UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE MATEMÁTICA

DEPARTAMENTO DE ESTATISTICA

ANÁLISE DE CORRESPONDÊNCIA

Autor: NORMA MARTINEZ DE SOUZA

Orientador: Professora JANDYRA M. G. FACHEL

Monografia apresentada para obtenção do título de Bacharel em Estatística.

Porto Alegre, dezembro de 1990.

5220

UFRGS
SISTEMAS DE BIBLIOTECAS
BIBLIOTECA SETORIAL DE MATEMÁTICA

A G R A D E C I M E N T O S

Gostaria de agradecer a todos aqueles que, de um modo ou de outro, colaboraram na execução deste trabalho. Agradeço em particular:

À professora Jandyra M. G . Fachel pela orientação prestada à presente monografia.

Aos colegas Alexandre L. Amaro e Márcio M. da Silva pela dedicação ao trabalho de digitação e Ana L. Ebling, pelo trabalho de tradução.

Manifesto também meus agradecimentos à professora Sídia M. C. Jacques e a Nilton C. M. de Araújo pela autorização para usar os arquivos de dados apresentados nesta monografia.

UFRGS
SISTEMAS DE BIBLIOTECAS
BIBLIOTECA SETORIAL DE MATEMÁTICA

I N D I C E

UFRGS
SISTEMAS DE BIBLIOTECAS
BIBLIOTECA SETORIAL DE MATEMÁTICA

1. INTRODUÇÃO

1.1. Apresentação

Pelo amplo campo de aplicação que as técnicas de análise de dados possuem, cada vez mais suas idéias básicas são difundidas. Basta um simples folhear em qualquer publicação estatística atual, para notarmos que os métodos de análise multivariada encontram-se em constante desenvolvimento.

Esta monografia trata de uma técnica de análise de dados especialmente útil para análise de várias variáveis categóricas, a análise de correspondência.

A análise de correspondência é uma técnica exploratória cujo produto final é a representação gráfica das linhas e das colunas de uma tabela de contingência.

Verificada a dependência entre as linhas e as colunas da tabela, o método de análise de correspondência procura a "melhor" representação gráfica simultânea de nuvens de pontos (linhas e colunas), nos planos de projeção formados pelos primeiros eixos fatoriais tomados dois a dois. A interpretação dos resultados é feita com a ajuda de alguns coeficientes como a contribuição absoluta e a contribuição relativa, que relacionam os fatores encontrados na análise com as variáveis. Entretanto, pode ser muito ampla e subjetiva e só o treino e a

UFRGS
SISTEMAS DE BIBLIOTECAS
BIBLIOTECA SETORIAL DE MATEMÁTICA

prática do estatístico e a experiência e o conhecimento do pesquisador podem auxiliar e tornar plausível esta interpretação.

A estrutura deste trabalho é simples e a linguagem é bastante informal. Não há pretensão de aprofundamento matemático e sim de divulgar aos pesquisadores e interessados as idéias mínimas para compreender e saber utilizar o método de análise de correspondência.

1.2. Nota Histórica

É difícil fazer um histórico preciso da técnica de análise de correspondência.

Em 1935, Hirschfeld publicou uma solução algébrica para a correlação entre linhas e colunas de tabelas de contingência. Entre 1940 e 1950 Guttman e Hayashi adotaram essas idéias para o aperfeiçoamento de técnicas escalonadas baseadas na análise de componentes principais e na quantificação de dados qualitativos.

O desenvolvimento dado por Hirschfeld não foi citado por Fisher em 1940 quando este desenvolveu uma técnica semelhante sob o nome de "análise canônica para tabelas de contingência" e assim Fisher é frequentemente apontado como o introdutor do método.

A análise de correspondência foi difundida na comunidade científica pelo estatístico francês Benzécri (1973) sob o nome geral de Análise de Dados e suas idéias foram baseadas em desenvolvimentos anteriores. Grande parte dos trabalhos nesta área são devidos a Benzécri e seus colaboradores entre os quais destacam-se: Lebart, Morineau, Tabard (1977).

Por algum tempo, a bibliografia disponível foi publicada por franceses. Mais recentemente, já existe publicações na literatura inglesa, como os livros de Nishisato (1980), Gifi (1981), Greenacre (1984), Lebart, Morineau & Warwick (1984) e capítulos específicos nos livros de Análise Multivariada. Em portugués, a literatura disponível é a tese de mestrado apresentada no Instituto de Matemática e Estatística da USP por Ana Musetti R. de Souza (1982).

Benzécri (1973, 1977b) introduziu a análise de correspondência (múltipla) como uma técnica de análise de componentes principais simultânea de duas nuvens de pontos ponderados, usando o critério de inércia com a métrica qui-quadrado. As características da abordagem explicam em grande parte o sucesso deste método na França.

Duas concepções simplificam nosso entendimento da abordagem de Benzécri. Uma é a análise de componentes principais (PCA) de uma matriz de dados Z (sujeitos x variáveis) associada com duas métricas M (no espaço dos sujeitos) e N (no espaço das variáveis. A outra é o diagrama de dualidade da tripla (Z, M, N).

