UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS NATURAIS E EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
E MODELAGEM QUANTITATIVA

ANÁLISE MULTIVARIADA: DA TEORIA À PRÁTICA

Lorena Vicini

MONOGRAFIA DE ESPECIALIZAÇÃO

PPGEMQ

Santa Maria, RS, Brasil
2005

ANÁLISE MULTIVARIADA:
DA TEORIA À PRÁTICA

---

por

Lorena Vicini

Monografia apresentada ao Curso de Especialização do Programa de Pós-Graduação em Estatística e Modelagem Quantitativa, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de
Especialista em Estatística e Modelagem Quantitativa.

Orientador Prof. Adriano Mendonça Souza

Santa Maria, RS, Brasil
2005

Vicini, Lorena, 1974-

V635a

Análise multivariada : da teoria à prática / por Lorena Vicini ; orientador Adriano Mendonça de Souza . – Santa Maria, 2005

140 f. : il., tabs.

Monografia (especialização) – Universidade Federal de Santa Maria, Centro de Ciências Naturais e Exatas, Programa de Pós-Graduação em Estatística e Modelagem Quantitativa, RS, 2005.

1. Estatística 2. Estatística multivariada3. Metodologia 4. Análise de agrupamentos 5. Análise fatorial 6. Análise de componentes pricipais I. Souza, Adriano Mendonça de II. Título

CDU: 311.21

Ficha catalográfica elaborada por
Luiz Marchiotti Fernandes – CRB 10/1160
Biblioteca Setorial do Centro de Ciências Rurais/UFSM

Fone (0xx) 54 9961-8410 ou (0xx) 54 504 6636;
End. Eletr: lorenavicini@pop.com.br

Universidade Federal de Santa Maria
Centro de Ciências Naturais e Exatas
Programa de Pós-Graduação em Estatística
e Modelagem Quantitativa

A Comissão Examinadora, abaixo assinada,
aprova a Monografia de Especialização

ANÁLISE MULTIVARIADA:
DA TEORIA Á PRÁTICA

elaborada por
Lorena Vicini

como requisito parcial para obtenção do grau de
Especialista em Estatística e Modelagem Quantitativa

COMISSÃO EXAMINADORA:

_____
Adriano Mendonça Souza, Dr.
(Presidente/Orientador)

_____
Luiz Felipe Dias Lopes
(Membro)

_____
José Vanderlei Prestes de Oliveira
(Membro)

Santa Maria, 11 de outubro de 2005.

AGRADECIMENTOS

À Universidade Federal de Santa Maria;

Ao professor Adriano Mendonça Souza, pelo tempo dedicado na orientação deste trabalho;

Aos membros da banca examinadora, pela contribuição e sugestões dadas a este trabalho;

Ao meu namorado, pelo carinho e incentivo para que este trabalho se concretizasse;

Ao grupo de pesquisa, coordenado pelo prof. Odorico e os bolsista Paulo e Luis, por cederem o banco de dados aplicado na *ACP* e *AF*.

A minha família, que, mesmo à distância, sempre torceram para que este sonho se transforma-se em realidade;

Aos meus colegas e amigos, que sempre foram parte fundamental desta caminhada, deixo a minha sincera gratidão, pois em muitos momentos de tropeços fizeram com que o caminho se tornasse menos difícil de ser seguido, e que a palavra desistir não fazia parte desta caminhada e acima de tudo a Deus, pela graça de ter vencido mais esta etapa, da minha vida.

RESUMO

Monografia de Especialização
Programa de Pós-Graduação em Estatística e Modelagem Quantitativa
Universidade Federal de Santa Maria, RS, Brasil

ANÁLISE MULTIVARIDA: DA TEORIA Á PRÁTICA

AUTORA: LORENA VICINI
ORIENTADOR: ADRIANO MENDONÇA SOUZA

Data e Local da Defesa: Santa Maria, 03 de outubro de 2005

Grande parte das informações divulgadas pelos meios de comunicação, atuais, provém de pesquisas e estudos estatísticos. Os índices da inflação, de emprego e desemprego, divulgados e analisados pela mídia, são exemplos de aplicação da estatística no nosso dia a dia. Os conceitos estatísticos têm exercido profundas influências na maioria dos campos do conhecimento humano. Métodos estatísticos vêm sendo utilizados no aprimoramento de produtos agrícolas, no desenvolvimento de equipamentos espaciais, no controle do tráfego, na previsão de surtos epidêmicos, bem como em melhorias de processos de gerenciamento, tanto na área governamental como nos negócios, de um modo geral. O crescente uso das técnicas estatísticas, principalmente as que se referem à área da análise multivariada, vem ao encontro da necessidade de aprendizagem destas, que têm se tornado uma ferramenta imprescindível na resolução de problemas do espaço *p-dimensional*, comum em nosso dia a dia. Na prática, essas técnicas podem ser empregadas como ferramenta fundamental em várias áreas da ciência, tais como: medicina, psicologia, ciências rurais, ciências biológicas, dentre outras. Devido a isso, no presente trabalho tem-se como objetivo um estudo teórico-prático sobre as técnicas da análise multivariada, desenvolvendo-se a metodologia e algumas aplicações das mesmas. As técnicas apresentadas foram: Análise de Agrupamentos, Análise Fatorial e Análise de Componentes Principais. Juntamente com as técnicas, apresentou-se os procedimentos operacionais para executá-los, com o auxílio de recursos computacionais, utilizando-se o *software STATISTICA*, detalhando-se os passos de como proceder na execução da análise. Devido a isso, espera-se fornecer um material que sirva de suporte, aos profissionais que atuam nas mais diversas áreas do conhecimento, para que estes possam utilizar este material, nas diversas situações que possam vir a surgir, devido às particularidades de cada área, proporcionando, dessa forma, um melhor entendimento dessas técnicas e uma maior facilidade em sua aplicação.

Palavras-chave: Análise multivariada, Metodologia, Análise de Agrupamentos, Análise Fatorial, Análise de Componentes Principais.

ABSTRACT

Monograph Of Specialization
Program of Masters degree in Statistics and Quantitative Modelling
Federal university of Santa Maria

ANÁLISE MULTIVARIDA: DA TEORIA Á PRÁTICA

AUTHOR: LORENA VICINI
ADVISOR: ADRIANO MENDONÇA SOUZA
Date and place of defense: Santa Maria, 03 de outubro de 2005.

Much of the information broadcast by the means of communication nonbdays comes from statistical researches. Inflation index, employment and unemployment, rates published and analyzed by the media, are examples of daily statistical applications. Statistical concepts have exert ed deep influences in most branches of human being knowledge. Statistical methods have been used to refine agricultural products, to develop special equipment, to traffic control, to forecast epidemic, outbreaks as well as to get better management processes, in the governmental area as much as in private negotiations. The growing employment of statistical techniques, mainly multivariate techniques, have become as necessary as the complexity of problems that have arisen in our daily life. In practice, multivariate analysis can be employed in many branches of science, like medicine, psychology, animal science, biological science and others. Due to this importance, the main purpose of this research was to develop a theoretical practical study about cluster analysis, principal component analysis and factorial analysis. Jointly with these, operational procedures were shown in order to apply these techniques with and without computers. Statistics software was used to explain the procedures step by step. The discussion and procedures presented in this study will be useful to give support to researchers who use these techniques in their area of knowledge helping them with better formulation and solutions for the application.

key words: Multivariate analysis, methodology, cluster analysis, factorial analysis, principal component analysis

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTAS DE REDUÇÕES

Símbolos com letras gregas

Σ = Matriz de variância-covariância populacional

$\hat{\Lambda}$ = Autovalores

Símbolos com letras romanas

F = Matriz fenética

C = Matriz cofenética

S = Matriz de variância-covariância amostral

R = Matriz de correlação

I = Matriz identidade

D = Matriz de distância

KMO = Medida de adequação dos dados

$\vec{x}$ = autovetores da matriz de variância-covariância amostral

$\hat{e}$ = autovetores da matriz de correlação

Siglas

AA – Análise de Agrupamentos

ACP – Análise de componentes principais

AF – Análise fatorial

CP **–** Componentes principais

1 INTRODUÇÃO

A análise multivariada é um vasto campo, no qual até os estatísticos experientes movem-se cuidadosamente, devido esta ser uma área recente da ciência, pois já se descobriu muito sobre esta técnica estatística, mas muito ainda está para se descobrir (MAGNUSSON, 2003).

Na vida, sempre que for necessário tomar uma decisão, deve-se levar em conta um grande número de fatores. Obviamente, nem todos esses pesam da mesma maneira na hora de uma escolha. Às vezes, por se tomar uma decisão usando a intuição, não se identifica, de maneira sistemática, esses fatores, ou essas variáveis, ou seja, não são identificadas quais as variáveis que afetaram a tomada de decisão.

Quando se analisa o mundo que nos cerca, identifica-se que todos os acontecimentos, sejam eles culturais ou naturais, envolvem um grande número de variáveis. As diversas ciências têm a pretensão de conhecer a realidade, e de interpretar os acontecimentos e os fenômenos, baseadas no conhecimento das variáveis intervenientes, consideradas importantes nesses eventos.

Estabelecer relações, encontrar, ou propor, leis explicativas, é papel próprio da ciência. Para isso, é necessário controlar, manipular e medir as variáveis que são consideradas relevantes ao entendimento do fenômeno analisado. Muitas são as dificuldades em traduzir as informações obtidas em conhecimento, principalmente quando se trata da avaliação estatística das informações.

Os métodos estatísticos, para analisar variáveis, estão dispostos em dois grupos: um que trata da estatística, que olha as variáveis de maneira isolada – a estatística univariada, e outro que olha as variáveis de forma conjunta – a estatística multivariada.

Até o advento dos computadores, a única forma de se analisar as variáveis era de forma isolada, e a partir dessa análise fazer inferências sobre a realidade. Sabe-se que essa simplificação tem vantagens e desvantagens. Quando um fenômeno depende de muitas variáveis, geralmente esse tipo de análise falha, pois não basta conhecer informações estatísticas isoladas, mas é necessário, também, conhecer a totalidade dessas informações fornecidas pelo conjunto das variáveis e suas relações. Quando as relações existentes entre as variáveis não são

percebidas, efeitos desconhecidos, entre variáveis, dificultam a interpretação do fenômeno a partir das variáveis consideradas.

O desenvolvimento tecnológico, oriundo das descobertas científicas, tem apoiado o próprio desenvolvimento científico, ampliando, em várias ordens de grandeza, a capacidade de obter informações de acontecimentos e fenômenos que estão sendo analisados. Uma grande massa de informação deve ser processada antes de ser transformada em conhecimento. Portanto, cada vez mais necessita-se de ferramentas estatísticas que apresentem uma visão mais global do fenômeno, do que aquela possível numa abordagem univariada. A denominação “Análise Multivariada” corresponde a um grande número de métodos e técnicas que utilizam, simultaneamente, todas as variáveis na interpretação teórica do conjunto de dados obtidos (NETO, 2004).

Existem vários métodos de análise multivariada, com finalidades bem diversas entre si. Portanto, volta-se ao passo inicial, que é saber que conhecimento se pretende gerar. Ou melhor, que tipo de hipótese se quer gerar a respeito dos dados.

Os pesquisadores devem ter cautela ao trabalhar com as técnicas de análise multivariada, pois a arte do seu uso está na escolha das opções mais apropriadas para detectar os padrões esperados nos seus dados, e as opções mais apropriadas podem não estar no programa de seu computador. Leva-se algum tempo até escolher as opções menos ruins em análises multivariadas, recomenda-se que os leitores exercitem, com cautela, durante o tempo necessário, para apreenderem as limitações dessas análises, antes de tentarem explorar suas grandes potencialidades (MAGNUSSON, 2003).

Os métodos multivariados são escolhidos de acordo com os objetivos da pesquisa, pois sabe-se que a análise multivariada é uma análise exploratória de dados, prestando-se a gerar hipóteses, e não tecer confirmações a respeito dos mesmos, o que seria uma técnica confirmatória, como nos testes de hipótese, nos quais se tem uma afirmação a respeito da amostra em estudo. Embora, às vezes, possa ser utilizada para confirmação dos eventos (HAIR *et al*., 2004). Portanto, a estatística multivariada, com os seus diferentes métodos, difere de uma prateleira de supermercado abarrotada de produtos com a mesma função, pois cada método tem sua fundamentação teórica e sua aplicabilidade. Quando o interesse é verificar como as amostras se relacionam, ou seja, o quanto estas são semelhantes, segundo as

variáveis utilizadas no trabalho, destacam-se dois métodos, que podem ser utilizados: a análise de agrupamento hierárquico e a análise de componentes principais com análise fatorial.

1.1 Tema da pesquisa

O tema da pesquisa refere-se às técnicas estatísticas multivariadas, tais como análise de agrupamentos, análise de componentes principais e análise fatorial, suas aplicações e, principalmente, o seu procedimento de análise.

1.2 Justificativa e importância da pesquisa

A estatística mostra-se, cada vez mais, como uma poderosa ferramenta para a análise e avaliação de dados, em várias áreas do conhecimento, sendo, muitas vezes, um tanto difícil, para os profissionais, trabalharem conceitos e elaborarem exemplos práticos, devido à limitação de materiais didáticos que expressem, com simplicidade e clareza, métodos e procedimentos da aplicação de certas técnicas multivariadas, que só passaram a ser utilizadas, em larga escala, a partir do advento dos computadores.

A presente pesquisa trará significativa contribuição, devido ao fato da carência de material didático na área da estatística multivariada, no que se refere à teoria e à prática descrita passo a passo, em um mesmo material, bem como sua aplicabilidade nas diversas áreas do conhecimento.

1.3 Objetivos

1.3.1 Objetivo Geral

Elaborar um material didático contemplando a teoria e a prática das técnicas de agrupamentos, componentes principais e análise fatorial, voltado às necessidades de atender pesquisadores dos cursos de graduação e pós-graduação, que necessitem dessa ferramenta estatística em suas pesquisas para seu trabalho.

1.3.2 Objetivos Específicos

Para que se cumpra esse objetivo geral, os seguintes objetivos específicos serão realizados:

- Apresentar os cálculos necessários para a aplicação das técnicas e análise de agrupamentos, análise de componentes principais e análise fatorial, manualmente;
- Apresentar os resultados das aplicações dessas técnicas multivariadas, utilizando o *software STATISTICA*;
- Desenvolver exemplos práticos, com banco de dados reais, a fim de mostrar a aplicabilidade das técnicas, bem como as interpretações pertinentes, de forma a elucidar os exemplos.

1.4 Metodologia de pesquisa

Para o desenvolvimento desta pesquisa, serão apresentadas as técnicas de: agrupamentos, componentes principais e análise fatorial.

Será feito uma revisão detalhada, junto a livros especializados, revistas, artigos, e publicações em geral, a fim de que o material elaborado seja de fácil entendimento, possuindo toda a teoria necessária para o bom desenvolvimento do trabalho das pessoas que dele fizerem uso.

Após a apresentação detalhada das teorias, buscar-se-á exemplificar essa metodologia através de análise de dados reais. Inicialmente, para que haja uma melhor compreensão dessa técnica, as análises serão desenvolvidas de forma prática, passo a passo, com as devidas exemplificações. Em um segundo momento, os exemplos são desenvolvidos no programa *Statistica*. A análise e a interpretação dos resultados numéricos e gráficos serão fornecidos após cada etapa do processo.

1.5 Delimitação da pesquisa

O tema abrangerá parte da estatística multivariada, que representa um conjunto de métodos estatísticos, em que as amostras em estudo são caracterizadas

por uma multiplicidade de variáveis. Esses métodos permitem, simultaneamente, analisar alterações em várias propriedades que caracterizem as amostras em análise.

Os métodos abordados são: análise de agrupamentos, análise fatorial e análise de componentes principais.

Não serão abordadas, aqui, outras técnicas da análise multivariada, tais como: análise de correspondência, análise de discriminante, análise de regressão múltipla, dentre outras técnicas multivariadas, utilizando dados amostrais. E, também, não serão utilizados outros *softwares,* para se realizar comparações.

1.6 Organização do trabalho

Este trabalho está organizado em 5 capítulos. No capítulo 1, aborda-se a importância do trabalho, a forma na qual este será desenvolvido, sua justificativa, a metodologia utilizada e a delimitação do estudo proposto.

No capítulo 2, apresenta-se revisão de literatura, que aborda as técnicas da análise multivariada, propostas no estudo.

No capítulo 3, mostra-se as técnicas com exemplos práticos, desenvolvidos passo a passo.

No capítulo 4, apresenta-se a aplicação da análise com dados reais, utilizando o *software statistica,* e no capítulo 5, apresenta-se a conclusão do trabalho desenvolvido.

2 REVISÃO DE LITERATURA

Neste capítulo, será apresentada a revisão de literatura, dividida em itens, servindo estes de suporte para o desenvolvimento deste trabalho. No item 2.1, será discutida a técnica de análise de agrupamentos. No item 2.2, será apresentada a técnica de análise de componentes principais. No item 2.3, apresenta-se a análise fatorial, abordando, nestas técnicas, a teoria e a prática.

2.1 Análise de agrupamentos - *AA*

> *Todos nós acreditamos que qualquer população é composta de segmentos distintos. Se trabalhamos com as variáveis adequadas, a análise de conglomerados nos ajudará a ver se existem grupos que são mais semelhantes entre si do que com membros de outros grupos (Tom Myers, consultor Burke Customer, Satisfaction Associates).*

A *AA,* em sua aplicação, engloba uma variedade de técnicas e algoritmos, sendo que o objetivo é encontrar e separar objetos em grupos similares. Essa técnica pode ser observada, por exemplo, se se tiver vários produtos em uma determinada prateleira de um supermercado, e distribuir esses produtos, na prateleira, segundo suas características, de um mesmo composto, ou o mesmo princípio ativo, por exemplo. Aí está-se a praticar *AA*. Agora, se esses produtos estiverem espalhados por toda a prateleira, significa que se terá mais de uma característica, e, para que se possa uní-los por características comuns, será muito trabalhoso, exigindo conceitos mais sofisticados de semelhança, e procedimentos mais científicos para juntá-los. É em relação a esse procedimento multidimensional que se trabalhará.

Em alguns estudos, torna-se necessário conhecer algumas características de determinado grupo de um conjunto de elementos amostrais, principalmente quando é resultante de uma ou mais variáveis. Quando se obtém mensuração de diferente natureza, pode-se observar se há similaridades no conjunto de dados. Um dos métodos a *AA*, que poderá ser utilizado para tais objetivos.

A análise de agrupamentos estuda todo um conjunto de relações interdependentes. Ela não faz distinção entre variáveis dependentes e independentes, isto é, variáveis do tipo causa e efeito, como na regressão.

Conforme Everitt (1974 apud BUSSAB, 1990), a *AA* pretende resolver o seguinte problema: “dada uma amostra de *n* objetos (ou indivíduos), cada um deles medindo segundo *p* variáveis, procurar um esquema de classificação que agrupe os objetos em *g* grupos. Deve ser determinado, também, o número de variáveis desses grupos”. Portanto, a finalidade dessa técnica é reunir os objetos (indivíduos, elementos) verificados nos grupos em que exista homogeneidade dentro do grupo e heterogeneidade entre os grupos, objetivando propor classificações. Os objetos em um grupo são relativamente semelhantes, em termos dessas variáveis, e diferentes de objetos de outros grupos. Quando utilizada dessa forma, a *AA* é o inverso da análise de fatores, pelo fato de reduzir o número de objetos, e não o número de variáveis, concentrando-os em um número muito menor de grupos.

A *AA* constitui uma metodologia numérica multivariada, com o objetivo de propor uma estrutura classificatória, ou de reconhecimento da existência de grupos, objetivando, mais especificamente, dividir o conjunto de observações em um número de grupos homogêneos, segundo algum critério de homogeneidade (REGAZZI, 2001). Muitas vezes, nessa técnica, são feitas afirmativas empíricas, que nem sempre têm respaldo teórico. Muitas técnicas são propostas, mas não há, ainda, uma teoria generalizada e amplamente aceita. Devido a isso, deve-se utilizar vários métodos e comparar os resultados, para que a análise dos dados seja realizada pela técnica mais adequada.

A *AA* é um método simples, calcada nos cálculos de distância, no entanto, não requerem conhecimento estatístico para a sua aplicação, como é o caso quando se aplica análise de variância, de regressão, ou fatorial. O primeiro caso, *AA* não requer o uso de um modelo, os demais casos necessitam. Para a aplicação da *AA,* as estatísticas e os conceitos, a seguir, serão utilizados:

> ***Esquema de aglomeração:*** Informa sobre objetos, ou casos a serem combinados em cada estágio de um processo hierárquico de aglomeração.
> ***Centróide do agrupamento:*** Representam os valores médios das variáveis para todos os casos, ou objetos em um agrupamento particular.
> ***Centros de agrupamentos:*** São os pontos iniciais em um agrupamento não-hierárquico. Os agrupamentos são construídos em torno desses centros.
> ***Composição de um Agrupamento:*** Indica o agrupamento ao qual pertence cada objeto, ou caso (MALHOTRA, 2001, p.528).

Dendograma ou Fenograma: Também chamado de gráfico em árvore. Este, representa uma síntese gráfica do trabalho desenvolvido, sintetizando a informação, ocasionando uma pequena perda da mesma, pelo fato de ser uma síntese. Embora aconteça essa perda de informação, esse gráfico é de grande utilidade para a classificação, comparação e discussão de agrupamentos.

Há duas formas de se representar um dendograma: horizontal e verticalmente.

No dendograma horizontal, as linhas verticais, ou o eixo y, representam os grupos unidos por ordem decrescente de semelhança, e a posição da reta, na escala ou o eixo x, indica as distâncias entre os grupos que foram formados. O dendograma é lido de cima para baixo, quando for feito na forma horizontal.

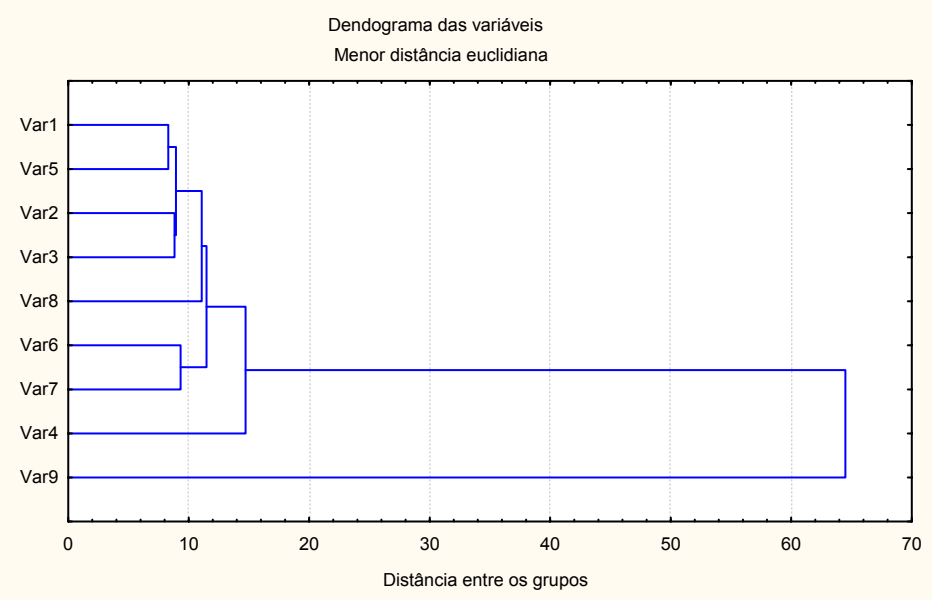


Figura 01 – Dendograma horizontal.

Verifica-se, na Figura 01, que as variáveis Var 1 e Var 5 são as que possuem a maior semelhança, no dendograma, por possuírem a menor distância euclidiana, sendo essas a formarem o primeiro grupo. Logo, em seguida, vêm as variáveis Var 2, Var 3, Var 8, e, assim, sucessivamente, as variáveis serão agrupadas, por ordem decrescente de semelhança, ou seja, a Var 9 formou o último grupo do dendograma, o qual manteve-se distinto dos demais grupos formados, pelo fato de essa variável possuir pouca semelhança em relação às outras.

Como hoje, ainda, não existe uma teoria que diga em qual altura deve-se fazer um corte no gráfico, é o pesquisador quem decide. Fazendo um corte entre as alturas 20 e 30, obter-se-á dois grupos homogêneos distintos, o primeiro e maior, que é formado pelas variáveis Var 1, Var 5, Var 2, Var 3, Var 8, Var 6, Var 7e Var 4, já o segundo grupo é formado apenas pela Var 9.

No dendograma vertical, a leitura é feita da direita para esquerda, no qual as linhas verticais, ou o eixo y, indicam as distâncias entre os grupos foram formados, e a posição da reta na escala, ou o eixo x, representa os grupos unidos por ordem decrescente de semelhança, conforme Figura 02.

A interpretação desta Figura 02 é análoga à Figura 01, apenas muda no eixo em que as variáveis estão representadas.

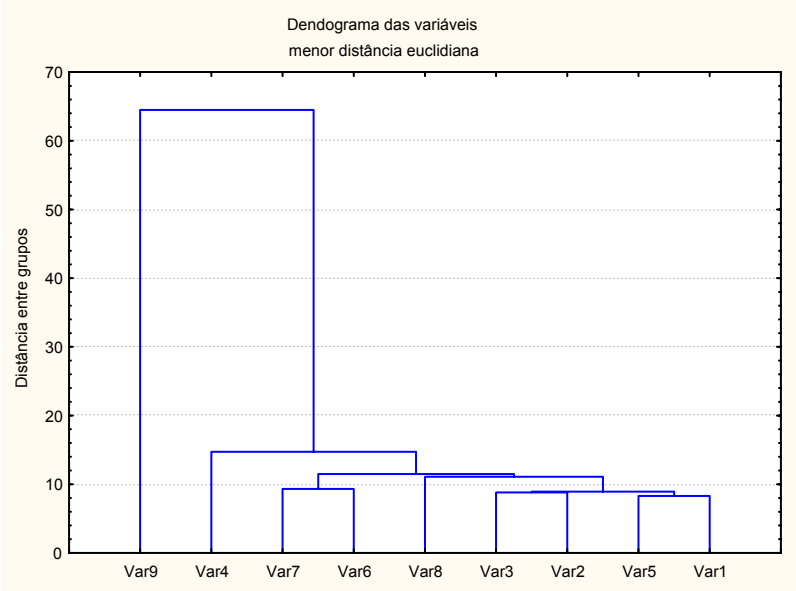


Figura 02 - Dendograma vertical.

Distância entre centros de conglomerados. Indica a distância que separa os pares individuais de conglomerados. Sendo que os conglomerados que se apresentam bem separados são distintos. São esses os desejáveis para a análise.
Matriz de coeficientes de semelhança ou distância. É o triângulo inferior ,ou superior, de uma matriz que contém distâncias emparelhadas entre objetos ou casos (MALHOTRA, 2001, p.528).

O primeiro passo, para realizar a *AA,* consiste em formular o problema de aglomeração, definindo as variáveis sobre as quais se baseará o agrupamento. Logo após, faz-se a coleta dos dados, que serão reunidos numa tabela com *m* colunas (variáveis) e *n* linhas (objetos). Antes de escolher a medida de distância para a

análise dos dados, é necessário verificar se os mesmos encontram-se com a mesma unidade de medida. Caso contrário, deve-se fazer a padronização dos mesmos. Escolhe-se, então, uma medida apropriada de distância, que irá determinar o quão semelhantes, ou diferentes, são os objetos que estão sendo agrupados. Dentre vários processos de aglomeração, o pesquisador deve escolher aquele que é mais apropriado ao problema estudado.

Um método é melhor do que um outro quando o dendograma fornece uma imagem menos distorcida da realidade. É possível avaliar o grau de deformação provocado pela construção do dendograma calculando-se o coeficiente de correlação cofenético (VALENTIN, 2000). Ou seja, o menor grau de distorção, será refletido pelo maior coeficiente cofenético, fornecido pela matriz fenética *F*, na qual seus valores foram obtidos junto à matriz de distâncias inicial e pela matriz cofenética *C*, sendo estes os valores obtidos junto à matriz final das distâncias. O maior coeficiente cofenético possui a capacidade de evidenciar melhor a estrutura dos dados, isto é, a existência de grupos.

A decisão sobre o número total de conglomerados, a constarem na análise, caberá ao pesquisador, pois esta dependerá de cada pesquisa.

A estrutura básica da aplicação da *AA* pode ser representada em etapas, conforme mostra a Figura 03:

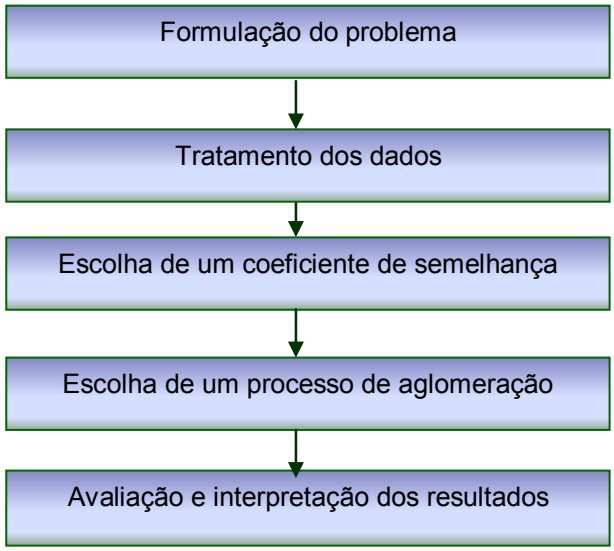


Figura 03 - Etapas para a realização da análise de agrupamentos.

Deve-se observar que essas etapas não são independentes. Algumas vezes, será necessário voltar a etapas anteriores para corrigir e aprimorar algumas etapas posteriores. Considera-se que as etapas descritas na Figura 03 formam um procedimento metodológico muito útil para a realização da *AA*.

Acredita-se que a formulação do problema seja a parte mais importante da análise de agrupamentos, ou seja, a escolha das variáveis nas quais se baseará o processo de aglomeração. A inclusão de uma, ou duas variáveis, sem importância, poderá vir a distorcer o resultado final da análise. O conjunto de variáveis escolhido deve descrever a semelhança entre objetos, em termos relevantes para o problema em pesquisa. Esta fase é importante para a *AA*, pois é onde se fixa o critério de homogeneidade. Segundo Bussab *et al*. (1190, p. 2), “critérios distintos levam a grupos homogêneos distintos, e o tipo de homogeneidade depende dos objetivos a serem alcançados”.

Ao analisar os dados, em primeiro lugar deve-se verificar se eles devem ser tratados. Por exemplo, deve-se observar se as variáveis foram medidas em unidades muito diferentes entre si. A solução por aglomerado será influenciada pelas unidades de medida. Nesse caso, deve-se, antes de aglomerar as amostras, padronizar os dados. Embora a padronização possa remover a influência da unidade de medida, poderá também reduzir as diferenças entre grupos em variáveis que melhor descrevam os conglomerados, pois as unidades associadas às variáveis podem, arbitrariamente, afetar o grau de similaridade entre os objetos, e a padronização dos dados faz com que esse efeito da arbitrariedade seja eliminado, fazendo com que as variáveis possuam a mesma contribuição no cálculo do coeficiente de similaridade entre os objetos.

Para que seja possível padronizar as variáveis, é necessário ter-se uma matriz de dados com *p* variáveis (j = 1, 2, ..., p) e *n* objetos (i = 1, 2, ..., n). Sendo que, na matriz de dados, o valor do i-ésimo objeto e j-ésima variável será denotado por $\mathrm{X_{ij}}$, no qual o valor padronizado será representado por $\mathrm{Z_{ij}}$. Onde as variáveis padronizadas terão média 0 e variância constante 1, sendo esta a mais utilizada na prática, e é representada pela seguinte função:

$$Z_{ij} = \frac{X_{ij} - \overline{X}_j}{S_j} \qquad (2.1)$$

sendo cada i fixo, no qual i = 1, 2, ..., n e j = 1, 2, ..., p

Para aplicar a *AA,* em um conjunto de dados, é muito importante a escolha de um coeficiente que quantifique o quão parecidos dois objetos são. Esse coeficiente pode ser dividido em duas categorias, que dizem respeito à estimação de uma medida de similaridade, ou dissimilaridade, entre os indivíduos, ou populações, a serem agrupados. Na medida de similaridade, quanto maior for o valor observado, mais parecido serão os objetos. Já na medida de dissimilaridade, quanto maior for o valor observado, menos parecido serão os objetos. Um exemplo de medida de similaridade é o coeficiente de correlação, pois quanto maior seu valor, maior a associação e de dissimilaridade a distância euclidiana, pois quanto menor o valor mais próximo os objetos estão uns dos outros.

Para que seja possível a escolha do melhor coeficiente de semelhança, é necessário ter-se uma matriz $X_{(n\ x\ p)} = X_{ij}$. Assim, cada vetor linha representa uma unidade amostral (indivíduos, tratamentos, espécies), e cada vetor coluna, uma variável (REGAZZI, 2001), como apresenta-se na Tabela 01.

Tabela 01 – Matriz de dados *n* indivíduos e *p* variáveis.

Indivíduos	Variáveis							
	X_1	X_2	X_3	X_4	...	X_j	...	X_p
1	X_{11}	X_{12}	X_{13}	X_{14}	...	X_{1j}	...	X_{1p}
2	X_{21}	X_{22}	X_{23}	X_{24}	...	X_{2j}	...	X_{2p}
3	X_{31}	X_{32}	X_{33}	X_{34}	...	X_{3j}	...	X_{3p}
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	...	X_{ij}	.	X_{ip}
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
n	X_{n1}	X_{n2}	X_{n3}	X_{n4}	...	X_{nj}	...	X_{np}

Fonte: Regazzi (2001)

O primeiro estágio, em muitos métodos da análise de agrupamentos, é a conversão da matriz $n_x p$ de dados em uma matriz quadrada, onde *n* é o número de

indivíduos, de similaridade ou dissimilaridade, que são medidas da relação entre pares de indivíduos, ou populações. Dado o valor de um conjunto de *p* variáveis, em cada intersecção da i-ésima fila, e da k-ésima coluna dessa matriz, coloca-se a medida de similaridade, ou dissimilaridade, entre o i-ésimo e k-ésimo indivíduo. A alta similaridade indica que dois indivíduos são comuns em relação ao conjunto de variáveis, enquanto que a alta dissimilaridade indica o contrário (MAXWEL, 1977 apud REGAZZI, 2001).

Algumas medidas de similaridade e dissimilaridade, que são utilizadas em análise de agrupamento, são citadas aqui. Ressalta-se que as expressões matemáticas, usadas na determinação dos coeficientes de distância, serão dadas em função das variáveis originais. Se forem usadas as variáveis transformadas, utilizam-se as mesmas fórmulas, trocando X_{ij} por Z_{ij}.

Como o objetivo da análise de agrupamento é reunir objetos semelhantes, torna-se necessário alguma medida para avaliar o quão semelhantes, ou diferentes são os objetos. Geralmente, costuma-se avaliar a semelhança em termos de distância entre pares de objetos. Os objetos que possuem a menor distância entre si são mais semelhantes, um do outro, do que os objetos com a maior distância. Essa medida de semelhança é fornecida pela distância euclidiana.

Um grande problema da *AA* é a escolha da medida de proximidade mais adequada, sendo que as técnicas são baseadas em diferentes medidas de proximidade, e nem sempre chegam ao mesmo resultado. Devido a isso, é importante testar mais de uma medida de distância, para que possa ser utilizada a mais adequada para a análise.

Segundo Regazzi (2001), “embora a distância euclidiana seja uma medida de dissimilaridade, às vezes ela é referida como uma medida de semelhança, pois quanto maior seu valor, menos parecidos são os indivíduos ou unidades amostrais”.

A distância entre dois pontos do plano pode ser definida como uma função *d,* que, a cada par de pontos P_1 e P_2, associa um número real positivo, $d(P_1, P_2)$, com as seguintes propriedades:

i) se $0 \leq d(P_1, P_2)$ *e* $d(P_2, P_1) = 0$, se e somente se, $P_1 = P_2$

ii) $d(P_1, P_2) = d(P_2, P_1)$ (Simetria)

iii) $d(P_1, P_2) \leq d(P_1, P_3) + d(P_3, P_2)$, onde P_3 é um ponto qualquer do plano (Desigualdade Triangular).

Essas condições somente traduzem, em linguagem matemática, as propriedades que, intuitivamente, espera-se de uma função que sirva para medir distâncias, isto é, a distância entre dois pontos deve ser sempre positiva, e só se deve anular quando os pontos coincidirem.

A distância medida de um ponto P_1 até um ponto P_2 deve ser a mesma, quer essa medida seja feita de P_1 a P_2, ou de P_2 a P_1.

A terceira propriedade diz simplesmente que, dados três pontos no plano, a medida de qualquer dos lados do triângulo, determinado por estes pontos, é menor que a soma da medida dos outros dois. Por isso, a desigualdade, que traduz essa condição, é chamada *desigualdade triangular*.

A expressão dissimilaridade surgiu em função de que, à medida que $d(P_1, P_2)$ cresce, diz-se que a divergência entre P_1 e P_2 aumenta, ou seja, torna-se cada vez mais dissimilar.

Conforme Malhotra (2001, p. 529), “a utilização de diferentes medidas de distância pode levar a resultados diferentes de aglomeração. Assim, é conveniente utilizar medidas diferentes e comparar os resultados”.

As medidas de distância consideram que, se dois indivíduos são similares, eles estão próximos um do outro, ou seja, eles são comuns ao conjunto de variáveis e vice-versa.

O coeficiente de associação pode ser chamado de cálculo da matriz, denominada de matriz de similaridade, ou dissimilaridade, podendo esta ser denominada de matriz de proximidade entre os elementos observados (similaridade, distância, dependência). Exemplificando, pode-se considerar a distância euclidiana como uma medida de dissimilaridade, e o coeficiente de correlação como uma medida de similaridade.

A seguir, estão apresentados alguns coeficientes de similaridade, usados para estabelecer o conceito de distância entre os objetos.

2.1.1 Alguns coeficientes de medida de distância

- **Distância Euclidiana**

A distância euclidiana é, sem dúvida, a medida de distância mais utilizada para a análise de agrupamentos.

Considerando o caso mais simples, no qual existem *n* indivíduos, onde cada um dos quais possuem valores para *p* variáveis, a distância euclidiana entre eles é obtida mediante o teorema de Pitágoras, para um espaço multidimensional.

Segundo Manly (1986), "a distância euclidiana, quando for estimada a partir das variáveis originais, apresenta a inconveniência de ser influenciada pela escala, de medida pelo número de variáveis e pela correlação existente entre as mesmas". Para contornar as escalas, faz-se a padronização das variáveis em estudo, para que possuam a variância igual à unidade.

Considerando dois indivíduos i e i', a distância entre eles é dada por

$$d_{ii'} = \left[\sum_{j=1}^{p} (X_{ij} - X_{i'j})^2 \right]^{\frac{1}{2}} \qquad (2.2)$$

- **Distância euclidiana média**

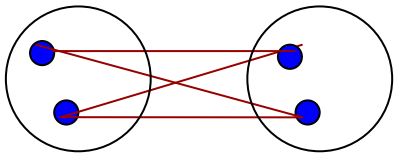


Figura 4 - Distância média.

A distância entre dois agrupamentos é obtida pela média das distâncias. Aqui, é possível encontrar o valor da distância através da média aritmética. Através dessa, a distância entre cada conglomerado tem o mesmo peso. A distância euclidiana média é dada por:

$$d = \sum_{j=1}^{p} \left\{ \frac{(X_{ij} - X_{i'j'})^2}{X_{ij}} \right\} \qquad (2.3)$$

- **Distância de Mahalanobis – D^2**

A similaridade entre as unidades amostrais (tratamentos, indivíduos, populações), com relação a um conjunto de características correlacionadas, e a distância entre quaisquer pares de unidades amostrais, deve considerar o grau de dependência entre as variáveis. A medida mais utilizada, para a quantificação das distâncias entre duas populações, quando existe repetição de dados, é à distância de Mahalanobis (D^2).

Conforme Cruz (1990), "a distância de Mahalanobis, considera a variabilidade de cada unidade amostral, sendo recomendada para dados provenientes de delineamento experimentais, e, principalmente, quando as variáveis são correlacionadas". Quando as correlações entre as variáveis forem nulas, considera-se as variáveis padronizadas, e a distância de Mahalanobis D^2 é equivalente à distância euclidiana.

A forma mais simples de explicar como obter tal medida é a forma matricial, sendo que essa medida entre duas unidades amostrais (tratamentos, indivíduos, populações), i e i', é fornecida pela notação:

$$D^2_{ii'} = \left(\vec{\overline{X}}_i - \vec{\overline{X}}_{i'}\right)' S^{-1} \left(\vec{\overline{X}}_i - \vec{\overline{X}}_{i'}\right) \qquad (2.4)$$

em que :

$\vec{\overline{X}}_i = \left[\overline{X}_{i1},\ \overline{X}_{i2}, \ldots..,\ \overline{X}_{ip}\right]'$

$\vec{\overline{X}}_{i'} = \left[\overline{X}_{i'1},\ \overline{X}_{i'2}, \ldots..,\ \overline{X}_{i'p}\right]'$

$\vec{\overline{X}}_i \ e \ \vec{\overline{X}}_{i'}$, são os vetores p-dimensionais de médias $i \ e \ i'$, respectivamente, com $i \neq i'$ e i, i' = 1, 2, ..., n.

onde *S* é a matriz de dispersão amostral comum a todas as unidades que, no caso de delineamentos experimentais, trata-se da matriz de variâncias e covariâncias residuais.

Embora $D^2_{ii'}$ seja o quadrado da distância de Mahalanobis, será chamado de distância de Mahalanobis.

Admitindo-se distribuição multinormal p-dimensional, e homogeneidade na matriz de variância-covariância nas unidades amostrais, pode-se chamar distância generalizada de Mahalanobis.

- **Coeficiente de Pearson**

Outra forma de estabelecer o conceito de distância, entre os objetos, é através do Coeficiente de Correlação de Pearson.

A medida de similaridade entre dois objetos *R* e *T*, denotada por *S(R,T)*, deve satisfazer as seguintes propriedades:

i) S(R,T) = S(T,R);

ii) $|S(R,T)| \geq 0$;

iii) *S(R,T)* cresce à medida em que a semelhança entre *R* e *T* cresce.

O coeficiente de Pearson, entre os objetos *R* e *T*, é dado pela seguinte equação:

$$r_{ii'} = \frac{\sum_j X_{ij}X_{i'j} - \frac{1}{p}(\sum_j X_{ij})(\sum_j X_{i'j})}{\sqrt{\left[\sum_j X^2_{ij} - \frac{1}{p}\left(\sum_j X_{ij}\right)^2\right]\left[\sum_j X^2_{i'j} - \frac{1}{p}\left(\sum_j X_{i'j}\right)^2\right]}}. \qquad (2.5)$$

Deve-se atentar para o fato de que o valor de $r_{ii'}$ varia de –1 a +1.

Escolhida uma medida de distância, ou de semelhança, passa-se a escolher um processo de agrupamento, ou aglomeração.

A escolha do método de agrupamento é tão difícil quanto a escolha do coeficiente de associação. Dessa escolha dependerá a correta classificação de uma amostra estar dentro de um grupo, ou de outro, que já tenha sido formado. Os métodos de agrupamento foram desenvolvidos com base nos modelos e dados diversos.

Há grande quantidade de métodos de agrupamento. As diferenças entre os métodos existem em função de diferentes formas de definir proximidade entre um indivíduo em um grupo, contendo vários indivíduos, ou entre grupos de indivíduos.

Na *AA*, não se pode dizer que existe um método que seja melhor para se aplicar. O pesquisador deve decidir qual será o mais adequado para o desenvolvimento do seu trabalho, pois cada método leva a um resultado. Os métodos de agrupamento mais utilizados são os hierárquicos.

Como se pode observar na Figura 05, os processos de agrupamento podem ser divididos em dois grupos: hierárquicos ou não-hierárquicos. Conforme Malhotra (2001, p. 529), a **aglomeração hierárquica** caracteriza-se pelo estabelecimento de uma hierarquia, ou estrutura em forma de árvore, sendo esta a mais utilizada. Os métodos hierárquicos são divididos em *aglomerativos e divisivos.*

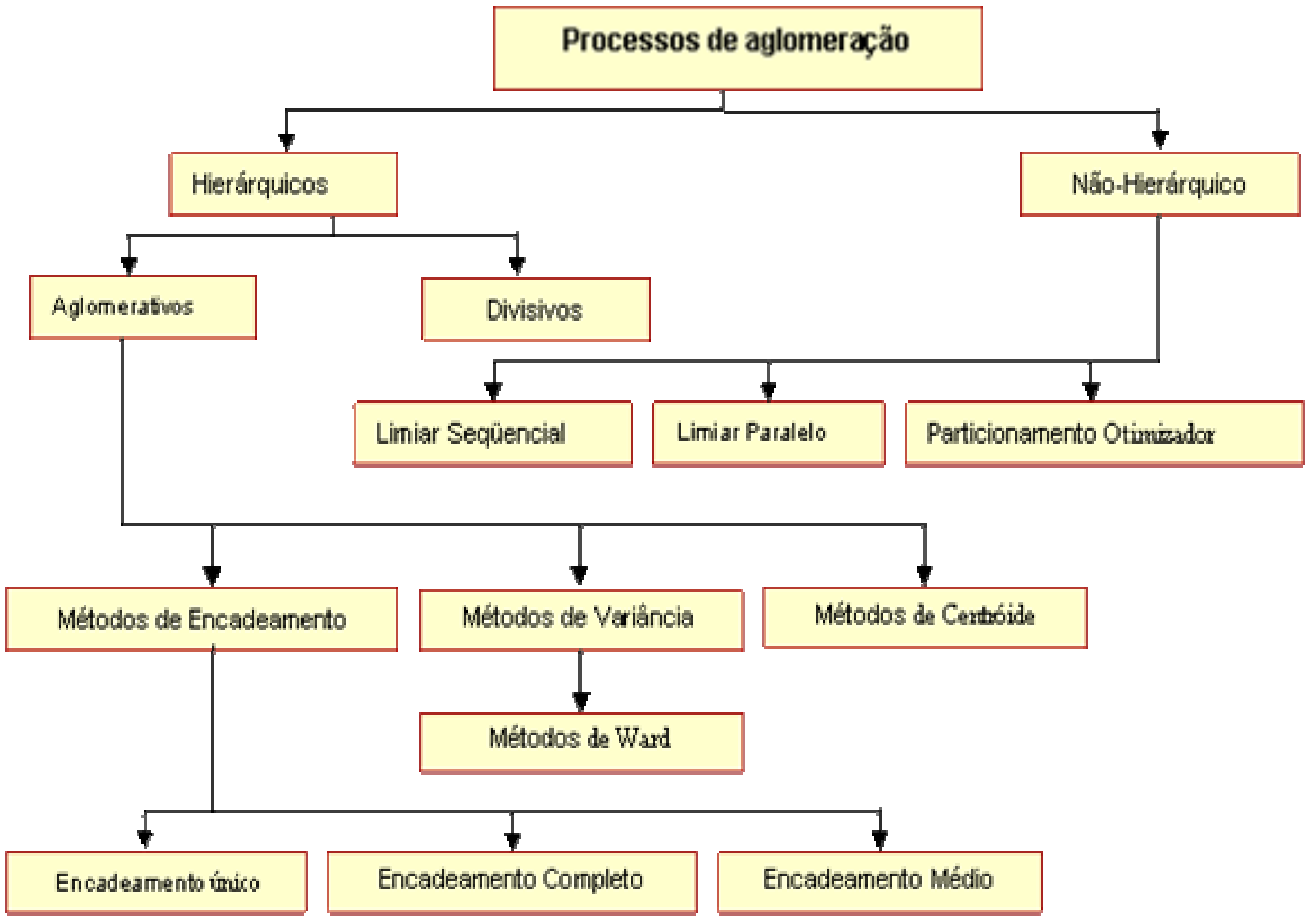


Figura 05 - Classificação dos processos de aglomeração.
Fonte: Malhotra (2001, p. 531).

