UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

ANÁLISE DE CORRELAÇÃO CANÔNICA: ESTRUTURAÇÃO TEÓRICA E APLICAÇÕES EM ESTATÍSTICA AMBIENTAL

Autora: Mariana Mizutani Ribeiro
Orientadora: Jandyra Maria Guimarães Fachel

Monografia apresentada para a obteção
do grau de Bacharel em estatística

Porto Alegre, Janeiro de 2004

AGRADECIMENTOS

O tempo que passei na universidade ficará guardado em minha alma para sempre com muito carinho, pois foi uma época em que cresci em todos os sentidos e na qual conheci pessoas muito importantes para mim e para o meu futuro. Quero agradecer a todas as pessoas que contribuíram de alguma forma para eu me tornar quem sou hoje.

Às pesquisadoras Elba C. Teixeira, Daniela M. Migliavacca e Claúdia Braga da FEPAM que cederam os dados, me auxiliaram na parte ambiental e conviveram comigo durante o meu estágio curricular.

À todos os professores que contribuiram para o meu aprendizado.

Aos colegas de aula, que passaram comigo muitos fins-de-semana estudando e que agora, ou muito em breve, serão meus colegas de trabalho.

Às minhas queridas amigas Renata e Krisley que sempre estiveram presentes, mesmo nos momentos difíceis.

Aos meus amados irmãos por serem meus eternos amigos e agüentarem meus chiliques.

Aos meus pais pelo apoio e incentivo que me deram durante todo o curso e mesmo antes de ingressar na universidade.

Ao professor João Riboldi que com muita paciência respondeu às minhas muitas perguntas e que sempre teve boa vontade em passar o seu conhecimento.

À minha inesquecível orientadora, professora e amiga Jandyra Maria Guimarães Fachel pelo incentivo incondicional que sempre me deu, desde o início do curso.

RESUMO

A Análise de Correlação Canônica é uma técnica estatística multivariada utilizada quando se quer identificar e quantificar a relação entre dois conjuntos de variáveis. A técnica é apropriada para variáveis quantitativas e tem sido utilizada, principalmente, na área ambiental.

Não era uma técnica muito utilizada por ser de difícil interpretação e por necessitar de recursos computacionais mais avançados, mas com o atual desenvolvimento de tais recursos o uso da técnica vem ganhando espaço em outras áreas do conhecimento.

Com o crescimento de sua utilização, pesquisadores estão estudando cada vez mais a técnica e desenvolvendo melhorias para a mesma, aprimorando e facilitando as formas de interpretação dos resultados. Também pesquisadores usuários da técnica têm tornado público um número maior de aplicações não só na área ambiental mas também em Psicologia, Sociologia, Engenharia, Administração, além da Biologia, Ecologia e Meteorologia.

As correlações canônicas são calculadas a partir de operações matriciais onde são encontradas combinações lineares de cada conjunto de variáveis de forma que as correlações entre os conjuntos sejam maximizadas.

Apresentamos neste trabalho o desenvolvimento teórico da técnica e um exemplo na área ambiental. Também apresentamos os principais *softwares* estatísticos que disponibilizam a técnica de Análise de Correlação Canônica.

SUMÁRIO

1 INTRODUÇÃO

Cada vez mais o Homem vem percebendo que os fenômenos que ocorrem na natureza envolvem múltiplas variáveis e que estas podem estar predizendo não só um fenômeno de interesse mas muitos outros, ou seja, para predizer um acontecimento muitas variáveis podem estar presentes e estas podem estar influenciando vários outros acontecimentos. Estendendo esta forma de pensar é possível imaginar que outros tipos de ocorrência, não só os que ocorrem na Natureza, também se dão desta forma, envolvendo conjuntos de variáveis.

No início do séc. XX, mais ou menos na década de 30, H. Hotteling começou a desenvolver uma técnica para resolver este tipo de problema, que foi chamada de Análise de Correlação Canônica (CANCORR) e possui grande importância prática.

A motivação para realização deste trabalho foi o pouco conhecimento que se tem sobre essa técnica e principalmente a dificuldade de interpretação da mesma.

O objetivo principal deste trabalho é explorar a teoria através de um exemplo didático e tentar deixar mais clara a interpretação dos resultados através de um problema aplicado. Como objetivo secundário temos a de apresentar alguns softwares estatísticos que realizam este tipo de análise.

Para fazer um estudo teórico sobre a CANCORR foi feita uma revisão bibliográfica da literatura para saber o que já existe sobre a teoria, aplicação e interpretação da CANCORR. Foi possível perceber que não existem muitos materiais sobre a técnica e os que existem são incompletos em algum aspecto.

Johnson e Wichern (2002) explica de forma completa a teoria mas não aborda a questão das suposições do modelo e da interpretação da técnica. Já Hair *et al.* (1998) traz uma explicação dos termos necessários sem se preocupar muito com a teoria deixando, muitas vezes, um pouco vagos os termos utilizados. Outros livros foram utilizados para auxílio na construção da estruturação teórica do trabalho. São eles: Cooley e Lohnes (1971), Press (1982), Davies (1986) e Kendall e Stuart (1967). Legendre e Legendre (1998) traz a aplicação de técnicas multivariadas à ecologia e pouco aborda da técnica em si. Trata sobre análises canônicas em geral.

Artigos aplicados trazem exemplos principalmente em áreas que envolvam dados ambientais, onde muitos deles relacionam dados meteorológicos com dados físico-químicos.

As revistas que possuem esse tipo de artigo são, geralmente, revistas de climatologia, como por exemplo Friederichs e Hence (2003).

Outros artigos explicam detalhes não encontrados nos livros, como por exemplo, que é possível utilizar a correlação canônica em relações que envolvam variáveis quantitativas e qualitativas. Neste caso, a análise passa a ser uma Análise Canônica Parcial, baseada na matriz generalizada de covariâncias ajustadas, que possui algumas restrições. Mas, é bom salientar que no caso de termos apenas variáveis categóricas a Análise de Correlação Canônica não é aplicável.

A estrutura deste trabalho apresenta no Capítulo 1 a Introdução. No Capítulo 2 a teoria e a ilustração da mesma com um exemplo. Já no Capítulo 3 foi abordada a aplicação da técnica em um problema ambiental. O mesmo banco de dados foi utilizado no exemplo do Capítulo 2 e na aplicação do Capítulo 3. Por finalidades didáticas, apenas algumas das variáveis que foram utilizadas no Capítulo 3 foram utilizadas nos exemplos do Capítulo 2 com intuito de servir apenas como auxílio para compreensão da técnica, podendo ser generalizado no caso de mais variáveis. No capítulo 4 foram comparados *softwares* estatísticos com o objetivo de auxiliar os usuários. No capítulo 5 foram apresentadas as conclusões do trabalho.

Para entender a análise de correlação canônica é preciso conhecer alguns termos-chave como: variáveis e funções canônicas, índice de redundância, cargas canônicas, pesos canônicos, etc. Esses termos serão devidamente explicados e para tal foi utilizado um exemplo prático envolvendo dados gentilmente fornecidos pela Fepam – Fundação Estadual de Proteção Ambiental – que se referem a um estudo sobre qualidade de água de chuva em POA. O conjunto de dados $X^{(1)}$ representa o conjunto de variáveis químicas obtidas de coletas de água de chuvas e o conjunto $X^{(2)}$ representa as variáveis meteorológicas fornecidas pelo Aeroporto Salgado Filho nos dias em que foram feitas as coletas de água.

2 ANÁLISE DE CORRELAÇÃO CANÔNICA (CANCORR)

Esse capítulo foi baseado em Johnson e Wichern (2002) e Hair *et al.* (1998). Para os cálculos envolvidos nos exemplos de cada seção foi utilizado o *software* de Matemática Maple versão 5.

2.1 Definição

A Análise de Correlação Canônica (CANCORR) tem como objetivo principal identificar e quantificar a relação entre dois conjuntos de variáveis e como objetivo secundário predizer múltiplas variáveis através de outras múltiplas variáveis.

Para compreendermos o que é a Análise de Correlação Canônica podemos pensá-la como uma extensão da análise de regressão múltipla que tem como equação básica a Eq(1). Na Análise de Correlação Canônica os pesos das combinações lineares são encontrados de forma que maximizem a correlação entre os dois conjuntos de variáveis. Esta técnica é uma técnica multivariada geral, envolvendo princípios utilizados em outras técnicas multivariadas. É correspondente à Análise Fatorial no que diz respeito à criação da composição das variáveis. Também lembra a Análise Discriminante na habilidade de determinar dimensões independentes (similar às funções discriminantes) para cada conjunto de variáveis, nesta situação com o objetivo de produzir a máxima correlação entre as dimensões. Apesar de ser um modelo multivariado geral, só pode ser utilizada para identificar relacionamentos lineares

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_n X_n \qquad (1)$$

entre as variáveis.

Como em qualquer análise estatística, este modelo possui algumas suposições que devem ser atendidas para se obter um resultado confiável.

Na correlação canônica, o tamanho da amostra e a necessidade de um número suficiente de observações por variável podem ter um grande impacto sobre a análise. Pesquisadores tendem a colocar muitas variáveis nos dois conjuntos de dados sem levar em conta que isso afeta o tamanho de amostra necessário. Tamanhos de amostra muito pequenos

não irão representar a correlação de forma correta, confundindo qualquer relação importante. Tamanhos de amostra muito grandes irão ter uma tendência a sempre indicar significâncias estatísticas, mesmo quando não existir significância na prática. O pesquisador deve ser encorajado a manter pelo menos 10 observações para cada variável a fim de ter um modelo adequado.

A CANCORR é uma técnica de difícil interpretação, relativamente desconhecida e pouco explorada nos livros. Um dos motivos de ser desconhecida é por necessitar de recursos computacionais que não eram disponíveis até algum tempo atrás. Com as constantes melhorias de tais recursos, o seu uso vêm crescendo cada vez mais em pesquisas científicas. Mas pela dificuldade de interpretação, muitos pesquisadores viam a correlação canônica como a última escolha, ou seja, só a utilizavam quando todas as outras técnicas não pudessem ser utilizadas. Mas, em situações com múltiplas variáveis dependentes e independentes, a correlação canônica é a mais apropriada e poderosa técnica.

O princípio básico da CANCORR é encontrar uma combinação linear em cada conjunto de variáveis, atribuindo pesos a cada uma delas, de tal forma que a correlação entre as combinações lineares seja maximizada, ou seja, se aplicarmos esses pesos às variáveis chegaremos a um escore para o primeiro conjunto de variáveis e um escore para o segundo conjunto de variáveis que terão correlação máxima. Para melhor compreensão podemos atribuir um nome a cada conjunto, por exemplo: variáveis químicas e meteorológicas, satisfação e condição, etc.

A CANCORR vem sendo aplicada principalmente nas áreas que tratam de dados ambientais, em engenharia e em pesquisas mercadológicas e vêm ganhando espaço em outras áreas.

2.2 Suposições do Modelo

Sempre que se fala em análise estatística deve-se estar ciente das suposições que precisam ser satisfeitas para que os resultados sejam confiáveis. Não poderia ser diferente com a CANCORR. Por ser um modelo mais sofisticado possui as seguintes suposições:

2.2.1 Linearidade do modelo

O coeficiente de correlação entre quaisquer duas variáveis é baseado em relações lineares. Se a relação não for linear, uma ou ambas as variáveis deverão ser transformadas.

A correlação canônica é a correlação entre duas variáveis canônicas. Se as relações entre essas variáveis não for linear, a técnica não conseguirá captar o relacionamento correto entre as variáveis. Apesar da CANCORR ser um modelo multivariado geral, só pode ser utilizado para identificar e quantificar relações lineares.

2.2.2 Multicolinearidade

Quando existe multicolinearidade entre as variáveis utilizadas na CANCORR, a técnica começa a ter problemas na hora de isolar o impacto de uma única variável, uma vez que duas ou mais variáveis estão explicando a mesma coisa, o que também ocorre em Regressão Linear múltipla.

2.2.3 Normalidade

A normalidade das variáveis não é necessária na Correlação Canônica, mas o é para realização de testes de significância. Pela dificuldade de testar se a distribuição é normal multivariada uma alternativa é testar se cada variável é normal univariada. As variáveis que não possuem distribuição normal poderão ser transformadas.

2.3 Variáveis, Funções e Correlações Canônicas Populacionais:

Correlação canônica é o grau da relação entre dois conjuntos de variáveis. Essa correlação é medida entre duas variáveis sintetizadas a partir das variáveis originais, chamadas de ***variáveis canônicas***, que nada mais são do que combinações lineares das

variáveis originais, ou seja, são atribuídos pesos às variáveis originais de tal forma que a correlação entre essas combinações lineares seja a maior possível. Cada variável será relacionada a todas as variáveis dos dois conjuntos fazendo com que a retirada de uma variável afete a solução.

As ***Funções canônicas*** são compostas de duas variáveis canônicas, uma representando o primeiro conjunto de variáveis e a outra representado o segundo conjunto. Por exemplo, as variáveis canônicas $a_1X_1^{(1)} + a_2X_2^{(1)} + ... + a_pX_p^{(1)}$ e $b_1X_1^{(2)} + b_2X_2^{(2)} + ... + b_qX_q^{(2)}$ formam a primeira função canônica, onde a_i e b_i são os pesos atribuídos a cada uma das variáveis.

Mas como encontrar matematicamente esses pesos?

O nosso objetivo é demonstrar a relação entre os dois conjuntos de variáveis através de um número de correlações menor do que as pq correlações (ou covariâncias) iniciais.

Para encontrar a relação entre os dois conjuntos devemos encontrar as combinações lineares (variáveis canônicas) que maximizem a correlação entre os conjuntos .

Digamos que o grupo $\mathbf{X}^{(1)}$ possui p variáveis e o grupo $\mathbf{X}^{(2)}$ possui q variáveis e $p \leq q$. Nos vetores $\mathbf{X}^{(1)}$ e $\mathbf{X}^{(2)}$ temos:

$$E(\mathbf{X}^{(1)}) = \boldsymbol{\mu}^{(1)} \qquad Cov(\mathbf{X}^{(1)}) = \sum_{11}$$
$$E(\mathbf{X}^{(2)}) = \boldsymbol{\mu}^{(2)} \qquad Cov(\mathbf{X}^{(2)}) = \sum_{22}$$
$$Cov(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \sum_{12} = \sum_{21}^{'}$$

Se considerarmos $\mathbf{X}^{(1)}$ e $\mathbf{X}^{(2)}$ conjuntamente teremos:

$$\underset{((p+q)\times 1)}{\mathbf{X}} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix} = \begin{bmatrix} X_1^{(1)} \\ X_2^{(1)} \\ \vdots \\ X_p^{(1)} \\ X_1^{(2)} \\ X_2^{(2)} \\ \vdots \\ Y_q^{(2)} \end{bmatrix}$$

que possui médias:

$$\underset{((p+q)\times 1)}{\boldsymbol{\mu}} = \begin{bmatrix} E(\mathbf{X}^{(1)}) \\ E(\mathbf{X}^{(2)}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}$$

e matriz de covariância

$$\underset{(p+q)\times(p+q)}{\sum} = E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})' = \begin{bmatrix} \underset{(p\times p)}{\sum_{11}} & \underset{(p\times q)}{\sum_{12}} \\ \underset{(q\times p)}{\sum_{21}} & \underset{(q\times q)}{\sum_{22}} \end{bmatrix} \qquad (2)$$

As combinações lineares são dadas por:

$$\begin{aligned} \mathrm{U} &= \boldsymbol{a}' \boldsymbol{X}^{(1)} \\ \mathrm{V} &= \boldsymbol{b}' \boldsymbol{X}^{(2)} \end{aligned} \qquad (3)$$

onde $\boldsymbol{a}$ e $\boldsymbol{b}$ são vetores dos pesos.

Seja A = CX onde X é a variável aleatória e C é uma constante, temos:

$$\begin{aligned} \mu_A &= E(A) = E(\mathbf{CX}) = C\mu_X \\ \sum_A &= Cov(A) = Cov(\mathbf{CX}) = C\sum_X C' \end{aligned} \tag{4}$$

Utilizando Eq(3) e Eq(4) obtemos:

$$\begin{aligned} Var(U) &= \boldsymbol{a}' Cov(\mathbf{X}^{(1)})\boldsymbol{a} = \boldsymbol{a}'\sum_{11}\boldsymbol{a} \\ Var(V) &= \boldsymbol{b}' Cov(\mathbf{X}^{(2)})\boldsymbol{b} = \boldsymbol{b}'\sum_{22}\boldsymbol{b} \\ Cov(U,V) &= \boldsymbol{a}' Cov(\mathbf{X}^{(1)},\mathbf{X}^{(2)})\boldsymbol{b} = \boldsymbol{a}'\sum_{12}\boldsymbol{b} \end{aligned}$$

A correlação, que deve ser maximizada, é dada pela covariância dividida pelos desvios padrões então:

$$Corr(U,V) = \frac{\boldsymbol{a}'\sum_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}'\sum_{11}\boldsymbol{a}}\sqrt{\boldsymbol{b}'\sum_{22}\boldsymbol{b}}} \tag{5}$$

$$\max_{a,b} Corr(U_i,V_i) = \rho_i^*$$

Como $p<q$, teremos p correlações canônicas que ao serem maximizadas irão gerar p variáveis canônicas em $X^{(1)}$ chamadas de U_i e p variáveis canônicas em $X^{(2)}$ chamadas de V_i.

Então, o primeiro par de variáveis canônicas (primeira função canônica), é o par de combinações lineares U_1 e V_1 com variâncias unitárias, que maximiza a correlação da Eq(5).

A segunda função canônica, é o par de combinações lineares U_2 e V_2 com variâncias unitárias, que maximiza a correlação da Eq(5) dentre todas as possíveis funções canônicas não correlacionadas com a primeira.

E assim segue até o par p, pois são encontradas tantas funções canônicas quantos forem o menor número entre p e q. Note que a p-ésima função canônica deve ser não correlacionada com as *(p-1)* funções canônicas anteriores.

Res 1:

Suponhamos que $p \leq q$ e $\mathbf{X}^{(1)}$ e $\mathbf{X}^{(2)}$ são os conjuntos de variáveis aleatórias com $Cov(\mathbf{X}^{(1)}) = \underset{(p\times p)}{\sum_{11}}$, $Cov(\mathbf{X}^{(2)}) = \underset{(q\times q)}{\sum_{22}}$ e $Cov(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \underset{(p\times q)}{\sum_{12}}$, que são partes de $\sum$, onde $\sum$ é uma matriz de *rank* completo. Os vetores de coeficientes $\underset{(p\times 1)}{\boldsymbol{a}}$ e $\underset{(q\times 1)}{\boldsymbol{b}}$ formam as combinações lineares $U = \boldsymbol{a}'\mathbf{X}^{(1)}$ e $V = \boldsymbol{b}'\mathbf{X}^{(2)}$. Então

$$\max_{a,b} Corr(U,V) = \rho_1^*$$

O k-ésimo par de variáveis canônicas (k-ésima função canônica) é dado por:

$$U_k = \underbrace{\boldsymbol{e}_k' \sum_{11}^{-1/2}}_{a'_k} \mathbf{X}^{(1)} \quad e \quad V_k = \underbrace{\boldsymbol{f}_k' \sum_{22}^{-1/2}}_{b'_k} \mathbf{X}^{(2)} \tag{6}$$

$$\max_{a,b} Corr(U_k, V_k) = \boldsymbol{\rho}_k^*$$

Então $\rho_1^{*2} \geq \rho_2^{*2} \geq .... \geq \rho_p^{*2}$ são os p maiores autovalores da matriz $\sum_{11}^{-1/2} \sum_{12} \sum_{22}^{-1} \sum_{21} \sum_{11}^{-1/2}$ com $e_1, e_2, ..., e_p$ autovetores associados e $f_1, f_2, ..., f_p$ calculados a partir da relação $\sum_{22}^{-1/2} \sum_{21} \sum_{11}^{-1/2} e_i$.

As matrizes $\sum_{11}^{-1} \sum_{12} \sum_{22}^{-1} \sum_{21}$ e $\sum_{22}^{-1} \sum_{21} \sum_{11}^{-1} \sum_{12}$ geralmente, não são simétricas.

Algumas propriedades das variáveis canônicas:

$Var(U_k) = Var(V_k) = 1$
$Cov(U_k, U_l) = Corr(U_k, U_l) = 0, \; k \neq l$
$Cov(V_k, V_l) = Corr(V_k, V_l) = 0, \; k \neq l$
$Cov(U_k, V_l) = Corr(U_k, V_l) = 0, \; k \neq l$

para k e l = 1,2,....,p.

Mas o que acontece se as variáveis forem padronizadas?

Se isso acontece, no lugar de $\mathbf{X}^{(1)}$ e $\mathbf{X}^{(2)}$ temos $Z^{(1)} = [Z_1^{(1)}, Z_2^{(1)}, ..., Z_p^{(1)}]'$ e $Z^{(2)} = [Z_1^{(2)}, Z_2^{(2)}, ..., Z_p^{(2)}]'$ respectivamente. Então, as variáveis canônicas ficam

$$\begin{aligned} U_k &= \boldsymbol{a}_k' \mathbf{Z}^{(1)} = \boldsymbol{e}_k' \rho_{11}^{1/2} \mathbf{Z}^{(1)} \\ V_k &= \boldsymbol{b}_k' \mathbf{Z}^{(2)} = \boldsymbol{f}_k' \rho_{22}^{-1/2} \mathbf{Z}^{(2)} \end{aligned} \tag{7}$$

onde, $Cov(\mathbf{Z}^{(1)}) = \boldsymbol{\rho}_{11}$. $Cov(\mathbf{Z}^{(2)}) = \boldsymbol{\rho}_{22}$, $Cov(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}) = \boldsymbol{\rho}_{12} = \boldsymbol{\rho}_{21}'$ e e_k e f_k são os autovetores de $\rho_{11}^{-1/2}\rho_{12}\rho_{22}^{-1}\rho_{21}\rho_{11}^{-1/2}$ e $\rho_{22}^{-1/2}\rho_{21}\rho_{11}^{-1}\rho_{12}\rho_{22}^{-1/2}$ respectivamente e satisfazem $Corr(U_k, V_k) = \rho_k^*$, $k=1,2,...,p$.

