UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ

CENTRO DE CIÊNCIAS EXATAS E TÉCNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AGRÍCOLA

Gabriela Medeiros

Marília Melo Favalesso

1ª PROVA

DISCIPLINA: ANÁLISE MULTIVARIADA

PROFESSORA: DRA. LUCIANA PAGLIOSA C. GUEDES

CASCAVEL-PR

2016

1. **Os dados do Quadro 1 representam uma pesquisa sobre o peso (em kg) de depósitos de cascas de 28 árvores (árvore de cortiça) em 4 direções: Norte (N), Sul (S), Leste (L) e Oeste (O).**

**Quadro 1: Massa (kg) de depósitos de cascas de 28 árvores de cortiça em nas direções Norte (N), Sul (S), Leste (L) e Oeste (O).**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Indivíduo**  | **N** | **L** | **S** | **O** |
| 1 | 72 | 66 | 76 | 77 |
| 2 | 60 | 53 | 66 | 63 |
| 3 | 56 | 57 | 64 | 58 |
| 4 | 41 | 29 | 36 | 38 |
| 5 | 32 | 32 | 35 | 36 |
| 6 | 30 | 35 | 34 | 26 |
| 7 | 39 | 39 | 31 | 27 |
| 8 | 42 | 42 | 31 | 25 |
| 9 | 37 | 40 | 31 | 25 |
| 10 | 33 | 29 | 27 | 36 |
| 11 | 32 | 30 | 34 | 28 |
| 12 | 63 | 45 | 74 | 63 |
| 13 | 54 | 46 | 60 | 52 |
| 14 | 47 | 51 | 52 | 43 |
| 15 | 91 | 79 | 100 | 75 |
| 16 | 56 | 68 | 47 | 50 |
| 17 | 79 | 65 | 70 | 61 |
| 18 | 81 | 80 | 68 | 58 |
| 19 | 78 | 55 | 67 | 60 |
| 20 | 46 | 38 | 37 | 38 |
| 21 | 39 | 35 | 34 | 37 |
| 22 | 32 | 30 | 30 | 32 |
| 23 | 60 | 50 | 67 | 54 |
| 24 | 35 | 37 | 48 | 39 |
| 25 | 39 | 36 | 39 | 31 |
| 26 | 50 | 34 | 37 | 40 |
| 27 | 43 | 37 | 39 | 50 |
| 28 | 48 | 54 | 57 | 43 |

Critérios para aplicação de Análise Fatorial e Análise dos Componentes Principais:

1. Analise do tamanho amostral:

Os números de observações cumprem com os critérios de tamanho de amostra (Quadro 2).

**Quadro 2: Critérios de tamanho amostral para aplicação de analise fatorial e análise dos componentes principais.**

|  |  |  |
| --- | --- | --- |
| Critério | Valor observado | Critério aceito? |
| ≥ 50 observações (Hair,2009) | 112 | Sim |
| ≤ 200 observações (Meyer 2006) | 112 | Sim |
| Observações / N° de Variáveis ≥ 10 (Hair, 2009 e Meyer, 2006) | 28 | Sim |

1. Correlação de Pearson:

Para a realização da análise de componentes principais (PCA) e análise fatorial, foi utilizada uma matriz de correlação de Pearson (tabela 1). Segundo a classificação de Callegari-Jacques (2003), com exceção de Oeste e Leste que apresentaram r=0,77 (correlação forte), as demais direções apresentaram correlações muito fortes (0,80 ≥ r < 1 (tabela 1; figura 1). Todas as correlações são significativas (p < 0,01) (tabela 1).

**Tabela 1. Matriz de correlação de Pearson para a variável massa de casca de árvore (kg) por direção geográfica. \*P < 0,001.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **N** | **L** | **S** |
| **L** | **0,89\*** |  |  |
| **S** | **0,90\*** | **0,83\*** |  |
| **O** | **0,88\*** | **0,77\*** | **0,92\*** |

**Figura 1: Gráficos de dispersão para massa de casca de árvore por direção geográfica.**

1. Teste de esfericidade de Bartlett:

H0: A matriz de correlação (R) é igual a matriz identidade (I) (R = I).