A análise de correspondência é muito popular na literatura francesa e atualmente na inglesa, onde o método tem recebido muitas atenções. Existe uma técnica de análise de dados multivariados chamada Análise Fatorial. Devido a grande confusão entre termos, a análise de correspondência é, às vezes citada por este nome pois anteriormente constava na literatura francesa como Análise Fatorial de Correspondência.

Existem vários métodos propostos com estruturas similares que podem ser usados com os mesmos objetivos, desenvolvidos nos diferentes países. Na América existem os métodos de Escalonamento ótimo (Optimal Scaling), Escore ótimo (Optimal Scoring), Escore Apropriado (Appropriate Scoring); no Canadá, Escalonamento Dual (Dual Scaling); na Holanda, Análise de Homogeneidade (Homogeneity Analysis); na França , Análise de Correspondência Múltipla; em Israel, Análise de Escalograma (Scalogram Analysis) e no Japão, Método de Quantificação (Quantification Method).

O histórico da análise de correspondência encontra-se formalmente descrito em Benzécri (1977a, 1977b).

1.3. Noções Sobre Tabelas de Contingência

A análise de correspondência é tipicamente utilizada para analisar uma tabela de contingência, também chamada de tabela cruzada.

A tabela de contingência nada mais é do que o cruzamento de várias categorias onde se conta o número de indivíduos em cada casela. De uma maneira mais formal, uma tabela de contingência a duas dimensões com I linhas e J colunas é simplesmente o resultado da amostragem de um vetor aleatório bidimensional (A,B). Tal vetor deve ser discreto (i.é., categórico) ou ter sido previamente discretizado (i.é., categorizado) em um número finito de ocorrências possíveis através de uma classificação. Ao conjunto de valores de A, convenciona-se chamar conjunto de linhas e o denotaremos por A_i, quanto ao de B, será o conjunto B_j de colunas. Assim a tabela a duas dimensões é uma matriz I x J onde em cada casela (i,j) coloca-se o número de vezes que as categorias A_i, B_j ocorrem simultaneamente.

A tabela de contingência pode ser representada da seguinte maneira:

critério B / critério A	B_1	. . .	B_j	. . .	B_J	marginal de A
A_1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
⋮						
A_i	n_{i1}		n_{ij}		n_{iJ}	$n_{i.}$
⋮						
A_I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
marginal de B	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

onde:

n_{ij} = número de indivíduos classificados simultaneamente em A_i e B_j ,

$n_{i.} = \sum_{j=1}^{J} n_{ij}$ = número de indivíduos classificados em A_i ,

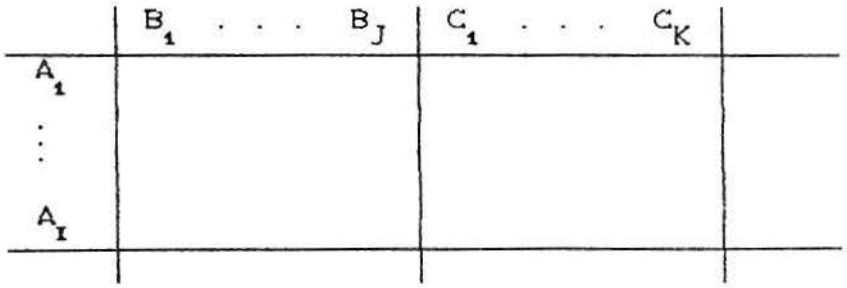
$n_{.j} = \sum_{i=1}^{I} n_{ij}$ = número de indivíduos classificados em B_j .

Então, temos:

$$n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} = \sum_{i=1}^{I} n_{i.} = \sum_{j=1}^{J} n_{.j}$$

1.3.1. Tabelas de Contingência Justapostas

Suponhamos que temos uma tabela de contingência a três dimensões A x B x C . Ao invés de analisarmos da maneira usual, mostramos uma nova abordagem em que as tabelas se justapoem . Podemos estar interessados em explicar uma variável resposta A, então podemos considerar as tabelas A x B e A x C justapostas como na figura



Podemos estender os conceitos acima a tabelas de dimensões maiores. Consideremos então o caso em que temos justaposição de tabelas de contingência bivariadas.

Abaixo segue um exemplo em que a variável A são as carreiras x tipo de escola superior (pública ou particular) nas quais os candidatos classificados no vestibular entraram. As outras variáveis consideradas são : B^1 = nível de instrução do pai; B^2 = nível de instrução da mãe; B^3 = ocupação do pai ; B^4 = renda familiar e B^5 = turno em que o candidato cursou o 2º grau.

Consideremos então, a variável A com categorias ou classes $A_i = i = 1, \ldots, I$ e as variáveis B^q, $q = 1, \ldots, Q$. Cada variável B^q tem J_q categorias ou classes B^q_j. Formaremos a tabela abaixo que é a justaposição das tabelas de contingência, $A \times B^q$.