O *agrupamento aglomerativo* tem início em um grupo separado. Formam-se os grupos reunindo-se os objetos em grupos cada vez maiores. O processo continua até que todos os objetos sejam membros de um único grupo, sendo esse método seqüencial, onde os objetos são reunidos um após o outro, respeitando uma determinada seqüência de aglomeração. O critério básico da fusão entre um objeto e um grupo, ou entre dois grupos, é sempre o mesmo: serão reunidos os grupos que têm maior similaridade entre si. O problema é: como calcular esta similaridade? O método de cálculo depende do método de aglomeração escolhido (VALENTIN, 2000).

No *agrupamento divisivo,* todos os objetos partem de um grupo gigante, e estes são subdivididos em dois subgrupos, de tal forma que exista o máximo de semelhança entre os objetos dos mesmos subgrupos e a máxima dissimilaridade entre elementos de subgrupos distintos. Esses subgrupos são, posteriormente, subdivididos em outros subgrupos dissimilares. O processo é repetido até que haja tantos subgrupos quantos objetos (MALHOTRA, 2001).

O procedimento básico, de todos os métodos aglomerativos de agrupamento, é similar. Inicia-se com o cálculo de uma matriz de distâncias entre as variáveis e finaliza-se com um dendograma, no qual é possível verificar as fusões sucessivas dos indivíduos, até os indivíduos formarem um único grupo (REGAZZI, 2001).

Os métodos aglomerativos são de uso comum. Estes são constituídos de métodos de encadeamento, métodos de erros de somas de quadrados, ou métodos de variância e métodos centróides.

Os **métodos de encadeamento** compreendem:

O *método do encadeamento único (Single Linkage)*, que se baseia na distância mínima, regra do vizinho mais próximo. Os dois primeiros objetos agrupados são os que apresentam menor distância entre si. Identifica-se a menor distância agrupando-se o terceiro objeto com os dois primeiros, ou formando um novo grupo de dois objetos. Em cada estágio a distância entre dois grupos é definida como a distância entre seus dois pontos mais próximos. Dois grupos podem incorporar-se em cada estágio por meio do encadeamento mais curto entre eles. Continua-se o processo até que todos os objetos, estejam em um único grupo.

O *método do encadeamento completo (Complete Linkage)* é semelhante ao encadeamento único, embora se baseie na distância máxima entre os objetos ou o método do vizinho mais afastado. Neste, a distância entre dois grupos é calculada entre seus dois pontos mais afastados.

> O ***método do encadeamento médio*** é semelhante aos métodos anteriores, embora a distância entre dois grupos se defina como a média da distância entre todos os pares de objetos, onde cada membro de um par provém de cada um dos grupos. No método de encadeamento médio são utilizadas informações sobre todos os pares de distâncias, e não apenas da distância mínima ou máxima. Devido a este fato, é perfeito em relação aos métodos de encadeamento único e completo.
> Os ***métodos de variância*** buscam gerar grupos que possam minimizar a variância dentro destes grupos. Dentre estes métodos, está o de *Ward,* que minimiza o quadrado da distância euclidiana às médias dos grupos. Um grupo será reunido a um outro se essa reunião proporcionar o menor aumento da variância intragrupo. Este método de variância calcula as médias de todas as variáveis para cada grupo, escolhendo a que proporciona a menor variância. Calcula-se então, para cada objeto, o quadrado da distância euclidiana, as médias do agrupamento, conforme Figura 04. Somam-se essas distâncias para todos os objetos. Em cada estágio, combinam-se os dois grupos que apresentar menor aumento na soma global de quadrados dentro dos agrupamentos. Este método é altamente eficiente na formação de grupos.
> Outro método de variância utilizado é o do ***Centróide,*** que considera que a distância entre dois aglomerados é a distância entre seus centróides, que nada mais é que a média para todas as variáveis. A cada agrupamento novo de objetos, deve-se calcular um novo centróide. Dentre os métodos hierárquicos, os que têm se revelado superior em relação aos outros são o do encadeamento médio e o de *Ward*.
> A segunda forma de processo de aglomeração está nos **métodos não-hierárquicos**, que se caracterizam por procurar maximizar a homogeneidade intragrupo, sem considerar a hierarquia entre grupos. Estes métodos costumam ser chamados de k médias ou *k-means clustering*. *k-means clustering* compreendem o ***limiar seqüencial, o limiar paralelo e o particionamento otmizador.***
> O método ***limiar seqüencial*** consiste em escolher um centro de aglomeração, e todos os objetos a menos de um valor pré-determinado a contar do centro são agrupados juntamente. A partir daí, escolhe-se então um novo centro de aglomeração, ou repete-se o processo para os pontos não aglomerados.
> O método ***limiar paralelo*** escolhe de uma só vez vários centros de aglomeração e os objetos dentro do limiar são agrupados com o centro mais próximo. Todos os objetos que estão a menos de um valor pré-determinado do centro são agrupados juntamente.
> O método do ***particionamento otmizador*** *difere* dos anteriores, pois permite a redistribuição posterior de objetos no agrupamento de modo a otimizar um critério global, tal como a distância média dentro do grupo para um dado número de agrupamentos.
> A escolha de um método de aglomeração e a escolha de uma medida de distância estão inter-relacionadas. Por exemplo, deve-se usar os quadrados das distâncias euclidiana com os métodos de Ward e dos centróides (MALHOTRA, 2001, p.530 e 531).

Neste trabalho, são abordados apenas dois métodos, ou algoritmos de agrupamento, que são:

- Método do encadeamento único (*Single Linkage*), ou, ainda, método do vizinho mais próximo.
- Método do encadeamento completo (*Complete Linkage*), ou, ainda, método do vizinho mais distante.

2.2 Análise de Componentes Principais - *ACP*

A análise de componentes principais tem por objetivo descrever os dados contidos num quadro indivíduos-variáveis numéricas: *p* variáveis serão mediadas com *n* indivíduos. Esta é considerada um método fatorial, pois a redução do número de variáveis não se faz por uma simples seleção de algumas variáveis, mas pela construção de novas variáveis sintéticas, obtidas pela combinação linear das variáveis inicias, por meio dos fatores (BOUROCHE, 1982).

A *ACP* é uma técnica matemática da análise multivariada, que possibilita investigações com um grande número de dados disponíveis. Possibilita, também, a identificação das medidas responsáveis pelas maiores variações entre os resultados, sem perdas significativas de informações. Além disso, transforma um conjunto original de variáveis em outro conjunto: os componentes principais (*CP*) de dimensões equivalentes. Essa transformação, em outro conjunto de variáveis, ocorre com a menor perda de informação possível, sendo que esta também busca eliminar algumas variáveis originais que possua pouca informação. Essa redução de variáveis só será possível se as *p* variáveis iniciais não forem independentes e possuírem coeficientes de correlação não-nulos.

A meta da análise de componentes principais é abordar aspectos como a geração, a seleção e a interpretação das componentes investigadas. Ainda pretende-se determinar as variáveis de maior influência na formação de cada componente, que serão utilizadas para estudos futuros, tais como de controle de qualidade, estudos ambientais, estudos populacionais entre outros.

> Conforme Souza (2000, p.23 e 24), a idéia matemática do método é conhecida há muito tempo, apesar do cálculo das matrizes dos autovalores e autovetores não ter sido possível até o advento da evolução dos computadores. O seu desenvolvimento foi conduzido, em parte, pela

> necessidade de se analisar conjuntos de dados com muitas variáveis correlacionadas.
> Inicialmente, o objetivo da *ACP* foi o de encontrar linhas e planos que melhor se ajustassem a um conjunto de pontos em um espaço *p-dimensional* (PEARSON, 1901). Posteriormente, um trabalho sobre o desempenho de estudantes foi avaliado por meio de uma seqüência de testes escolares, onde as variáveis utilizadas na sua maioria eram correlacionadas. Então, a matriz de correlação e a matriz de covariância foram utilizadas para que fosse feita uma análise simultânea. Na época, quando um estudante apresentava boas notas nos testes aplicados, pensava-se que era porque ele possuía algum componente psicológico mais desenvolvido do que os outros, facilitando assim algumas tarefas. Na Psicologia moderna, as variáveis que apresentavam uma maior influência foram chamadas de *fatores mentais*. Na Matemática, foram denominadas de *fatores* e, depois, elas receberam o nome de *componentes* para não serem confundidas com o mesmo termo usado na matemática. A componente era determinada pela combinação linear das variáveis que apresentassem a maior variabilidade na matriz de covariância. Mais tarde, a análise que encontrava estas componentes e que maximizava a variância dos dados originais foi denominada por Hotelling de "*Principal Component Analysis"* (HOTELLING, 1933).
> Atualmente, um dos principais usos da *ACP* ocorre quando as variáveis são originárias de processos em que diversas características devem ser observadas ao mesmo tempo. Esta técnica vem sendo estudada por autores como MORRISON (1976), SEBER (1984), REINSEL (1993), JACKSON (1980, 1981) e JOHNSON & WICHERN (1992, 1998).

A idéia central da análise baseia-se na redução do conjunto de dados a ser analisado, principalmente quando os dados são constituídos de um grande número de variáveis inter-relacionadas. Conforme Regazzi (2001, p.1), "procura-se redistribuir a variação nas variáveis (eixos originais) de forma a obter o conjunto ortogonal de eixos não correlacionados". Essa redução é feita transformando-se o conjunto de variáveis originais em um novo conjunto de variáveis que mantém, ao máximo, a variabilidade do conjunto. Isto é, com a menor perda possível de informação. Além disso, esta técnica nos permite o agrupamento de indivíduos similares mediante exames visuais, em dispersões gráficas no espaço bi ou tridimensional, de fácil interpretação geométrica. A redução de dimensionalidade é chamada de *transformação* de *karhunnen-Loéve, ou Análise de Componentes Principal,* no qual os autovalores são chamados de principal.

Na prática, o algoritmo baseia-se na matriz de variância-covariância, ou na matriz de correlação, de onde são extraídos os autovalores e os autovetores.

A análise de componentes principais tem a finalidade de substituir um conjunto de variáveis correlacionadas por um conjunto de novas variáveis não-correlacionadas, sendo essas combinações lineares das variáveis iniciais, e

colocadas em ordem decrescente por suas variâncias, VAR CP_1 > VAR CP_2 > .... > VAR CP_p (VERDINELLI, 1980).

As novas variáveis geradas denominam-se *CP,* e possuem independência estatística e são não correlacionadas. Isso significa que, se as variáveis originais não estão correlacionadas, as *ACP* não oferece vantagem alguma. Variáveis dependentes quer dizer que o conhecimento de uma variável importa para o conhecimento da outra (SOUZA, 2000).

Para a determinação das componentes principais, é necessário calcular a matriz de variância-covariância (Σ), ou a matriz de correlação (R), encontrar os autovalores e os autovetores e, por fim, escrever as combinações lineares, que serão as novas variáveis, denominadas de componentes principais, sendo que cada componente principal é uma combinação linear de todas as variáveis originais, independentes entre si e estimadas com o propósito de reter, em ordem de estimação e em termos da variação total, contida nos dados iniciais, (REGAZZI, 2001).

O esquema descrito na Figura 06 servirá de base para a aplicação da *ACP*.

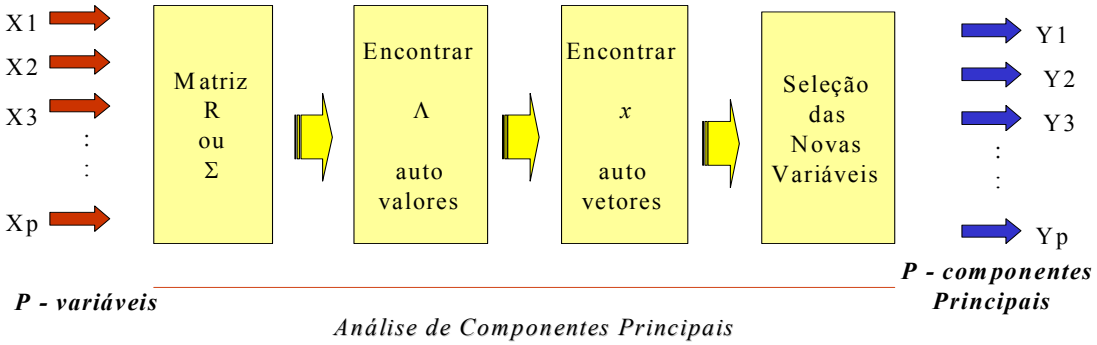


Figura 06 - Esquema da aplicação da análise de componentes principais.
Fonte: SOUZA, Adriano Mendonça (2000, p.25).

Supondo-se que na análise que se está realizando exista apenas duas variáveis X_1 e X_2, conforme a Figura 07, observa-se o elipsóide de densidade de probabilidade constante.

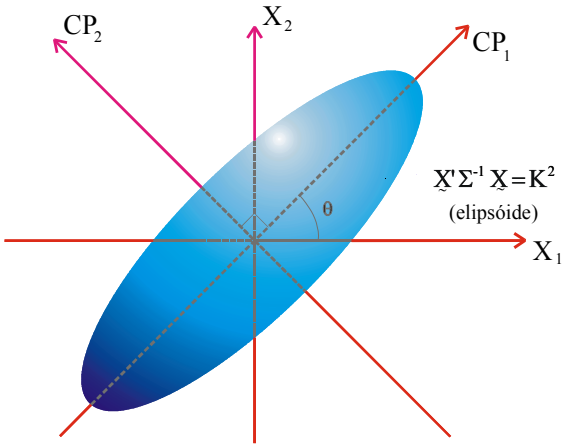


Figura 07 - Elipsóide de densidade constante.
Fonte: LOPES (2001, p.31).

O primeiro componente corresponde ao maior eixo da elipse (CP_1), e o comprimento desse eixo é proporcional a $\sqrt{\Lambda_1}$. O eixo de menor variância (CP_2) é perpendicular ao eixo maior. Esse eixo chama-se segundo componente principal, e seu comprimento é proporcional a $\sqrt{\Lambda_2}$. Assim, a análise das componentes principais toma os eixos X_1 e X_2 e os coloca na direção de maior variabilidade (JOHNSON & WICHERN, 1992).

Para a geração das componentes principais, deve-se ter uma matriz de dimensão nxp, na qual observa-se que $X_1, X_2,...,X_P$ representam as variáveis, e cada uma das *n* unidades experimentais representam os indivíduos, tratamentos, etc. O conjunto de nxp medida origina uma matriz X, conforme mostrado na Tabela 02.

Tabela 02 – Matriz de dados de *n* indivíduos e *p* variáveis.

Indivíduos	Variáveis							
	X_1	X_2	X_3	X_4	...	X_j	...	X_p
1	X_{11}	X_{12}	X_{13}	X_{14}	...	X_{1j}	...	X_{1p}
2	X_{21}	X_{22}	X_{23}	X_{24}	...	X_{2j}	...	X_{2p}
3	X_{31}	X_{32}	X_{33}	X_{34}	...	X_{3j}	...	X_{3p}
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	...	X_{ij}	.	X_{ip}
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
n	X_{n1}	X_{n2}	X_{n3}	X_{n4}	...	X_{nj}	...	X_{np}

Fonte: Regazzi 2001.

O primeiro estágio da *ACP* é a conversão da matriz nxp de dados em uma matriz quadrada, onde *n* é o número de indivíduos e *p* representa um conjunto de variáveis.

Intuitivamente, percebe-se que, quanto maior for o número de variáveis, e quanto mais estas forem interdependentes entre si (algumas têm variância grande, algumas têm variância média, e outras têm variância pequena, e as correlações entre elas assumem valores muito diferentes entre si), será mais fácil comparar indivíduos baseando-se nos valores dessas variáveis, originais (REGAZZI, 2001). Essa interdependência é representada pela matriz de variância-covariância Σ, ou pela matriz de correlação *R*.

Seja Σ a matriz de variância-covariância associada ao vetor aleatório $\vec{X}$ = [X_1, X_2,...,X_P]. Se Σ possuir o par de autovalores e autovetores estimados da amostra analisada, serão representados por $(\hat{\Lambda}_1, X_1)$, $(\hat{\Lambda}_2, X_2)$, ... ,$(\hat{\Lambda}_p, X_p)$, onde $\hat{\Lambda}_1 \geq \hat{\Lambda}_2 \geq ... \geq \hat{\Lambda}_p \geq 0$, e fornecerão o i-ésimo componente principal dado por:

$$Y_i = \vec{x}_i X = \vec{x}_{1i} X_1 + \vec{x}_{2i} X_2 + ... + \vec{x}_{pi} X_p, \ onde\ i = 1, 2, ..., p.$$

Com as escolhas de que:

$$Var(Y_i) = x_i' \sum x_i = \hat{\Lambda}_i \quad i = 1, 2, ..., p$$

$$Cov(Y_i, Y_k) = x_i' \sum x_k = 0 \ \ i, k = 1, 2, ..., p$$

Se algum $\hat{\Lambda}_i$ é igual, a escolha do coeficiente do vetor correspondente $\vec{X}_i$ também será, e, então, Y_i não é único.

Essa definição mostra que os *CP,* são não correlacionados e possuem variâncias iguais ao autovalor de Σ (JOHNSON & WICHERN, 1992)**.**

Para proceder a *ACP,* em casos populacionais, utiliza-se a matriz de variância covariância Σ. Porém, quando se tratar de um conjunto de dados amostrais, a matriz será estimada através da matriz de variância-covariância amostral *S*, e o vetor média por $\vec{\overline{X}}$ =[$\overline{X}_1, \overline{X}_2, ..., \overline{X}_p$] (SOUZA, 2000).

É importante lembrar que, embora as técnicas multivariadas que constam na literatura tenham sido criadas com o objetivo de resolver problemas específicos, como na área de biologia e psicologia, essas podem ser utilizadas para resolver muitos outros problemas práticos nas diversas áreas do conhecimento. Na maioria

das vezes, os objetivos desses problemas práticos só são resolvidos mediante aplicação de mais de uma técnica multivariada, aplicadas em uma seqüência.

Dessa forma, é interessante ter-se uma visão global de todas, ou quase todas, técnicas multivariadas. Entre as técnicas multivariadas mais utilizadas estão: análise de agrupamentos, análise de componentes principais, análise de discriminante, análise de correspondência, dentre outras.

Conforme Reis (1997), a aplicação da *ACP* e *AF* deve incluir:

- As variáveis incluídas na análise;
- As percentagens da variância explicadas por cada uma das componentes principais;
- O número de componentes retidas e a proporção de variância total por elas explicada;
- Uma tabela com a contribuição de cada variável para cada componente (*factor loadings*), antes e depois de ser aplicado um método de rotação de fatores.
- Fazer a interpretação de cada componente principal retido.

2.3 Análise Fatorial - *AF* - relacionada à análise de componentes principais - *ACP*

A *AF* é formada por um conjunto de técnicas estatísticas, e possui como objetivo reduzir o número de variáveis iniciais com a menor perda possível de informação. Em outras palavras, pode-se dizer que *AF* é aplicada à busca de identificação de fatores num conjunto de medidas realizadas, sendo que esses fatores identificados pela *AF* são uma descoberta feita pelo pesquisador. Inicialmente, ele tem várias medidas e não será possível identificar quais variáveis poderão ser reunidas num fator. A *AF* é quem vai descobrir isso, pois ela permite identificar novas variáveis, em um número reduzido em relação às variáveis iniciais, sem uma perda significativa de informação contida nos dados originais.

A versão clássica da *AF* determina os fatores ortogonais que descrevem aproximadamente e sucessivamente os vetores-resposta de *n* indivíduos a um conjunto constituído por *m* testes psicológicos. As primeiras pesquisas realizadas nesta área foram desenvolvidas por Karl Pearson (1901) e por Charles Spearman (1904). Sperman estudou a hipótese da existência de um só fator de inteligência e

da impossibilidade de medi-lo diretamente, ele desenvolveu esta análise para que fosse possível estudar o fator inteligência indiretamente a partir das correlações entre diferentes testes. Em 1947 Thurstone partiu da idéia inicial de Spearman e desenvolveu a *AF,* por acreditar que existe mais de um fator de inteligência. Essa análise permite identificar mais de um fator nos dados iniciais.

A *AF* não se refere, apenas, a uma técnica estatística, mas a um conjunto de técnicas relacionadas, para tornar os dados observados mais claros para a interpretação. Isso é feito analisando-se os inter-relacionamentos entre as variáveis, de tal modo que essas possam ser descritas convenientemente por um grupo de categorias básicas, em número menor que as variáveis originais, chamado fatores.

Os fatores podem ser denominados como um constructo, que pode ser uma variável não observada, escalas, itens, ou uma medida de qualquer espécie. Na análise, fatores explicam a variância das variáveis observadas, tal como se revelam pelas correlações entre as variáveis que estão sendo analisadas.

Um dos métodos mais conhecidos, para a extração dos fatores, é feito por meio da análise de componentes principais, que é baseado no pressuposto que se pode definir $\vec{X}$ vetores estatisticamente não correlacionados, a partir de combinações lineares dos *p* indicadores iniciais.

A *ACP* permite transformar um conjunto de variáveis iniciais, correlacionadas entre si, num outro conjunto de variáveis não correlacionadas (ortogonais), que são as componentes principais, que resultam das combinações lineares do conjunto inicial.

Tanto a análise de componentes principais, quanto a análise fatorial, são técnicas da análise multivariada, que são aplicadas a um conjunto de variáveis, para descobrir quais dessas são mais relevantes, na composição de cada fator, sendo estes independentes um dos outros. Os fatores, que são gerados, são utilizados de maneira representativa do processo em estudo e utilizados para análises futuras.

O objetivo da *ACP* não é explicar as correlações existentes entre as variáveis, mas encontrar funções matemáticas, entre as variáveis iniciais, que expliquem o máximo possível da variação existente nos dados e permita descrever e reduzir essas variáveis. Já a *AF* explica a estrutura das covariâncias, entre as variáveis, utilizando um modelo estatístico casual e pressupondo a existência de *p* variáveis não-observadas e subjacentes aos dados. Os fatores expressam o que existe de comum nas variáveis originais (REIS, 1997).

A *AF* é uma técnica que é aplicada para identificar fatores num determinado conjunto de medidas realizadas, sendo utilizada, também, como uma ferramenta na tentativa de reduzir um grande conjunto de variáveis para um conjunto mais significativo, representado pelos fatores. Esse método determina quais variáveis pertencem a quais fatores, e o quanto cada variável explica cada fator.

Essas duas técnicas, *ACP* e *AF,* são sensíveis a correlações pobres entre variáveis, pois, neste caso, as variáveis não apresentarão uma estrutura de ligação entre elas. Logo, a correlação será fraca e prejudicará as análises, inviabilizando o uso da técnica, que tem como objetivo principal o estudo de conjuntos de variáveis correlacionadas.

Quando se trabalha com *AF,* deve-se levar em consideração que coeficientes de correlação tendem a ser de menor confiança quando se faz cálculos de estimativas de amostra pequenas. Em geral, o mínimo é ter cinco casos, pelo menos, para cada variável observada.

O primeiro passo a ser realizado, quando se aplica *AF,* é verificar as relações entre as variáveis, que pode ser feito utilizando-se o coeficiente de correlação linear como medida de associação entre cada par de variáveis. Conforme Reis (1997), “a matriz de correlação poderá permitir identificar subconjuntos de variáveis que estão muito correlacionadas entre si no interior de cada subconjunto, mas pouco associadas a variáveis de outros subconjuntos”. Nesse caso, utilizar a técnica de *AF* permitirá concluir se é possível explicar esse padrão de correlações mediante um menor número de variáveis.

A *AF* é exploratória, pois é utilizada com o objetivo de reduzir a dimensão dos dados, podendo, também, ser confirmatória, se for utilizada para testar uma hipótese inicial de que os dados poderão ser reduzidos a uma determinada dimensão e de qual a distribuição de variáveis, segundo essa dimensão (REIS, 1997).

A *ACP* e a *AF,* quando utilizadas na forma direta, servem para a identificação de grupos de variáveis inter-relacionadas e para a redução do número de variáveis. Em seu uso indireto é um método que serve para transformar dados. A transformação de dados ocorre através da reescrita dos mesmos, com propriedades que os dados originais não tinham.

Antes de aplicar a *AF*, deve-se levar em consideração certas premissas sobre a natureza dos dados. Primeiramente, o pesquisador deve analisar a

distribuição de freqüência das variáveis através de testes de ajuste da normalidade (Kolmogorov-Smirnov), ou, até, fazer um simples exame de curvas da distribuição. O pesquisador pode, ainda, fazer um gráfico de dispersão (*scatterplot*), fazendo um contraste em relação aos valores observados com os esperados numa distribuição normal (PEREIRA, 2001).

Há, também, uma medida de adequação dos dados, muito importante, sugerida por *Kaiser-Meyer-Olkin Measure of Adequacy (KMO)*. O *KMO* serve para avaliar o valor de entrada das variáveis para o modelo, sendo que seu valor possibilita prover resultados no alcance de 0,5 a 0,9, se se obtiver valores nesse intervalo, então as variáveis podem ser utilizadas para realizar a *AF*. Para encontrar o valor do *KMO*, utiliza-se a expressão:

$$KMO = \frac{\sum_i \sum_j r_{ij}^2}{\sum_i \sum_j r_{ij}^2 + \sum_i \sum_j a_{ij}^2}, \qquad (2.6)$$

sendo a razão da soma dos quadrados das correlações de todas as variáveis dividida por essa mesma soma, acrescida da soma dos quadrados das correlações parciais de todas as variáveis.

Onde:

r_{ij} = é o coeficiente de correlação observado entre as variáveis *i* e *j.*

a_{ij} = é o coeficiente de correlação parcial entre as mesmas variáveis, que é, simultaneamente, uma estimativa das correlações entre os fatores. Os a_{ij} deverão estar próximos de zero, pelo fato de os fatores serem ortogonais entre si.

> Quando as correlações parciais forem muito baixas, o *KMO* terá valor mínimo próximo a 1 e indicará perfeita adequação dos dados para análise fatorial“. O teste do *KMO* possui valores que são considerados críticos como se pode observar:
>
> - para valores na casa dos 0,90: a adequação é considerada ótima para os dados da *AF*;
> - para valores na casa dos 0,80: a adequação é considerada boa para os dados da *AF*;
> - para valores na casa dos 0,70: a adequação é considerada razoável para os dados da *AF*;
> - para valores na casa dos 0,60: a adequação é considerada medíocre para os dados da *AF*;

> - para valores na casa dos 0,50 ou inferiores: a adequação é considerada imprópria para os dados da *AF*;
>
> O *KMO* é uma medida de adequação que verifica o ajuste dos dados, utilizando todas as variáveis simultaneamente, e o seu resultado é uma informação sintética sobre os dados.
>
> Outro teste que poderá ser utilizado para análise fatorial, que também verifica as premissas é o de *Bartlett Test of Sphericity (BTS),* que testa a hipótese da matriz de correlação ser uma matriz identidade, ou seja, a diagonal principal igual a 1 e todos os outros valores serem zero, isto é, seu determinante é igual a 1. Isso significa que não há correlação entre as variáveis. A hipótese nula poderá ser rejeitada caso o α adotado for igual a 5% e o valor encontrado for inferior ao valor de α. O teste de *Bartlett* na aplicação da *ACP* pressupõe que se rejeite a hipótese nula:
>
> H0 = P = I ou H0 = $\hat{\Lambda}_1 = \hat{\Lambda}_2 = ... = \hat{\Lambda}_P$ (PEREIRA 2001, p. 124 e 125).

A análise de correspondência, a análise canônica e a análise fatorial discriminante são, também, métodos fatoriais, que levam a representações gráficas e terão, por isso, traços comuns com *ACP.* O que diferencia a *ACP* é que ela trata, exclusivamente, de variáveis numéricas, que desempenham, todas, o mesmo papel, enquanto a análise de correspondência trata de variáveis qualitativas, nas análises canônicas e discriminante as variáveis são repartidas em grupos bem distintos (BOUROCHE & SAPORTA, 1982).

A *AF* possui, como princípio, cada variável pode ser decomposta em duas partes: uma parte comum e uma parte única. A primeira é a parte da sua variação partilhada com outras variáveis, enquanto a segunda é específica da sua própria variação. Dessa forma, uma diferença entre os dois métodos parte do montante de variância analisada, na qual a *ACP* considera a variação total presente no conjunto das variáveis originais. Na *AF,* só é retida a variação comum, partilhada por todas as variáveis (REIS, 1997).

A base fundamental para a análise de fator comum *ACP* e *AF* é que as variáveis escolhidas podem ser transformadas em combinações lineares de um conjunto de componentes (fatores) hipotéticos, ou despercebidos. Os fatores podem ser associados com uma variável individual (fatores únicos), ou, ainda, associados com duas ou mais das variáveis originais (fatores comuns). As cargas são responsáveis por relacionar a associação específica entre os fatores e as variáveis originais. Logo, pode-se concluir que o primeiro passo é encontrar as cargas e a solução para os fatores, que aproximarão a relação entre as variáveis originais e fatores encontrados, sendo que as cargas são derivadas dos autovalores, que estão associados às variáveis individuais.

Para ter-se uma melhor visualização das variáveis, que melhor representem cada fator, é realizada uma rotação nos eixos, pois a *AF* busca colocar os fatores em uma posição mais simples, com respeito às variáveis originais, que ajudam na interpretação de fatores. Essa rotação coloca os fatores em posições em que serão associadas só às variáveis relacionadas distintamente a um fator. Existem várias rotações que podem ser realizadas para a matriz fatorial, *varimax, quartimax* e *equimax.* São todas as rotações ortogonais, enquanto as rotações oblíquas são não-ortogonais. A rotação *varimax rotation* busca minimizar o número de variáveis com altas cargas num fator, ou seja, maximiza a variância da carga e é, também, o mais utilizado. Conforme Pereira (2001), "a rotação da matriz não afeta a inércia (comunalidades) das variáveis nem a percentagem de variações explicadas pelos fatores".

Antes de aplicar *ACP* e *AF,* o pesquisador deve tomar duas decisões importantes que são: o método a ser utilizado para a extração dos fatores e o número de fatores para serem extraídos.

Antes se falar da interpretação da *AF,* é importante ter claro dois o conceitos: o de ortogonalidade e o de carga fatorial.

O primeiro está relacionado com independência, no qual e deve haver dissociação entre variáveis. E isso é conseguido quando se realiza a *ACP,* onde cada componente é independente da outra. Por isso, a *ACP* é, geralmente, utilizada como uma técnica para se extrair fatores.

O segundo conceito importante é o de carga fatorial. A matriz de cargas fatoriais é um dos passos finais da análise fatorial. A carga fatorial é um coeficiente: um número decimal, positivo ou negativo, geralmente menor do que um, que expressa o quanto um teste, ou variável, observada, está carregado, ou saturado, em um fator. Entre outras palavras, pode-se dizer que: quanto maior for a carga em cima de um fator, mais a variável se identifica com o que quer que seja o fator.

Em resumo, a *AF* é um método para determinar o número de fatores existente em um conjunto de dados, e serve para determinar quais testes, ou variáveis, pertencem a quais fatores.

A *AF,* em seus resultados, apresenta alguns conceitos que devem ser entendidos, para que haja uma interpretação correta dos dados. Como neste trabalho utiliza-se o *software statistica,* os resultados são apresentados com conceitos em língua inglesa. Conforme Pereira (2001), conceitos da *AF*:

- *eigenvalue* corresponde aos autovalores e à variância total, que pode ser explicada pelo fator. Ou seja, avalia a contribuição do fator ao modelo construído pela análise fatorial. Se a explicação da variância pelo fator for alta, existe uma alta explicação desse fator ao modelo, se for baixa, existe uma baixa explicação do fator ao modelo.
- *factor loading* é a proporção de variação da variável, que é explicada pelo fator, ou, ainda, o quanto cada variável contribui na formação de cada componente.
- *factor score* são os autovetores que definem as direções dos eixos da máxima variabilidade. Representam a medida assumida pelos objetos estudados na função derivada da análise.
- *Communality,* é a medida de quanto da variância, de uma variável, é explicada pelos fatores derivados pela análise fatorial. Avalia a contribuição da variável ao modelo construído pela *AF*, ou seja, o quanto cada variável participa na formação da outra. Nas *communality,* os valores mais altos são os mais importantes para análise.
- *factor matrix* é a matriz de correlação entre as variáveis originais e os fatores encontrados.

Para que se possa nomear os fatores, deve-se olhar a pontuação dos mesmos, individualmente, e ver quais variáveis possuem as pontuações mais altas. Deve-se olhar, também, a pontuação do fator, para ver se as interpretações iniciais são confirmadas pela pontuação do fator.

A *ACP* adota a premissa de que a relação entre variáveis e fatores é linear. Dessa forma, pode-se tentar interpretar um eixo, seja graficamente, por regressão linear, entre as coordenadas das amostras e os autovetores de cada variável, ou seja, pelo cálculo de um coeficiente de correlação não-paramétrico (Spearman, por exemplo).

Para que se possa resolver a equação característica, em *AF*, é necessário fazer a inversão de matriz, o que não é possível com uma matriz singular.

A multicolinearidade e singularidade são assuntos derivados de uma matriz de correlação, com alto grau de correlação entre as variáveis. A multicolinearidade acontece quando variáveis são altamente correlacionadas, ou seja, acima de 0.90, o que é muito bom para a *AF*, e a singularidade acontece quando as variáveis são perfeitamente correlacionadas. Com multicolinearidade, os efeitos são aumentados,

as variáveis independentes estão inter-relacionadas. Se a variável é perfeitamente relacionada às outras variáveis, então a singularidade está presente.

Raramente os resultados da *AF* são todos publicados, pois nem todos possuem uma contribuição significativa para a interpretação dos dados e à elaboração de conclusões para o assunto que está sendo abordado.

Conforme Valentin (2000), as informações, que devem constar nas publicações, são:

- as dimensões da matriz de dados: número de variáveis e indivíduos;
- a natureza dos dados e as transformações eventuais;
- as figuras dos planos fatoriais;
- a necessidade de análises preliminares para testar a estabilidade e, se for preciso, eliminar certas variáveis ou observações.

Comentários deste capítulo

Nesse capítulo 2, abordou-se os conceitos de análise de agrupamentos, análise de componentes principais e análise fatorial, que servirão de base para o pleno desenvolvimento da aplicação prática.

No capítulo 3, apresenta-se como estas técnicas são desenvolvidas manualmente.

3 METODOLOGIA

No capítulo 3, item 3.1, apresenta-se o desenvolvimento de exemplo práticos da análise de agrupamentos, que consiste na reunião de elementos semelhantes. No item 3.2, mostra-se conceitos e aplicação de exemplos práticos da análise de componentes principais, sendo que a principal meta, desta análise, é a redução de dimensão do problema e a análise fatorial, que busca fatores abstratos para a representação do conjunto de dados.

3.1 Análise de agrupamentos

Muitos algoritmos existem para formar os agrupamentos. Devido a existência de vários critérios, para conceituar esses grupos, o pesquisador deve optar por aquele que for mais adequado à análise em estudo.

Para aplicar a análise de agrupamento, neste trabalho, optou-se por apresentar os métodos de agrupamento hierárquicos aglomerativos, que tem início com um grupo separado. Primeiramente, os objetos mais similares são agrupados formando um único grupo. Eventualmente, o processo é repetido, e com o decréscimo da similaridade, todos os subgrupos são agrupados, formando um único grupo com todos os objetos.

O desenvolvimento da *AA* será concentrado nos métodos hierárquicos aglomerativos (*Linkage Methods*). Serão discutidos os métodos de ligação simples (mínima distância ou vizinho mais próximo) e ligação completa (máxima distância, ou vizinho mais distante).

Conforme Ferreira (1996), nas etapas a seguir, apresenta-se um algoritmo geral para os agrupamentos hierárquicos aglomerativos com *n* objetos (itens, ou variáveis)

- Iniciar o agrupamento com *n* grupos, cada um com um único elemento, e com uma matriz simétrica $n_x n$ de dissimilaridades (distâncias) $D = \{d_{hi}\}$.
- Buscar na matriz *D* o par de grupos mais similar (menor distância), e fazer a distância entre os grupos mais similares *U* e *V* igual à d_{uv}.
- Fundir os grupos *U* e *V* e nomeá-los por *(UV)*. Recalcular e rearranjar as distâncias na matriz *D*:

(a) eliminando as linhas e colunas correspondentes a *U* e *V* e

(b) acrescentando uma linha e coluna com as distâncias, entre o grupo *(UV)* e os demais grupos.

- Repetir os passos 2 e 3 num total de (n-1) vezes, até que todos os objetos estejam em único grupo. Anotar a identidade dos grupos, que vão sendo agrupados, e os respectivos níveis (distâncias) nas quais isto ocorre.

A seguir, está o desenvolvimento da *AA*, pelos métodos referentes à ligação simples e de ligação completa.

3.1.1 Método de encadeamento único, ou por ligação simples

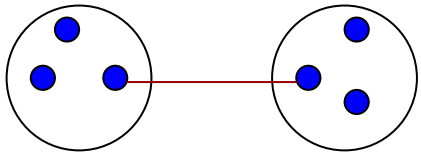


Figura 08 - Distância mínima entre os grupos.

O método de encadeamento único foi introduzido em taxonomia numérica por Florek *et al.* (1951, apud REGAZZI, 2001), no qual os grupos são, inicialmente, constituídos cada um de um indivíduo, simplesmente, e são reunidos de acordo com a proximidade dos elementos, e, então, os indivíduos mais próximos são fundidos. Esse método, que pode ser chamado, também, de salto mínimo, ou vizinho mais próximo, é de concepção simples, podendo ser realizado sem ajuda do computador.

Na Tabela 03 apresenta-se cinco variáveis e quatro indivíduos. Desenvolve-se um exemplo prático do método de encadeamento único.

Para que seja possível formar grupos com características semelhantes, com os valores da Tabela 03, faz-se necessário estabelecer a medida de distância que será utilizada na análise.

Tabela 03 – Número de indivíduos com suas respectivas variáveis.

Indivíduos	Variável 1	Variável 2	Variável 3	Variável 4	Variável 5
1	20	5	11	7	49
2	18	9	10	2	45
3	11	35	30	15	7
4	10	3	7	4	26

Neste exemplo, utilizar-se-á o método do encadeamento único, sendo este uma medida da distância euclidiana, que é um algoritmo de agrupamento. Para saber quais são as menores distâncias, e dar início a formação dos grupos, faz-se necessário calcular estes valores conforme item 2.2:

$$d_{\text{var1, var1}} = \sqrt{(20-20)^2 + (18-18)^2 + (11-11)^2 + (10-10)^2} = 0$$

$$d_{\text{var 1, var 2}} = \sqrt{(5-20)^2 + (9-18)^2 + (35-11)^2 + (3-10)^2} = 30{,}5$$

$$d_{\text{var1, var3}} = \sqrt{(11-20)^2 + (10-18)^2 + (30-11)^2 + (7-10)^2} = 22{,}7$$

$$d_{\text{var1, var4}} = \sqrt{(7-20)^2 + (2-18)^2 + (15-11)^2 + (4-10)^2} = 21{,}8$$

$$d_{\text{var1, var5}} = \sqrt{(49-20)^2 + (45-18)^2 + (7-11)^2 + (26-10)^2} = 49{,}9$$

$$d_{\text{var2, var1}} = \sqrt{(20-5)^2 + (18-9)^2 + (11-35)^2 + (10-3)^2} = 30{,}5$$

$$d_{\text{var2, var3}} = \sqrt{(11-5)^2 + (10-9)^2 + (30-35)^2 + (7-3)^2} = 8{,}8$$

$$d_{\text{var2, var4}} = \sqrt{(7-5)^2 + (2-9)^2 + (15-35)^2 + (4-3)^2} = 21{,}3$$

$$d_{\text{var2, var5}} = \sqrt{(49-5)^2 + (45-9)^2 + (7-35)^2 + (26-3)^2} = 67{,}4$$

As demais distâncias serão obtidas analogamente.

Com todas as distâncias calculadas, obteve-se a seguinte matriz de distâncias euclidiana:

$$D_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0{,}0 & 30{,}5 & 22{,}7 & 21{,}8 & 42{,}9 \\ - & 0{,}0 & 8{,}8 & 21{,}3 & 67{,}4 \\ - & - & 0{,}0 & 17{,}7 & 59{,}7 \\ - & - & - & 0{,}0 & 64{,}5 \\ - & - & - & - & 0{,}0 \end{bmatrix} \end{array}$$

Para ilustrar o método da ligação simples, os objetos menos distantes devem, inicialmente, ser agrupados. Então, com essa matriz das distâncias, é possível dar início à formação dos grupos, sendo que a menor distância existente entre as duas variáveis distintas é 8,8, ou seja, este será o primeiro grupo a ser formado.

$$D_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0{,}0 & 30{,}5 & 22{,}7 & 21{,}8 & 42{,}9 \\ - & 0{,}0 & 8{,}8 & 21{,}3 & 67{,}4 \\ - & - & 0{,}0 & 17{,}7 & 59{,}7 \\ - & - & - & 0{,}0 & 64{,}5 \\ - & - & - & - & 0{,}0 \end{bmatrix} \end{array}$$

Como se pode verificar na matriz acima, a menor distância está na linha 2 e coluna 3, e será representada por $d_{23} = 8{,}8$, logo esses serão os primeiros indivíduos a serem agrupados, 2 e 3.

A Figura 09 refere-se ao primeiro grupo formado da análise referente às variáveis 2 e 3.

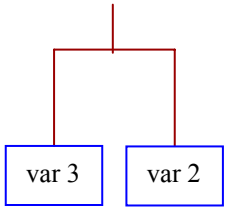


Figura 09- Primeiro grupo formado do agrupamento.

A distância existente entre esse grupo, e os grupos individuais 1, 4 e 5, será obtida pelo método do vizinho mais próximo, como segue:

$$d_{(23)1} = \min\{d_{21}, d_{13}\} = \min\{30{,}5;\ 22{,}7\} = \min d_{13} = 22{,}7$$

$$d_{(23)4} = \min\{d_{24}, d_{43}\} = \min\{21{,}3;\ 17{,}7\} = \min d_{43} = 17{,}7$$

$$d_{(23)5} = \min\{d_{25}, d_{53}\} = \min\{67{,}4;\ 59{,}7\} = \min d_{53} = 59{,}7$$

Logo D_2 será:

$$D_2 = \begin{array}{c} \\ 1 \\ (23) \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & (23) & 4 & 5 \end{array} \\ \begin{bmatrix} 0{,}0 & 22{,}7 & 21{,}8 & 42{,}9 \\ - & 0{,}0 & 17{,}7 & 59{,}7 \\ - & - & 0{,}0 & 64{,}5 \\ - & - & - & 0{,}0 \end{bmatrix} \end{array}$$

A segunda menor distância está na linha 23 e coluna 4, representada em D_3 por $d_{(23)4} = 17{,}7$, logo o indivíduo 4 será incluído no grupo 2 e 3.

A Figura 08 refere-se ao segundo grupo, formado da análise, no qual está sendo adicionada a variável 4 ao grupo de variáveis já formado anteriormente, 23.

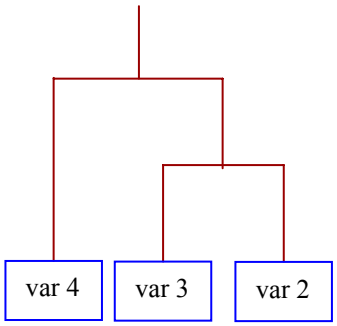


Figura 10 - Segundo grupo formado do agrupamento.

As distâncias serão obtidas pelo método do vizinho mais próximo, de forma análoga aos anteriores:

$$d_{(234)1} = \min\{d_{(23)1}, d_{14}\} = \min\{22{,}7;\ 21{,}8\} = \min d_{14} = 21{,}8$$

$$d_{(234)5} = \min\{d_{(23)5}, d_{45}\} = \min\{59{,}7;\ 64{,}5\} = \min d_{(23)5} = 59{,}7$$

$$D_3 = \begin{array}{c} \\ 1 \\ (234) \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & (234) & 5 \end{array} \\ \begin{bmatrix} 0{,}0 & 21{,}8 & 42{,}9 \\ - & 0{,}0 & 59{,}7 \\ - & - & 0{,}0 \end{bmatrix} \end{array}$$

A terceira menor distância está na linha 1 e coluna 234, e será representada pela matriz D_4 por $d_{(234)1} = 21{,}8$. Incluindo o indivíduo 1 no grupo (234).

A Figura 11 refere-se ao terceiro grupo, formado da análise, no qual está sendo adicionada a variável 1 ao grupo de variáveis já formado anteriormente (234).

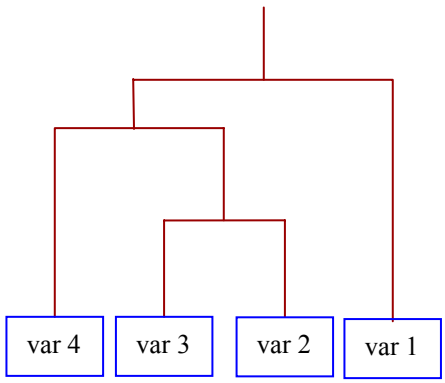


Figura 11 - Terceiro grupo formado do agrupamento.

As distâncias serão obtidas de forma análoga às anteriores:

$$d_{(1234)5} = \min\{d_{15}, d_{(234)5}\} = \{42{,}9;\ 59{,}7\} = \min d_{15} = 42{,}9$$

$$D_4 = \begin{matrix} \\ (1234) \\ 5 \end{matrix} \begin{matrix} \begin{matrix} (1234) & 5 \end{matrix} \\ \begin{bmatrix} 0{,}0 & 42{,}9 \\ - & 0{,}0 \end{bmatrix} \end{matrix}$$

A Figura 12 refere-se ao quarto grupo, formado da análise, no qual está sendo adicionada a variável 5 ao grupo de variáveis já formado anteriormente (1234).

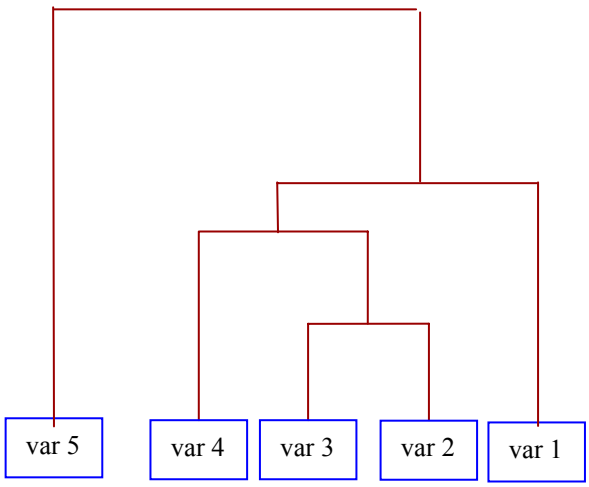


Figura 12 - Quarto e último grupo formado do agrupamento.

Dessa forma, agrupa-se (1234) e 5, formando, assim, o último grupo da análise.

Segundo Valentin (2000, p. 56), o dendograma será formado de acordo com os itens que seguem:

- no eixo vertical são colocados os valores das distâncias, sendo que este dendograma inicia na distância 5 e vai até à distância 45;
- a Figura 13, chamado de dendograma, ou árvore de aglomerados, representa as variáveis que estão em estudo.
- para compor o dendograma, deve-se buscar na matriz de distâncias euclidianas o menor valor, ou a menor distância, isto é, uma maior similaridade entre os elementos. Como já calculado anteriormente, a menor distância encontrada nessa matriz é 8,8. Está entre as variáveis 2 e 3, que serão reunidas no dendograma na altura 8,8 formando, assim, o primeiro grupo I;
- a segunda menor distância é 17,7, que está entre as variáveis 2 e 3, que já pertence ao grupo I anteriormente formado, e a variável 4. A variável 4 deve,

então, ser reunida no primeiro grupo, ao nível de distância de 17,7, formando, assim, o grupo II;

- a próxima distância é 21,8, que está entre as variáveis 2, 3 e 4, que já pertence ao primeiro grupo I, e a variável 1. Como a variável 3 pertence ao grupo I, já ligado com a variável 4, agrupa as variáveis do grupo I e do grupo II, formando, assim, o grupo III;
- a próxima, e última distância, é 42,9, que está entre as variáveis 1, 2, 3 e 4, e a variável 5, como a variável 1, já está ligada a outros grupos. Vai agrupar todos os grupos existentes, deixando, dessa forma, o dendograma completo, com um grupo único, agrupando, assim, todas as variáveis.

No dendograma da Figura 13, a escala vertical indica o nível de similaridade, e no eixo horizontal são marcados os indivíduos, na ordem em que são agrupados. As linhas verticais partem dos indivíduos, e têm altura correspondente ao nível em que os indivíduos são considerados semelhantes.

