UM PROCEDIMENTO PARA CALCULAR ÍNDICES A PARTIR DE UMA BASE DE DADOS MULTIVARIADOS

Lucia Silva Kubrusly
Instituto de Economia /UFRJ
Av. Pasteur, 250 – URCA
CEP 22290-240 – Rio de Janeiro – RJ

Resumo

Este trabalho trata do problema de se estabelecer um índice *I* que possibilite ordenar um conjunto de *n* objetos, segundo critério definido por um conjunto de *m* variáveis. De um modo geral, é necessário escolher o conjunto de variáveis adequadas, e também, os pesos atribuídos a cada variável. A Análise de Grupamento é usada para seleção de variáveis, e a Análise de Componentes Principais é usada para fornecer as ponderações. São apresentadas duas aplicações do procedimento proposto.

Palavras-chave: números índices, análise de grupamento, análise de componentes principais, análise multivariada.

Abstract

This paper presents a solution to the problem of finding an index number *I* for ranking a set of *n* objects, according to criteria defined by a set of *m* variables. In general, it is necessary to choose a suitable set of variables, and weights for each variable. Cluster Analysis is used for variable selection, and Principal Component Analysis is used to attain weights. Two applications are presented.

Keywords: index-number, cluster analysis, principal component analysis, multivariate analysis.

ISSN 0101-7438

Introdução

O problema de se estabelecer índices para ordenar objetos pode ser facilmente entendido através dos dois exemplos a seguir:

1. Considere um conjunto de pacientes em um hospital sobre os quais são observadas variáveis relacionadas a sua saúde. Suponha que se deseja ordenar esses pacientes segundo seu "grau de saúde".
2. Considere que foram observadas variáveis relacionadas com poluição da água em diversos pontos do litoral do Estado do Rio de Janeiro, e que se deseja ordenar esses pontos segundo seu "grau de poluição".

Esses são exemplos em que surge a necessidade da construção de um índice para ordenar objetos (pacientes, pontos no litoral, etc.) segundo um conjunto de variáveis observadas. O problema de construção de índices tem sido resolvido buscando uma forma adequada para essa ordenação. A abordagem desse trabalho tem base nos trabalhos de Kendall (1939), e posteriormente Theil (1960), que estabeleceu um índice linear ótimo no sentido da quantidade de informação contida nos dados. Aqui também foi utilizada a técnica de Análise de Componentes Principais para obter a participação de cada variável na construção do índice. Além disso, neste trabalho, é tratado também o problema de seleção das variáveis usadas para construção do índice. Esta seleção é feita com auxilio de outra técnica de análise multivariada, a técnica de Análise de Grupamento, que é usada no sentido de descartar variáveis consideradas pouco importantes do ponto de vista estatístico. Além desses aspectos, é dada uma ênfase especial na interpretação do índice, isto é, no significado da ordenação obtida.

O trabalho foi dividido em duas partes. Na seção 1 são detalhados os passos para a construção do índice. Na seção 2 o procedimento proposto é aplicado em um problema de ordenação de países da América Latina, tendo como critério variáveis econômicas, e em um problema de ordenação das regiões brasileiras tendo como critério os Indicadores Sociais Mínimos do IBGE.

1. Definição do Índice

Seja um conjunto de objetos $O_1,\ldots, O_n$ que se deseja ordenar, segundo características associadas à um conjunto de variáveis $X_1,\ldots, X_p$. Assim, a cada objeto O_j associamos um valor

$$I_j = \sum_i a_i x_{ij} \qquad \text{onde:}$$

x_{ij} é o valor da i-ésima variável observada para o j-ésimo objeto;

a_i é o peso da i-ésima variável (importância da variável na construção do índice I);

Na tentativa de construção do índice I, é necessário primeiramente selecionar variáveis, e posteriormente ponderá-las. Esses dois aspectos do problema serão abordados nesse trabalho, usando duas técnicas de análise multivariada. A Análise de Grupamento será usada para seleção das variáveis, e a Análise de Componentes Principais será usada para ponderá-las.