A \ B	B^1_1	...	$B^1_{J_1}$	B^2_1	...	$B^2_{J_2}$	...	B^Q_1	...	$B^Q_{J_Q}$	
A_1	n^1_{11}		$n^1_{1J_1}$	n^2_{11}		$n^2_{1J_2}$		n^Q_{11}		$n^Q_{1J_Q}$	$n_{1.}$
⋮											
A_I	n^1_{I1}		$n^1_{IJ_1}$	n^2_{I1}		$n^2_{IJ_2}$		n^Q_{I1}		$n^Q_{IJ_Q}$	$n_{I.}$
	$n^1_{.1}$		$n^1_{.J_1}$	$n^2_{.1}$		$n^2_{.J_2}$		$n^Q_{.1}$		$n^Q_{.J_Q}$	n

onde:

n^q_{ij} = número de elemento na classe A_i e classe B^q_j

$n_{i.}$ = número de elemento na classe A_i

$n^q_{.j}$ = número de elementos na classe B^q_j

1.3.2. Tabela na Forma Disjuntiva Completa ou de Incidência

Suponhamos I indivíduos A_i e Q critérios ou perguntas $B^1, \ldots, B^Q$. A pergunta B^q é composta de J_q ítens exclusivos, isto é, cada indivíduo só pode responder a um ítem da pergunta e tem que responder a um deles.

Temos então a tabela

A \ B	B^1_1	. . .	$B^1_{J_1}$	. . .	B^Q_1	. . .	$B^Q_{J_Q}$	
A_1	n^1_{11}	. . .	$n^1_{1J_1}$	. . .	n^Q_{11}	. . .	$n^Q_{1J_Q}$	$n_{1.}$
⋮								
A_I	n^1_{I1}		$n^1_{IJ_1}$		n^Q_{I1}		$n^Q_{IJ_Q}$	$n_{I.}$
	$n^1_{.1}$		$n^1_{.J_1}$		$n^Q_{.1}$		$n^Q_{.J_Q}$	n

onde:

$$n^q_{ij} \begin{cases} 1 \text{ se } A_i \text{ responde } B^q_j \\ 0 \text{ se } A_i \text{ não responde } B^q_j \end{cases}$$

2. ANÁLISE DE CORRESPONDÊNCIA

2.1. Introdução

A análise de correspondência é um algoritmo de redução de dados qualitativos ou categóricos, apresentados em tabelas de contingência ou em tabelas de incidência. Consiste na obtenção de eixos fatoriais, ou fatores, em geral dois ou três, que contenham o máximo possível de informações das variáveis e o objetivo final e fundamental do método é obter a melhor representação simultânea de dois ou mais conjuntos de variáveis através de gráficos, representando cada variável nos planos de projeção formados pelos primeiros eixos fatoriais.

A análise de correspondência permite o estudo de grupos de variáveis que constituem as linhas e as colunas da tabela de contingência. Estamos interessados nos perfis de linha e de coluna, ou seja, nas magnitudes relativas. O uso de uma distância Euclidiana, denominada a distância χ^2, nos permite tratar as linhas e colunas de forma idêntica. Um subproduto é que as projeções de linhas e colunas no novo espaço podem ser ambas delineadas no mesmo gráfico.

Os dados categorizados podem ser codificados pelo escore 1 (presença) ou 0 (ausência) para cada uma das categorias possíveis. Esta codificação será a "codificação

disjuntiva completa". A análise desta tabela é referida como a análise de correspondência múltipla.

A análise de correspondência é descrita abaixo inicialmente com referência a dados apresentados em tabela de contingência. A seguir, mostraremos como o método será estendido para dados disjuntivos completos.

2.2. Propriedades da Análise de Correspondência

Vamos supor sempre que a tabela ou matriz de dados **X** tem m linhas e n colunas, $n \leq m$, e no cruzamento da linha i com a coluna j está o valor x_{ij}, que representa o número de elementos pertencentes a linha i e a coluna j, onde i varia de 1 até m e j varia de 1 até n. (No capítulo anterior utilizamos n_{ij} para representar o número de elementos da linha i e coluna j, por ser esta a notação usada para representar tabelas de contingência. Neste capítulo usaremos a notação x_{ij} ao invés de n_{ij}, visto que esta é a notação usual para Análise de Correspondência).

Consideremos o espaço $\mathbb{R}^n$: neste espaço temos m vetores cada um com n coordenadas (cada linha constitui um vetor de $\mathbb{R}^n$). Se os componentes dos vetores de $\mathbb{R}^n$ são os próprios elementos x_{ij}, as proximidades entre eles podem ficar deturpadas pela falta de padronização dos dados. Não são os

valores brutos que interessam na análise de correspondência e sim os perfis (pesos) das linhas, que serão dados pelas probabilidades condicionais do indivíduo aparecer na coluna j, dado que pertence a linha i, ou seja, x_{ij} deve ser dividido pelo total da linha i. Portanto, cada ponto-linha é considerado como tendo um peso associado. O peso do i-ésimo ponto-linha é dado por $x_i = \sum_j x_{ij}$.