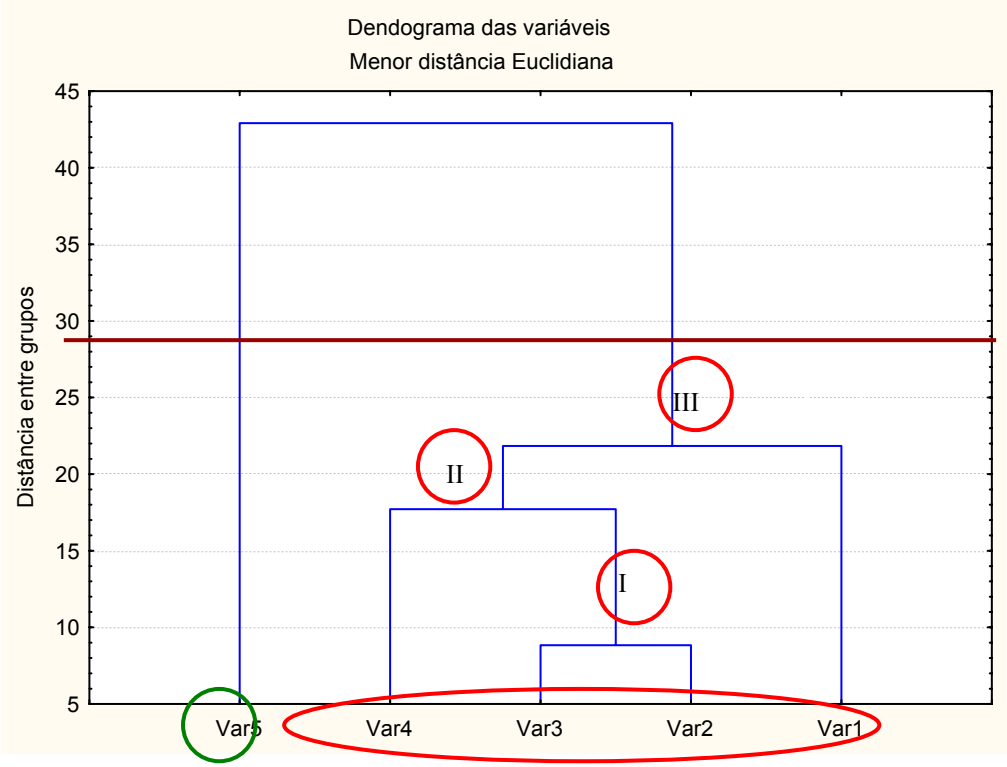


Figura 13 - Dendograma da matriz de distâncias pelo método de ligação simples, representado utilizando o programa computacional *statistica.*

Observando a Figura 13, é possível verificar que o maior salto encontra-se entre as distâncias 21,8 e 42,9. Se fizer um corte no gráfico, entre essas distâncias, existirão dois grupos homogêneos distintos: o primeiro grupo, formado pelas

variáveis de um, dois, três e quatro, que é representado pelo círculo em vermelho e o segundo grupo, formado pela quinta variável, representado pelo círculo em verde, sendo essa variável distinta das demais, pelo fato de ter formado um grupo isolado, isso significa dizer que esta variável é heterogênea em relação às outras.

Esses grupos foram definidos pelo traçado de uma linha paralela ao eixo horizontal, denominada “Linha Fenon”. Optou-se por traçar essa linha entre as alturas 21,8 e 42,9, que representam as distâncias euclidianas de ligação entre as vaiáveis.

O método do vizinho mais próximo pode ser resumido da seguinte forma, como mostra a Tabela 04:

Tabela 04 – Resultado da análise de agrupamentos, pelo método do vizinho mais próximo.

Passo	Junção	Níveis
1	2,3	8,8
2	23,4	17,7
3	234,1	21,8
4	1234,5	42,9

Em razão da sua simplicidade, esse método apresenta grande desvantagem. O fato de reunir um objeto ao elemento “mais próximo” do grupo já formado, faz com que os objetos intermediários entre os grupos sejam rapidamente aglomerados a esses. Ocorre, então, um encadeamento de objetos que dificulta a separação dos grupos. Nos estudos, ecológicos em que as amostras de características intermediárias são geralmente numerosas, esse método deve ser evitado (VALENTIN, 2000).

3.1.2 Método de encadeamento completo ou por ligação completa

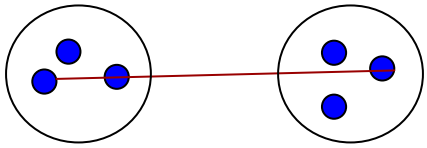


Figura 14 - Distância máxima entre grupos.

Esse método foi introduzido em 1948, sendo exatamente o oposto ao método do vizinho mais próximo, em que a distância entre grupos será definida como a distância entre os pares de indivíduos mais distantes.

Aqui, a distância entre dois grupos é definida pelos objetos de cada grupo que estão mais distantes. Ou seja, formam-se todos os pares com um membro de cada grupo. A distância entre os grupos é definida pelo par que possuir maior distancia (BUSSAB *et al*, 1990).

É importante ressaltar que a união ainda é feita com os grupos mais parecidos, ou seja, a menor distância. Para ilustrar, serão utilizados neste exemplo os dados referentes a Tabela 03, considerando-se a mesma matriz de dissimilaridade *D* do exemplo anterior. Inicialmente, serão agrupados os dois objetos menos distantes. Então, o dendograma será construído através do método do encadeamento completo, ou do vizinho mais distante.

$$D_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0{,}0 & 30{,}5 & 22{,}7 & 21{,}8 & 42{,}9 \\ - & 0{,}0 & 8{,}8 & 21{,}3 & 67{,}4 \\ - & - & 0{,}0 & 17{,}7 & 59{,}7 \\ - & - & - & 0{,}0 & 64{,}5 \\ - & - & - & - & 0{,}0 \end{bmatrix} \end{array}$$

Observando a matriz D_1, a menor distância está no elemento da linha 2 e coluna 3. Esta distância é representado por $d_{23} = 8{,}8$, logo, esses serão os primeiros indivíduos a serem agrupados 2 e 3. A distância existente entre esse grupo, e os grupos individuais 1, 4 e 5, serão obtidas pelo método do vizinho mais distante, conforme segue:

$$d_{(23)1} = \max\{d_{21}, d_{13}\} = \max\{30{,}5,\ 22{,}7\} = \max d_{21} = 30{,}5$$

$$d_{(23)4} = \max\{d_{24}, d_{43}\} = \max\{21{,}3,\ 17{,}7\} = \max d_{24} = 21{,}3$$

$$d_{(23)5} = \max\{d_{25}, d_{53}\} = \max\{67{,}4,\ 59{,}7\} = \max d_{25} = 67{,}4$$

Logo D_2 será:

$$D_2 = \begin{array}{c} \\ 1 \\ (23) \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & (23) & 4 & 5 \end{array} \\ \begin{bmatrix} 0{,}0 & 30{,}5 & 21{,}8 & 42{,}9 \\ - & 0{,}0 & 21{,}3 & 67{,}4 \\ - & - & 0{,}0 & 64{,}5 \\ - & - & - & 0{,}0 \end{bmatrix} \end{array}$$

A menor distância em D_2 é o elemento que está localizado na linha 23 e coluna 4. Este elemento é representado pela distância $d_{(23)4} = 21{,}3$, logo o indivíduo 4 será incluído no grupo 2 e 3. As distâncias serão obtidas pelo método do vizinho mais distante, de forma análoga ao anterior:

$$d_{(234)1} = \max\{d_{(23)1}, d_{14}\} = \max\{30{,}5,\ 21{,}8\} = \max d_{(23)1} = 30{,}5$$

$$d_{(234)5} = \max\{d_{(23)5}, d_{45}\} = \max\{67{,}4,\ 64{,}5\} = \max d_{(23)5} = 67{,}4$$

$$D_3 = \begin{array}{c} \\ 1 \\ (234) \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & (234) & 5 \end{array} \\ \begin{bmatrix} 0{,}0 & 30{,}5 & 42{,}9 \\ - & 0{,}0 & 67{,}4 \\ - & - & 0{,}0 \end{bmatrix} \end{array}$$

A menor distância da matriz D_3 é o elemento da linha 1 e coluna 234. Essa distância é dada por $d_{(234)1} = 30{,}5$ incluindo, assim, o indivíduo 1 no grupo (234), e as distâncias serão obtidas pelo método do vizinho mais distante, da mesma forma que as anteriores:

$$d_{(1234)5} = \max\{d_{15}, d_{(234)5}\} = \max\{42{,}9,\ 67{,}4\} = \max d_{(234)5} = 67{,}4$$

$$D_4 = \begin{array}{c} \\ (1234) \\ 5 \end{array} \begin{array}{c} \begin{array}{cc} (1234) & 5 \end{array} \\ \begin{bmatrix} 0{,}0 & 67{,}4 \\ - & 0{,}0 \end{bmatrix} \end{array}$$

Dessa forma, agruparam-se os indivíduos (1234) e 5, formando, assim, o último grupo do dendograma. A Figura 15 representa o dendograma vertical da matriz de distâncias, pelo método de ligação completa.

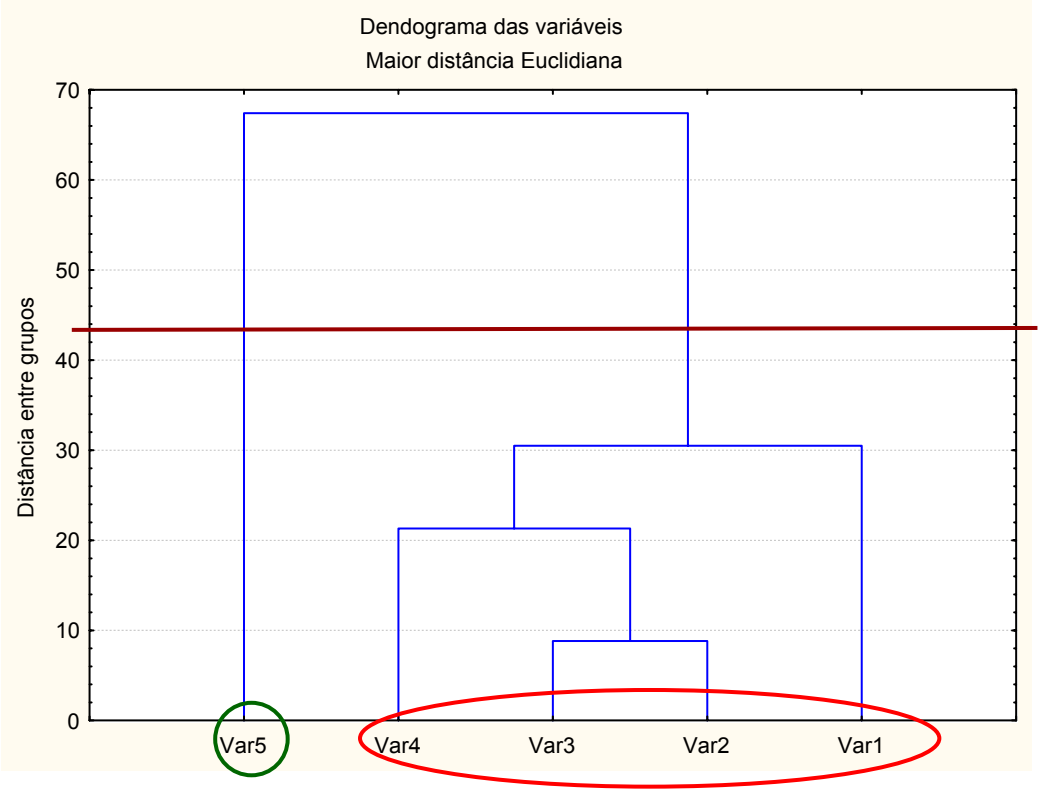


Figura 15 - Dendograma da matriz de distâncias pelo método de ligação completa.

Para analisar esse dendograma, deve-se ter cuidado, pois a união de dois grupos depende do par de objetos mais distantes. Pode-se dizer que um elemento unir-se-á a um grupo unicamente se for ligado a todos os elementos desse grupo.

Observando-se a Figura 15, é possível verificar que o maior salto está na última etapa, se se fizer um corte no gráfico entre a altura 30,5 e 67,4 ter-se-á dois grupos homogêneos distintos. O primeiro grupo será formado pelas variáveis de um a quatro, representado pelo círculo em vermelho, o segundo grupo será formado pela quinta variável, representado pelo círculo em verde, sendo que esta variável é distinta das demais, pelo fato de ter formado um grupo isolado.

Comparando-se os resultados alcançados, e apresentados nas Figuras 13 e 15, pode-se notar que os dendrogramas, para o método do vizinho mais próximo e do vizinho mais distante, não diferem na alocação dos objetos, para esse exemplo em particular.

Os algoritmos vistos produzem grupos que constituem uma proposição sobre a organização básica e desconhecida dos dados. Entretanto, eles esbarram em uma dificuldade, que é a determinação do número ideal de grupos a serem formados (REGAZZI, 2001).

Tabela 05 – Resumo do método do vizinho mais distante.

Passo	Junção	Nível
1	2,3	8,8
2	23,4	21,3
3	234,1	30,5
4	1234,5	67,4

3.1.3 Como escolher o melhor método?

Até hoje não se sabe muito a respeito de qual técnica é a mais adequada para aplicar para certo tipo de dados. Independente do método usado para resumir os dados, é importante que sejam efetuadas medidas do grau de ajuste entre a matriz original dos coeficientes de distância e a matriz resultante do processo de agrupamento ROHLF (1970, *apud* REGAZZI, 2001). Sendo que, quanto maior for o grau de ajuste, menor será a distorção ocasionada pelo método. Alguns autores consideram que acima de 7,0 o grau é considerado bom, e que abaixo de 7,0 existe inadequação no método de agrupamento, para resumir a informação do conjunto de dados.

Segundo Valentin (2000, p.60), “um método é melhor que outro quando o dendograma fornece uma imagem menos distorcida da realidade”. Pode-se avaliar o grau de deformação provocado pela construção do dendograma através do “coeficiente de correlação cofenético”, que serve para medir o grau de ajuste entre a matriz de dissimilaridade (matriz fenética *F*) e a matriz resultante da simplificação proporcionada pelo método de agrupamento (matriz cofenética *C*).

Esse coeficiente de correlação cofenético é o coeficiente *r* de Pearson, sendo calculado entre índices de similaridade da matriz original e os índices reconstituídos com base no dendograma. Logo, quanto maior for o *r,* menor será a distorção. Conforme Valentim (2000, p.60), “há sempre um certo grau de distorção, pois o *r* nunca será igual a 1”.

O coeficiente de correlação momento produto é dado pela seguinte expressão:

$$r_{nm} = \frac{\sum_{j=1}^{n-1} \sum_{j'=j+1}^{n} \left(c_{jj'} - \bar{c}\right)\left(f_{jj'} - \bar{f}\right)}{\sqrt{\sum_{j=1}^{n-1} \sum_{j'=j+1}^{n} \left(c_{jj'} - \bar{c}\right)^2} \sqrt{\sum_{j=1}^{n-1} \sum_{j'=j+1}^{n} \left(f_{jj'} - \bar{f}\right)^2}}, \tag{3.1}$$

onde $\bar{c}$ *e* $\bar{f}$ são as médias aritméticas, definidas por:

$$\bar{c} = \frac{\sum_{i=1}^{n} c_i}{n}, \tag{3.2}$$

$$\bar{f} = \frac{\sum_{j=1}^{n} f_j}{n}. \tag{3.3}$$

A Tabela 06 mostra o rendimento de quatro variedades de milho em quatro colheitas diferentes. Utilizar-se-á estes dados para desenvolver um exemplo prático do coeficiente de correlação cofenético.

Tabela 06 – Rendimento de quatro variedades de milho em quatro colheitas.

	Indivíduos			
Características	1ª colheita	2ª colheita	3ª colheita	4ª colheita
Premium	22,00	24,00	20,00	26,00
AG_9020	20,00	19,00	22,00	25,00
AG_9090	24,00	20,00	28,00	23,00
Agroeste	21,00	26,00	24,00	25,00

Para que seja possível calcular os valores da matriz cofenética *C*, faz-se necessário estabelecer a medida de distância que será utilizada na análise.

Neste exemplo, utilizar-se-á o método do encadeamento único, sendo este uma medida da distância euclidiana média, que é um algoritmo de agrupamento. Para calcular os valores da distância euclidiana média, utiliza-se a expressão do item 2.3.

$$d_{11} = \sqrt{\frac{1}{4}\left[(22-22)^2 + (20-20)^2 + (24-24)^2 + (21-21)^2\right]} = 0$$

$$d_{12}=\sqrt{\frac{1}{4}\left[(24-22)^2+(19-20)^2+(20-24)^2+(26-21)^2\right]}=3{,}39$$

$$d_{13}=\sqrt{\frac{1}{4}\left[(20-22)^2+(22-20)^2+(28-24)^2+(24-21)^2\right]}=2{,}87$$

$$d_{14}=\sqrt{\frac{1}{4}\left[(26-22)^2+(25-20)^2+(23-24)^2+(25-21)^2\right]}=3{,}81$$

As demais distâncias são obtidas de forma análoga, sendo que a matriz de distâncias D_1, ou seja, a matriz fenética de *F* é dada por:

$$D_1 = F = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 3{,}39 & 2{,}87 & 3{,}81 \\ - & 0 & 4{,}82 & 3{,}54 \\ - & - & 0 & 4{,}21 \\ - & - & - & 0 \end{bmatrix} \end{array}$$

Na matriz D_1, a menor distância está localizado na linha 1 e coluna 3. Essa distância é dada por $d_{13}=2{,}87$, logo, os indivíduos 1 e 3 irão formar um grupo, sendo que as distâncias serão dadas por:

$$d_{(13)2}=\min\{d_{21},d_{23}\}=\{3{,}39,\ 4{,}82\}=\min d_{21}=3{,}39$$

$$d_{(13)4}=\min\{d_{41},d_{43}\}=\{3{,}81,\ 4{,}21\}=\min d_{41}=3{,}81$$

Logo a matriz D_2 será:

$$D_2 = \begin{array}{c} \\ 13 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} 13 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & 3{,}39 & 3{,}81 \\ - & 0 & 3{,}54 \\ - & - & 0 \end{bmatrix} \end{array}$$

Observando-se a matriz D_2, é possível verificar que a menor distância é o elemento localizado na linha 13 e coluna 2, sendo que esta é dada por $d_{(13)2}=3{,}39$. Logo, o indivíduo 2 será incluído no grupo de 1 e 3. Nesta etapa serão agrupadas as variáveis (123) e 4, formando, dessa maneira, um único grupo.

$$d_{(123)4}=\min\{d_{(13)2},d_{42}\}=\min\{3{,}81,\ 3{,}54\}=\min d_{42}=3{,}54.$$

Logo:

$$D_3 = \begin{matrix} \\ (123) \\ 4 \end{matrix} \begin{matrix} \begin{matrix} (123) & 4 \end{matrix} \\ \begin{bmatrix} 0 & 3{,}54 \\ - & 0 \end{bmatrix} \end{matrix}$$

Pode-se fazer um resumo desse método, do vizinho mais próximo, através da Tabela 07.

Tabela 07 – Resumo do método do vizinho mais próximo.

Passos	Junção	Nível
1	1,3	2,87
2	13,2	3,39
3	123,4	3,54

O dendograma da Figura 16 mostra os grupos formados com os dados da Tabela 06:

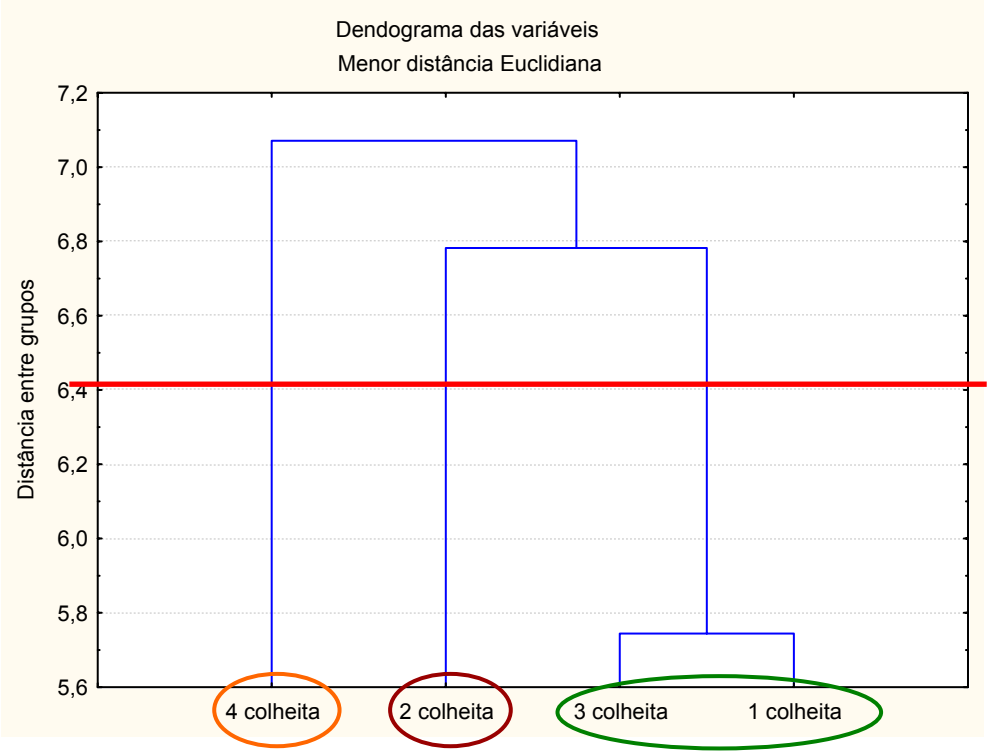


Figura 16 - Dendograma da matriz de distâncias pelo método de ligação simples.

Esse dendograma formou três grupos distintos, no qual o grupo representado pelo círculo em verde engloba a primeira e a terceira colheita. Devido a

isso, pode-se dizer que essas duas variáveis são semelhantes entre si. Já as variáveis que representam a segunda e a quarta colheita formaram dois grupos distintos entre si e entre o primeiro grupo formado, por se manterem isoladas das demais.

As menores distâncias encontradas, através do método do vizinho mais próximo, serão utilizadas para compor a matriz cofenética. Essas distâncias encontradas passam a formar as linhas e as colunas dessa matriz. Logo, o elemento 2,87 estará localizado na linha 1 e coluna 3 da matriz cofenética. Já o elemento da 3,39 estará localizado na linha 1 e coluna 2, e na linha 2 e coluna 3 da matriz cofenética. O elemento 3,54 estará localizado nas seguintes linhas e seguintes colunas: linha 1 e coluna 4, linha 2 e coluna 4, linha 3 e coluna 4, formando, assim, a matriz cofenética C.

(1,3) = 2,87

(13,2) = 1,2 e 2,3 = 3,39

(123,4) = 1,4; 2,4; 3,4 = 3,54.

Logo, a matriz cofenética C é composta pelos seguintes elementos:

$$C = \begin{bmatrix} - & 3{,}39 & 2{,}87 & 3{,}54 \\ - & - & 3{,}39 & 3{,}54 \\ - & - & - & 3{,}54 \\ - & - & - & - \end{bmatrix}$$

A partir dos valores da matriz cofenética *C*, passa-se a calcular o coeficiente de correlação cofenética dado por:

Tabela 08 – Valores correspondentes à matriz fenética e cofenética.

F	C
3,39	3,39
2,87	2,87
3,81	3,54
4,82	3,39
3,54	3,54
4,21	3,54

onde:

F = matriz fenética, na qual seus valores foram obtidos junto à matriz inicial das distâncias.

C = matriz cofenética, na qual os valores são obtidos junto à matriz final das distâncias, pelo método do vizinho mais próximo.

Para obter o coeficiente de correlação cofenético, deve-se calcular os valores da média e desvio padrão das matrizes fenética e cofenética.

A média da matriz fenética, é calculada mediante a expressão do item 3.3.

$$\bar{f} = \frac{3{,}39 + 2{,}87 + 3{,}81 + 4{,}82 + 3{,}54 + 4{,}21}{6} = 3{,}77.$$

A expressão 3.5 refere-se à variância da matriz fenética.

$$S_F^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1} = \frac{\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \ldots + \left(X_n - \overline{X}\right)^2}{n-1} \qquad (3.4)$$

$$S_F^2 = \frac{(3{,}39 - 3{,}77)^2 + (2{,}87 - 3{,}77)^2 + \ldots\ldots + (4{,}21 - 3{,}77)^2}{6-1} = 0{,}46.$$

O desvio padrão da matriz fenética será dado por:

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}} = \sqrt{\frac{\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \ldots + \left(X_n - \overline{X}\right)^2}{n-1}} \qquad (3.5)$$

$$S_F = \sqrt{0{,}46} = 0{,}68.$$

A média da matriz cofenética, é calculada mediante a expressão do item 3.2.

$$\bar{c} = \frac{3{,}39 + 2{,}87 + 3{,}54 + 3{,}39 + 3{,}54 + 3{,}54}{6} = 3{,}38.$$

Variância da matriz cofenética.

$$S_F^2 = \frac{(3{,}39 - 3{,}38)^2 + (2{,}87 - 3{,}38)^2 + \ldots\ldots + (3{,}54 - 3{,}38)^2}{6-1} = 0{,}07.$$

O desvio padrão da matriz cofenética será dado por:

$$S_C = \sqrt{0{,}07} = 0{,}26.$$

A medida de correlação é dada pela covariância entre as duas variáveis, definida por:

$$\hat{Cov}_{FC} = \frac{1}{n-1}\left[\sum x.y - \frac{\sum x.\sum y}{n}\right] \quad (3.6)$$

$\sum xy =$ 3,39.3,39 + 2,87.2,87 + 3,81.3,54 + 4,82.3,39 + 3,54.3,54 + 4,21.3,54

$\sum xy =$ 76,99

$\sum x =$ 22,64

$\sum y =$ 20,27,

logo a Cov_{FC} é dada por:

$$Cov_{FC} = \frac{1}{6-1}\left[76,99 - \frac{22,64.20,27}{6}\right] = 0,10.$$

Sendo mais conveniente usar, para medida de correlação cofenética, o coeficiente de correlação linear de Pearson, definida por:

$$r_{cof} = r_{FC} = \frac{Cov(F,C)}{\sqrt{\hat{V}(F).\hat{V}(C)}} \quad (3.7)$$

$$r_{cof} = \frac{0,10}{\sqrt{(0,46)(0,07)}} \cong 0,56.$$

Como $r_{cof} = 0,56 <$ 0,7, pode-se concluir que o método utilizado não foi adequado para resumir a informação ao conjunto de dados. Logo, deve-se utilizar outros métodos para fazer a análise dos dados.

3.1.4 Interpretação do dendograma

Existem três *regras de bolso,* que se deve utilizar para interpretar um dendograma, Valentim (2000, p.61).

- escrever no próprio dendograma, em frente de cada amostra, as suas características, tudo o que poderá revelar os aspectos comuns entre as amostras de um mesmo grupo e as diferenças com as amostras de outro grupo;

- Começar a “ler” o dendograma dos baixos valores de similaridade, para os maiores. Assim, deverão ser interpretados, em primeiro lugar, os “grandes grupos”, geralmente poucos numerosos, pois seria em vão tentar explicar os grupos menores sem ter conseguido formular, antes, uma hipótese plausível sobre os grandes;
- Quando é possível, desenvolver, paralelamente, com os mesmos dados, uma análise de ordenação, que evidenciará os fatores responsáveis pelos agrupamentos.

3.2 Análise de Componentes Principais

Para aplicar a análise de componentes principais, deve-se seguir algumas etapas até obter-se o resultado final.

Inicialmente, calcula-se a matriz *S,* ou a matriz *R,* e verifica-se se as variáveis estão correlacionadas umas em relação as outras. Caso não estejam, deve-se aplicar o teste do *KMO*, ou fazer um teste que verifique se as correlações entre as variáveis são significativas, ou não, para verificar se é possível proceder a análise dos dados aplicando esta técnica.

O pesquisador deve verificar, também, se as variáveis foram medidas em escalas diferentes. Deve-se proceder a padronização das mesmas, para evitar erros nos resultados.

Na etapa seguinte, decide-se pelo número total de componentes que melhor explicarão o conjunto de variáveis originais. Existem duas formas de selecionar esses componentes:

- Mediante os autovalores, pelo critério sugerido por KAISER (1960) apud MARDIA (1979), que consiste em incluir somente aquelas componentes cujos valores próprios sejam superiores a 1. Este critério tende a incluir poucas componentes quando o número de variáveis originais é inferior a vinte e, em geral, utiliza-se aquelas componentes que conseguem sintetizar uma variância acumulada em torno de 70%.
- Através do método gráfico, este critério considera as componentes anteriores ao ponto de inflexão da curva. Foi sugerido por CATTEL (1966) e exemplificado por PLA (1986).

Decidido o número de componentes, passa-se a encontrar os autovetores que irão compor as combinações lineares, que irão formar as novas variáveis.

A última etapa será fazer normalização e a ortogonalização dos autovetores, para garantir solução única as componentes principais e, também, que estas sejam independentes umas das outras.

Matriz de variância-covariância

A matriz de variância-covariância é expressa pelas ligações realizadas entre as *p* variáveis, tomadas duas a duas sendo, resumidas por suas covariâncias s_{ij} .

Conforme Regazzi (2001), considerando as variáveis $X_1, X_2, ..., X_P$, denota-se a matriz de covariância por *S* da seguinte forma:

$$S = \begin{bmatrix} Vâr(X_1) & Côv(X_1,X_2) & ..... & Côv(X_1,X_p) \\ Côv(X_1,X_2) & Vâr(X_2) & ..... & Côv(X_2,X_p) \\ ..... & ..... & ..... & ..... \\ Côv(X_1,X_p) & Côv(X_2,X_p) & ..... & Vâr(X_p) \end{bmatrix} \text{ ou } S = \begin{bmatrix} S_1^2 & S_{12} & . & . & . & S_{1p} \\ & S_2^2 & . & . & . & S_{2p} \\ & & . & . & . & S_{3p} \\ & & & . & . & . \\ & & & & . & . \\ & & & & & S_P^2 \end{bmatrix} \quad (3.8)$$

sendo que o conjunto de variância-covariância está representado na matriz *S,* chamada matriz de variância-covariância das *p* variáveis. O termo situado na intercessão da i-ésima linha e da j-ésima coluna é a covariância de $\left(s_{ij}\right)$, e os termos da diagonal principal são as variâncias $\left(s_i^2\right)$.

$$Vâr(X_j) = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_{ij'}{}^2 \frac{(\sum_{i=1}^{n} X_{ij'})^2}{n}\right] \quad (3.9)$$

$$Côv(X_j, X_{j'}) = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_{ij} X_{ij'} - \frac{\left(\sum_{i=1}^{n} X_{ij}\right)\left(\sum_{i=1}^{n} X_{ij'}\right)}{n}\right] \quad (3.10)$$

Observando-se a matriz *S,* pode-se concluir que é uma matriz quadrada de ordem pxp, simétrica, pois $s_{ij} = s_{ji}$.

A seguir, representa-se um exemplo prático dos procedimentos, para calcular a matriz *S,* utilizando-se os dados da Tabela 09, referentes a duas variáveis X e Y, sendo estas mensuradas em uma amostra constituída de cinco observações (indivíduos).

Tabela 09 – Observações relativas a duas variáveis X e Y avaliadas em cinco indivíduos.

Observações	Método X	Método Y
1	10,0	10,7
2	10,4	9,8
3	9,7	10,0
4	9,7	10,1
5	11,7	11,5

O primeiro procedimento a ser realizado será a análise descritiva nas duas variáveis, sendo que os resultados obtidos serão utilizados na análise subseqüente, para constituir a matriz *S.*

A Tabela 10 refere-se à estatística descritiva relativa as duas variáveis que estão sendo utilizadas na análise.

Tabela 10 – Estatística descritiva relativa a duas variáveis, avaliadas em cinco indivíduos.

	Método X	Método Y
Média aritmética das variáveis	10,3	10,42
Somatório ao quadrado das variáveis	533,23	544,79
Somatório das variáveis	51,5	52,1
Variância amostral das variáveis	0,70	0,48
Desvio padrão amostral das variáveis	0,84	0,69

A matriz de variância e covariância *S* é estimada conforme item 3.8.

Como pela estatística descritiva já foram encontrados os valores de S_x^2 e S_y^2, deve-se calcular o valor da covariância entre x e y, que serão fornecidos através do item 3.10.

Substituindo-se os dados na expressão, tem-se que:

$$Côv(x,y) = \frac{1}{5-1}\left[538{,}44 - \frac{51{,}5.52{,}1}{5}\right]$$

$$Côv(x,y) = \frac{1}{4}\left[538{,}4 - 536{,}63\right] = 0{,}45.$$

Logo, a matriz ***S*** é assim constituída:

$$S = \begin{bmatrix} 0{,}69 & 0{,}45 \\ 0{,}45 & 0{,}48 \end{bmatrix}$$

- **Matriz de correlação**

A matriz de correlação é utilizada quando se necessita de uma padronização dos dados, evitando-se problemas como a influência da magnitude das variáveis SOUZA (2000, *apud* JACKSON, 1981).

Considerando-se $X_1, X_2, ..., X_p$, as variáveis originais, a estimativa da matriz de correlação (que é igual à estimativa da matriz de variância-covariância entre as variáveis padronizadas $Z_1, Z_2, ..., Z_p$) é denotada por *R,* da seguinte forma:

$$R = \begin{bmatrix} 1 & r_{12} & . & . & . & r_{1p} \\ r_{12} & 1 & . & . & . & r_{2p} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ r_{1p} & r_{2p} & . & . & . & 1 \end{bmatrix} \qquad (3.11)$$

na qual:

$$r_{jj'} = r(X_j, X_{j'}) = Côv(Z_j, Z_{j'}) = \frac{Côv(X_j, X_{j'})}{\sqrt{Vâr(X_j).Vâr(X_{j'})}} \qquad (3.12)$$

Como é possível de se observar, os termos da diagonal principal na matriz de correlação *R* valem, todos, 1, pois a correlação entre r_{11}, r_{22}, ..., r_{np} é igual a 1.

para j = 1, 2, ..., p.

A matriz *R* é uma matriz quadrada de ordem pxp, simétrica em relação a diagonal principal, pois $r_{ij} = r_{ji}$.

Ainda utilizando os dados da Tabela 09, faz-se um exemplo prático com todos os procedimentos necessários para constituir a matriz de correlação *R,* referente ao item 3.11:

Para ilustrar os cálculos, apresenta-se, a seguir, a correlação entre X e Y, utilizando-se a expressão do item 3.12.

Substituindo-se, na expressão, os valores da covariância entre X e Y e S_x, S_y, já calculados anteriormente, junto ao exemplo da matriz de *S,* obtém-se a correlação de $r_{12}\ e\ r_{21}$:

$r_{12} = \dfrac{0,45}{0,83.0,69}$ = 0,79.

Como a correlação entre $r_{12} = r_{21} = r_{xy}$, logo $r_{11} = r_{22} = r_{xy}$ também são equivalentes, calculando-se, apenas uma das correlações, obtém-se o valor da outra.

$$r_{11} = \frac{Côv(X_1, X_1)}{S_{x_1}.S_{x_1}} = \frac{S_{x_1}^{\ 2}}{S_{x_1}^{\ 2}}, \qquad (3.13)$$

$r_{11} = \dfrac{0,83^2}{0,83^2} = 1$.

Logo, a matriz de correlação *R* será assim constituída:

$$R = \begin{bmatrix} 1 & 0,79 \\ 0,79 & 1 \end{bmatrix}.$$

A solução, utilizando-se a matriz de correlação, é recomendada quando as variáveis são medidas em escalas muito diferentes entre si, pois essa matriz é equivalente à matriz das variáveis padronizadas, (JOHNSON & WICHERN, 1992).

Detalha-se a partir de agora um exemplo numérico para o cálculo das componentes principais, mediante a matriz *S* e *R.*

Segundo Magnusson & Maurão (2003, p.106), “estabelecendo-se algumas premissas importantes e usualmente improváveis, é possível determinar a posição dos eixos no espaço multidimensional usando-se a álgebra de matrizes”.

As análises baseadas nesse princípio são chamadas de análises de “auto-vetores”, sendo que “Eigen” é uma palavra da língua alemã, que significa “característica”.

O escalar $\hat{\Lambda}$ será chamado de autovalor, e o vetor $\vec{x}$ um autovetor.

Seja *S* a matriz de variância-covariância quadrada pxp, e *I* a matriz identidade pxp, então os escalares $\hat{\Lambda}_1, \hat{\Lambda}_2, ..., \hat{\Lambda}_p$ satisfazem a equação polinomial.

$$\left| S - \hat{\Lambda} I \right| = 0 \qquad (3.14)$$

são chamados autovalores, ou raízes características, da matriz *S*.

Seja *S* a matriz de variância-covariância de dimensão pxp, e seja $\hat{\Lambda}$ um autovalor de *S*. Logo $\vec{x}$ é um vetor não nulo ($x \neq 0$), tal que:

$$S\vec{X} = \hat{\Lambda}\,\vec{X}, \qquad (3.15)$$

no qual, $\vec{X}$ é uma matriz pxp de todos autovetores, e $\hat{\Lambda}$ é uma matriz pxp de todos autovalores.

Então $\vec{x}$ é dito autovetor ou vetor característico da matriz *S,* associada com o valor $\hat{\Lambda}$.

Para determinar as componentes principais, a partir da matriz *S,* procede-se da seguinte forma:

a) Resolve-se a seguinte equação característica para obter a solução:

$\left| S - \hat{\Lambda} I \right| = 0$**,** isto é,

$\left| S - \hat{\Lambda} I \right| = 0$.

Conforme Regazzi (2001), “se o posto de *S* é igual a *p*, a equação $\left| S - \hat{\Lambda} I \right| = 0$ terá *p* raízes, chamadas de autovalores, ou raízes características da matriz *S”.*

Sejam $\hat{\Lambda}_1, \hat{\Lambda}_2, ..., \hat{\Lambda}_p$ as *p* soluções, temos que a cada autovalor $\hat{\Lambda}_i$ corresponde um autovetor característico.

$\vec{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ . \\ . \\ x_{ip} \end{bmatrix}$ com $\sum_{j=1}^{p} x_{ij}^2 = 1 \quad (\vec{x}_i^t . \vec{x}_i = 1)$, sendo esta a condição de normalidade.

e $\sum_{j=1}^{p} x_{ij} x_{kj} = 0$ para i ≠ k $(\vec{x}_i^t . \vec{x}_k = 0 \text{ para i} \neq k)$, sendo esta a condição de ortogonalidade dos vetores.

A normalidade é a primeira restrição feita para que o sistema tenha solução única, e a segunda restrição é a ortogonalidade, que garante que as componentes principais são independentes.

Isso significa dizer que cada autovetor é normalizado, ou seja, a soma dos quadrados dos coeficientes é igual a 1, sendo, ainda, ortogonais entre si.

b) Para cada autovalor $\hat{\Lambda}_i$ determina-se o autovetor normalizado $\vec{x}_i$, a partir da solução do sistema de equações dado a seguir:

$$\left| S - \hat{\Lambda} I \right| \vec{x}_i = 0$$

$\vec{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ . \\ . \\ x_{ip} \end{bmatrix}$, é um autovetor não normalizado.

$\vec{o}$ é um vetor nulo, de dimensão p×1.

O autovetor normalizado é dado por:

$$\vec{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ . \\ . \\ x_{ip} \end{bmatrix} = \frac{1}{\sqrt{x_{i1}^2 + x_{i2}^2 + ... + x_{ip}^2}} \begin{bmatrix} x_{i1} \\ x_{i2} \\ . \\ . \\ x_{ip} \end{bmatrix} = \frac{\vec{x}_i}{\left\| \vec{x}_i \right\|} \quad \vec{x}_i^t \vec{x}_i = 1 . \tag{3.16}$$

Conforme Regazzi (2001), tomando os elementos do vetor $\vec{x}_i$, assim determinados como os coeficientes de Y_i, tem-se que o i-ésimo componente principal é dado por:

$$Y_i = x_{i1}X_1 + \cdots + x_{i2}X_2 + \ldots.. + x_{ip}X_p.$$

Tem-se, ainda:

i) $\hat{Var}(Y_i) = \hat{\Lambda}_i$ logo $\hat{Var}(Y_1) > \hat{Var}(Y_2) > \ldots\ldots \hat{Var}(Y_p)$;

ii) $\sum \hat{Var}(X_i) = \sum \hat{\Lambda}_i = \sum \hat{Var}(Y_i)$;

iii) $\hat{Cov}(Y_i, Y_j) = 0, \ desde \ que \sum_{j=i}^{p} x_{ij}x_{kj} = 0$.

Deve-se observar que, nesta metodologia, a contribuição de cada componente principal Y_i é medida em termos de variância. Logo, tem-se que o quociente é expresso em percentagem:

$$\frac{\hat{Var}(Y_i)}{\sum_{i=1}^{p} \hat{Var}(Y_i)}.100 = \frac{\hat{\Lambda}_i}{\sum_{i=1}^{p} \hat{\Lambda}_i}.100 = \frac{\hat{\Lambda}_i}{traço(S)}.100, \qquad (3.17)$$

sendo que esta expressão representa a proporção da variância total explicada pela componente Y_i.

Ao se estudar um conjunto de *n* observações de *p-variáveis*, é possível encontrar novas variáveis denominadas de $\hat{Y}_k$, *k = 1, ..., p,* que são combinações lineares (*CL)* das variáveis originais X_p, não correlacionados, e apresentam um grau de variabilidade diferente umas das outras, também apresentados em ordem decrescente de valores. É importante lembrar que, em componentes principais, a unidade de medida são combinações lineares não correlacionadas, por isso são de difícil interpretação, e também é por esse motivo que as variáveis originais devem estar na mesma unidade de medida.

A soma dos *k* autovalores, dividida pela soma de todos os *p* autovalores $(\hat{\Lambda}_1 + \ldots + \hat{\Lambda}_k)/(\hat{\Lambda}_1 + \ldots. + \hat{\Lambda}_p)$, representa a proporção total explicada pelos primeiros *k* componentes principais. Isto é, a proporção da informação retida na redução de *p* para *k* dimensões. Com isso, pode-se decidir quantos componentes principais serão utilizados no estudo para diferenciar os indivíduos.

Portanto, para se fazer uma interpretação correta de quais componentes utilizar no estudo, basta selecionar as primeiras componentes que acumulam uma percentagem de variância explicada, igual ou superior a 70%. Ou seja, fica-se com $Y_1,...,Y_k$ tal que:

$$\frac{\hat{Var}(Y_1) + ... + \hat{Var}(Y_k)}{\sum_{i=1}^{p} \hat{Var}(Y_i)}.100 \geq 70\% \text{ no qual k < p.} \qquad (3.18)$$

O sucesso da metodologia é medido pelo valor de *k*. Se *k* = 1, dire-se-á que o método está reduzindo ao máximo, à dimensão inicial. Nesse caso, pode-se comparar os indivíduos em uma escala linear. Se *k* = 2, é possível localizar cada indivíduo em um plano cartesiano, sendo que os dois eixos representam as duas componentes. Se *k* for maior do que dois, a comparação dos indivíduos passa a ser mais complicada (REGAZZI, 2001).

A partir da matriz *S* é possível encontrar os valores $\hat{\Lambda}_1 \geq \hat{\Lambda}_2 \geq ... \geq \hat{\Lambda}_p \geq 0$, que são as raízes características, todas distintas e apresentadas em ordem decrescente de valores e, como *S* é positiva definida, todos os autovalores são não negativos.

Os eixos principais são os autovetores das matrizes *SI* ou *RI*, sendo que são os autovetores que fornecem a direção dos eixos na análise.

A Figura 17 mostra a elipse que possui dois eixos perpendiculares, cujas coordenadas estão representadas pelos autovetores I e II da matriz *S,* ou da matriz *R.* Os elementos desses vetores definem sua posição, isto é, o ângulo que eles formam com os eixos originais de Y_1 e Y_2. O comprimento desses vetores são os autovalores correspondentes a $\hat{\Lambda}$ dessa matriz, que representa a variância dos novos eixos (VALENTIN, 2000).

A Figura 17 é a representação gráfica dos autovalores e autovetores.

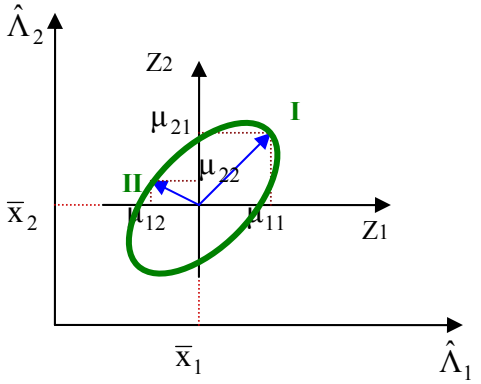


Figura 17 - Representação gráfica dos autovalores e autovetores.
Fonte : Valentin 2000.

Os eixos fatoriais *CP são* definidos pela direção e comprimento, através da seguinte equação característica: $\left|S - \hat{\Lambda}I\right| = 0$

S = matriz de variância-covariância, ou *R* a matriz de correlação.

$\hat{\Lambda}$ = autovalor de *S,* ou *R*.

I = matriz identidade.

Mostra-se, a seguir, um exemplo numérico para o cálculo dos autovalores e autovetores, utilizando-se os dados da Tabela 09.

Seja *S* a matriz de variância e covariância amostral, dada por:

$$S = \begin{bmatrix} 0{,}69 & 0{,}45 \\ 0{,}45 & 0{,}48 \end{bmatrix},$$

para encontrar os autovalores e autovetores, deve-se partir da seguinte equação característica:

$$\left|S - \hat{\Lambda}I\right| = 0.$$

Substituindo-se essa equação pelas matrizes *S* e *I,* obtém-se a seguinte expressão:

$$\left| \begin{bmatrix} 0{,}69 & 0{,}45 \\ 0{,}45 & 0{,}48 \end{bmatrix} - \hat{\Lambda} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0.$$

Multiplicando-se o autovalor $\hat{\Lambda}$ à matriz identidade, obtém-se as seguintes matrizes:

$$\left| \begin{bmatrix} 0{,}69 & 0{,}45 \\ 0{,}45 & 0{,}48 \end{bmatrix} - \begin{bmatrix} \hat{\Lambda} & 0 \\ 0 & \hat{\Lambda} \end{bmatrix} \right| = 0.$$

Realizando-se a subtração entre as matrizes, obtém-se a matriz:

$$\left\| \begin{bmatrix} 0,69 - \hat{\Lambda} & 0,45 \\ 0,45 & 0,48 - \hat{\Lambda} \end{bmatrix} \right\| = 0.$$

Resolvendo-se o determinante dessa matriz, encontra-se o seguinte resultado:

$$(0,69 - \hat{\Lambda})(0,48 - \hat{\Lambda}) - (0,45)^2 = 0.$$

Unindo-se os termos semelhantes, encontra-se uma equação do segundo grau:

$$0,33 - 0,69\hat{\Lambda} - 0,48\hat{\Lambda} + \hat{\Lambda}^2 - 0,20 = 0.$$

Resolvendo-se essa equação, encontra-se os autovalores correspondentes à matriz *S*.

$$\hat{\Lambda}^2 - 1,17\hat{\Lambda} + 0,13 = 0.$$

Os autovalores (raízes características) são obtidos da seguinte equação:

$\hat{\Lambda} = \dfrac{1,17 \pm \sqrt{(-1,17)^2 - (4)(1)(0,13)}}{(2)(1)}$, logo, os dois autovalores resultantes da equação são: $\hat{\Lambda}_1$ = 1,05 e $\hat{\Lambda}_2$ = 0,13.

Após encontrado os autovalores, passa-se a calcular os autovetores, correspondentes à matriz *S*. Na expressão que segue, $\vec{x}_1$ é um autovetor que será associado ao autovalor $\hat{\Lambda}_1$.

$S\vec{X} = \hat{\Lambda}\vec{X}$, para $\hat{\Lambda}_1$ = 1,05.