1.1 Seleção de Variáveis

Considere um conjunto inicial de k variáveis, a partir do qual um subconjunto de $n < k$ será escolhido. As n variáveis escolhidas deverão guardar as principais características observadas nas k variáveis iniciais. Isto é, serão excluídas as variáveis "pouco importantes do ponto de

vista estatístico". Serão consideradas "pouco importantes" aquelas variáveis que, se excluídas de uma análise estatística, não alteram seu resultado. A análise estatística escolhida foi a Análise de Grupamento, pois esta fornece uma estrutura de semelhança entre os objetos que poderá servir de auxílio na avaliação do índice obtido. Essa abordagem para seleção de variáveis já foi utilizada anteriormente em Kubrusly (1992). Outra abordagem para o problema de seleção de variáveis pode ser encontrado em Tanaka & Mori (1997).

A Análise de Grupamento é uma técnica de análise estatística multivariada que procura grupar objetos semelhantes segundo o critério dado pelo conjunto de variáveis observadas. O modelo de Análise de Grupamento foi descrito por Lucas (1982) da seguinte forma:

Seja $X = \{X_1,...,X_p\}$ um conjunto de variáveis, e

$O = \{O_1,...,O_n\}$ o conjunto de objetos que se deseja grupar.

Com base no conjunto *X*, determinar uma partição de *O* em grupos g_i tal que:

$se\ O_r\ e\ O_s \in g_i \Rightarrow O_r\ e\ O_s\ são\ semelhantes,$

$se\ O_r \in g_i \quad e \quad O_s \in g_j \Rightarrow O_r\ e\ O_s\ são\ distintos.$

Para solução desse problema é necessário calcular as distâncias entre os objetos no espaço das variáveis. Essas distâncias fornecem as medidas de similaridade entre os objetos. Mais detalhes sobre a escolha das métricas, e escolha de métodos para análise de grupamento, pode ser visto em Barros (1992).

Voltando a questão da seleção de variáveis, o procedimento adotado nesse trabalho pode ser assim descrito:

1. Obter uma solução de análise de grupamento para o conjunto inicial de k variáveis;
2. Faça $i = 1$;
3. Excluir X_i do conjunto inicial, e repetir a análise. Comparar com o resultado obtido no passo 1;
4. Se os resultados são similares, exclua X_i do conjunto de variáveis selecionadas. Senão mantenha;
5. Faça $i = i + 1$, se $i = k + 1$ pare, senão volte ao passo 3.

1.2 As Ponderações

Na construção de índices, muitos métodos são utilizados para ponderar as variáveis. O objetivo é obter pesos que traduzam a importância das variáveis. Em análise estatística, uma medida de importância muito usada é a variância. De certa forma, a variância traduz o a informação contida na variável. Ao construirmos um índice como uma combinação linear de variáveis, é desejável que este tenha a maior variância possível, ou seja, que contenha o máximo de informação fornecida pelo conjunto de variáveis selecionadas. Um método que cria combinações lineares com essa propriedade (máxima variância), é a Análise de Componentes Principais. Por isso esta será a técnica utilizada neste trabalho para construção do índice.

O modelo de Análise de Componentes Principais será empregado na sua forma clássica como foi descrito em Johnson & Wichern (1992). Pesquisas mais recentes desse modelo podem ser encontradas em Cazes, Chouakria, Diday & Schektman (1997) ou em Le Cercle Factoriel (1997).

Seja $X = (X_1,\ldots, X_p)$ um conjunto de variáveis observadas sobre n objetos. As componentes principais C_i são definidas como:

$C_i = \sum_j a_{ij} X_j$, sujeito a:

var(C_i) = máxima

$\sum_j a_{ij}^2 = 1$

cor $(C_i, C_{i'}) = 0$ para $i \neq i'$, $i = 1,\ldots, p$.