Nota-se que

$$\begin{aligned} \boldsymbol{a}_k'(\mathbf{X}^{(1)} - \boldsymbol{\mu}^{(1)}) &= a_{k_1}(X_1^{(1)} - \mu_1^{(1)}) + a_{k_2}(X_2^{(1)} - \mu_2^{(1)}) + ... + a_{k_p}(X_p^{(1)} - \mu_p^{(1)}) \\ &= a_{k_1}\sqrt{\sigma_{11}^2}\frac{(X_1^{(1)} - \mu_1^{(1)})}{\sqrt{\sigma_{11}^2}} + a_{k_2}\sqrt{\sigma_{22}^2}\frac{(X_2^{(1)} - \mu_2^{(1)})}{\sqrt{\sigma_{22}^2}} + ... + a_{k_p}\sqrt{\sigma_{pp}^2}\frac{(X_p^{(1)} - \mu_p^{(1)})}{\sqrt{\sigma_{pp}^2}} \\ &= a_{k_1}\sqrt{\sigma_{11}^2}Z_1^{(1)} + a_{k_2}\sqrt{\sigma_{22}^2}Z_2^{(1)} + ... + a_{k_p}\sqrt{\sigma_{pp}^2}Z_p^{(1)} \end{aligned}$$

onde, $Var(X_i^{(1)}) = \sigma_{ii}^2$, $i = 1,2,...,p$. Portanto, os coeficientes canônicos para as variáveis padronizadas, $Z_i^{(1)} = (X_i^{(1)} - \mu_i^{(1)})/\sqrt{\sigma_{ii}^2}$, são simplesmente relacionados aos coeficientes canônicos das variáveis originais. Especificamente se $\boldsymbol{a}_k'$ é o vetor dos coeficientes da k-ésima variável canônica U_k, então $\boldsymbol{a}_k'\mathbf{V}_{11}^{1/2}$ é o vetor dos coeficientes da k-ésima variável canônica padronizada $\mathbf{Z}^{(1)}$. Onde $\mathbf{V}_{11}^{1/2}$ é a matriz diagonal de elementos $\sqrt{\sigma_{ii}^2}$. Similarmente $\boldsymbol{b}_k'\mathbf{V}_{11}^{1/2}$ é o vetor de coeficientes das variáveis canônicas padronizadas $\mathbf{Z}^{(2)}$.

O que se conclui é que a correlação canônica não é alterada pela padronização e o sistema para encontrar os coeficientes $\boldsymbol{a}_k, \boldsymbol{b}_k$ terá uma única solução se a matriz $\sum$ tiver rank completo.

Para facilitar os cálculos alguns autores sugerem que se obtenha as correlações canônicas a partir de

$$|\sum_{11}^{-1}\sum_{12}\sum_{22}^{-1}\sum_{21} - \rho^{*2} I| = 0 \qquad (8)$$

e daí obter os autovetores.

Então podemos reescrever as variáveis canônicas como

$$\boldsymbol{a}_k = \sum_{11}^{-1/2} \boldsymbol{e}_k \quad \text{e} \quad \boldsymbol{b}_k = \sum_{22}^{-1/2} \boldsymbol{f}_k$$

2.4 Variáveis, Funções e Correlações Canônicas Amostrais:

Podemos representar uma amostra aleatória de n observações em cada uma das $(p+q)$ variáveis de $\mathbf{X}^{(1)}$ e $\mathbf{X}^{(2)}$ através do vetor $n \times (p+q)$

$$\mathbf{X} = \begin{bmatrix} x_{11}^{(1)} & x_{12}^{(1)} & \cdots & x_{1p}^{(1)} & x_{11}^{(2)} & x_{12}^{(2)} & \cdots & x_{1q}^{(2)} \\ x_{21}^{(1)} & x_{22}^{(1)} & \cdots & x_{2p}^{(1)} & x_{21}^{(2)} & x_{22}^{(2)} & \cdots & x_{2q}^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1}^{(1)} & x_{n2}^{(1)} & \cdots & x_{np}^{(1)} & x_{n1}^{(2)} & x_{n2}^{(2)} & \cdots & x_{nq}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathrm{x}_1^{(1)'} & \mathrm{x}_1^{(2)'} \\ \vdots & \vdots \\ \mathrm{x}_n^{(1)'} & \mathrm{x}_n^{(2)'} \end{bmatrix}$$

O vetor de médias amostrais pode ser representado como:

$$\underset{(p+q)\times 1}{\bar{\mathrm{x}}} = \begin{bmatrix} \bar{\mathrm{x}}^{(1)} \\ \\ \bar{\mathrm{x}}^{(2)} \end{bmatrix} \quad \text{onde} \quad \begin{aligned} \bar{\mathrm{x}}^{(1)} &= \frac{1}{n}\sum_{j=1}^{n} \mathrm{x}_j^{(1)} \\ \bar{\mathrm{x}}^{(2)} &= \frac{1}{n}\sum_{j=1}^{n} \mathrm{x}_j^{(2)} \end{aligned}$$

Similarmente, a matriz de covariância amostral pode ser arranjada de forma análoga a populacional Eq(3)

$$\underset{(p+q)\times(p+q)}{\mathbf{S}} = \begin{bmatrix} \underset{p\times p}{\mathbf{S}_{11}} & \underset{p\times q}{\mathbf{S}_{12}} \\ \underset{q\times p}{\mathbf{S}_{21}} & \underset{q\times q}{\mathbf{S}_{22}} \end{bmatrix}$$

onde

$$\mathbf{S}_{kl} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathrm{x}_j^{(k)} - \bar{\mathrm{x}}^{(k)})(\mathrm{x}_j^{(l)} - \bar{\mathrm{x}}^{(l)})' \quad k, l = 1, 2 \tag{9}$$

As combinações lineares

$$\begin{aligned} \hat{U} &= \hat{\boldsymbol{a}}'\mathrm{x}^{(1)} \\ \hat{V} &= \hat{\boldsymbol{b}}'\mathrm{x}^{(2)} \end{aligned} \tag{10}$$

tem correlação amostral

$$r_{\hat{U},\hat{V}} = \frac{\hat{\boldsymbol{a}}'\mathbf{S}_{12}\hat{\boldsymbol{b}}}{\sqrt{\hat{\boldsymbol{a}}'\mathbf{S}_{11}\hat{\boldsymbol{a}}}\sqrt{\hat{\boldsymbol{b}}'\mathbf{S}_{22}\hat{\boldsymbol{b}}}} \tag{11}$$

O primeiro par (amostral) de variáveis canônicas (primeira função canônica amostral) é o par de combinações lineares $\hat{U}_1, \hat{V}_1$ que possui variância amostral unitária e maximiza a correlação da Eq(11).

Em geral, a k-ésima função canônica (amostral) é o par de combinações lineares $\hat{U}_k, \hat{V}_k$, com variâncias amostrais unitárias, que maximiza a correlação da Eq(11) dentre todas as possíveis funções canônicas (amostrais) não correlacionadas com as *k-1* funções canônicas anteriores. A correlação entre $\hat{U}_k, \hat{V}_k$ é chamada de k–ésima correlação canônica amostral.

As variáveis canônicas amostrais e as correlações amostrais podem ser obtidas das matrizes de covariâncias amostrais $\mathbf{S}_{11}$, $\mathbf{S}_{12}=\mathbf{S}_{21}'$ e $\mathbf{S}_{22}$ e se parecem muito com o caso populacional descrito no Res 1.

Res 2:

Sejam $\hat{\rho}_1^{*2} \geq \hat{\rho}_2^{*2} \geq \ldots . \hat{\rho}_p^{*2}$ os primeiros p autovalores ordenados de $\mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1/2}$ com autovetores associados $\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \ldots, \hat{\boldsymbol{e}}_p$, onde $\mathbf{S}_{kl}$ está definido na Eq(9) e $p \leq q$. Sejam $\hat{\boldsymbol{f}}_1, \hat{\boldsymbol{f}}_2, \ldots, \hat{\boldsymbol{f}}_p$ os autovetores associados a $\mathbf{S}_{22}^{-1/2}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1/2}$, onde $\hat{\boldsymbol{f}}_k$ deve ser obtido da relação $\hat{\boldsymbol{f}}_k = \left(1/\hat{\rho}_k^*\right)\mathbf{S}_{22}^{-1/2}\mathbf{S}_{21}\mathbf{S}_{11}^{-1/2}\hat{\boldsymbol{e}}_k$, $k= 1,2,\ldots,p$.

Então o k-ésimo par de variáveis canônicas amostrais é:

$$\hat{U}_k = \underbrace{\hat{\boldsymbol{e}}_k'\mathbf{S}_{11}^{-1/2}}_{\hat{a}_k'}\mathbf{x}^{(1)} \quad e \quad \hat{V}_k = \underbrace{\hat{\boldsymbol{f}}_k'\mathbf{S}_{22}^{-1/2}}_{\hat{b}_k'}\mathbf{x}^{(2)} \tag{12}$$

onde $\mathbf{x}^{(1)}$ e $\mathbf{x}^{(2)}$ são os valores de $\mathbf{X}^{(1)}$ e $\mathbf{X}^{(2)}$ para uma particular amostra. Da mesma forma, o primeiro par de variáveis canônicas tem correlação amostral máxima.

$$r_{\hat{U}_1,\hat{V}_1} = \hat{\rho}_1^*$$

e o k-ésimo par

$$r_{\hat{U}_k,\hat{V}_k} = \hat{\rho}_k^*$$

que é a maior correlação possível dentre as combinações lineares não-correlacionadas com as *k-1* funções canônicas anteriores.

As quantidades $\hat{\rho}_1^* \geq \hat{\rho}_2^* \geq \ldots . \hat{\rho}_p^*$ são as correlações canônicas amostrais.

Se $p>rank\ (\mathbf{S}_{12})= p_1$, as correlações canônicas não nulas serão $\hat{\rho}_1^* \geq \hat{\rho}_2^* \geq \ldots . \hat{\rho}_{p_1}^*$.

Prova: A prova do Res 2 segue a prova do Res 1 com $\mathbf{S}_{kl}$ substituído por $\sum_{kl}$, k, l =1,2. Ambas encontram-se em Johnson e Wichern (2002).

As variáveis canônicas amostrais possuem variâncias unitárias

$$\mathbf{S}_{\hat{U}_k,\hat{U}_k} = \mathbf{S}_{\hat{V}_k,\hat{V}_k} = 1 \tag{13}$$

e suas correlações amostrais são:

$$r_{\hat{U}_k,\hat{U}_l} = r_{\hat{V}_k,\hat{V}_l} = 0 \;\; e \;\; r_{\hat{U}_k,\hat{V}_l} = 0 \quad k \neq l \tag{14}$$

A interpretação de $\hat{U}_k, \hat{V}_k$ pode ser auxiliada por outros procedimentos que serão explicados mais adiante.

2.5 Exemplo:

Utilizando-se a matriz de dados meteorológicos e de dados químicos da água de chuva de POA fornecidos pela Fundação Estadual de Proteção Ambiental (FEPAM) e fazendo com que $\mathbf{X}^{(1)}$ seja o conjunto de variáveis químicas (SO_4 e Na) e $\mathbf{X}^{(2)}$ o conjunto de variáveis meteorológicas (Velocidade dos ventos, Temperatura e Umidade) temos:

As covariâncias amostrais S:

$$\underset{5\times5}{\mathbf{S}} = \begin{bmatrix} \underset{2\times2}{\mathbf{S}_{11}} & \underset{2\times3}{\mathbf{S}_{12}} \\ \underset{3\times2}{\mathbf{S}_{21}} & \underset{3\times3}{\mathbf{S}_{22}} \end{bmatrix} = \begin{bmatrix} 285.561 & 58.780 & -10.330 & -1.999 & -11.057 \\ 58.780 & 263.120 & 9.771 & 18.121 & -35.589 \\ -10.330 & 9.771 & 14.410 & 6.483 & -16.843 \\ -1.999 & 18.121 & 6.483 & 12.822 & -12.317 \\ -11.057 & -35.589 & -16.843 & -12.317 & 49.401 \end{bmatrix}$$

A partir da matriz **S** podemos calcular:

$$\mathbf{S}_{11}^{1/2} = \begin{bmatrix} 0.060 & -0.007 \\ -0.007 & 0.063 \end{bmatrix} \qquad \mathbf{S}_{22}^{-1} = \begin{bmatrix} 0.123 & -0.029 & 0.035 \\ -0.029 & 0.109 & 0.017 \\ 0.035 & 0.017 & 0.036 \end{bmatrix}$$

A partir da matriz de covariâncias amostrais para as variáveis padronizadas (que é igual a matriz de correlações amostrais) , chamada de matriz **R**:

$$\underset{5\times5}{\mathbf{R}} = \begin{bmatrix} \underset{2\times2}{\mathbf{R}_{11}} & \underset{2\times3}{\mathbf{R}_{12}} \\ \underset{3\times2}{\mathbf{R}_{21}} & \underset{3\times3}{\mathbf{R}_{22}} \end{bmatrix} = \begin{bmatrix} 1.000 & 0.214 & -0.161 & -0.033 & -0.093 \\ 0.214 & 1.000 & 0.159 & 0.312 & -0.312 \\ -0.161 & 0.159 & 1.000 & 0.477 & -0.631 \\ -0.033 & 0.312 & 0.477 & 1.000 & -0.489 \\ -0.093 & -0.312 & -0.631 & -0.489 & 1.000 \end{bmatrix}$$

Podemos calcular:

$$\mathbf{R}_{11}^{-1/2} = \begin{bmatrix} 1.018 & -0.110 \\ -0.110 & 1.018 \end{bmatrix} \qquad \mathbf{R}_{22}^{-1} = \begin{bmatrix} 1.772 & -0.391 & 0.927 \\ -0.391 & 1.401 & 0.439 \\ 0.927 & 0.439 & 1.799 \end{bmatrix}$$

Calculando as correlações canônicas por ambas matrizes obtivemos:

$$\mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1/2} = \begin{bmatrix} .086 & .015 \\ .015 & .138 \end{bmatrix}$$

$$\mathbf{R}_{11}^{-1/2}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1/2} = \begin{bmatrix} .086 & .015 \\ .015 & .138 \end{bmatrix}$$

Ambas produzem os seguintes autovalores:

$$\hat{\rho}_1^{*2} = 0.142$$
$$\hat{\rho}_2^{*2} = 0.082$$

e, portanto as correlações canônicas são

$$\hat{\rho}_1^{*} = 0.377$$
$$\hat{\rho}_2^{*} = 0.286$$

Calculando os autovetores associados obtemos os U_i's:

$$\begin{matrix} \hat{e}_1' = [-0.259, -0.966] \\ \hat{e}_2' = [-0.966, 0.259] \end{matrix} \quad \text{pela Eq(12)} \quad \begin{matrix} \hat{U}_1 = [-0.157, \ -0.955] \\ \hat{U}_2 = [-1.011, \ 0.369] \end{matrix}$$

E a partir da relação $\hat{f}_k = (1/\hat{\rho}_k^{*})\mathbf{S}_{22}^{-1/2}\mathbf{S}_{21}\mathbf{S}_{11}^{-1/2}\hat{e}_k$ obtemos:

$$\begin{matrix} \hat{f}_1 = [0.049, \ -0.649, \ 0.759] \\ \hat{f}_2 = [0.846, \ 0.429, \ 0.314] \end{matrix} \quad \text{pela Eq(12)} \quad \begin{matrix} \hat{V}_1 = [0.479, \ -0.594, \ 0.842] \\ \hat{V}_2 = [1.102, \ 0.394, \ 0.814] \end{matrix}$$

Esses resultados também foram calculados aplicando os pesos às variáveis e calculando a correlação entre esses escores.

2.6 Interpretação das variáveis canônicas

Em geral, as variáveis canônicas são artificiais, isto é, elas geralmente não possuem significado físico. Se as variáveis canônicas forem obtidas das variáveis originais ($\mathbf{X}^{(1)}$ e $\mathbf{X}^{(2)}$),

os coeficientes canônicos ***a*** e ***b*** possuirão unidades proporcionais às unidades dos conjuntos $\mathbf{X}^{(1)}$ e $\mathbf{X}^{(2)}$. É mais aconselhável que se utilize variáveis padronizadas, pois estas passam a ter média zero e variância unitária, e os coeficientes canônicos ***a*** e ***b*** não terão unidade de medida, e deverão ser interpretados em termos das variáveis padronizadas.

Existem três formas de interpretação das variáveis canônicas. O nível de significância das raízes canônicas, a magnitude das correlações canônicas e as relações entre as variâncias das variáveis canônicas e variáveis originais. Essas relações podem ser de três tipos: os próprios ***pesos canônicos*** que estamos chamando de ***coeficientes canônicos***, *as* ***cargas canônicas*** ou ***estrutura de correlação canônica*** e as ***cargas canônicas cruzadas***.

Segundo *Hair et al. (1998)* as cargas canônicas cruzadas são preferidas e mesmo que alguns softwares não façam esse cálculo, devemos calculá-las utilizando o esquema apresentado na Figura 2.1 ou teremos de confiar nos outros métodos de interpretação. Se, ainda assim não pudermos utilizá-las, o uso da estrutura de correlação canônica é mais confiável do que o uso dos pesos canônicos por se tratar da correlação entre cada variável canônica e suas respectivas variáveis originais. Por essa razão, sempre que possível o uso das cargas canônicas cruzadas é recomendado como melhor alternativa para a interpretação do dados. Como segunda opção são as cargas canônicas e por último o os pesos canônicos. Os pesos canônicos e as cargas canônicas são medidas que estão condicionadas à amostra, ou seja, estão sujeitas a consideráveis instabilidades de uma amostra para outra e, além disso, sofrem as mesmas críticas dos coeficientes da regressão múltipla. Por exemplo, pesos (coeficientes) pequenos, numa escala padronizada, podem tanto significar que a variável é irrelevante para o modelo como pode indicar que existem altos graus de multicolinearidade entre as variáveis utilizadas.

2.6.1 Estrutura canônica e cargas cruzadas canônicas populacionais:

Apesar de serem variáveis artificiais, ou sintéticas, as variáveis canônicas podem, muitas vezes, ser "identificadas" em termos de sua importância para as variáveis originais. Essa identificação é auxiliada calculando as correlações entre as variáveis canônicas e as variáveis originais. As correlações entre as variáveis canônicas e seus respectivos conjuntos

originais são chamadas de ***estrutura canônica*** e as correlações entre as variáveis canônicas e os conjuntos de variáveis opostas são chamadas de ***cargas cruzadas***. Essas correlações devem ser interpretadas com cuidado porque elas trazem informações univariadas, ou seja, não indicam como a variável original contribui conjuntamente para a análise canônica. Por essa razão, muitos pesquisadores preferem avaliar a contribuição das variáveis originais diretamente pelos ***coeficientes canônicos calculados*** à partir das variáveis padronizadas (***coeficientes canônicos padronizados***), que não é o caso de *Hair et al.* (1998).

Seja $\underset{p\times p}{A} = [a_1, a_2, ..., a_p]'$ e $\underset{q\times q}{B} = [b_1, b_2, ..., b_q]'$, então os vetores de variáveis canônicas são:

$$\begin{aligned} \underset{(p\times 1)}{U} &= A\mathbf{X}^{(1)} \\ \underset{(q\times 1)}{V} &= B\mathbf{X}^{(2)} \end{aligned} \tag{15}$$

onde estamos interessados nas p variáveis canônicas de $\mathbf{V}$. Então:

$$Corr(\mathbf{U}_i, \mathbf{X}_k^{(1)}) = \frac{\mathrm{A}\sum_{11}}{\sqrt{Var(\mathbf{X}_k^{(1)})}} = Cov(\mathrm{A}\mathbf{X}^{(1)}, \sigma_{kk}^{-1/2}\mathbf{X}_k^{(1)}) = \frac{\mathrm{A}\sum_{11}}{\sqrt{\sigma_{kk}}}$$

Introduzindo a matriz diagonal $\underset{p\times p_{11}}{V^{-1/2}}$ com k elementos na diagonal principal ($\sigma_{kk}^{-1/2}$) temos, em termos matriciais:

$$\underset{p\times p}{\rho_{\mathrm{U},\mathrm{X}^{(1)}}} = Corr(\mathbf{U}, \mathbf{X}^{(1)}) = Cov(\mathbf{U}, \mathbf{V}_{11}^{-1/2}\mathbf{X}^{(1)}) = Cov(A\mathbf{X}^{(1)}, \mathbf{V}_{11}^{-1/2}\mathbf{X}^{(1)}) = A\sum_{11}\mathbf{V}_{11}^{-1/2}$$

Cálculos similares dos pares $(\mathbf{U},\mathbf{X}^{(2)}),(\mathbf{V},\mathbf{X}^{(1)}),(\mathbf{V},\mathbf{X}^{(2)})$ produzem:

$$\begin{aligned} \underset{p\times p}{\rho_{\mathrm{U},\mathrm{X}^{(1)}}} &= A\sum_{11}V_{11}^{-1/2} & \underset{q\times p}{\rho_{\mathrm{V},\mathrm{X}^{(1)}}} &= B\sum_{21}V_{11}^{-1/2} \\ \underset{p\times q}{\rho_{\mathrm{U},\mathrm{X}^{(2)}}} &= A\sum_{12}V_{22}^{-1/2} & \underset{q\times q}{\rho_{\mathrm{U},\mathrm{X}^{(2)}}} &= B\sum_{22}V_{22}^{-1/2} \end{aligned} \tag{16}$$

As correlações que, muitas vezes são calculadas no auxílio da interpretação das variáveis canônicas, são dadas por:

$$\begin{aligned} \rho_{U,Z^{(1)}} &= A_Z \rho_{11} \qquad & \rho_{V,Z^{(1)}} &= B_Z \rho_{21} \\ \rho_{U,Z^{(2)}} &= A_Z \rho_{12} \qquad & \rho_{V,Z^{(2)}} &= B_Z \rho_{22} \end{aligned} \tag{17}$$

onde A_Z e B_Z são as matrizes cujas linhas contém os coeficientes canônicos para os conjuntos $\mathbf{Z}^{(1)}$ e $\mathbf{Z}^{(2)}$ respectivamente. Os valores das correlações na Eq(16) e Eq(17) são os mesmos,

indicando que a padronização não afeta as correlações entre as variáveis canônicas e as variáveis originais. Ou seja, $\rho_{U,X^{(1)}} = \rho_{U,Z^{(1)}}$ e assim por diante.