Consideramos os pontos-linhas como tendo as coordenadas $\frac{x_{ij}}{x_i}$ (ou seja $\frac{x_{ij}}{x_i}$ é a j-ésima componente do i-ésimo vetor de $\mathbb{R}^n$), permitindo, portanto, pontos de mesmo perfil serem idênticos (i. é. sobrepostos). É natural que cada ponto tenha um peso proporcional a sua frequência, para que não haja uma falsa idéia da repartição real da população. Em $\mathbb{R}^n$ a distância Euclidiana clássica entre dois pontos linha i e i' será:

$$d^2(i,i') = \sum_j^n \left[\frac{x_{ij}}{x_i} - \frac{x_{i'j}}{x_{i'}} \right]^2$$

Mas se um dos valores, suponhamos o valor na coluna j, for muito grande em relação aos outros, este valor, na distância relativa a esta coluna, será muito grande em relação aos outros e poderá alterar os resultados.

A seguinte distância Euclidiana ponderada, denominada a distância χ^2, é então usada entre os pontos-linha:

$$d^2(i,i') = \sum_j \frac{1}{x_j} \left(\frac{x_{ij}}{x_i} - \frac{x_{i'j}}{x_{i'}} \right)^2$$

UFRGS
SISTEMAS DE BIBLIOTECAS
BIBLIOTECA DE INSTITUTO DE MATEMÁTICA

Consideremos agora o espaço $\mathbb{R}^m$: temos n vetores cada um com m coordenadas (cada coluna constitui um vetor de $\mathbb{R}^m$).

Seguindo o mesmo raciocínio anterior, devemos usar na análise os perfis (pesos) das colunas, que serão dados pelas probabilidades condicionais do indivíduo aparecer na linha i dado que pertence a coluna j , ou seja, x_{ij} deve ser dividido pelo total da coluna j .

As tabelas 2.1. e 2.2. resumem a situação entre os espaços dual, ou seja, mostra as propriedades dos espaços $\mathbb{R}^n$ e $\mathbb{R}^m$ na análise de correspondência.

ESPAÇO $\mathbb{R}^n$

1. m pontos-linha, cada um com n coordenadas.
2. A j-ésima coordenada é $\frac{x_{ij}}{x_i}$.
3. O peso do ponto i é x_i .
4. A distância χ^2 entre i e i' é:

$$d^2(i,i') = \sum_j \frac{1}{x_j} \left(\frac{x_{ij}}{x_i} - \frac{x_{i',j}}{x_{i'}} \right)^2$$

   Portanto esta é a distância Euclidiana com respeito a ponderação $\frac{1}{x_j}$ (para todo j).

5. O critério a ser otimizado: a soma ponderada dos quadrados das projeções, onde a ponderação é dada por x_i (para todo i).

Tabela 2.1.- Propriedades do espaço $\mathbb{R}^n$ na Análise de Correspondência.

ESPAÇO $\mathbb{R}^m$

1. n pontos-colunas, cada um com m coordenadas.
2. A i-ésima coordenada é $\frac{x_{ij}}{x_j}$.
3. O peso do ponto j é x_j .
4. A distância χ^2 entre pontos-colunas j e j' é

$$d^2(j,j') = \sum_i \frac{1}{x_i} \left(\frac{x_{ij}}{x_j} - \frac{x_{ij'}}{x_{j'}} \right)^2$$

   Portanto esta é a distância Euclidiana com respeito à ponderação $\frac{1}{x_i}$ (para todo i) .

5. O critério a ser otimizado: a soma ponderada dos quadrados das projeções, onde a ponderação é dada por x_j (para todo j).

Tabela 2.2.- Propriedades do espaço $\mathbb{R}^m$ na Análise de Correspondência.

O ítem 4, nas tabelas acima, indica que a distância Euclidiana entre os pontos não é a distância Euclidiana clássica mas, é a distância com relação aos pesos especificados.

O ítem 5 indica que a inércia, ao invés da variância, será estudada, isto é, as massas dos pontos são incorporadas no critério a ser otimizado.

O ponto-linha médio é dado pela média ponderada de todos os pontos-linhas:

$$\sum_i x_i \frac{x_{ij}}{x_i} = x_j$$

para $j = 1,2,\ldots,n$. Igualmente o perfil coluna médio tem como i-ésima coordenada x_i .

2.3. Determinação dos Eixos Fatoriais e dos Fatores da Análise

Seja **u** um vetor (coluna) unitário, isto é, $\mathbf{u'u} = 1$. Cada linha de **X** é um vetor de $\mathbb{R}^n$ e o produto **Xu** é uma matriz coluna (vetor) com m componentes, onde cada componente é o produto escalar de uma linha **X** por **u** . Logo, as m componentes de **Xu** são as m projeções da nuvem de pontos sobre **u** .

