108 & 0.2137 & -0.3108 \end{bmatrix}, \\
 X_2 &= \begin{bmatrix} -0.1245 & 0.1126 \\ 0.8898 & -0.8051 \\ \vdots & \vdots \\ 0.0789 & -0.0714 \end{bmatrix} \text{ e} \\
 X_3 &= \begin{bmatrix} -0.0093 & 0.1112 & -0.1289 \\ -0.1963 & 0.3204 & -0.0946 \\ \vdots & \vdots & \vdots \\ 0.1042 & -0.5701 & 0.5682 \end{bmatrix}.
 \end{aligned}$$

Agora utilizando-se a decomposição de valor singular generalizada, dada na seção (5.2.3), nas X_t tabelas, lembrando que o peso das linhas e

Tabela 17 Tabela de frequência relativa dos dados.

| Café Bourbon | | | Total | Café Acaiá | | Total | ... |
|--------------|--------|--------|--------|------------|--------|--------|-----|
| Paris | NI | Tóquio | Grupo | NI | Tóquio | Grupo | |
| 0.0250 | 0.0214 | 0.0036 | 0,0499 | 0.0250 | 0.0303 | 0,0553 | ... |
| 0.0125 | 0.0232 | 0.0018 | 0,0374 | 0.0214 | 0.0071 | 0,0285 | |
| 0.0018 | 0.0089 | 0.0000 | 0,0107 | 0.0071 | 0.0036 | 0,0107 | |
| 0.0143 | 0.0232 | 0.0125 | 0,0499 | 0.0071 | 0.0160 | 0,0232 | |
| 0.0053 | 0.0250 | 0.0018 | 0,0321 | 0.0018 | 0.0143 | 0,0160 | |
| 0.0000 | 0.0125 | 0.0000 | 0,0125 | 0.0053 | 0.0036 | 0,0089 | |
| 0,0588 | 0,1141 | 0,0196 | 0,1925 | 0,0677 | 0,0749 | 0,1426 | ... |

| Café Catuaí | | | Total | Total |
|-------------|--------|--------|--------|--------|
| Paris | NI | Tóquio | Grupo | Geral |
| 0.0196 | 0.0196 | 0.0089 | 0,0481 | 0,1533 |
| 0.0196 | 0.0303 | 0.0143 | 0,0642 | 0,1301 |
| 0.0570 | 0.0428 | 0.0303 | 0,1301 | 0,1515 |
| 0.0392 | 0.0463 | 0.0089 | 0,0945 | 0,1676 |
| 0.0570 | 0.0446 | 0.0392 | 0,1408 | 0,1889 |
| 0.0838 | 0.0357 | 0.0677 | 0,1872 | 0,2086 |
| 0,2763 | 0,2193 | 0,1693 | 0,6649 | 1,0000 |

das colunas é dado por

$$\begin{aligned}
 M &= \begin{bmatrix} 0.1533 & 0.1301 & 0.1515 & 0.1676 & 0.1889 & 0.2086 \end{bmatrix}', \\
 C_1 &= \begin{bmatrix} 0.0588 & 0.1141 & 0.0196 \end{bmatrix}, C_2 = \begin{bmatrix} 0.0677 & 0.0749 \end{bmatrix} \text{ e} \\
 C_3 &= \begin{bmatrix} 0.2763 & 0.2193 & 0.1693 \end{bmatrix},
 \end{aligned}$$

que são obtidos, respectivamente, pelo total geral das linhas e o total das colunas da Tabela 17. Agora, fazendo

$$Q_1 = \text{diag}(\sqrt{M}) \times X_1 \times \text{diag}(\sqrt{C_1}),$$

em que *diag* corresponde à matriz com os valores na diagonal, e aplicando a decomposição de valor singular usual em Q_1 , tem-se o seguinte primeiro autovalor $\lambda_{x_1} = 0.0239$ e, de modo análogo, para X_2 e X_3 , tem-se $\lambda_{x_2} =$

0.0221 e, $\lambda_{x_3} = 0.0450$, respectivamente. A matriz global será uma matriz $I \times J$, denotada por

$$Z = \begin{bmatrix} X_1 & X_2 & X_2 \end{bmatrix} = \left[\begin{array}{ccc|ccc} 1.0762 & \cdots & \cdots & 0.1126 & \cdots & -0.1289 \\ 0.1358 & \cdots & \cdots & -0.8051 & \cdots & -0.0946 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -0.3108 & \cdots & \cdots & -0.0714 & \cdots & 0.5682 \end{array} \right]$$