A solução do modelo é dada pela decomposição da matriz de covariância (ou correlação) em seus auto-valores e auto-vetores. O índice I será identificado com a primeira componente principal C_1, pois esta é a combinação linear das variáveis que possui a maior variância.

Na aplicação desse método para construção de índices, a solução será tanto melhor quanto maior for a proporção da variância total contida na primeira componente C_1. Em alguns casos é possível a construção de índices bidimensionais, desde que facilmente interpretáveis, identificados com as duas primeiras componentes principais (veja Theil, 1960). Nesses casos é necessário lembrar que a primeira componente (que fornece o primeiro índice) será mais importante que a segunda (que fornece o segundo índice) devido à sua maior variância. Intervalos de confiança para os índices podem ser estudados a partir dos intervalos de confiança definidos para o modelo de análise de componentes principais. Um ponto de partida para esse estudo pode ser encontrado em Jolliffe (1986).

2. Aplicações

2.1 Ordenação de Países da América Latina

Esta aplicação tem como objetivo ordenar países da América Latina, segundo algumas variáveis sócio-econômicas. O conjunto de dados iniciais é formado por 8 variáveis sócio-econômicas observadas em 15 países. Os demais países da América Latina foram excluídos pois não apresentavam dados suficientes nas fontes pesquisadas. Abaixo listamos os países e as variáveis utilizadas.

Os Países:

Argentina	Costa Rica	México
Bolívia	El Salvador	Paraguai
Brasil	Equador	Peru
Chile	Guatemala	Uruguai
Colômbia	Honduras	Venezuela

As Variáveis Iniciais:

PIB – Taxa média de crescimento anual do produto nacional bruto *per capita* (1991 – 1996).

INFLAÇÃO – Taxa média da variação anual do INPC (1995 – 1996).

DESEMPREGO – Taxas médias anuais dos indicadores de desemprego urbano (1995 – 1996).

JUROS EXTERNOS – Relação entre os juros externos totais e as exportações de bens e serviços (1995 – 1996).

CÂMBIO – Índices de taxas de câmbio efetivo real, ponderado pela importância relativa das exportações para cada parceiro comercial.

JUROS – Taxas reais de juros deflacionadas pela variação dos preços ao consumidor.

RENDA – Renda *per capita* ($).

GASTOS SOCIAIS – Gasto Social *per capita* ($); média entre 1994 e 1995.

Fontes:

CEPAL – Panorama Social de América Latina, 1996.

WORLD BANK – World Development Report, 1996.

A Seleção de Variáveis usando Análise de Grupamento

Para a solução de Análise de Grupamento tendo como critério as oito variáveis sócio-econômicas, foi usado o *software* SPSS 8.0, com o método Ward, e distância euclidiana. A representação pelo dendograma está dada abaixo:

