**Serviço Público Federal**

**Universidade Federal do Pará**

**Instituto de Ciências Exatas e Naturais**

**Faculdade de Estatística**

Dionisio Alves da Silva Neto

Matrícula: 202007840008

**Atividade 5 de Análise Multivariada II**:

Análise de Correlação Canônica

Belém, PA

2022

1. **Introdução**

O presente trabalho visa aplicar a técnica multivariada de Análise de Correlação Canônica, em dados sobre os fatores que influenciam a localização de sítios habitacionais maias pré-históricos, os quais estão localizados no distrito de Corozal Belize, na América Central. O banco de dados pode ser interpretado como a divisão de dois blocos, um relacionado a um conjunto de variáveis independentes (X) e, o outro associado a um conjunto de variáveis dependentes (Y).

Para o conjunto X, temos elementos que ajudam a explicar características do solo:

* **X1**: O percentual do solo com enriquecimento constante de cal;
* **X2**: O percentual do solo de prado com água subterrânea de cálcio;
* **X3**: O percentual do solo com coral leito rochoso sob condições de enriquecimento constante de cal ; e
* **X4**: O percentual de solos aluviais e orgânicos adjacentes a rios e solo orgânico salino na costa.

Para o conjunto Y, temos características relacionadas à efetividade da vegetação:

* **Y1**: O percentual de floresta latifoliada decídua sazonal;
* **Y2**: O percentual de pântano de floresta alta e baixa, pântano herbáceo, e brejo;
* **Y3**: O percentual de floresta de cohume; e
* **Y4**: O percentual de floresta mista.

1. **Análise Descritiva**

Como de costume, antes da aplicação de qualquer técnica multivariada ou modelo de predição, é sempre adequado realizar uma Análise Exploratória de Dados (AED), de modo a avaliar a qualidade, peculiaridades e características únicas que possam estar presentes no conjunto de dados. Dessa forma, para verificar os conjuntos X e Y do estudo Green (1973), iremos explorar alguns fatores que os dados revelam através de gráficos e medidas descritivas.

De acordo com a **Tabela 1**, percebe-se que para as variáveis X2, X4, 50% dos percentuais estão acima de 0, o que entra em contraste com X1, que tem 50% dos valores acima de 20 por cento e X3 que acumula valores acima de 0 acima do Quartil 3, o que indicaria que apenas 25% de seus valores estão acima de 0. Ademais, as médias e a variabilidade para as variáveis entram em discordância, com X1 tendo um percentual mediano bem distante das demais (47,71), assim como o maior valor para o desvio-padrão (32,39).

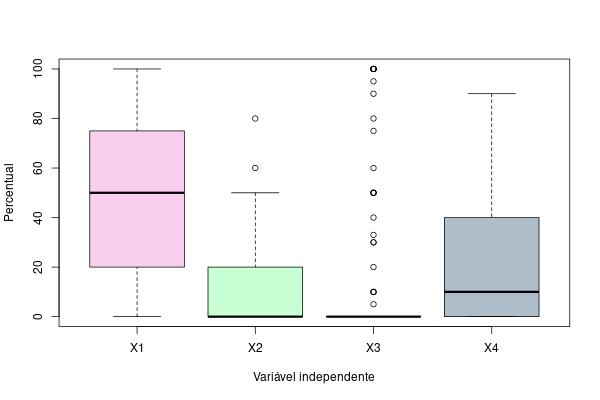
**Tabela 1**: Estatísticas do Mínimo, Primeiro Quartil (Q1), Mediana, Média, Desvio-Padrão, Terceiro Quartil (Q3) e Máximo para cada variável independente.

|  | Variável independente | | | |
| --- | --- | --- | --- | --- |
| Estatística | X1 | X2 | X3 | X4 |
| Mínimo | 0,00 | 0,00 | 0,00 | 0,00 |
| Q1 | 20,00 | 0,00 | 0,00 | 0,00 |
| Mediana | 50,00 | 0,00 | 0,00 | 10,00 |
| Média | 47,71 | 11,32 | 9,85 | 20,15 |
| Desvio-Padrão | 32,39 | 17,25 | 26,24 | 24,14 |
| Q3 | 75,00 | 20,00 | 0,00 | 40,00 |
| Máximo | 100,00 | 80,00 | 100,00 | 90,00 |

Fonte: Construído pelo autor, 2022.

A **Figura 1** elucida a diferença entre as distribuições de cada variável, com X3 com valores muito próximos a zero e exibindo alguns outliers, X1 tendo conflito de seus quartis em relação aos das outras variáveis por concentrar altos valores, X2 com 75% dos valores abaixo de 20 e, X3 exibindo 75% dos valores abaixo de 40.