H1: A matriz de correlação (R) é diferente da matriz identidade (I) (R ≠ I).

Segundo o teste de esfericidade de Bartlett, a aplicação da análise de componentes principais, bem como a análise fatorial, é adequada (qui-quadrado: 131,3517, df = 6, P < 0,001). Uma vez que a hipótese nula de esfericidade foi rejeitada, é justificável a aplicação de alguma forma de redução de dimensão. Ao rejeitar a hipótese nula de igualdade foi permitido ver que na realidade as variáveis apresentam correlações verdadeiras e que a matriz de correlação é diferente da matriz identidade.

1. Critério de Kayser-Meyer-Olkin (índice KMO).

Levou-se em consideração a classificação do índice KMO segundo Alexandre Pereira (1999) onde:

**Quadro 3: classificação dos intervalos do índice KMO:**

|  |  |
| --- | --- |
| **KMO** | **Classificação** |
| 1,00- 0,90 | Muito boa |
| 0,80-0,90 | Boa |
| 0,70-0,80 | Média |
| 0,60-0,70 | Razoável |
| 0,50-0,60 | Má |
| < 0,50 | Inaceitável |

Segundo o índice de Kayser-Meyer-Olkin a adequação amostral é aceitável e meritória (KMO = 0,82).

1. **Faça a análise dos dados pela técnica de componentes principais. Interprete os resultados obtidos.**

Foi realizada a análise de componentes principais (PCA) a partir da matriz de correlação de Pearson e os resultados são apresentados em quatro tabelas e um gráfico. Para decisão do número de componentes principais a serem utilizados na PCA, foram considerados os critérios abordados por Hair (2009), sendo esses o scree plot (figura 2), o critério de raíz latente (autovalor > 1) e o critério de variação total dos dados (proporção da variância acumulada ser ≥ 95%). Dois dos três critérios considerados (gráfico scree e raiz latente) apontam apenas o primeiro fator como adequado para prosseguimento da PCA (Figura 1, Tabela 2). Porém, para a demonstração gráfica da PCA, o componente principal 2 também foi considerado nesta análise, além deste ser considerado como adequado pelo critério de variação total apresentado por Hair (2009).

Foram definidos dois componentes principais na PCA, sendo o primeiro componente formado pelas variáveis ‘Norte’ e ‘Sul’ (autovalor = 1,90, explicabilidade = 90%) e o segundo pelas variáveis ‘Leste’ e ‘Oeste’ (autovalor = 0,50, explicabilidade = 6%) (Tabela 2).

**Figura 2: Gráfico de autovalor para o critério de teste scree.**

**Tabela 2: Autovalores dos componentes principais para massa de casca de árvore (kg) por direção geográfica.**

|  |  |  |
| --- | --- | --- |
|  | **Comp.1** | **Comp.2** |
| **Autovalores** | 1,8973 | 0,5004 |
| **Proporção de variância** | 0,8999 | 0,0626 |
| **Proporção acumulada** | 0,8999 | 0,9625 |

No componente principal 1, os escores negativos indicam maiores valores (Kg) de depósito de casca de árvore para as direções norte e sul (tabela 2). No componente principal 2, escores positivos indicam maior depósito de casca de árvore na direção Leste, enquanto escores negativos indicam maior depósito de casca de árvore na direção oeste (Tabela 3).

**Tabela 3: Autovetores calculados pela PCA para a variável massa de casca de árvore por direção geográfica.**

|  |  |  |
| --- | --- | --- |
|  | **Comp.1** | **Comp.2** |
| **Norte** | **-0,511** | 0,129 |
| **Leste** | -0,483 | **0,759** |
| **Sul** | **-0,508** | -0,301 |
| **O** | -0,497 | **-0,563** |

A maioria dos indivíduos analisados apresentaram maior associação com o componente principal 1, ou seja, apresentaram maior depósito de casca de árvore nas regiões Norte e Sul (Tabela 4).