Isso segue porque, por exemplo,

$$\rho_{U,X^{(1)}} = A\sum\nolimits_{11} V_{11}^{-1/2} = AV_{11}^{1/2} V_{11}^{-1/2} \sum\nolimits_{11} V_{11}^{-1/2} = A_Z \rho_{11} = \rho_{U,Z^{(1)}}.$$

Ainda de Eq(16) e Eq(17) podemos classificar as correlações como:

- Estrutura canônica:

$$\underset{p\times p}{\rho_{U,X^{(1)}}} = A\sum\nolimits_{11} V_{11}^{-1/2} \quad \text{e} \quad \underset{q\times q}{\rho_{U,X^{(2)}}} = B\sum\nolimits_{22} V_{22}^{-1/2}$$

ou

$$\rho_{U,Z^{(1)}} = A_Z \rho_{11} \qquad \text{e} \qquad \rho_{V,Z^{(2)}} = B_Z \rho_{22}$$

- Cargas canônicas cruzadas:

$$\underset{p\times q}{\rho_{U,X^{(2)}}} = A\sum\nolimits_{12} V_{22}^{-1/2} \qquad \text{e} \qquad \underset{q\times p}{\rho_{V,X^{(1)}}} = B\sum\nolimits_{21} V_{11}^{-1/2}$$

ou

$$\rho_{U,Z^{(2)}} = A_Z \rho_{12} \quad \text{e} \quad \rho_{V,Z^{(1)}} = B_Z \rho_{21}$$

2.6.2 Estrutura canônica e cargas cruzadas canônicas amostrais:

O caso amostral é bem parecido com o populacional, onde também podemos calcular as correlações entre as variáveis canônicas e as variáveis originais para auxiliar na interpretação das variáveis canônicas $\hat{U}_k, \hat{V}_k$.

Definimos as matrizes

$$\begin{aligned} \underset{p\times p}{\hat{A}} &= \left[\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p\right]' \\ \underset{q\times q}{\hat{B}} &= \left[\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_p\right]' \end{aligned} \tag{18}$$

cujas colunas são os vetores dos coeficientes das variáveis canônicas amostrais. Analogamente a Eq(15) temos

$$\begin{aligned} \underset{(p\times 1)}{\hat{\mathbf{U}}_k} &= \hat{\mathbf{A}}\mathrm{x}^{(1)} \\ \underset{(q\times 1)}{\hat{\mathbf{V}}_k} &= \hat{\mathbf{B}}\mathrm{x}^{(2)} \end{aligned} \tag{19}$$

e podemos definir

$\mathbf{R}_{\hat{\mathrm{U}},\mathrm{x}^{(1)}}$ = matriz de correlações amostral de $\hat{\mathrm{U}}$ com $\mathrm{x}^{(1)}$.

$\mathbf{R}_{\hat{\mathrm{U}},\mathrm{x}^{(2)}}$ = matriz de correlações amostral de $\hat{\mathrm{U}}$ com $\mathrm{x}^{(2)}$.

$\mathbf{R}_{\hat{\mathrm{V}},\mathrm{x}^{(1)}}$ = matriz de correlações amostral de $\hat{\mathbf{V}}$ com $\mathrm{x}^{(1)}$.

$\mathbf{R}_{\hat{\mathrm{V}},\mathrm{X}^{(2)}}$ = matriz de correlações amostral de $\hat{\mathbf{V}}$ com $\mathrm{x}^{(2)}$.

Correspondendo a Eq(17), temos

$$\begin{aligned}
\mathbf{R}_{\hat{\mathrm{U}},\mathrm{X}^{(1)}} &= \hat{A}\mathbf{S}_{11}D_{11}^{-1/2} \\
\mathbf{R}_{\hat{\mathrm{U}},\mathrm{X}^{(2)}} &= \hat{B}\mathbf{S}_{22}D_{22}^{-1/2} \\
\mathbf{R}_{\hat{\mathrm{V}},\mathrm{X}^{(1)}} &= \hat{B}\mathbf{S}_{21}D_{11}^{-1/2} \\
\mathbf{R}_{\hat{\mathrm{V}},\mathrm{X}^{(2)}} &= \hat{A}\mathbf{S}_{12}D_{22}^{-1/2}
\end{aligned} \tag{20}$$

onde $D_{11}^{-1/2}$ é a matriz diagonal pxp com o i-ésimo elemento da diagonal $(Var\ \mathrm{x}_i^{(1)})^{-1/2}$ e $D_{22}^{-1/2}$ é a matriz diagonal qxq com o i-ésimo elemento da diagonal $(Var\ \mathrm{x}_i^{(2)})^{-1/2}$.

Comentário: Se as observações forem padronizadas a matriz de dados passa a ser

$$\mathbf{Z} = \begin{bmatrix} \mathrm{z}_1^{(1)\prime} & \mathrm{z}_1^{(2)\prime} \\ \vdots & \vdots \\ \mathrm{z}_n^{(1)\prime} & \mathrm{z}_n^{(2)\prime} \end{bmatrix}$$

e as variáveis canônicas amostrais se tornam

$$\begin{aligned}
\underset{(p\times 1)}{\hat{\mathbf{U}}} &= \hat{A}_Z \mathrm{z}^{(1)} \\
\underset{(q\times 1)}{\hat{\mathbf{V}}} &= \hat{B}_Z \mathrm{z}^{(2)}
\end{aligned} \tag{21}$$

onde $\hat{A}_Z = \hat{A}D_{11}^{-1/2}$ e $\hat{B}_Z = \hat{B}D_{22}^{-1/2}$. As correlações canônicas amostrais não são afetadas pela padronização (como no caso populacional). As correlações calculadas a partir de Eq(20) permanecem inalteradas e devem ser calculadas substituindo $\hat{A}_Z$ por $\hat{A}$, $\hat{B}_Z$ por $\hat{B}$ e $\mathbf{R}$ por $\mathbf{S}$. Note que $\mathbf{D}_{11}^{-1/2} = \underset{p\times p}{\mathbf{I}}$ e $\mathbf{D}_{22}^{-1/2} = \underset{q\times q}{\mathbf{I}}$ para as distribuições padronizadas.

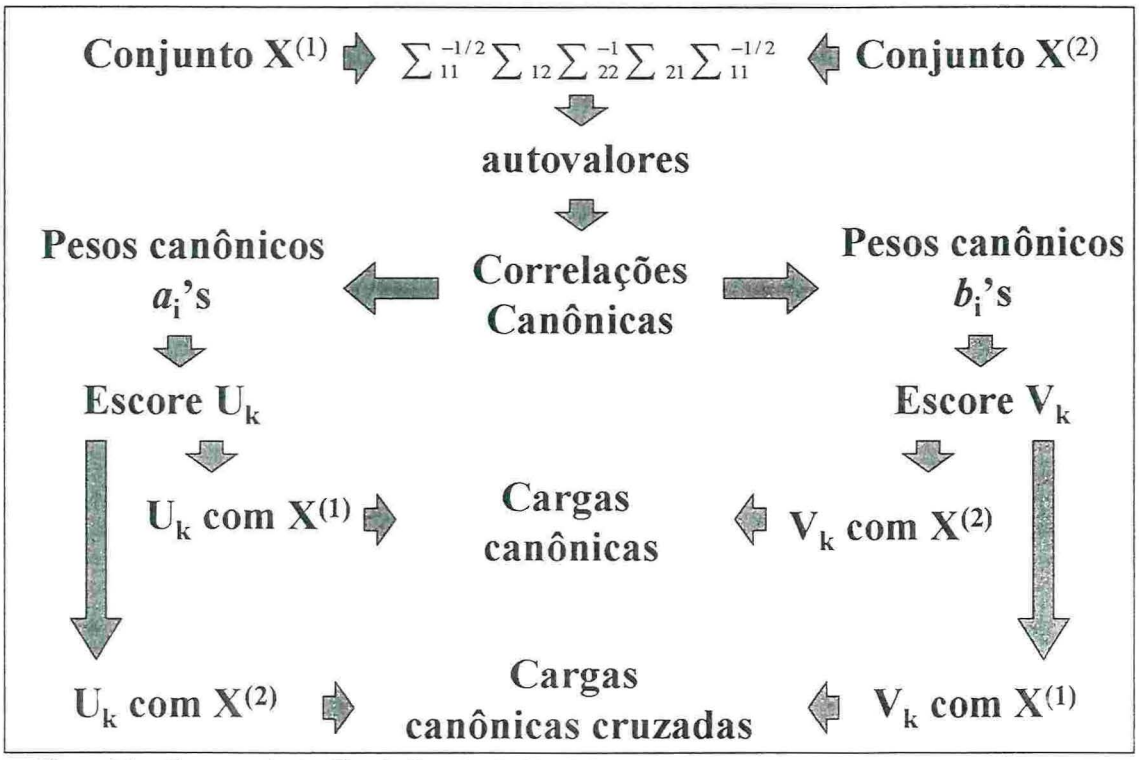


Figura 2.1 – Esquema da Análise de Correlação Canônica

2.7 Exemplo:

Já encontramos a matriz **R**, e as funções canônicas. Queremos agora encontrar as correlações entre as variáveis originais (padronizadas) e as variáveis canônicas. Do exemplo anterior temos:

$$\underset{5\times5}{\mathbf{R}} = \begin{bmatrix} \underset{2\times2}{\mathbf{R}_{11}} & \underset{2\times3}{\mathbf{R}_{12}} \\ \underset{3\times2}{\mathbf{R}_{21}} & \underset{3\times3}{\mathbf{R}_{22}} \end{bmatrix} = \begin{bmatrix} 1.000 & 0.214 & -0.161 & -0.033 & -0.093 \\ 0.214 & 1.000 & 0.159 & 0.312 & -0.312 \\ -0.161 & 0.159 & 1.000 & 0.477 & -0.631 \\ -0.033 & 0.312 & 0.477 & 1.000 & -0.489 \\ -0.093 & -0.312 & -0.631 & -0.489 & 1.000 \end{bmatrix}$$

$$\hat{U}_1 = [-0.157,\ -0.955] \quad \text{e} \quad \hat{V}_1 = [0.479,\ -0.594,\ 0.842]$$
$$\hat{U}_2 = [-1.011,\ 0.369] \qquad \hat{V}_2 = [1.102,\ 0.394,\ 0.814]$$

Estruturas canônicas:

Para a 1ª variável canônica (p=1):

$$\rho_{U_1,z^{(1)}} = A_z \rho_{11} = [-0.157,\ -0.955] \begin{bmatrix} 1.000 & 0.214 \\ 0.214 & 1.000 \end{bmatrix} = \begin{bmatrix} -0.361,\ -0.988 \end{bmatrix}$$

$$\rho_{V_1,z^{(2)}} = B_z \rho_{22} = [0.479,\ -0.594,\ 0.842] \begin{bmatrix} 1.000 & 0.477 & -0.631 \\ 0.477 & 1.000 & -0.489 \\ -0.631 & -0.489 & 1.000 \end{bmatrix} = \begin{bmatrix} -0.335,\ -0.777,\ 0.830 \end{bmatrix}$$

Para a 2ª variável canônica (p=2):

$$\rho_{U_2,z^{(1)}} = A_z \rho_{11} = [-1.011,\ 0.369] \begin{bmatrix} 1.000 & 0.214 \\ 0.214 & 1.000 \end{bmatrix} = \begin{bmatrix} -0.933,\ 0.153 \end{bmatrix}$$

$$\rho_{V_{21},z^{(2)}} = B_z \rho_{22} = [1.102,\ 0.394,\ 0.814] \begin{bmatrix} 1.000 & 0.477 & -0.631 \\ 0.477 & 1.000 & -0.489 \\ -0.631 & -0.489 & 1.000 \end{bmatrix} = \begin{bmatrix} 0.776,\ 0.521,\ -0.074 \end{bmatrix}$$

Chegando, finalmente, as matrizes de estrutura canônica:

Estrutura canônica química

	U_1	U_2
SO_4	-0,361	-0,933
Na	-0,988	0.153

Estrutura canônica meteorológica

	V_1	V_2
Velocidade	-0,335	0,776
Temperatura	-0,777	0.521
Umidade	0,830	-0,074

Figura 2.2 – estrutura canônica das variáveis químicas e meteorológicas

Podemos concluir que das variáveis químicas (padronizadas), o composto *Na* está mais associado a variável canônica U_2 enquanto que o composto SO_4 à variável canônica U_1.

Já no conjunto meteorológico, a *Velocidade dos ventos* está mais associada a V_2, *Temperatura* está mais associada a V_1 apesar de também estar associada a V_2 e a *Umidade* está associada a V_1.

Cargas canônicas cruzadas

Para a 1ª variável canônica (p=1):

$$\boldsymbol{\rho}_{U_1,z^{(2)}} = \boldsymbol{A}_z\boldsymbol{\rho}_{12} = [-0.157,\ -0.955]\begin{bmatrix} -0.161 & -0.033 & -0.093 \\ 0.159 & 0.312 & 0.312 \end{bmatrix} = \begin{bmatrix} -0.012, & -0.293, & 0.313 \end{bmatrix}$$

$$\boldsymbol{\rho}_{V_1,z^{(1)}} = \boldsymbol{B}_z\boldsymbol{\rho}_{21} = [0.479,\ -0.594,\ 0.842]\begin{bmatrix} -0.161 & 0.159 \\ -0.033 & 0.312 \\ -0.093 & 0.312 \end{bmatrix} = \begin{bmatrix} -0.136, & -0.372 \end{bmatrix}$$

Para a 2ª variável canônica (p=2):

$$\boldsymbol{\rho}_{U_2,z^{(2)}} = \boldsymbol{A}_z\boldsymbol{\rho}_{12} = [-1.012,\ 0.369]\begin{bmatrix} -0.161 & -0.033 & -0.093 \\ 0.159 & 0.312 & 0.312 \end{bmatrix} = \begin{bmatrix} 0.222, & 0.149, & -0.021 \end{bmatrix}$$

$$\boldsymbol{\rho}_{V_2,z^{(1)}} = \boldsymbol{B}_z\boldsymbol{\rho}_{21} = [1.102,\ 0.394,\ 0.814]\begin{bmatrix} -0.161 & 0.159 \\ -0.033 & 0.312 \\ -0.093 & 0.312 \end{bmatrix} = \begin{bmatrix} -0.266, & 0.044 \end{bmatrix}$$

Como no exemplo anterior também podemos conferir os resultados calculando a correlação entre as variáveis originais e os escores obtidos através dos pesos.

2.8 As primeiras r variáveis canônicas como um resumo da variabilidade

Já sabemos que os vetores de coeficientes $\boldsymbol{a}_i$, $\boldsymbol{b}_i$ são selecionados de forma que maximizem as correlações entre as variáveis canônicas sendo que essas correlações são os

autovalores da matriz $\sum_{11}^{-1/2}\sum_{12}\sum_{22}^{-1}\sum_{21}\sum_{11}^{-1/2}$. Se tomarmos algumas poucas funções canônicas (pares de variáveis canônicas) e estas explicarem um quantidade pequena da variabilidade em $\sum_{11}$ e $\sum_{22}$, não fica claro como devemos interpretar a mais alta correlação canônica pois mesmo ela sendo alta, não explica uma porcentagem significativa da variabilidade de cada conjunto de variáveis.

Mas se as variáveis canônicas representam "bem" os respectivos conjuntos de variáveis originais, então as associações entre as variáveis podem ser descritas em termos das variáveis canônicas e suas correlações. Para interpretação, é útil saber quanto cada variável canônica está contribuindo para a explicação da variabilidade do seu respectivo conjunto e também saber qual a proporção da variância de um conjuntos de variáveis que é explicada pelas variáveis canônicas do outro conjunto.

2.9 Matrizes de resíduos

As matrizes de resíduos do modelo (também chamadas de erros de aproximação) podem ser interpretadas como descrição de quão bem as primeiras *r* variáveis canônicas amostrais reproduzem as matrizes de covariâncias amostrais. Modelos com muitas linhas e/ou colunas nessas matrizes de resíduos indicam um ajuste pobre da(s) variável(is) correspondente(s).

É preciso salientar que para calcular as matrizes de resíduos precisamos ter o mesmo número de variáveis em ambos conjuntos. Isso acontece porque para o cálculo das matrizes de erros aproximados necessitamos inverter as matrizes dos coeficientes canônicos e, se o número de variáveis em cada conjunto for diferente irá produzir matrizes de coeficientes não quadradas, que não são inversíveis.

As primeiras *r* variáveis são melhores na reprodução de $\mathbf{S}_{12}=\mathbf{S}_{21}'$ do que na reprodução dos elementos $\mathbf{S}_{11}$ e $\mathbf{S}_{22}$. Essas correlações usualmente são próximas de zero. De outra forma, as matrizes residuais associadas com as aproximações das matrizes $\mathbf{S}_{11}$ e $\mathbf{S}_{22}$ dependem somente do último *p-r* e *q-r* vetores. Os elementos destes vetores devem ser relativamente grandes e, dessa forma, produzem matrizes grandes.

A partir das matrizes $\hat{A}$ e $\hat{B}$ definidas em Eq(18), seja $\hat{a}^{(i)}$ e $\hat{b}^{(i)}$ denotando a i-ésima coluna de $\hat{A}^{-1}$ e $\hat{B}^{-1}$, respectivamente e sabendo que $\hat{\mathbf{U}} = \hat{A}\mathrm{x}^{(1)}$ e $\hat{\mathbf{V}} = \hat{B}\mathrm{x}^{(2)}$ podemos escrever

$$\begin{aligned} \underset{(p\times 1)}{\mathrm{x}^{(1)}} &= \underset{(p\times p)}{\hat{A}}^{-1}\hat{\mathbf{U}} \\ \underset{(q\times 1)}{\mathrm{x}^{(2)}} &= \underset{(q\times q)}{\hat{B}}^{-1}\hat{\mathbf{V}} \end{aligned} \tag{22}$$

Porque na amostra $Cov(\hat{\mathbf{U}}, \hat{\mathbf{V}}) = \hat{A}\mathbf{S}_{12}\hat{B}'$, $Cov(\hat{\mathbf{U}}) = \hat{A}\mathbf{S}_{11}\hat{A}' = \underset{(p\times p)}{\mathbf{I}}$ e $Cov(\hat{\mathbf{V}}) = \hat{B}\mathbf{S}_{22}\hat{B}' = \underset{(q\times q)}{\mathbf{I}}$.

Matricialmente temos

$$\mathbf{S}_{12} = \hat{A}^{-1}\begin{bmatrix} \rho_1^* & 0 & 0 & \cdots & 0 & \vdots & \\ 0 & \rho_2^* & 0 & \cdots & & \vdots & \\ 0 & 0 & \rho_3^* & \cdots & 0 & \vdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & 0 & \vdots & \\ 0 & 0 & 0 & \cdots & \rho_p^* & \vdots & \end{bmatrix}\left(\hat{B}^{-1}\right)' = \rho_1^*\hat{a}^{(1)}\hat{b}^{(1)\prime} + \ldots\ldots + \rho_p^*\hat{a}^{(p)}\hat{b}^{(p)\prime} \tag{23}$$

$$\mathbf{S}_{11} = \left(\hat{A}^{-1}\right)\left(\hat{A}^{-1}\right)' = \hat{a}^{(1)}\hat{a}^{(1)\prime} + \hat{a}^{(2)}\hat{a}^{(2)\prime} + \ldots\ldots + \hat{a}^{(p)}\hat{a}^{(p)\prime} \tag{24}$$

$$\mathbf{S}_{22} = \left(\hat{B}^{-1}\right)\left(\hat{B}^{-1}\right)' = \hat{b}^{(1)}\hat{b}^{(1)\prime} + \hat{b}^{(2)}\hat{b}^{(2)\prime} + \ldots\ldots + \hat{b}^{(p)}\hat{b}^{(p)\prime} \tag{25}$$

desde que $\mathbf{x}^{(1)} = \hat{A}^{-1}\hat{\mathbf{U}}$ e $\hat{\mathbf{U}}$ possui covariância amostral = I, as primeiras r colunas de $\hat{A}^{-1}$ contém a covariância amostral das primeiras r variáveis canônicas $\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_r$ com as variáveis $X_1^{(1)}, X_2^{(1)}, \ldots, X_p^{(1)}$. Similarmente, as primeiras r colunas de $\hat{B}^{-1}$ contém a covariância amostral de $\hat{V}_1, \hat{V}_2, \ldots, \hat{V}_r$ com as variáveis $X_1^{(2)}, X_2^{(2)}, \ldots, X_p^{(2)}$.