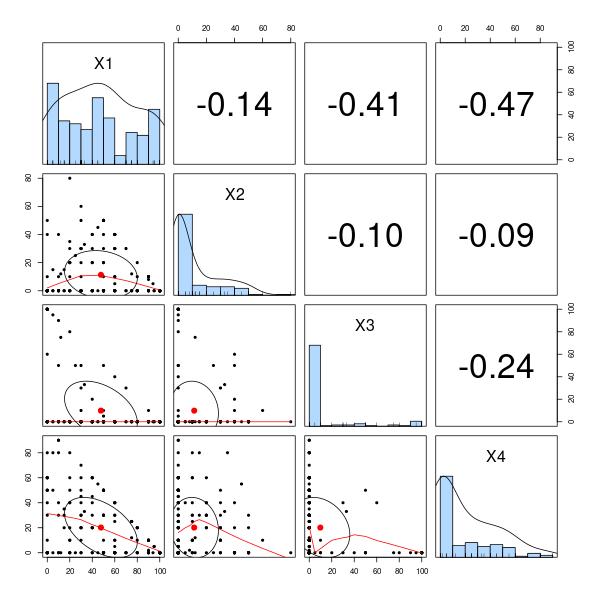
**Figura 1**: Diagrama de caixa para cada variável independente.



Fonte: Construído pelo autor, 2022.

Em última análise exploratória do conjunto X, a **Figura 2** nos mostra que X1 se distribui bem entre os valores 0 e 1; X2 está concentrado valores bem próximos a 0 e 20 por cento, com alguns valores chegando próximo a 40 por cento; X3 comprime bastante valores dentro do intervalo de 0 a 10 por cento e; X4 tem valores próximos ao início, mas exibe valores para os outros intervalos até 100 por cento. As correlações entre as variáveis apontam relações negativas entre as variáveis, com algumas sendo consideradas fracas e outras como moderadas. Entre X1 e X4 ocorre a maior correlação negativa e, entre X2 e X4 ocorre a menor correlação negativa, considerada quase inexistente.

**Figura 2**: Diagrama de dispersão com ajuste, histograma com densidade e correlações para o conjunto de variáveis independentes (X).

****

Fonte: Construído pelo autor, 2022.

A **Tabela 2** resume os dados do conjunto de variáveis dependentes a medidas descritivas. Desse modo, percebe-se que para Y3 e Y4, 75% dos valores estão abaixo de 0%, enquanto que no primeiro quartil das variáveis Y1 e Y2, já observamos valores sendo expressados. Para os valores médios dos percentuais, temos também uma discrepância ao avaliar que as variáveis Y1 (43,31) e Y2 (43,05) com as variáveis Y3 (1,03) e Y4 (2,35), ou seja, para o segundo par temos valores médios bem baixos com relação ao primeiro par.

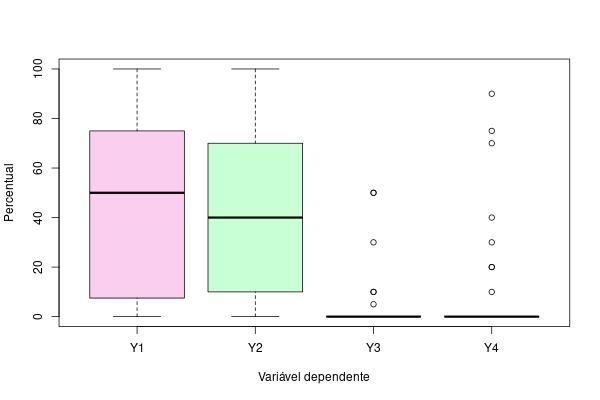
**Tabela 2**: Estatísticas do Mínimo, Primeiro Quartil (Q1), Mediana, Média, Desvio-Padrão, Terceiro Quartil (Q3) e Máximo para cada variável dependente.

|  | Variável independente | | | |
| --- | --- | --- | --- | --- |
| Estatística | Y1 | Y2 | Y3 | Y4 |
| Mínimo | 0,00 | 0,00 | 0,00 | 0,00 |
| Q1 | 7,50 | 10,00 | 0,00 | 0,00 |
| Mediana | 50,00 | 40,00 | 0,00 | 0,00 |
| Média | 43,31 | 43,05 | 1,03 | 2,35 |
| Desvio-Padrão | 34,88 | 33,29 | 6,31 | 11,89 |
| Q3 | 75,00 | 70,00 | 0,00 | 0,00 |
| Máximo | 100,00 | 100,00 | 50,00 | 90,00 |

Fonte: Construído pelo autor, 2022.

A **Figura 3** exibe diagramas de dispersão para cada variável dependente. Dessa maneira, temos uma alta concentração de valores próximos a zero por cento para as variáveis Y3 e Y4, assim modo uma certa igualdade nos quartis das variáveis Y1 e Y4. Também, é evidente a diferença entre as distribuições das variáveis Y1 e Y2 para as variáveis Y3 e Y4.