**Tabela 4: Escores fatoriais segundo indivíduos amostrados.**

|  |  |  |
| --- | --- | --- |
| **Árvores** | **Comp.1** | **Comp.2** |
| **1** | **-3,1138** | -0,4442 |
| **2** | **-1,5681** | -0,5165 |
| **3** | **-1,3550** | -0,1156 |
| **4** | **1,4799** | -0,4688 |
| **5** | **1,7500** | -0,2892 |
| **6** | **2,0761** | 0,2492 |
| **7** | **1,7180** | 0,5379 |
| **8** | **1,5944** | 0,7935 |
| **9** | **1,8133** | 0,6508 |
| **10** | **2,0403** | -0,3067 |
| **11** | **2,1135** | -0,0721 |
| **12** | **-1,6155** | -1,0413 |
| **13** | **-0,6163** | -0,4099 |
| **14** | -0,0441 | **0,2708** |
| **15** | **-4,7213** | 0,0622 |
| **16** | -0,9802 | **1,0409** |
| **17** | **-2,5892** | 0,2656 |
| **18** | **-2,9915** | 1,2095 |
| **19** | **-2,1103** | -0,1758 |
| **20** | **1,0011** | 0,0222 |
| **21** | **1,4309** | -0,1007 |
| **22** | **2,0894** | -0,1590 |
| **23** | **-1,1931** | -0,3462 |
| **24** | **1,0319** | -0,3329 |
| **25** | **1,4615** | 0,0981 |
| **26** | **0,9444** | -0,2317 |
| **27** | **0,6662** | -0,5431 |
| **28** | -0,3124 | **0,3529** |

A Figura 3 mostra um gráfico de cargas dos dois primeiros componentes principais, cada vetor representa um par de correlações: a correlação da variável original (direção geográfica) com cada um dos dois componentes principais. O gráfico revela quais das medidas originais das direções estão mais proximamente associadas a cada um dos dois componentes, bem como quais medidas são mais semelhantes. Além disso, na mesma figura, vemos associação de cada um dos indivíduos/árvores (pontos) com os componentes 1 e 2.

Na representação gráfica da PCA vemos a formação de três grupos: O primeiro (vermelho – Figura 3) é formado pelos indivíduos que apresentam maiores massas de casca de árvore (kg) nas direções Norte e Sul (ligados ao componente 1), o segundo (amarelo – Figura 3) é formado pelos indivíduos que têm massa intermediaria entre os grupos verde e vermelho ou seja, apresentou valores constantes entre as direções e relação positiva com o componente 2 (Leste e Oeste), e o terceiro (verde – Figura 3) é formado pelos indivíduos que apresentam os menores valores de massa de casca de árvore (kg) nas direções Norte e Sul. Ainda fora dos grupos são observados indivíduos discrepantes – indivíduo 15 (que apresentou maior depósito, em kg, de casca de árvore quando comparado com os demais, destacando valores mais elevados nas direções Sul e Norte), indivíduo 18 (que apresentou o maior depósito de casca de árvore para a direção Leste), o indivíduo 16 (que apesar de apresentar variação similar ao indivíduo 15, tem valores de massa de casca de árvore mais elevados nas direções Leste e Norte) e o indivíduo 12 (que apresentou relação negativa com ambos componentes e massas iguais para as direções Norte e Oeste, maior para a Direção Sul, e menor para Leste).