Agora segue à análise global em Z . Usando a NI, sabe-se que $Z = \tilde{U}\tilde{\Lambda}\tilde{V}'$, com $\tilde{U}'M\tilde{U} = \tilde{V}'W\tilde{V} = \mathbf{I}$, sendo \tilde{U} e \tilde{V} as matrizes de autovetores, $\tilde{\Lambda} = \text{diag}(\sqrt{\lambda_i})$ é a matriz de valores singulares e \mathbf{I} é a matriz identidade, lembrando que M e W são as restrições das linhas e colunas, respectivamente. O peso da coluna é obtido dividindo-se o peso das colunas pelo respectivo primeiro autovalor de cada grupo. Assim, tem-se que

$$\begin{aligned} W &= \begin{bmatrix} \frac{0.0588}{0.0239} & \frac{0.1141}{0.0239} & \frac{0.0196}{0.0239} & \frac{0.0677}{0.0221} & \frac{0.0749}{0.0221} & \cdots & \frac{0.1693}{0.0450} \end{bmatrix} \\ &= \begin{bmatrix} 2.4655 & 4.7816 & 0.8218 & 3.0592 & 3.3812 & \cdots & 3.7627 \end{bmatrix}. \end{aligned}$$

Novamente fazendo

$$Q = \text{diag}(\sqrt{M}) \times Z \times \text{diag}(\sqrt{W}),$$

em que diag corresponde à matriz com os valores na diagonal. Agora, aplicando a decomposição de valor singular usual em Q , tem-se

$$\tilde{U} = M^{-\frac{1}{2}}P = \begin{bmatrix} 0.5229 & \cdots & -0.0323 \\ -0.6405 & \cdots & 1.0572 \\ \vdots & \ddots & \vdots \\ -1.1206 & \cdots & 0.2775 \end{bmatrix},$$

$$\tilde{V} = W^{-\frac{1}{2}}Q = \begin{bmatrix} 0.1031 & \cdots & 0.0133 \\ -0.1633 & \cdots & -0.0358 \\ \vdots & \ddots & \vdots \\ -0.2287 & \cdots & 0.3154 \end{bmatrix}.$$

e as variâncias são dadas por

$$\text{diag}(\tilde{\Lambda}^2) = \begin{bmatrix} 1.7828 & 1.1944 & 0.6206 & 0.2547 & 0.0276 \end{bmatrix}.$$

Os escores dos fatores globais são dados por

$$F = \tilde{U}\tilde{\Lambda} = \begin{bmatrix} 0.6981 & \cdots & -0.0054 \\ -0.8552 & \cdots & 0.1757 \\ \vdots & \ddots & \vdots \\ -1.4962 & \cdots & 0.0461 \end{bmatrix}.$$

Em F , cada linha representa um indivíduo (homem ou mulher) e cada coluna um componente, o que gera a Figura 13. Pode-se interpretar o primeiro componente como sendo que as mulheres com 30 anos e os homens menos de 30 anos forma uma oposição aos outros indivíduos. Ainda em relação ao segundo componente, pode-se interpretar que os homens e mulheres são diferentes. Agora, usando a equação (1) para encontrar a proporção da variância total explicada pelos componentes principais, segue a Tabela 18.

Tabela 18 Proporção explicada pelas variâncias.

| Componentes | λ_i | % explicação | % explicação acumulada |
|-------------|-------------|--------------|------------------------|
| 1 | 1.7828 | 45.95 | 45.95 |
| 2 | 1.1944 | 30.78 | 76.73 |
| 3 | 0.6206 | 15.99 | 92.72 |
| 4 | 0.2547 | 6.56 | 99.29 |
| 5 | 0.0276 | 0.71 | 100.00 |

Análise parcial

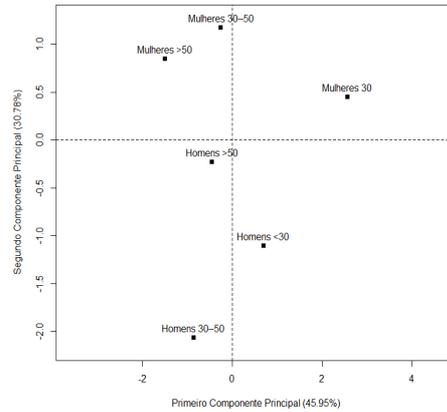


Figura 13 Gráfico da análise global dos homens e mulheres.

A análise global revela a estrutura comum do espaço dos indivíduos (homens e mulheres), para ver como cada café interpreta este espaço. Isto é alcançado projetando-se o conjunto de dados de cada variável café sobre a análise global. Isto é implementado utilizando-se

$$F_t = T \times X_t \times \text{diag}(W_t) \times \tilde{V}_t,$$

em que T é o número de tabelas, W_t é o peso das colunas do grupo t em W , e \tilde{V}_t são as linhas correspondentes ao grupo t em \tilde{V} . Assim, os cálculos dos escores do primeiro grupo da variável café, tem-se que $T = 3$,