Para encontrar **u** , devemos maximizar a forma quadrática **u'X'Xu** em relação a **u** , sujeito à condição $\mathbf{u'u} = 1$. Seja $\mathbf{u_1}$ o vetor encontrado e λ_1 a maior raiz característica. O subespaço de duas dimensões que se ajusta "melhor" a nuvem, contém o subespaço formado por $\mathbf{u_1}$. Devemos procurar, então, um vetor $\mathbf{u_2}$, segundo vetor de base do subespaço, que seja ortogonal a $\mathbf{u_1}$ e que maximize $\mathbf{u_2'X'Xu_2}$.

Para achar o máximo de **u'X'Xu** , definindo de uma maneira mais geral, devemos ter :

$$\frac{d}{du}\left[\, u'X'Xu - \lambda\ (\, u'Iu - 1\,)\,\right] = 0$$

onde λ é o multiplicador de Lagrange.

Derivando em relação a u, temos:

$$2\,X'Xu - 2\,\lambda\,Iu = 0 \qquad \text{ou} \qquad X'Xu = \lambda\,Iu$$

Como $u'u = 1$ temos:

$$\lambda = u'X'Xu \ ,$$

onde λ é a máxima raiz característica procurada.

Como I é inversível, $I^{-1}X'Xu = \lambda u$, ou seja u é o vetor característico da matriz $I^{-1}X'X$ correspondente a maior raiz característica λ.

De modo análogo, se estende para um vetor unitário u_α ($u'_\alpha\, I\, u_\alpha = 1$). Chegaremos a:

$$I^{-1}X'X\,u_\alpha = \lambda_\alpha\,u_\alpha$$

e u_α será o α-ésimo vetor característico relativo à raiz característica λ_α.

Em $\mathbb{R}^m$, a procura de um vetor unitário v que se ajusta "melhor" a nuvem de $\mathbb{R}^m$, é obtida de maneira análoga a anterior. A maximização da soma de quadrados das projeções dos m pontos sobre v, que são as m componentes de $X'v$.

Devemos então maximizar $v'XX'v$ com a condição $v'v = 1$. Temos que encontrar os primeiros vetores característicos de XX' correspondentes as maiores raízes características. Então v_α é o α-ésimo vetor característico de XX' correspondente a raiz característica λ_α.

Segundo Murtagh e Heck (1987), as projeções de m perfis em $\mathbb{R}^n$ em um eixo **u** , é dado por :

$$w_i = \sum_j \frac{x_{ij}}{x_i} \quad \frac{1}{x_j} \quad u_j$$

para todo i (note como o produto escalar, usado aqui, está proximamente relacionado com a definição de distância do ítem 4 da tabela 2.1.). Seja w_i a notação para a expressão acima.

A soma ponderada das projeções usa pesos x_i (i. é., as massas das linhas), pois a inércia das projeções é que será maximizada. Portanto a quantidade a ser maximizada é :

$$\sum_i x_i \, w_i^2$$

sujeito ao vetor **u** ser de comprimento unitário :

$$\sum_j \frac{1}{x_j} \quad u_j^2 = 1$$

Pode, então, ser verificado usando multiplicadores de Lagrange que o **u** ótimo é um vetor da matriz de dimensões n x n cujo termo j,j'-ésimo é:

$$\sum_i \frac{x_{ij}}{x_i} \quad \frac{x_{ij'}}{x_{i'}}$$

onde $1 \leq j,j' \leq n$ (note que esta matriz não é simétrica, e que uma matriz simétrica relacionada pode ser construída por "eigenredução"). O outro valor associado λ, indica a importância do eixo melhor ajustado .

Os resultados da análise de correspondência são centrados, ou seja, x_j e x_i são as j-ésimas e i-ésimas coordenadas (perfis médios) da origem da representação gráfica resultante. O primeiro autovalor resultante da análise de correspondência é trivial, de valor 1 e o autovetor associado é um vetor de 1s.

Vimos então que as projeções de pontos no eixo **u**, são em relação à métrica Euclidiana ponderada $\frac{1}{x_i}$. Isto torna a interpretação das projeções muito difícil do ponto de vista visual, portanto é mais natural apresentar resultados de tal maneira que as projeções possam ser vistas de forma mais simples.

Segundo Murtagh e Heck (1987), os fatores são definidos, tais que as projeções dos vetores-linha sobre o vetor ϕ associado com o eixo **u** são dados por :

$$\sum_j \frac{x_{ij}}{x_i} \phi_j$$

para todo i . Fazendo

$$\phi_j = \frac{1}{x_j} u_j$$

garantimos a proposição anterior e as projeções sobre ϕ são em relação à distância Euclidiana não ponderada.

Um conjunto análogo de relações vale em $\mathbb{R}^m$ onde é procurado o melhor eixo ajustado **v**. Uma simples relação matemática vale entre **u** e **v**, e entre ϕ e ψ (sendo ψ o fator

associado com o autovetor v). Podemos escrever as coordenadas dos pontos como :

$$\begin{cases} \sqrt{\lambda}\ \psi_i = \sum_j \dfrac{x_{ij}}{x_i}\ \phi_j \\ \quad (\text{ projeção do ponto-linha } i \text{ no eixo } u_\alpha\ ) \\ \sqrt{\lambda}\ \phi_j = \sum_i \dfrac{x_{ij}}{x_j}\ \psi_i \\ \quad (\text{ projeção do ponto-coluna } j \text{ no eixo } v_\alpha\ ) \end{cases}$$

2.4. INTERPRETAÇÃO DA ANÁLISE DE CORRESPONDÊNCIA

A interpretação dos gráficos representando as nuvens de pontos nos planos de projeção formados pelos primeiros eixos fatoriais dois a dois é o principal e mais difícil objetivo da análise de correspondência. Temos que verificar quais são os fatores da análise mais representativos e interpretar as proximidades entre os elementos de uma mesma nuvem de pontos.