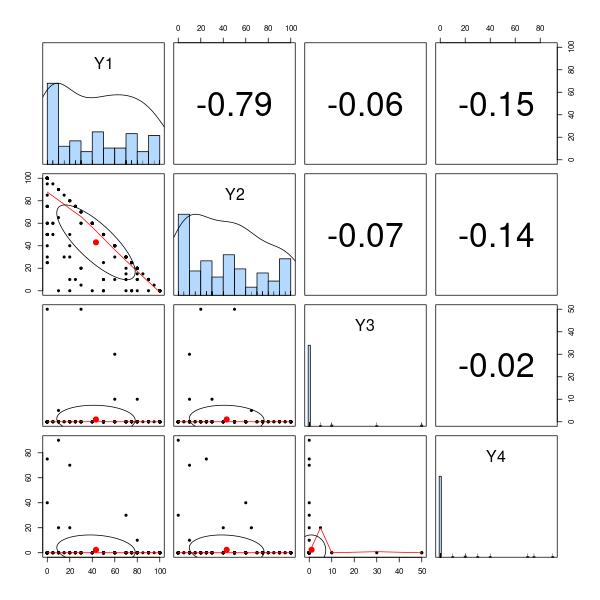
**Figura 3**: Diagrama de caixa para cada variável dependente.



Fonte: Construído pelo autor, 2022.

Em último, na análise descritiva das variáveis dependentes, a **Figura 4** mostra que as variáveis Y3 e Y4 concentram muitos valores em torno de 0% a 10%; enquanto que as variáveis Y1 e Y2 já englobam outros valores para além de 10%. As correlações lineares entre todas as variáveis dependentes são negativas, dado que o par contendo as variáveis Y1 e Y2 tem uma forte correlação linear negativa e as variáveis Y3 e Y4 formam o para com a menor correlação negativa, a qual é bem próxima de zero.

**Figura 4**: Diagrama de dispersão com ajuste, histograma com densidade e correlações para o conjunto de variáveis dependentes (Y).

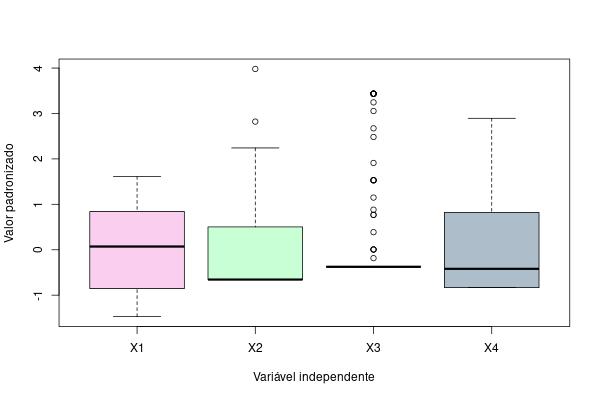


Fonte: Construído pelo autor, 2022.



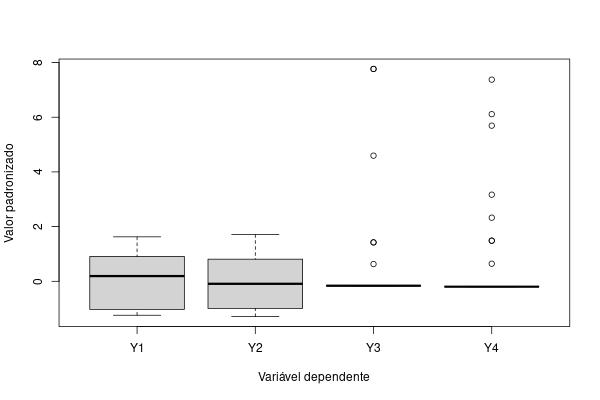
Após verificar algumas incongruências nos valores dos conjuntos X e Y, devemos realizar a padronização das variáveis para a adequação da técnica de Análise de correlação Canônica. Neste viés, para o presente trabalho, escolhemos a metodologia na qual consiste em transformar as variáveis de modo que as suas médias sejam igual a zero e a variância seja igual a 1. Os diagramas de caixa para os valores padronizados dos conjuntos X e Y estão presentes nas **Figuras 5** e **6**, nesta respectiva ordem. Nestas podemos perceber uma melhor adequação para os conjuntos de independência e dependência.

**Figura 5**: Diagrama de caixa para cada variável dependente padronizada.



Fonte: Construído pelo autor, 2022.

**Figura 6**: Diagrama de caixa para cada variável dependente padronizada.

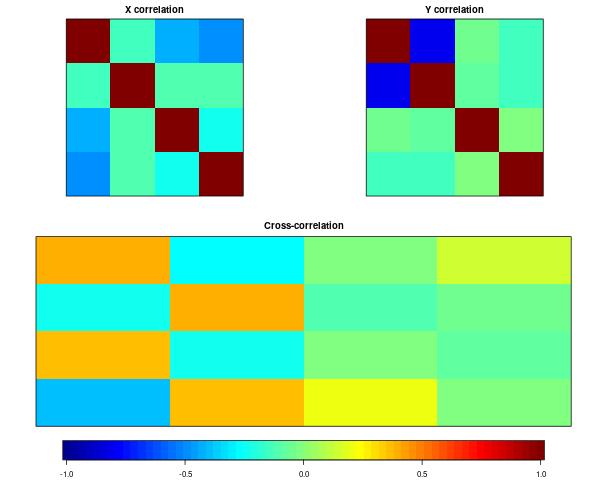


Fonte: Construído pelo autor, 2022.