$$W_1 = \begin{bmatrix} 2.4655 & 4.7816 & 0.8218 \end{bmatrix} \text{ e } \tilde{V}_1 = \begin{bmatrix} 0.1031 & \cdots & 0.0133 \\ -0.1633 & \cdots & -0.0358 \\ 0.6412 & \cdots & 0.1684 \end{bmatrix},$$

logo;

$$F_1 = T \times X_1 \times \text{diag}(W_1) \times \tilde{V}_1 = \begin{bmatrix} 1.1185 & \cdots & 0.1367 \\ -1.3106 & \cdots & -0.3513 \\ \vdots & \ddots & \vdots \\ -1.2288 & \cdots & -0.2695 \end{bmatrix},$$

$$F_2 = T \times X_2 \times \text{diag}(W_2) \times \tilde{V}_2 = \begin{bmatrix} 0.3369 & \cdots & 0.0573 \\ -2.4090 & \cdots & -0.4096 \\ \vdots & \ddots & \vdots \\ -0.2135 & \cdots & -0.0363 \end{bmatrix},$$

$$F_3 = T \times X_3 \times \text{diag}(W_3) \times \tilde{V}_3 = \begin{bmatrix} 0.6389 & \cdots & -0.2101 \\ 1.1540 & \cdots & 1.2882 \\ \vdots & \ddots & \vdots \\ -3.0463 & \cdots & 0.4442 \end{bmatrix}.$$

A Figura 14 mostra os dois primeiros componentes principais da análise global, juntamente com as projeções de homens e mulheres para os cafés. Note que a posição de cada homem ou mulher na análise global é o baricentro (isto é, centroide) de suas posições para os cafés. Para facilitar a interpretação, puxaram-se linhas unindo os Cafés com a posição global dos homens e mulheres. Isto mostra uma relativa semelhança entre as três representações parciais de cada grupo. Infelizmente, não sugere muita coisa. A priori, pode-se concluir que não existe diferença entre os grupos de variáveis.

Correlações entre os componentes principais e as variáveis originais

Para obter as correlações entre os componentes principais e as variáveis originais basta apenas utilizar a equação

$$\rho_{Y,X} = \tilde{V} \times \tilde{\Lambda},$$

o que resulta na Tabela 19.

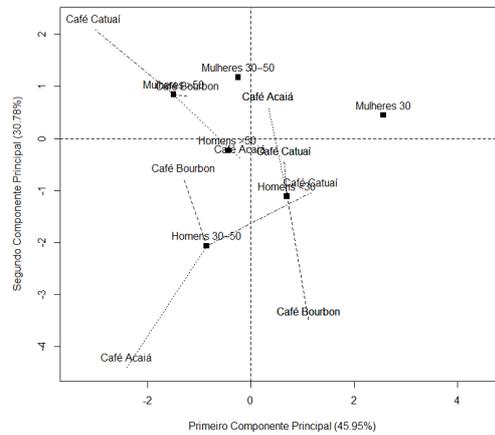


Figura 14 Gráfico dos Cafés com a Análise Global dos Homens e Mulheres.

Verifica-se que a cidade de Tóquio, do grupo café Bourbon, possui o maior coeficiente de correlação (0.8561) com o componente principal de maior explicação e, em geral, correlaciona-se de forma fraca com as outras variáveis. As variáveis do segundo componente principal correlacionam-se de forma fraca e inversa. Segue a Figura 15.

Inércia parcial entre os cafés

As relações entre os grupos de cafés e a solução global são analisadas calculando-se a inércia parcial de cada café por cada dimensão da análise global. Isto é calculado, do seguinte modo

$$G = \tilde{V} \times \text{diag}(\sqrt{\tilde{\Lambda}}) = \begin{bmatrix} 0.1191 & \cdots & 0.0054 \\ -0.1887 & \cdots & -0.0146 \\ \vdots & \ddots & \vdots \\ -0.2642 & \cdots & 0.1286 \end{bmatrix},$$

Tabela 19 Correlações entre os componentes principais (Eixos) e as variáveis originais.

| | Variáveis | Eixos 1 | Eixos 2 | Eixos 3 |
|--------------|-----------|---------|---------|---------|
| Café Bourbon | Paris | 0.1376 | -0.3338 | -0.3211 |
| | NI | -0.2181 | 0.1191 | 0.1080 |
| | Tóquio | 0.8561 | 0.3088 | 0.3345 |
| Café Acaiá | NI | -0.2068 | -0.3089 | 0.1664 |
| | Tóquio | 0.1871 | 0.2795 | -0.1505 |
| Café Catuaí | Paris | -0.0116 | 0.0585 | -0.0236 |
| | NI | 0.2505 | -0.1690 | 0.0847 |
| | Tóquio | -0.3053 | 0.1234 | -0.0711 |

$$H = (G_{ij})^2 \times \text{diag} \left(\frac{1}{\Lambda} \right) = \begin{bmatrix} 0.0106 & \cdots & 0.0002 \\ 0.0267 & \cdots & 0.0013 \\ \vdots & \ddots & \vdots \\ 0.0523 & \cdots & 0.0995 \end{bmatrix},$$