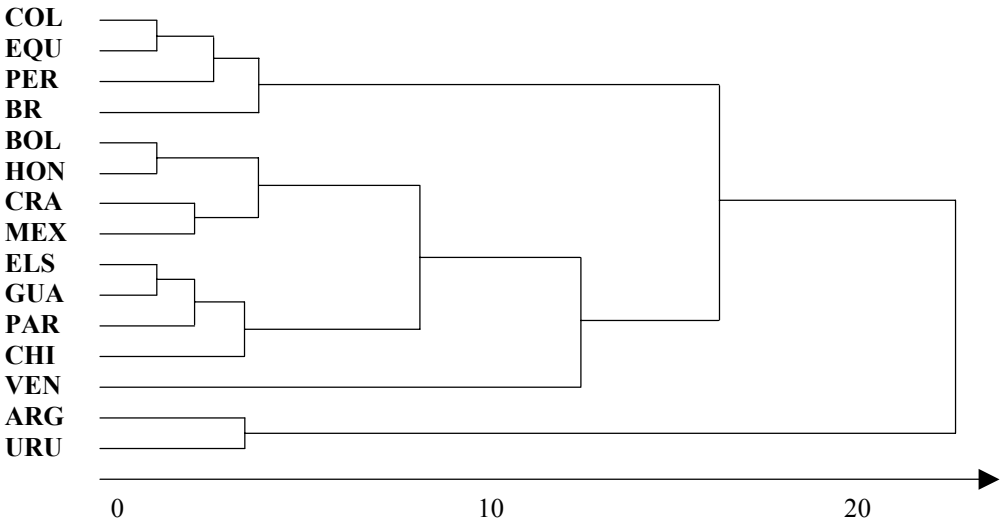

Figura 1 – A.G. variáveis iniciais

Nesta análise destaca-se três grupos de países semelhantes. O primeiro contém COLÔMBIA, EQUADOR, PERU, BRASIL. No segundo grupo estão BOLÍVIA, HONDURAS, COSTA RICA, MÉXICO, EL SALVADOR, GUATEMALA, PARAGUAI, CHILE, VENEZUELA, sendo que esta última aparece um pouco afastada dos demais. O terceiro grupo é formado por ARGENTINA e URUGUAI.

Conforme o procedimento descrito anteriormente, foram retiradas uma a uma as variáveis e foi primeiramente obtida uma solução de Análise de Grupamento similar a anterior com sete variáveis (excluindo "JUROS"), e a seguir foi possível ainda retirar "DESEMPREGO" sem

que a solução de Análise de Grupamento fosse fortemente afetada. Abaixo mostramos a solução com as seis variáveis selecionadas: PIB, INFLAÇÃO, JUROS EXTERNOS, CÂMBIO, RENDA, GASTOS SOCIAIS.

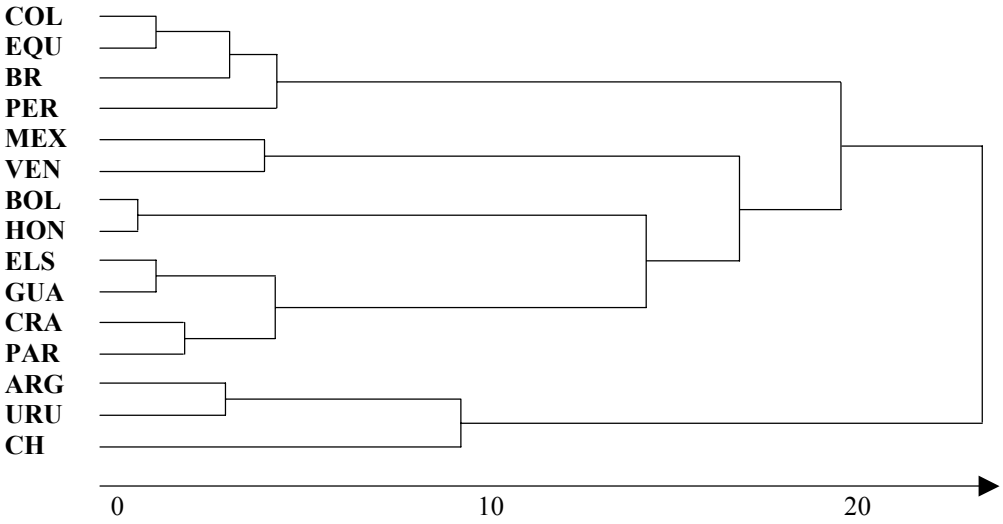

Figura 2 – A.G. variáveis selecionadas

Comparando-se essa solução com a anterior observa-se os mesmos grupos de países semelhantes, exceto o CHILE que nessa última solução vem se juntar a ARGENTINA e URUGUAI.