1. **Análise de Correlação Canônica**

O primeiro passo, antes de estudar as correlações canônicas, é avaliar as correlações lineares dos conjuntos originais. Desta maneira, a **Figura 7** nos mostra as correlações de Pearson avaliadas individualmente entre o conjunto dependente e independente, tal como a correlação cruzada entre as variáveis destes conjuntos. Em suma, têm-se que, para o conjunto X, as maiores correlações são para os pares de variáveis (X1, X3) e (X1, X4); para o conjunto Y, as maiores correlações são entre as variáveis Y1 e Y2. Na avaliação cruzada, temos uma alta correlação entre X1 e Y4.

**Figura 7**: Gráfico de correlações lineares entre os conjuntos X e Y, de forma singular e conjunta.



Fonte: Construído pelo autor, 2022.

**3.1 Estimativas dos coeficientes canônicos**

As **Tabelas 3** e **4** mostram as estimativas dos pesos canônicos para os conjuntos de variáveis independentes e dependentes, respectivamente. Desse modo, é válido ressaltar que cada variável canônica é formada pela combinação linear de cada variável original do seu respectivo conjunto associado ao peso que a mesma tem. Por exemplo, para formar o primeiro par de variáveis canônicas temos que:

Esses pesos têm uma interpretação similar aos coeficientes de regressões lineares. Exemplificando, no conjunto de variáveis canônicas a respeito do solo, ao acrescentarmos uma unidade padronizada da variável X1 (percentual do solo com enriquecimento constante de cal), teremos uma diminuição de, aproximadamente, 0,0409 na dimensão 1 do conjunto de variáveis canônicas do solo, mantendo as demais condições constantes. A exemplo do conjunto de variáveis canônicas da vegetação, ao aumentarmos um percentual padronizado da variável Y3 (percentual de floresta de cohume), teremos um decréscimo perto de 0,0320 da primeira dimensão 1 do conjunto de variáveis canônicas a respeito da vegetação, mantendo todas as outras condições constantes.

**Tabela 3**: Estimativas dos pesos canônicos para o conjunto de variáveis independentes.

|  | U1 | U2 | U3 | U4 |
| --- | --- | --- | --- | --- |
| X1 | -0.04092638 | -0.01540472 | 0.011846070 | -0.01352513 |
| X2 | -0.01663907 | -0.05100292 | -0.03305513 | -0.00100801 |
| X3 | -0.04272641 | -0.01111547 | 0.005440905 | 0.027383169 |
| X4 | -0.02303790 | -0.04003337 | 0.035895661 | 0.006279409 |

Fonte: Construído pelo autor, 2022.

**Tabela 4**: Estimativas dos pesos canônicos para o conjunto de variáveis dependentes.

|  | V1 | V2 | V3 | V4 |
| --- | --- | --- | --- | --- |
| Y1 | -0.04808048 | -0.01997836 | -0.00633285 | 0.003450043 |
| Y2 | -0.03009303 | -0.04505894 | -0.00915159 | 0.000329835 |
| Y3 | -0.03419738 | -0.05029829 | 0.145338822 | 0.041645885 |
| Y4 | -0.04275721 | -0.02576262 | 0.016868520 | -0.07864085 |

Fonte: Construído pelo autor, 2022.

**3.2 Teste para as Correlações Canônicas**

As estimativas para as correlações entre os pares variáveis canônicas são representadas por:

* Corr() = 0,7610692
* Corr() = 0,5324149
* Corr() = 0,2382218
* Corr() = 0,1214940

Dessa maneira, as **Tabelas 5, 6, 7** e **8** abordam os testes estatísticos de Wilks, Hotelling, Pillai e Roy para avaliar a significância das correlações lineares entre os pares de variáveis canônicas. Eles funcionam de forma que a cada nova estatística de teste e, consequentemente, um p-valor associado, exclua as maiores correlações canônicas anteriores. Com exceção do de Roy, o qual utiliza a maior correlação canônica diretamente.

Como pode ser observado nas **Tabelas 5, 6,** e **7**, os testes de Wilks, Hotelling, Pillai mostram que as correlações canônicas são significativas até a dimensão 3, pois os níveis descritivos encontrados foram menores do que 0,05, isto é, as correlações canônicas se mostraram significativas quando combinamos as dimensões 1 e 4, 2 e 4, 3 e 4. Apenas a quarta dimensão unicamente não se mostrou estatisticamente significativa nos testes, pois retornou um valor maior do que 0,05. Para o teste da Raiz de Roy, a maior correlação canônica foi determinada como estatisticamente significativa, como resume a **Tabela 8**.