Se apenas os primeiros r pares canônicos forem utilizados, então:

$$\tilde{x}^{(1)} = \begin{bmatrix} \hat{a}^{(1)} & \hat{a}^{(2)} & \dots & \hat{a}^{(r)} \end{bmatrix} \begin{bmatrix} \hat{U}_1 \\ \hat{U}_2 \\ \vdots \\ \hat{U}_r \end{bmatrix} \quad \text{e} \quad \tilde{x}^{(2)} = \begin{bmatrix} \hat{b}^{(1)} & \hat{b}^{(2)} & \dots & \hat{b}^{(r)} \end{bmatrix} \begin{bmatrix} \hat{V}_1 \\ \hat{V}_2 \\ \vdots \\ \hat{V}_r \end{bmatrix}$$

então S_{12} é aproximado por $Cov(\tilde{x}^{(1)}, \tilde{x}^{(2)})$.

Na Análise de Regressão, quando queremos encontrar os resíduos fazemos:

$$res = Y - \hat{Y}$$

No caso da Análise de Correlação Canônica, utilizando as equações Eq(23), Eq(24) e Eq(25) temos:

$$S_{11} - \left(\hat{a}^{(1)}\hat{a}^{(1)\prime} + \hat{a}^{(2)}\hat{a}^{(2)\prime} + \dots + \hat{a}^{(r)}\hat{a}^{(r)\prime}\right) = \hat{a}^{(r+1)}\hat{a}^{(r+1)\prime} + \dots + \hat{a}^{(p)}\hat{a}^{(p)\prime} \tag{26}$$

$$S_{22} - \left(\hat{b}^{(1)}\hat{b}^{(1)\prime} + \hat{b}^{(2)}\hat{b}^{(2)\prime} + \dots + \hat{b}^{(r)}\hat{b}^{(r)\prime}\right) = \hat{b}^{(r+1)}\hat{b}^{(r+1)\prime} + \dots + \hat{b}^{(p)}\hat{b}^{(p)\prime} \tag{27}$$

$$S_{12} - \left(\rho_1^*\hat{a}^{(1)}\hat{b}^{(1)\prime} + \dots + \rho_p^*\hat{a}^{(r)}\hat{b}^{(r)\prime}\right) = \rho_{r+1}^*\hat{a}^{(r+1)}\hat{b}^{(r+1)\prime} + \dots + \rho_p^*\hat{a}^{(p)}\hat{b}^{(p)\prime} \tag{28}$$

Se utilizarmos observações padronizadas devemos trocar S_{kl} com R_{kl} e $\hat{a}^{(k)}, \hat{b}^{(l)}$ por $\hat{a}_Z^{(k)}, \hat{b}_Z^{(l)}$ nas equações Eq(23), Eq(24) e Eq(25).

2.10 Exemplo:

Utilizando os mesmos dados anteriores vamos considerar que só fossemos utilizar a primeira função canônica então:

$$\hat{A}_Z^{-1} = \begin{bmatrix} -0.157 & -0.955 \\ -1.011 & 0.369 \end{bmatrix}^{-1} = \begin{bmatrix} -0.361 & -0.933 \\ -0.988 & 0.153 \end{bmatrix}$$

A matriz de resíduos de $\mathbf{R}_{11}$ será:

$$\mathbf{R}_{11} - Cov(\tilde{z}^{(1)}) = \begin{bmatrix} -0.933 \\ 0.153 \end{bmatrix} \begin{bmatrix} -0.933 & 0.153 \end{bmatrix} = \begin{bmatrix} 0.869 & -0.143 \\ -0.143 & 0.234 \end{bmatrix}$$

2.11 Proporção da variância amostral explicada e índice de redundância

A correlação canônica elevada ao quadrado representa uma estimativa da ***variância conjunta*** (ou compartilhada) entre as variáveis canônicas. É uma medida que pode ser mal interpretada uma vez que mede a variância compartilhada entre as variáveis canônicas e não entre as variáveis originais.

A ***proporção da variância explicada*** é a proporção da variância de um conjunto de variáveis que é explicada pelas respectivas variáveis canônicas.

Dessa forma, mesmo que tenhamos correlações canônicas fortes podemos não ter extraído quantidades significativas da variância das variáveis originais.

O ***índice de redundância*** foi proposto para facilitar as interpretações. Ele é equivalente ao coeficiente de determinação da análise de regressão. É a média simples dos coeficientes de correlação múltiplo de um conjunto de variáveis com cada uma das variáveis do outro conjunto, que resulta num coeficiente de determinação médio. Essa medida mede a porcentagem da variância em um conjunto de variáveis que é explicada pelo outro conjunto. Devemos notar que o valor máximo desse coeficiente não é 100% e sim a variância compartilhada entre os dois conjuntos.

Podemos calcular o ***índice de redundância*** para cada variável canônica e depois calcular o ***índice de redundância total***, que nada mais é do que a quantidade de variância em um conjunto explicada pelas *r* primeiras variáveis canônicas do outro.

Quando as observações são padronizadas, as matrizes de covariâncias amostrais $\mathbf{S}_{kl}$ são matrizes de correlação $\mathbf{R}_{kl}$. Os vetores de coeficientes canônicos são as linhas das matrizes

$\hat{A}_Z$ e $\hat{B}_Z$ e as colunas de $\hat{A}_Z^{-1}$ e $\hat{B}_Z^{-1}$ são as correlações amostrais entre as variáveis canônicas e suas respectivas variáveis.

Especificamente,

$$\begin{aligned} Cov\left(\mathrm{z}^{(1)}, \hat{\mathbf{U}}\right) &= Cov(\hat{A}_Z^{-1}\hat{\mathbf{U}}, \hat{\mathbf{U}}) = \hat{A}_Z^{-1} \\ Cov\left(\mathrm{z}^{(2)}, \hat{\mathbf{V}}\right) &= Cov(\hat{B}_Z^{-1}\hat{\mathbf{V}}, \hat{\mathbf{V}}) = \hat{B}_Z^{-1} \end{aligned} \tag{29}$$

então

$$\begin{aligned} \hat{A}_Z^{-1} &= \left[\hat{\boldsymbol{a}}_Z^{(1)}, \hat{\boldsymbol{a}}_Z^{(2)}, \ldots, \hat{\boldsymbol{a}}_Z^{(p)}\right] = \begin{bmatrix} r_{\hat{U}_1, Z_1^{(1)}} & r_{\hat{U}_{21}, Z_1^{(1)}} & \cdots & r_{\hat{U}_p, Z_1^{(1)}} \\ r_{\hat{U}_1, Z_2^{(1)}} & r_{\hat{U}_2, Z_2^{(1)}} & \cdots & r_{\hat{U}_p, Z_2^{(1)}} \\ \vdots & \vdots & \ddots & \vdots \\ r_{\hat{U}_1, Z_p^{(1)}} & r_{\hat{U}_2, Z_p^{(1)}} & \cdots & r_{\hat{U}_p, Z_p^{(1)}} \end{bmatrix} \\ \hat{B}_Z^{-1} &= \left[\hat{\boldsymbol{b}}_Z^{(1)}, \hat{\boldsymbol{b}}_Z^{(2)}, \ldots, \hat{\boldsymbol{b}}_Z^{(p)}\right] = \begin{bmatrix} r_{\hat{V}_1, Z_1^{(2)}} & r_{\hat{V}_{21}, Z_1^{(2)}} & \cdots & r_{\hat{V}_p, Z_1^{(2)}} \\ r_{\hat{V}_1, Z_2^{(2)}} & r_{\hat{V}_2, Z_2^{(2)}} & \cdots & r_{\hat{V}_p, Z_2^{(2)}} \\ \vdots & \vdots & \ddots & \vdots \\ r_{\hat{V}_1, Z_p^{(2)}} & r_{\hat{V}_2, Z_p^{(2)}} & \cdots & r_{\hat{V}_p, Z_p^{(2)}} \end{bmatrix} \end{aligned} \tag{30}$$

onde $r_{\hat{U}_i, Z_k^{(1)}}$ e $r_{\hat{V}_i, Z_k^{(2)}}$ são os coeficientes de correlação amostrais entre as variáveis originais e as variáveis canônicas, ou seja, os termos que compõem a matriz de estrutura canônica.

Usando as equações Eq(23), Eq(24) e Eq(25) para observações padronizadas, obtemos:

- Variância total amostral (padronizada) no primeiro conjunto

$$= tr(\mathbf{R}_{11}) = tr\left(\hat{\boldsymbol{a}}_Z^{(1)}\hat{\boldsymbol{a}}_Z^{(1)\prime} + \hat{\boldsymbol{a}}_Z^{(2)}\hat{\boldsymbol{a}}_Z^{(2)\prime} + \ldots + \hat{\boldsymbol{a}}_Z^{(p)}\hat{\boldsymbol{a}}_Z^{(p)\prime}\right) = p \tag{31}$$

- Variância total amostral (padronizada) no segundo conjunto

$$= tr(\mathbf{R}_{22}) = tr\left(\hat{\boldsymbol{b}}_Z^{(1)}\hat{\boldsymbol{b}}_Z^{(1)\prime} + \hat{\boldsymbol{b}}_Z^{(2)}\hat{\boldsymbol{b}}_Z^{(2)\prime} + \ldots + \hat{\boldsymbol{b}}_Z^{(p)}\hat{\boldsymbol{b}}_Z^{(p)\prime}\right) = q \tag{32}$$

Desde que as correlações nas primeiras $r < p$ colunas de $\hat{A}_Z^{-1}$ e $\hat{B}_Z^{-1}$ envolvem só as variáveis canônicas amostrais, $\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_r$ e $\hat{V}_1, \hat{V}_2, \ldots, \hat{V}_r$ respectivamente, definimos a contribuição das primeiras r variáveis canônicas à variância total amostral (padronizada) como:

$$tr\left(\hat{\boldsymbol{a}}_Z^{(1)}\hat{\boldsymbol{a}}_Z^{(1)\prime} + \hat{\boldsymbol{a}}_Z^{(2)}\hat{\boldsymbol{a}}_Z^{(2)\prime} + \ldots + \hat{\boldsymbol{a}}_Z^{(r)}\hat{\boldsymbol{a}}_Z^{(r)\prime}\right) = \sum_{i=1}^{r}\sum_{k=1}^{p} r^2_{\hat{U}_i, z_k^{(1)}}$$

$$tr\left(\hat{\boldsymbol{b}}_Z^{(1)}\hat{\boldsymbol{b}}_Z^{(1)\prime} + \hat{\boldsymbol{b}}_Z^{(2)}\hat{\boldsymbol{b}}_Z^{(2)\prime} + \ldots + \hat{\boldsymbol{b}}_Z^{(r)}\hat{\boldsymbol{b}}_Z^{(r)\prime}\right) = \sum_{i=1}^{r}\sum_{k=1}^{p} r^2_{\hat{V}_i, z_k^{(2)}}$$

As proporções da variância total amostral (padronizada) explicada pelas r primeiras variáveis canônicas se torna

$$\mathbf{R}^2_{z^{(1)}|\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_r} = \frac{tr\left(\hat{\boldsymbol{a}}_Z^{(1)}\hat{\boldsymbol{a}}_Z^{(1)\prime} + \hat{\boldsymbol{a}}_Z^{(2)}\hat{\boldsymbol{a}}_Z^{(2)\prime} + \ldots + \hat{\boldsymbol{a}}_Z^{(r)}\hat{\boldsymbol{a}}_Z^{(r)\prime}\right)}{tr(\mathbf{R}_{11})} = \frac{\sum_{i=1}^{r}\sum_{k=1}^{p} r^2_{\hat{U}_i, z_k^{(1)}}}{p} \tag{33}$$

= (proporção da variância total amostral padronizada no 1° conjunto de variáveis explicada por $\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_r$).

= a soma de cada valor da matriz de estrutura canônica elevado ao quadrado dividida por p (que é o número de variáveis no primeiro conjunto).

Da mesma forma para o 2° conjuntos de variáveis:

$$\mathbf{R}^2_{z^{(2)}|\hat{V}_1, \hat{V}_2, \ldots, \hat{V}_r} = \frac{tr\left(\hat{\boldsymbol{b}}_Z^{(1)}\hat{\boldsymbol{b}}_Z^{(1)\prime} + \hat{\boldsymbol{b}}_Z^{(2)}\hat{\boldsymbol{b}}_Z^{(2)\prime} + \ldots + \hat{\boldsymbol{b}}_Z^{(r)}\hat{\boldsymbol{b}}_Z^{(r)\prime}\right)}{tr(\mathbf{R}_{22})} = \frac{\sum_{i=1}^{r}\sum_{k=1}^{p} r^2_{\hat{V}_i, z_k^{(2)}}}{q} \tag{34}$$

= (proporção da variância total amostral padronizada no 2° conjunto de variáveis explicada por $\hat{V}_1, \hat{V}_2, \ldots, \hat{V}_r$).

= a soma de cada valor da matriz de estrutura canônica elevado ao quadrado dividida por q (que é o número de variáveis no segundo conjunto).

Para calcular a redundância de cada uma das combinações lineares encontradas temos

$$IR_{U_i} = \hat{\rho}_i^{*2} \times \frac{\sum_{i=1}^{p} r^2_{\hat{U}_i, z_i^{(1)}}}{p} = \lambda_i \times \frac{\sum_{i=1}^{p} r^2_{\hat{U}_i, z_i^{(1)}}}{p} \tag{35}$$

ou seja, é o autovalor vezes a proporção da variância explicada pelo respetivo autovalor.
Da mesma forma para o segundo conjunto:

$$IR_{V_i} = \hat{\rho}_i^{*2} \times \frac{\sum_{i=1}^{p} r^2_{\hat{V}_i, z_i^{(2)}}}{q} = \lambda_i \times \frac{\sum_{i=1}^{p} r^2_{\hat{V}_i, z_i^{(2)}}}{q} \tag{36}$$

O índice de redundância total nada mais é do que a soma dos índices de redundância de cada combinação linear do mesmo conjunto e dados, ou seja:

$$IRT_U = \sum_{i=1}^{p} \hat{\rho}_i^{*2} \times \frac{\sum_{i=1}^{p} r^2_{\hat{U}_i, z_i^{(1)}}}{p} = \sum_{i=1}^{p} \lambda_i \times \frac{\sum_{i=1}^{p} r^2_{\hat{U}_i, z_i^{(1)}}}{p} \tag{37}$$

$$IRT_V = \sum_{i=1}^{p} \hat{\rho}_i^{*2} \times \frac{\sum_{i=1}^{p} r^2_{\hat{V}_i, z_i^{(2)}}}{q} = \sum_{i=1}^{p} \lambda_i \times \frac{\sum_{i=1}^{p} r^2_{\hat{V}_i, z_i^{(2)}}}{q} \tag{38}$$

As medidas descritivas Eq(33) e Eq(34) produzem algumas indicações de quão bem as variáveis canônicas representam seus respectivos conjuntos. Elas produzem descrições das matrizes de erros. Em particular,

$$\frac{1}{p} tr\left[R_{11} - \hat{a}_Z^{(1)}\hat{a}_Z^{(1)\prime} + \hat{a}_Z^{(2)}\hat{a}_Z^{(2)\prime} + \ldots + \hat{a}_Z^{(r)}\hat{a}_Z^{(r)\prime}\right] = 1 - R^2_{Z^{(1)}|\hat{U}_1,\hat{U}_2,\ldots,\hat{U}_r}$$

$$\frac{1}{q} tr\left[R_{22} - \hat{b}_Z^{(1)}\hat{b}_Z^{(1)\prime} + \hat{b}_Z^{(2)}\hat{b}_Z^{(2)\prime} + \ldots + \hat{b}_Z^{(r)}\hat{b}_Z^{(r)\prime}\right] = 1 - R^2_{Z^{(2)}|\hat{V}_1,\hat{V}_2,\ldots,\hat{V}_r}$$

de acordo com Eq(31), Eq(32), e Eq(33).

2.12 Exemplo:

No exemplo consideraremos VE a variância extraída e ID o índice de redundância.
Continuando a analisar os dados do exemplo anterior já sabemos que as matrizes de estrutura canônica são:

Estrutura canônica química

	U_1	U_2
SO_4	-0,361	-0,933
Na	-0,988	0.153

Estrutura canônica meteorológica

	V_1	V_2
Velocidade	-0,335	0,776
Temperatura	-0,777	0.521
Umidade	0,830	-0,074

Para saber a quantidade de variância que é extraída por cada raiz canônica do conjunto original precisamos calcular a média simples de cada peso elevado ao quadrado:

$$VE_{U_1} = \frac{(-0,361)^2 + (-0,988)^2}{2} = 0,554$$

$$VE_{U_2} = \frac{(-0,933)^2 + (0,153)^2}{2} = 0,446$$

$$VE_{V_1} = \frac{(-0,335)^2 + (-0,777)^2 + (0,830)^2}{3} = 0,469$$

$$VE_{V_2} = \frac{(0,776)^2 + (0,521)^2 + (-0,074)^2}{3} = 0,293$$

Se multiplicarmos a correlação canônica correspondente elevada ao quadrado pelas variâncias extraídas temos o índice de redundância:

$$IR_{U_1} = 0,142 \times 0,554 = 0,079$$

$$IR_{U_2} = 0,019 \times 0,446 = 0,036$$

$$IR_{V_1} = 0,142 \times 0,469 = 0,066$$

$$IR_{U_1} = 0,019 \times 0,293 = 0,024$$

Que podem ser reorganizados nas tabelas:

	Variância Extraída	Índice de Redundância		Variância Extraída	Índice de Redundância
U_1	0,554	0,079	V_1	0,469	0,066
U_2	0,446	0,036	V_2	0,293	0,024

O índice de redundância total, que é a soma dos índices de redundância fica:

$$IRUT = 0{,}079 + 0{,}036 = 0{,}115$$
$$IRVT = 0{,}066 + 0{,}024 = 0{,}090$$

2.13 Testes de significância global

Quando temos amostras grandes podemos estar interessados a fazer inferências dos resultados da CANCORR. Como qualquer teste estatístico precisamos saber qual a significância de cada correlação canônica. Existem testes de significância global e o mais utilizado é o teste de Rao. Já para testar separadamente cada função canônica existem alguns testes. São eles: Lambda de Wilk (*Wilk's Lambda*), traço de Hotteling-Lawley (*Hotteling's trace*), Traço de Pillai (*Pillai's trace*) e maior raiz característica de Roy (*Roy's gnc*).

Quando $\sum_{12} = \mathbf{0}$, $\boldsymbol{a}'\mathbf{X}^{(1)}$ e $\boldsymbol{b}'\mathbf{X}^{(2)}$ tem variância $\boldsymbol{a}'\sum_{12}\boldsymbol{b} = 0$ para todos os vetores $\boldsymbol{a}$ e $\boldsymbol{b}$. Conseqüentemente, todas as correlações canônicas serão zero, e não existe porque propor uma análise de correlação canônica. O próximo resultado é uma forma de testar se $\sum_{12} = \mathbf{0}$, para amostras grandes, que significa que os conjuntos de variáveis não são linearmente correlacionados ou relacionados.

Res 3:

Seja $\mathbf{X}_j = \begin{bmatrix} \mathbf{X}_j^{(1)} \\ \mathbf{X}_j^{(2)} \end{bmatrix}$, $j=1,2,\ldots,n$ uma amostra aleatória de uma população $N_{(p+q)}(\boldsymbol{\mu}, \sum)$

com

$$\sum = \begin{bmatrix} \underset{p\times p}{\sum_{11}} & \underset{p\times q}{\sum_{12}} \\ \underset{q\times p}{\sum_{21}} & \underset{q\times q}{\sum_{22}} \end{bmatrix}$$

então o teste $H_0 : \underset{p\times q}{\sum_{12}} = \mathbf{0}$ versus $H_1 : \underset{p\times q}{\sum_{12}} \neq \mathbf{0}$ rejeita H_0 para valores grandes de

$$-2\ln \Lambda = n\ln\left(\frac{|\mathbf{S}_{11}||\mathbf{S}_{22}|}{|\mathbf{S}|}\right) = -n\ln\prod_{i=1}^{p}(1-\rho_i^{*2}) \qquad (39)$$

onde $\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}$ é o estimador não viciado de $\sum$. Para n grande, a estatística da Eq(39) tem distribuição aproximadamente χ^2_{pq}.

A estatística da Eq(39) compara a variância amostral generalizada sob H_0, isto é:

$\begin{vmatrix} \mathbf{S}_{11} & \mathbf{0} \\ \mathbf{0}' & \mathbf{S}_{22} \end{vmatrix} = |\mathbf{S}_{11}||\mathbf{S}_{22}|$ com a variância generalizada irrestrita $|\mathbf{S}|$.