A Construção do Índice via Análise de Componentes Principais

A Análise de Componentes Principais para as seis variáveis observadas sobre os países foi realizada, e foram mantidas as duas primeiras componentes que correspondem a 62% da variância total. Uma discussão sobre o número de componentes mantidas na Análise de Componentes Principais, pode ser encontrada em (Johnson & Wichern, 1992). O resultado, obtido com o auxilio do *software* SPSS 8.0, está na Tabela 1:

Tabela 1 – Componentes Principais para as variáveis econômicas

	C1	*C2*
PIB	0,720	-0,495
INFL	-0,318	0,781
CAM	0,461	-0,254
JUREX	0,428	-0,091
RENDA	0,851	0,436
GASTSOC	0,819	0,476
VARIÂNC.	2,41	1,34
%VAR	40%	22%

De acordo com esse resultado a primeira componente mantém 40% da informação contida nos dados (medida pela variância). Os coeficientes da tabela medem as correlações entre as variáveis e as componentes. Dessa forma a primeira componente fornecerá um índice apresentando valores altos para países com maior PIB, RENDA, GASTO SOCIAIS, e menor INFLAÇÃO. As ponderações das variáveis serão os *coeficientes de escore* de C1, isto é, seu auto-vetor dividido pela raiz quadrada do auto-valor correspondente. Assim as variáveis recebem os seguintes pesos:

PIB	0,299
INFL	-0,132
CAM	0,178
JUREX	0,191
RENDA	0,353
GASTSOC	0,340

Finalmente, usando a definição para o índice:

$I_j = \sum_i a_i x_{ij}$ onde:

x_{ij} é o valor da i-ésima variável observada para o j-ésimo país,

a_i é o peso da i-ésima variável,

é possível ordenar os países conforme a lista abaixo:

ARGENTINA	2,4
URUGUAI	1,42
CHILE	0,97
BRASIL	0,47
PERU	0,40
COLÔMBIA	0,07
COSTA RICA	-0,13
EL SALVADOR	-0,27
MÉXICO	-0,33
EQUADOR	-0,35
GUATEMALA	-0,70
BOLÍVIA	-0,86
VENEZUELA	-0,86
PARAGUAI	-1,10
HONDURAS	-1,13

Observando esse resultado é possível perceber os três grupos apontados pela Análise de Grupamento. Confirma-se a semelhança de ARGENTINA, URUGUAI, e CHILE, e também BRASIL, PERU, e mais fracamente COLÔMBIA. Os demais países, que formavam um grande grupo na Análise de Grupamento, aparecem aqui todos com índices negativos. Dessa forma, as duas análises fornecem resultados coerentes.

2.2 Ordenação de Regiões Brasileiras

Esta aplicação tem como objetivo ordenar as cinco regiões geográficas brasileiras, segundo variáveis extraídas dos Indicadores Sociais Mínimos do IBGE (1998). Abaixo listamos as regiões e as variáveis utilizadas.

As Regiões

NORTE – Rondônia, Acre, Amazonas, Roraima, Pará, Amapá, Tocantins.

NORDESTE – Maranhão, Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Sergipe, Bahia.

SUDESTE – Minas Gerais, Espírito Santo, Rio de Janeiro, São Paulo.

SUL – Paraná, Santa Catarina, Rio Grande do Sul.

CENTRO OESTE – Mato Grosso, Mato Grosso do Sul, Goiás, Distrito Federal.

As Variáveis Iniciais

RMEDIA – Rendimento médio mensal familiar.

IGINI – Índice de Gini.

TXATIV – Taxa de atividade das pessoas de 15 a 65 anos.

ESTUDOS – Média de anos de estudo das pessoas de 10 anos ou mais de idade.

TXANALF – Taxa de analfabetismo das pessoas de 15 anos ou mais de idade.

NPESSOAS – Número médio de pessoas por domicílio.

AGTRAT – Percentual de domicílios com água tratada.

LXCOLET – Percentual de domicílios com lixo coletado.