**Figura 3: Análise de componentes principais para depósito de casca de árvore por direção geográfica**.

1. **Faça a análises desses dados pela técnica de análise fatorial. Interprete os resultados**

Com a finalidade de condensar as variáveis de direção geográficas em um número menor de variáveis (fatores), foi aplicada a análise fatorial para os dados amostrados. Para seleção dos fatores que melhor combinam as variáveis em uma relação linear, foram considerados os critérios abordados por Hair (2009), sendo esses o *scree* plot (figura 1), o critério de raíz latente (autovalor > 1) e o critério de variação total dos dados (proporção da variância acumulada ser ≥ 95%). Dois dos três critérios considerados (gráfico *scree* e raiz latente) apontam apenas o primeiro fator como adequado para prosseguimento da análise fatorial (Figura 4, Tabela 5). Pelo critério de percentagem de variância abordado por Hair (2009), o fator 2 poderia ser incluso na análise, pois a soma da proporção da variância do fator 1 (90%) em conjunto com o fator 2 resultaria em um total aproximado de 96% da variação total dos dados (>95%), mas como os outros critérios analisados selecionaram apenas o fator 1, esse acabou sendo mantido. Dessa forma, o conjunto de variáveis aqui selecionados compõem apenas um fator segundo a análise fatorial, demonstrando ser um conjunto de dados unidimensional.

**Figura 4: Gráfico de autovalor para o critério de teste *scree*.**

**Tabela 5: Autovalores das variáveis ‘direção geográfica’.**

|  |  |
| --- | --- |
|  | **Fator 1** |
| **Autovalores** | 3,5996 |
| **Proporção de variância** | 0,8999 |
| **Proporção acumulada** | 0,8999 |

As cargas fatoriais das variáveis apresentam correlação satisfatória com o Fator 1, pois apresentam valores >0,40 (critério de Hair, 2009) (Tabela 6). Além disso, quando as cargas fatoriais são elevadas ao quadrado para encontrar os valores de comunalidade, encontramos valores >0,80, indicando que uma grande quantia da variância em cada uma das variáveis foi extraída pela solução fatorial (Tabela 6). Dessa forma, não foi necessário nenhum tipo de rotação para melhorar a interpretação das variáveis dentro dos fatores.

De maneira geral, as diferentes variáveis direcionais (Norte, Sul, Leste e Oeste) influenciam o Fator 1 (nomeado como “Direções geográficas”) quase que da mesma forma, logo que apresentam valores entre -0,94 e -0,97 (Tabela 6). A variável Oeste foi a menos influente (carga = -0,94) e a Norte a mais influente (carga = -0,97) (Tabela 6).

**Tabela 6: Cargas fatoriais, comunalidade e variância única para o fator 1 “Direções geográficas”.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Carga fatorial** | **Comunalidade** | **Variância única** |
| **Norte** | -0,9687 | 0,9385 | 0,0615 |
| **Leste** | -0,9172 | 0,8412 | 0,1588 |
| **Sul** | -0,9641 | 0,9295 | 0,0705 |
| **Oeste** | -0,9436 | 0,8990 | 0,1096 |

Em conclusão, as variáveis puderam ser condensadas em um único fator denominado “direção geográfica”, sendo os coeficientes do modelo fatorial apresentados na Tabela 7.

**Tabela 7: Coeficientes do modelo fatorial.**

|  |  |
| --- | --- |
|  | **Fator 1: Direção geográfica** |
| **Norte** | -0,5106 |
| **Leste** | -0,4834 |
| **Sul** | -0,5082 |
| **Oeste** | -0,4973 |

1. **Faça uma análise de agrupamento das árvores, considerando como medida de dissimilaridade a distância euclidiana e o método de ligação do vizinho mais próximo. Interprete os resultados obtidos.**

O cluster, com base nos dados de massa de casca de árvore (kg), dividiu os indivíduos em quatro grupos: a, b, c e d (Figura 4). O grupo a é composto por somente ‘árvore 15’, que apresentou a maior perda de massa de casca de árvore (kg) quando comparado aos demais indivíduos (figura 5). O grupo b é composto pela árvore 16, sendo esta similar aos grupos c e d. O grupo c foi o que apresentou indivíduos com valores de perca de massa de árvore entre 40-80kg (Figura 4, Figura 5), enquanto o último grupo (d) foi composto de indivíduos com as menores percas de casca de árvore (< 40 kg) (figura 4).