$$U = w_i \times H_i = \begin{bmatrix} 2.4655 \times \begin{bmatrix} 0.0106 & \cdots & 0.0002 \\ 0.0267 & \cdots & 0.0013 \\ \vdots & & \end{bmatrix} \\ 4.7816 \times \begin{bmatrix} 0.0106 & \cdots & 0.0002 \\ 0.0267 & \cdots & 0.0013 \\ \vdots & & \end{bmatrix} \\ 3.7627 \times \begin{bmatrix} 0.0106 & \cdots & 0.0002 \\ 0.0267 & \cdots & 0.0013 \\ \vdots & & \end{bmatrix} \\ 0.0523 & \cdots & 0.0995 \end{bmatrix} = \begin{bmatrix} 0.0262 & \cdots & 0.0004 \\ 0.1276 & \cdots & 0.0061 \\ \vdots & \ddots & \vdots \\ 0.1968 & \cdots & 0.3744 \end{bmatrix}.$$

As relações entre os grupos de cafés e a solução global são analisadas calculando-se a inércia parcial de cada grupo por cada dimensão da análise global, isto é, usando a equação

$$R_{j,i} = \lambda_i \times \sum_j^{J_k} U_{j,i},$$

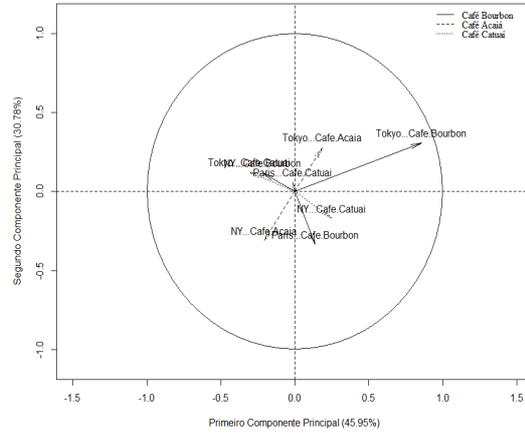


Figura 15 Circulo das correlações entre os componentes principais (eixos) e as variáveis originais.

obtém-se

$$R_{1,1} = \lambda_1 \times \sum_j^{J_k} U_{1,1} = 1.7828 \times (0.0262 + 0.1276 + 0.3379) = 0.8765$$

$$R_{1,2} = \lambda_1 \times \sum_j^{J_k} U_{1,2} = 1.7828 \times (0.0734 + 0.0664) = 0.2491$$

⋮

$$R_{3,5} = \lambda_5 \times \sum_j^{J_k} U_{3,5} = 0.0276 \times (0.5185 + 0.0732 + 0.3744) = 0.0276.$$

Cálculos semelhantes dão os valores informados na Tabela 20.

A partir das inércias obtidas em cada café (Tabela 20), visando uma melhor interpretação, o gráfico das inércias (Figura 16) é gerado, mostrando que há diferenças entre os grupos de variáveis.

Tabela 20 Inércia dos cafés.

| | Eixo 1 | Eixo 2 | Eixo 3 | Eixo 4 | Eixo 5 |
|--------------|-------------|-------------|-------------|-------------|-------------|
| Café Bourbon | 0.8765 | 0.4209 | 0.4019 | 0.1219 | 0.0008 |
| Café Acaiá | 0.2491 | 0.5562 | 0.1613 | 0.0333 | 0.0001 |
| Café Catuai | 0.6572 | 0.2173 | 0.0574 | 0.0996 | 0.0267 |
| Σ | 1.7828 | 1.1944 | 0.6206 | 0.2547 | 0.0276 |
| | λ_1 | λ_2 | λ_3 | λ_4 | λ_5 |

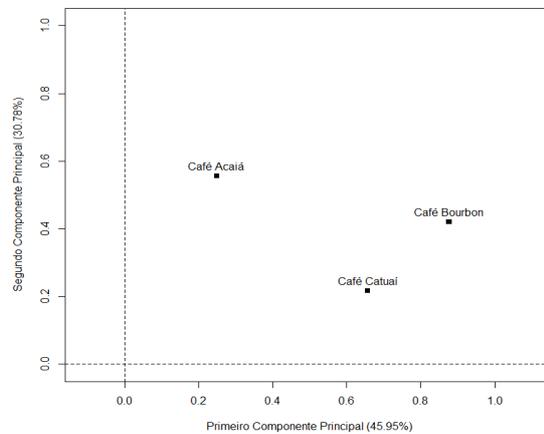


Figura 16 Gráfico das inércias dos cafés.

B) Exemplo aplicado do método MFA para dados categóricos

Para ilustrar o método MFA para dados categóricos, tomemos a tabela abaixo com dados simulados de dois grupos de degustadores de cafés sendo que cada grupo foi constituído de seis homens e seis mulheres, onde cada Grupo possui características próprias (Tabela 21).