A Seleção de Variáveis Usando Análise de Grupamento

Para a solução de Análise de Grupamento tendo como critério as variáveis iniciais, foi usado o SPSS 8.0, com o método Ward, e distância euclidiana. A representação pelo dendograma está dada abaixo:

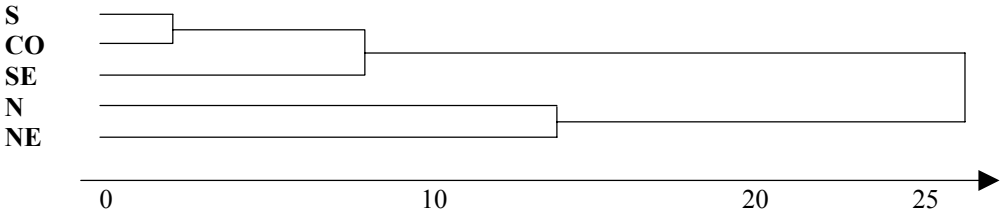

Figura 3 – A.G. variáveis iniciais

Esse resultado apresenta dois grupos bem distintos, separando as regiões Norte e Nordeste das demais. Nota-se também que as semelhanças entre as regiões Sul, Centro Oeste, e Sudeste, é mais acentuada que a semelhança entre as regiões Norte e Nordeste.

Conforme o procedimento descrito anteriormente, foram retiradas uma a uma as variáveis até reduzir o conjunto de variáveis iniciais para o seguinte conjunto de variáveis selecionadas: RMEDIA, TXATIV, ESTUDO, AGTRAT. A nova solução para Análise de Grupamento, idêntica a solução anterior, está mostrada abaixo:

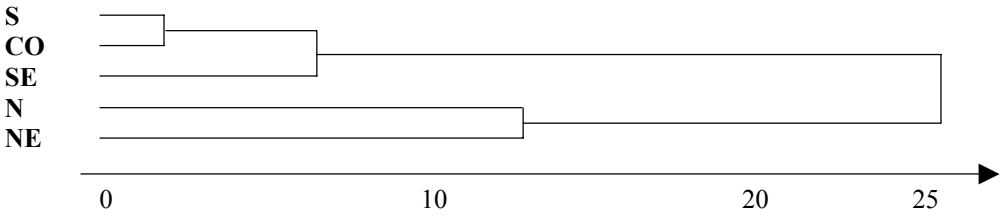

Figura 4 – A.G. variáveis selecionadas

A Construção do Índice via Análise de Componentes Principais

Para a solução do modelo de Análise de Componentes Principais foi usado o *software* SPSS 8.0. A ACP para as quatro variáveis observadas sobre as regiões geográficas forneceu o seguinte resultado:

Tabela 2 – Componentes Principais para os indicadores sociais

	C1
RMED	0,988
TXATIV	0,557
ESTUDO	0,944
AGTRAT	0,934
VARIÂNC	3,05
%VAR	76%

Nesta solução, a primeira componente é responsável por 76% da variância total, e fornecerá um índice com as seguintes ponderações para as variáveis:

RMEDIA	0,323
TXATIV	0,183
ESTUDO	0,309
AGTRAT	0,306

Utilizando essas ponderações para construção do índice, foi possível ordenar as regiões geográficas brasileiras conforme está mostrado abaixo:

SUDESTE	0,99
SUL	0,84
CENTRO OESTE	0,16
NORTE	-0,63
NORDESTE	-1,37

Também nessa aplicação o índice obtido está coerente com o resultado da Análise de Grupamento, que separa as regiões SE, S, CO das regiões N e NE.