Substituindo-se os valores da expressão pelos seus respectivos dados tem-se:

$$\begin{bmatrix} 0,69 & 0,45 \\ 0,45 & 0,48 \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} = 1,05 \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}.$$

Realizando a multiplicação da matriz *S* com o autovetor $\vec{x}$ e o autovalor $\hat{\Lambda}_1$, obtem-se o seguinte sistema linear:

$$\begin{cases} 0,69x_{11} + 0,45x_{12} = 1,05x_{11} \\ 0,45x_{11} + 0,48x_{12} = 1,05x_{12} \end{cases}.$$

Unindo-se os termos semelhantes no sistema, obtem-se o seguinte:

$$\begin{cases} -0,36x_{11} + 0,45x_{12} = 0 \\ 0,45x_{11} - 0,57x_{12} = 0 \end{cases}$$

Resolvendo o sistema, foi possível calcular os dois autovetores associados ao autovalor $\hat{\Lambda}_1$, no qual $x_{11} = 1$ e $x_{12} = 0{,}8$ e o (autovetor $\neq$ 0), logo o autovetor associado ao autovalor 1,05 é:

$$\vec{x}_1 = \begin{bmatrix} 1 \\ 0{,}8 \end{bmatrix}.$$

Para obter os autovetores associados ao autovalor $\hat{\Lambda}_2$ = 0,13, faz-se os cálculos de forma análoga ao autovalor $\hat{\Lambda}_1$:

$S\vec{X} = \hat{\Lambda}\vec{X}$, para $\hat{\Lambda}_2$ = 0,13.

Substituindo-se os valores da expressão pelos seus respectivos dados tem-se:

$$\begin{bmatrix} 0{,}69 & 0{,}45 \\ 0{,}45 & 0{,}48 \end{bmatrix}\begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} = 0{,}13\begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix}.$$

Realizando a multiplicação da matriz *S* com o autovetor $\vec{x}_2$ e o autovalor $\hat{\Lambda}_2$, obtem-se o seguinte sistema linear:

$$\begin{cases} 0{,}69x_{21} + 0{,}45x_{22} = 0{,}13x_{21} \\ 0{,}45x_{21} + 0{,}48x_{22} = 0{,}13x_{22} \end{cases}.$$

Unindo-se os termos semelhantes no sistema, obtem-se o seguinte:

$$\begin{cases} 0{,}56x_{21} + 0{,}45x_{22} = 0 \\ 0{,}32x_{21} + 0{,}48x_{22} = 0 \end{cases}.$$

Resolvendo o sistema, foi possível calcular os dois autovetores associados ao autovalor $\hat{\Lambda}_2$, no qual $x_{21} = 1$ e $x_{22} = -1{,}25$ e o (autovetor $\neq$ 0), logo o autovetor associado ao autovalor 0,13 é:

$$\vec{x}_2 = \begin{bmatrix} 1 \\ -1{,}25 \end{bmatrix}.$$

Ao realizar uma análise de componentes principais, é muito importante saber o significado de cada componente no estudo que está sendo realizado.

A interpretação de uma componente principal é feita mediante o grau de importância, ou, ainda, a influência que cada variável tem sobre cada componente, sendo que esta importância é dada pela correlação entre cada variável X_j e o componente Y_i que estiver sendo interpretado (REGAZZI, 2001).

Dessa forma, para a componente Y_1 tem-se que:

$$\mathrm{Corr}(X_j, Y_1) = r_{X_j Y_1} = x_{1j}\frac{\sqrt{\hat{\mathrm{Var}}(Y_1)}}{\sqrt{\hat{\mathrm{Var}}(X_j)}} = \sqrt{\hat{\Lambda}_1}\,\frac{x_{1j}}{\sqrt{\hat{\mathrm{Var}}(X_j)}}, \tag{3.19}$$

logo, para se comparar a importância de $X_1, X_2, ..., X_p$ sobre Y_1, basta fazer:

$$\frac{x_{11}}{\sqrt{\hat{Var}(X_1)}}, \frac{x_{12}}{\sqrt{\hat{Var}(X_2)}}, \ldots\ldots\ldots, \frac{x_{1p}}{\sqrt{\hat{Var}(X_p)}} \tag{3.20}$$

e, assim, com todas as componentes em estudo.

A Tabela 11 mostra um resumo da análise de componentes principais, quais são os componentes principais, seus autovalores, seus autovetores, a correlação das variáveis, a percentagem de variância, explicada por cada componente, e a percentagem total da variância acumulada pelos componentes principais.

Tabela 11 – Componentes principais obtidas da análise de *p* variáveis $X_1, X_2, ..., X_p$.

	Variância explicada pelos	Coeficientes de ponderação associados às variáveis	Correlação entre X_j e Y_i	Percentagem da variância de Y_i	Percentagem acumulada da variância dos Y_i
Componentes Principais	Autovalores $\hat{\Lambda}_i$	$X_1 \quad X_2 \quad .. X_p$	$X_1 \qquad X_2 \qquad ..... X_P$		
Y_1	$\hat{\Lambda}_1$	$x_{11} \; x_{12} .. x_{1p}$	$\sqrt{\hat{\Lambda}_1}\frac{x_{11}}{s_1} \; \sqrt{\hat{\Lambda}_1}\frac{x_{12}}{s_2} .. \sqrt{\hat{\Lambda}_1}\frac{x_{1p}}{s_p}$	$\left(\hat{\Lambda}_1 / \sum_{i=1}^{p}\hat{\Lambda}_i\right).100$	$\left(\hat{\Lambda}_1 / \sum_{i=1}^{p}\hat{\Lambda}_i\right).100$
Y_2	$\hat{\Lambda}_2$	$x_{21} \; x_{22} .. x_{2p}$	$\sqrt{\hat{\Lambda}_2}\frac{x_{21}}{s_1} \; \sqrt{\hat{\Lambda}_2}\frac{x_{22}}{s_2} .. \sqrt{\hat{\Lambda}_2}\frac{x_{2p}}{s_p}$	$\left(\hat{\Lambda}_2 / \sum_{i=1}^{p}\hat{\Lambda}_i\right).100$	$\left(\hat{\Lambda}_1 + \hat{\Lambda}_2 / \sum_{i=1}^{p}\hat{\Lambda}_i\right).100$
.	.	. . .	. . .	.	.
.	.	. . .	. . .	.	.
.	.	. . .	. . .	.	.
.	.	. . .	. . .	.	.
Y_p	$\hat{\Lambda}_p$	$x_{p1} \; x_{p2} .. x_{pp}$	$\sqrt{\hat{\Lambda}_p}\frac{x_{p1}}{s_1} \; \sqrt{\hat{\Lambda}_p}\frac{x_{p2}}{s_2} .. \sqrt{\hat{\Lambda}_p}\frac{x_{pp}}{s_p}$	$\left(\hat{\Lambda}_p / \sum_{i=1}^{p}\hat{\Lambda}_i\right).100$	$\left(\hat{\Lambda}_1 + \hat{\Lambda}_2 + ... + \hat{\Lambda}_p / \sum_{i=1}^{p}\hat{\Lambda}_i\right).100$

Fonte: Regazzi (2001)

Se o objetivo da análise for comparar os indivíduos, ou agrupá-los, deve-se calcular, para cada indivíduo, os seus valores (escores), para cada componente principal, que será utilizado na análise. Isso equivale a substituir a matriz de dados originais de dimensão $n_x p$ por outra matriz $n_x k$, sendo que *k* é o número de componentes principais selecionados (REGAZZI, 2001).

A Tabela 12 ilustra a substituição da matriz de dados originais (variáveis) por uma nova matriz, gerada após a análise, das componentes principais (escores para os componentes).

Tabela 12 – Escores relativos a *n* objetos (indivíduos), obtidos em relação aos *k* primeiros componentes principais.

Objetos (indivíduos)	Variáveis	Escores para os componentes
	$X_1 \quad X_2 \quad ..... X_p$	$Y_1 \quad Y_2 \quad ..... Y_k$
1	$x_{11} \quad x_{12} .... x_{1p}$	$y_{11} \quad y_{12} .... y_{1k}$
2	$x_{21} \quad x_{22} .... x_{2p}$	$y_{21} \quad y_{22} .... y_{2k}$
.	. . .	. . .
.	. . .	. . .
.	. . .	. . .
n	$x_{n1} \quad x_{n2} .... x_{np}$	$y_{n1} \quad y_{n2} .... y_{nk}$

Fonte: Regazzi (2001)

Para obter as *CP* é necessário formar as combinações lineares das variáveis originais. Para formar essas *CP* utiliza-se o seguinte procedimento:

$$Y_{11} = x_{11}X_{11} + x_{12}X_{12} + ... + x_{1p}X_{1p}$$

$$Y_{21} = x_{11}X_{21} + x_{12}X_{22} + ... + x_{1p}X_{2p}$$

. . . .

. . . .

. . . .

$$Y_{n1} = x_{11}X_{n1} + x_{12}X_{n2} + ... + x_{1p}X_{np}$$

Assim, faz-se, sucessivamente, até encontrar todos os componentes da análise.

Os componentes são combinações lineares não correlacionados de $Y_1, Y_2, ..., Y_p$, cuja variância é a maior possível.

Na prática, se forem utilizados os dados da Tabela 09, as componentes serão representadas da seguinte forma:

$$Y_1 = (autovetor\ x_{11})(\text{var}\,iável\ X) + (autovetor\ x_{12})(\text{var}\,iável\ Y)$$

$$Y_{11} = 1.10,0 + 0,8.10,7 = 18,56$$

$$Y_{12} = 1.10,4 + 0,8.9,8 = 18,24$$

$$Y_{13} = 1.9,7 + 0,8.10,0 = 17,7$$

$$Y_{14} = 1.9,7 + 0,8.10,1 = 17,78$$

$$Y_{15} = 1.11,7 + 0,8.11,5 = 20,9$$

$$Y_2 = (autovetor\ x_{21})(\text{var}\,iável\ X) + (autovetor\ x_{22})(\text{var}\,iável\ Y)$$

$$Y_{21} = 1.10,0 - 1,25.10,7 = -3,38$$

$$Y_{22} = 1.10,4 - 1,25.9,8 = -1,85$$

$$Y_{23} = 1.9,7 - 1,25.10,0 = -2,8$$

$$Y_{24} = 1.9,7 - 1,25.10,1 = -2,93$$

$$Y_{25} = 1.11,7 - 1,25.11,5 = -2,68$$

Dessa forma, encontrara-se as duas componentes referentes à Tabela 09. Como pode-se verificar, acima, em um número reduzido de combinações lineares é possível sintetizar a maior parte da informação contida nos dados originais.

Caso seja necessário padronizar as variáveis, utiliza-se a expressão do item 2.1.

A Tabela 13 mostra um exemplo das variáveis padronizadas.

Tabela 13 – Matriz de variáveis padronizados de ***n*** indivíduos e ***p*** variáveis.

Indivíduos	Variáveis							
	Z_1	Z_2	Z_3	Z_4	...	Z_j	...	Z_p
1	Z_{11}	Z_{12}	Z_{13}	Z_{14}	...	Z_{1j}	...	Z_{1p}
2	Z_{21}	Z_{22}	Z_{23}	$Z24$	...	Z_{2j}	...	Z_{2p}
3	Z_{31}	Z_{32}	Z_{33}	Z_{34}	...	Z_{3j}	...	Z_{3p}
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
i	Z_{i1}	Z_{i2}	Z_{i3}	Z_{i4}	...	Z_{ij}	.	Z_{ip}
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
n	Z_{n1}	Z_{n2}	Z_{n3}	Z_{n4}	...	Z_{nj}	...	Z_{np}

Fonte: Regazzi 2001

Pode-se afirmar que a matriz *R* das variáveis X_j é igual à matriz *S* das variáveis padronizadas Z_j.

Desta forma, utilizando os dados padronizados garante-se que todas as variáveis tenham o mesmo grau de importância, portanto, trabalha-se com o conjunto de dados padronizados. Neste caso, faz-se necessário estimar a matriz *R* para se calcular os autovalores e autovetores que darão origem às componentes principais, cujo procedimento para a estimação dos autovalores e autovetores será o mesmo mostrado anteriormente, apenas substituindo *S* por *R.* Os autovetores passarão a ser denominados de $\hat{e}_p$, pois esta nova representação indica que o conjunto amostral dos dados foi padronizado. Logo, os pares de autovalores e autovetores estimados da amostra analisada serão representados por $(\hat{\Lambda}_1, \hat{e}_1)$, $(\hat{\Lambda}_2, \hat{e}_2)$, ... ,$(\hat{\Lambda}_p, \hat{e}_p)$; onde $\hat{\Lambda}_1 \geq \hat{\Lambda}_2 \geq ... \geq \hat{\Lambda}_p \geq 0$; e fornecerão as novas combinações lineares (JOHNSON & WICHERN, 1992) expressas por $Y_1 = x_1' X,\ Y_2 = x_2' X ,...,\ Y_p = x_p' X$ os *CP* então:

$$S_{11}^2 + S_{22}^2 + ... + S_{pp}^2 = \sum_{i=1}^{p} Var(X_i) = \hat{\Lambda}_1 + \hat{\Lambda}_2 + ... + \hat{\Lambda}_p = \sum_{i=1}^{p} Var(Y_i)$$

$$S_{11}^2 + S_{22}^2 + ... + S_{pp}^2 = tr(S)$$

Já a proporção explicada pelo k – ésimo componente principal é dada pela expressão:

$$\frac{\hat{\Lambda}_1}{\hat{\Lambda}_1 + \hat{\Lambda}_2 + ... + \hat{\Lambda}_p} \quad k = 1, 2, ..., p$$

Ao utilizar-se a matriz *R* ao invés da matriz *S* para a extração das componentes principais, a soma da diagonal principal da matriz *R*

corresponderá ao número total de variáveis que representa a variabilidade total do sistema padronizado, conforme mostra a relação a seguir:

$$tr\,R = p$$

Como se pode verificar, o traço da matriz *R* será igual ao número de variáveis que estão envolvidas na formação das componentes principais, e a proporção da explicação fornecido pela *j-ésima* componente será dada por:

$$\frac{\hat{\Lambda}_j}{tr\,R}$$

pois, ao se utilizar a matriz *R*, teremos na sua diagonal principal somente elementos unitários, facilitando a determinação da proporção de variância explicada de cada componente.

As combinações lineares obtidas através das *CP´s,* segundo JACKSON (1980), possuem a característica de que nenhuma combinação linear das variáveis originais irá explicar mais que a primeira componente e, sempre que se trabalhar com a matriz de correlação, as variáveis não sofrerão influência da magnitude de suas unidades medidas.

Resolvendo a matriz de correlação, pode-se observar se existe correlação entre as variáveis; se algumas variáveis iniciais forem linearmente dependentes umas das outras, alguns dos valores próprios serão nulos na matriz de correlação. Neste caso, a variação total poderá ser explicada pelas primeiras componentes principais.

É difícil encontrar em um problema a existência de dependência linear exata, a menos que esta seja introduzida propositalmente nas variáveis redundantes. Na *ACP* pode ocorrer a dependência linear aproximada entre algumas variáveis. Neste caso, os valores próprios menores são muito próximos de zero e a sua contribuição para explicar a variância será muito pequena (REIS, 1997). Por isso, deve-se retirar da análise aquelas componentes que possuem pouca informação, isso não implica em uma perda significativa de informação.

Com isso, pode-se reduzir os dados e tornar os resultados mais fáceis de serem interpretados. Dentre vários critérios que excluem componentes que possuem pouca informação, cita-se estes:

A definição do número de componentes a serem utilizadas é feita por meio de dois critérios. O primeiro, denominado de método gráfico, representa graficamente a porcentagem de variação explicada pela componente nas ordenadas e os autovalores em ordem decrescente nas abscissas. Quando esta percentagem diminui e a curva passa a ser praticamente paralela ao eixo das abscissas, exclui-se as componentes que restam, pois possuem pouca informação. Este critério, que considera as componentes anteriores ao ponto de inflexão da curva, foi sugerido por CATTEL (1966) e exemplificado por PLA (1986), que considera quatro situações distintas, conforme mostra Tabela 14.

Tabela 14 – Variação explicada pela componente.

Situações	Percentual da variação total explicada pela componente					
	CP_1	CP_2	CP_3	CP_4	CP_5	Total
Caso 1	35	30	28	4	3	100
Caso 2	45	30	9	8	8	100
Caso 3	75	7	7	6	5	100
Caso 4	22	21	20	19	18	100

Na Figura 18 a seguir, visualiza-se melhor a seleção dos componentes principais através do método gráfico.

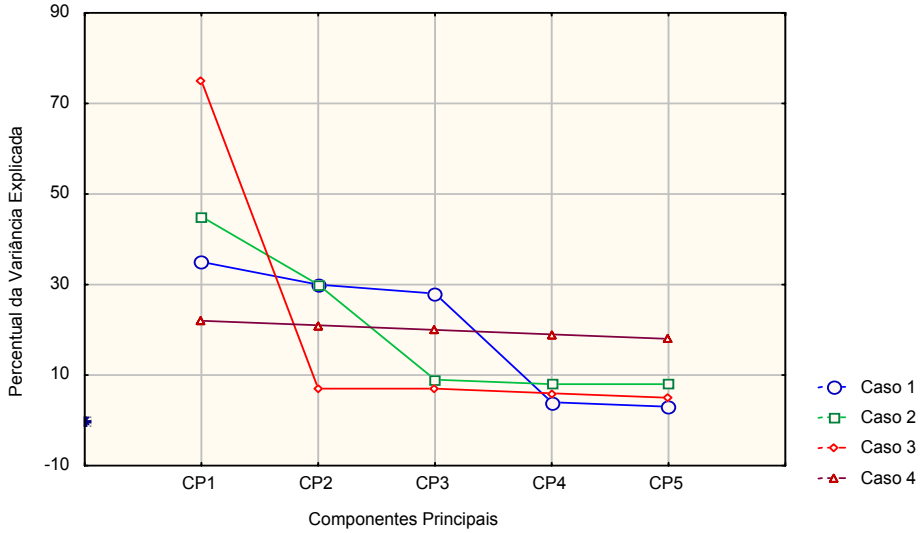


Figura 18 - Proporção da variação explicada pela componente. Exemplo retirado de Analisis multivariado: método de componentes principales; PLA (1986).

No caso 1, as três primeiras componentes explicam 93% da variância total, havendo uma quebra brusca depois da quarta componente, sendo consideradas as três primeiras. No caso 2, as duas primeiras componentes explicam 75% da variabilidade total e a quebra brusca, neste caso, ocorre na terceira componente, considerando-se as duas primeiras. Este mesmo procedimento ocorre para os demais casos, podendo-se observar, também, que as outras componentes apresentam uma baixa explicação.

O segundo critério de seleção consiste em incluir somente aquelas componentes cujos valores próprios sejam superiores a 1. Este critério é sugerido por KAISER (1960) apud MARDIA (1979). Ele tende a incluir poucas componentes quando o número de variáveis originais é inferior a vinte e, em geral, utilizam-se aquelas componentes que conseguem sintetizar uma variância acumulada em torno de 70%.

Além do uso na redução da dimensionalidade, a técnica de *ACP* pode ser utilizada como apoio à busca da variável de maior prevalência no sistema responsável, servindo-se do estudo dos coeficientes de correlação entre as componentes e as variáveis originais.

Quando se fala em avaliar a estabilidade de um processo produtivo, as dificuldades que porventura existam devem-se à complexidade do processo e não aos métodos multivariados. A *ACP* é um recurso adicional de apoio para verificar a estabilidade do sistema (TELHADA, 1995). O problema existente em um conjunto multivariado é que, às vezes, uma observação pode não ser extrema para uma determinada variável, mas pode ser considerada uma observação extrema por não ser semelhante à estrutura de correlação fornecida pelo restante dos dados.

A equação $r_{\hat{Y}_i, X_k} = \sqrt{\hat{\Lambda}_i}\,\dfrac{\hat{e}_{ki}}{\sqrt{s_{kk}}}$ deve ser utilizada quando os autovetores são derivados da matriz de variância *S*, e a equação $r_{\hat{Y}_i, Z_k} = \hat{e}_{ki}\sqrt{\hat{\Lambda}_i}$ quando os autovetores são derivados da matriz de correlação *R*.

> Quando duas ou mais componentes apresentam-se fora dos limites de controle, deve-se estabelecer uma ordem hierárquica entre as componentes principais para auxiliar na solução de conflitos quanto à variável de maior influência sobre a perda de controle. Pois, neste caso, pode-se ficar em dúvida quanto a dar mais atenção a uma componente em detrimento da outra. Deve-se, então, levar em consideração o maior autovalor que originou a componente, optando-se por esta (SOUZA, 2000, p.30 a 35).

3.3 Aplicação da análise de componentes principais, exemplos práticos

Neste item serão desenvolvidos dois exemplos práticos, utilizando-se no ex. 1 para o cálculo da matriz *S*, e no exemplo 2 a matriz *R*.

Exemplo 1:

Considere os dados da Tabela 15, referentes a duas variáveis X_1 e X_2, sendo estas mensuradas em uma amostra constituída de cinco observações (indivíduos). Os componentes principais serão calculados a partir da matriz de variância-covariância.

Tabela 15 – Observações relativas a duas variáveis, avaliadas em cinco indivíduos.

Observações	(Variável) X_1	(Variável) X_2
1	100	76
2	93	82
3	102	81
4	95	68
5	90	62

Realizando uma estatística descritiva nas duas variáveis, tem-se os seguintes resultados na Tabela 16:

Tabela 16 – Estatística descritiva relativa a duas variáveis, avaliadas em cinco indivíduos.

	Variável X_1	Variável X_2
Média aritmética das variáveis	96	73,8
Somatório ao quadrado das variáveis	46178	27529
Somatório das variáveis	480	369
Variância amostral das variáveis	24,5	74,2
Desvio padrão amostral das variáveis	4,95	8,61

A matriz *S* é estimada pela expressão do item 3.8, e a covariância entre as variáveis pela equação do item 3.10, conforme segue o exemplo:

$$\hat{Cov}(x_1, x_2) = \frac{1}{5-1}\left[35528 - \frac{480.369}{5}\right]$$

$$\hat{Cov}(x_1, x_2) = \frac{1}{4}\left[35528 - 35424\right]$$

$$\hat{Cov}(x_1, x_2) = 26,$$

logo, a matriz *S* é assim constituída:

$$S = \begin{bmatrix} 24{,}5 & 26 \\ 26 & 74{,}2 \end{bmatrix}.$$

Para encontrar os autovalores, deve-se partir da equação característica abaixo, utilizando a matriz *S*:

$$\left|S - \hat{\Lambda} I\right| = 0.$$

Substituindo-se essa equação pelas matrizes *S* e *I*, obtém-se a seguinte expressão:

$$\left|\begin{bmatrix} 24{,}5 & 26 \\ 26 & 74{,}2 \end{bmatrix} - \hat{\Lambda}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right| = 0.$$

Multiplicando-se o autovalor $\hat{\Lambda}$ à matriz *I,* obtém-se as seguintes matrizes:

$$\left|\begin{bmatrix} 24{,}5 & 26 \\ 26 & 74{,}2 \end{bmatrix} - \begin{bmatrix} \hat{\Lambda} & 0 \\ 0 & \hat{\Lambda} \end{bmatrix}\right| = 0.$$

Realizando-se a subtração entre as matrizes, obtém-se a matriz:

$$\begin{vmatrix} 24{,}5 - \hat{\Lambda} & 26 \\ 26 & 74{,}2 - \hat{\Lambda} \end{vmatrix} = 0.$$

Resolvendo o determinante dessa matriz, encontra-se o seguinte resultado:

$$(24{,}5 - \hat{\Lambda})(74{,}2 - \hat{\Lambda}) - (26)^2 = 0.$$

Unindo-se os termos semelhantes, encontra-se uma equação do segundo grau:

$$1817{,}9 - 24{,}5\hat{\Lambda} - 74{,}2\hat{\Lambda} + \hat{\Lambda}^2 - 676 = 0.$$

Resolvendo essa equação, encontra-se os autovalores correspondentes à matriz *S*.

$$\hat{\Lambda}^2 - 98{,}7\hat{\Lambda} + 1141{,}9 = 0.$$

Os autovalores (raízes características) são obtidos da seguinte equação:

$$\hat{\Lambda} = \frac{-b \pm \sqrt{(-b)^2 - 4(a)(c)}}{2(a)}$$

$\hat{\Lambda} = \frac{98,7 \pm \sqrt{(-98,7)^2 - 4(1)(1141,9)}}{(2)(1)}$, logo, os dois autovalores resultantes da equação são: $\hat{\Lambda}_1$ = 85,32 e $\hat{\Lambda}_2$ = 13,38.

Como pode-se observar, a soma dos autovalores corresponde ao traço e ao determinante da matriz *S*.

$\hat{\Lambda}_1 + \hat{\Lambda}_2 + ... + \hat{\Lambda}_p =$ traço da matriz *S*. Ou seja,

13,38 + 85,32 = 98,7 = traço da matriz *S.*

$(\hat{\Lambda}_1).(\hat{\Lambda}_2)\ ....(\hat{\Lambda}_p) =$ determinante da matriz *S*.

(13,38).(85,32) = 1141.6.

Se se resolver a seguinte expressão $\frac{\hat{\Lambda}_1}{traço\,S}.100$, será obtida a proporção da variância total, explicada por cada componente principal. Observa-se que a primeira componente explica $\frac{85,32}{98,7}.100 = 86,44\%$, e a segunda componente explica $\frac{13,38}{98,7}.100 = 13,56\%$.

Ou seja, a primeira componente relativa à raiz $\hat{\Lambda}_1$, explica 86,44% da variação total dos dados.

Já a segunda componente, relativa à raiz $\hat{\Lambda}_2$, explica 13,56% da variação total dos dados.

Essa variância será distribuída entre $\hat{\Lambda}_1$ = 85,32 e $\hat{\Lambda}_2$ = 13,38, ou seja, 86,44% da variância é explicada pelo primeiro eixo fatorial, e 13,56% pelo segundo.

Como pode-se observar, acima, cada componente principal sintetiza a máxima proporção de variância contida nos dados.

Deve-se observar, também, que a adição de duas raízes características dá 98,7, que nada mais é que o segundo termo da equação.

O cálculo da primeira componente referente, a $\hat{\Lambda}_1$ = 85,32, será dado pelo autovetor associado a $\hat{\Lambda}_1$, sendo que a equação característica dos autovetores é $\left|S - \hat{\Lambda}_1 I\right|\vec{X}_1 =$ 0. Existe um vetor $\vec{x}$ para cada valor de $\hat{\Lambda}$.

As coordenadas de $x_{11}\ e\ x_{12}$ do autovetor $\vec{X}_1$ são calculadas pela equação matricial:

$\left|S - \hat{\Lambda}_1 I\right|\vec{X}_1 = 0.$

Substituindo-se essa equação pelas matrizes *S*, *I*, pelo primeiro autovalor $\hat{\Lambda}_1$ = 85,32 e pela matriz de incógnitas, obtém-se a seguinte expressão:

$$\left\|\begin{bmatrix} 24{,}5 & 26 \\ 26 & 74{,}2 \end{bmatrix} - 85{,}32\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right\|\begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Multiplicando-se o autovalor $\hat{\Lambda}_1$ à matriz *I* e subtraindo da matriz *S,* obtém-se as seguintes matrizes:

$$\begin{bmatrix} 24{,}5 - 85{,}32 & 26 \\ 26 & 74{,}2 - 85{,}32 \end{bmatrix}\begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Multiplicando-se essas matrizes, encontra-se o seguinte sistema:

$$\begin{cases} -60{,}82x_{11} + 26x_{12} = 0 \\ 26x_{11} - 11{,}12x_{12} = 0 \end{cases}.$$

Esse sistema de equações é indeterminado, em virtude de $\left|S - \hat{\Lambda}I\right| = 0$

$$\begin{vmatrix} -60{,}82 & 26 \\ 26 & -11{,}12 \end{vmatrix} = 0,$$

ou, ainda, por $x_{11} = x_{12} = 0$, ou seja, o vetor passando pela origem.

Devido a isso, pode-se deixar uma das equações (neste caso a segunda), e atribuir um valor qualquer, que não seja nulo, a uma das incógnitas ($x_{12} = 1$). Dessa forma, tem-se:

- 60,82 x_{11} + 26.(1) = 0

- 60,82 x_{11} = -26, logo o valor da incógnita x_{11} será:

x_{11} = 0,43,

e o autovetor associado ao primeiro autovalor $\hat{\Lambda}_1$ = 85,32, será:

$\vec{x}_1 = \begin{bmatrix} 0{,}43 \\ 1 \end{bmatrix}$ e, sua norma será de:

$\left\|\vec{x}_1\right\| = \sqrt{(0{,}43)^2 + (1)^2} = 1{,}09.$

Para que esse vetor seja unitário, é necessário normalizar o autovetor a 1, da seguinte forma:

$$x_1 = \frac{1}{\left\|\vec{x}_1\right\|}.\vec{x}_1 .$$

Substituindo-se essa expressão pelos seus respectivos valores têm-se:

$$x_1 = \frac{1}{1{,}09}\begin{bmatrix} 0{,}43 \\ 1 \end{bmatrix},$$

logo, o primeiro autovetor normalizado será:

$$x_1 = \begin{bmatrix} 0{,}39 \\ 0{,}92 \end{bmatrix},$$

e a sua norma será:

$$\left\|x_1\right\| = \sqrt{(0{,}39)^2 + (0{,}92)^2} \text{ = 1.}$$

Como pode-se observar $x_1^t x_1 = 1$, sendo esta a primeira restrição feita por Morrison (1976), para que o sistema tenha solução única.

Logo, o primeiro componente principal será:

$$Y_1 = 0{,}39X_1 + 0{,}92X_2 .$$

O segundo componente principal é dado pela outra raiz $\hat{\Lambda}_2 = 13{,}38$:

$$\left|S - \hat{\Lambda}_2 I\right|\vec{X}_2 = 0.$$

Substituindo-se essa equação pelas matrizes *S*, *I*, pelo segundo autovetor $\hat{\Lambda}_2 = 13{,}38$, e pela matriz de incógnitas, obtém-se a seguinte expressão:

$$\begin{bmatrix} 24{,}5 & 26 \\ 26 & 74{,}2 \end{bmatrix} - 13{,}38\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Multiplicando-se o autovalor $\hat{\Lambda}_2$ à matriz *I* e subtraindo da matriz *S,* obtém-se as seguintes matrizes:

$$\begin{bmatrix} 24{,}5 - 13{,}38 & 26 \\ 26 & 74{,}2 - 13{,}38 \end{bmatrix}\begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Multiplicando-se essas matrizes, encontra-se o seguinte sistema:

$$\begin{cases} 11{,}12x_{21} + 26x_{22} = 0 \\ 26x_{21} + 60{,}82x_{22} = 0 \end{cases}.$$

Esse sistema de equações é indeterminado, em virtude de $\left|S - \hat{\Lambda}I\right| = 0$

$$\begin{vmatrix} 11{,}12 & 26 \\ 26 & 60{,}82 \end{vmatrix} = 0,$$

ou, ainda, por $x_{21} = x_{22} = 0$, ou seja, o vetor passando pela origem.

Devido a isso, pode-se deixar uma das equações (neste caso a segunda), e atribuir um valor qualquer, que não seja nulo, a uma das incógnitas ($x_{22} = 1$). Dessa forma, tem-se:

$11{,}12x_{21} + 26.(1) = 0$, logo a incógnita x_{21}, será:

$$x_{21} = -\frac{26}{11{,}12} = -2{,}34$$

e o autovetor, associado ao segundo autovalor $\hat{\Lambda}_2 = 13{,}38$, será:

$$x_2 = \begin{bmatrix} -2{,}34 \\ 1 \end{bmatrix},$$

e sua norma será de:

$\left\|x_2\right\| = \sqrt{(-2{,}34)^2 + (1)^2}$ = 2,54.

Para que esse vetor seja unitário, é necessário normalizar o autovetor a 1, da seguinte forma:

$$x_2 = \frac{1}{\left\|\vec{x}_2\right\|}\vec{x}_2 = \frac{1}{2{,}54}\begin{bmatrix} -2{,}34 \\ 1 \end{bmatrix},$$

logo, o segundo autovetor normalizado será:

$$x_2 = \begin{bmatrix} -0{,}92 \\ 0{,}39 \end{bmatrix},$$

e sua norma será de:

$\left\|x_2\right\| = \sqrt{(-0{,}92)^2 + (0{,}39)^2} = 1.$

Como pode-se observar, $x_2' x_2 = 1$ é a primeira restrição feita por Morrison (1976), para que o sistema tenha solução única (SOUZA, 2001).

Os elementos desses dois vetores de norma 1 são os cossenos-diretores dos ângulos que eles fazem com o sistema de origem.

Logo, a segunda componente principal será:

$Y_2 = -0{,}92X_1 + 0{,}39X_2$.

Outra restrição é que, nesse exemplo, os dois vetores são ortogonais, pois $x_1^t x_2 = 0$ (o produto escalar é igual à zero), que é a segunda restrição feita por Morrison (1976).

Para que esta restrição seja satisfeita, deve-se multiplicar o primeiro autovetor normalizado transposto pelo segundo autovetor normalizado, procedendo-se da seguinte forma:

$$x_1^t x_2 = \begin{bmatrix} 0,39 & 0,92 \end{bmatrix} \begin{bmatrix} -0,92 \\ 0,39 \end{bmatrix}.$$

Multiplicando-se os autovetores normalizados, têm-se a seguinte expressão:

$$x_1^t x_2 = (0,39)(-0,92) + (0,92)(0,39),$$

logo, têm-se que:

$$x_1^t x_2 = -0,36 + 0,36 = 0.$$

Conforme Regazzi (2001), “cada componente admite duas soluções, pois cada uma delas é obtida da outra pela multiplicação de seu segundo membro por (-1)”. Um exemplo disso pode ser a primeira componente principal:

$$Y_1 = 0,39X_1 + (0,92)(-1)X_2$$
$$Y_1 = 0,39X_1 - 0,92X_2.$$

O passo a seguir é realizado para encontrar o valor de cada componente principal, procede-se da seguinte forma:

$$Y_1 = 0,39X_1 + 0,92X_2$$
$$Y_{11} = 0,39(100) + 0,92(76) = 108,92$$
$$Y_{12} = 0,39(93) + 0,92(82) = 111,71$$
$$Y_{13} = 0,39(102) + 0,92(81) = 114,3$$
$$Y_{14} = 0,39(95) + 0,92(68) = 99,61$$
$$Y_{15} = 0,39(90) + 0,92(62) = 92,14$$
$$Y_2 = -0,92X_1 + 0,39X_2$$
$$Y_{21} = -0,92(100) + 0,39(76) = -62,36$$
$$Y_{22} = -0,92(93) + 0,39(82) = -53,58$$
$$Y_{23} = -0,92(102) + 0,39(81) = -62,25$$
$$Y_{24} = -0,92(95) + 0,39(68) = -60,88$$
$$Y_{25} = 0,92(90) - 0,39(62) = -58,62$$

Na Tabela 16 mostra-se as observações, e as variáveis originais utilizadas na análise e as novas componentes geradas a partir das combinações lineares, formadas na análise.

Tabela 17 – Mostra a substituição da matriz dos dados originais por uma nova matriz, gerada a partir das combinações lineares.

Observações	Variáveis originais		Novas variáveis geradas para as componentes principais	
	X_1	X_2	Y_1	Y_2
1	100	76	108,22	-62,36
2	93	82	111,71	-53,58
3	102	81	114,3	-62,25
4	95	68	99,61	-60,88
5	90	62	92,14	-58,62

Para completar a análise de componentes principais, é necessário fazer a correlação entre as variáveis X_j e Y_i, como se pode verificar a seguir:

$$r_{x_1y_1} = \sqrt{\hat{\Lambda}_1}.\frac{x_{11}}{\sqrt{Vâr(x_1)}}$$

$$r_{x_1y_1} = \sqrt{85,32}.\frac{0,39}{\sqrt{24,5}} = 0,73$$

$$r_{x_2y_1} = \sqrt{\hat{\Lambda}_1}.\frac{x_{12}}{\sqrt{Vâr(x_2)}}$$

$$r_{x_2y_1} = \sqrt{85,32}.\frac{0,92}{\sqrt{74,2}} = 0,99$$

$$r_{x_1y_2} = \sqrt{\hat{\Lambda}_2}.\frac{x_{21}}{\sqrt{Vâr(x_1)}}$$

$$r_{x_1y_2} = \sqrt{13,39}.\frac{-0,92}{\sqrt{24,5}} = -0,68$$

$$r_{x_2y_2} = \sqrt{\hat{\Lambda}_2}.\frac{x_{22}}{\sqrt{Vâr(x_2)}}$$

$$r_{x_2y_2} = \sqrt{13,39}.\frac{0,39}{\sqrt{74,2}} = 0,17.$$

A Tabela 18 mostra os componentes principais encontrados na análise, os autovalores, os autovetores, a correlação existente entre as variáveis, a percentagem de explicação de cada componente e a percentagem total de variância acumulada pelas componentes principais.

Tabela 18 – Resumo da análise de componentes principais.

Componentes principais	Autovalor	Coeficiente de ponderação associado às variáveis		Correlação entre X_j Y_i		Percentagem da variância de Y_i	Percentagem acumulada da variância dos Y_i
		X_1	X_2	X_1	X_2		
Y_1	85,32	0,39	0,92	0,73	0,99	86,44%	86,44%
Y_2	13,39	-0,92	0,39	-0,68	0,17	13,56%	100%

Como pode-se observar na Tabela 17, a componente Y_1 possui a maior correlação, sendo essa variável a de maior importância para o estudo.

Exemplo 2:

Considerando-se os dados do exemplo 01, referentes a duas variáveis X_1 e X_2, sendo estas mensuradas em uma amostra constituída de cinco observações (indivíduos), passa-se a desenvolver este exemplo, da Tabela 19, a partir da matriz de correlação.

Na Tabela 19 mostra-se as observações e as variáveis originais utilizadas na análise, e as variáveis padronizadas.

Tabela 19 – Observações relativas a duas variáveis, avaliadas em cinco indivíduos e com as respectivas variáveis padronizadas.

Observações	Variáveis originais		Variáveis padronizadas	
	X_1	X_2	Z_1	Z_2
1	100	76	0,81	0,26
2	93	82	-0,61	0,95
3	102	81	1,21	0,84
4	95	68	-0,20	-0,67
5	90	62	-1,21	-1,37

Para se obter as variáveis padronizadas, pode-se utilizar a expressão do item 2.1:

$$Z_{11} = \frac{100 - 96}{4{,}95} = 0{,}81 \qquad Z_{21} = \frac{76 - 73{,}8}{8{,}61} = 0{,}26$$

$$Z_{12} = \frac{93 - 96}{4{,}95} = -0{,}61 \qquad Z_{22} = \frac{82 - 73{,}8}{8{,}61} = 0{,}95$$

$$Z_{13} = \frac{102 - 96}{4{,}95} = 1{,}21 \qquad Z_{23} = \frac{81 - 73{,}8}{8{,}61} = 0{,}84$$

$$Z_{14} = \frac{95 - 96}{4{,}95} = -0{,}20 \qquad Z_{24} = \frac{68 - 73{,}8}{8{,}61} = -0{,}67$$

$$Z_{15} = \frac{90 - 96}{4{,}95} = -1{,}21 \qquad Z_{25} = \frac{62 - 73{,}8}{8{,}61} = -1{,}37$$

Realizando-se uma estatística descritiva, nas duas variáveis, têm-se os seguintes resultados:

Tabela 20 – Estatística descritiva relativa a duas variáveis, avaliadas em cinco indivíduos.

	Variável X_1	Variável X_2
Média aritmética das variáveis	96	73,8
Somatório ao quadrado das variáveis	46178	27529
Somatório das variáveis	480	369
Variância amostral das variáveis	24,5	74,2
Desvio padrão amostral das variáveis	4,9497	8,6139
Desvio padrão amostral das variáveis padronizadas	1	1

A matriz de correlação *R*, que é extraída das variáveis originais, *será* calculada pela expressão do item 3.11, e as correlações entre as variáveis serão obtidas pela equação do item 3.12*:*

$$r_{x_1 x_2} = \frac{26}{4{,}95.8{,}61} = 0{,}61.$$

A correlação entre a variável, em relação a ela mesma, será fornecida pela expressão do item 3.13:

$$r_{11} = \frac{24{,}5^2}{24{,}5^2} = 1,$$

logo, a matriz de correlação será assim constituída:

$$R = \begin{bmatrix} 1 & 0{,}61 \\ 0{,}61 & 1 \end{bmatrix}.$$

Para encontrar os autovalores, a partir da matriz de correlação *R*, deve-se partir da seguinte equação característica:

$$\left| R - \hat{\Lambda} I \right| = 0.$$

Substituindo-se essa equação pelas matrizes *R* e *I*, obtém-se a seguinte expressão:

$$\left\| \begin{bmatrix} 1 & 0{,}61 \\ 0{,}61 & 1 \end{bmatrix} - \hat{\Lambda} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\| = 0.$$

Multiplicando-se o autovalor $\hat{\Lambda}$ à matriz *I,* obtém-se as seguintes matrizes:

$$\left\| \begin{bmatrix} 1 & 0{,}61 \\ 0{,}61 & 1 \end{bmatrix} - \begin{bmatrix} \hat{\Lambda} & 0 \\ 0 & \hat{\Lambda} \end{bmatrix} \right\| = 0.$$

Realizando-se a subtração entre as matrizes, obtém-se a matriz:

$$\begin{vmatrix} 1 - \hat{\Lambda} & 0{,}61 \\ 0{,}61 & 1 - \hat{\Lambda} \end{vmatrix} = 0.$$

Resolvendo o determinante dessa matriz, encontra-se o seguinte resultado:

$$(1 - \hat{\Lambda})(1 - \hat{\Lambda}) - (0{,}61)^2 = 0.$$

Unindo-se os termos semelhantes, encontra-se uma equação do segundo grau:

$$1 - \hat{\Lambda} - \hat{\Lambda} + \hat{\Lambda}^2 - 0{,}37 = 0.$$

Resolvendo essa equação, encontra-se os autovalores correspondentes à matriz *R*.

$$\hat{\Lambda}^2 - 2\hat{\Lambda} + 0{,}63 = 0.$$

Os autovalores (raízes características) são obtidos da seguinte equação:

$\hat{\Lambda} = \dfrac{2 \pm \sqrt{(-2)^2 - 4(1)(0{,}63)}}{2(1)}$ logo, os dois autovalores resultantes da equação são:

$\hat{\Lambda}_1$ = 1,61 e $\hat{\Lambda}_2$ = 0,39.

Como pode-se observar, a adição de duas raízes características dá 2, que nada mais é que o segundo termo da equação.

Deve-se observar, também, que a soma dos autovalores corresponde ao traço e ao determinante da matriz *R*.

$\hat{\Lambda}_1 + \hat{\Lambda}_2 + ... + \hat{\Lambda}_p =$ traço da matriz *R*.

ou seja, 1,61 + 0,39 = 2 = traço da matriz *R.*

$(\hat{\Lambda}_1).(\hat{\Lambda}_2) ....(\hat{\Lambda}_p) =$ determinante da matriz *R*.

(1,61).(0,39) = 0,63.

Se a seguinte expressão for resolvida $\frac{\hat{\Lambda}_1}{traço\, R}.100$, tem-se a proporção da variância total, explicada por cada componente principal. Observa-se que a primeira componente explica $\frac{1,61}{2}.100 = 80,50\%$, e a segunda componente explica $\frac{0,39}{2}.100 = 19,50\%$.

Ou seja, a primeira componente relativa à raiz $\hat{\Lambda}_1$, explica 80,50% da variação total dos dados.

A segunda componente, relativa à raiz $\hat{\Lambda}_2$, explica 19,50% da variação total dos dados.

Essa variância será distribuída entre $\hat{\Lambda}_1$ = 1,61 e $\hat{\Lambda}_2$ = 0,39, ou seja, 80,50% da variância é explicada pelo primeiro eixo fatorial, e 19,50% pelo segundo.

O cálculo da primeira componente, referente a $\hat{\Lambda}_1$ = 1,61, será dado pelo autovetor associado a $\hat{\Lambda}_1$, conforme a equação:

$\left|R - \hat{\Lambda}_1 I\right|\hat{e}_1 =$ 0.

Substituindo-se essa equação pelas matrizes *R*, *I*, pelo primeiro autovetor $\hat{\Lambda}_1$ = 1,61 e pela matriz de incógnitas, obtém-se a seguinte expressão:

$$\left|\begin{bmatrix} 1 & 0,61 \\ 0,61 & 1 \end{bmatrix} - 1,61\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right|\begin{bmatrix} \hat{e}_{11} \\ \hat{e}_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Multiplicando-se o autovalor $\hat{\Lambda}_1$ à matriz *I* e subtraindo da matriz *R,* obtém-se as seguintes matrizes:

$$\begin{bmatrix} 1-1,61 & 0,61 \\ 0,61 & 1-1,61 \end{bmatrix}\begin{bmatrix} \hat{e}_{11} \\ \hat{e}_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Multiplicando-se essas matrizes encontra-se o seguinte sistema:

$$\begin{cases} -0,61\hat{e}_{11} + 0,61\hat{e}_{12} = 0 \\ 0,61\hat{e}_{11} - 0,61\hat{e}_{12} = 0 \end{cases}.$$

Esse sistema de equações é indeterminado em virtude de $\left|R - \hat{\Lambda}I\right| = 0$

$$\begin{vmatrix} -0,61 & 0,61 \\ 0,61 & -0,61 \end{vmatrix} = 0.$$

Devido a isso, pode-se deixar uma das equações (neste caso a segunda) e atribuir um valor qualquer, que não seja nulo, a uma das incógnitas ($\hat{e}_{12} = 1$). Dessa forma, tem-se:

- 0,61 $\hat{e}_{11}$ + 0,61.(1) = 0

- 0,61 $\hat{e}_{11}$ = - 0,61, logo $\hat{e}_{11}$ será:

$\hat{e}_{11}$ = 1,

e o autovetor associado ao primeiro autovalor $\hat{\Lambda}_1$ = 1,61, será:

$\hat{e}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ e, sua norma será:

$\left\|\hat{e}_1\right\| = \sqrt{(1)^2 + (1)^2}$ = 1,41.

Para que esse vetor seja unitário, é necessário normalizar o autovetor a 1, da seguinte forma:

$$e_1 = \frac{1}{\left\|\hat{e}_1\right\|}.\hat{e}_1.$$

Substituindo-se essa expressão, pelos seus respectivos valores, têm-se:

$$e_1 = \frac{1}{1,41}\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Portanto, o primeiro autovetor normalizado será:

$$e_1 = \begin{bmatrix} 0,71 \\ 0,71 \end{bmatrix},$$

e a sua norma será:

$\left\|e_1\right\| = \sqrt{(-0,71)^2 + (0,71)^2}$ = 1.

Como pode-se observar $e_1^t e_1 = 1$, sendo esta a primeira restrição feita por Morrison (1976), para que o sistema tenha solução única.

Logo, o primeiro componente principal será:

$Y_1 = 0{,}71Z_1 + 0{,}71Z_2$.

O segundo componente principal é dado pela outra raiz $\hat{\Lambda}_2 = 0{,}39$:

$$\left|R - \hat{\Lambda}^2 I\right|\hat{e}_2 = 0.$$

Substituindo-se essa equação pelas matrizes *R*, *I*, pelo segundo autovalor $\hat{\Lambda}_2 = 13{,}38$, e pela matriz de incógnitas, obtém-se a seguinte expressão:

$$\left\|\begin{bmatrix} 1 & 0{,}61 \\ 0{,}61 & 1 \end{bmatrix} - 0{,}39\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right\|\begin{bmatrix} \hat{e}_{21} \\ \hat{e}_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Multiplicando-se o autovalor $\hat{\Lambda}_2$ à matriz *I* e subtraindo da matriz *R,* obtém-se as seguintes matrizes:

$$\begin{bmatrix} 1 - 0{,}39 & 0{,}61 \\ 0{,}61 & 1 - 0{,}39 \end{bmatrix}\begin{bmatrix} \hat{e}_{21} \\ \hat{e}_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Multiplicando-se essas matrizes encontra-se o seguinte sistema:

$$\begin{cases} 0{,}61\hat{e}_{21} + 0{,}61\hat{e}_{22} = 0 \\ 0{,}61\hat{e}_{21} + 0{,}61\hat{e}_{22} = 0 \end{cases}.$$

Fazendo-se o procedimento análogo ao anterior, tem-se:

$0{,}61\hat{e}_{21} + 0{,}61(1) = 0,$ logo a incógnita $\hat{e}_{21}$, será:

$\hat{e}_{21} = -1,$

e o autovetor associado ao segundo autovalor $\hat{\Lambda}_2 = 0{,}39$, será:

$$\hat{e}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix},$$

e sua norma será de:

$\left\|\hat{e}_2\right\| = \sqrt{(-1)^2 + (1)^2}$ = 1,41.