Tem-se por objetivo geral obter a topologia dos homens e mulheres e saber se existe diferença entre os grupos de variáveis (características que definem cada grupo).

Como estamos trabalhando com variáveis categóricas, é necessário converter os dados em variáveis binárias. Logo, segue a matriz indicadora dos dados (Tabela 22).

Tabela 21 Tabela de dados.

| Sexo | Grupo 1 | | Grupo 2 | | |
|----------|------------|---------|--------------|------------|-----------|
| | Idade | Fumante | Estado civil | Esportista | Estudante |
| Homem 1 | Jovem | Não | Casado | Sim | Não |
| Homem 2 | Meia-idade | Não | Solteiro | Não | Não |
| Homem 3 | Velho | Não | Solteiro | Não | Não |
| Homem 4 | Jovem | Sim | Casado | Sim | Sim |
| Homem 5 | Velho | Sim | Solteiro | Não | Não |
| Homem 6 | Meia-idade | Sim | Casado | Não | Sim |
| Mulher 1 | Velho | Sim | Casado | Não | Não |
| Mulher 2 | Velho | Sim | Casado | Não | Sim |
| Mulher 3 | Jovem | Não | Solteiro | Não | Não |
| Mulher 4 | Meia-idade | Não | Solteiro | Sim | Não |
| Mulher 5 | Jovem | Não | Solteiro | Sim | Não |
| Mulher 6 | Meia-idade | Sim | Casado | Sim | Sim |

Chamemos de P_1 e P_2 as matrizes indicadoras de cada grupo. Assim, tem-se que $T = 2$ tabelas:

$$P_1 = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} \quad P_2 = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Usando a análise de correspondência usual, deve-se encontrar os respectivos autovalores de P_1 e P_2 , que são dados por $\lambda_{P_1} = 0.7041241$ e $\lambda_{P_2} = 0.5872716$, que serão usados mais adiante.

Inicialmente, deve-se equilibrar a influência dos grupos de variáveis em cada grupo P_j , para que possamos executar posteriormente uma análise

Tabela 22 Matriz indicadora dos dados categóricos.

| Idade | | | Fumante | | Estado civil | | ... |
|-------|------------|-------|---------|-----|--------------|----------|-----|
| Jovem | Meia-idade | Velho | Não | Sim | Casado | Solteiro | ... |
| 1 | 0 | 0 | 1 | 0 | 1 | 0 | |
| 0 | 1 | 0 | 1 | 0 | 0 | 1 | |
| 0 | 0 | 1 | 1 | 0 | 0 | 1 | |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ... |
| 0 | 1 | 0 | 0 | 1 | 1 | 0 | |

| Esportista | | Estudante | |
|------------|-----|-----------|-----|
| Não | Sim | Não | Sim |
| 0 | 1 | 1 | 0 |
| 1 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 0 | 1 | 0 | 1 |

global. Assim para melhor compreensão, tomemos P_1 e façamos

$$\begin{aligned}
 C_1 &= mP_1 \\
 &= \begin{bmatrix} \frac{1}{12} & 0.00 & 0.00 & \frac{1}{12} & 0.00 \\ 0.00 & \frac{1}{12} & 0.00 & \frac{1}{12} & 0.00 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.00 & \frac{1}{12} & 0.00 & 0.00 & \frac{1}{12} \end{bmatrix}
 \end{aligned}$$

sendo m a massa das observações, ou seja, $m = \frac{1}{12}$. Cálculo semelhante é feito em P_2 . Em seguida, centralizam-se os dados na média por coluna. Segue que as médias de cada coluna são dadas por $\bar{C}_1 = [\frac{1}{36}, \frac{1}{36}, \frac{1}{36}, \frac{1}{24}, \frac{1}{24}]$.

Assim, centralizando-se os dados, tem-se

$$D_1 = C_1 - \bar{C}_1 = \begin{bmatrix} 0.0556 & -0.0278 & \cdots & -0.0417 \\ -0.0278 & 0.0556 & \cdots & -0.0417 \\ \vdots & \vdots & \ddots & \vdots \\ -0.0278 & 0.0556 & \cdots & 0.0417 \end{bmatrix}.$$

Agora divide-se D_1 pela raiz quadrada da soma ao quadrado de cada elemento das observações da tabela t , ou seja

$$S_t = \sum_{i=1}^I \sum_{j=1}^J \sqrt{x_{ijt}^2}.$$

Assim, para a $t = 1$, tem-se que

$$S_1 = \left[\sqrt{0.0185} \quad \sqrt{0.0185} \quad \sqrt{0.0185} \quad \sqrt{0.0208} \quad \sqrt{0.0208} \right].$$