Bartlett (1939) [apud Johnson e Wichern (2002), pg. 616] sugere que se troque o fator multiplicativo de n (n likelihood ratio) por $n-1-\frac{1}{2}(p+q+1)$ a fim de melhorar a aproximação da distribuição amostral de $(-2\ln\Lambda)$ pela distribuição χ^2. Desta forma, para n e $n-(p+q)$ grandes, se rejeitarmos $H_0 : \sum_{12} = \mathbf{0}$ $\left(\rho_1^* = \rho_2^* = \ldots = \rho_p^* = 0\right)$ é natural que se examine a significância de cada uma das correlações canônicas separadamente. Como as mesmas se encontram ordenadas da maior para a menor, podemos começar a testá-las

assumindo que a primeira correlação canônica é diferente de zero e as *(p-1)* demais são iguais a zero. Se essa hipótese é rejeitada, assumimos que as duas primeiras correlações canônicas são diferentes de zero e as demais são iguais a zero, e assim sucessivamente.

$$H_0^k : \rho_1^* \neq 0, \rho_2^* \neq 0, \ldots, \rho_k^* \neq 0, \rho_{k+1}^* = 0, \ldots, \rho_p^* \neq 0$$
$$H_1^k : \rho_i^* \neq 0, \quad \text{para algum } i > k+1 \tag{40}$$

Bartlett (1938) [apud Johnson e Wichern (2002), pg. 617] argumentava que a k-ésima hipótese da Eq(40) pode ser testada pelo critério da razão de verossimilhança (ratio). Especificamente,

Rejeitamos H_0 a um nível α de significância se

$$-\left(n-1-\frac{1}{2}(p+q+1)\right)\ln\prod_{i=k+1}^{p}\left(1-\rho_i^{*2}\right) > \chi^2_{(p-k)(q-k)}(\alpha) \tag{41}$$

Se a seqüência $H_0, H_0^{(1)}, H_0^{(2)}, \ldots$ são testadas uma a uma desde que $H_0^{(k)}$ não tenha sido rejeitada para uma certo k, a significância global não é mais α e, de fato, seria difícil

determiná-la. Um outro defeito deste procedimento é a tendência que introduz na conclusão de que a hipótese nula é correta simplesmente porque não é rejeitada.

Resumindo, a significância global do resultado 3 é útil para dados com distribuição normal multivariada. Os testes seqüenciais utilizados na Eq(41) devem ser interpretados com cuidado, e são, muitas vezes, utilizados para selecionar o número de variáveis canônicas importantes, isto é, que devem ser analisadas.

2.14 Exemplo:

Sabemos que $n = 177$ e que as raízes canônicas (autovalores) são *0,1418 e 0,0815*.
Se utilizarmos as duas raízes canônicas (autovalores) o teste qui-quadrado fica:

H_0: $\rho_1^* = 0 \quad e \quad \rho_2^* = 0$

H_1: $\rho_i^* \neq 0$ para algum i.

$$\chi^2_{calc} = 177 - 1 - \frac{1}{2}(2+3+1) \times \ln \prod_{i=1}^{2} (1 - \rho_i^{*2}) = 173 \times \ln[(1-0{,}1418)(1-0{,}0815)] = 41{,}164$$

$$\chi^2_{5;(6)} = 12{,}59$$

Com nível descritivo (valor de probabilidade) $2{,}71723 \times 10^{-7}$.

Já que $\chi^2_{calc} > \chi^2_{tab}$ rejeitamos H_0, ou seja, rejeitamos a hipótese de que as duas correlações canônicas são nulas. Temos, agora que testar se a retirada de uma raiz muda essa significância. O novo teste será:

H_0: $\rho_2^* = 0$

H_1: $\rho_2^* \neq 0$

$$\chi^2_{calc} = 177 - 1 - \frac{1}{2}(2+3+1) \times \ln(1-0{,}0815) = 14{,}716$$

$$\chi^2_{5;(2)} = 5{,}99$$

Com nível descritivo (valor de probabilidade) $6{,}3858 \times 10^{-4}$.

Já que $\chi^2_{calc} > \chi^2_{tab}$ rejeitamos H_0, ou seja, rejeitamos a hipótese de que a segunda correlação canônica seja nula, ou seja, a segunda correlação canônica também é significativa.

Num caso onde se tenham mais variáveis canônicas, esse procedimento deve ser seguido até que se encontre uma correlação canônica não significativa.

3 APLICAÇÃO EM UM ESTUDO AMBIENTAL

O exemplo que será utilizado aqui foi cedido pela Fundação Estadual de Proteção Ambiental (FEPAM), setor de Projetos de Pesquisas, Parte Integrande do Projeto CNPq Plano Sul intitulado "CARACTERIZAÇÃO DOS POLUENTES ATMOSFÉRICOS NA REGIÃO DA BACIA HIDROGRÁFICA DO GUAÍBA - RIO GRANDE DO SUL", onde realizei meu estágio obrigatório. Este trabalho teve a participação de Elba Calesso Texeira, Daniela Montanari Migliavacca e Cláudia Braga. O software utilizado foi o STATISTICA for windows versão 4.3.

A coleta dos dados foi feita de janeiro a dezembro de 2002 em três pontos diferentes da Bacia Hidrográfica do Guaíba com dois tipos de amostradores.

Os amostradores utilizados foram:

Amostrador tipo Bulk que era composto por um funil de polietileno, com 21,5 cm de diâmetro, acoplado a um frasco coletor de 5 litros do mesmo material, fixados a 2 m da superfície do solo, livre de obstáculos. O funil foi recoberto com uma tela de nylon para impedir a contaminação das amostras por agentes externos, tais como folhas, insetos.

Amostrador de precipitação úmida que era constituído de uma caixa metálica de proteção e um frasco coletor de polietileno, com capacidade de 5 litros, acoplada a um funil de acrílico com tampa do mesmo material, cuja tampa abre-se apenas na presença de precipitação úmida e fechando-se após o término da mesma. Sendo alimentado por rede elétrica, e na ausência desta alimenta-se por uma bateria de 12 volts.

Os locais selecionados para a coleta das amostras foram: estação Charqueadas no município de Charqueadas e estações Ceasa e 8° Distrito localizadas no município de Porto Alegre. Os amostradores eram colocados nestes três pontos e as amostras (água da chuva) eram retiradas imediatamente após o final da chuva.

As estações 8° Distrito e CEASA encontram-se localizadas em áreas urbanizadas e de intensa atividade veicular. A sudeste da estação 8°Distrito está localizado o Hospital São Lucas (HSL), principal fonte de contribuição de particulados por conseqüência da incineração do lixo hospitalar. A estação CEASA localiza-se dentro da área das Centrais de Abastecimento do RS (CEASA), nas proximidades das Rodovias BR-116 e BR-290, cujo tráfego de veículos é bastante intenso e a nordeste da estação está localizada a Refinaria Alberto Pasqualini. A estação Charqueadas está localizada a aproximadamente 60 km de

Porto Alegre e as principais vias de acesso são as Rodovias Federal BR-290 e Estadual RS-401. A estação foi instalada à noroeste da siderúrgica Aços Finos Piratini (AFP) e à sudeste da Termoelétrica Charqueadas (TERMOCHAR) e a aproximadamente 10 km a sudoeste está localizada a Usina Termoelétrica Salto do Jacuí (UTSJ), no município de São Jerônimo, com uma capacidade instalada de 20 MW.

A tabela 3.1 mostra a localização dos pontos de coleta.

Estações de Amostragem	Município	Coordenadas geográficas (X, Y)	
Charqueadas	Charqueadas	438289	6682698
CEASA	Porto Alegre	483682	6682545
8°Distrito de Meteorologia	Porto Alegre	482729	6675003

Tabela 3.1: Estações de amostragem de precipitação atmosférica da Bacia Hidrográfica do Guaíba.

Das amostras coletadas eram medidas os parâmetros físico-químicos: concentração de hidrogênio (H^+), Condutividade, Alcalinidade, Fluoreto (F^-), Cloreto (Cl^-), Nitrato (NO_3^-), Sulfato (SO_4^{2-}), Sódio (Na^+), Amônio (NH_4^+), Potássio (K^+), Magnésio (Mg^{2+}) e Cálcio (Ca^{2+}) E em cada dia de coleta tínhamos algumas variáveis meteorológicas que foram fornecidas pelo aeroporto Salgado Filho que foram: Pluviometria, Precipitação Total, direção do vento nordeste (NE), direção do vento sudeste (SE), direção do vento sudoeste (SO) e direção do vento noroeste (NO), velocidade dos ventos, pressão, temperatura e umidade relativa .

Utilizamos a CANCORR com o objetivo de ver se as variáveis meteorológicas influenciam no resultado das variáveis físico-químicas.

O primeiro passo da análise foi analisar os "*outliers*", isto é, valores atípicos no banco de dados, que são muito comuns em dados ambientais e podem influenciar o resultado da análise. Analisando estes valores discrepantes foi possível perceber que não se tratavam de erros de medidas ou de digitação. Os valores foram considerados corretos e não foram retirados da análise.

O segundo passo da análise foi testar a normalidade de cada variável. Como possuímos muitos *outliers* podemos supor que as variáveis não terão distribuição normal. Mas segundo o capítulo anterior sabemos que a normalidade é importante porém não é necessária para a CANCORR. Apenas deveremos ter mais cuidado na interpretação dos sucessivos testes de significância das raízes canônicas.

Outra suposição que deve ser analisada é a de multicolinearidade. Uma forma bastante prática de sabermos quais variáveis são multicolineares é fazer anteriormente uma análise de cluster baseada na matriz de correlação de Pearson. Esta metodologia é proposta por *Ter Braak* (1986) que diz que quando possuímos multicolinearidade, os efeitos das diferentes variáveis ambientais não podem ser visualizados em separado e, por conseqüência, os coeficientes canônicos tornam-se instáveis. O dendograma mostra quais as variáveis mais correlacionadas. Estas devem ser retiradas e colocadas novamente até que se encontre um modelo adequado que não altere de forma drástica o valor da correlação canônica. Portanto, devemos encontrar o melhor modelo através das várias combinações entre as variáveis colineares.

3.1 Aplicação

Químicas	Meteorológicas
Condutividade (Cond)	Pluviometria (Pluv)
Alcalinidade (Alca)	Precipitação Total (PPT)
H^+ (H)	Q1 – NE
F^- (F)	Q2 – SE
Cl^- (CL)	Q3 – SO
NO_3^- (NO3)	Q4 – NO
SO_4^{2-} (SO4)	Velocidade dos ventos (Vel)
Na^+ (NA)	Pressão (Pres)
NH_4^+ (NH4)	Temperatura (Temp)
K^+ (K)	Umidade relativa (Umid)
Mg^{2+} (MG)	
Ca^{2+} (CA)	

Tabela 3.2 – Variáveis físico-químicas e meteorológicas coletadas.

A tabela 3.2 traz a relação das variáveis físico-químicas e meteorológicas analisadas em cada uma das amostras. A tabela 3.3 é a matriz de correlações de Pearson entre todas as variáveis envolvidas no estudo. Pelas correlações é possível perceber que H^+ possui

correlações extremamente baixas com as demais variáveis. Por esse fato decidimos retirá-la da análise. Podemos perceber também que algumas variáveis possuem altas correlações o que pode indicar colinearidade entre elas, o que será confirmado pela análise de cluster realizada e mostrada na figura 3.1.

	Pluv	H	Cond	Alca	F	CL	NO3	SO4	NA	NH4	K	MG	CA	PPT	Q1	Q2	Q3	Q4	Vel	Pres	Temp	Umid
Pluv	1	0,13	-0,35	-0,10	-0,24	-0,29	-0,23	-0,29	-0,34	-0,29	-0,20	-0,34	-0,27	0,88	-0,09	-0,11	0,24	-0,01	-0,11	0,21	-0,36	0,26
H	0,13	1	-0,02	-0,29	0,29	-0,14	-0,03	-0,11	-0,18	-0,05	-0,16	-0,30	-0,28	0,10	-0,21	0,07	0,19	-0,12	-0,29	0,23	-0,27	0,18
Cond	-0,35	-0,02	1	0,49	0,55	0,44	0,28	0,72	0,36	0,75	0,61	0,65	0,62	-0,30	0,20	0,11	-0,07	-0,09	-0,05	-0,01	0,06	-0,13
Alca	-0,10	-0,29	0,49	1	0,08	0,38	0,01	0,32	0,37	0,40	0,47	0,47	0,45	-0,12	0,07	0,10	-0,06	-0,10	0,13	-0,14	0,22	-0,19
F	-0,24	0,29	0,55	0,08	1	0,15	0,24	0,54	0,07	0,40	0,18	0,40	0,46	-0,23	0,02	0,09	0,01	-0,08	-0,15	0,05	-0,01	-0,15
CL	-0,29	-0,14	0,44	0,38	0,15	1	0,10	0,31	0,91	0,06	0,35	0,61	0,36	-0,30	0,17	0,27	-0,07	-0,24	0,09	-0,02	0,23	-0,24
NO3	-0,23	-0,03	0,28	0,01	0,24	0,10	1	0,32	0,04	0,29	0,08	0,18	0,19	-0,19	-0,10	-0,22	0,01	0,25	0,07	-0,23	0,26	-0,26
SO4	-0,29	-0,11	0,72	0,32	0,54	0,31	0,32	1	0,21	0,42	0,25	0,74	0,80	-0,25	0,21	-0,02	-0,10	0,01	-0,16	0,05	-0,03	-0,09
NA	-0,34	-0,18	0,36	0,37	0,07	0,91	0,04	0,21	1	0,02	0,30	0,57	0,32	-0,33	0,12	0,32	-0,09	-0,27	0,16	-0,13	0,31	-0,31
NH4	-0,29	-0,05	0,75	0,40	0,40	0,06	0,29	0,42	0,02	1	0,56	0,20	0,22	-0,23	0,11	-0,01	-0,04	0,03	-0,07	-0,04	0,02	-0,05
K	-0,20	-0,16	0,61	0,47	0,18	0,35	0,08	0,25	0,30	0,56	1	0,40	0,25	-0,18	0,04	-0,09	0,04	0,14	0,21	-0,14	0,11	-0,24
MG	-0,34	-0,30	0,65	0,47	0,40	0,61	0,18	0,74	0,57	0,20	0,40	1	0,89	-0,31	0,12	0,07	-0,06	-0,04	0,18	-0,19	0,26	-0,33
CA	-0,27	-0,28	0,62	0,45	0,46	0,36	0,19	0,80	0,32	0,22	0,25	0,89	1	-0,24	0,11	-0,01	-0,06	0,02	0,12	-0,14	0,18	-0,29
PPT	0,88	0,10	-0,30	-0,12	-0,23	-0,30	-0,19	-0,25	-0,33	-0,23	-0,18	-0,31	-0,24	1	0,07	-0,11	0,19	0,03	-0,04	0,22	-0,33	0,25
Q1	-0,09	-0,21	0,20	0,07	0,02	0,17	-0,10	0,21	0,12	0,11	0,04	0,12	0,11	0,07	1	0,22	-0,65	-0,11	-0,15	0,32	-0,08	0,22
Q2	-0,11	0,07	0,11	0,10	0,09	0,27	-0,22	-0,02	0,32	-0,01	-0,09	0,07	-0,01	-0,11	0,22	1	-0,41	-0,92	-0,32	0,04	0,25	0,20
Q3	0,24	0,19	-0,07	-0,06	0,01	-0,07	0,01	-0,10	-0,09	-0,04	0,04	-0,06	-0,06	0,19	-0,65	-0,41	1	0,23	0,22	0,03	-0,28	-0,13
Q4	-0,01	-0,12	-0,09	-0,10	-0,08	-0,24	0,25	0,01	-0,27	0,03	0,14	-0,04	0,02	0,03	-0,11	-0,92	0,23	1	0,39	-0,16	-0,13	-0,29
Vel	-0,11	-0,29	-0,05	0,13	-0,15	0,09	0,07	-0,16	0,16	-0,07	0,21	0,18	0,12	-0,04	-0,15	-0,32	0,22	0,39	1	-0,58	0,48	-0,63
Pressão	0,21	0,23	-0,01	-0,14	0,05	-0,02	-0,23	0,05	-0,13	-0,04	-0,14	-0,19	-0,14	0,22	0,32	0,04	0,03	-0,16	-0,58	1	-0,82	0,54
Temp	-0,36	-0,27	0,06	0,22	-0,01	0,23	0,26	-0,03	0,31	0,02	0,11	0,26	0,18	-0,33	-0,08	0,25	-0,28	-0,13	0,48	-0,82	1	-0,49
Umid	0,26	0,18	-0,13	-0,19	-0,15	-0,24	-0,26	-0,09	-0,31	-0,05	-0,24	-0,33	-0,29	0,25	0,22	0,20	-0,13	-0,29	-0,63	0,54	-0,49	1

Tabela 3.3 – Tabela de Correlação de Pearson de todas as variáveis.

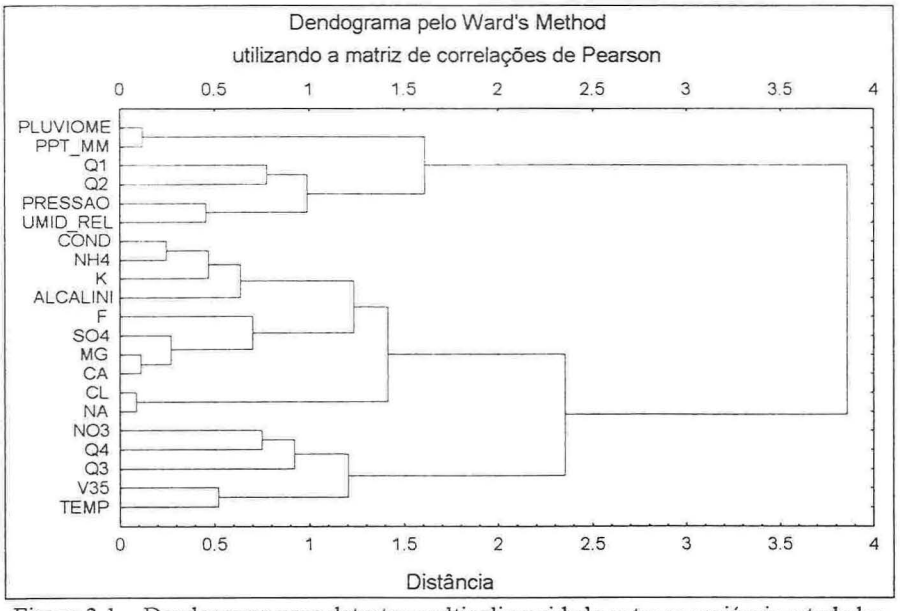


Figura 3.1 – Dendograma para detectar multicolinearidade entre as variáveis estudadas.

As variáveis que se mostraram colineares, tanto pelo dendograma como pelas correlações, foram: pluviometria e precipitação total (PPT), Mg^{2+} e Ca^{2+} e Cl^{-} e Na^{+}.

Agora que sabemos quais variáveis são colineares foram procedidas análises canônicas com todas as combinações possíveis dessas variáveis colineares e as demais variáveis.

A tabela 3.4 traz um resumo das análises feitas com as combinações das variáveis colineares contendo o valor da correlação canônica, o índice de redundância total das variáveis químicas e o índice de redundância total das variáveis meteorológicas.

	Correlação Canônica	Redundância Total Química	Redundância Total Meteorológica
Todas as variáveis	0.681	22.708%	25.674%
1) PPT – Ca^{2+} – Na^{+}	0.672	20.095%	23.586%
2) PPT – Ca^{2+} – Cl^{-}	0.673	19.752%	22.213%
3) PPT – Mg^{2+} – Na^{+}	0.662	20.835%	23.086%
4) PPT – Mg^{2+} – Cl^{-}	0.665	20.429%	22.903%
5) Pluv – Ca^{2+} – Na^{+}	0.673	20.805%	24.214%
6) Pluv – Ca^{2+} – Cl^{-}	0.673	20.387%	22.746%
7) Pluv – Mg^{2+} – Na^{+}	0.660	21.445%	23.693%
8) Pluv – Mg^{2+} – Cl^{-}	0.666	21.028%	23.443%

Tabela 3.4 – Resumo das combinações possíveis para escolha do modelo final a ser analisado.

Pelos resultados obtidos optamos pela combinação 5 que foi a combinação que teve menor diminuição da correlação canônica e ao mesmo tempo dos índices de redundância. Outras combinações também obtiveram pequenas diminuições no valor da correlação canônica, mas obtiveram diminuições maiores nos índices de redundância.

Dessa forma optamos por retirar da análise as variáveis PPT, Mg^{2+} e Cl^{-} da análise.

3.2 Análise completa dos dados

As análises descritivas das variáveis meteorológicas e químicas que serão consideradas na análise, aparecem, respectivamente, nas tabelas 3.5 e 3.6. A tabela 3.7 traz o resumo da CANCORR para as variáveis selecionadas.

Estatísticas Descritivas Meteorológicas		
Variável	Média	Desvio Padrão
Pluv	46,352	27,851
Q1	34,885	23,938
Q2	63,392	25,380
Q3	41,402	23,812
Q4	27,123	23,346
Vel	14,846	3,796
Pres	1014,281	3,949
Temp	19,643	3,581
Umid	83,195	7,029

Tabela 3.5 – Estatísticas Descritivas Meteorológicas

Estatísticas Descritivas Químicas		
Variável	Média	Desvio Padrão
Cond	11,315	6,766
Alca	16,095	25,871
F	6,145	5,782
NO3	4,243	4,994
SO4	24,545	16,899
NA	16,016	16,221
NH4	40,619	37,437
K	6,349	9,907
CA	23,908	27,273

Tabela 3.6 – Estatística Descritivas Físico-Químicas

Resumo da CANCORR		
R Canônico: 0.67267		
Qui-quadrado(81)=306.58 p=0.0000		
	Meteorológicas	Químicas
Nº de variáveis	9	9
Variância extraída	100.000%	100.000%
Redundância Total	24.2144%	20.8047%
Variáveis		
1	Pluv	Cond
2	Q1 - NE	Alca
3	Q2 - SE	F
4	Q3 - SO	NO3
5	Q4 - NO	SO4
6	Vel	NA
7	Pres	NH4
8	Temp	K
9	Umid	CA

Tabela 3.7 – Resumo da CANCORR.