Para que esse vetor seja unitário, é necessário normalizar o autovetor a 1, da seguinte forma:

$$e_2 = \frac{1}{\left\|\hat{e}_2\right\|}\hat{e}_2 = \frac{1}{1{,}41}\begin{bmatrix} -1 \\ 1 \end{bmatrix},$$

logo, o segundo autovetor normalizado será:

$$e_2 = \begin{bmatrix} -0{,}71 \\ 0{,}71 \end{bmatrix},$$

e sua norma será:

$\|e_2\| = \sqrt{(-0,71)^2 + (0,71)^2} = 1.$

Como pode-se observar, $e_2^t e_2 = 1$ é a primeira restrição feita por Morrison (1976), para que o sistema tenha solução única.

Logo, a segunda componente principal será:

$Y_2 = -0,71Z_1 + 0,71Z_2.$

Outra observação é que, neste exemplo, os componentes principais são ortogonais, pois $e_1^t e_2 = 0,$ que é a segunda restrição feita por Morrison (1976).

Para que esta restrição seja satisfeita deve-se multiplicar o primeiro autovetor normalizado transposto pelo segundo autovetor normalizado, procedendo-se da seguinte forma:

$$e_1^t e_2 = \begin{bmatrix} 0,71 & 0,71 \end{bmatrix} \begin{bmatrix} -0,71 \\ 0,71 \end{bmatrix}.$$

Multiplicando-se os autovetores normalizados, têm-se a seguinte expressão:

$e_1^t e_2 = (0,71)(-0,71) + (0,71)(0,71),$

tem-se que:

$e_1^t e_2 = -0,50 + 0,50 = 0.$

O passo a seguir é encontrar o valor de cada componente principal, procedendo-se de forma análoga ao exemplo 1:

Tabela 21 – Mostra os escores para análise de componentes principais.

Observações	Variáveis		Escores para os componentes principais	
	X_1	X_2	Y_1	Y_2
1	100	76	0,76	-0,39
2	93	82	0,24	1,10
3	102	81	1,46	-0,26
4	95	68	-0,62	0,34
5	90	62	-1,83	-0,11

Para completar a análise de componentes principais, é necessário fazer a correlação entre as variáveis Z_j e Y_i, como se pode verificar a seguir:

$r_{z_1 y_1} = e_{11}\sqrt{\hat{\Lambda}_1}$

$r_{z_1 y_1} = 0.71\sqrt{1,61}$ = 0,90

$r_{z_2 y_1} = e_{12}\sqrt{\hat{\Lambda}_1}$

$r_{z_2 y_1} = 0.71\sqrt{1,61}$ = 0,90

$r_{z_1 y_2} = e_{21}\sqrt{\hat{\Lambda}_2}$

$r_{z_1 y_2} = -0.71\sqrt{0,39}$ = -0,44

$r_{z_2 y_2} = e_{22}\sqrt{\hat{\Lambda}_2}$

$r_{z_2 y_2} = 0.71\sqrt{0,39}$ = 0,44

A Tabela 22 mostra as principais informações de uma análise de componentes principais.

Tabela 22 – Componentes principais obtidos da análise de duas variáveis padronizadas Z_1 *e* Z_2.

Componentes principais	Autovalor	Coeficiente de ponderação		Correlação entre Z_j Y_i		Percentagem da variância de Y_i	Percentagem acumulada da variância dos Y_i
		Z_1	Z_2	Z_1	Z_2		
Y_1	1,61	0,71	0,71	0,90	0,90	80,50%	80,50%
Y_2	0,39	-0,71	0,71	-0,44	0,44	19,50%	100%

Como pode-se observar novamente, a primeira componente Y_1 possui a maior correlação, sendo esta a de maior importância para o estudo.

Deve-se observar que os valores obtidos dos componentes principais, através da matriz *S*, em geral não são os mesmos que os obtidos da matriz *R*.

Comentário desse capítulo

Nesse capítulo foi possível verificar como as análises se comportam quando realizadas normalmente, o que mostrou o inter-relacionamento das variáveis, sendo óbvio que, quando o número de variáveis é grande, o uso de um pacote

computacional é indispensável. No capítulo 4, desenvolveram-se dois exemplos com dados reais, utilizando-se um programa específico.

4 APLICAÇÃO

Neste capítulo 4, aplica-se técnicas multivariadas utilizandon-se o *software statistica*.

Para desenvolver esta pesquisa, utilizou-se dois bancos de dados. O primeiro, utilizado para desenvolver o exemplo da análise de agrupamentos, refere-se à produção de grãos do setor agroindustrial brasileiro, no período de 1995 a 2002, e o segundo para desenvolver o exemplo da análise fatorial de componentes principais, refere-se a 30 coletas da fauna edáfica do solo, no período de 06 de junho de 2004 a 04 de janeiro de 2005, com coletas semanais.

4.1 Análise de Agrupamentos

Detalha-se, a partir de agora, os procedimentos para realização da *AA,* utilizando-se o método de agrupamento do vizinho mais próximo, no qual serão salientados alguns princípios gerais de interpretação dos resultados numéricos e gráficos de uma *AA,* utilizando-se o *software Statistica*.

Conforme Figura 19, para encontrar os grupos de variáveis com as mesmas características, que constituem o dendograma na análise, deve-se proceder da seguinte forma: Acessar a barra de tarefas e clicar em Iniciar/Programas/*Statistica /Statistica*, conforme a seguinte caixa do programa:

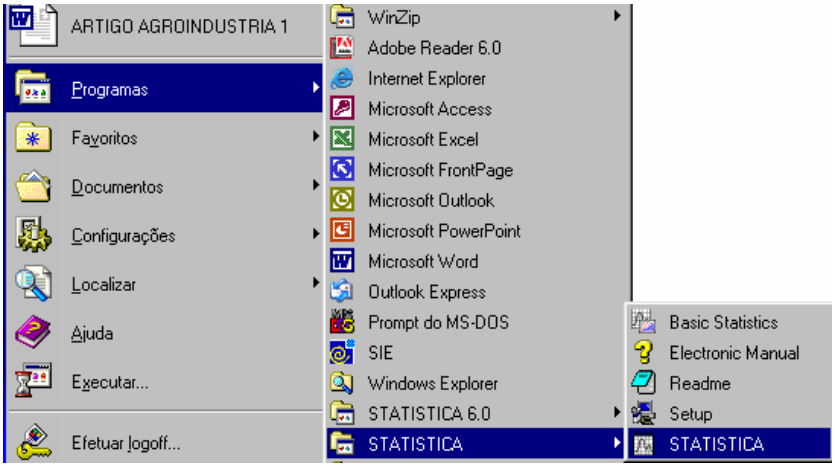


Figura 19 - Caixa de seleção das análises estatísticas.

A Figura 20 mostra como transportar o banco de dados do excel para o programa *statistica* sem que seja necessário copiar as variáveis de forma individual.

Deve-se *clicar* na opção abrir *Arquivos do tipo*: selecionar *Excel Files (*.xls)*, na opção *Examinar* selecionar a pasta em que está arquivo do excel, na opção *Nome do arquivo:* selecionar a o banco de dados do excel e clicar em *Abrir.*

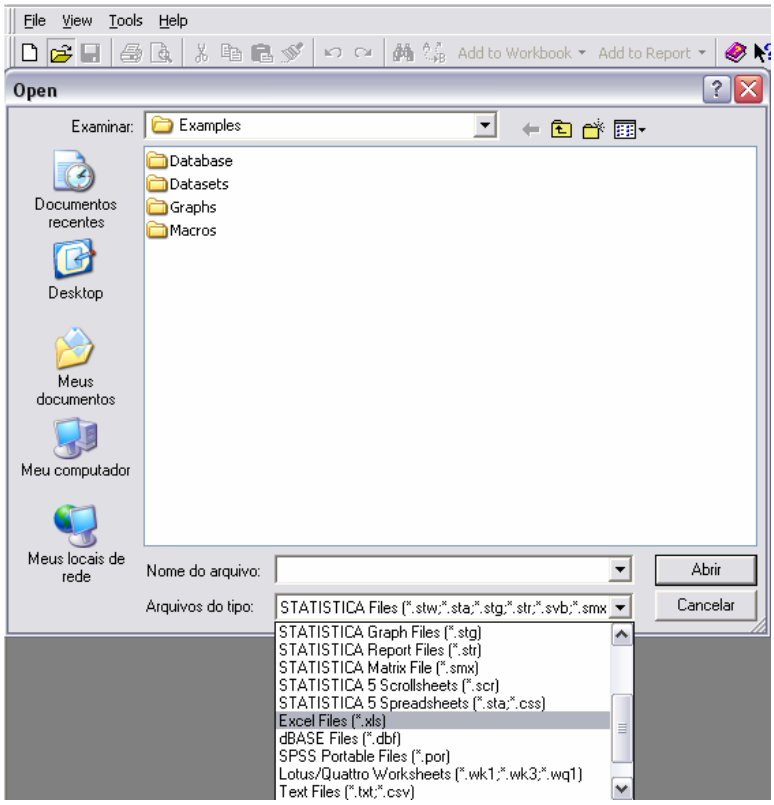


Figura 20 - Caixa de seleção para importar os dados do excel para o programa *statistica.*

Na Figura 21 selecionando a primeira opção *Import all sheets to a Workbook*, importa-se todas as planilhas para área de trabalho, selecionando a segunda opção, *Import selected to a Spreadsheet*, importa-se todas as planilhas selecionadas.

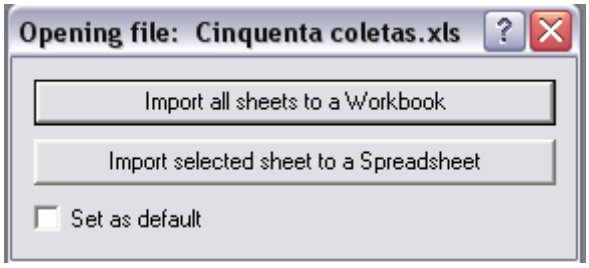


Figura 21 - Caixa de seleção para importar os todos os dados do excel para o programa *statistica.*

A Figura 22 mostra que selecionando a primeira opção serão importados os nomes da primeira coluna, que geralmente são variáveis qualitativas, selecionando a segunda opção serão importados os nomes das variáveis que estão na primeira linha de uma planilha excel e selecionando a terceira opção serão importados no formato em que foram importados os dados.

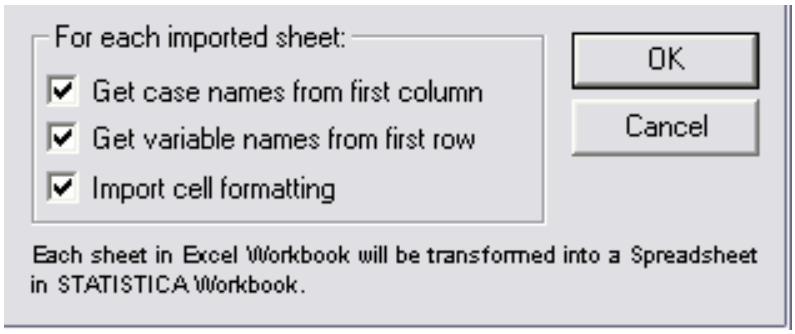
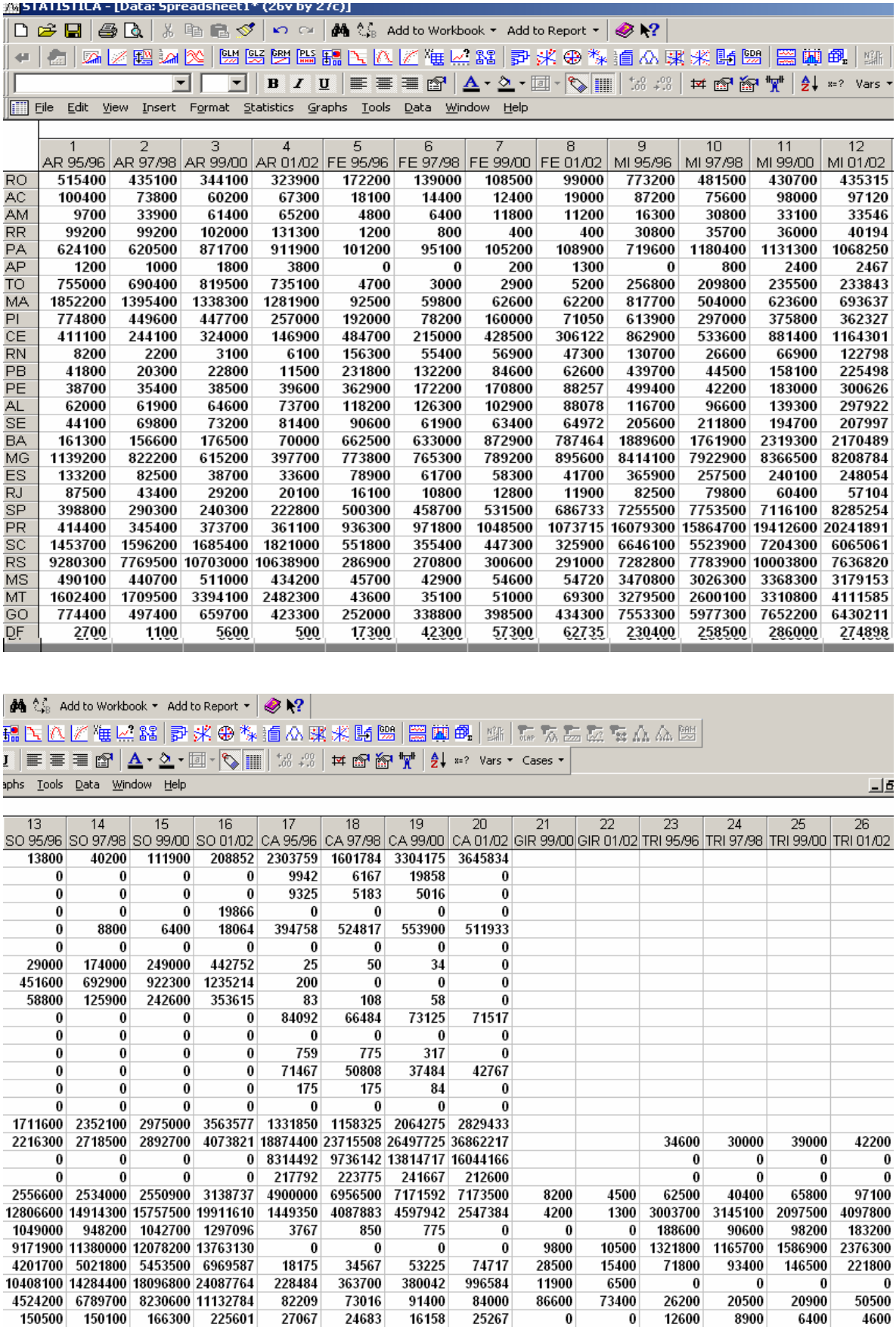


Figura 22 - Caixa de seleção para importar os dados do excel para o programa *statistica,* por linhas e por colunas.

A amostra, utilizada para este exemplo, refere-se à produção de grãos do setor agrícola brasileiro, no período de 1995 a 2002, sendo que esta técnica possibilitará fazer uma síntese da produção de grãos neste período, bem como identificar os estados que possuíram médias semelhantes de produção, através dos grupos formados e, conseqüentemente, os estados que apresentaram a maior produção.

O banco de dados é constituído pelos 27 estados brasileiros, que são os casos, e pela produção das seguintes culturas: soja, milho, café, trigo, girassol, feijão e arroz, entre outras, perfazendo um total de 26 variáveis, num período de oito anos. As culturas em estudo são constituídas pelos produtos de maior expressão de produção nos 27 estados, com coletas anuais medidas em toneladas. Para efetuar a análise, foi realizada uma média bianual das produções, pois esta possibilitou uma melhor visualização das variáveis, não sobrepondo, graficamente, as culturas analisadas.

Inicialmente, elaborou-se o banco de dados com as variáveis representadas nas colunas, e os objetos nas linhas, como mostra a Figura 23.



	1 AR 95/96	2 AR 97/98	3 AR 99/00	4 AR 01/02	5 FE 95/96	6 FE 97/98	7 FE 99/00	8 FE 01/02	9 MI 95/96	10 MI 97/98	11 MI 99/00	12 MI 01/02
RO	515400	435100	344100	323900	172200	139000	108500	99000	773200	481500	430700	435315
AC	100400	73800	60200	67300	18100	14400	12400	19000	87200	75600	98000	97120
AM	9700	33900	61400	65200	4800	6400	11800	11200	16300	30800	33100	33546
RR	99200	99200	102000	131300	1200	800	400	400	30800	35700	36000	40194
PA	624100	620500	871700	911900	101200	95100	105200	108900	719600	1180400	1131300	1068250
AP	1200	1000	1800	3800	0	0	200	1300	0	800	2400	2467
TO	755000	690400	819500	735100	4700	3000	2900	5200	256800	209800	235500	233843
MA	1852200	1395400	1338300	1281900	92500	59800	62600	62200	817700	504000	623600	693637
PI	774800	449600	447700	257000	192000	78200	160000	71050	613900	297000	375800	362327
CE	411100	244100	324000	146900	484700	215000	428500	306122	862900	533600	881400	1164301
RN	8200	2200	3100	6100	156300	55400	56900	47300	130700	26600	66900	122798
PB	41800	20300	22800	11500	231800	132200	84600	62600	439700	44500	158100	225498
PE	38700	35400	38500	39600	362900	172200	170800	88257	499400	42200	183000	300626
AL	62000	61900	64600	73700	118200	126300	102900	88078	116700	96600	139300	297922
SE	44100	69800	73200	81400	90600	61900	63400	64972	205600	211800	194700	207997
BA	161300	156600	176500	70000	662500	633000	872900	787464	1889600	1761900	2319300	2170489
MG	1139200	822200	615200	397700	773800	765300	789200	895600	8414100	7922900	8366500	8208784
ES	133200	82500	38700	33600	78900	61700	58300	41700	365900	257500	240100	248054
RJ	87500	43400	29200	20100	16100	10800	12800	11900	82500	79800	60400	57104
SP	398800	290300	240300	222800	500300	458700	531500	686733	7255500	7753500	7116100	8285254
PR	414400	345400	373700	361100	936300	971800	1048500	1073715	16079300	15864700	19412600	20241891
SC	1453700	1596200	1685400	1821000	551800	355400	447300	325900	6646100	5523900	7204300	6065061
RS	9280300	7769500	10703000	10638900	286900	270800	300600	291000	7282800	7783900	10003800	7636820
MS	490100	440700	511000	434200	45700	42900	54600	54720	3470800	3026300	3368300	3179153
MT	1602400	1709500	3394100	2482300	43600	35100	51000	69300	3279500	2600100	3310800	4111585
GO	774400	497400	659700	423300	252000	338800	398500	434300	7553300	5977300	7652200	6430211
DF	2700	1100	5600	500	17300	42300	57300	62735	230400	258500	286000	274898

13 SO 95/96	14 SO 97/98	15 SO 99/00	16 SO 01/02	17 CA 95/96	18 CA 97/98	19 CA 99/00	20 CA 01/02	21 GIR 99/00	22 GIR 01/02	23 TRI 95/96	24 TRI 97/98	25 TRI 99/00	26 TRI 01/02
13800	40200	111900	208852	2303759	1601784	3304175	3645834						
0	0	0	0	9942	6167	19858	0						
0	0	0	0	9325	5183	5016	0						
0	0	0	19866	0	0	0	0						
0	8800	6400	18064	394758	524817	553900	511933						
0	0	0	0	0	0	0	0						
29000	174000	249000	442752	25	50	34	0						
451600	692900	922300	1235214	200	0	0	0						
58800	125900	242600	353615	83	108	58	0						
0	0	0	0	84092	66484	73125	71517						
0	0	0	0	0	0	0	0						
0	0	0	0	759	775	317	0						
0	0	0	0	71467	50808	37484	42767						
0	0	0	0	175	175	84	0						
0	0	0	0	0	0	0	0						
1711600	2352100	2975000	3563577	1331850	1158325	2064275	2829433						
2216300	2718500	2892700	4073821	18874400	23715508	26497725	36862217			34600	30000	39000	42200
0	0	0	0	8314492	9736142	13814717	16044166			0	0	0	0
0	0	0	0	217792	223775	241667	212600			0	0	0	0
2556600	2534000	2550900	3138737	4900000	6956500	7171592	7173500	8200	4500	62500	40400	65800	97100
12806600	14914300	15757500	19911610	1449350	4087883	4597942	2547384	4200	1300	3003700	3145100	2097500	4097800
1049000	948200	1042700	1297096	3767	850	775	0	0	0	188600	90600	98200	183200
9171900	11380000	12078200	13763130	0	0	0	0	9800	10500	1321800	1165700	1586900	2376300
4201700	5021800	5453500	6969587	18175	34567	53225	74717	28500	15400	71800	93400	146500	221800
10408100	14284400	18096800	24087764	228484	363700	380042	996584	11900	6500	0	0	0	0
4524200	6789700	8230600	11132784	82209	73016	91400	84000	86600	73400	26200	20500	20900	50500
150500	150100	166300	225601	27067	24683	16158	25267	0	0	12600	8900	6400	4600

Figura 23 - Caixa das variáveis para *AA*.

Analisando-se a Figura 23, pode-se concluir que nem todos os estados produzem todos os produtos, ou seja, alguns produtos são característicos de algumas regiões, apenas. A descrição das variáveis envolvidas neste estudo é a seguinte: V_1 representará a variável 1, V_2 representará a variável 2 e assim sucessivamente, com a demais variáveis:

V_1 = produção de arroz, nos anos de 1995/1996.

V_2 = produção de arroz, nos anos de 1997/1998.

V_3 = produção de arroz, nos anos de 1999/2000.

V_4 = produção de arroz, nos anos de 2001/2002.

V_5 = produção de feijão, nos anos de 1995/1996.

V_6 = produção de feijão, nos anos de 1997/1998.

V_7 = produção de feijão, nos anos de 1999/2000.

V_8 = produção de feijão, nos anos de 2001/2002.

V_9 = produção de milho, nos anos de 1995/1996.

V_{10} = produção de milho, nos anos de 1997/1998.

V_{11} = produção de milho, nos anos de 1999/2000.

V_{12} = produção de milho, nos anos de 2001/2002.

V_{13} = produção de soja, nos anos de 1995/1996.

V_{14} = produção de soja, nos anos de 1997/1998.

V_{15} = produção de soja, nos anos de 1999/2000.

V_{16} = produção de soja, nos anos de 2001/2002.

V_{17} = produção de café, nos anos de 1995/1996.

V_{18} = produção de café, nos anos de 1997/1998.

V_{19} = produção de café, nos anos de 1999/2000.

V_{20} = produção de café, nos anos de 2001/2002.

V_{21} = produção de girassol, nos anos de 1999/2000.

V_{22} = produção de girassol, nos anos de 2001/2002.

V_{23} = produção de trigo, nos anos de 1995/1996.

V_{24} = produção de trigo, nos anos de 1997/1998.

V_{25} = produção de trigo, nos anos de 1999/2000.

V_{26} = produção de trigo, nos anos de 2001/2002.

Para a realização da análise, seleciona-se, no menu de opções, o módulo principal do *STATISTICA*, a opção *Multivariate Exploratory Techniques – Cluster Analysis*, conforme a caixa de seleção mostrada na Figura 24.

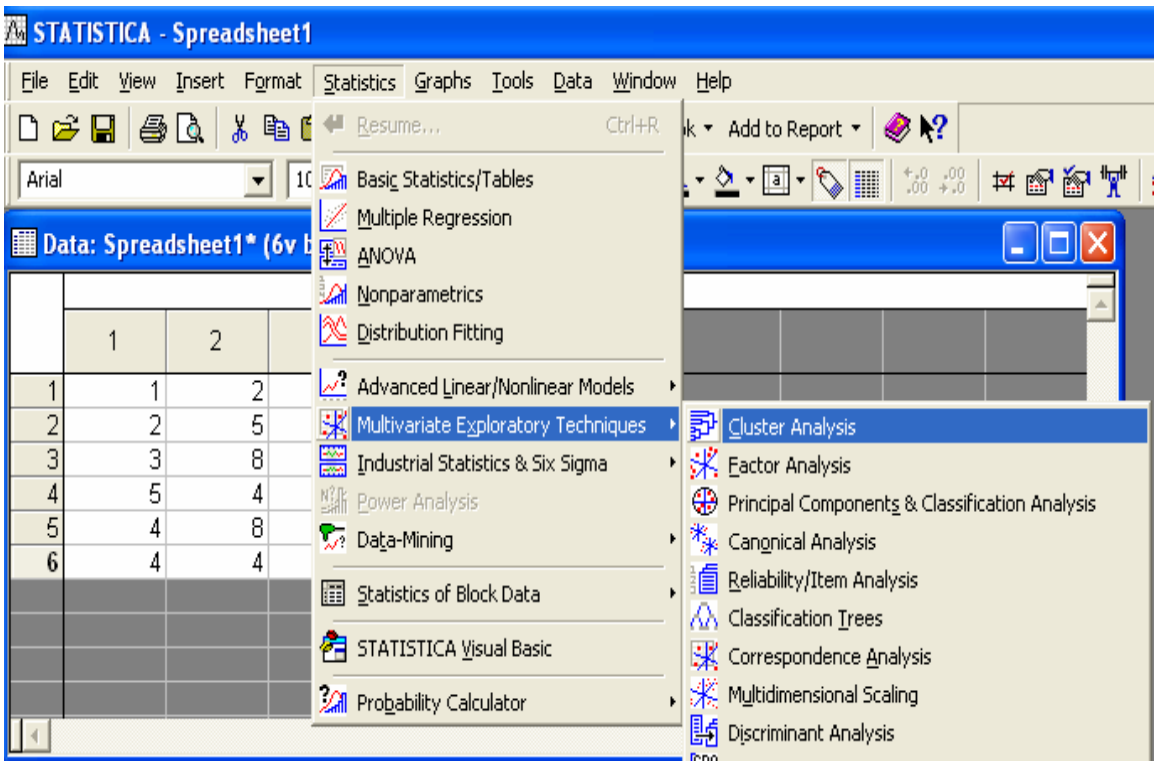


Figura 24 - Caixa de seleção da *AA*.

A Figura 26 mostra a caixa de seleção de opções, para se realizar uma análise de agrupamentos. Selecionando *Joning (tree clustering),* é possível encontrar o dendograma, o qual mostrará o número de grupos formados pelas mesmas características. Outra opção é selecionar *K-means clustering,* que irá definir o número de grupos a serem utilizados na análise. Esses grupos são definidos pelas médias encontradas no banco de dados inicial. E ainda existe outra forma de realizar a análise, através da opção *Two-way joining,* que torna possível fazer um mapa associativo entre cada variável e a unidade amostral, permitindo, através da inspeção visual, qual variável possui uma maior representatividade para o conjunto de dados, mas estas não foram citadas no trabalho.

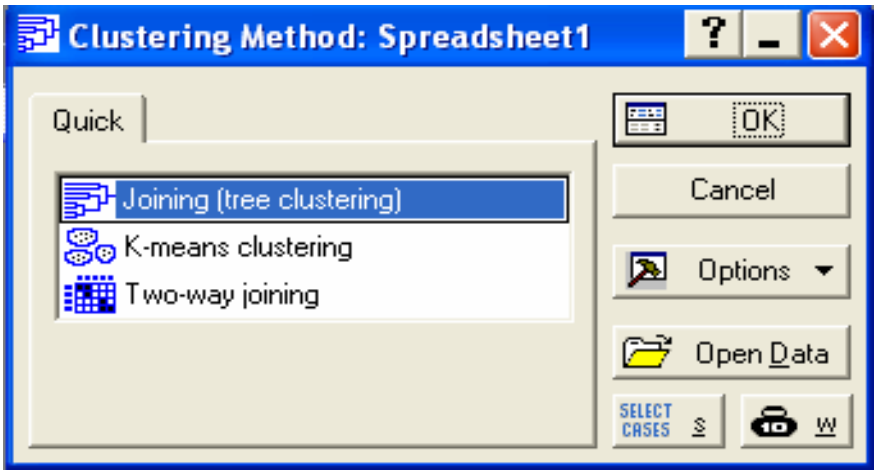


Figura 25 - Caixa de seleção para análise de agrupamentos.

A Figura 25 mostra a caixa de diálogo das variáveis para *AA.* Nesta caixa existem várias opções para a realização da análise. Selecionando a opção V*ariables,* é possível visualizar e selecionar as variáveis que o pesquisador deseja incluir na análise. Na opção *Imput in file* encontra-se as opções *Raw data,* que é utilizada para os dados brutos do banco de dados. Outra opção desta caixa de diálogo é *Cluster,* que possibilita realizar a análise de duas formas: se selecionar *variables,* o agrupamento será feito por colunas e se for selecionado *cases* o agrupamento será realizado por linhas.

A caixa de seleção mostra, ainda, a opção *Amalgamation (linkage) rule,* na qual se encontra os métodos de encadeamento: *Single Linkage,* que se baseia na distância mínima; *Complete Linkage*, que se baseia na distância máxima entre objetos, dentre outras distâncias que se encontram dispostas para serem utilizadas na análise. A última opção desta caixa de diálogo é *Distance measure,* na qual o pesquisador poderá selecionar o tipo de distância que deseja utilizar em seu trabalho. É importante lembrar que a distância mais utilizada é a *Euclidean distances,* ou seja, a distância euclidiana.

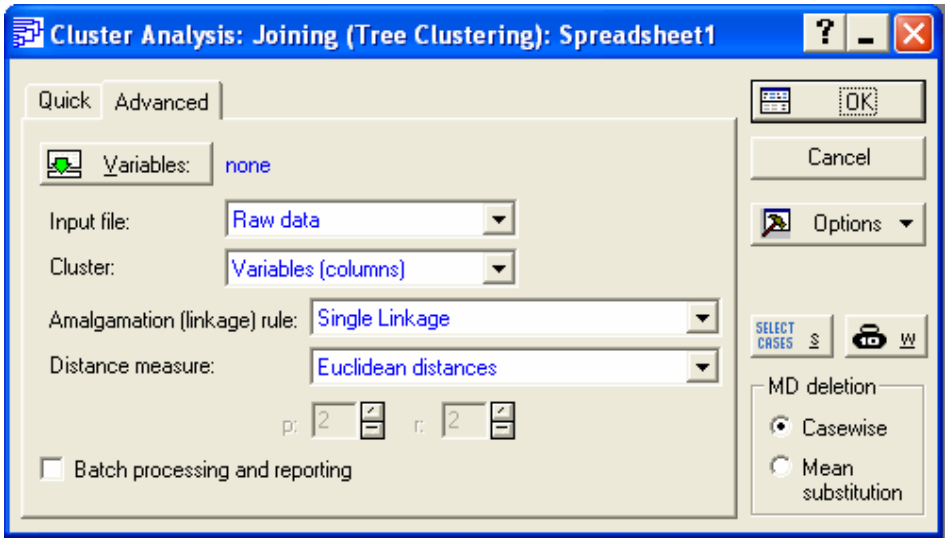


Figura 26 - Caixa de seleção, para análise de agrupamento.

Para selecionar todas as variáveis, basta clicar em *Select All,* e *OK,* conforme Figura 27. Se desejar selecionar apenas algumas variáveis, deve-se utilizar a tecla *ctrl,* e clicar nas variáveis desejadas.

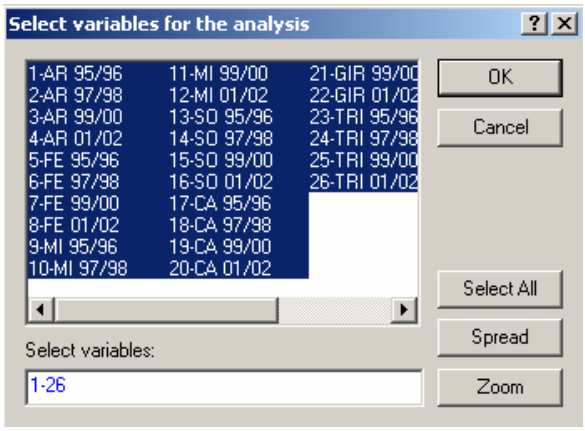


Figura 27- Caixa de seleção das variáveis, para a análise de agrupamentos.

A Figura 28 mostra a caixa de seleção de comandos para a *AA*, selecionando *Advanced/Horizontal hierarchical tree plot,* tem-se o dendograma horizontal, e escolhendo-se a opção *Vertical icicle plot,* tem-se o dendograma vertical. A caixa de seleção ainda traz a opção da matriz de distâncias entre as variáveis *Distance matrix,* e possibilita, ainda, realizar uma estatística descritiva nos dados, selecionando a opção *Descriptive statistics,* que pode ser de interesse do pesquisador. Vale lembrar que estas estatísticas são referentes às variáveis originais.

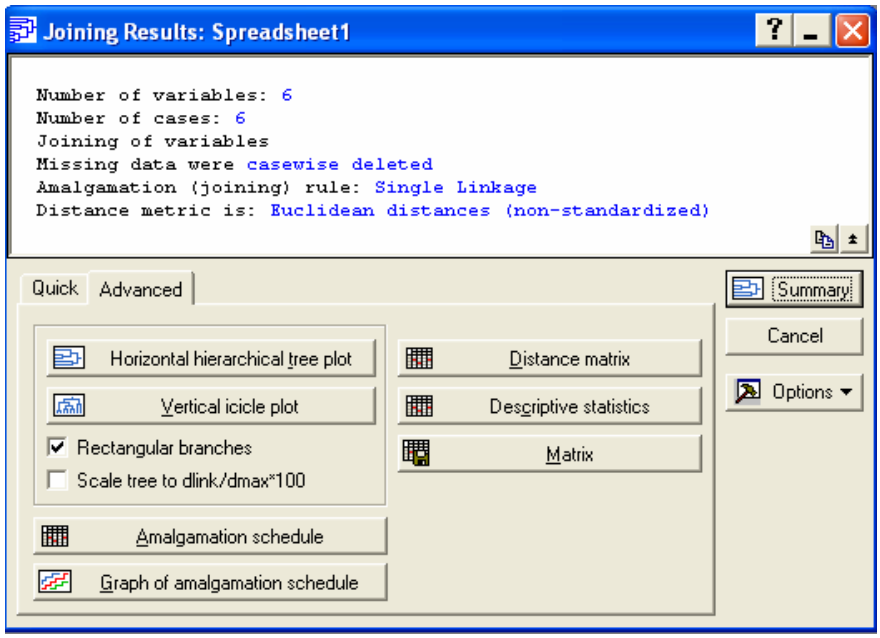


Figura 28 - Caixa de seleção do dendograma, matriz de distâncias e estatística descritiva, para a análise de agrupamento.

A Figura 29, mostra o dendograma considerando o método do vizinho mais próximo, como o algoritmo de agrupamento dos dados, e será considerada a distância euclidiana como medida de dissimilaridade.

O dendograma, a seguir, é formado com base nos pares de objetos mais similares, ou seja, com a menor distância entre eles. Logo após, estes objetos, ou grupos já formados, vão reunir-se em razão de similaridade decrescente.

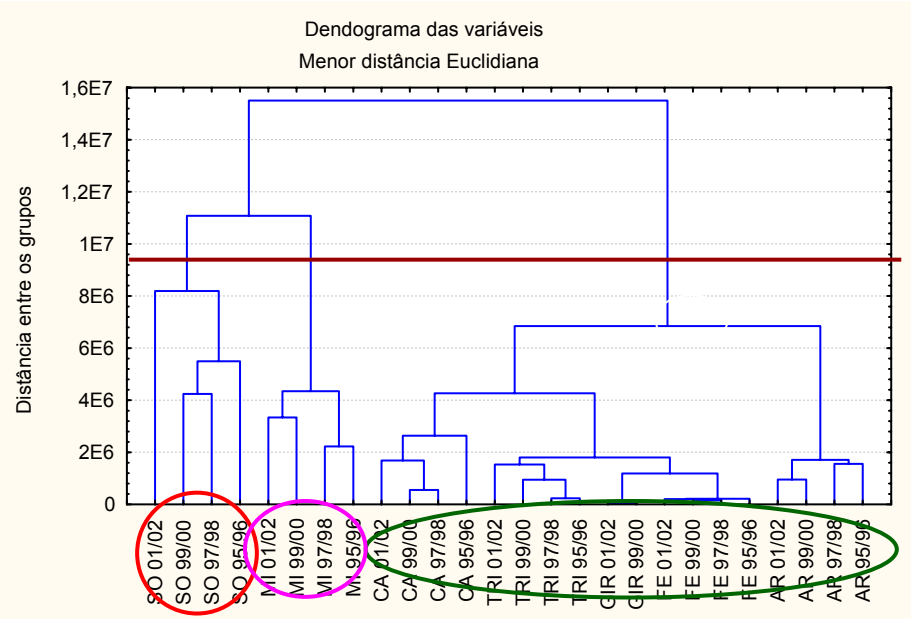


Figura 29 - Dendograma da matriz de distâncias, pelo método de agrupamento por ligação simples.

No dendograma da Figura 29, a escala vertical indica o nível de similaridade, e no eixo horizontal são marcados os indivíduos, na ordem em que são agrupados. As linhas verticais partem dos indivíduos, e têm altura correspondente ao nível em que os indivíduos são considerados semelhantes.

Observando a Figura 29, verifica-se que o maior salto encontra-se entre a distância $8x10^6$ e $1x10^7$ no gráfico referido como 8E6 e 1E7 respectivamente. Se se fizer um corte no gráfico, entre essas distâncias, ter-se-á, três grupos homogêneos distintos. O primeiro grupo é formado pelas variáveis: arroz, feijão, girassol, trigo e café, que está sendo representado pelo círculo em verde, sendo que as variáveis, que formam esse grupo, representam a menor produção de grãos em todo o período; o segundo grupo é formado pela variável milho, que está sendo representada pelo círculo em rosa, esta variável manteve sua produção constante no período de 1995 a 1998 e teve um aumento significativo no ano de 1999, mantendo-se constante até o ano de 2002.

O terceiro grupo é formado pela variável soja, que está sendo representado pelo círculo em vermelho. Essa variável formou, no dendograma, um grupo isolado, devido a sua produção ser superior às demais, embora que esta tenha tido várias oscilações ocorridas no período. Nos anos de 1995 e 1996 representou uma produção significativa, ocorrendo um decréscimo no ano de 1997, mantendo-se instável até o ano de 2000. Só tornou a aumentar no ano de 2001 e 2002, os quais se destacaram pela alta produção ocorrida.

Antes de concluir a análise sobre o dendograma, é pertinente lembrar que o corte, no gráfico, geralmente, é realizado em relação às maiores distâncias em que os grupos foram formados, levando-se, sempre, em consideração os critérios adotados por cada pesquisador.

O gráfico da Figura 30 serve de auxílio para o pesquisador, caso no dendograma não esteja claro entre quais distâncias ocorra o maior salto. Analisando-se este gráfico, é possível ver que o corte deve ser realizado no dendograma entre as distâncias $8x10^6$ e $1x10^7$, no qual ocorre o maior salto, conforme indicado no gráfico pelo círculo em vermelho.

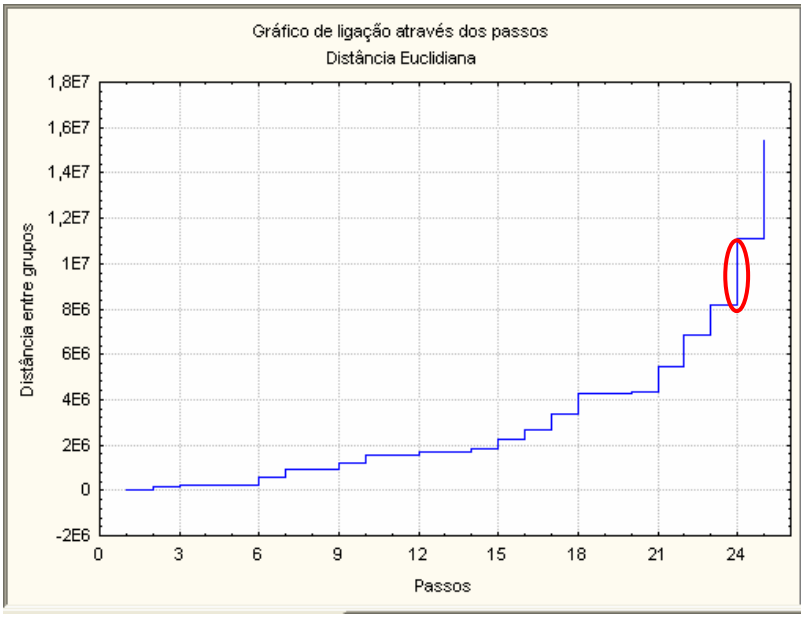


Figura 30- Gráfico das distâncias nas quais os grupos foram formados.

Como pode-se observar na Figura 31, os indivíduos que estão em um mesmo grupo possuem médias de produção semelhantes, e os que possuíam médias diferentes formaram outros grupos, isso comprova a existência de homogenidade dentro do grupo e heterogenidade entre os grupos.

Aplicando-se a *AA,* por linhas, encontra-se o dendograma referente aos estados que constituiram a amostra.

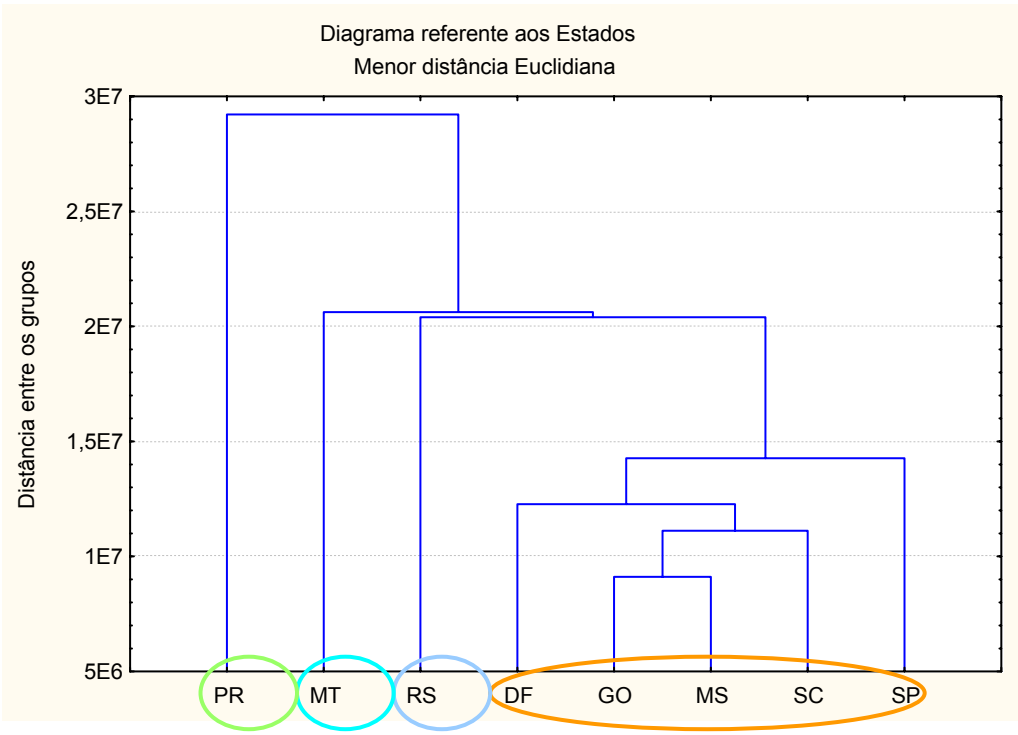


Figura 31 - Dendograma referente aos estados, utilizando o método de agrupamento de ligação simples.

Analisando-se o dendograma da Figura 31, pode-se concluir que nos estados do DF, GO, MS, SC e SP, no período de 1995 a 2002, a produção de grãos manteve-se semelhante, a qual foi inferior em relação aos estados do RS, MT e o PR, que formaram grupos distintos no dendograma, ou seja, no decorrer do período, a produção de grãos, nesses estados, teve uma característica própria, uma maior representatividade, formando, assim, grupos distintos dos demais. Pode-se observar, também, que o estado de GO e MS possuem a menor produção de grãos, seguidos de SC, DF e SP. Os demais estados não foram representados no dendograma, devido ao fato de exercerem outras atividades econômicas. Pode-se dizer, também, que GO e MS são os estados que possuem a maior semelhança no dendograma, por ter sido o primeiro grupo formado, ao contrário do PR que foi o ultimo grupo a ser formado, mantendo-se distinto dos demais. Esses três estados foram os mais distintos no dendograma.

4.2 Aplicação da análise fatorial - *AF* - e análise de componentes principais – *ACP*

Neste exemplo serão apresentados alguns princípios gerais de interpretação dos resultados numéricos, e gráficos da *AF* com *ACP*.

A amostra utilizada, para este trabalho, refere-se a 30 coletas da fauna edáfica do solo. As coletas foram realizadas na área experimental do Departamento de Solos, em uma área de campo nativo da UFSM/RS. O período, no qual os dados foram coletados, é de 06 de junho de 2004 a 04 de janeiro de 2005, com coleta semanal, sendo que essa técnica possibilitará verificar a influência das variáveis suplementares: temperatura e umidade, sobre a quantidade e diversidade de organismos existentes no solo.

Para realizar a *ACP,* faz-se necessário o auxílio de um *software*, pois a amostra em estudo possui a dimensão R^{15}, ou seja, tem-se 15 variáveis.

Essas variáveis suplementares são utilizadas quando o pesquisador busca identificar o comportamento destas, em relação às demais variáveis.

Descrição das variáveis envolvidas neste estudo:

V_1 = Colêmbolos V_2 = Isópteros

V_3 = Hymenópteros V_4 = Hemípteros

V_5 = Dípteros V_6 = Coleópteros

V_7 = Aranae V_8 = Diplópodes

V_9 = Chilópodas V_{10} = Crustáceos

V_{11} = Ácaros V_{12} = Anelídeos

V_{13} = Moluscos V_{14} = Umidade (H_2O)

V_{15} = Temperatura

A Figura 32 mostra o banco de dados com as variáveis 15 representadas nas colunas, e as 32 coletas que representam os objetos nas linhas.

	1 COLEM.	2 ISOP.	3 HYMENOF	4 HEMIP.	5 DIP.	6 COLEOP.	7 ARANAE	8 DIPLOP.	9 CHILOP.	10 CRUSTACE	11 ÁCAROS	12 ANELID.	13 MOLUSC.	14 H2O	15 Temp
C1	5,5	0	0,5	0,25	0,75	2,5	0,25	0	0,25	0,75	4,75	2	0	13,02	15,5
C2	4	0,25	0,75	0	0	0,5	0,25	0,5	0,75	0,25	2,5	7,5	0,5	14,17	12,5
C3	1	0	2,25	0	0,25	1	0	0	0	0	3	0,25	0	11,56	16
C4	1,25	0,25	0,25	0	0,25	0,5	0,75	0	0	0	0	0,5	0	14,4	18,15
C5	0,25	0	0,5	0	0,25	0,75	0	0	0	0	1	0,5	0	15,19	14,9
C6	1	0,5	1	0	0	0,5	0	0	0	0	1	0,5	0,25	14,17	11
C7	2	1	1,5	0	0	0,25	0	0	0	0	0,75	1,5	0	14,37	12,8
C8	1,25	3	7	0	0,25	0,75	0	0,5	0	0	0	1,5	0	12,25	14,55
C9	1	0,5	7	0	0,25	0,25	0,25	0,5	0,25	0	0,75	5,5	0	13,64	18,1
C10	1,75	0	0	0	0,25	0,25	0	0	0	0	0,5	4,75	0	14,56	18,1
C11	0,25	0	3,25	0	0	0	0,75	0	0	0	0,25	2,5	0,25	18,45	15,9
C12	0,75	0	0	0	0	0,25	0	0	0,25	0,25	1	3,5	0	14,19	15,2
C13	1,75	0	0	0	0	0	0	0	3	0,5	1,5	7	0,25	10,78	15,95
C14	2	0	1,25	0	0	0,25	0,5	0	0	0	1	4,75	0,25	9,47	15,9
C15	0,5	0	1	0	0	0,25	0	0	1,75	0	0	2,5	0	14,36	15,9
C16	1,5	0	0,5	0	1,25	0,25	0,25	0	0	0	1	3,75	0	14,77	16,9
C17	0,5	0,75	11,75	0	0	0,25	0	0	0	0	1	2,25	0	12,25	17,9
C18	0,5	0	4,5	0	0,5	0	0,25	0	0	0	0,75	0,25	0	9,8	18,2
C19	0,25	0	0	0	0,25	0,25	0	0	0	0	0,5	1,75	0	11,24	18,9
C20	0		0,25	0	0	0	0,25	0	0,25	0	0	1,25	0	13,75	19,3
C21	0	0	6,25	0	0	0	0,25	0	0	0	0,25	1	0	12,78	18
C22	0,75	0	2,75	0	0,5	0	0,25	0	0	0	0,5	3,5	0	15,11	21,2
C23	2,5	0	0,75	0	0	0,25	0	0	0	0	0,25	0,75	0	16,27	19,85
C24	0,25	0	9	0	0,5	0,25	0	0	0	0	0	3	0	12,67	21,1
C25	1	0	5	0	0,5	0	0	0	0	0	0,5	3,75	0	6,86	24,1
C26	0,25	0	10	0	0	0	0	0	0	0	0	0,25	0	6,54	24,7
C27	0	0	2,5	0	0,25	0	0	0	0,5	0	0,25	2,25	0	12,64	23,5
C28	0	0	1	0	0,25	0	0	0	0	0	0	0	0,25	8,08	24,1
C29	0	0	13,5	0	0,25	0,25	0,5	0	0	0	0,5	1	0	10,07	24,7
C30	0	0	5	0	0,25	0,25	0	0	0,25	0	0,5	0,25	0	4,25	24,5

Figura 32 - Caixa de seleção das variáveis e os objetos, para *AF e ACP.*

Para a realização da análise, seleciona-se, no menu de opções o módulo principal do *STATISTICA*, a opção: *MultivariateExploratory Techniques – Factor Analysis*, conforme a janela mostrada na Figura 33.