**Figura 4: Grupos definidos no dendrograma gerado a partir da análise de agrupamento das árvores, considerando como medida de dissimilaridade a distância euclidiana e o método de ligação do vizinho mais próximo**

O coeficiente de correlação cofenética pode ser utilizado para avaliar a consistência do padrão de agrupamento, sendo que valores próximos à unidade indicam melhor representação (CRUZ e CARNEIRO, 2003), sendo esse coeficiente a correlação linear de Pearson entre os elementos da matriz de dissimilaridade (distância euclidiana) e os elementos da matriz cofenética (matriz de distâncias entre as massas de cascas de árvore, obtida a partir do dendrograma). Assim, pode-se assumir que houve correlação linear significativa e de alta magnitude (r=0,7789491, t = 14,09, P < 0,001) entre as estimativas das distâncias euclidiana e os elementos da matriz cofenética.

**Figura 5: Boxplot da variação da massa de casca de árvore nos indivíduos (árvores) amostrados.**

1. **Faça a análise de agrupamento das variáveis (direções), considerando como medida de similaridade a correlação linear de Pearson (e de dissimilaridade uma função do coeficiente de correlação linear de Pearson) e o método de ligação do vizinho mais próximo. Interprete os resultados obtidos e compare-os com os resultados obtidos nos itens (a) e (b).**

A Figura 6a apresenta o agrupamento das variáveis de ‘direção geográfica’ mais similares entre sí, enquanto a Figura 6b apresenta o agrupamento das variáveis de ‘direção geográfica’ mais dissimilares entre sí. No cluster formado pela matriz de dissimilaridade da função 1-R², temos a formação dos grupos: ‘Sul’ e ‘Oeste’ como os mais similares entre sí (grupo 1), seguido da variável ‘Norte’ (grupo 2) e então da variável ‘Leste’ (grupo 3) (Figura 6a). No cluster formado pela matriz de similaridade de Pearson, temos a formação dos grupos: ‘Leste’ e ‘Oeste’ como os mais dissimilares entre sí (grupo 1), seguido de ‘Sul’ (grupo 2) e então de ‘Norte’ (grupo 3) (Figura 6b).

Apesar do cluster da Figura 6a apresentar uma medida de dissimilaridade como medida de distância entre as variáveis de ‘direção geográfica’, com o *método de ligação do vizinho mais próximo* para a formação dos grupos cluster, as variáveis agrupadas acabaram sendo as mais similares entre sí. Ao contrário deste, quando considerada a matriz de correlação de Pearson como medida de distância entre as variáveis (Figura 6b), e com o *método de ligação do vizinho mais próximo*, as variáveis agrupadas acabaram sendo as mais dissimilares entre sí.

O *método do vizinho mais próximo* é um tipo de método hierárquico aglomerativo que tem como objetivo agrupar indivíduos tendo como base uma matriz de distâncias pré-definida. Esse método cria ‘links’ entre indivíduos que apresentem os menores valores de distância entre sí (ou seja, indivíduos com distâncias mais baixas são ‘linkados’). Como a função do coeficiente de correlação de Pearson (1-R²) apresenta valores mais baixos para indivíduos que são mais proximamente associados entre sí, esses acabam por serem ‘linkados’ pela metodologia do vizinho mais próximo para formação dos grupos no cluster de similaridade (Figura 6a). Com a matriz de distâncias construída com base na correlação de Pearson acaba por ter a mesma situação, porém nesta os valores menores representam indivíduos menos correlacionados, e esses acabam sendo ‘linkados’ formando grupos dissimilares (Figura 6b).

Ambas os agrupamentos apresentaram coeficiente cofenético não significativos, mostrando inadequação na formação dos grupos (P > 0,05).