Na tabela 3.7 são apresentados o valor da primeira correlação canônica, o valor do teste qui-quadrado com sua respectiva significância, a variância extraída por todas as variáveis canônicas criadas (neste caso 9 variáveis canônicas pois é o menor número de variáveis entre os dois conjuntos de variáveis), o índice de redundância total extraído pelas 9

variáveis canônicas e as variáveis que compõem cada conjunto. Na tabela 3.8 aparecem os autovalores.

	Autovalores								
	Raiz 1	Raiz 2	Raiz 3	Raiz 4	Raiz 5	Raiz 6	Raiz 7	Raiz 8	Raiz 9
Valor	0,452	0,436	0,315	0,129	0,076	0,051	0,014	0,003	0,001

Tabela 3.8 – Autovalores encontrados (raízes canônicas).

Os autovalores são uma estimativa da variância compartilhada entre as variáveis canônicas, ou seja, as duas primeiras variáveis canônicas possuem 45,25% das suas variâncias compartilhadas e assim para as demais funções. Percebe-se que até a 4ª raiz canônica o valor da variância compartilhada é alto. Mas devemos ter cuidado pois essa variância compartilhada é entre as variáveis canônicas e não entre as variáveis originais. Como a correlação canônica é simplesmente a raiz quadrada do autovalor, até a 4ª raiz a correlação canônica também é alta.

Para escolher quais variáveis canônicas devem ser interpretadas o capítulo 2 descreve três formas de escolha das mesmas. Pelo valor das correlações acredita-se que as 4 primeiras raízes devam ser analisadas. Mas sabemos que temos que olhar o significado prático das mesmas e suas significâncias.

Pelo teste qui-quadrado da tabela 3.9 as três primeiras raízes canônicas são as raízes que devem ser interpretadas. Esse teste testa globalmente todas as raízes, num primeiro estágio. Dado que o resultado foi significativamente diferente de zero, sabemos que existe pelo menos uma raiz não nula. Dessa forma, num segundo estágio, a primeira raiz canônica é retirada (pois é a maior correlação) e novamente é feito o teste. Dado que o resultado foi significativo, sabemos que a segunda raiz canônica também é significativa. Dessa forma o teste procede até que sobre apenas uma raiz canônica.

Não podemos esquecer que como não temos normalidade dos dados não podemos confiar totalmente nos resultados do teste.

Teste Qui-quadrado para remoção sucessiva das raízes						
	R Canônico	R^2 Canônico	Qui quadrado	gl	p	Lambda Prime
0	0,673	0,452	306,585	81	0,000	0,159
1	0,660	0,436	206,290	64	0,000	0,290
2	0,562	0,315	111,027	49	0,000	0,513
3	0,360	0,129	47,919	36	0,089	0,750
4	0,276	0,076	24,841	25	0,471	0,861
5	0,225	0,051	11,686	16	0,765	0,932
6	0,118	0,014	3,017	9	0,964	0,982
7	0,058	0,003	0,678	4	0,954	0,996
8	0,027	0,001	0,118	1	0,731	0,999

Tabela 3.9 – Teste Qui-quadrado para remoção sucessiva de raízes canônicas.

Outra forma de escolher quais variáveis canônicas devem ser interpretadas é pelo significado real das cargas canônicas cruzadas que são a correlação entre cada variável com sua variável canônica oposta. Essa terceira forma de seleção das variáveis canônicas faz com que selecionemos as três primeiras variáveis. As tabelas 3.10, 3.11 e 3.12 trazem os pesos canônicos, as cargas canônicas (estrutura canônica) e as cargas canônicas cruzadas respectivamente.

Os pesos canônicos são os pesos encontrados que maximizam a correlação canônica, ou seja, se aplicarmos esses pesos às variáveis originais iremos obter um escore para as variáveis químicas e outro para as variáveis meteorológicas (variáveis canônicas químicas e meteorológicas) que possuirão correlação máxima. Como os pesos canônicos são difíceis de analisar, optamos por calcular os pesos canônicos cruzados e selecionar o número de raízes a partir deles.

a'_k ; b'_k ;

Pesos Canônicos Meteorológicos				Pesos Canônicos Químicos			
	Raiz 1	Raiz 2	Raiz 3		Raiz 1	Raiz 2	Raiz 3
Pluv	0,169	-0,610	-0,381	Cond	-0,768	0,158	-0,930
Q1 - NE	-0,322	0,586	0,390	Alca	0,114	-0,171	-0,281
Q2 - SE	0,027	0,149	-1,028	F	0,073	0,271	-0,086
Q3 - SO	-0,003	0,510	0,098	NO3	0,724	-0,129	0,381
Q4 - NO	0,453	-0,449	-0,122	SO4	-0,942	0,020	1,046
Vel	0,168	-0,342	-0,962	NA	0,207	0,847	-0,200
Pres	0,636	-0,088	-0,587	NH4	0,110	0,404	0,595
Temp	1,127	-0,041	-0,097	K	0,434	-0,307	0,314
Umid	-0,296	-0,611	-0,240	CA	1,031	0,088	-0,003

Tabela 3.10 – Pesos canônicos meteorológicos e físico-químicos

Cargas Canônicas Meteorológicas			
	Raiz 1	Raiz 2	Raiz 3
Pluv	-0,176	-0,672	-0,317
Q1 - NE	-0,358	0,282	0,060
Q2 - SE	-0,280	0,521	-0,621
Q3 - SO	0,120	-0,168	-0,025
Q4 - NO	0,369	-0,461	0,607
Vel	0,720	-0,052	-0,230
Pres	-0,680	-0,039	-0,063
Temp	0,746	0,290	-0,120
Umid	-0,762	-0,354	-0,099

Cargas Canônicas Químicas			
	Raiz 1	Raiz 2	Raiz 3
Cond	-0,096	0,673	0,314
Alca	0,229	0,307	-0,090
F	-0,064	0,532	0,314
NO3	0,499	0,129	0,623
SO4	-0,173	0,530	0,637
NA	0,274	0,804	-0,295
NH4	-0,099	0,398	0,471
K	0,237	0,257	0,162
CA	0,225	0,512	0,307

Tabela 3.11 – Cargas canônicas meteorológicas e químcas.

Cargas Canônicas Cruzadas Meteorológicas			
	Raiz 1	Raiz 2	Raiz 3
Pluv	-0,118	-0,444	0,178
Q1 - NE	-0,241	0,186	-0,034
Q2 - SE	-0,189	0,344	0,349
Q3 - SO	0,081	-0,111	0,014
Q4 - NO	0,248	-0,304	-0,341
Vel	0,484	-0,034	0,129
Pres	-0,457	-0,026	0,035
Temp	0,502	0,192	0,068
Umid	-0,513	-0,234	0,055

Cargas Canônicas Cruzadas Químicas			
	Raiz 1	Raiz 2	Raiz 3
Cond	-0,065	0,444	-0,176
Alca	0,154	0,203	0,051
F	-0,043	0,351	-0,176
NO3	0,335	0,085	-0,350
SO4	-0,116	0,350	-0,358
NA	0,184	0,531	0,166
NH4	-0,066	0,263	-0,265
K	0,159	0,170	-0,091
CA	0,151	0,338	-0,172

Tabela 3.12 – Cargas canônicas cruzadas meteorológicas e químicas.

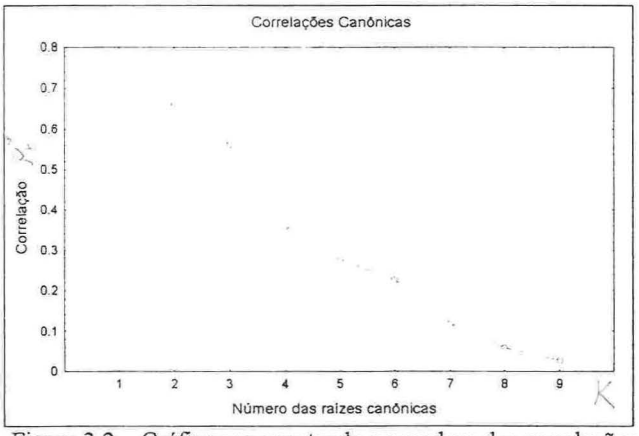


Figura 3.2 – Gráfico representando a grandeza das correlações canônicas

Pelo gráfico 3.2 podemos ver que da terceira para a quarta há um decréscimo grande, o que mais uma vez comprova que devemos analisar as três primeiras raízes. Na tabela 3.13 aparecem as variâncias extraídas e índices de redundância para cada variável.

Meteorológicas			Químicas		
	Variância Extraída	Redundância		Variância Extraída	Redundância
Raiz 1	0,279	0,126	Raiz 1	0,060	0,027
Raiz 2	0,140	0,061	Raiz 2	0,251	0,109
Raiz 3	0,104	0,033	Raiz 3	0,159	0,050
Raiz 4	0,091	0,012	Raiz 4	0,081	0,010
Raiz 5	0,082	0,006	Raiz 5	0,074	0,006
Raiz 6	0,053	0,003	Raiz 6	0,085	0,004
Raiz 7	0,087	0,001	Raiz 7	0,044	0,001
Raiz 8	0,101	0,000	Raiz 8	0,128	0,000
Raiz 9	0,063	0,000	Raiz 9	0,118	0,000

Tabela 3.13 – Variância extraída e índices de redundância para cada variável

Os índices de redundância extraídos pelas três primeiras raízes foram 0.220 para as variáveis meteorológicas que representam 90,8% da redundância total das variáveis meteorológicas e 0,187 para as variáveis químicas que representam 89.6% da redundância total das variáveis químicas.

Os gráficos das cargas canônicas cruzadas (Figura 3.3) mostra como se dão as relações entre as variáveis para cada raiz canônica. O *scatterplot* entre as variáveis canônicas duas a duas mostram como se dão as relações entre as mesmas.

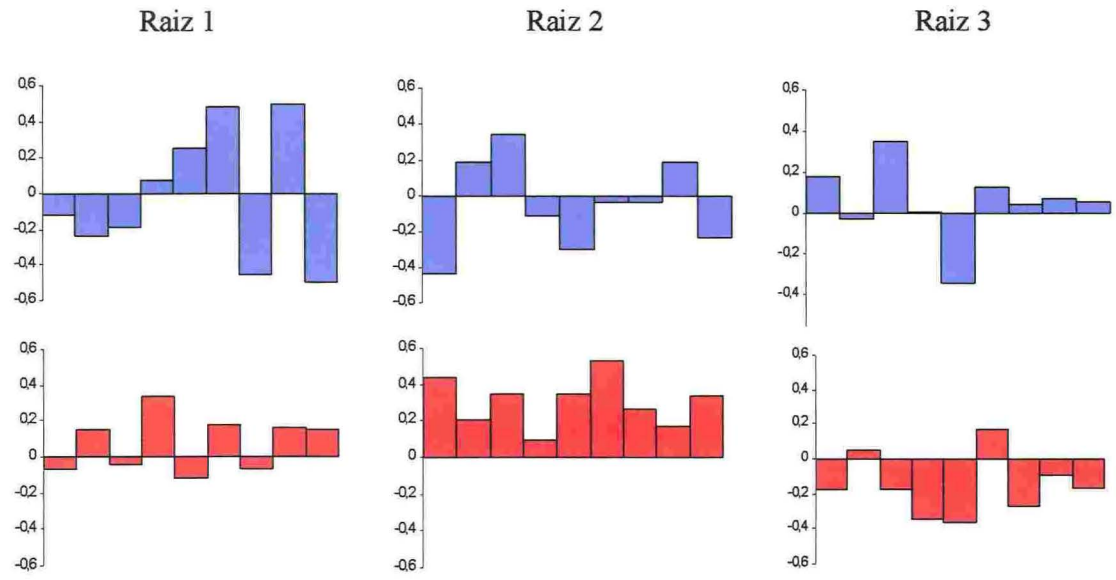
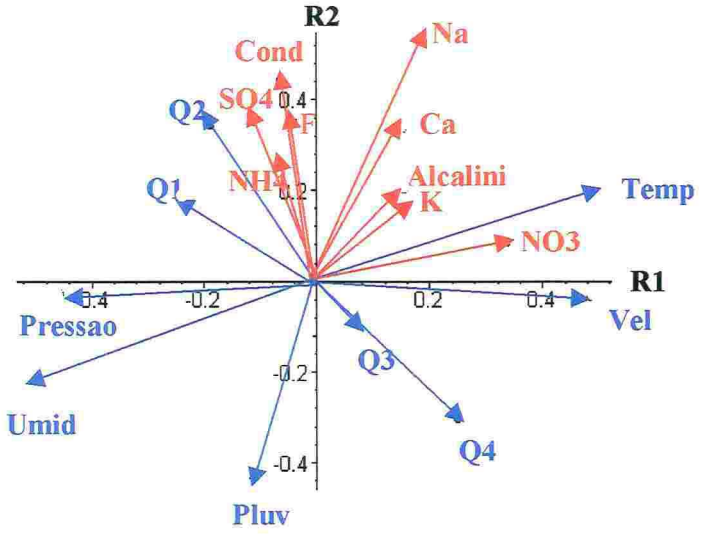


Figura 3.3 - cargas canônicas cruzadas para as variáveis meteorológicas e ambientais. As variáveis representadas são: Pluvi, Q1, Q2, Q3, Q4, Vel, Pres, Temp e Umid para as meteorológicas e Cond, Alcalini, F, NO3, SO4, Na, NH4, K e Ca para as químicas.

Nas figuras 3.4, 3.5 e 3.6, faz-se a representação das cargas canônicas cruzadas considerando as raízes canônicas, combinadas 2 a 2.

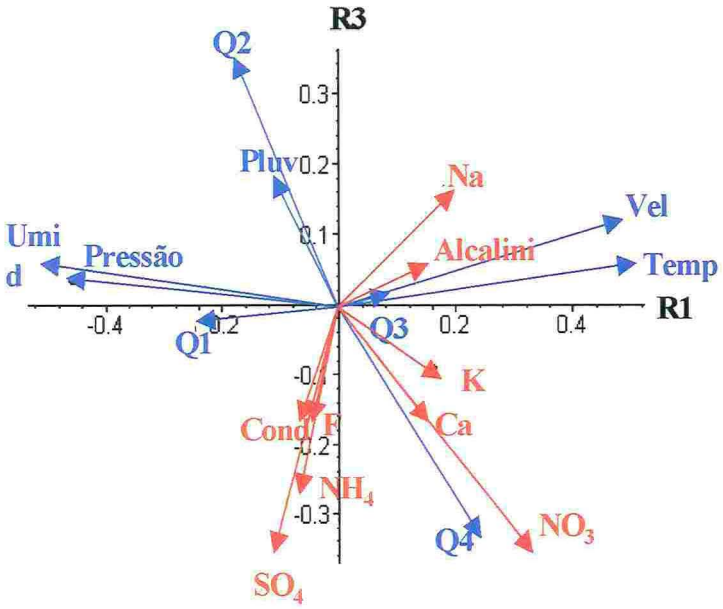
Cargas cruzadas R1 e R2



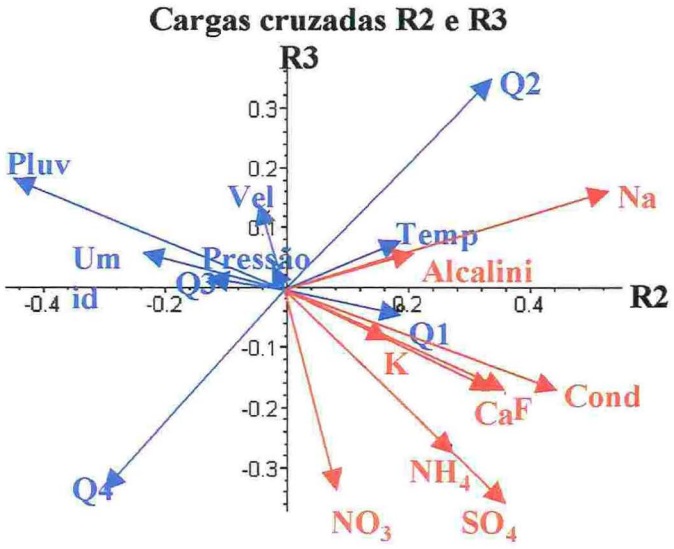
O gráfico mostra quais variáveis se relacionam nas funções canônicas 1 e 2.

Existe uma relação inversamente proporcional entre as variáveis Na e Ca com a Pluviometria e também entre Condutividade, SO4 e F com os quadrantes 3 e 4.

Cargas cruzadas R1 e R3



Este gráfico mostrou uma associação do NO3 positiva com o quadrante 4 e negativa com Q2. Na Raiz 3, SO_4 e NH_4 se associa inversamente com Q2.



O gráfico das raízes 2 e 3 apresentou associação direta entre Q2 e Na e inversa entre Pluv e SO4, NH4, Cond, etc.

3.3 Interpretação dos resultados

Foram extraídas para a interpretação da análise as três primeiras raízes canônicas como explicado anteriormente na seção 3.2.

Pela tabela 3.12 vemos que a Raiz 1 apresentou cargas canônicas cruzadas significativas para as variáveis meteorológicas (Vel, Pres, Temp, Umid) não apresentando cargas canônicas cruzadas significativas nas variáveis químicas. A ausência de correlação entre as variáveis físico-químicas e meteorológicas na Raiz 1 pode estar relacionada ao conjunto de dados analisados, uma vez que as amostras foram coletadas a partir de dois tipos de amostradores (Bulk e Automático). Outro motivo que pode ter causado essa anomalia é o fato de os dados meteorológicos não terem sido coletados nos locais das amostragens e sim no Aeroporto Salgado Filho.

As raízes canônicas 2 e 3 apresentaram cargas canônicas mais significativas nas variáveis químicas onde podemos associar a raiz 2 química, com cargas elevadas para Cond, F, Na e Ca, a fontes naturais, principalmente o Na que pode estar relacionado à presença de sais marinhos na precipitação atmosférica coletada na região de estudo. O Ca pode ter origem da resuspensão da poeira do solo.

A correlação inversa entre a pluviometria e as variáveis químicas sugere que quanto maior a pluviometria menor a concentração das variáveis químicas correlacionadas.

A raiz canônica 3 apresentou cargas canônicas elevadas para as variáveis químicas NO_3 e SO_4, que são provenientes de fontes antrópicas na região, ou seja, são provenientes de ações do Homem. Essa variáveis com origem antropogênica podem estar relacionadas à queima de combustíveis fósseis (o carvão mineral) para a geração de energia elétrica e à emissões veiculares.

O NH_{4+} presente na atmosfera pode ser emitido por diversas fontes, incluindo volatização de resíduos animais (fezes de animais que se decompõe e vão para a atmosfera), excrementos humanos, perda natural pela vegetação, queima de biomassa e também pode ser decorrente de processos industriais, como o uso ou fabricação de fertilizantes e de emissões da combustão de combustíveis fósseis..

Estas variáveis químicas correlacionaram-se inversamente com a variável meteorológica Q2 – SE (direção do vento sudeste) e diretamente com Q4 – NO (direção do vento noroeste). Isto pode ser explicado pelas circulações atmosféricas que ocorrem na Região Metropolitana de Porto Alegre (RMPA) onde os ventos provenientes do quadrante NO são ventos pré-frontais que podem coincidir com as primeiras ocorrências de pluviometria associadas a frentes frias e ou linhas de instabilidade que atuam na região. Isto explica a correlação direta desta direção (NO) com as variáveis químicas SO4 e NO3.

Já o vento do quadrante SE ocasiona ocorrência de chuva pós-frontal a qual é caracterizada por uma atmosfera limpa. O vento de SE na RMPA normalmente são acompanhados de chuva no inverno, com aproximação das massas polar marítima que atuam na região. Desta forma, podemos explicar a correlação inversa da direção com as variáveis de fonte antropogênica.

4 COMPARAÇÃO DE *SOFTWARES* ESTATÍSTICOS:

A fim de saber quais as diferenças entre alguns *softwares* estatísticos, foi utilizado o banco de dados do capítulo anterior. Os *softwares* comparados foram o STATISTICA for Windows versão 4.3, SAS System versão 8.2 e SPSS versão 8.0. Claro que existem outros *softwares* estatísticos que realizam este tipo de análise como é o caso do R.

Os três *softwares* possuem a possibilidade de importar dados diretamente do Excel, o que facilita na construção de bancos de dados. Será mostrado como importar o banco de dados do Excel. Mas os *softwares* STATISTICA e SPSS exigem que o banco de dados no Excel, esteja salvo na versão 4.0 e que o arquivo não possua mais nenhuma planilha. Os três *softwares* têm dificuldades em ler os nomes das variáveis que possuem acentos ou símbolos, portanto devemos ter o cuidado de dar nomes simples às variáveis.

Para cada um dos softwares serão apresentados comentários, o caminho a seguir para obtenção da CANCORR e os resultados. Como os dados utilizados já foram analisados serão apresentadas sempre as três primeiras raízes canônicas.

4.1 STATISTICA

Este software tem um *layout* não usual em relação aos demais, isto é, a forma de utilização dos comandos é bastante diferente. Qualquer tipo de análise é feita a partir de uma janela que possui várias opções (Figura 4.1). Este menu será explicado com maiores detalhes.