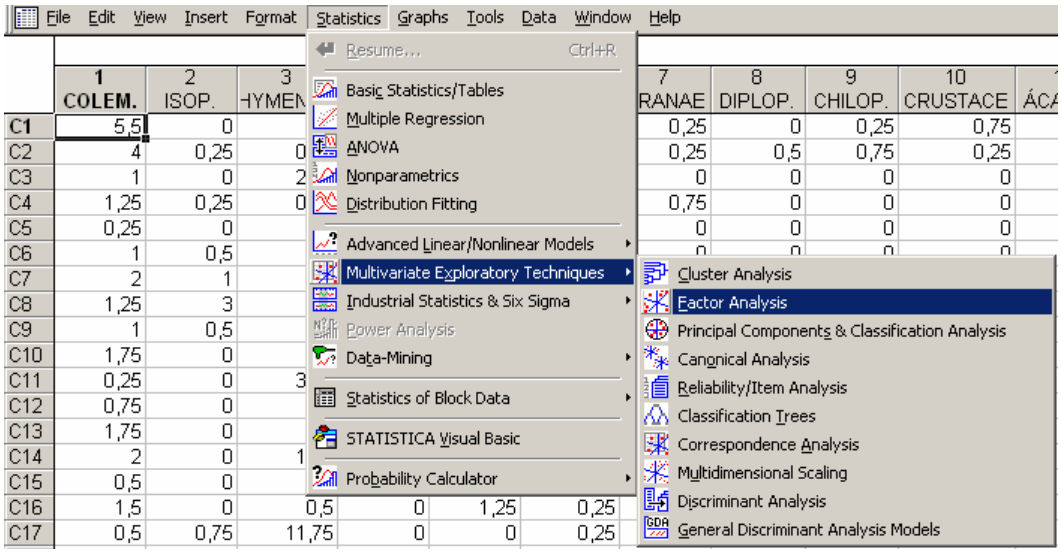


Figura 33 - Caixa de seleção da análise fatorial.

Na Figura 34, apresenta-se a janela na qual são apresentadas as variáveis para análise. Nessa janela, seleciona-se todas as variáveis clicando em *Select All*, isso se não houver variáveis suplementares para serem analisadas. Se houver variáveis suplementares, essas devem ser analisadas apenas no círculo unitário, o qual oferece a opção de análise para as mesmas. Deve-se proceder da seguinte forma: manter o *ctrl* pressionado e selecionar, apenas, as variáveis desejadas, com o *mouse.*

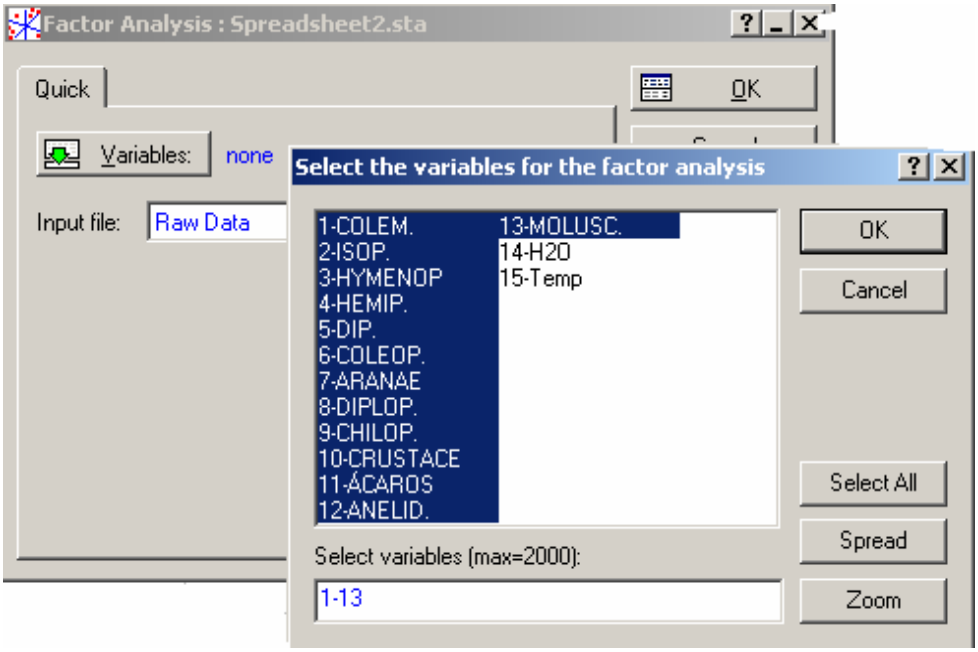


Figura 34 - Caixa de seleção das variáveis.

Na Figura 35, após selecionadas as variáveis, deve-se informar na opção da janela *input file*, se os dados são os originais, conforme coletados, seleciona-se, *Raw Data* e *Ok.*

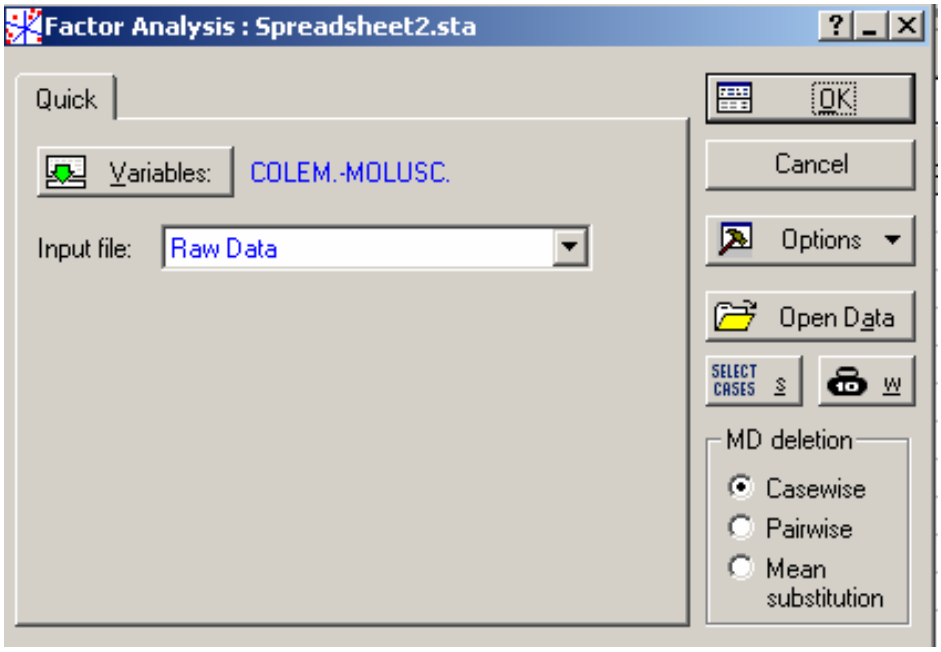


Figura 35 – Caixa de seleção para *ACP.*

Na Figura 36, determina-se o número de fatores que se deseja ter, na análise, da seguinte forma: coloca-se no *Maximum no. of factors* o número desejado. Neste caso, optou-se pelo número total de variáveis que é 13, pois não poderá haver número de fatores superior ao número de variáveis. Em *minimum* eingevalue, aconselha-se informar um valor bem baixo do tipo 0,001, pois, assim, obtém-se o maior número possível de autovalores, o que possibilita fazer uma investigação melhor do estudo, caso contrário pode-se informar um valor igual a 1 e obtém-se, então, somente os autovalores superiores a 1 e, desta forma, segue-se a regra de KAISER (1960, *apud* MARDIA, 1979).

Deve-se lembrar que nem sempre o pesquisador está interessado nas primeiras componentes, às vezes as componentes com menor grau de explicação são as mais estáveis, merecendo a devida atenção. Realizado isso, clica-se em *Ok*.

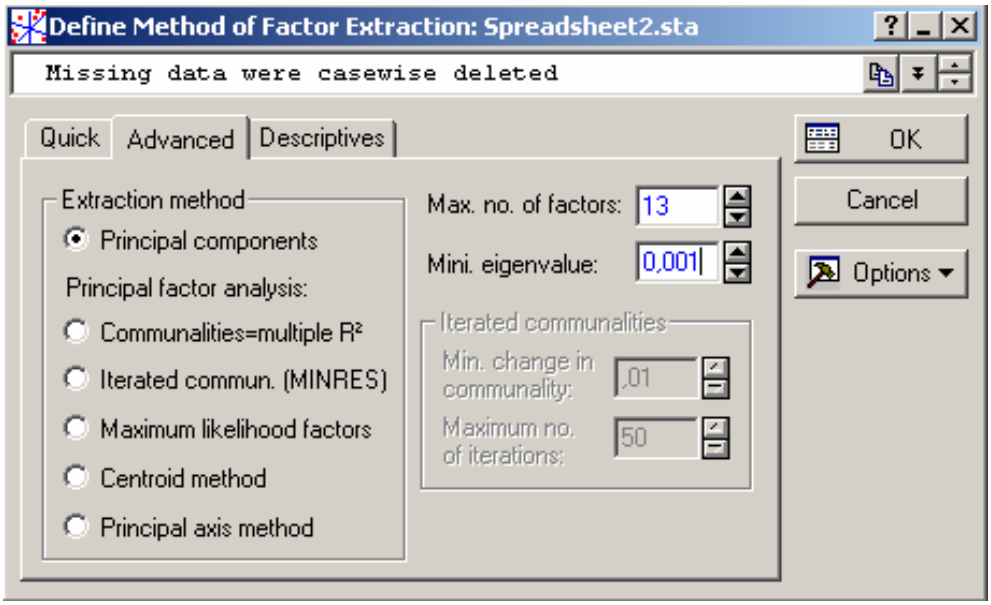


Figura 36 - Janela de seleção do número de fatores, para AF e ACP.

A Figura 37 mostra a caixa de seleção de comandos para a extração dos autovalores seleciona-se *Explained variance/Eigenvalues.* Nesta janela tem-se a opção de verificar o método gráfico *Scree plot,* que representa, graficamente, a porcentagem de variação explicada pela componente nas ordenadas e os autovalores, em ordem decrescente, nas abscissas, sugerido por CATTEL (1966) e exemplificado por PLA (1986), as comunalidades, a proporção de contribuição de cada variável *factor loadings* e outros valores de interesse.

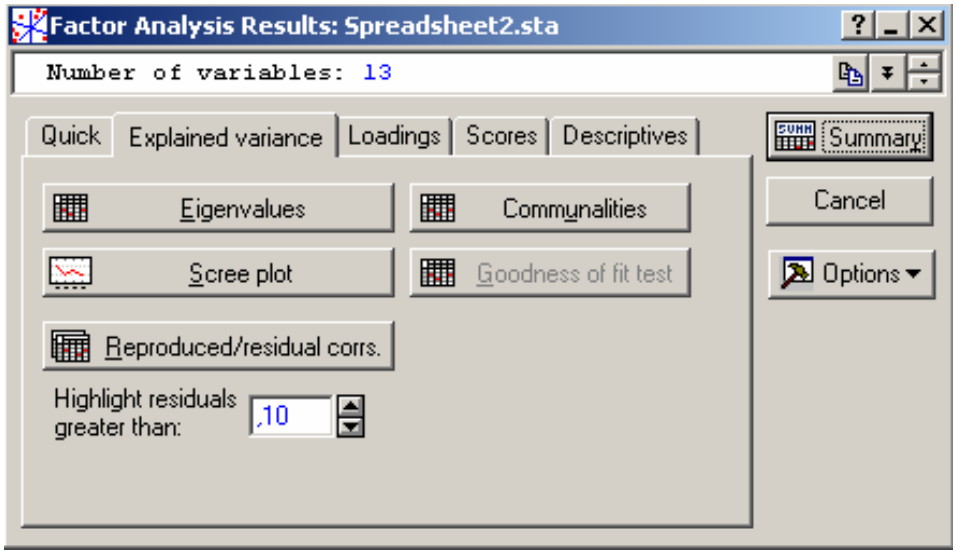


Figura 37- Caixa de seleção para extração dos autovalores.

Na Tabela 23 apresenta-se o resultado dos autovalores, bem como a porcentagem de variância explicada por cada componente, e também a variância acumulada pelas mesmas.

Numa análise fatorial, considerando-se 13 variáveis, poder-se-ia ter 13 fatores que corresponderiam às variáveis originais. A escolha do número de fatores pode levar em conta diferentes critérios. Um deles está em incluir, na análise, aquelas componentes que conseguem sintetizar uma variância acumulada em torno de 70%. Como se pode observar, na Tabela 23, quatro primeiros autovalores representam cerca de 74,31% da variância. Portanto, os dados serão resumidos pelas quatro primeiras componentes principais. Pode-se, também, fazer, esta seleção, incluindo-se somente aquelas componentes cujos valores próprios são superiores a 1. Neste caso, são quatro autovalores, este critério foi sugerido por KAISER (1960) apud MARDIA (1979).

Tabela 23 – Autovalores e percentual da variância explicada de cada componente.

Número de componentes	Autovalores			
	Extração dos componentes principais			
	Autovalores	% da variância explicada	Autovalores acumulados	% da variância explicada acumulada
1	**4,30**	33,05	4,30	**33,05**
2	**2,35**	18,10	6,65	**51,15**
3	**1,78**	13,66	8,43	**64,82**
4	**1,23**	9,49	9,66	**74,31**
5	0,94	7,27	10,60	81,58
6	0,83	6,42	11,44	87,99
7	0,52	3,98	11,96	91,97
8	0,35	2,66	12,30	94,63
9	0,26	1,99	12,56	96,62
10	0,19	1,43	12,75	98,05
11	0,13	0,99	12,88	99,04
12	0,09	0,66	12,96	99,70
13	0,04	0,30	13,00	100,00

Olhando para a Tabela 23, pode-se observar que os quatro primeiros fatores possuem autovalores, que correspondem a 33,05%, 18,10%, 13,66%, e 9,49% da variância total, explicada pelos autovalores do modelo, ou seja, explicam juntos 74,31% das variações das medidas originais. Decidindo-se por estes quatro fatores, o pesquisador sabe qual o nível de explicação está conseguindo de seus dados, e decide se vale a pena a síntese fornecida por essa redução de dimensionalidade, ou se deve considerar todas as variáveis. Conforme Pereira (2001), “essa é uma medida de ajuste do modelo à análise de dados: no exemplo, o modelo com quatro fatores terá 74,31% de representação real”.

A Figura 38 mostra a seleção dos componentes principais através do método gráfico *Scree Plot*, sendo que a porcentagem de variação explicada pela componente está no eixo das ordenadas, e os autovalores estão representados em ordem decrescente no eixo das abscissas. Como se pode observar, na Figura 35, as quatro primeiras componentes explicam 74,31% da variância total, havendo uma estabilização do gráfico após a quinta componente, sendo consideradas as quatro primeiras. Pode-se observar, também, que as outras componentes apresentam uma baixa explicação, não sendo aconselhável inclui-las na análise.

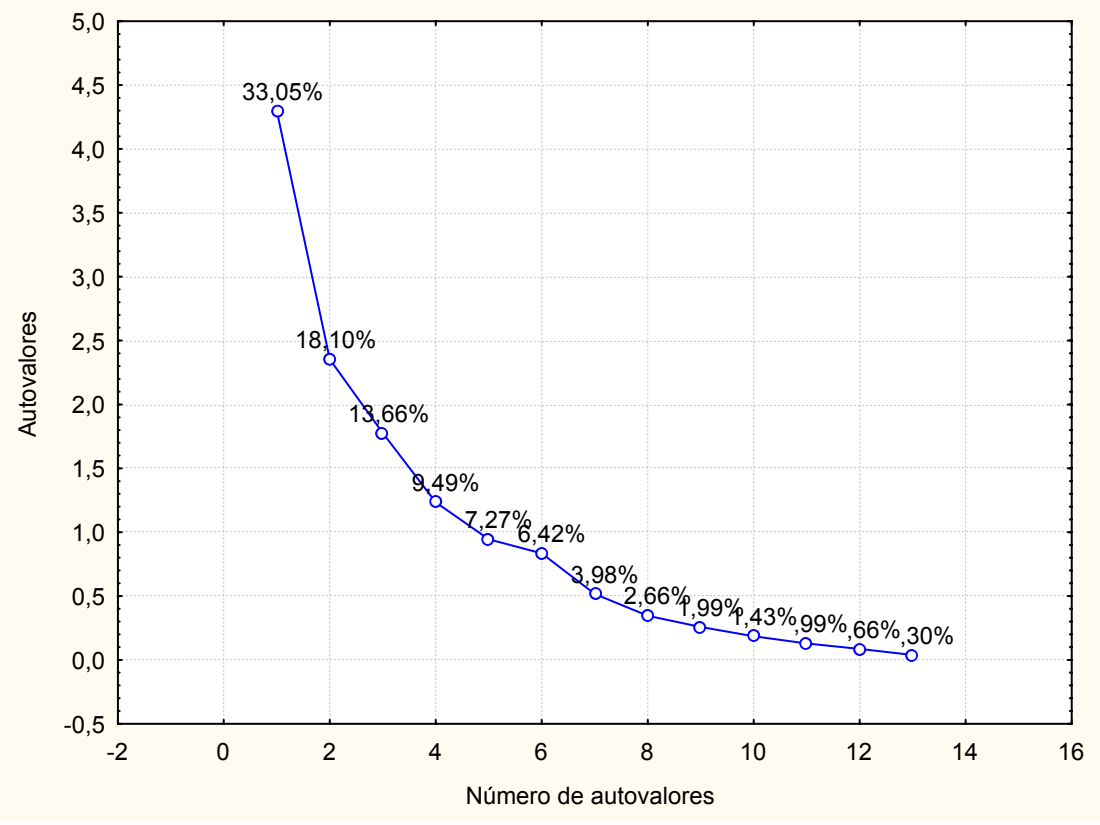


Figura 38- Gráfico de explicação da proporção de variação de cada componente principal.

A Figura 39 mostra a caixa de seleção e comandos das análises estatísticas que possam ser de interesse do pesquisador. Vale lembrar, aqui, que essas estatísticas são referentes às variáveis originais, e não aos valores derivados das componentes principais.

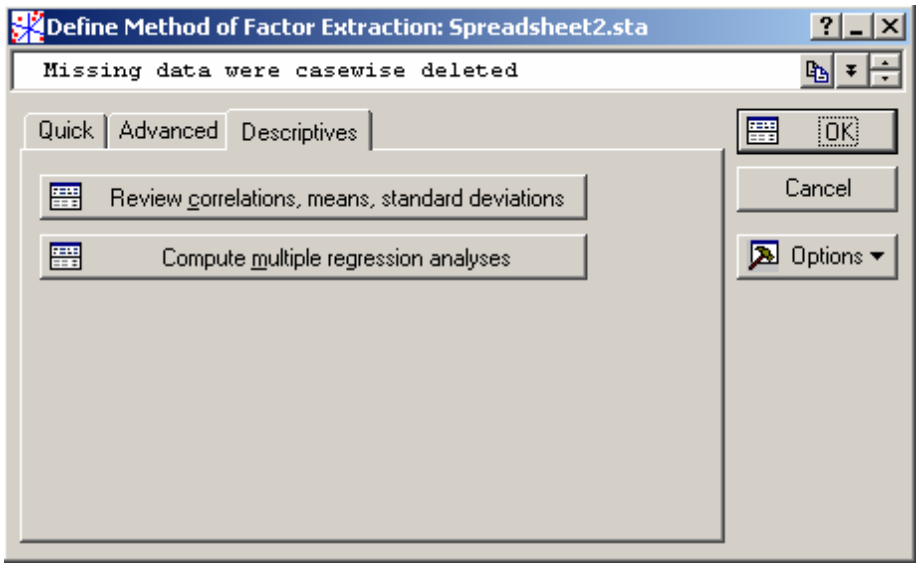


Figura 39 - Caixa de seleção das análises estatísticas.

A Figura 40 mostra uma caixa de seleção na qual mais ferramentas estatísticas são disponibilizadas, para se fazer uma análise complementar a *AF* e *ACP*.

Como a *AF* e a *ACP* são técnicas exploratórias de dados, é importante que se realize uma estatística descritiva nas variáveis, para que haja uma melhor compreensão nos resultados obtidos.

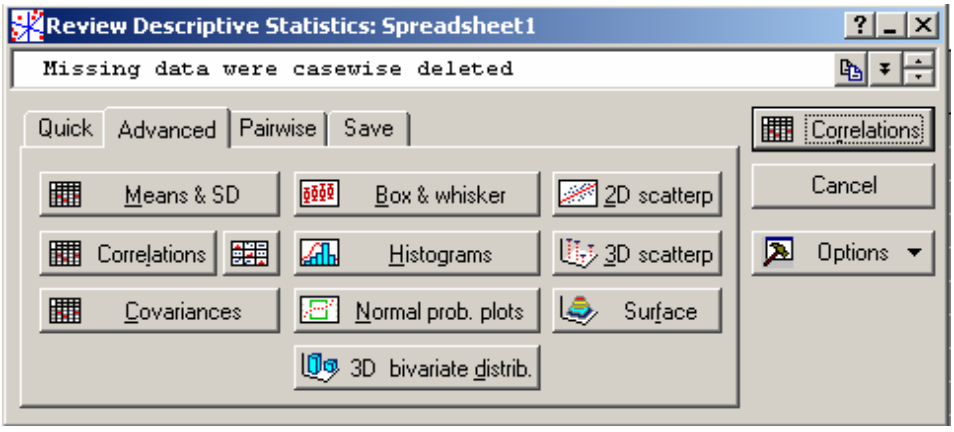


Figura 40 - Caixa de comandos para análise descritiva dos dados.

A Figura 41 mostra a média e o desvio padrão de cada uma das variáveis originais, que se obtém selecionando-se *Means & SD* na Figura 40.

Variáveis	Means and Standard Deviations Casewise deletion of MD N=29	
	Média	Desvio padrão
COLEM.	1,09	1,25
ISOP.	0,22	0,59
HYMENOP	3,41	3,83
HEMIP.	0,01	0,05
DIP.	0,23	0,28
COLEOP.	0,34	0,49
ARANAE	0,15	0,23
DIPLOP.	0,05	0,15
CHILOP.	0,24	0,64
CRUSTACE	0,06	0,17
ÁCAROS	0,83	1,03
ANELID.	2,36	2,05
MOLUSC.	0,06	0,13

Figura 41 - Caixa de resultados da estatística descritiva.

Na Figura 42, apresenta-se o resultado da matriz de correlação entre as variáveis, a qual é obtida selecionando-se, *Advanced*/*Correlations,* conforme Figura 40.

Variáveis	Correlação												
	COLEM.	ISOP.	HYMENOP	HEMIP.	DIP.	COLEOP.	ARANAE	DIPLOP.	CHILOP.	CRUSTACE	ACAROS	ANELID.	MOLUSC.
COLEM.	1,00	0,08	-0,40	**0,68**	0,16	**0,68**	0,10	0,27	0,15	**0,71**	**0,74**	0,40	0,30
ISOP.		1,00	0,22	-0,07	-0,11	0,15	-0,14	**0,60**	-0,11	-0,11	-0,13	-0,06	-0,06
HYMENOP			1,00	-0,15	-0,05	-0,20	0,04	0,14	-0,25	-0,29	-0,27	-0,19	-0,29
HEMIP.				1,00	0,35	**0,85**	0,09	-0,06	0,00	**0,77**	**0,73**	-0,03	-0,09
DIP.					1,00	0,29	0,08	-0,08	-0,24	0,11	0,23	0,00	-0,34
COLEOP.						1,00	0,02	0,12	-0,08	**0,63**	**0,80**	-0,11	-0,09
ARANAE							1,00	0,03	-0,18	-0,01	0,01	0,07	0,22
DIPLOP.								1,00	0,05	0,05	0,09	0,42	0,29
CHILOP.									1,00	**0,49**	0,13	**0,50**	0,31
CRUSTACE										1,00	**0,75**	0,36	0,23
ACAROS											1,00	0,21	0,20
ANELID.												1,00	0,45
MOLUSC.													1,00

Figura 42 - Caixa de resultados da matriz de correlação.

Com a matriz de correlação, da Figura 42, é possível observar que existe um número representativo de valores superiores a 0,7, o que significa que a

correlação entre as variáveis é boa. Sendo assim, pode-se concluir que as variáveis estão interligadas umas com as outras. Isso mostra que o estudo das variáveis não deve ser feito de forma isolada, mas, sim, de maneira conjunta, com a utilização de uma técnica adequada, neste estudo a *ACP.*

A Figura 43 mostra a caixa de seleção de comandos para *ACP,* seleciona-se: *Scores/Factor score coefficients,* para extrair os autovetores, que definam a direção dos eixos, para *AF* e *ACP*.

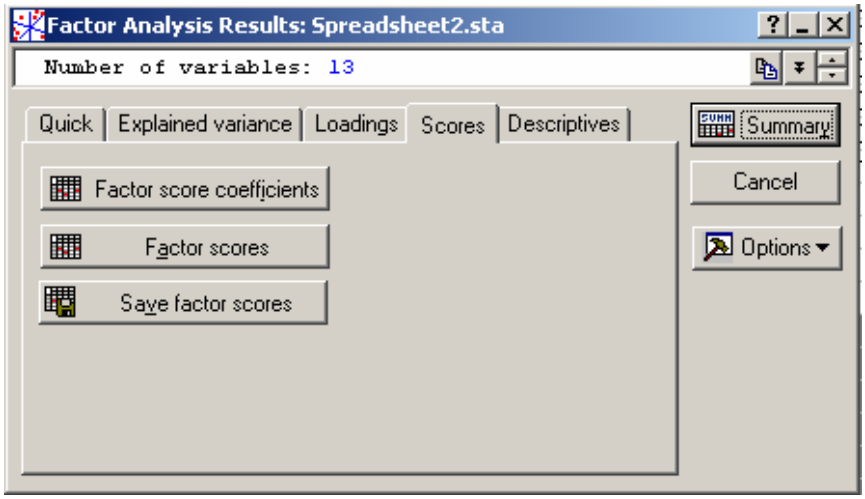


Figura 43 - Caixa de seleção dos autovetores.

Na Figura 44, são apresentados os resultados dos *factor Score coefficientes* (autovetores), que definem a direção dos eixos para *ACP*.

Factor Score Coefficients (Sattistica ACP e AF)
Rotation: Unrotated
Extraction: Principal components

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10	Factor 11	Factor 12	Factor 13
COLEM.	-0,21	0,04	-0,08	-0,08	0,04	-0,18	0,06	-0,62	0,70	-0,14	1,90	0,49	0,46
ISOP.	0,01	0,03	-0,49	0,17	0,09	-0,15	-0,55	0,27	0,66	-0,89	-0,54	0,09	-0,70
HYMENOP	0,10	-0,09	-0,24	-0,04	-0,11	0,86	0,46	0,14	0,31	-0,23	0,57	-0,20	0,51
HEMIP.	-0,19	-0,18	-0,01	0,01	0,07	0,21	-0,13	-0,02	0,46	1,03	-0,36	-1,03	-3,02
DIP.	-0,06	-0,22	0,02	-0,16	-0,75	-0,25	-0,03	0,82	0,17	-0,07	0,31	0,11	0,27
COLEOP.	-0,19	-0,18	-0,13	0,05	0,19	-0,03	-0,06	0,07	-0,60	-0,22	-0,09	-1,93	2,67
ARANAE	-0,02	0,01	0,02	-0,73	0,09	0,25	-0,67	-0,10	-0,28	-0,30	-0,05	0,24	-0,11
DIPLOP.	-0,04	0,19	-0,45	-0,07	-0,16	-0,08	0,12	0,01	-0,94	1,15	0,20	0,57	-0,23
CHILOP.	-0,07	0,27	0,17	0,34	-0,15	0,33	-0,63	0,43	-0,47	-0,22	1,03	-0,50	-0,63
CRUSTACE	-0,21	0,04	0,07	0,13	-0,01	0,31	-0,19	0,14	0,39	0,48	-1,03	1,79	2,15
ACAROS	-0,21	-0,05	0,01	0,01	0,08	0,06	0,51	0,13	-0,75	-1,23	-0,26	0,99	-1,91
ANELID.	-0,08	0,31	-0,00	-0,09	-0,52	0,06	0,20	-0,69	0,26	-0,43	-1,01	-1,17	0,05
MOLUSC.	-0,06	0,31	0,03	-0,26	0,32	-0,11	0,44	1,01	0,52	0,13	0,02	-0,63	0,17

Figura 44 - Caixa de resultados dos autovetores.

No exemplo, que segue, é mostrado o cálculo manual das componentes principais:

CP_1 = (Autovetor 11)(Variável 11) + (Autovetor 21)(Variável 12) + (Autovetor 31)(Variável 13) + ... + (Autovetor 131)(Variável 113)

CP_{11} = (-0,21)(5,5) + (0,01)(0) + (0,10)(0,5) + (-0,19)(0,25) + (-0,06)(0,75) + (-0,19)(2,5) + (-0,02)(0,25) + (-0,04)(0) + (-0,07)(0,25) + (-0,21)(0,75) + (-0,21)(4,75) + (-0,08)(2) + (-0,06)(0)

CP_{11} = -3,01

CP_{12} = (-0,21)(4) + (0,25)(0) + (0,10)(0,75) + (-0,19)(0) + (-0,06)(0) + (-0,19)(0,5) + (-0,02)(0,25) + (-0,04)(0,5) + (-0,07)(0,75) + (-0,21)(0,25) + (-0,21)(2,5) + (-0,08)(7,5) + (-0,06)(0,5)

CP_{12} = -2,15

Como pode-se observar, o valor da primeira componente principal, realizando-se os cálculos de forma manual, é -3,01, e o valor encontrado pelo *software* é de -4,35, conforme Figura 50. Isso ocorre devido à transformação realizada pelo programa ao rodar os dados, ou seja, o valor das componentes principais, encontradas de forma manual, não será o mesmo que o fornecido pela análise.

Para encontrar os componentes principais, através do *software,* deve-se selecionar a opção do programa *statistica*, referente a esta análise. Para isso seleciona-se: *Multivariate Exploratory Techniques – Principal Components & Classification Analysis*, conforme a Figura 45:

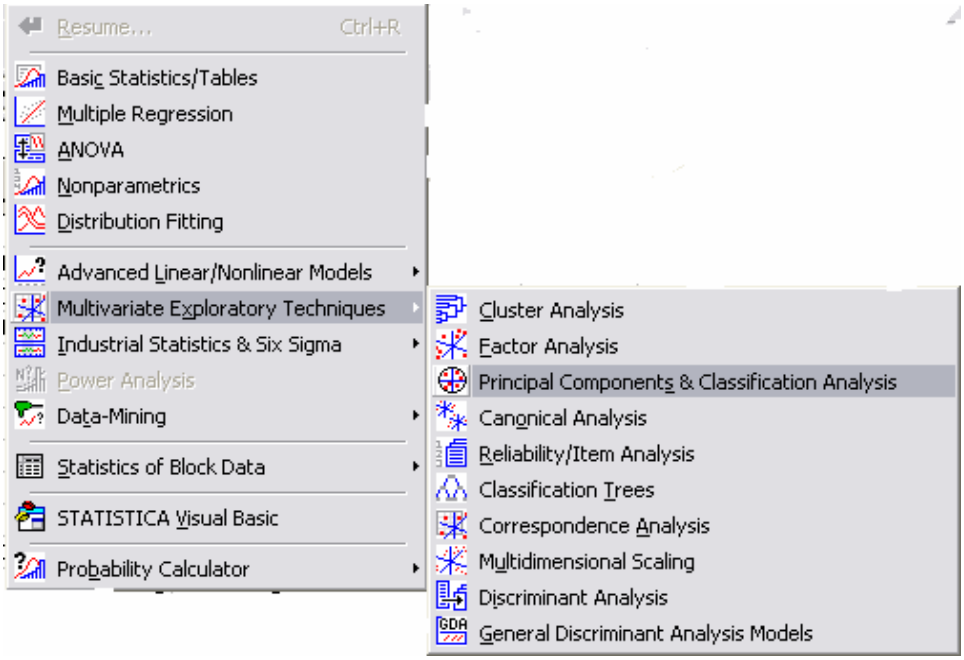


Figura 45 - Caixa de seleção da *ACP.*

A Figura 46 mostra a caixa de seleção de variáveis e comandos para *ACP.* Clica-se em *Variables* e o programa mostrará todas as variáveis, e é só clicar em *Ok.*

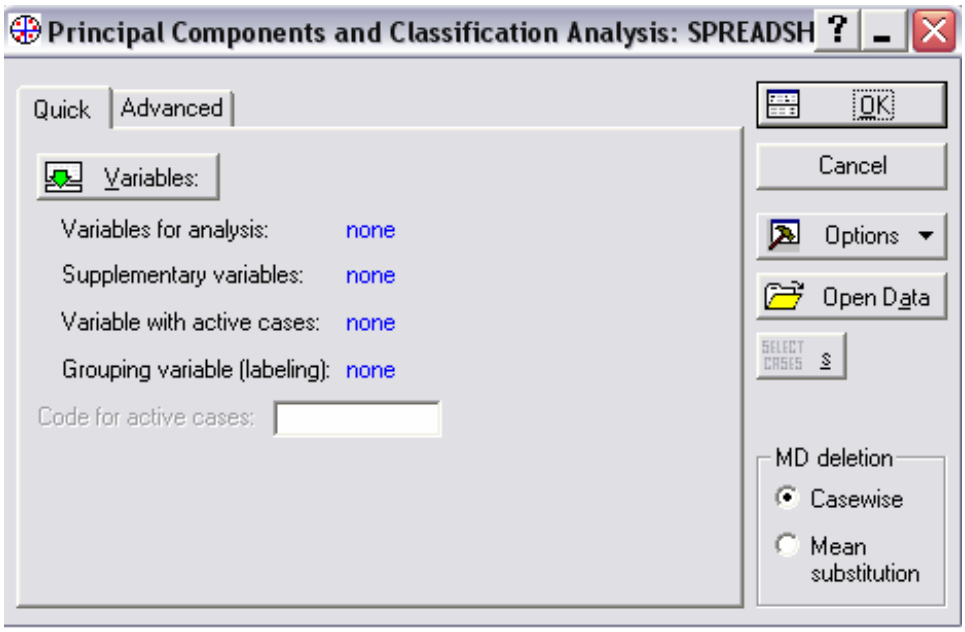


Figura 46 - Caixa de seleção da *ACP.*

Na Figura 47, apresenta-se a totalidade de variáveis para análise. Neste caso, após selecionadas as variáveis, clica-se em *Ok*.

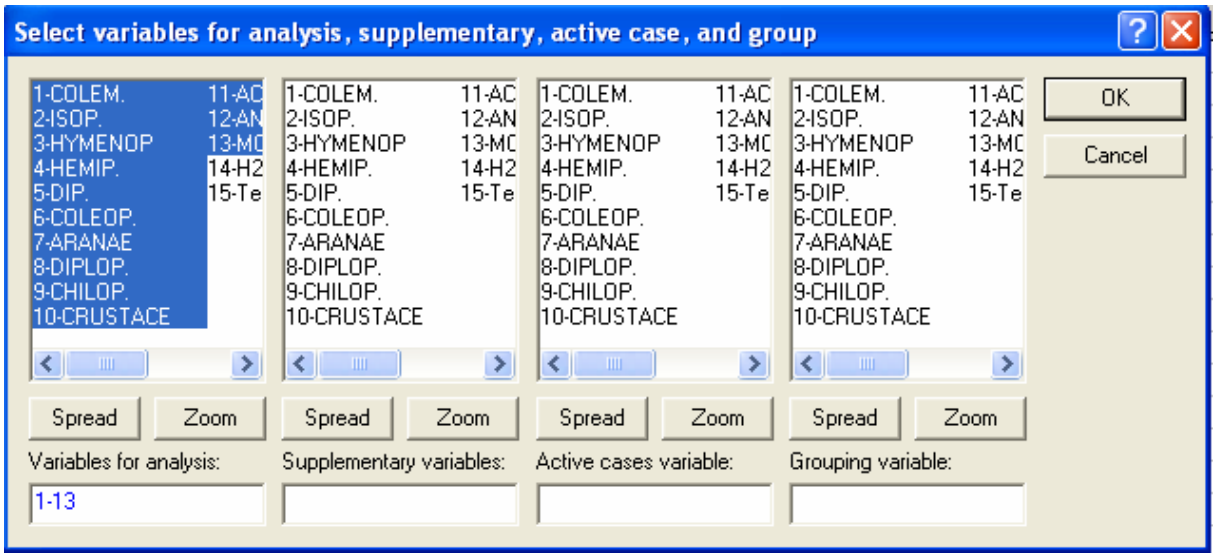


Figura 47 - Caixa de seleção das variáveis para *ACP.*

A Figura 48, na opção *Variables for analysis:* mostra que todas as variáveis foram selecionadas, não existindo variáveis suplementares para o estudo, basta clicar em *Ok*.

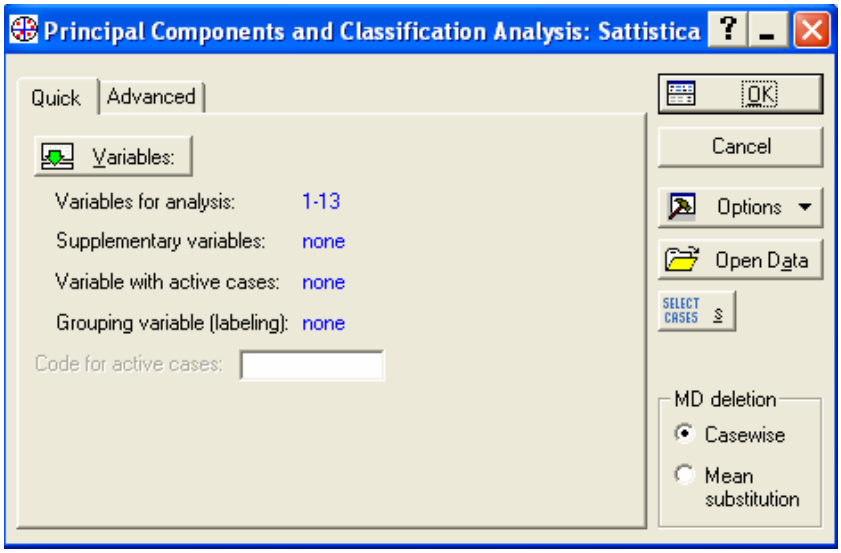


Figura 48 - Caixa de seleção da *ACP*.

A Figura 49 mostra a caixa de seleção para encontrar os componentes principais, seleciona-se *Cases/Factor scores,* e clica-se em *Ok.*

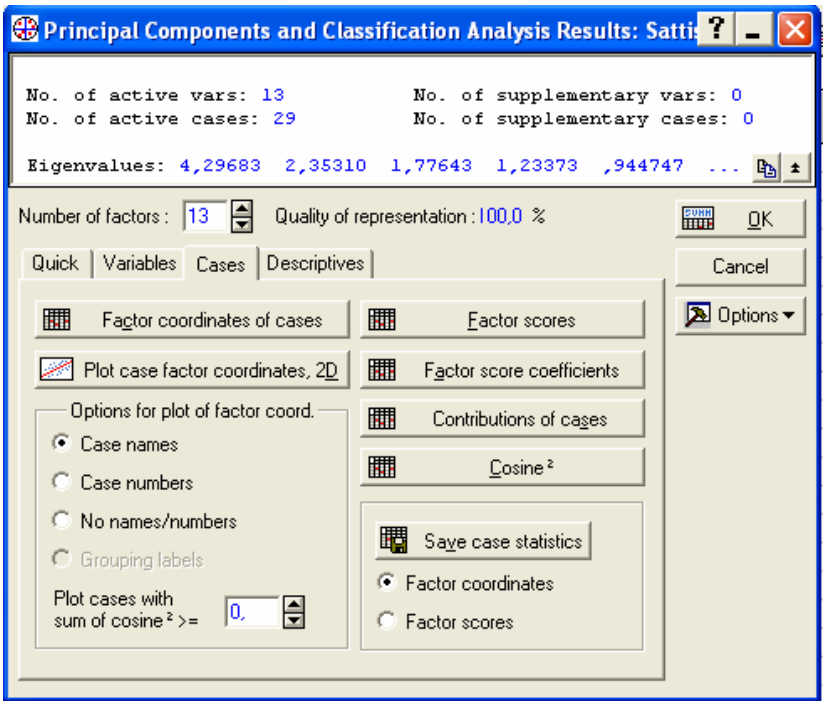


Figura 49 - Caixa de seleção dos componentes principais.

A Figura 50 refere-se aos componentes principais encontrados na análise. É importante observar que, pelo fato de existir 13 variáveis, foram encontrados 13 componentes, mas pela análise fatorial, seguindo o critério sugerido por KAISER (1960) apud MARDIA (1979), deve-se considerar apenas as primeiras quatro componentes principais.

	Componentes Principais												
Case	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Fact.10	Fact.11	Fact.12	Fact.13
C1	-4,35	-2,24	-0,05	0,05	0,32	0,92	-0,34	-0,04	0,61	1,00	-0,24	-0,46	-0,60
C2	-1,68	2,85	-1,08	-1,23	0,24	-0,54	2,00	0,29	-0,53	0,49	1,14	0,59	0,64
C3	-0,45	-0,96	0,14	0,56	0,80	-0,54	1,35	0,42	-2,84	-2,57	0,25	0,37	-0,25
C4	0,20	-0,50	0,20	-1,86	0,72	-0,40	-2,72	-0,56	-0,69	-0,10	0,68	0,37	1,76
C5	0,12	-0,71	0,35	0,57	0,52	-0,91	0,17	0,35	-1,61	0,06	-0,72	-0,91	1,59
C6	0,05	0,21	-0,04	0,27	1,79	-1,02	0,73	1,45	0,59	-0,38	-0,15	-0,91	0,07
C7	0,14	-0,09	-0,55	0,79	0,89	-0,93	-0,49	-1,20	1,34	-1,33	0,58	0,94	-1,30
C8	0,23	0,10	-4,09	1,03	-0,02	-0,67	-1,82	1,00	0,62	-0,25	-0,55	-0,08	0,43
C9	0,05	0,83	-1,76	-0,56	-1,52	0,60	0,35	-1,26	-2,21	1,61	-0,36	-0,08	-1,14
C10	-0,01	0,19	0,41	0,25	-0,73	-1,12	0,42	-1,98	0,72	-0,14	-0,37	-1,23	0,36
C11	0,44	0,65	0,47	-2,40	1,17	0,59	-0,90	0,49	0,27	0,03	-1,36	-0,48	-0,98
C12	-0,17	0,31	0,63	0,83	0,19	-0,18	0,00	-1,34	-0,14	0,30	-2,79	1,87	1,67
C13	-1,19	2,75	1,52	1,74	-0,75	1,48	-1,45	1,06	0,39	-1,04	-0,11	0,94	0,55
C14	-0,23	0,95	0,40	-1,76	0,77	-0,30	0,25	-0,96	0,84	-1,19	-0,28	-1,51	-0,46
C15	0,28	0,74	0,87	1,47	-0,02	0,33	-1,69	-0,17	-1,26	0,46	1,64	-2,45	-1,06
C16	-0,25	-0,77	0,47	-1,06	-3,01	-1,58	-0,25	1,34	0,43	-1,11	0,74	0,37	0,23
C17	0,56	-0,30	-0,88	0,68	0,33	1,69	1,10	-0,40	0,96	-1,85	-0,38	-0,41	-0,64
C18	0,46	-0,73	0,31	-0,49	-0,41	0,03	-0,16	0,93	-0,22	0,08	0,81	2,21	-1,31
C19	0,36	-0,30	0,51	0,47	-0,02	-0,99	0,05	-0,23	-0,51	0,65	-1,19	-0,11	-0,26
C21	0,77	-0,27	0,20	-0,23	0,62	0,93	-0,09	-0,52	-0,21	0,60	-0,48	0,81	-0,80
C22	0,30	-0,19	0,40	-0,63	-1,19	-0,32	-0,16	-0,39	0,37	-0,22	-0,60	0,31	-0,92
C23	0,14	-0,19	0,30	0,51	0,92	-0,96	0,06	-1,73	0,73	0,92	2,65	0,97	0,87
C24	0,58	-0,50	-0,05	0,19	-1,30	0,82	0,97	0,35	0,90	0,39	-0,05	-1,66	2,14
C25	0,32	-0,20	0,22	0,13	-1,41	-0,12	0,88	-0,40	1,04	-0,10	0,05	-0,12	-0,40
C26	0,92	-0,46	-0,07	0,56	0,58	1,43	0,92	-0,17	0,63	1,13	0,95	0,64	0,36
C27	0,54	0,02	0,56	0,68	-0,46	-0,11	-0,20	0,09	-0,26	0,66	-0,53	-0,27	-1,40
C28	0,57	0,13	0,59	0,02	0,87	-1,00	0,64	2,41	0,91	1,94	-0,35	0,02	-0,36
C29	0,73	-0,73	-0,28	-1,22	-0,04	2,61	0,10	0,42	-0,27	-0,64	0,71	0,06	1,19
C30	0,56	-0,55	0,28	0,64	0,14	0,26	0,24	0,76	-0,61	0,60	0,32	0,20	0,04

Figura 50 - Componentes principais, referente às treze variáveis.

Quando os dados estiverem dispostos em unidades de medidas diferentes, deve-se eliminar a influência que uma variável poderá causar sobre a outra na formação das componentes. Deve-se fazer a padronização dos dados.

Para que seja possível padronizar os dados, utilizando-se o *software statistica,* deve-se selecionar o banco de dados inicial, conforme a Figura 51.