**Tabela 5**: Resultados do teste de Wilks para a significância das correlações canônicas.

| Par | Estatística | Aproximação | df1 | df2 | P-valor |
| --- | --- | --- | --- | --- | --- |
| 1 para 4 | 0.2801910 | 14.125336 | 16 | 437.5098 | 0.000000e+00 |
| 2 para 4 | 0.6658949 | 7.084095 | 9 | 350.6088 | 1.950724e-09 |
| 3 para 4 | 0.9293272 | 2.706205 | 4 | 290.0000 | 3.060939e-02 |
| 4 para 4 | 0.9852392 | 2.187364 | 1 | 146.0000 | 1.413016e-01 |

Fonte: Construído pelo autor, 2022.

**Tabela 6**: Resultados do teste de Hotelling para a significância das correlações canônicas.

| Par | Estatística | Aproximação | df1 | df2 | P-valor |
| --- | --- | --- | --- | --- | --- |
| 1 para 4 | 1.84732716 | 16.337300 | 16 | 566 | 0.000000e+00 |
| 2 para 4 | 0.47075226 | 7.505883 | 9 | 574 | 2.018451e-10 |
| 3 para 4 | 0.07514585 | 2.733430 | 4 | 582 | 2.829234e-02 |
| 4 para 4 | 0.01498195 | 2.209837 | 1 | 590 | 1.376673e-01 |

Fonte: Construído pelo autor, 2022.

**Tabela 7**: Resultados do teste de Pillai para a significância das correlações canônicas.

|  | Estatística | Aproximação | df1 | df2 | P-valor |
| --- | --- | --- | --- | --- | --- |
| 1 para 4 | 0.93420242 | 11.122192 | 16 | 584 | 0.000000e+00 |
| 2 para 4 | 0.35497603 | 6.405866 | 9 | 592 | 1.050175e-08 |
| 3 para 4 | 0.07151043 | 2.730455 | 4 | 600 | 2.840306e-02 |
| 4 para 4 | 0.01476080 | 2.251952 | 1 | 608 | 1.339651e-01 |

Fonte: Construído pelo autor, 2022.

**Tabela 8**: Resultados do teste de Roy para a significância das correlações canônicas.

|  | Estatística | Aproximação | df1 | df2 | P-valor |
| --- | --- | --- | --- | --- | --- |
| 1 para 1 | 0.5792264 | 50.24498 | 4 | 146 | 0.000000e+00 |

Fonte: Construído pelo autor, 2022.

**3.3 Correlação entre as variáveis originais e variáveis canônicas**

As correlações lineares de Pearson entre os conjuntos de variáveis originais e variáveis dependentes são também chamadas de *loadings*. Os *loadings* canônicos representam a variância captada das variáveis originais para o novo conjunto de variáveis canônicas formadas e, portanto, os *loadings* podem ser interpretados como a contribuição de cada variável original para a formação da função canônica.

Desse modo, a **Tabela 9** mostra a representação das variáveis originais do conjunto X para a formação das variáveis canônicas independentes. Logo, percebe-se que as variáveis canônicas U3 e U4 conseguem representar bastante acerca da variabilidade original do conjunto independente. Para X1 e X3, temos uma forte correlação em U4 e, para X3 e X4, temos uma alta correlação em U3.

**Tabela 9**: Correlações lineares entre as variáveis originais do conjunto independente e as variáveis canônicas do conjunto independente.

|  | U1 | U2 | U3 | U4 |
| --- | --- | --- | --- | --- |
| X1 | -0.56500103 | 0.1999069 | 0.0003682975 | -0.80050667 |
| X2 | 0.06308064 | -0.6888603 | -0.7211485117 | -0.03791011 |
| X3 | -0.41898770 | 0.2274037 | -0.166291258 | 0.86318254 |
| X4 | 0.36068909 | -0.5793048 | 0.7065760097 | 0.18724236 |

Fonte: Construído pelo autor, 2022.

Na **Tabela 10**, temos as correlações lineares entre o conjunto original Y e as variáveis canônicas independentes. Dessa forma, no geral, temos fracos e moderados valores de correlações, sendo que U1 engloba a variabilidade de Y1 e Y2; U2 também representa moderadamente Y2 e; as variáveis Y3 e Y4 não encontram-se tão bem representadas nas variáveis canônicas.

**Tabela 10**: Correlações lineares entre as variáveis originais do conjunto dependente e as variáveis canônicas do conjunto independente.

|  | U1 | U2 | U3 | U4 |
| --- | --- | --- | --- | --- |
| Y1 | -0.60750533 | 0.291989873 | -0.01591164 | 0.02915940 |
| Y2 | 0.30469446 | -0.473287301 | -0.05263193 | 0.00317759 |
| Y3 | -0.02711553 | -0.088525530 | 0.22530719 | 0.03361905 |
| Y4 | -0.08224247 | 0.007152211 | 0.06064267 | -0.11674429 |

Fonte: Construído pelo autor, 2022.

A **Tabela 11** resume os *loadings* das variáveis originais do conjunto X (independente) e as variáveis canônicas dependentes. Destarte, Como foi observado anteriormente, por serem de naturezas opostas, verifica-se que as variáveis originais do conjunto independente são fracamente representadas nas novas variáveis canônicas dependentes. Apenas V1 consegue representar moderadamente a variável X1.