A seguir, são definidos três coeficientes que auxiliam na interpretação dos resultados : taxa de inércia, contribuição absoluta e contribuição relativa.

Taxa de Inércia

Taxa de inércia ou porcentagem da variância de um eixo fatorial α é definida pelo quociente entre a α-ésima raiz característica da matriz **XX'** e a soma total das raízes características, ou seja:

$$I_\alpha = \frac{\lambda_\alpha}{\sum_\alpha \lambda_\alpha}$$

$100\ I_\alpha$ indica a porcentagem de variação explicada pelo eixo α.

Contribuição Absoluta

A contribuição absoluta exprime a parte da inércia explicada por uma dada variável em cada fator, ou seja, permite saber quais são as variáveis realmente responsáveis na construção do fator em questão.

A contribuição absoluta de um ponto j de $\mathbb{R}^n$ ao eixo α é defido por:

$$ca_\alpha(j) = x_j \phi^2_{\alpha j}$$

e $$\sum_{i=1}^{m} ca_\alpha(j) = 1$$

UFRGS
SISTEMAS DE BIBLIOTECAS
MATEMÁTICA

Analogamente, a contribuição absoluta, para um ponto i de $\mathbb{R}^m$ ao eixo α, é definida por:

$$ca_{\alpha}(i) = x_i \ \psi_{\alpha i}^2$$

e $$\sum_{j=1}^{n} ca_{\alpha}(i) = 1$$

Contribuição Relativa

A contribuição relativa, ou correlação entre variáveis e fator, exprime a parte tomada por um fator na explicação da dispersão de uma variável. Exibe quais são as características exclusivas do fator.

Os eixos fatoriais formam em cada espaço uma base ortogonal. O quadrado da distância de uma variável j ao centro de gravidade G se decompõe na soma de quadrados da projeção do ponto sobre cada um dos eixos.

Em $\mathbb{R}^n$, o quadrado da projeção do ponto j sobre o eixo α é:

$$d_{\alpha}^2(j,a) = \left(\sqrt{\lambda_{\alpha}} \ \phi_{\alpha j} \right)^2$$

e o quadrado das distâncias da variável j ao centro de gravidade é:

$$\sum_{\alpha} d_{\alpha}^2(j,a) = \sum_{i} \frac{1}{x_i} \left(\frac{x_{ij}}{x_j} - x_i \right)^2$$

Então :

$$\sum_{\alpha} d_{\alpha}^{2}(j,a) = d^{2}(j,a)$$

A contribuição relativa do fator α na explicacão da variância da variável j é definida por :

$$cr_{\alpha}(j) = \frac{d_{\alpha}^{2}(j,a)}{d^{2}(j,a)}$$

com

$$\sum_{\alpha} cr_{\alpha}(j) = 1$$

Analogamente, em $\mathbb{R}^{m}$ a contribuição relativa do fator α na explicação da variância da variável i é:

$$cr_{\alpha}(i) = \frac{d_{\alpha}^{2}(i,a)}{d^{2}(i,a)}$$

onde :

$$d_{\alpha}^{2}(i,a) = \left(\sqrt{\lambda_{\alpha}}\ \psi_{\alpha i}\right)^{2} \quad e$$

$$d^{2}(i,a) = \sum_{j} \frac{1}{x_{j}} \left(\frac{x_{ij}}{x_{i}} - x_{j} \right)^{2}$$

UFRGS
SISTEMAS DE BIBLIOTECAS
BIBLIOTECA SETORIAL DE MATEMÁTICA

3. Análise de Correspondência Múltipla

Consideremos uma série de variáveis cada uma com um certo número de respostas possíveis (modalidades respostas) respondidas por um conjunto de indivíduos. É muito comum que as categorias ou modalidades de resposta de um atributo apareçam na forma de zeros ou uns, isto é, para cada variável o indivíduo escolhe uma modalidade de resposta e esta recebe o valor 1 (um) se o sujeito escolher esta modalidade, e as demais modalidades da variável recebe o valor 0 (zero). A este tipo de codificação chamamos codificação "disjuntiva completa". As modalidades de respostas se excluem mutuamente e somente uma das modalidades é escolhida (Tabela 3.1).

Quando a matriz de dados está na forma disjuntiva completa, a análise de correspondência é denominada Análise de Correspondência Múltipla.