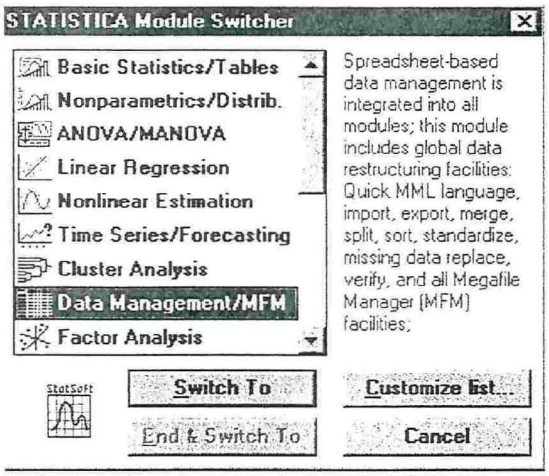


Figura 4.1 – Menu principal do STATISTICA

Para importar uma banco de dados devemos clicar em **Data Management/MFM** e no botão **Switch To** da janela inicial do programa (Figura 4.1). Ao clicar em **Switch To** outra janela será aberta com algumas opções que aparecem na Figura 4.2. Nessa janela, devemos clicar em **Import forein data file** e **OK**.

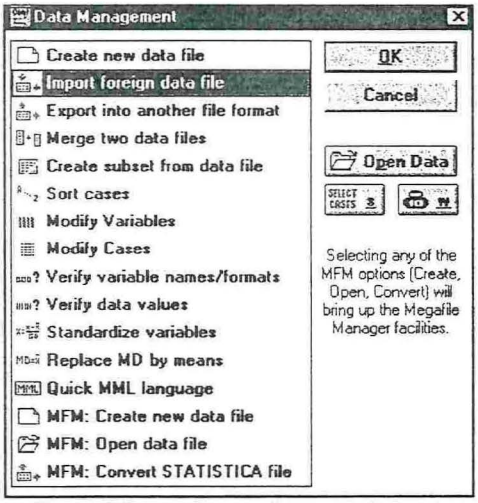


Figura 4.2 – opção para importar dados de outros *softwares*

Após apertar **OK** devemos selecionar o arquivo que queremos abrir. Depois de escolher o arquivo outra janela será aberta, onde devemos selecionar o formato do arquivo que será importado e clicar em **Options** para selecionar algumas opções possíveis para a leitura do banco de dados como, por exemplo, ler o nome das variáveis na primeira linha do banco, importar o número (ou nome) dos casos, etc: A Figura 4.3 apresenta esses

procedimentos. Note que, se o banco que for importado estiver no Excel (como na Figura 4.3), deverá ser salvo na versão 4.0 e não deverá ter acento em nenhuma de suas variáveis..

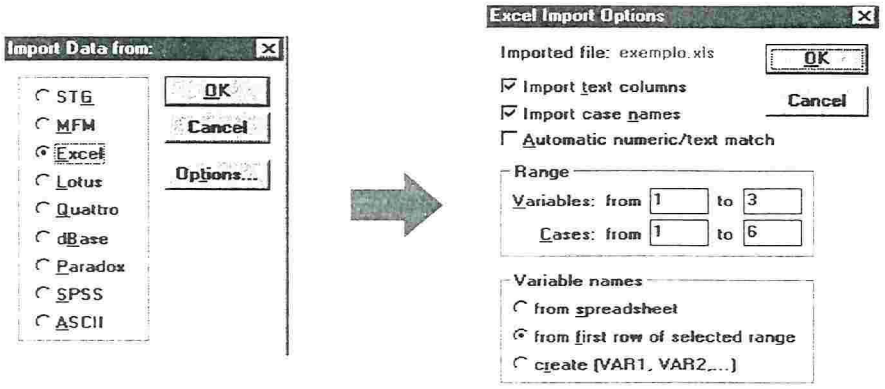


Figura 4.3 – opções para abrir um banco de dados do Excel.

Após todas essas etapas, o banco de dados será aberto, mas antes o programa irá abrir uma janela para salvar o arquivo no seu formato (.STA). Finalmente o banco de dados será aberto, agora já com a terminação .STA.

Para calcular a CANCORR devemos entrar no menu Analysis da barra de ferramentas e clicar em **Other statistics** (como mostra a figura 4.4) que abrirá uma janela com as possibilidades de análises estatísticas que o software apresenta (Figura 4.5).

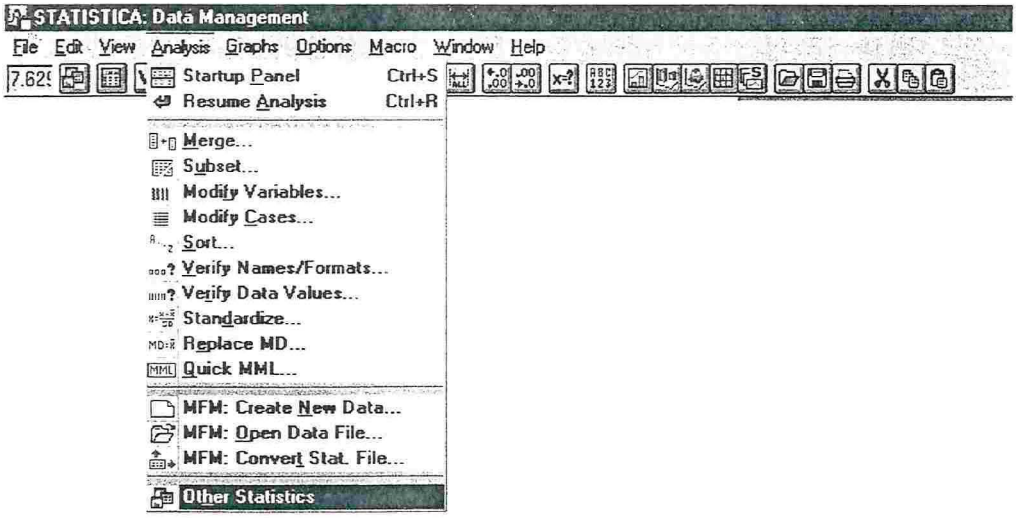


Figura 4.4 – Abrindo o menu de análises estatísticas.

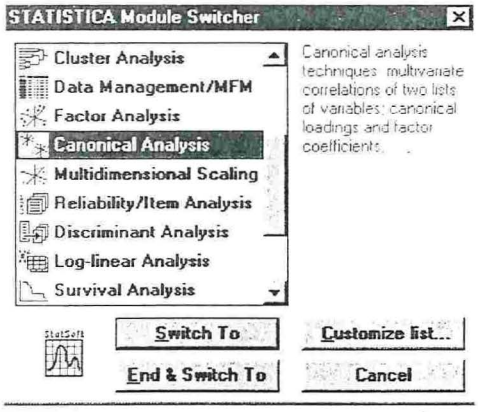


Figura 4.5 – Opções de análises estatísticas

A Figura 4.5 mostra o mesmo menu utilizado para abrir o banco de dados, mas agora iremos utilizá-lo para calcular a CANCORR através do comando **Canonical Analysis** e **Switch To.**.

O caminho a seguir é dado pela Figuras 4.6 e 4.7. Na Figura 4.6 devemos marcar todas as variáveis que entrarão na análise. A figura 4.7 mostra o menu onde devemos selecionar quais das variáveis marcadas na figura 4.6 irão compor cada um dos conjuntos de dados. Ainda neste menu é possível pedir as médias e desvios padrões de todas as variáveis, a matriz de correlações de Pearson, os *boxplots* e uma matriz de dados que traz o histograma de cada variável e seu *scatterplot* com todas as demais variáveis. Esse gráfico é muito interessante e está exemplificado na figura 4.8.

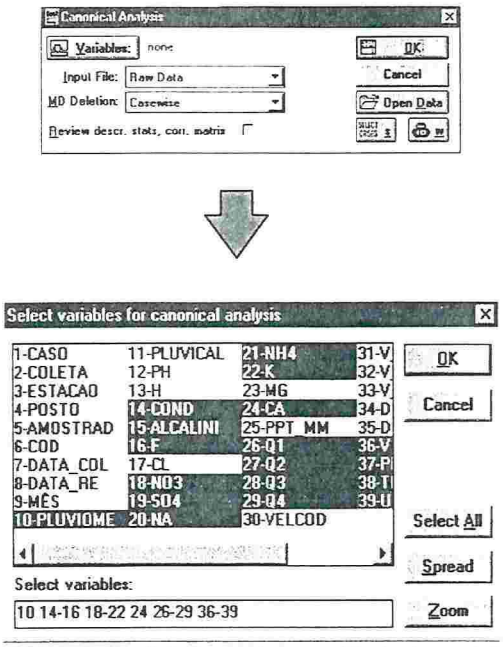


Figura 4.6 – Selecionando as variáveis do modelo.

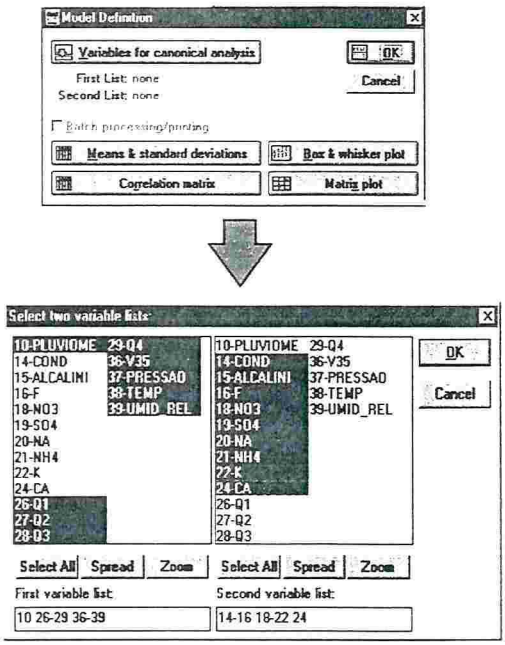


Figura 4.7 – Selecionando as variáveis de cada conjunto de dados

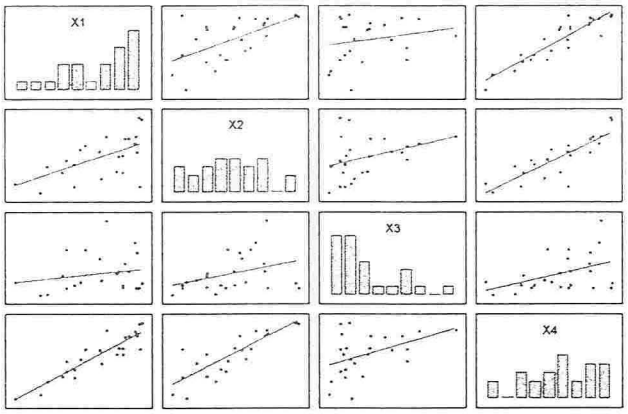


Figura 4.8 – Histograma de cada variável e seu *scatterplot* com as demais variáveis.

Depois selecionar quais variáveis farão parte de cada conjunto de dados, apertando **OK** chegamos a um novo menu (Figura 4.9) que é o menu da análise de correlação canônica que já apresenta um resumo dos resultados da análise. Cada uma das opções deste menu será explicada e serão apresentados os resultados da análise.

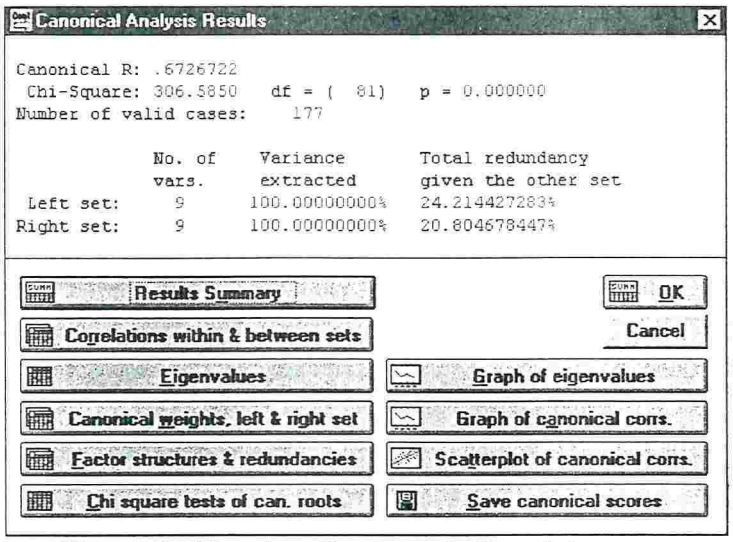


Figura 4.9 – Menu da análise de correlação canônica.

Results Summary: Tabela resumo da análise de correlação canônica.

Correlations with & between sets:. Matriz de correlação de Pearson. É possível escolher quais as variáveis que se quer as correlações.

Eigenvalues: Autovalores, que são as correlações canônicas elevadas ao quadrado.

Canonical weights, left & right set: Pesos canônicos.

Factor structure & redundancies: Cargas canônicas, ou estrutura canônica.

Chi square tests of can. roots: Teste qui-quadrado para remoção sucessiva de raízes canônicas.

Graph os eigenvalues: Gráfico dos autovalores.

Graph of canonical corrs: Gráfico da magnitude das correlações canônicas.

Scatterplot of canonical corrs: Scatterplots entre variáveis canônicas.

Save canonical scores: essa opção salva os escores canônicos obtidos através dos pesos canônicos. Quando pedimos para salvar esses escores, podemos escolher quais variáveis do banco de dados original queremos que fiquem salvas junto. O programa criará um novo banco de dados com os escores e as variáveis selecionadas. Devemos nomear esse novo banco de dados. A partir desses escores podemos calcular as cargas canônicas cruzadas.

Resultados do STATISTICA:

Resumo da CANCORR		
R Canônico: 0.67267		
Qui-quadrado(81)=306.58 p=0.0000		
	Meteorológicas	Químicas
Nº de variáveis	9	9
Variância extraída	100.000%	100.000%
Redundância Total	24.2144%	20.8047%
Variáveis		
1	Pluv	Cond
2	Q1 - NE	Alca
3	Q2 - SE	F
4	Q3 - SO	NO3
5	Q4 - NO	SO4
6	Vel	NA
7	Pres	NH4
8	Temp	K
9	Umid	CA

Autovalores									
	Raiz 1	Raiz 2	Raiz 3	Raiz 4	Raiz 5	Raiz 6	Raiz 7	Raiz 8	Raiz 9
Valor	0,452	0,436	0,315	0,129	0,076	0,051	0,014	0,003	0,001

Pesos Canônicos Meteorológicos			
	Raiz 1	Raiz 2	Raiz 3
Pluv	0,169	-0,610	-0,381
Q1 - NE	-0,322	0,586	0,390
Q2 - SE	0,027	0,149	-1,028
Q3 - SO	-0,003	0,510	0,098
Q4 - NO	0,453	-0,449	-0,122
Vel	0,168	-0,342	-0,962
Pres	0,636	-0,088	-0,587
Temp	1,127	-0,041	-0,097
Umid	-0,296	-0,611	-0,240

Pesos Canônicos Químicos			
	Raiz 1	Raiz 2	Raiz 3
Cond	-0,768	0,158	-0,930
Alca	0,114	-0,171	-0,281
F	0,073	0,271	-0,086
NO3	0,724	-0,129	0,381
SO4	-0,942	0,020	1,046
NA	0,207	0,847	-0,200
NH4	0,110	0,404	0,595
K	0,434	-0,307	0,314
CA	1,031	0,088	-0,003

Cargas Canônicas Meteorológicas			
	Raiz 1	Raiz 2	Raiz 3
Pluv	-0,176	-0,672	-0,317
Q1 - NE	-0,358	0,282	0,060
Q2 - SE	-0,280	0,521	-0,621
Q3 - SO	0,120	-0,168	-0,025
Q4 - NO	0,369	-0,461	0,607
Vel	0,720	-0,052	-0,230
Pres	-0,680	-0,039	-0,063
Temp	0,746	0,290	-0,120
Umid	-0,762	-0,354	-0,099

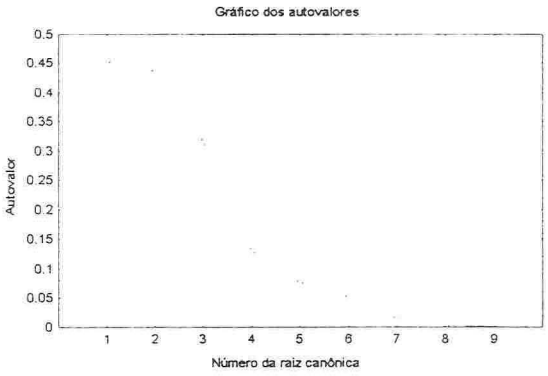
Cargas Canônicas Químicas			
	Raiz 1	Raiz 2	Raiz 3
Cond	-0,096	0,673	0,314
Alca	0,229	0,307	-0,090
F	-0,064	0,532	0,314
NO3	0,499	0,129	0,623
SO4	-0,173	0,530	0,637
NA	0,274	0,804	-0,295
NH4	-0,099	0,398	0,471
K	0,237	0,257	0,162
CA	0,225	0,512	0,307

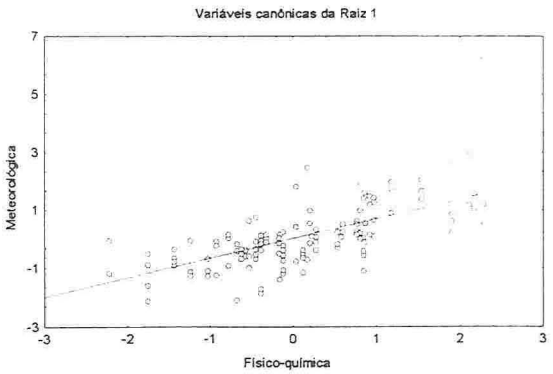
Meteorológicas		
	Variância Extraída	Redundância
Raiz 1	0,279	0,126
Raiz 2	0,140	0,061
Raiz 3	0,104	0,033
Raiz 4	0,091	0,012
Raiz 5	0,082	0,006
Raiz 6	0,053	0,003
Raiz 7	0,087	0,001
Raiz 8	0,101	0,000
Raiz 9	0,063	0,000

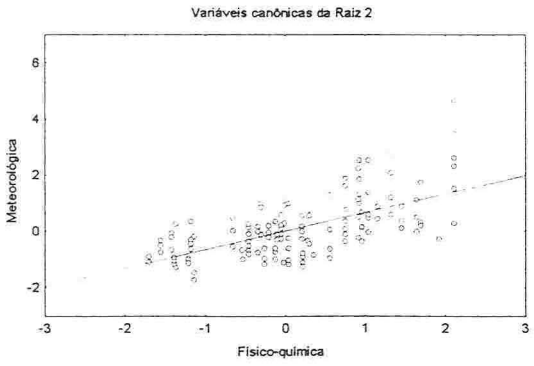
Químicas		
	Variância Extraída	Redundância
Raiz 1	0,060	0,027
Raiz 2	0,251	0,109
Raiz 3	0,159	0,050
Raiz 4	0,081	0,010
Raiz 5	0,074	0,006
Raiz 6	0,085	0,004
Raiz 7	0,044	0,001
Raiz 8	0,128	0,000
Raiz 9	0,118	0,000

Teste Qui-quadrado para remoção sucessiva das raízes

	R Canônico	R^2 Canônico	Qui quadrado	gl	p	Lambda Prime
0	0,673	0,452	306,585	81	0,000	0,159
1	0,660	0,436	206,290	64	0,000	0,290
2	0,562	0,315	111,027	49	0,000	0,513
3	0,360	0,129	47,919	36	0,089	0,750
4	0,276	0,076	24,841	25	0,471	0,861
5	0,225	0,051	11,686	16	0,765	0,932
6	0,118	0,014	3,017	9	0,964	0,982
7	0,058	0,003	0,678	4	0,954	0,996
8	0,027	0,001	0,118	1	0,731	0,999







4.2 SAS System:

Esse *software* possui um menu fácil de ser utilizado uma vez que o banco de dados esteja aberto. Para abrí-lo temos que, primeiramente, abrir uma planilha da seguinte forma: **Solutions / Analysis / Analyst** mostrada na Figura 4.10. Uma vez aberta a planilha, abrir o banco de dados do Excel é simples, basta ir na barra de ferramentas no menu **File / Open** e encontrar o arquivo que se quer abrir no formato escolhido. Mas note que o banco de dados não pode ter variáveis que possuem acentos. Antes de abrir o banco de dados aparece uma janela para escolhermos qual a planilha que desejamos utilizar. Essa é uma diferença deste *software* para os demais, é possível ter mais de uma planilha ativa no banco de dados do Excel (Figura 4.11).

Para realizar a CANCORR segue-se o seguinte comando: **Statistics / Multivariate / Canonical Correlation...** (Figura 4.12).

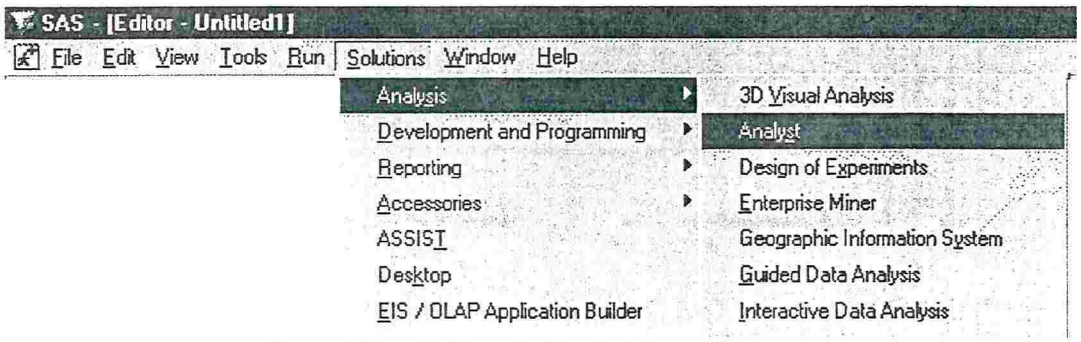


Figura 4.10 – Abridno planilha de dados no SAS.