**Tabela 11**: Correlações lineares entre as variáveis originais do conjunto independente e as variáveis canônicas do conjunto dependente.

|  | V1 | V2 | V3 | V4 |
| --- | --- | --- | --- | --- |
| X1 | -0.43000491 | 0.1064334 | 8.773648e-05 | -0.097256789 |
| X2 | 0.04800874 | -0.3667595 | -1.717933e-01 | -0.004605852 |
| X3 | -0.31887865 | 0.1210731 | -3.961420e-02 | 0.104871532 |
| X4 | 0.27450937 | -0.3084305 | 1.683218e-01 | 0.022748830 |

Fonte: Construído pelo autor, 2022.

A **Tabela 12** nos mostra as correlações de Pearson entre as variáveis do conjunto dependente original e as variáveis canônicas dependentes. Em vista disso, nota-se que a variável canônica V1 explica a variabilidade da variável X1; a variável canônica V2 explica fortemente sobre a variabilidade de Y2 e moderadamente da variabilidade de Y1; a variável canônica V3 engloba variabilidade de Y3 e; a variável canônica V4 explica fortemente a variabilidade de Y4 e fracamente a variabilidade de Y1 e Y3.

**Tabela 12**: Correlações lineares entre as variáveis originais do conjunto dependente e as variáveis canônicas do conjunto dependente.

|  | V1 | V2 | V3 | V4 |
| --- | --- | --- | --- | --- |
| Y1 | -0.79822609 | 0.54842546 | -0.06679338 | 0.24000680 |
| Y2 | 0.40035051 | -0.88894455 | -0.22093666 | 0.02615429 |
| Y3 | -0.03562821 | -0.16627171 | 0.94578747 | 0.27671359 |
| Y4 | -0.10806174 | 0.01343353 | 0.25456389 | -0.96090552 |

Fonte: Construído pelo autor, 2022.

**3.4 Índice de Redundância**

O índice de redundância é uma medida mais direta para sabermos o quando a variabilidade do conjunto canônico dependente é representada nas variáveiscanônicas dependentes. Desta maneira, de acordo com a **Tabela 12**, temos que as variáveis canônicas independentes conseguem englobar cerca de 17,51% da variabilidade do conjunto canônico dependente.

**Tabela 12**: Redundâncias das variáveis canônicas independentes para o conjunto variáveis canônicas dependentes.

| U1 | U2 | U3 | U4 | Total X|Y |
| --- | --- | --- | --- | --- |
| 0.091062 | 0.063907 | 0.014854 | 0.005249 | 0.175072 |

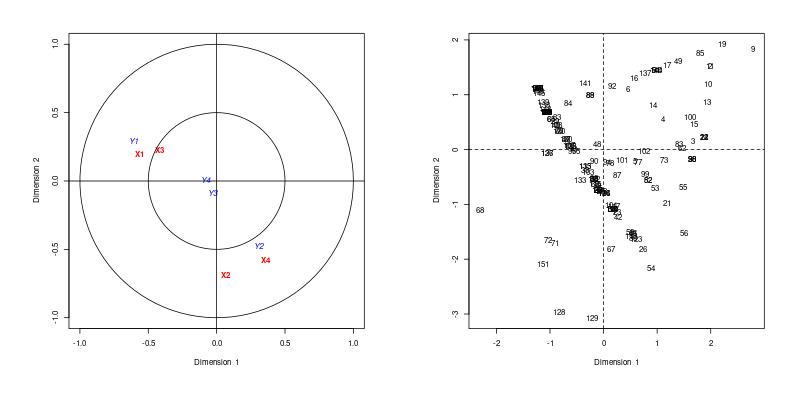
Fonte: Construído pelo autor, 2022.

**3.5 Gráfico dos *loadings* canônicos e indivíduos**

O gráfico dos *loadings* canônicos serve para avaliarmos os pares de variáveis canônicas entre as dimensões. Dessa maneira, o gráfico do lado esquerdo na **Figura 8**  nos mostra que, na dimensão 1, Y1, X1 e X3 são representadas moderadamente; na dimensão 2, conseguimos explicar X2 e X4, de forma moderada.

O gráfico de indivíduos, representado no lado esquerdo da **Figura 8**, nos mostra a representação dos elementos da base entre as dimensões 1 e 2.

**Figura 8**: Gráfico das dos loadings canônicos e dos indivíduos nos primeiro para o par canônico (U1, V1) de primeira dimensão.



Fonte: Construído pelo autor, 2022.