Geralmente a análise de correspondência é usada para análise de tabelas de contingência. Tais tabelas podem ser derivadas de uma tabela de forma disjuntiva completa tomando a matriz produto entre sua transposta e ela mesma. A tabela simétrica, obtida desta forma é chamada tabela de Burt (Tabela 3.2.). A análise de correspondência de ambas tabelas dão resultados similares, diferindo apenas os autovalores.

A seguir algumas características da análise de tabelas na forma disjuntiva completa :

. As modalidades (ou categorias de resposta) de cada atributo na análise de correspondência múltipla, tem o seu centro de gravidade na origem.

. O número de autovalores diferentes de zero encontrados é menor ou igual que o número total de modalidades menos o número total de atributos.

. Devido à grande dimensão do espaço a ser analisado, os autovalores tendem a ser muito pequenos na análise de correspéndencia múltipla. É comum achar que os primeiros poucos fatores podem ser adequadamente interpretados e também explicar apenas uma pequena porcentagem da inércia total.

A interpretação dos resultados da análise de correspondência múltipla é similar à interpretação da análise de correspondência.

Os eixos são interpretados em ordem de importância decrescente usando as modalidades as quais mais contribuem em termos de inércia, aos eixos (i. é., massa vezes distância quadrada projetada). As coordenadas projetadas servem para indicar a distância da modalidade em relação aos eixos.

As representações gráficas (projeções de pontos-linha e pontos-coluna no plano formado pelos fatores 1 e 2 e por outros pares de fatores) são examinadas.

Tipo			Idade			Propriedade				
T_1	T_2	T_3	A_1	A_2	A_3	P_1	P_2	P_3	P_4	P_5
1	0	0	0	1	0	0	0	0	0	1
0	1	0	0	0	1	0	0	0	0	1
1	0	0	0	0	1	0	0	1	0	0
1	0	0	0	1	0	0	0	1	0	0
1	0	0	1	0	0	1	0	0	0	0
0	0	1	0	1	0	1	0	0	0	0
0	0	1	1	0	0	0	0	0	0	1

Tabela 3.1. - Tabela da forma disjuntiva completa.

	T_1	T_2	T_3	A_1	A_2	A_3	P_1	P_2	P_3	P_4	P_5
T_1	4	0	0	1	2	1	1	0	2	0	1
T_2	0	1	0	0	0	1	0	0	0	0	1
T_3	0	0	2	1	1	0	1	0	0	0	1
A_1	1	0	1	2	0	0	1	0	0	0	1
A_2	2	0	1	0	3	0	1	0	1	0	1
A_3	1	1	0	0	0	2	0	0	1	0	1
P_1	1	0	1	1	1	0	2	0	0	0	0
P_2	0	0	0	0	0	0	0	0	0	0	0
P_3	2	0	0	0	1	1	0	0	2	0	0
P_4	0	0	0	0	0	0	0	0	0	0	0
P_5	1	1	1	1	1	1	0	0	0	0	3

Tabela 3.2. - Tabela de Burt associada à tabela disjuntiva completa (Tabela 2.3.)

Notas:

. Atributos : Tipo, Idade, Propriedades

. Modalidades (categorias) : T_1, T_2, ..., P_5.

. A soma das linhas da tabela na forma disjuntiva completa são constantes e igual ao número de atributos.

. Cada submatriz atributo x atributo da tabela de Burt é necessariamente diagonal, com total de colunas da tabela na forma disjuntiva completa somando valores diagonais.

4. APRESENTAÇÃO DO SOFTWARE SPHINX E EXEMPLO

4.1. Introdução

Nesta seção é apresentado o programa utilizado para fazer os cálculos de análise de correspondência da presente monografia.

O programa SPHINX foi desenvolvido por dois franceses, J. Moscarola e J. de Lagarde e existem várias opções de entrada de dados, inclusive via questionário. Possui também várias opções para o tratamento dos dados. Nesta monografia é apresentada somente as etapas referentes à análise de correspondência.

A fim de facilitar a apresentação do SPHINX, foi utilizado um exemplo com a finalidade de verificar a associação entre marcas de carros e critérios para escolha da marca ,com os dados apresentados em uma tabela de contingência (digitados em ASCII), onde as linhas (variável independente) são as marcas dos carros e as colunas (variável dependente) os critérios. Os dados são apresentados na tabela 4.1.

	Renault	Peugeot	Citroen	Talbot	Ford	Fiat	Volkswagen	Opel	Japonnaise	
VELOCIDADE	5	8	10	1	6	0	14	9	8	61
CONFORTO	0	9	4	0	1	1	13	5	11	44
SEGURANCA	3	4	0	1	10	0	5	5	1	29
CONSUMO	21	7	10	2	1	5	0	5	4	55
PRECO	5	3	13	3	1	5	0	4	3	37
PUBLICIDADE	5	1	0	2	1	1	6	5	3	24
DISTRIBUIDOR	0	0	1	2	0	0	1	5	0	9
ATENDIMENTO	5	0	5	0	5	0	0	0	1	16
	44	32	43	11	25	12	39	38	31	275

UFRGS Dprt Stat. PORTO ALEGRE

Tabela 4.1. - Marcas de Carros x Critérios

Uma condição do programa é que o número de modalidades linhas não ultrapasse a 70 e o número de modalidades colunas, não ultrapasse a 30.