	1 COLEM.	2 ISOP.	3 IYMENOI	4 HEMIP.	5 DIP.	6 COLEOP	7 ARANAE	8 DIPLOP.	9 CHILOP.	10 RUSTAC	11 ÁCAROS	12 ANELID.	13 MOLUSC	14 H2O	15 Temp
C1	5,5	0	0,5	0,25	0,75	2,5	0,25	0	0,25	0,75	4,75	2	0	13,02	15,5
C2	4	0,25	0,75	0	0	0,5	0,25	0,5	0,75	0,25	2,5	7,5	0,5	14,17	12,5
C3	1	0	2,25	0	0,25	1	0	0	0	0	3	0,25	0	11,56	16
C4	1,25	0,25	0,25	0	0,25	0,5	0,75	0	0	0	0	0,5	0	14,4	18,15
C5	0,25	0	0,5	0	0,25	0,75	0	0	0	0	1	0,5	0	15,19	14,9
C6	1	0,5	1	0	0	0,5	0	0	0	0	1	0,5	0,25	14,17	11
C7	2	1	1,5	0	0	0,25	0	0	0	0	0,75	1,5	0	14,37	12,8
C8	1,25	3	7	0	0,25	0,75	0	0,5	0	0	0	1,5	0	12,25	14,55
C9	1	0,5	7	0	0,25	0,25	0,25	0,5	0,25	0	0,75	5,5	0	13,64	18,1
C10	1,75	0	0	0	0,25	0,25	0	0	0	0	0,5	4,75	0	14,56	18,1
C11	0,25	0	3,25	0	0	0	0,75	0	0	0	0,25	2,5	0,25	18,45	15,9
C12	0,75	0	0	0	0	0,25	0	0	0,25	0,25	1	3,5	0	14,19	15,2
C13	1,75	0	0	0	0	0	0	0	3	0,5	1,5	7	0,25	10,78	15,95
C14	2	0	1,25	0	0	0,25	0,5	0	0	0	1	4,75	0,25	9,47	15,9
C15	0,5	0	1	0	0	0,25	0	0	1,75	0	0	2,5	0	14,36	15,9
C16	1,5	0	0,5	0	1,25	0,25	0,25	0	0	0	1	3,75	0	14,77	16,9
C17	0,5	0,75	11,75	0	0	0,25	0	0	0	0	1	2,25	0	12,25	17,9
C18	0,5	0	4,5	0	0,5	0	0,25	0	0	0	0,75	0,25	0	9,8	18,2
C19	0,25	0	0	0	0,25	0,25	0	0	0	0	0,5	1,75	0	11,24	18,9
C20	0	0	0,25	0	0	0	0,25	0	0,25	0	0	1,25	0	13,75	19,3
C21	0	0	6,25	0	0	0	0,25	0	0	0	0,25	1	0	12,78	18
C22	0,75	0	2,75	0	0,5	0	0,25	0	0	0	0,5	3,5	0	15,11	21,2
C23	2,5	0	0,75	0	0	0,25	0	0	0	0	0,25	0,75	0	16,27	19,85
C24	0,25	0	9	0	0,5	0,25	0	0	0	0	0	3	0	12,67	21,1
C25	1	0	5	0	0,5	0	0	0	0	0	0,5	3,75	0	6,86	24,1
C26	0,25	0	10	0	0	0	0	0	0	0	0	0,25	0	6,54	24,7
C27	0	0	2,5	0	0,25	0	0	0	0,5	0	0,25	2,25	0	12,64	23,5
C28	0	0	1	0	0,25	0	0	0	0	0	0	0	0,25	8,08	24,1
C29	0	0	13,5	0	0,25	0,25	0,5	0	0	0	0,5	1	0	10,07	24,7
C30	0	0	5	0	0,25	0,25	0	0	0,25	0	0,5	0,25	0	4,25	24,5

Figura 51 - Seleção das variáveis para a padronização dos dados.

Logo após, clicar, com o botão auxiliar, no meio da tela, na qual estão as variáveis selecionadas. Abrirá a caixa de seleção da Figura 52, na qual existem duas opções de padronização: por colunas, sendo esta a utilizada neste trabalho, selecionando *Fill/Standardize Block/Standardize Columns,* ou por linhas, selecionando *Fill/Standardize Block/Standardize Rows.*

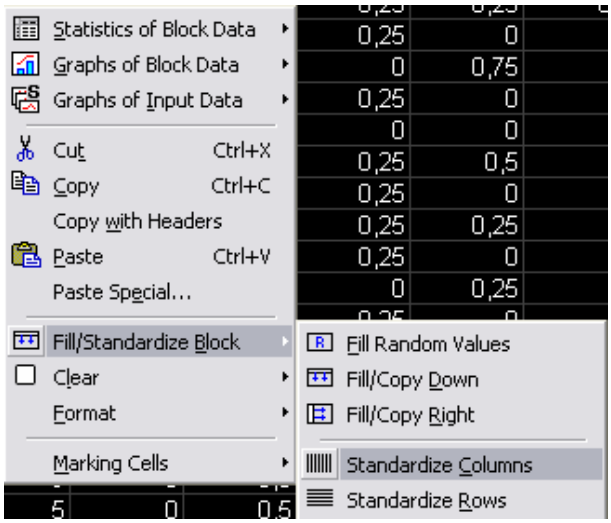


Figura 52 - Caixa de seleção para a padronização das variáveis.

A Figura 53 mostra as variáveis padronizadas.

	1 COLEM.	2 ISOP.	3 IYMENOI	4 HEMIP.	5 DIP.	6 COLEOP	7 ARANAE	8 DIPLOP.	9 CHILOP.	10 RUSTAC	11 ÁCAROS	12 ANELID.	13 MOLUSC	14 H2O	15 Temp
C1	3,58219	-0,3568	-0,7356	5,29465	1,86675	4,50198	0,44721	-0,3277	0,01327	4,07516	3,86434	-0,1602	-0,463	0,20398	-0,7172
C2	2,37244	0,07137	-0,6699	-0,1826	-0,8	0,36223	0,44721	2,94958	0,80951	1,12926	1,66313	2,55046	3,50524	0,57553	-1,5007
C3	-0,047	-0,3568	-0,2759	-0,1826	0,08889	1,39717	-0,6708	-0,3277	-0,3848	-0,3437	2,15229	-1,0226	-0,463	-0,2677	-0,5867
C4	0,15458	0,07137	-0,8013	-0,1826	0,08889	0,36223	2,68328	-0,3277	-0,3848	-0,3437	-0,7827	-0,8994	-0,463	0,64985	-0,0252
C5	-0,6519	-0,3568	-0,7356	-0,1826	0,08889	0,8797	-0,6708	-0,3277	-0,3848	-0,3437	0,19566	-0,8994	-0,463	0,90509	-0,8739
C6	-0,047	0,49956	-0,6042	-0,1826	-0,8	0,36223	-0,6708	-0,3277	-0,3848	-0,3437	0,19566	-0,8994	1,52114	0,57553	-1,8924
C7	0,75945	1,35595	-0,4729	-0,1826	-0,8	-0,1552	-0,6708	-0,3277	-0,3848	-0,3437	-0,0489	-0,4066	-0,463	0,64015	-1,4223
C8	0,15458	4,78151	0,97204	-0,1826	0,08889	0,8797	-0,6708	2,94958	-0,3848	-0,3437	-0,7827	-0,4066	-0,463	-0,0448	-0,9653
C9	-0,047	0,49956	0,97204	-0,1826	0,08889	-0,1552	0,44721	2,94958	0,01327	-0,3437	-0,0489	1,56478	-0,463	0,4043	-0,0383
C10	0,55783	-0,3568	-0,867	-0,1826	0,08889	-0,1552	-0,6708	-0,3277	-0,3848	-0,3437	-0,2935	1,19514	-0,463	0,70154	-0,0383
C11	-0,6519	-0,3568	-0,0131	-0,1826	-0,8	-0,6727	2,68328	-0,3277	-0,3848	-0,3437	-0,5381	0,08625	1,52114	1,95837	-0,6128
C12	-0,2487	-0,3568	-0,867	-0,1826	-0,8	-0,1552	-0,6708	-0,3277	0,01327	1,12926	0,19566	0,57909	-0,463	0,582	-0,7956
C13	0,55783	-0,3568	-0,867	-0,1826	-0,8	-0,6727	-0,6708	-0,3277	4,39258	2,60221	0,68482	2,30404	1,52114	-0,5197	-0,5997
C14	0,75945	-0,3568	-0,5386	-0,1826	-0,8	-0,1552	1,56525	-0,3277	-0,3848	-0,3437	0,19566	1,19514	1,52114	-0,943	-0,6128
C15	-0,4503	-0,3568	-0,6042	-0,1826	-0,8	-0,1552	-0,6708	-0,3277	2,40199	-0,3437	-0,7827	0,08625	-0,463	0,63692	-0,6128
C16	0,3562	-0,3568	-0,7356	-0,1826	3,6446	-0,1552	0,44721	-0,3277	-0,3848	-0,3437	0,19566	0,7023	-0,463	0,76939	-0,3517
C17	-0,4503	0,92776	2,21994	-0,1826	-0,8	-0,1552	-0,6708	-0,3277	-0,3848	-0,3437	0,19566	-0,037	-0,463	-0,0448	-0,0905
C18	-0,4503	-0,3568	0,31526	-0,1826	0,97782	-0,6727	0,44721	-0,3277	-0,3848	-0,3437	-0,0489	-1,0226	-0,463	-0,8364	-0,0122
C19	-0,6519	-0,3568	-0,867	-0,1826	0,08889	-0,1552	-0,6708	-0,3277	-0,3848	-0,3437	-0,2935	-0,2834	-0,463	-0,3711	0,17061
C20	-0,8535	-0,3568	-0,8013	-0,1826	-0,8	-0,6727	0,44721	-0,3277	0,01327	-0,3437	-0,7827	-0,5298	-0,463	0,43984	0,27506
C21	-0,8535	-0,3568	0,77501	-0,1826	-0,8	-0,6727	0,44721	-0,3277	-0,3848	-0,3437	-0,5381	-0,653	-0,463	0,12644	-0,0644
C22	-0,2487	-0,3568	-0,1445	-0,1826	0,97782	-0,6727	0,44721	-0,3277	-0,3848	-0,3437	-0,2935	0,57909	-0,463	0,87924	0,77122
C23	1,1627	-0,3568	-0,6699	-0,1826	-0,8	-0,1552	-0,6708	-0,3277	-0,3848	-0,3437	-0,5381	-0,7762	-0,463	1,25403	0,41869
C24	-0,6519	-0,3568	1,49747	-0,1826	0,97782	-0,1552	-0,6708	-0,3277	-0,3848	-0,3437	-0,7827	0,33267	-0,463	0,0909	0,7451
C25	-0,047	-0,3568	0,44661	-0,1826	0,97782	-0,6727	-0,6708	-0,3277	-0,3848	-0,3437	-0,2935	0,7023	-0,463	-1,7863	1,52851
C26	-0,6519	-0,3568	1,76019	-0,1826	-0,8	-0,6727	-0,6708	-0,3277	-0,3848	-0,3437	-0,7827	-1,0226	-0,463	-1,8897	1,68519
C27	-0,8535	-0,3568	-0,2102	-0,1826	0,08889	-0,6727	-0,6708	-0,3277	0,41139	-0,3437	-0,5381	-0,037	-0,463	0,0812	1,37183
C28	-0,8535	-0,3568	-0,6042	-0,1826	0,08889	-0,6727	-0,6708	-0,3277	-0,3848	-0,3437	-0,7827	-1,1459	1,52114	-1,3921	1,52851
C29	-0,8535	-0,3568	2,67969	-0,1826	0,08889	-0,1552	1,56525	-0,3277	-0,3848	-0,3437	-0,2935	-0,653	-0,463	-0,7491	1,68519
C30	-0,8535	-0,3568	0,44661	-0,1826	0,08889	-0,1552	-0,6708	-0,3277	0,01327	-0,3437	-0,2935	-1,0226	-0,463	-2,6295	1,63296

Figura 53 - Variáveis padronizadas.

Após ter-se realizado a padronização das variáveis, deve-se encontrar a contribuição de cada variável, em relação aos fatores formados nos *Factor Loading.*

Existem duas formas de encontrar esta contribuição:

1º) Uma forma é através da matriz de correlação entre as variáveis originais e as componentes principais. Para verificar a correlação existente entre as variáveis originais e as componentes principais, deve-se selecionar, na Figura 54, a opção *Save case statistics* e a opção *Factor Scores* deve estar selecionada, Ok.

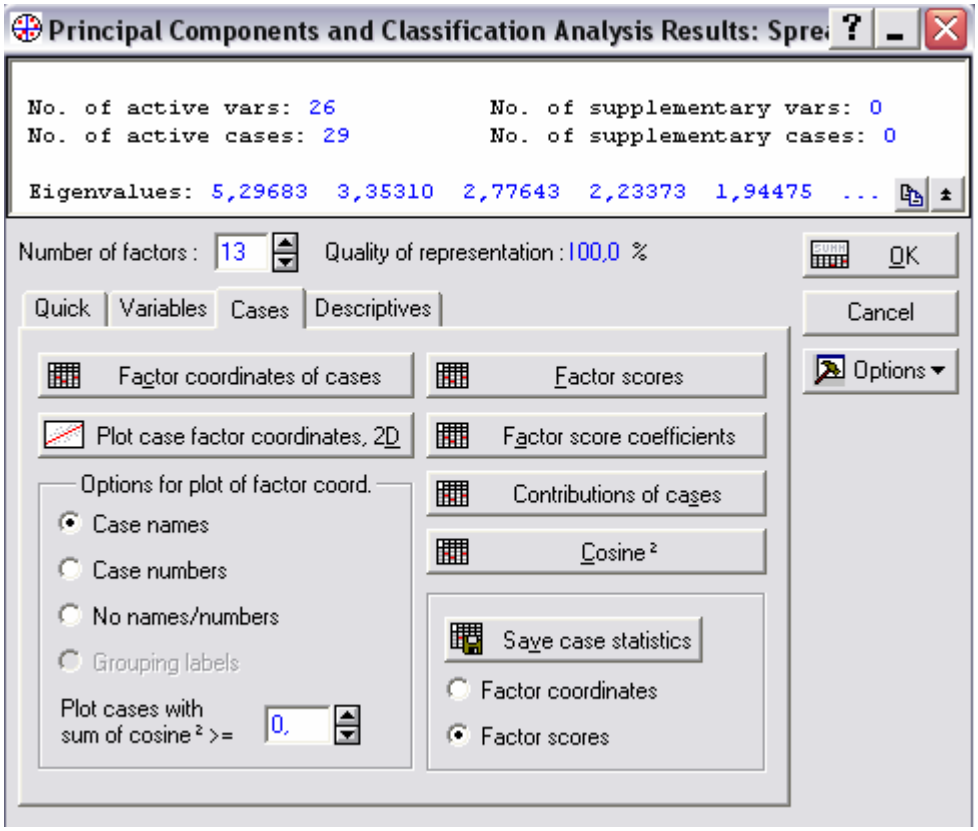


Figura 54 - Caixa de seleção para análise de componentes principais.

Selecionar as variáveis, que se deseja salvar, e *Ok*, conforme Figura 55:

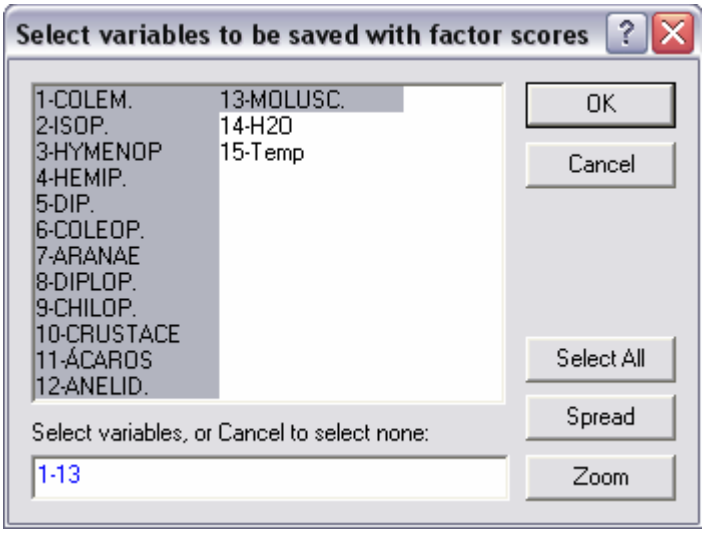


Figura 55 - Caixa de variáveis para análise de componentes principais.

A Figura 56 mostra as variáveis originais, e as componentes principais, que serão utilizadas para compor as correlações, dentro de cada fator.

	1 COLEM.	2 ISOP.	3 HYMENOP	4 HEMIP.	5 DIP.	6 COLEOP.	7 ARANAE	8 DIPLOP.	9 CHILOP.	10 CRUSTACE	11 ÁCAROS	12 ANELID.	13 MOLUSC.	14 CP1
C1	5,5	0	0,5	0,25	0,75	2,5	0,25	0	0,25	0,75	4,75	2	0	-4,35152
C2	4	0,25	0,75	0	0	0,5	0,25	0,5	0,75	0,25	2,5	7,5	0,5	-1,67742
C3	1	0	2,25	0	0,25	1	0	0	0	0	3	0,25	0	-0,44905
C4	1,25	0,25	0,25	0	0,25	0,5	0,75	0	0	0	0	0,5	0	0,203059
C5	0,25	0	0,5	0	0,25	0,75	0	0	0	0	1	0,5	0	0,124351
C6	1	0,5	1	0	0	0,5	0	0	0	0	1	0,5	0,25	0,054192
C7	2	1	1,5	0	0	0,25	0	0	0	0	0,75	1,5	0	0,136403
C8	1,25	3	7	0	0,25	0,75	0	0,5	0	0	0	1,5	0	0,227277
C9	1	0,5	7	0	0,25	0,25	0,25	0,5	0,25	0	0,75	5,5	0	0,052301
C10	1,75	0	0	0	0,25	0,25	0	0	0	0	0,5	4,75	0	-0,01327
C11	0,25	0	3,25	0	0	0	0,75	0	0	0	0,25	2,5	0,25	0,436391
C12	0,75	0	0	0	0	0,25	0	0	0,25	0,25	1	3,5	0	-0,17335
C13	1,75	0	0	0	0	0	0	0	3	0,5	1,5	7	0,25	-1,18612
C14	2	0	1,25	0	0	0,25	0,5	0	0	0	1	4,75	0,25	-0,22629
C15	0,5	0	1	0	0	0,25	0	0	1,75	0	0	2,5	0	0,284703
C16	1,5	0	0,5	0	1,25	0,25	0,25	0	0	0	1	3,75	0	-0,24751
C17	0,5	0,75	11,75	0	0	0,25	0	0	0	0	1	2,25	0	0,556165
C18	0,5	0	4,5	0	0,5	0	0,25	0	0	0	0,75	0,25	0	0,463913
C19	0,25	0	0	0	0,25	0,25	0	0	0	0	0,5	1,75	0	0,356826
C21	0	0	6,25	0	0	0	0,25	0	0	0	0,25	1	0	0,7668
C22	0,75	0	2,75	0	0,5	0	0,25	0	0	0	0,5	3,5	0	0,298823
C23	2,5	0	0,75	0	0	0,25	0	0	0	0	0,25	0,75	0	0,144101
C24	0,25	0	9	0	0,5	0,25	0	0	0	0	0	3	0	0,58071
C25	1	0	5	0	0,5	0	0	0	0	0	0,5	3,75	0	0,321285
C26	0,25	0	10	0	0	0	0	0	0	0	0	0,25	0	0,917563
C27	0	0	2,5	0	0,25	0	0	0	0,5	0	0,25	2,25	0	0,537283
C28	0	0	1	0	0,25	0	0	0	0	0	0	0	0,25	0,574028
C29	0	0	13,5	0	0,25	0,25	0,5	0	0	0	0,5	1	0	0,730301
C30	0	0	5	0	0,25	0,25	0	0	0,25	0	0,5	0,25	0	0,558045

15 CP2	16 CP3	17 CP4	18 CP5	19 CP6	20 CP7	21 CP8	22 CP9	23 CP10	24 CP11	25 CP12	26 CP13
-2,24423	-0,04868	0,045066	0,320024	0,916503	-0,33603	-0,03837	0,613131	0,997849	-0,23859	-0,4574	-0,6027
2,85367	-1,07679	-1,23137	0,240871	-0,53661	2,003616	0,287993	-0,53059	0,491669	1,138933	0,585468	0,637425
-0,96215	0,135435	0,555941	0,802186	-0,53972	1,351973	0,418824	-2,83757	-2,57494	0,24924	0,370741	-0,24767
-0,5011	0,204612	-1,86302	0,71915	-0,39949	-2,72227	-0,55702	-0,68897	-0,09961	0,683187	0,36656	1,759111
-0,71444	0,348211	0,571147	0,516068	-0,91219	0,173266	0,347733	-1,60666	0,06204	-0,72186	-0,91182	1,593186
0,212946	-0,03653	0,270617	1,793056	-1,01625	0,730813	1,445644	0,591418	-0,38153	-0,151	-0,9122	0,070253
-0,09378	-0,54786	0,789884	0,886793	-0,9333	-0,48736	-1,19769	1,341815	-1,3254	0,578567	0,938216	-1,30112
0,09651	-4,08632	1,033191	-0,02392	-0,67252	-1,81771	0,999085	0,621389	-0,24875	-0,55184	-0,08268	0,425007
0,828292	-1,76432	-0,56051	-1,52054	0,599324	0,347053	-1,25742	-2,20568	1,613523	-0,36041	-0,07633	-1,14056
0,18928	0,406905	0,248519	-0,72785	-1,1188	0,420075	-1,98009	0,719606	-0,14068	-0,37072	-1,22845	0,360953
0,653891	0,471652	-2,40481	1,172183	0,591603	-0,90048	0,494169	0,272328	0,028245	-1,36122	-0,4806	-0,98147
0,30661	0,632828	0,834404	0,189025	-0,18023	0,002259	-1,34448	-0,14316	0,299065	-2,78742	1,872977	1,665305
2,745219	1,520445	1,73795	-0,75091	1,482996	-1,44972	1,059697	0,38551	-1,03848	-0,10627	0,940144	0,549472
0,954897	0,399012	-1,7574	0,774815	-0,29522	0,246199	-0,95972	0,842315	-1,19098	-0,27796	-1,50823	-0,45746
0,742304	0,871166	1,474595	-0,02031	0,328036	-1,68502	-0,16954	-1,26089	0,457791	1,63853	-2,44684	-1,05589
-0,77324	0,470669	-1,05849	-3,00962	-1,58151	-0,2455	1,335581	0,432769	-1,10593	0,737058	0,374236	0,233389
-0,30032	-0,88169	0,684006	0,330228	1,68742	1,096151	-0,40244	0,962936	-1,85223	-0,3839	-0,40829	-0,64226
-0,73382	0,31204	-0,48735	-0,41109	0,028993	-0,15909	0,926354	-0,22258	0,082945	0,81239	2,214123	-1,3122
-0,30478	0,508751	0,47104	-0,0197	-0,98569	0,051624	-0,22535	-0,50769	0,651174	-1,18833	-0,10623	-0,25644
-0,26756	0,203524	-0,22961	0,620509	0,92987	-0,08684	-0,5205	-0,21318	0,596961	-0,48363	0,81487	-0,8021
-0,18817	0,400938	-0,63025	-1,18621	-0,32341	-0,16171	-0,3883	0,369252	-0,2151	-0,60293	0,307074	-0,916
-0,19317	0,297769	0,509216	0,918769	-0,96359	0,063213	-1,7345	0,726933	0,917555	2,645428	0,97035	0,867584
-0,50422	-0,04607	0,191155	-1,30166	0,823897	0,974518	0,34507	0,900745	0,385304	-0,05022	-1,66413	2,13826
-0,20391	0,224557	0,128931	-1,40957	-0,11944	0,884531	-0,39809	1,038903	-0,10018	0,047607	-0,12277	-0,39946
-0,45565	-0,06778	0,558077	0,584202	1,429244	0,91781	-0,17093	0,634205	1,127169	0,948085	0,63825	0,355407
0,016831	0,564525	0,680477	-0,46368	-0,11135	-0,19766	0,094293	-0,25762	0,66068	-0,52698	-0,26614	-1,40357
0,125854	0,588238	0,017231	0,873435	-1,00047	0,638288	2,409713	0,91015	1,940671	-0,34742	0,018365	-0,36173
-0,73461	-0,28382	-1,2167	-0,03769	2,609354	0,103262	0,422418	-0,27381	-0,64098	0,713183	0,064152	1,189192
-0,55117	0,278592	0,638052	0,141445	0,262567	0,244729	0,757866	-0,61499	0,602138	0,3185	0,1966	0,036099

Figura 56 - Caixa com variáveis originais e as componentes principais.

Para fazer a matriz de correlação, seleciona-se *Statistics/Basic Statistics/Tables,* conforme Figura 57:

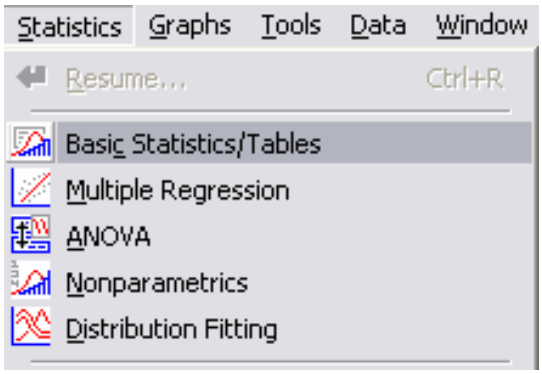


Figura 57 - Caixa de seleção da estatística descritiva.

Selecionando, na Figura 58, *Correlation matrices* e Ok, abre-se uma caixa de opções para encontrar a matriz de correlação entre as variáveis originais e as componentes principais.

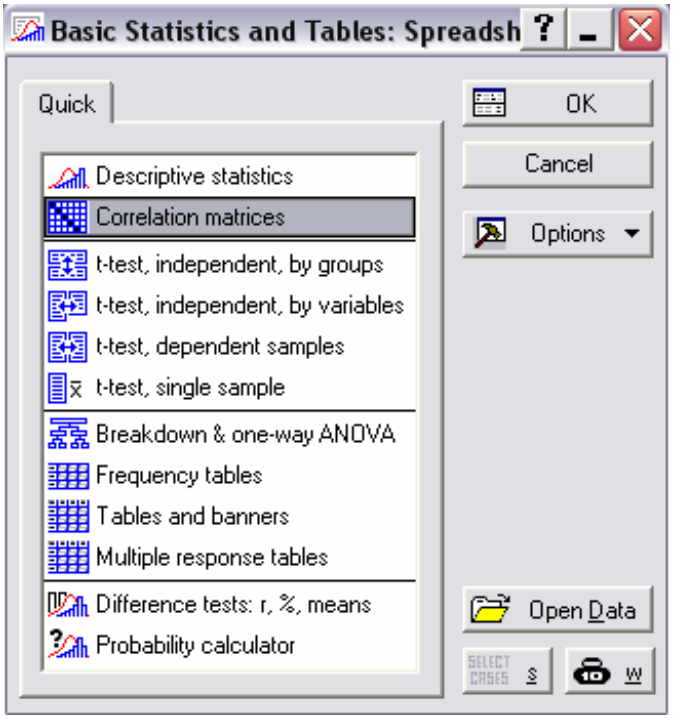


Figura 58 - Caixa de seleção para matriz de correlação entre variáveis originais e as componentes principais.

Selecionando a opção *Two lists (rect. matrix)*, é possível visualizar todas as variáveis e as componentes que se deve selecionar, para que seja possível verificar as correlações, conforme Figura 59.

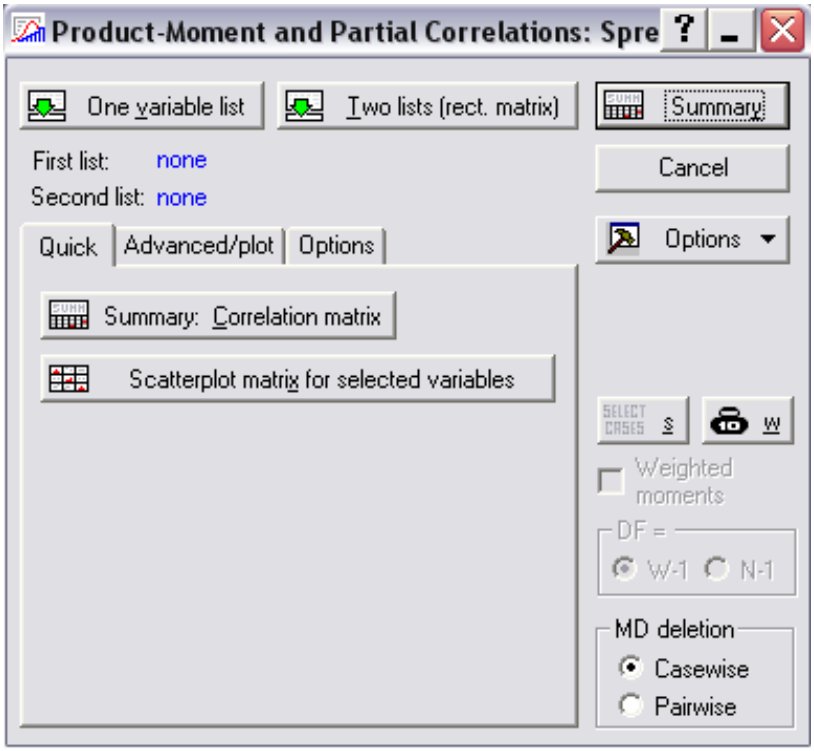


Figura 59 - Caixa de seleção das variáveis que irão compor a matriz de correlação.

A Figura 60 mostra as variáveis e as componentes a serem selecionadas.

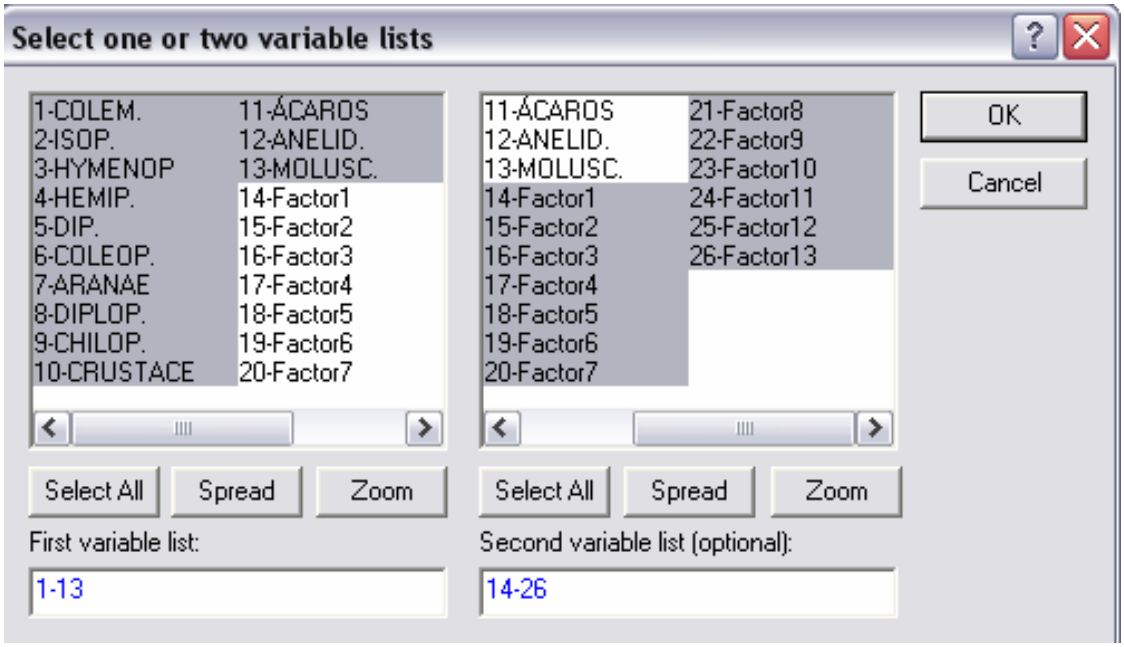


Figura 60 - Caixa com as variáveis e as componentes selecionadas.

Na Figura 61, selecionando a opção *Summary: Correlation matrix*, encontra-se a matriz de correlação.

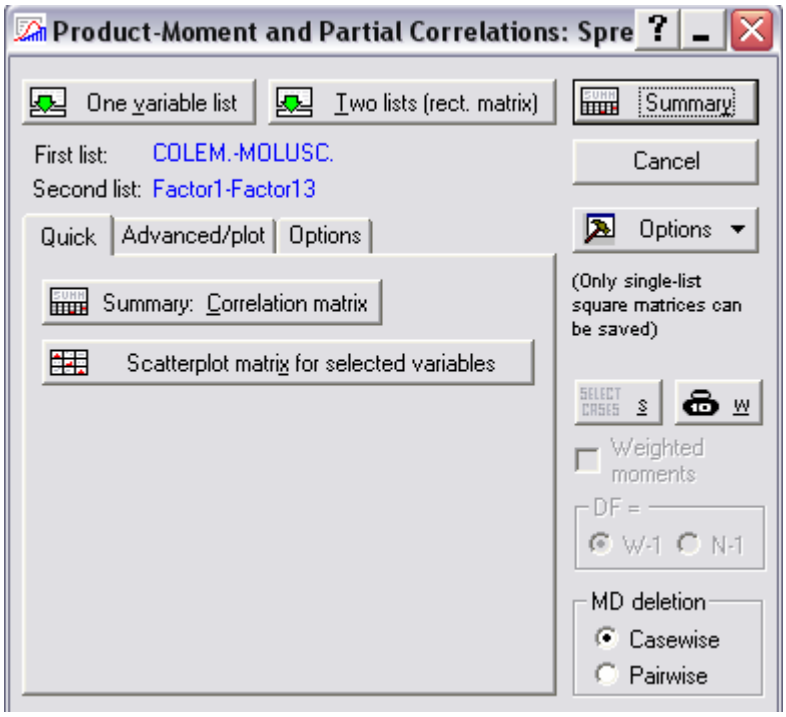


Figura 61 - Caixa de seleção da matriz de correlação.

A Figura 62 mostra a matriz de correlação entre as variáveis originais e as componentes principais e a contribuição de cada variável em relação a cada fator.

Variáveis	Correlação entre os dados originais e as componentes principais As correlações significativas estão em vermelho e ocorrem quando p < ,05000 N=29 (número de observações)												
	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9	Factor10	Factor11	Factor12	Factor13
COLEM.	-0,89	0,09	-0,15	-0,09	0,03	-0,15	0,03	-0,21	0,18	-0,03	0,25	0,04	0,02
ISOP.	0,06	0,07	-0,88	0,21	0,09	-0,13	-0,29	0,09	0,17	-0,17	-0,07	0,01	-0,03
HYMENOP	0,41	-0,22	-0,43	-0,04	-0,11	0,72	0,24	0,05	0,08	-0,04	0,07	-0,02	0,02
HEMIP.	-0,84	-0,43	-0,01	0,01	0,06	0,18	-0,06	-0,01	0,12	0,19	-0,05	-0,09	-0,12
DIP.	-0,25	-0,52	0,03	-0,19	-0,71	-0,21	-0,02	0,28	0,05	-0,01	0,04	0,01	0,01
COLEOP.	-0,81	-0,43	-0,23	0,06	0,18	-0,03	-0,03	0,02	-0,16	-0,04	-0,01	-0,17	0,10
ARANAE	-0,07	0,01	0,03	-0,90	0,09	0,21	-0,35	-0,04	-0,07	-0,06	-0,01	0,02	-0,00
DIPLOP.	-0,16	0,44	-0,80	-0,09	-0,15	-0,07	0,06	0,00	-0,24	0,21	0,03	0,05	-0,01
CHILOP.	-0,28	0,63	0,30	0,42	-0,14	0,28	-0,33	0,15	-0,12	-0,04	0,13	-0,04	-0,02
CRUSTACE	-0,89	0,10	0,13	0,17	-0,01	0,26	-0,10	0,05	0,10	0,09	-0,13	0,15	0,08
ÁCAROS	-0,90	-0,12	0,01	0,01	0,07	0,05	0,26	0,05	-0,19	-0,23	-0,03	0,08	-0,07
ANELID.	-0,35	0,72	-0,00	-0,11	-0,49	0,05	0,10	-0,24	0,07	-0,08	-0,13	-0,10	0,00
MOLUSC.	-0,26	0,73	0,06	-0,32	0,30	-0,09	0,23	0,35	0,14	0,02	0,00	-0,05	0,01

Figura 62 - Matriz de correlação entre as variáveis originais e as componentes principais.

Na Figura 62, os valores que estão em vermelho representam a contribuição de cada variável em cada fator, ou seja, as variáveis que estão em vermelho, no fator 1, são as que melhor o explicam.

2º) Outra forma de encontrar a contribuição das variáveis em relação aos fatores formados, é mediante os *Factor loadings.* Aqui, o número de fatores a serem utilizados na análise é quatro, pois foram apenas esses os autovalores superiores a 1, encontrados na análise.

A Figura 63 mostra a caixa de seleção de comandos para a *ACP*. Retornando para a *AF*, seleciona-se: *Loadings/ Factor rotation* seleciona-se *unrotated/ Summary: Factor loadings,* para ver quanto cada variável contribui na formação de cada componente. Também nesta janela tem-se a opção de verificar o método gráfico *Plot of loadings, 2D,* que representa, graficamente, os planos fatoriais, mostrando a importância de cada variável no estudo. Nesta janela ainda há a opção do método gráfico *Plot of loadings, 3D*, que possibilita identificar a localização das variáveis num espaço tri-dimensional.

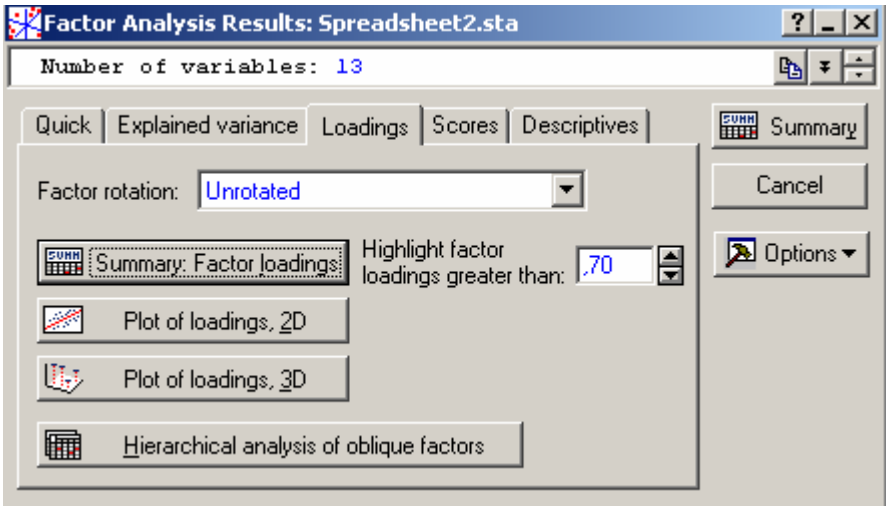


Figura 63 - Caixa de seleção dos *Factor Loadings.*

Conforme Pereira (2001), “o passo final da *AF* é verificar se os fatores, que são dimensões abstratas, podem ser interpretados de forma coerente com a natureza dos fenômenos estudados”. Para isso, deve-se analisar a matriz fatorial, na qual estão os *factor loadings*, e verificar quais as variáveis que melhor se correlacionam com cada fator.

Em *ACP,* a derivação de fatores se dá por várias rotações de eixos que melhor expressem a dispersão dos dados. No modelo fatorial final, as variações das medidas estão maximizadas, e as relações entre dimensões suavizadas. Devido a

isso, o pesquisador deverá buscar relação entre os fatores e as variáveis originais numa matriz fatorial rodada (PEREIRA, 2001).

A Figura 64 mostra o resultado dos *Factor Loadings*, antes da rotação nos eixos, e mostra a contribuição das variáveis na formação dos componentes.

Variáveis	Factor Loadings Extração das compentes principais Factor 1	Factor 2	Factor 3	Factor 4
COLEM.	-0,89	0,09	-0,14	-0,09
ISOP.	0,04	0,07	-0,88	0,21
HYMENOP	0,38	-0,21	-0,45	-0,04
HEMIP.	-0,83	-0,43	0,01	0,01
DIP.	-0,27	-0,51	0,00	-0,18
COLEOP.	-0,82	-0,43	-0,22	0,05
ARANAE	-0,06	0,01	0,05	-0,90
DIPLOP.	-0,17	0,44	-0,80	-0,09
CHILOP.	-0,28	0,63	0,30	0,42
CRUSTACE	-0,89	0,10	0,14	0,17
ÁCAROS	-0,90	-0,12	0,01	0,01
ANELID.	-0,36	0,72	-0,01	-0,10
MOLUSC.	-0,27	0,72	0,05	-0,32

Figura 64 - Composição dos fatores.

Na Figura 64, pode-se visualizar as ponderações de cada variável que irão compor a combinação linear. Observa-se que os valores em destaque são os que possuem uma significância maior que 0,7. Este valor de significância pode ser alterado segundo as necessidades do pesquisador.

O ideal é identificar, em cada combinação linear, um conjunto de variáveis que representa este fator e, a partir daí, atribuir-se um nome para o fator. Esta abstração, para o fator, passa a identificá-lo, representando um conjunto de variáveis. Quando esta identificação ficar difícil, por apresentar mais de um grupo de variáveis significativas no mesmo fator, ou em fatores diferentes, recorre-se à realização de rotações, pois, desta forma, mantem-se a mesma inércia no conjunto analisado, mas os eixos são rotacionados, possibilitando uma melhor visualização da disposição dos pontos. Existem diversos tipos de rotações, as quais devem ser estudadas para maior entendimento, e deve-se verificar em quais situações elas devem ser utilizadas. A rotação mais utilizada é a *Varimax normalizada*, pois esta mantem os eixos perpendiculares entre si, ou seja, ortogonais.

A Figura 65 mostra a caixa de seleção de comandos para *ACP,* seleciona-se: *Loadings/ no Factor rotation (Varimax normalized)/Summary:Factor loadings,* para se fazer a rotação nos eixos, possibilitando uma melhor visualização das variáveis mais representativas em cada componente.

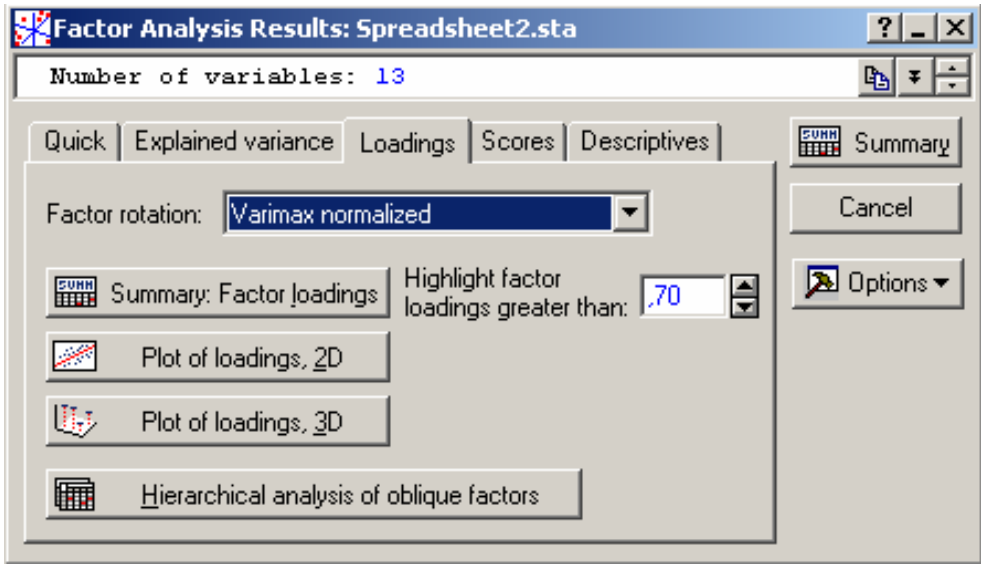


Figura 65 - Caixa de seleção para a rotação *varimax normalized.*

A Figura 66 mostra o resultado dos *Factor Loadings,* após a rotação *varimax normalized.*

Variáveis	Factor Loadings após a rotação dos eixos Extração das componentes principais			
	Factor 1	Factor 2	Factor 3	Factor 4
COLEM.	0,82	0,36	0,13	0,10
ISOP.	0,01	-0,09	0,88	-0,17
HYMENOP	-0,26	-0,38	0,42	0,07
HEMIP.	0,92	-0,13	-0,10	0,03
DIP.	0,41	-0,37	-0,10	0,22
COLEOP.	0,92	-0,17	0,13	-0,01
ARANAE	0,02	0,10	-0,08	0,89
DIPLOP.	0,08	0,35	0,85	0,10
CHILOP.	0,06	0,69	-0,19	-0,48
CRUSTACE	0,81	0,39	-0,13	-0,18
ÁCAROS	0,89	0,18	-0,05	0,00
ANELID.	0,12	0,79	0,12	0,05
MOLUSC.	0,01	0,79	0,05	0,27

Figura 66 - Composição dos fatores.

Observa-se, na Figura 66, que a rotação *varimax normalized* possibilitou uma melhor visualização dos fatores, nos quais a proporção de variação das variáveis está melhor representada. Observa-se que os valores que possuem uma significância igual, ou superior, a 0,7 estão em destaque em cada fator.

Neste estudo, utilizar-se-á todos os quatro fatores que possuem as variáveis explicativas, pois através do método gráfico sugerido por CATTEL (1966), esses fatores explicam a maior variância.

Pode-se concluir, ainda, que o fator 1 é o mais importante para o estudo, pois é derivado do maior autovalor e possui uma explicação de 33,05%, sendo que as variáveis, que mais contribuem neste, são representadas pelos seguintes organismos: Colêmbolos, Hemípteros, Coleópteros, Crustáceos e Ácaros; o fator 2 e o fator 3, são explicados por duas variáveis, apenas. O fator 2, pelas variáveis representadas pelos Anelídeos e Moluscos, e o fator 3 pelas variáveis Isópteros, Diplópodes. Já o fator 4 é explicado apenas por uma variável, representada pelo organismo Aranae.

Para que haja uma melhor visualização desses fatores, optou-se em utilizar os gráficos de dispersão, ou os planos fatoriais, que examinam a localização das variáveis num sistema de coordenadas criado pelos fatores.

Na Figura 63, ao selecionar a opção *Plot of loadings, 2D,* pode-se analisar todos os fatores encontrados, sendo que, apenas aqueles fatores que apresentarem variáveis explicativas, trarão a devida contribuição para o estudo, de forma que se possa identificar quais as variáveis possuem uma maior representatividade nos planos fatoriais.

Os fatores a serem relacionados, neste primeiro plano, são: *Factor 1* com *Factor 2,* clica-se em *Ok,* conforme Figura 67.

É importante salientar que o fator 1 é composto de cinco variáveis, que possuem variância igual, ou superior a 0,7. Sendo assim, esse é o fator mais importante para análise. Logo, ao fazer os planos fatoriais, o fator 1 será mantido fixo no eixo do x, e os fatores do eixo y serão modificados a cada plano, para que se possa verificar a importância de cada variável na formação de cada fator.

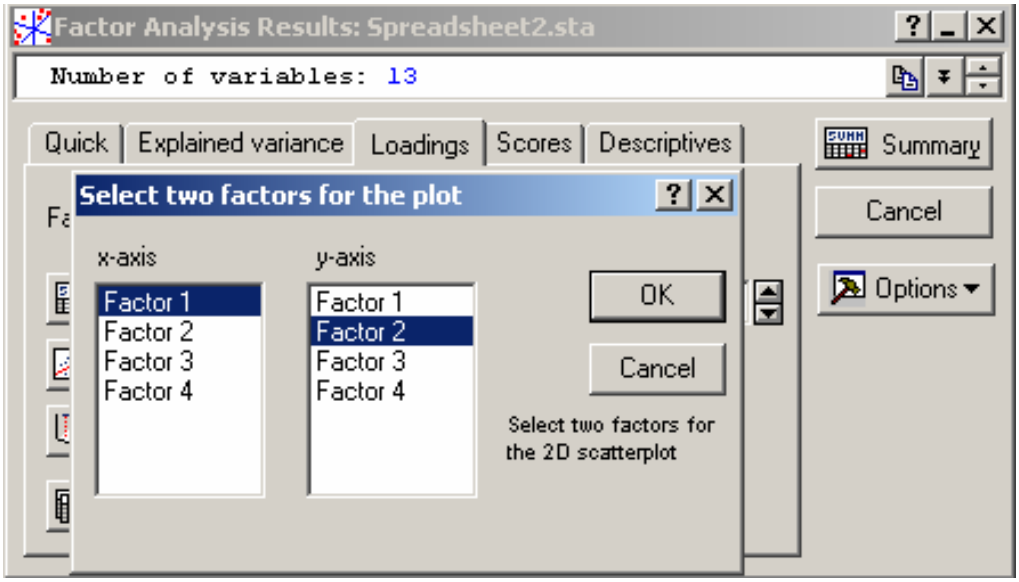


Figura 67 - Caixa de seleção dos fatores, para fazer planos fatoriais.