**ANEXO I**: Script em R desenvolvido na atividade

## ------

## Análise Multivariada II

## Lista 5

## Análise de Correlação Canônica

## ------

## ------

## Pacotes

## ------

if(!require(pacman)) install.packages("pacman"); library(pacman)

p\_load(CCA, CCP, readxl,psych, knitr, qcc, candisc, tibble, ggplot2, forcats)

## ------

## Dados

## ------

setwd('/cloud/project')

dados = read\_excel('Quadrados.xlsx')

head(dados)

dim(dados)

# X1 = % do solo com enriquecimento constante de cal;

# X2 = % do solo de prado com água subterrânea de cálcio;

# X3 = % do solo com coral leito rochoso sob condições de

# enriquecimento constante de cal; e

# X4 = % de solos aluviais e

# orgânicos adjacentes a rios e solo orgânico salino na costa.

X = as.matrix(dados[,2:5])

# Y1 = % de floresta latifoliada decídua sazonal;

# Y2 = % de pântano de floresta alta e baixa, pântano herbáceo, e brejo;

# Y3 = % de floresta de cohune; e

# Y4 = % de floresta mista.

Y = as.matrix(dados[,6:9])

## ---

## Analise descritiva

## ---

jpeg("descritva.conj.X.jpeg", width = 600, height = 600)

pairs.panels(X, method = "pearson", hist.col = "#b3d9ff", density = TRUE)

dev.off()

jpeg("descritva.conj.Y.jpeg", width = 600, height = 600)

pairs.panels(Y, method = "pearson", hist.col = "#b3d9ff", density = TRUE)

dev.off()

summary(X)

apply(X, MARGIN = 2, FUN = sd)

summary(Y)

apply(Y, MARGIN = 2, FUN = sd)

cores = c('#F9CEEE', '#C8FFD4', '#F29393', '#AEBDCA')

jpeg("boxplot.conj.X.jpeg", width = 600, height = 400)

boxplot(X, xlab = "Variável independente", col = cores,

ylab = 'Percentual')

dev.off()

jpeg("boxplot.conj.Y.jpeg", width = 600, height = 400)

boxplot(Y, xlab = "Variável dependente", col = cores,

ylab = 'Percentual')

dev.off()

## ---

## Realizacao da padronizacao

## ---

X.scaled = scale(X, center = TRUE, scale = T)

colMeans(X.scaled) ## verificando se a media para as colunas ficam zero.

apply(X.scaled, 2, sd) ## verificando se a variancia para as colunas ficam um.

Y.scaled = scale(Y, center = TRUE, scale = T)

colMeans(Y.scaled) ## verificando se a media para as colunas ficam zero.

apply(Y.scaled, 2, sd) ## verificando se a variancia para as colunas ficam um.

jpeg("boxplot.conj.X.padronizado.jpeg", width = 600, height = 400)

boxplot(X.scaled, xlab = "Variável independente",

ylab = "Valor padronizado", col = cores)

dev.off()

jpeg("boxplot.conj.Y.padronizado.jpeg", width = 600, height = 400)

boxplot(Y.scaled, xlab = "Variável dependente",

ylab = "Valor padronizado", color = cores)

dev.off()

## ----

## Realizacao da Analise de Correlacao Canonica

## ----

cor.can = matcor(X, Y)

jpeg("correlograma.matriz.canonica.jpeg", width = 600, height = 500)

img.matcor(cor.can, type = 2)

dev.off()

## ----

## Aplicacao do modelo de Correlacao Canonica

## ----

mod.cc = cc(X,Y)

mod.cc$cor

dim(X)

dim(Y)

# These are the “canonical correlations”, which indicate the overall

# relationship between X and Y in 4 dimensions. We have quatro since

# our smallest input matrix has 4 columns.

## ----

## Estimativas para as variaveis canonicas

## ----

mod.cc$xcoef

mod.cc$ycoef

mod.cc$scores

## ----

## Teste para ver quais variaveis canonicas sao estatisticamente

## significativas

## ----

correlacao = mod.cc$cor

n = nrow(X) ## numero de linhas em X

p = ncol(X) ## numero de colunas em X

q = ncol(Y) ## numero de colunas em Y

p.asym(correlacao, n, p, q, tstat="Wilks")

p.asym(correlacao, n, p, q, tstat="Hotelling")

p.asym(correlacao, n, p, q, tstat="Pillai")

p.asym(correlacao, n, p, q, tstat="Roy")

# An assymptotic test on the size of the correlations indicates that only the

# first two canonical correlations are significantly different from zero.

# Since the first canoncial correlation is so much larger than the second,

# the remainder of the analysis will focus on the first component.

# In simple terms, the rest of the analysis is geared towards understanding

# how the matrices relate to each other, via this variate.

## ----

## Correlacao entre as cargas canonicas e as variaveis originais

## ----

# Correlacao entre as variaveis independentes e suas variaveis canonicas

mod.cc$scores$corr.X.xscores

# Correlacao entre as variaveis dependentes e suas variaveis canonicas

mod.cc$scores$corr.Y.yscores

# cargas canonicas para a variavel estatistica independente

kable(mod.cc$scores$corr.X.xscores,

col.names = c("U1", " U2",

"U3", " U4"))

# cargas canonicas para a variavel estatistica dependente

kable(mod.cc$scores$corr.Y.yscores,

col.names = c("U1", " U2",

"U3", " U4"))