Segue que a matriz balanceada é dada por

$$X_1 = D_1 \times \text{diag} \left(\frac{1}{S_1} \right) = \begin{bmatrix} 0.4082 & -0.2041 & \cdots & -0.2887 \\ -0.2041 & 0.4082 & \cdots & -0.2887 \\ \vdots & \vdots & \ddots & \vdots \\ -0.2041 & 0.4082 & \cdots & 0.2887 \end{bmatrix}.$$

Do mesmo modo, tem-se que

$$X_2 = D_2 \times \text{diag} \left(\frac{1}{S_2} \right) = \begin{bmatrix} 0.2887 & -0.2887 & \cdots & -0.2041 \\ -0.2887 & 0.2887 & \cdots & -0.2041 \\ \vdots & \vdots & \ddots & \vdots \\ 0.2887 & -0.2887 & \cdots & 0.4082 \end{bmatrix}.$$

Agora como, os dados são qualitativos, multiplicam-se as X_j novas tabelas por

$$\varphi_x = \sqrt{\frac{1 - \sum_{i=1}^I \sum_{j=1}^J C_{ijt}}{\lambda_P \times N_v}},$$

sendo N_v o número de variáveis de cada grupo. Nesta equação, o termo $1 - \sum_{i=1}^I \sum_{j=1}^J C_{ijt}$ representa a proporção das observações que não possuem categoria k . Assim, no primeiro grupo temos $N_v = 2$ (idade, fumante) e no segundo $N_v = 3$ (estado civil, esportista, estudante). Como $\lambda_{P_1} = 0.7041241$

e $\lambda_{P_2} = 0.5872716$, segue que: $\varphi_{x_1} = \sqrt{\frac{1 - \left[\frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{2} \quad \frac{1}{2} \right]}{0.704124 \times 2}}$ e $\varphi_{x_2} = \sqrt{\frac{1 - \left[\frac{1}{2} \quad \frac{1}{2} \quad \frac{7}{12} \quad \frac{5}{12} \quad \frac{2}{3} \quad \frac{1}{3} \right]}{0.5872716 \times 3}}$. Portanto, temos

$$Z_1 = \varphi_{x_1} \times X_1 = \begin{bmatrix} 0.2809 & -0.1404 & \cdots & -0.172 \\ -0.1404 & 0.2809 & \cdots & -0.172 \\ \vdots & \vdots & \ddots & \vdots \\ -0.1404 & 0.2809 & \cdots & 0.172 \end{bmatrix}$$

Do mesmo modo segue $Z_2 = \varphi_{x_2} \times X_2$. A matriz global será uma matriz $I \times K$, denotada por

$$Z = \begin{bmatrix} Z_1 & Z_2 \end{bmatrix} = \left[\begin{array}{ccc|ccc} 0.2809 & \cdots & -0.172 & 0.1538 & \cdots & -0.1256 \\ -0.1404 & \cdots & -0.172 & -0.1538 & \cdots & -0.1256 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -0.1404 & \cdots & 0.172 & 0.1538 & \cdots & 0.2511 \end{array} \right]$$

Agora, como os dados já foram balanceados, segue a análise global em Z .

Usando a matriz de decomposição valor singular sabemos que $Z = \hat{U}\hat{\Lambda}\hat{V}^T$, com $\hat{U}^T\hat{U} = \hat{V}^T\hat{V} = I$ é a matriz de autovetores e $\hat{\Lambda}$ é a matriz de autovalores. Assim, extraindo-se os autovalores e autovetores da matriz Z obtém-se

$$\hat{U} = \begin{bmatrix} -0.2009 & 0.3291 & \cdots & 0.1414 \\ -0.2367 & -0.1075 & \cdots & -0.1926 \\ \vdots & \vdots & \ddots & \vdots \\ 0.3456 & 0.3368 & \cdots & 0.0961 \end{bmatrix},$$

$$\hat{V} = \begin{bmatrix} -0.2287 & 0.3753 & \cdots & -0.4299 \\ 0.0742 & 0.1837 & \cdots & -0.4299 \\ \vdots & \vdots & \ddots & \vdots \\ 0.4179 & 0.2204 & \cdots & -0.2173 \end{bmatrix}$$

e

$$diag(\hat{\Lambda}) = \begin{bmatrix} 1.2361 & 1.0516 & 0.8444 & 0.4829 & 0.3937 & \cdots & 0.00 \end{bmatrix}$$

e as variâncias são dadas por

$$diag(\tilde{\Lambda}^2) = \begin{bmatrix} 1.5281 & 1.1059 & 0.7130 & 0.2332 & 0.1550 & \cdots & 0.00 \end{bmatrix}.$$

Os escores dos fatores globais são dados pela equação (5), ou seja

$$F = M^{\frac{1}{2}}\hat{U}\hat{\Lambda}$$

$$= \begin{bmatrix} -0.8604 & 1.1988 & \cdots & 0.0317 \\ -1.0136 & -0.3916 & \cdots & 0.0931 \\ \vdots & \vdots & \ddots & \vdots \\ 1.4797 & 1.2268 & \cdots & -0.0320 \end{bmatrix}$$

sendo M a matriz diagonal onde $m_i = I$.