Comentários Finais

Na abordagem usada nesse trabalho para construção de índice, duas técnicas de análise multivariada foram utilizadas, a Análise de Grupamento e a Análise de Componentes Principais. Em muitos casos pode ser interessante aproveitar o poder altamente descritivo dessas duas técnicas para interpretação do índice resultante. Em particular, a Análise de Componentes Principais analisa a matriz de correlação das variáveis, e por seu resultado é possível saber se um único índice é adequado para a ordenação, ou se o conjunto de variáveis fornece duas ou mais dimensões igualmente importantes. Nas aplicações apresentadas neste trabalho, uma única componente foi escolhida para construção do índice devido a sua variância, sensivelmente maior que a da segunda componente. Além disso, foi mostrada a coerência dos resultados das duas técnicas em ambas aplicações. Na verdade, o índice obtido ordenou os grupos apontados pela Análise de Grupamento.

Na aplicação ordenando os países da América Latina, o índice obtido pode ser considerado como um índice de desenvolvimento econômico, já que as variáveis mais importantes (com maior peso) na sua construção foram PIB, RENDA, GASTO SOCIAL, e considerando também que a variável INFLAÇÃO teve peso negativo.

Na segunda aplicação, o índice obtido pode ser associado ao desenvolvimento econômico e social de cada região geográfica brasileira. Embora o resultado tenha sido o esperado, a análise de grupamento indicou claramente que o grupo com maior desenvolvimento sócio-econômico (SE, S, CO) tem um grau de semelhança bem maior que o grupo formado pelas regiões menos desenvolvidas (N, NE). Por outro lado, por meio do índice foi possível "quantificar" essas semelhanças e diferenças.

É importante ressaltar que embora o procedimento proposto seja inteiramente baseado em duas técnicas de análise, a sua utilização não deve ser desvinculada de uma interpretação adequada. Daí a importância da escolha das variáveis incluídas no índice. Na verdade não há dificuldade técnica em se construir um índice com 100 ou mais variáveis. Alguns analistas adotam essa postura, na esperança de "não perder informação alguma", mas o resultado será um índice cujo significado dificilmente será percebido.

Referências Bibliográficas

(1) Barros, A.C. (1992). Relações Intersetoriais em Matrizes de Insumo-Produto: uma abordagem da análise de agrupamento. Tese M. Sc., COPPE / UFRJ.

(2) Cazes; Chouakria; Diday & Schektman (1997). Extension de l'Analyse en Composantes Principales à des données de type intervale. *Revue de Statistique Appliquée*, vol. XLV, n. 3.

(3) CEPAL (1996). *Panorama Social de América Latina*.

(4) Instituto Brasileiro de Geografia e Estatística (1998). *Indicadores Sociais Mínimos*: *Educação e Condições de Vida, Trabalho e Rendimento.*

(5) Johnson, R.A. & Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis.* Prentice-Hall, Inc., A Simon & Schuster Company Upper Saddle River, New Jersey.

(6) Jolliffe, I.T. (1986). *Principal Component Analysis.* Springer Verlag, Berlin.

(7) Kendall, M.G. (1939). The geographical distribution of crop productivity in England. *J. Roy. Statist. Soc.*, **102**(21).

(8) Kubrusly, L.S. (1992). Utilização de Técnicas de Análise Multivariada para Redução de Variáveis num Problema de Controle Ecológico. *Revista Brasileira de Estatística*, ano 53, n. 199/200.

(9) Le Cercle Factoriel (1997). Exploitation Graphique des Plans Factoriel. *Revue de Statistique Appliquée*, vol. XLV, n. 3.

(10) Lucas, L.C.S. (1982). Análise de Grupamentos. *Revista Brasileira de Estatística*, ano 43, n. 172.

(11) SPSS Base 8.0 (1998). Applications Guide.

(12) Tanaka, Y. & Mori, Y. (1997). Principal Component Analysis based on a subset of variables: variable selection and sensitivity analysis. *American Journal of Mathematics and Management Science*, **17**(1).

(13) Theil, H. (1960). Best Linear Index Numbers of Prices and Quantities, *Econometrica*, **28**(2).

(14) WORLD BANK (1996). World Development Report.