## ----

## Total de variancia explicaa por cada variavel canonica

## ----

## cargas canonicas

loadings = comput(X, Y, mod.cc)

## cargas canonicas

## corelacao entre a variaveis canonicas e os bancos

loadings[3:6]

## pvte uk: independentes

pvte.u = (colSums((loadings$corr.X.xscores)^2)) / (dim(X)[2])\*100

## pvte vk: dependentes

pvte.v = (colSums((loadings$corr.Y.yscores)^2)) / (dim(Y)[2])\*100

print(pvte.u)

## PVTE U1 = 27.516544 ==> primeira variavel canonica U1 explica 27,61%

## da variancia original do grupo de variaveis independnetes

## PVTE U2 = 8.224648 ==> segunda variavel canonica U1 explica 8,22%

## da variancia original do grupo de variaveis independnetes

print(pvte.v)

## PVTE V1 = 85.4406 ==> primeira variavel canonica V1 explica 85,47% da

## variancia original do grupo de variaveis dependnetes

## PVTE V2 = 14.55938 ==> segunda variavel canonica V1 explica 14,56% da

## variancia original do grupo de variaveis dependnetes

## PVTE U1 = 27.516544 ==> primeira variavel canonica U1 explica 27,61%

## da variancia original do grupo de variaveis independnetes

## PVTE U2 = 8.224648 ==> segunda variavel canonica U1 explica 8,22%

## da variancia original do grupo de variaveis independnetes

porcent.u = sort(pvte.u, decreasing = T)

names(porcent.u) = c('1', '2', '3', '4')

jpeg("variancia.explicada.u.jpeg", width = 600, height = 400)

pareto.chart(porcent.u,

main = '',

ylab = "Percentual de variância explicada",

xlab = "Componente",

ylab2 = "Percentual de variância explicada acumulada")

dev.off()

porcent.v = sort(pvte.v, decreasing = T)

names(porcent.v) = c('1', '2', '3', '4')

jpeg("variancia.explicada.v.jpeg", width = 600, height = 400)

pareto.chart(porcent.v,

main = '',

ylab = "Percentual de variância",

xlab = "Componente",

ylab2 = "Percentual de variância acumulada")

dev.off()

## ----

## Grafico dos loadings e dos individuos

## ----

jpeg("loadins.individuos.jpeg", width = 800, height = 400)

plt.cc(mod.cc,var.label = TRUE)

dev.off()

## ----

## Indicie de redudancia

## ----

can\_redunds = redundancy(candisc::cancor(X,Y))

can\_redunds

# X Plot

t0<- tibble(Var= c(c("X1", "X2", "X3", "X4"), c("X1", "X2", "X3", "X4")),

X1= c(loadings$corr.X.xscores[,1], loadings$corr.X.yscores[,1]),

Loading= c(rep("X-X Loading", 4), rep("X-Y Cross-loading", 4)))

## Loadings and Cross-Loadings for X

ggplot(t0, aes(x= X1, y= fct\_reorder(Var, -X1), fill= Loading)) +

geom\_col(size= 0.2, color= "grey50") +

geom\_text(aes(label= sprintf("%2.1f", X1\*100)), hjust= -0.6)+

facet\_wrap(~ Loading, scales= "free\_x") +

scale\_x\_continuous(limits = c(-0.7, 0.6)) +

theme\_bw() +

theme(legend.position = "None",

strip.background = element\_rect(fill= "grey30"),

strip.text = element\_text(color="white", face= "bold")) +

scale\_fill\_manual(values= c("#a3dca3", "#b3d9ff")) +

labs(x="", y= "Variável")

ggsave("loadings.cross.loading.X.jpeg", width = 10,

height = 5, dpi = 300)

# Y Plot

t1<- tibble(Var= c(c("Y1", "Y2", "Y3", "Y4"), c("Y1", "Y2", "Y3", "Y4")),

X1= c(loadings$corr.Y.yscores[,1], loadings$corr.Y.xscores[,1]),

Loading= c(rep("Y-Y Loading", 4), rep("Y-X Cross-loading", 4)))

## Loadings and Cross-Loadings for Y (Liver)

ggplot(t1, aes(x= X1, y= fct\_reorder(Var, -X1), fill= Loading)) +

geom\_col(size= 0.2, color= "grey50") +

geom\_text(aes(label= sprintf("%2.1f", X1\*100)), hjust= -0.6)+

facet\_wrap(~ Loading, scales= "free\_x") +

scale\_x\_continuous(limits = c(-0.9, 0.6)) +

theme\_bw() +

theme(legend.position = "None",

strip.background = element\_rect(fill= "grey30"),

strip.text = element\_text(color="white", face= "bold")) +

scale\_fill\_manual(values= c( "#b3d9ff", "#a3dca3")) +

labs(x="", y= "Variável")

ggsave("loadings.cross.loading.Y.jpeg", width = 10,

height = 5, dpi = 300)