Em F , cada linha representa um indivíduo (homens ou mulheres), e cada coluna, um componente, o que gera a Figura 17. Pode-se interpretar que os homens e as mulheres estão bem distribuídos entre os componentes, não podendo interpretar mais.

É fácil verificar que

$$\sum_{i=1}^p \Lambda_{ii}^2 = \sum_{i=1}^p \lambda_i \Rightarrow 1.5281 + 1.1059 + 0.7130 + \dots + 0.0000 = 3.8332$$

Usando a equação (1) para encontrar a proporção da variância total explicada pelos componentes principais, segue que

Tabela 23 Proporção explicada pelas variâncias.

| Componentes | λ_i | % explicação | % explicação acumulada |
|-------------|-------------|--------------|------------------------|
| 1 | 1.5281 | 39.86 | 39.86 |
| 2 | 1.1059 | 28.85 | 68.72 |
| 3 | 0.7130 | 18.60 | 87.32 |
| 4 | 0.2332 | 6.08 | 93.40 |
| 5 | 0.1550 | 4.04 | 97.44 |
| 6 | 0.0980 | 2.56 | 100.00 |

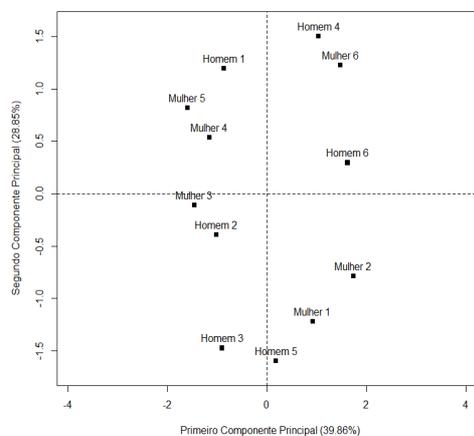


Figura 17 Gráfico da análise global dos homens e mulheres.

Análise parcial

A análise global revela a estrutura comum do espaço dos indivíduos (homens e mulheres). Agora para ver como cada variável interpreta este espaço, usa-se a equação (7). Logo tem-se que

$$F_1 = J \times Z_1 Z_1' P = \begin{bmatrix} -1.7478 & 1.2192 & \cdots & -0.5278 \\ -0.8636 & 0.6600 & \cdots & -0.5981 \\ \vdots & \vdots & \ddots & \vdots \\ 1.2968 & 0.4126 & \cdots & 1.4826 \end{bmatrix},$$

de modo análogo para o segundo grupo, segue que;

$$F_2 = J \times Z_2 Z_2' P = \begin{bmatrix} 0.0269 & 1.1784 & \cdots & 0.5911 \\ -1.1637 & -1.4432 & \cdots & 0.7842 \\ \vdots & \vdots & \ddots & \vdots \\ 1.6627 & 2.0411 & \cdots & -1.5467 \end{bmatrix}.$$

sendo $J = 2$ é o número de conjuntos de variáveis e P é dado pela equação (6).

A Figura 18 mostra os dois primeiros componentes principais da análise global juntamente com as projeções de homens e mulheres para os grupos de variáveis. Note que, a posição de cada homem ou mulher na análise global é o baricentro (isto é, centroide) de suas posições para as variáveis. Para facilitar a interpretação, puxamos linhas que unem as variáveis com a posição global dos homens e mulheres. Isto mostra uma relativa semelhança entre as duas representações parciais de cada grupo. Infelizmente não sugere muita coisa. A priori pode-se concluir que não existe diferença entre os grupos de variáveis.

Correlações entre os componentes principais e as variáveis originais

Para obter as correlações entre os componentes principais e as variáveis originais, basta apenas utilizar a Equação (3), o que resulta em

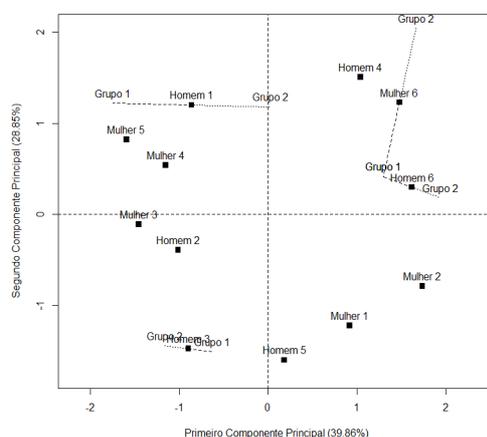


Figura 18 Gráfico dos grupos de variáveis com a análise global dos homens e mulheres.