Antes de interpretar a Figura 68, deve-se levar em consideração que, se a variância for nula, ou próxima de zero, significa todos os indivíduos estarem próximos, ou em cima, da origem do plano principal da nuvem de pontos, e possuem baixa representatividade. Pode-se, então, interpretar o plano principal da nuvem de pontos como sendo o plano que torna máxima a variância do conjunto dos ***n*** pontos projetados sobre ele.

A Figura 68 corresponde à relação entre as variáveis do fator 1 e fator 2, da *AF*. Analisando a Figura 68, observa-se que as variáveis formam grupos por similaridades de explicação, ou seja, estão agrupadas por fatores. As variáveis que melhor representam o fator 1 formam um grupo distinto dos demais, e são representadas pelos organismos: Colêmbolos, Hemípteros, Coleópteros, Crustáceos e Ácaros, estando localizadas distantes da origem, sendo estas que possuem uma maior representatividade em relação ao fator 1, pois se forem traçadas perpendiculares em relação a esse fator, pode-se verificar que essas variáveis são as que estão localizadas mais distante da origem. As variáveis que melhor representam o fator 2, e formam outro grupo distinto, são as seguintes: Anelídeos, Moluscos e Chilópodas. O restante das variáveis possuem baixa representatividade, por estarem localizadas próximas à origem do plano fatorial.

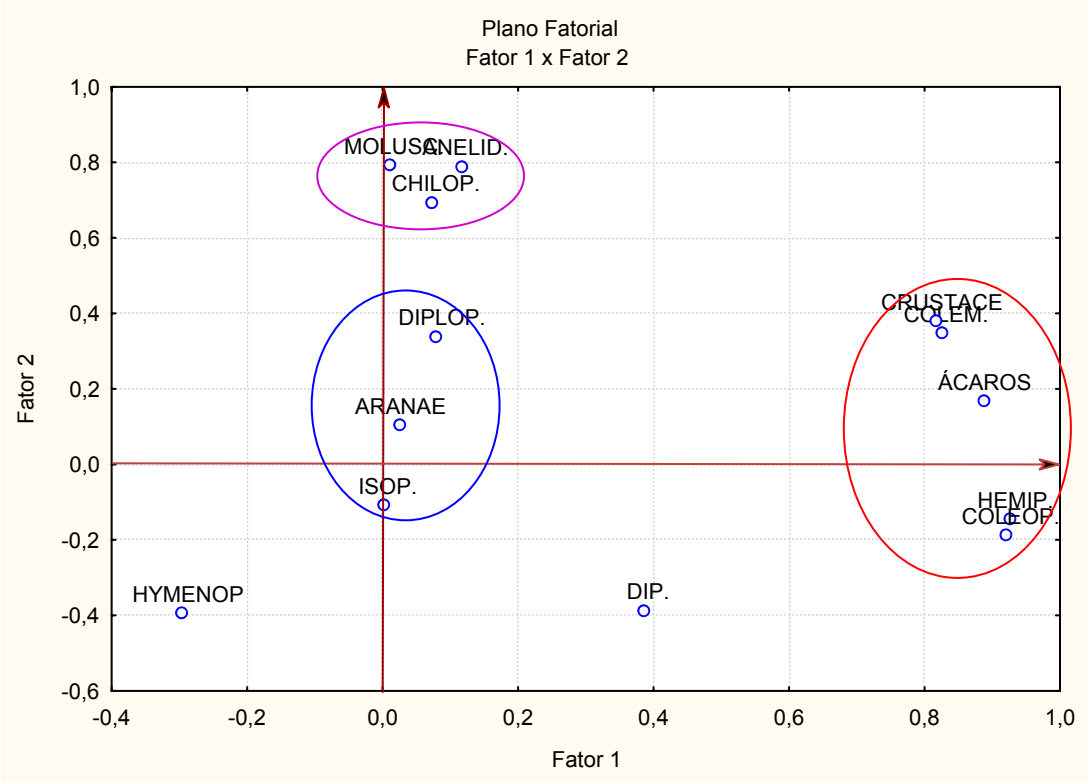


Figura 68 - Gráfico representando a relação entre fatores (fator 1 e fator 2) e variáveis segundo *factor loadings.*

Pode-se concluir ainda, na Figura 68, na qual fica evidente como as variáveis agrupam-se e como são suas relações com os eixos, os *factors loadings,* referentes aos fatores 1 e 2. As variáveis que melhor representam o fator 1 são as que melhor o explicam, ou seja, as que estão mais distantes da origem, em relação ao eixo do x, representadas pelo círculo em vermelho.

As variáveis que melhor representam o fator 2 são as que estão contidas no círculo em rosa, ou seja, as que estão mais distantes da origem, em relação ao eixo y, sendo as que melhor explicam esse fator.

As demais variáveis possuem baixa representatividade, devido ao fato de estarem próximas da origem, em relação aos dois eixos.

A análise que auxilia a interpretação dos planos fatoriais é análise de agrupamentos, pois esta serve para confirmar se as varáveis que estão num mesmo grupo são as mesmas que explicam determinado fator.

A Figura 69, que representa os planos fatoriais correspondentes ao fator 1 e fator 2 da *ACP*, neste plano, foram traçadas perpendiculares, como pode-se observar em relação ao fator 1.

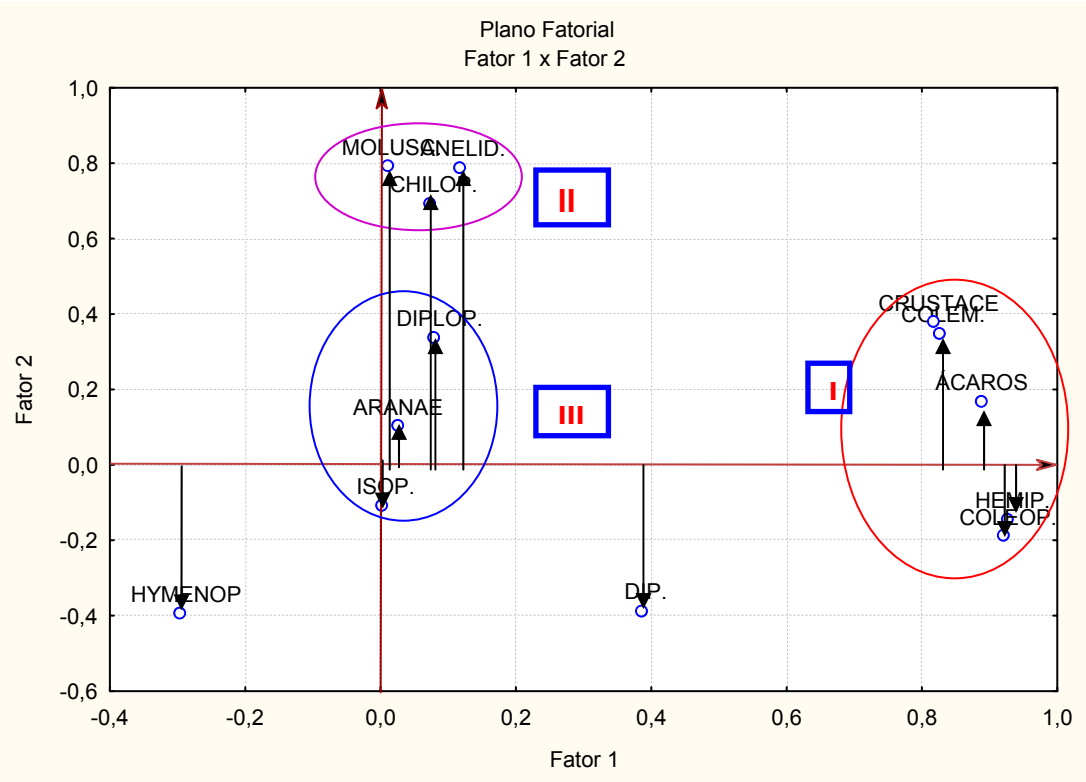


Figura 69 - Gráfico dos planos fatoriais, que representam as perpendiculares em relação ao fator 1.

Observando a Figura 69, pode-se concluir que o grupo I é o mais representativo, em relação ao fator 1, pois este é o que está localizado na extremidade do eixo x e, portanto, o mais distante da origem do eixo cartesiano, logo, possui a maior influência. Para se encontrar as distâncias de cada variável, traça-se um segmento de reta perpendicular ao eixo x, que representa o fator 1. Após realizada esta tarefa, verifica-se qual a variável, ou o conjunto de variáveis, que está localizado mais distante da origem. As variáveis que estiverem mais distantes possuirão maior influência em relação ao fator examinado.

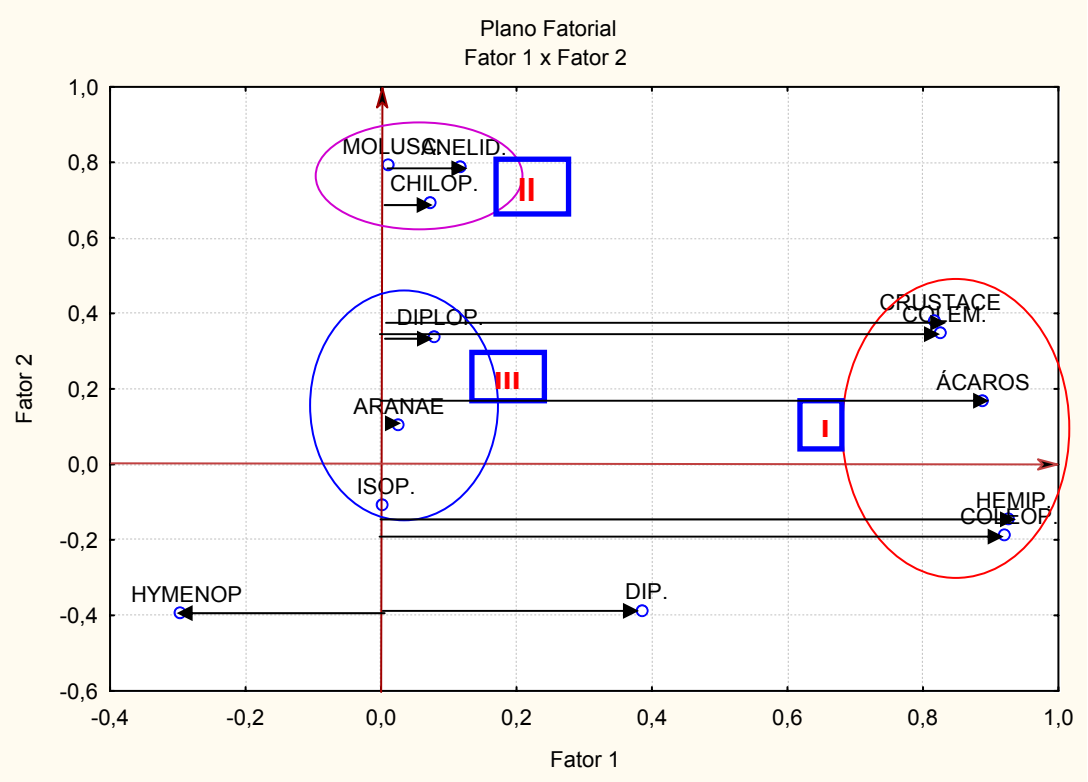


Figura 70 - Gráfico dos planos fatoriais, que representam as perpendiculares traçadas em relação ao fator 2.

A Figura 70 representa os planos fatoriais da relação entre o fator 1 e fator 2 da *AF*, Nesses planos, o segmento de reta será traçado perpendicular ao eixo y, que representa o fator 2. A análise é realizada de forma análoga ao fator 1, levando-se em consideração, neste caso, o fator 2.

Observando-se esse gráfico, o grupo II, das variáveis que estão contidas no círculo rosa, constata-se que são as variáveis que possuem uma maior representatividade em relação ao fator 2, pois estão localizadas distante da origem, sendo que as demais variáveis possuem baixa representatividade em relação a este fator.

A Figura 71 representa os planos fatoriais, da relação entre variáveis dos fatores 1 com 2 da *AF*. Nestes planos foram traçadas perpendiculares em relação a bissetriz dos planos.

Após, encontra-se o significado, isto é, atribui-se um nome para cada fator e pode-se verificar como as variáveis estão influenciando, concomitantemente, estes fatores. Para tal, traça-se a bissetriz, que passa pelo primeiro e terceiro quadrantes do plano fatorial, e, novamente, traça-se segmentos de reta perpendiculares à

bissetriz. Novamente, as variáveis mais distantes da origem serão as mais importantes.

Da Figura 71, pode-se concluir que as variáveis de maior expressão, em relação a esses dois planos, continuam sendo as que estão contidas nos círculos em vermelho e rosa, as quais possuem uma maior distância em relação à origem desses planos, sendo que as variáveis que melhor representam o fator 1 estão contidas no grupo I, e as que melhor representam o fator 2 estão contidas no grupo II.

Nos outros planos fatoriais, que correspondem ao fator 1 *x* fator 3 e fator 1 *x* fator 4, a análise é realizada de forma análoga a este exemplo.

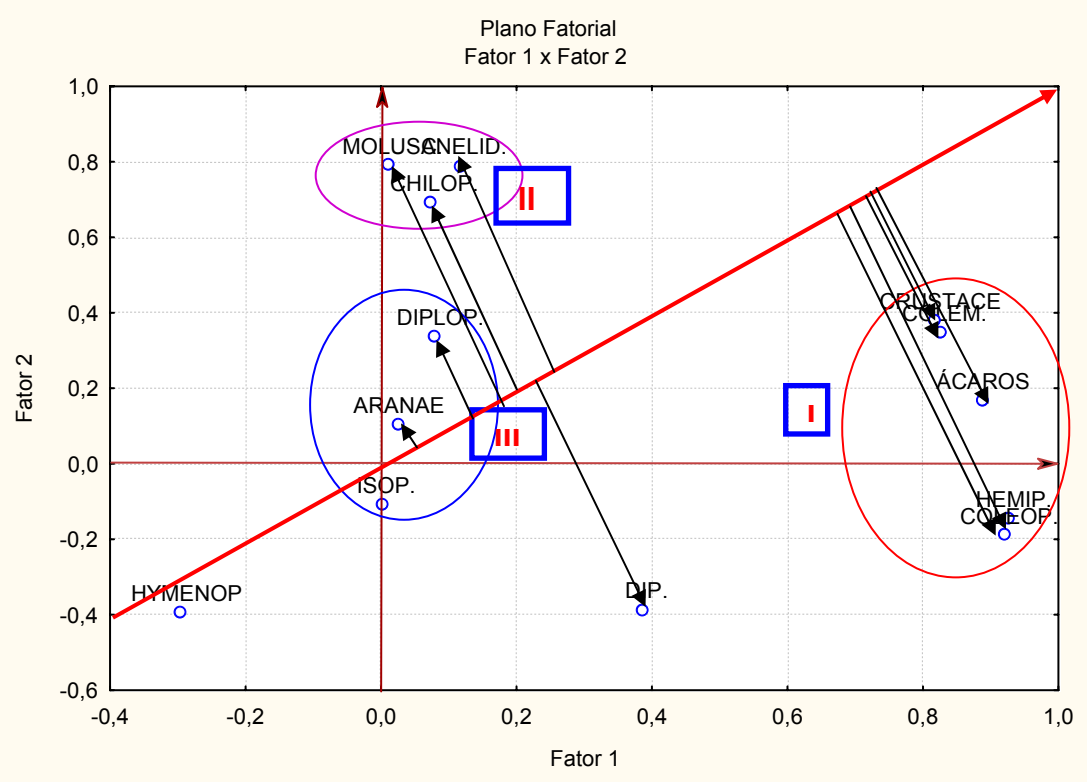


Figura 71 - Gráfico dos planos fatoriais, da relação entre variáveis do fator 1 com 2 em relação à bissetriz.

Ao selecionar a opção *Loadings/ Plot of loadings, 3D* na Figura 65, obtém-se a Figura 72, que mostra a localização das variáveis num espaço tri-dimensional.

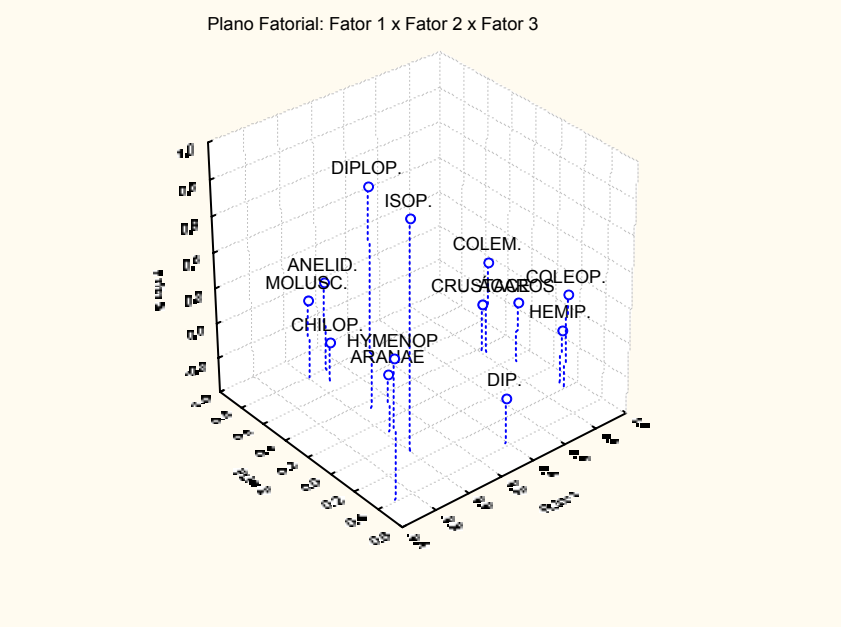


Figura 72 - Gráfico do plano tri-dimensional, da *ACP.*

A Figura 73 mostra o módulo principal do *STATISTICA*, para encontrar os planos principais, que possibilitarão visualizar a nuvem de variáveis que melhor representa cada plano, bem como a nuvem de pontos que mostra a localização de cada objeto (estado) em relação às variáveis nos planos principais, para isso seleciona-se: *Multivariate Exploratory Techniques – Principal Components & Classification Analysis*:

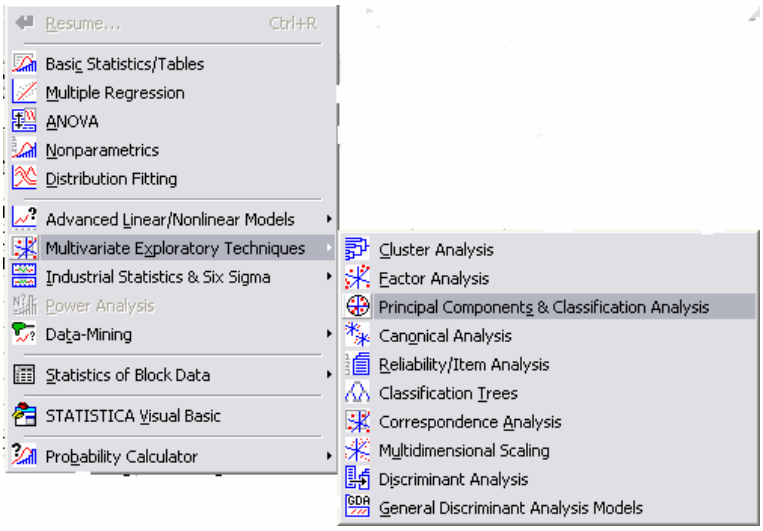


Figura 73 - Caixa de seleção da *ACP.*

A Figura 74 mostra a caixa de seleção de variáveis e comandos para *ACP.* Clica-se em *Variables,* e o programa mostrará todas as variáveis.

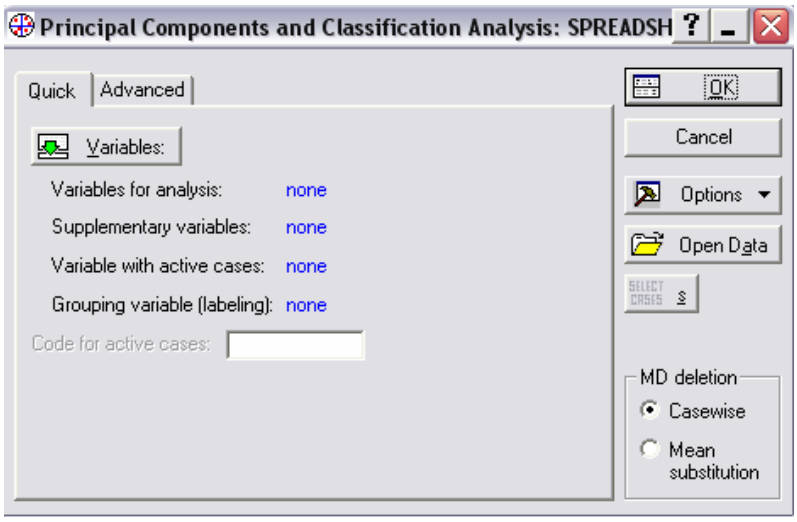


Figura 74 - Caixa de seleção da *ACP.*

Se o pesquisador quiser estudar todas, basta selecioná-las e clicar em *Ok*. Se no estudo tiver algumas variáveis suplementares, isto é, que o pesquisador busque identificar seu comportamento, em relação às outras variáveis, basta selecionar as variáveis que não são suplementares na primeira janela, que diz, logo abaixo, *Variables for analysis,* e na outra janela selecionar as variáveis suplementares, sendo que estas podem ser uma ou mais, na janela *Supplementary variables* e, a seguir, é só clicar em *Ok.*

Na Figura 75, apresenta-se a totalidade de variáveis para análise. Neste caso, após selecionadas todas as variáveis, clica-se em *Ok*.

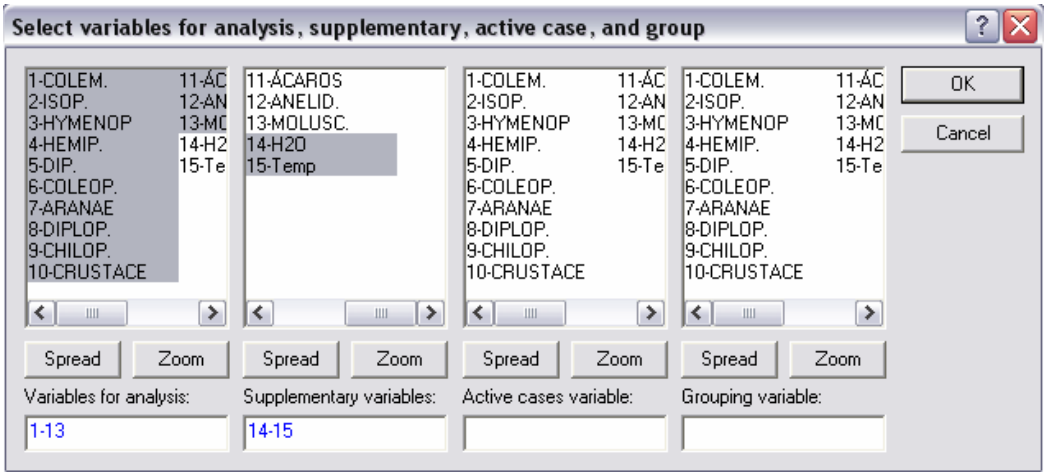


Figura 75 - Caixa de seleção das variáveis para *ACP.*

A Figura 76, na opção *Variables for analysis:* mostra que todas as variáveis foram selecionadas, inclusive as suplementares, basta clicar em *Ok*.

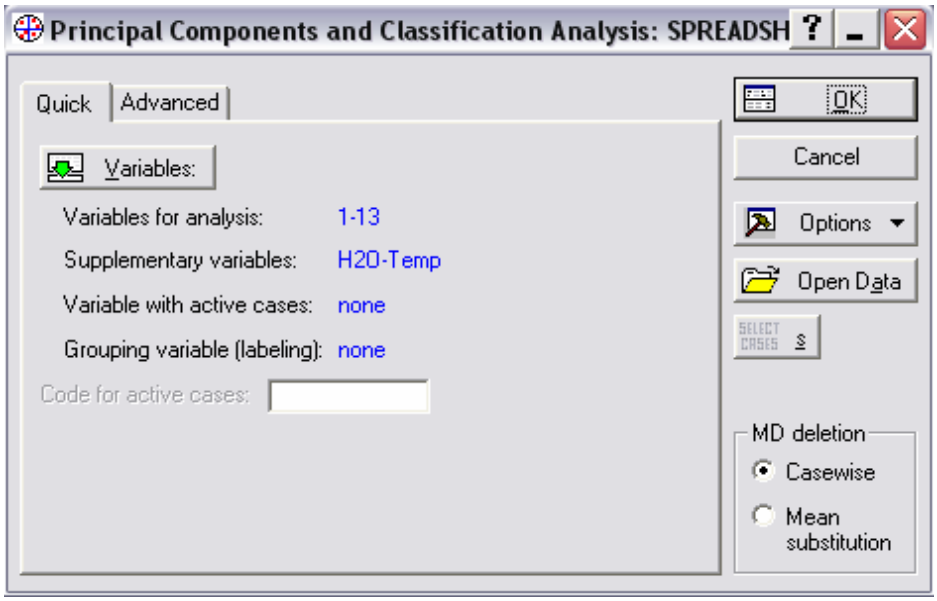


Figura 76 - Caixa de seleção da *ACP.*

A Figura 77 mostra a caixa de seleção de variáveis e comandos para *ACP.* Seleciona-se *Variables/Plot case factor coordinates, 2D,* e clica-se em *Ok,* para fazer os planos principais, com a nuvem de variáveis.

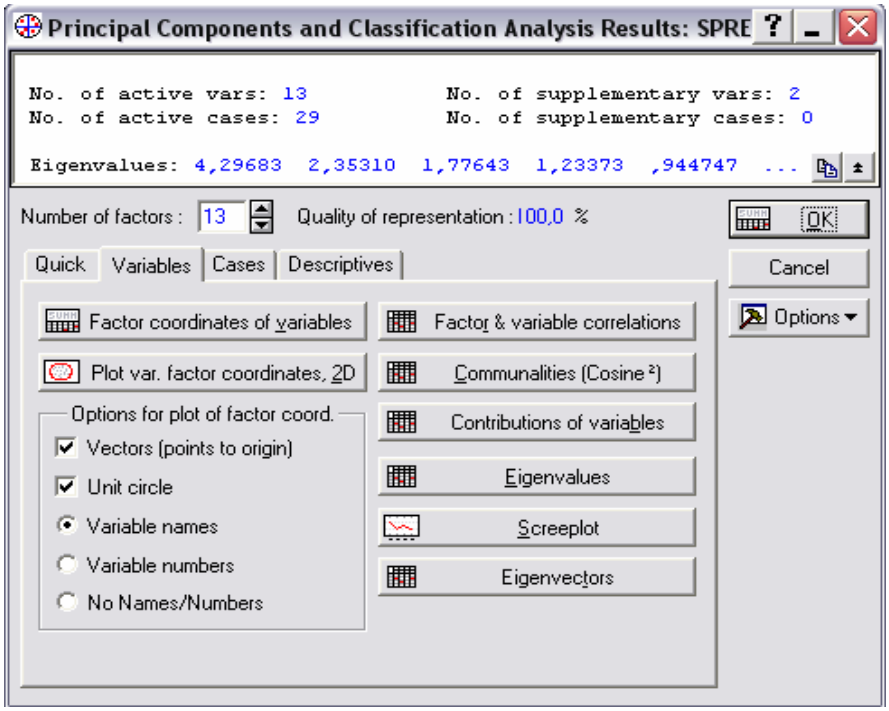


Figura 77 - Caixa de seleção da *ACP.*

A Figura 78 mostra os fatores a serem relacionados, neste primeiro plano principal, que são: *Factor 1 x Factor 2* e, em seguida, clica-se em *Ok.*

É importante lembrar que os fatores de um a quatro são os que possuem as variáveis explicativas. Portanto, aqui também os fatores serão relacionados de forma análoga aos planos fatoriais.

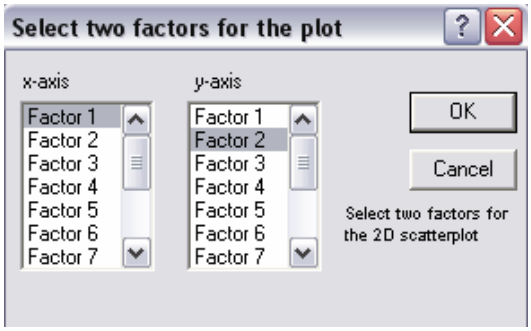


Figura 78 - Caixa de seleção dos fatores.

A interpretação dos componentes principais é, sem dúvida, um dos pontos mais delicados da análise. Aqui, duas abordagens devem ser utilizadas: considerar, de um lado, as correlações com as variáveis iniciais e, de outro, os indivíduos que estão sendo estudados.

A Figura 79 mostra o círculo de correlação, ou primeiro plano principal, com a nuvem de variáveis.

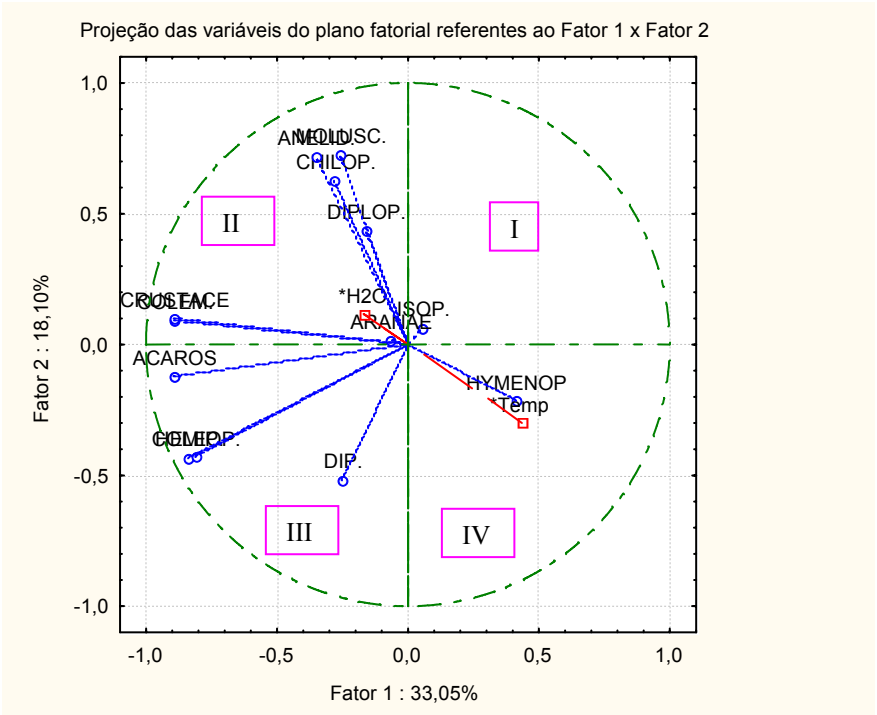


Figura 79 - Gráfico da distribuição da nuvem de variáveis.

Como pode-se observar, na Figura 79, algumas variáveis estão sobrepostas umas às outras. Isso mostra que essas possuem a mesma representatividade no gráfico. Outro fato importante, é que algumas variáveis estão bem próximas ao círculo unitário. Isso mostra que estas são mais representativas, em relação às variáveis que estão mais afastadas.

Conclui-se, também, que as variáveis localizadas nos quadrantes II e III não sofreram influência da umidade, pelo fato da umidade estar localizada no mesmo quadrante que estas variáveis, mas são influenciadas pela temperatura, que está localizada no quadrante oposto, a essas. As variáveis localizadas nos quadrantes I e IV possuem influência da umidade, apenas.

A Figura 80 mostra a caixa de seleção de variáveis e comandos para *ACP.* Seleciona-se *Cases/Plot case factor coordinates, 2D,* e clica-se em *Ok,* para fazer os planos principais, com a nuvem de pontos dos indivíduos (as coletas).

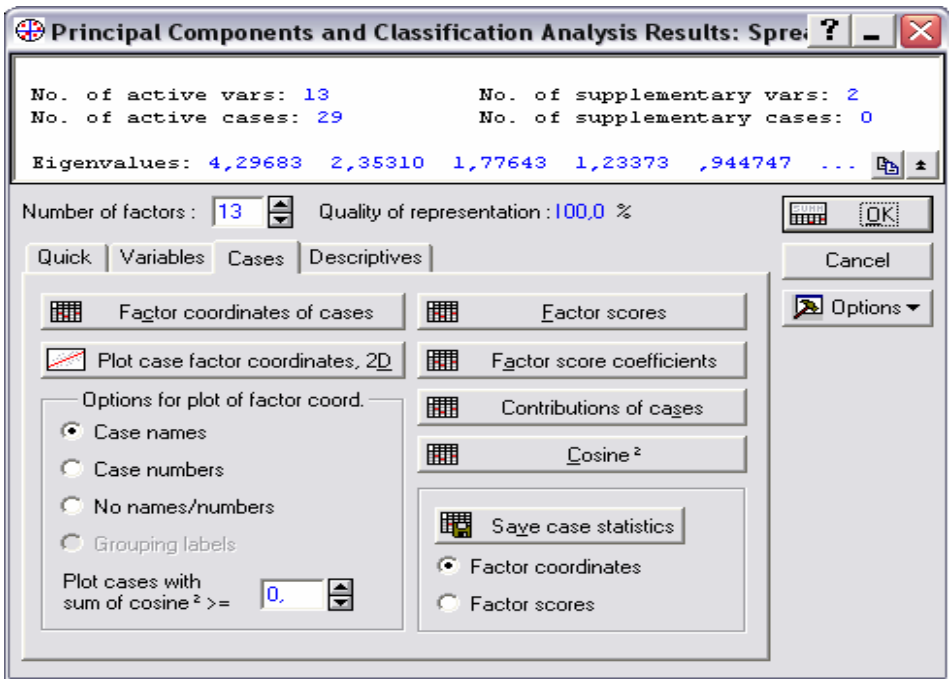


Figura 80 - Caixa de seleção da *ACP.*

A Figura 81 mostra os fatores a serem relacionados para a nuvem de pontos dos indivíduos. Neste caso, relaciona-se *Factor 1* com *Factor 2,* e clica-se em *Ok.*

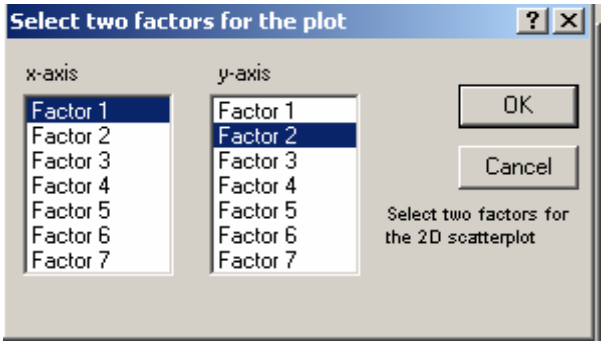


Figura 81 - Caixa de seleção dos fatores para *ACP*.

A Figura 82 mostra o primeiro plano principal, com a nuvem de pontos dos indivíduos.

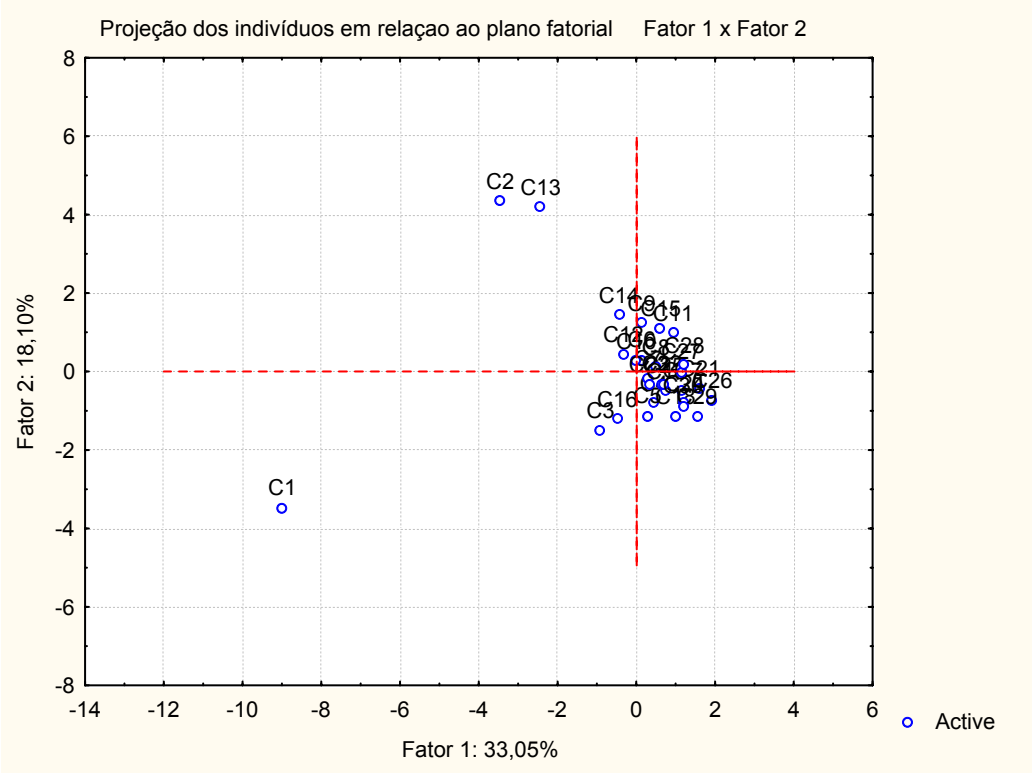


Figura 82 - Gráfico da distribuição da nuvem de pontos (os estados).

Analisando-se o gráfico da Figura 79, da distribuição da nuvem de variáveis em relação ao gráfico da Figura 82, da distribuição da nuvem de pontos, pode-se concluir que as variáveis Colêmbolos, Hemípteros, Coleópteros, Crustáceos e Ácaros são as mais representativas. Isto significa que foram encontradas em maior abundância no solo, em relação a estes dois fatores, e a coleta, que mais contribui na formação da combinação linear do fator 1, é a primeira (C1), pois está representando estas cinco variáveis. Pode-se concluir, ainda, que as coletas que

possuem uma maior contribuição, na formação da combinação linear do fator 2, é a segunda e a décima terceira coleta (C2 e C13), que representam as variáveis Anelídeos, Moluscos, Chilópodas e Diplópodes. O restante dos organismos e coletas não apresentam representatividade significativa, em relação a estes dois fatores.

Nos outros planos principais, que correspondem ao fator 1 *x* fator 3 e fator 1 *x* fator 4, a análise é realizada de forma análoga a esse exemplo.

É importante salientar que a interpretação da *ACP* consiste em definir o que representa cada eixo, em termos de fator, responsável pela ordenação das amostras, do assunto que está sendo estudado. Conforme Valentin (2000), "a interpretação de um eixo deve ser baseada nas coordenadas das variáveis neste eixo, a partir das quais foi elaborada a matriz de correlação que deu origem aos autovetores". Ao realizar *ACP,* deve-se observar os seguintes princípios:

- que uma proximidade maior, ou menor, entre dois pontos-variáveis, no plano, traduz uma maior, ou menor, correlação entre essas variáveis, principalmente quando elas são afastadas do centro do plano;
- a proximidade entre dois pontos-amostra (objeto) traduz uma certa similaridade entre essas duas amostras, em termos de variáveis.

Comentários desse capítulo

Nesse capítulo, foi possível, desenvolver dois exemplos numéricos, utilizando-se dados reais. O primeiro, aplicando-se a técnica de *AA,* utilizou-se os dados referentes aos principais produtos que compõe a produção nacional de grãos, no período de 1995 a 2002. O segundo exemplo aplicou-se a técnica de *ACP* e *AF,* cujos dados eram referentes a 30 coletas da fauna edáfica do solo, no período de junho de 2004 a janeiro de 2005. Encontra-se, também, descrito, neste capítulo 4, como realizar as interpretações pertinentes a cada etapa da análise. Consta, ainda, nesse, todas as etapas necessárias para que seja possível desenvolver as técnicas de análise de agrupamentos, análise de componentes principais e análise fatorial.

5 CONCLUSÕES E RECOMENDAÇÕES

Embora a estatística multivariada tenha surgido por volta de 1901, apenas nos dias de hoje consegue-se desenvolver e aplicar essa técnica, pois sem o auxilio de programas computacionais não seria possível realizar tão rápido, e com tanta clareza, os gráficos que possibilitam estudar o inter-relacionamento das variáveis.

Pode-se verificar, no decorrer da pesquisa, que as técnicas de análise de agrupamentos, e análise de componentes principais, são técnicas matemáticas, com grande fundamentação na álgebra e na geometria, o que muitas vezes faz com que os estatísticos não considerem como técnica estatística. Por outro lado, figuram, quase sempre, em congressos nacionais e revistas especializadas, que tratam de assuntos sobre estatística.

A análise fatorial, que muitas vezes é confundida com análise de componentes principais, pelo fato de um dos modos de extração de fatores ser a de componentes principais, é considerada uma técnica estatística, pois ela pressupõe a existência de um modelo, permite que se faça inferências e cumpre com algumas pressuposições básicas sobre as variáveis em análise, como a multinormalidade dos dados.

Nos dias atuais, o uso dessas técnicas está bastante consolidado, mas deve-se ter o cuidado de que não basta se observar um conjunto de variáveis e aplicar técnicas multivariadas, simplesmente, com o intuito de apresentar a técnica e valorizar a pesquisa que se está realizando. Há a necessidade de que exista uma estrutura de correlação entre as variáveis, pois, se as mesmas não estiverem ligadas entre si, tem-se que utilizar uma análise univariada, uma vez que esta, se bem aplicada, é capaz de fornecer um nível muito bom de informação.

A estatística univariada, em nenhum momento deve ser dispensada, quando se realiza um trabalho estatístico, pois é por meio da análise exploratória de dados que será possível conhecer as variáveis em estudo. Como se sabe, a análise multivariada é uma técnica exploratória e, devido a isso, a análise univariada será útil, também, para realizar um estudo confirmatório.

Com o material didático, que está sendo apresentado, fez-se uma ampla revisão de literatura, levando-se em consideração textos clássicos e atuais, pois procura-se revelar, ao máximo, essa técnica, que, muitas vezes, é obscura para os alunos, pesquisadores e profissionais que a utilizam. O uso do *software* foi

indispensável, pois sem ele não seria possível a realização dos estudos de caso. Vale lembrar que, trabalhando com programas diferentes existe, uma similaridade entre eles. Isto é, ao se saber bem interpretar os resultados, não se terá problemas ao utilizar outros programas.

Devido à crescente procura, que tem ocorrido no Departamento de Estatística, de vários cursos da UFSM, e de outras instituições, para o assessoramento em análise de dados, e em busca de material didático que esteja disponível para pesquisas na área de análise multivariada, desenvolve-se este material, que traz, passo a passo, as técnicas da análise multivariada, análise de agrupamentos, análise fatorial e análise de componentes principais, pois sabe-se que muitos materiais existem e mostram como aplicar as técnicas, mas poucos dizem como estas são desenvolvidas.

A estatística, por ser multidisciplinar, está inserida em várias áreas do conhecimento, por isso faz-se necessário a sua aplicação, o seu entendimento e sua interpretação como ferramenta de pesquisa.

O desenvolvimento deste material servirá de suporte, tanto para profissionais da área da educação, bem como de outras áreas do conhecimento e pesquisadores que se utilizam dessas técnicas.

O objetivo deste trabalho foi elaborar um material didático contemplando a teoria e a prática de algumas técnicas da análise multivariada, voltado às necessidades de todos aqueles estudantes, ou profissionais, que necessitem conhecer as possibilidades de aplicação dessa ferramenta estatística para seu trabalho. As conclusões, que são expostas a seguir, referem-se aos diferentes aspectos do trabalho desenvolvido. Em relação aos aspectos teóricos, aqui discutidos, servirão de suporte para as atividades desenvolvidas neste trabalho e em trabalhos subseqüentes.

Os exemplos práticos foram elaborados de forma clara, passo a passo, para que todos que fizerem uso deste material possam compreender em que condições e como poderão ser aplicadas as técnicas de análise de agrupamentos, análise fatorial e análise de componentes principais, bem como interpretar os resultados obtidos nas análises.

Este material poderá ser utilizado por todos que necessitem analisar base de dados relativamente complexas, ou seja, espaços de dimensão iguais ou superiores ao R^4, nos quais deve existir altas correlações entre as variáveis. Mostrou-se,

também, como interpretar essas variáveis, para que todos possam utilizar com segurança os métodos da estatística multivariada.

Em suma, o trabalho desenvolvido mostra ter atingido os objetivos aqui propostos, seja em relação à elaboração da teoria, seja em relação aos exemplos práticos, o que proporcionará uma maior agilidade e praticidade no que se refere ao conhecimento e à utilização da metodologia apresentada.

As aplicações das técnicas multivariadas são muito amplas. Logo, deixa-se como sugestão o desenvolvimento de pesquisas em outras áreas do conhecimento.

Em relação ao uso de programas utilizados, para aplicação da técnica, sugere-se que outros programas sejam utilizados, assim como os *softwares*, pois, desta forma, estimula-se o pesquisador a criar as suas próprias rotinas computacionais.

Técnicas como Análise de Discriminante, Análise de Correspondência e Correlação Canônica ainda carecem de maiores estudos e divulgação de um material didático, contemplando a teoria e a prática, material, este, que será de fundamental importância para a ciência.

6 BIBLIOGRAFIA

BUSSAB, W. O.; MIAZAKI, É. S.; ANDRADE, D. F. Introdução à análise de agrupamentos**:** In: SIMPÓSIO BRASILEIRO DE PROBABILIDADE E ESTATÍSTICA, 9**.,**1990, São Paulo. **Resumos**...São Paulo, 1990.

CRUZ, C. D. **Aplicação de algumas técnicas multivariadas no melhoramento de plantas.** 1990. Tese (Doutorado) – ESALQ, Piracicaba, 1990.

CATTEL, R. B. The scree test for the number of factors. In: **Multivariate behavior research.** v.1, p. 245-276, 1966.

FERREIRA, D. F. **Análise multivariada.** Lavras, 1996.

HAIR, J. F.; *et al*. **Análise multivariada de dados.** 5 ed. Porto Alegre, 2005.

JACKSON, J.E. Principal componets and factor analysis: Part I - principal componets. **Journal of Quality Technology**, October. v.12, n.4, p.201-213.

JOHONSON, R.A., WICHERN, D.W. **Applied multivariate statistical analysis.** 3 rd ed. New Jersey: Prentice-Hall, 1992.

LOPES, L. F. D. **Análise de componentes principais à confiabilidade de sistemas complexos.** 2001. Tese (Doutorado em Engenharia de Produção) – Universidade Federal Santa Catarina, 2001.

MAGNUSSON, W. E., MOURÃO, G. **Estatística sem matemática:** a ligação entre as questões e a análise. Curitiba: 2003.

MALHOTRA, N. K. **Pesquisa de marketing:** uma orientação aplicada. Porto Alegre: Bookman, 2001.

MANLY, B. F. J. **Multivariate statistical methods:** a primer. London: Chapman and Hall, 1986.

MARDIA, K.V.; KENT, J. T. and BIBBY, J. M. **Multivariate analysis**. London: Academic, 1979.

NETO, M. M. J. Estatística multivariada. **Revista de Filosofia e Ensino**. 9 maio 2004. Disponível em: http://www.criticanarede.com/cien_estatistica.html. Acesso em: 9 maio 2004.

PLA, L. E. **Analysis multivariado: Método de componentes principales.** Departamento de Producción Vegetal. Área de Ciências del agro Y del mar. Universidad Nacional Experimental Francisco de Miranda. Coro, Falcón, Venezuela. Secretaria General de la Organización de Los Estados Americanos, Washington, D. C. 1986.

PEREIRA, J. C. R. **Análise de dados qualitativos**: estratégias metodológicas para as ciências da saúde, humanas e sociais. São Paulo: Edusp, 2001.

REGAZZI, A. J. **INF 766 - Análise multivariada.** Viçosa: Universidade Federal de Viçosa, Centro de Ciências Exatas e Tecnológicas. Departamento de Informática, 2001. 166p. Apostila de disciplina.

REIS, E. **Estatística multivariada aplicada.** Lisboa, 1997.

SOUZA, A. M. **Monitoração e ajuste de realimentação em processos produtivos multivariados.** 2000.Tese (Doutorado em Engenharia de Produção) – Universidade Federal Santa Catarina, 2000.

VALENTIN, J. L. **Ecologia numérica:** uma introdução à análise multivariada de dados ecológicos. Rio de Janeiro: Interciência, 2000.