**Figura 6: a) Cluster para as variáveis de ‘direção geográfica’ tendo como medida de distância o coeficiente 1-R^2; b) Cluster para as variáveis de ‘direção geográfica’ tendo como medida de distância o coeficiente de correlação de Pearson.**

Diferença entre as análises realizadas:

De maneira geral, a análise de componentes principais (PCA), a análise fatorial e a análise de agrupamento (Cluster) apresentaram um mesmo objetivo geral: Reduzir as variáveis independentes em um número menor de variáveis– seja em ‘componentes principais’ pela PCA, seja em ‘fatores’ pela análise fatorial, ou seja em ‘grupos’ pela análise cluster. A diferença entre elas está no objetivo intrínseco de cada análise: Enquanto a PCA buscou reorientar as variáveis independentes, de forma que um pequeno número de componentes contendo algumas dessas variáveis explicassem a maior parte da variação dos dados, a análise fatorial objetivou reduzir as variáveis em um número menor de fatores comuns; e a análise cluster teve como objetivo agrupar as variáveis independentes tendo como medida de similaridade (ou dissimilaridade) o coeficiente de correlação de Pearson.

No boxplot da figura 7 é visto que as variáveis de ‘direção geográfica’ não apresentaram diferenças gritantes quanto a variação de seus valores de massa de casca de árvore (kg), e as medianas foram muito próximas. Além disso, todas apresentaram coeficientes de correlação de Pearson entre forte e muito forte (0,7>r≤0,92), demonstrando ampla associação entre as variáveis (tabela 1). Essa proximidade de valores (figura 7) demonstra não haver direção privilegiada para massa de casca de árvore (kg), além dos valores apresentarem associações lineares entre sí (tabela 1), o que contribuiu para o agrupamento de todas as variáveis em único fator na análise fatorial (Questão 1b). Pela PCA (Questão 1a), vemos que as variáveis ‘Norte’ e ‘Sul’ foram responsáveis por cerca de 90% da variação dos dados, e que ambas apresentam coeficientes negativos dentro do componente 1 (Figura 1, Tabela 3). Essas duas variáveis foram as que apresentaram os maiores coeficientes de correção com as demais variáveis, englobando assim a maior fração de variação quando comparada as demais ‘direções geográficas’ (Tabela 1).

Apesar da PCA mostrar as variáveis ‘Norte’ e ‘Sul’ como as responsáveis por grande parte da variação existentes nos dados, o cluster apontou o ‘Sul’ como mais similar a variável direcional ‘Oeste’ (questão 1c), sendo essa similaridade composta pela variação conjunta dos valores de casca de árvore (kg) (questão 1c – I parte). Sendo assim, ‘Sul’ e ‘Oeste’ apresentaram variações de valores mais proximamente relacionados quando comparados a ‘Sul’ e ‘Norte’. Enquanto isso, ‘Leste’ e ‘Oeste’ foram considerados os mais dissimilares entre sí, isso em razão dos indivíduos amostrados apresentarem percas de casca de árvore (kg) diferentes para essas direções.

Em conclusão, vemos que cada uma das metodologias apresenta as suas particularidades (objetivos intrínsecos), o que contribuiu para resultados de agrupamento diferenciados.

**Figura 7: Boxplot da variação da massa de casca de árvore nas direções analisadas**

**REFERÊNCIAS**

CALLEGARI-JACQUES, S.M**.. Bioestatística. Princípios e aplicações**. Porto Alegre: ArtMed, 2003. v. 1. 255p .

CRUZ, C.D.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético**. Viçosa: Editora UFV, 2003. 579p.

HAIR, Joseph F. et al. **Análise multivariada de dados**. Bookman Editora, 2009.

MEYER, K. To have your steak and eat it: genetic principal component analysis for beef cattle data. In**: WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTIO**N, 8., 2006, Belo Horizonte. Anais… Belo Horizonte: SBMA, 2006. p. 13-18. CDROM.

PEREIRA, Alexandre. **Guia Prático de Utilização do SPSS: Análise de Dados para Ciências Sociais e Psicologia**. 2. ed. Lisboa: Edições Silabo, 1999.