Pode-se observar que a Tabela 24 não é de muita ajuda, pois, como os dados são categóricos, as variáveis originais foram expandidas em variáveis binárias. Assim, não se pode dizer nada a respeito das variáveis.

Inércia parcial entre os peritos

As relações entre os grupos e a solução global são analisadas calculando-se a inércia parcial de cada grupo por cada dimensão da análise global, isto é, usando-se a equação (8), ou seja

$$R_{j,i} = \lambda_p \times \sum_j^{J_k} \hat{V}_{j,i}^2,$$

Tabela 24 Correlações entre os componentes principais (eixos) e as variáveis originais.

| | Variáveis | Eixos 1 | Eixos 2 | Eixos 3 |
|---------|------------------|---------|---------|---------|
| Grupo 1 | Jovem | -0.4109 | 0.5736 | 0.6454 |
| | Meia-Idade | 0.1333 | 0.2808 | -0.9462 |
| | Velho | 0.2776 | -0.8544 | 0.3009 |
| | Fumante (Não) | -0.9402 | 0.0916 | -0.1095 |
| | Fumante (Sim) | 0.9402 | -0.0916 | 0.1095 |
| Grupo 2 | Casado | 0.7994 | 0.3519 | 0.2432 |
| | Solteiro | -0.7994 | -0.3519 | -0.2432 |
| | Esportista (Não) | 0.1491 | -0.8504 | -0.0917 |
| | Esportista (Sim) | -0.1491 | 0.8504 | 0.0917 |
| | Estudante (Não) | -0.8397 | -0.3767 | 0.0630 |
| | Estudante (Sim) | 0.8397 | 0.3767 | -0.0630 |

em que j refere-se ao grupo, p é o componente principal (eixo), e λ_p é o autovalor da decomposição de valores singulares. Assim, obtêm-se

$$R_{1,1} = \lambda_1 \times \sum_j^{J_k} \hat{V}_{1,1}^2 = 1.5281 \times \left((-0.2287)^2 + \dots + (0.4532)^2 \right) = 0.7525$$

$$R_{1,1} = \lambda_1 \times \sum_j^{J_k} U_{1,2} = 1.5281 \times \left((0.3445)^2 + \dots + (0.4179)^2 \right) = 0.7755$$

$$\vdots$$

$$R_{2,6} = \lambda_6 \times \sum_j^{J_k} U_{2,6} = 0.0980 \times \left((0.0249)^2 + \dots + (-0.5461)^2 \right) = 0.0449.$$

Cálculos semelhantes dão os valores informados na Tabela 25.

Tabela 25 Inércia dos grupos de variáveis.

| | Eixo 1 | Eixo 2 | Eixo 3 | Eixo 4 | Eixo 5 | Eixo 6 |
|----------|-------------|-------------|-------------|-------------|-------------|-------------|
| Grupo 1 | 0.7525 | 0.5446 | 0.6724 | 0.0761 | 0.0316 | 0.0531 |
| Grupo 2 | 0.7755 | 0.5613 | 0.0406 | 0.1570 | 0.1234 | 0.0449 |
| Σ | 1.5281 | 1.1059 | 0.7130 | 0.2332 | 0.1550 | 0.0980 |
| | λ_1 | λ_2 | λ_3 | λ_4 | λ_5 | λ_6 |

A partir das inércias obtidas em cada grupo (Tabela 25), visando uma melhor interpretação, o gráfico das inércias (Figura 19) é gerado, mostrando que há relação moderada entre os grupos de variáveis.

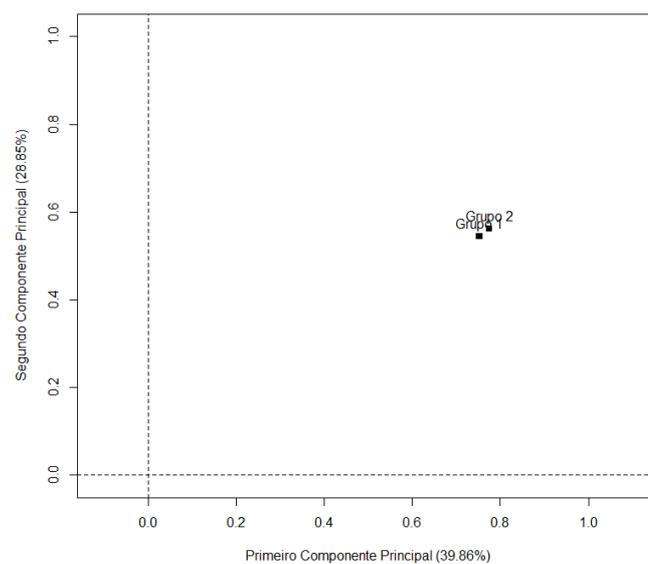


Figura 19 Gráfico das inércias dos grupos de variáveis.