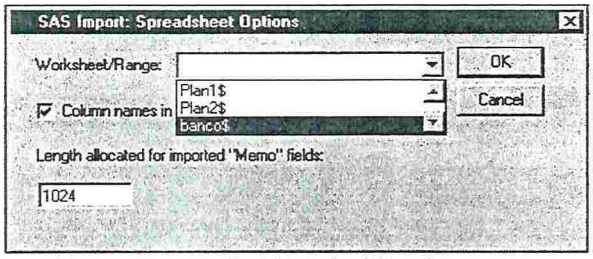


Figura 4.11 – escolhendo a planilha ativa

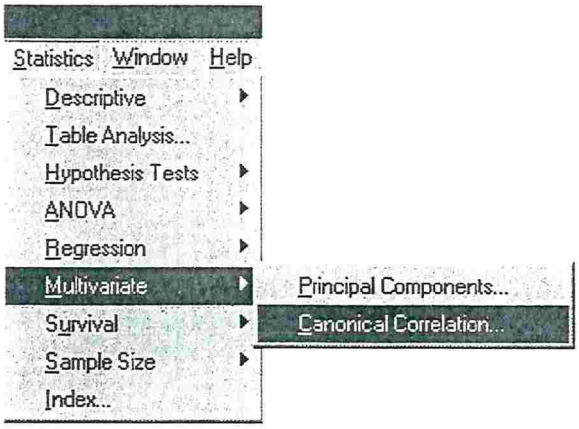


Figura 4.12 – menu da análise da CANCORR.

O próximo passo é colocar cada variável no seu conjunto correspondente (Figura 4.13).

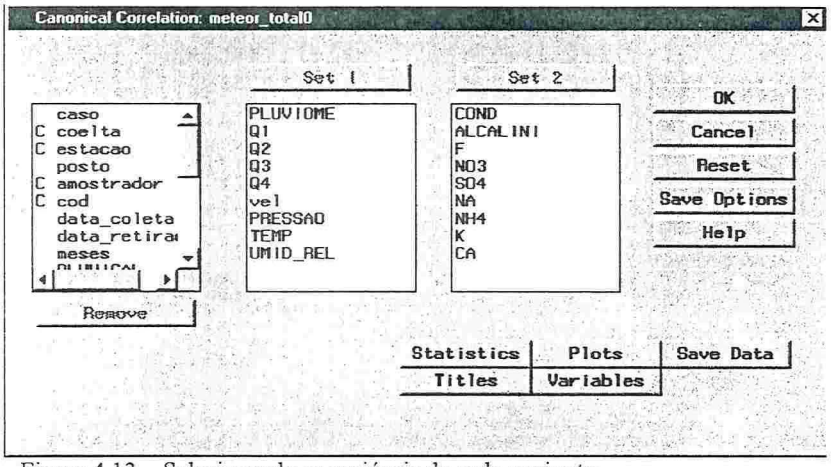


Figura 4.13 – Selecionando as variáveis de cada conjunto.

Statistics: Neste menu é possível atribuir um nome e um prefixo para cada conjunto, por exemplo meteorológico e químco. Isso ajuda na interpretação da saída (Figura 4.14).

Plots: Cria gráficos das variáveis canônicas. Pode-se escolher até qual relação se quer plotar.

Save data: Cria e salva os escores.

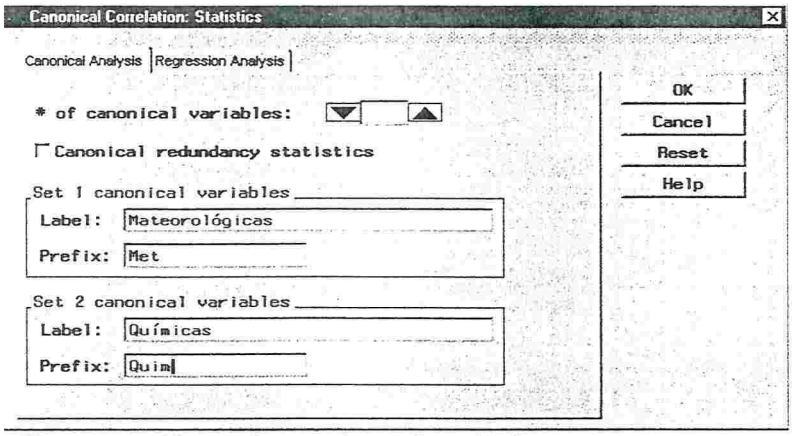


Figura 4.14 – Nomeando os conjuntos de variáveis.

Resultados do SAS:

The CANCORR Procedure

Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.672672	.	0.041270	0.452488
2	0.660067	.	0.042537	0.435689
3	0.561672	0.534211	0.051598	0.315476
4	0.359760	0.278187	0.065622	0.129427
5	0.275627	0.191916	0.069651	0.075970
6	0.225237	.	0.071554	0.050732
7	0.118114	.	0.074326	0.013951
8	0.057936	.	0.075125	0.003357
9	0.026640	.	0.075324	0.000710

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	0.8264	0.0544	0.3499	0.3499
2	0.7721	0.3112	0.3269	0.6768
3	0.4609	0.3122	0.1951	0.8719
4	0.1487	0.0665	0.0629	0.9348
5	0.0822	0.0288	0.0348	0.9697
6	0.0534	0.0393	0.0226	0.9923
7	0.0141	0.0108	0.0060	0.9983
8	0.0034	0.0027	0.0014	0.9997
9	0.0007		0.0003	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.15860296	4.22	81	1036.5	<.0001
2	0.28967936	3.48	64	929.35	<.0001
3	0.51333263	2.35	49	821.79	<.0001
4	0.74991174	1.34	36	714.15	0.0886
5	0.86140009	1.00	25	607.02	0.4715
6	0.93222111	0.73	16	501.67	0.7654
7	0.98204180	0.33	9	401.72	0.9636
8	0.99593610	0.17	4	332	0.9540
9	0.99929031	0.12	1	167	0.7310

Multivariate Statistics and F Approximations

S=9 M=-0.5 N=78.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.15860296	4.22	81	1036.5	<.0001
Pillai's Trace	1.47779873	3.65	81	1503	<.0001
Hotelling-Lawley Trace	2.36193871	4.59	81	666.97	<.0001
Roy's Greatest Root	0.82644362	15.34	9	167	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

The CANCORR Procedure
Canonical Correlation Analysis
Raw Canonical Coefficients for the Mateorológicas

	Met1	Met2	Met3
PLUVIOME	0.006	-0.022	0.014
Q1	-0.013	0.025	-0.016
Q2	0.001	0.006	0.041
Q3	-0.000	0.021	-0.004
Q4	0.019	-0.019	0.005
vel	0.044	-0.090	0.253
PRESSAO	0.161	-0.022	0.149
TEMP	0.315	-0.011	0.027
UMID_REL	-0.042	-0.087	0.034

Raw Canonical Coefficients for the Quimicas

	Quim1	Quim2	Quim3
COND	-0.114	0.023	0.137
ALCALINI	0.004	-0.007	0.011
F	0.013	0.047	0.015
NO3	0.145	-0.026	-0.076
SO4	-0.056	0.001	-0.062
NA	0.013	0.052	0.012
NH4	0.003	0.011	-0.016
K	0.044	-0.031	-0.032
CA	0.038	0.003	0.000

The CANCORR Procedure
Canonical Correlation Analysis
Standardized Canonical Coefficients for the Mateorológicas

	Met1	Met2	Met3
PLUVIOME	0.169	-0.607	0.381
Q1	-0.322	0.586	-0.389
Q2	0.027	0.149	1.029
Q3	-0.003	0.510	-0.098
Q4	0.453	-0.450	0.122
vel	0.168	-0.341	0.961
PRESSAO	0.636	-0.088	0.587
TEMP	1.127	-0.041	0.097
UMID_REL	-0.296	-0.611	0.240

Standardized Canonical Coefficients for the Químicas

	Quim1	Quim2	Quim3
COND	-0.768	0.158	0.930
ALCALINI	0.114	-0.171	0.281
F	0.073	0.271	0.086
NO3	0.724	-0.129	-0.381
SO4	-0.942	0.020	-1.046
NA	0.207	0.845	0.200
NH4	0.110	0.404	-0.595
K	0.435	-0.307	-0.314
CA	1.031	0.088	0.003

The CANCORR Procedure

Canonical Structure

Correlations Between the Mateorológicas and Their Canonical Variables

	Met1	Met2	Met3
PLUVIOME	-0.176	-0.672	0.317
Q1	-0.358	0.282	-0.060
Q2	-0.280	0.521	0.621
Q3	0.120	-0.168	0.025
Q4	0.370	-0.461	-0.607
vel	0.720	-0.052	0.231
PRESSAO	-0.680	-0.039	0.063
TEMP	0.746	0.290	0.120
UMID_REL	-0.762	-0.355	0.099

Correlations Between the Quimicas and Their Canonical Variables

	Quim1	Quim2	Quim3
COND	-0.096	0.673	-0.314
ALCALINI	0.229	0.307	0.090
F	-0.064	0.532	-0.314
NO3	0.499	0.129	-0.623
SO4	-0.173	0.530	-0.637
NA	0.274	0.804	0.295
NH4	-0.099	0.398	-0.471
K	0.237	0.257	-0.162
CA	0.225	0.512	-0.307

Correlations Between the Mateorológicas and the Canonical Variables of the Químicas

	Quim1	Quim2	Quim3
PLUVIOME	-0.118	-0.444	0.178
Q1	-0.241	0.186	-0.034
Q2	-0.189	0.344	0.349
Q3	0.081	-0.111	0.014
Q4	0.248	-0.304	-0.341
vel	0.484	-0.034	0.129
PRESSAO	-0.457	-0.026	0.036
TEMP	0.502	0.192	0.068
UMID_REL	-0.513	-0.234	0.055

Correlations Between the Químicas and the Canonical Variables of the Mateorológicas

	Met1	Met2	Met3
COND	-0.065	0.444	-0.176
ALCALINI	0.154	0.203	0.051
F	-0.043	0.351	-0.176
NO3	0.335	0.085	-0.350
SO4	-0.116	0.350	-0.358
NA	0.184	0.531	0.166
NH4	-0.066	0.263	-0.265
K	0.159	0.170	-0.091
CA	0.151	0.338	-0.172

4.3 SPSS

Para abrir o banco de dados no SPSS devemos salvar no Excel na versão 4.0 e abrir normalmente no SPSS. Ao abrir o banco, o programa pergunta se queremos ler os nomes das variáveis.

Para calcular a CANCORR pelo SPSS precisamos utilizar uma sintaxe. Que deve ser aberta em **File** / **New** / **Syntax**. Uma vez aberta a sintaxe devemos utilziar o seguinte comando:

```
        INCLUDE 'Canonical correlation.sps'.
CANCORR SET1= var list1 /
        SET2= var list2 / .
```

Este comando irá executar a análise quando clicarmos em **Run** / **All**.

Resultados do SPSS:

Utilizando a programação na sintaxe:

```
INCLUDE 'Canonical correlation.sps'.

CANCORR SET1= PLUVIOME Q1 Q2 Q3 Q4 v35 pressao temp umid_rel /
        SET2= cond alcalini f no3 so4 Na nh4 k Ca / .
```

Canonical Correlations

1	,673
2	,660
3	,562
4	,360
5	,276
6	,225
7	,118
8	,058
9	,027

Test that remaining correlations are zero:

	Wilk's	Chi-SQ	DF	Sig.
1	,159	306,585	81,000	,000
2	,290	206,290	64,000	,000
3	,513	111,027	49,000	,000
4	,750	47,919	36,000	,088
5	,861	24,841	25,000	,471
6	,932	11,686	16,000	,765
7	,982	3,017	9,000	,964
8	,996	,678	4,000	,954
9	,999	,118	1,000	,731

Standardized Canonical Coefficients for Set-1

Columns 1 - 3

	1	2	3
PLUVIOME	-,169	-,610	,381
Q1	,322	,586	-,390
Q2	-,027	,149	1,028
Q3	,003	,510	-,098
Q4	-,453	-,449	,122
V35	-,168	-,342	,962
PRESSAO	-,636	-,088	,587
TEMP	-1,127	-,041	,097
UMID_REL	,296	-,611	,240

Raw Canonical Coefficients for Set-1

Columns 1 - 3

	1	2	3
PLUVIOME	-,006	-,022	,014
Q1	,013	,024	-,016
Q2	-,001	,006	,041
Q3	,000	,021	-,004
Q4	-,019	-,019	,005
VEL	-,044	-,090	,253
PRESSAO	-,161	-,022	,149
TEMP	-,315	-,011	,027
UMID_REL	,042	-,087	,034

Standardized Canonical Coefficients for Set-2
Columns 1 - 3

	1	2	3
COND	,768	,158	,930
ALCALINI	-,114	-,171	,281
F	-,073	,271	,086
NO3	-,724	-,129	-,381
SO4	,942	,020	-1,046
NA	-,207	,847	,200
NH4	-,110	,404	-,595
K	-,434	-,307	-,314
CA	-1,031	,088	,003

Raw Canonical Coefficients for Set-2
Columns 1 - 3

	1	2	3
COND	,114	,023	,137
ALCALINI	-,004	-,007	,011
F	-,013	,047	,015
NO3	-,145	-,026	-,076
SO4	,056	,001	-,062
NA	-,013	,052	,012
NH4	-,003	,011	-,016
K	-,044	-,031	-,032
CA	-,038	,003	,000

Canonical Loadings for Set-1
Columns 1 - 3

	1	2	3
PLUVIOME	,176	-,672	,317
Q1	,358	,282	-,060
Q2	,280	,521	,621
Q3	-,120	-,168	,025
Q4	-,369	-,461	-,607
VEL	-,720	-,052	,230
PRESSAO	,680	-,039	,063
TEMP	-,746	,290	,120
UMID_REL	,762	-,354	,099

Cross Loadings for Set-1
Columns 1 - 3

	1	2	3
PLUVIOME	,118	-,444	,178
Q1	,241	,186	-,034
Q2	,189	,344	,349
Q3	-,081	-,111	,014
Q4	-,248	-,304	-,341
VEL	-,484	-,034	,129
PRESSAO	,457	-,026	,035
TEMP	-,502	,192	,068
UMID_REL	,513	-,234	,055

Canonical Loadings for Set-2
Columns 1 - 3

	1	2	3
COND	,096	,673	-,314
ALCALINI	-,229	,307	,090
F	,064	,532	-,314
NO3	-,499	,129	-,623
SO4	,173	,530	-,637
NA	-,274	,804	,295
NH4	,099	,398	-,471
K	-,237	,257	-,162
CA	-,225	,512	-,307

Cross Loadings for Set-2
Columns 1 - 3

	1	2	3
COND	,065	,444	-,176
ALCALINI	-,154	,203	,051
F	,043	,351	-,176
NO3	-,335	,085	-,350
SO4	,116	,350	-,358
NA	-,184	,531	,166
NH4	,066	,263	-,265
K	-,159	,170	-,091
CA	-,151	,338	-,172

Redundancy Analysis:

Proportion of Variance of Set-1 Explained by Its Own Can. Var.

	Prop Var
CV1-1	,279
CV1-2	,140
CV1-3	,104
CV1-4	,091
CV1-5	,082
CV1-6	,053
CV1-7	,087
CV1-8	,101
CV1-9	,063

Proportion of Variance of Set-1 Explained by Opposite Can.Var.

	Prop Var
CV2-1	,126
CV2-2	,061
CV2-3	,033
CV2-4	,012
CV2-5	,006
CV2-6	,003
CV2-7	,001
CV2-8	,000
CV2-9	,000

Proportion of Variance of Set-2 Explained by Its Own Can. Var.

	Prop Var
CV2-1	,060
CV2-2	,251
CV2-3	,159
CV2-4	,081
CV2-5	,074
CV2-6	,085
CV2-7	,044
CV2-8	,128
CV2-9	,118

Proportion of Variance of Set-2 Explained by Opposite Can. Var.

	Prop Var
CV1-1	,027
CV1-2	,109
CV1-3	,050
CV1-4	,010
CV1-5	,006
CV1-6	,004
CV1-7	,001
CV1-8	,000
CV1-9	,000

------ END MATRIX -----

O SPSS utiliza termos diferentes do STATISTICA. A variância explicada pela variável canônica oposta é a redundância.

4.4 Comparando os *softwares* estatísticos:

O STATISTICA é o *software* com alta qualidade gráfica mas tem um formato complexo por apresentar muitas janelas, criando múltiplos arquivos para os resultados. Além disso a saída é um pouco confusa em relação ao SAS pois não é possível nomear os conjuntos de variáveis. Dessa forma os conjuntos são nomeados conjunto da direita e da esquerda (*right set* e *left set*) fazendo com que tenhamos que voltar várias vezes ao resumo da análise, onde as variáveis estão todas descritas em seus respectivos conjuntos. O SPSS apresenta o mesmo problema chamando de Set-1 e Set-2.

O SAS e o SPSS apresentam mais resultados do que o STATISTICA. Eles calculam automaticamente as cargas cruzadas e os pesos canônicos não padronizados. Além disso, o SAS mostra o resultado de alguns testes de significância para a primeira correlação canônica. O STATISTICA e o SPSS possuem apenas um teste de significância e ambos apresentaram os mesmos resultados.

Quanto à proporção da variância explicada e o índice de redundância, o SPSS e o SAS os chamam de variância explicada pela sua própria variável canônica e pela variável canônica oposta, respectivamente. O SAS traz a proporção do total da variância explicada enquanto que o SPSS e o STATISTICA trazem a quantidade bruta. A forma de apresentação do SAS é mais completa. Se não forem utilizadas todas as variáveis canônicas para interpretação é mais fácil visualizar o quanto da variância das variáveis originais estão sendo explicadas pelas variáveis canônicas escolhidas para interpretação.

Quanto à parte gráfica, o SAS e o STATISTICA possuem a possibilidade de plotar os escores canônicos de cada raiz, já o SPSS não. Talvez versões mais recentes tragam uma linha de comando maior que inclua os gráficos. Mas, em termos de apresentação estética os gráficos do STATISTICA são melhores e mais fáceis de manipular (arrumar cores, títulos, etc). Além disso, ao colar o gráfico em outro programa (Word, Excel, etc.) se clicarmos duas vezes nele, automaticamente se abre a janela do STATISTICA para edição do gráfico.

5 CONCLUSÕES

Através desse estudo notamos que a Correlação Canônica realmente é uma técnica de difícil interpretação, mas que poderia ser mais utilizada dada sua utilidade. Acredita-se que por, causa das melhorias tecnológicas, pesquisadores venham a utilizá-la cada vez mais, o que faz com sejam desenvolvidas mais pesquisas sobre a técnica e conseqüentemente poderiam ser elucidados os problemas de interpretação.

O desenvolvimento teórico da técnica ajudou na compreensão dos termos e dos cálculos envolvidos, bem como na interpretação dos resultados. O exemplo didático também foi importante para a melhor compreensão da técnica.

A aplicação mostrou que muitas vezes, apesar de termos uma correlação canônica relativamente alta, seus índices de redundância podem ser baixos. Outro aspecto importante da aplicação no exemplo real foi a descoberta de novas formas gráficas de apresentação dos resultados. Ficou claro que dependendo das variáveis, o grau de dificuldade da interpretação das mesmas exige o trabalho conjunto de um estatístico e de uma pessoa ligada a área ambiental, no caso.

Quanto à comparação dos *softwares* estatísticos o SAS foi o que apresentou melhor desempenho na facilidade de manipulação e nos resultados obtidos. Já o STATISTICA apresentou melhor qualidade dos gráficos e maior facilidade na alteração do *layout* dos mesmos, isto é, alteração de cores, edição do título, etc.

Em resumo pode-se considerar que devemos aplicar cada vez mais essa técnica, por ser extremamente útil quando bem empregada e com um enorme potencial para aplicações na área ambiental devido às características das variáveis desta.

REFERÊNCIAS BIBLIOGRÁFICAS

Cooley, W. W. and Lohnes, P. R. (1971). Multivariate Data Analysis. New York: John Wiley & Sons, inc.

Davies, J. L. (1986). Statistics and Data Analysis in Geology. 2nd Ed. New York: Springer Verlang.

Friederichs, P. and Hence, A. (2003). Statistical Inference in Canonical Correlation Analysis Exemplified by Influence of North Atlantic SST on European Climate. Journal of Climate. 16 (3). 522-534.

Hair, J. F.; Anderson, R. E.; Tatham, R. L. and Black, W. C. (1998). Multivariate Data Analysis. 5th Ed. New Jersey: Prentice Hall

Johnson, R. A. and Wichern, D. W. (2002). Applied Multivariate Statistical Analysis. 5th Ed. New Jersey: Prentice Hall

Kendall, M. G. and A. Stuart (1967). Advanced theory of Statistics. 2nd Ed., Vol. 2. London: Charles Griffin & Co. Ltd.

Kowalski, J.; Tu, X. M.; Jia, G.; Perlis, A.; Frank, E.; Cristis-Cristoph, P. and Kupfer, D. J. (2003). Generalized Covariance-Adjusted Canonical Correlation Analysis with Application to Psychiatry. Statistical in Medicine. 22 (4). 595-610.

Legendre, p. and Legendre, L. (1998). Numerical Ecology. 2nd Ed. Amsterdam: Elsevier.

Press, S. J. (1982). Applied Multivariate Analysis: using Bayesian and frequentist methods of inference. 2nd Ed. Florida: Robert E. Krieger Publishing Company.

Ter Braak, C. J. S. (1986). Canonical Correspondence Analysis: A new eigenvector technique for multivariate direct gradient analysis. Ecology. 64 (5). 1167-1179.