A seguir serão abordados alguns aspectos de entrada e saída do programa SPHINX.

4.2. Entrada de Dados

A Tela 1 mostra todos os passos que devemos seguir para entrar com os dados.

Tela 1

```
LE SPHINX 1: Conception et traitements d'enquêtes et sondages
Version 6.0 * Copyright:J.de LAGARDE - J.MOSCAROLA 1985-87-89
                    Reproduction Interdite
Licence N°  190190   Bénéficiaire: UFRGS Dprt Stat. PORTO ALEGRE

                                    Données sur le disque :  c
Nom de l'enquête    :   exep

Année de l'enquête :   90
```

Vous commencez un nouveau travail .Si ce n'est pas le cas vous avez fait une erreur dans nom de l'enquête. Tapez C pour corriger ou O pour continuer

Pour voir les fichiers présents sur le disque: touche F1

Primeiramente coloca-se o drive a ser utilizado, o nome do arquivo de dados e o ano da pesquisa. Logo digita-se a letra O para continuar e na mesma tela aparecerá uma opção para colocar o título do trabalho (se não deseja colocar título digite ENTER). Ainda, na mesma tela, aparecerá duas opções de entrada:

1 CONCEPTION ET SAISIE 2 BOITE OUTILS

Temos que entrar na opção 2 para entrada de dados .

Tela 2

LA BOITE A OUTILS DU SPHINX
Copyright J.de Lagarde J.Moscarola Licence UFRGS Dprt Stat. PORTO ALEGRE

MENU 4

1 EXPORTER LES TRIS VERS TABLEURS
2 QUESTIONNAIRE VERS FICHIER TEXTE
3 RESULTATS TRIS VERS FICHIER TEXTE
4 MODIFIER STRUCTURES DONNEES SPHINX
5 IMPORTER DES DONNEES EXTERNES
6 AIDE A L'ECHANTILLONAGE
7 FIN

Nom de l'enquête (4 caractères maxi) EXEP
Année de l'enquête : 19 90

Na Tela 2 entraremos na opção 5 (importar dados externos) colocando novamente o nome do arquivo e o ano.

Tela 3

IMPORTATION DE TABLEAUX OU DONNES BRUTES

Tableau d'EFFECTIFS ou CONTINGENCE ==> A.F.C

Lignes et Colonnes: Modalités qualitatives ou Caractères
Cellules: Nombre d'individus ayant les caractères en ligne et colonne

Tableau de VALEURS MOYENNES ==> A.C.P.

Lignes: Modalités ou Caractères. Colonnes: Critères quantitatifs
Cellules: Moyenne de la Modalité ligne, évaluée par rapport au critère colonne

Tableau INDIVIDUS-CRITERES ==> SPHINX A.C.P. REGRESSION

En Lignes des observations individuelles, en Colonne des Critères
Cellule: Valeur prise par l'individu en ligne pour le critère en colonne
Tableau équivalent aux réponses à un questionnaire purement quantitatif:
Ne comportant que des questions de type 4

Définissez la nature des données à importer

1Importation FICHIERS 2Saisie CLAVIER 3Retour
1Tableaux EFFECTIFS 2Tableau VALEURS 3Observations INDIVIDUS CRITERES 4Retour

Passando para a Tela 3 digita-se o número 1 para a opção Tabela de Contingência e digita-se novamente o número 1 para a opção Importação de Fichários.

Tela 4

IMPORTATION D'UN TABLEAU D'EFFECTIFS OU DE CONTINGENCE ==> A.F.C.

MULTIPLAN Format SYLK (Option de Lit_Ecrit) ou ASCII (Imprime Fichier)

LOTUS 123 feuille sauvée en format texte ASCCI par Imprime Fichier

ASCII Toujours laisser un blanc pour séparer deux colonnes du tableau

Un seul tableau par fichier. 70 Lignes 30 Colonnes au MAXIMUM

Les tableaux doivent commencer au coin supérieur gauche du tableur ou du champ par une ligne de titre pour les colonnes ou directement par les chiffres
Les totaux marginaux ne doivent pas figurer dans le tableau

1Fichier SYLK 2Fichier ASCII 3Retour

Aparecerá a Tela 4 onde escolheremos a opção 2 FICHARIO ASCII (pois os dados estão digitados em ASCII). Logo digitaremos o drive que está sendo utilizado e o nome do arquivo e então colocaremos o nome de cada linha e de cada coluna da tabela.

Tela 5

	VELO	CONF	SEGU	CONS	PREC	PUBL	DIST	ATEN
Rena	5	0	3	21	5	5	0	5
Peug	8	9	4	7	3	1	0	0
Citr	10	4	0	10	13	0	1	5
Talb	1	0	1	2	3	2	2	0
Ford	6	1	10	1	1	1	0	5
Fiat	0	1	0	5	5	1	0	0
Volk	14	13	5	0	0	6	1	0
Opel	9	5	5	5	4	5	5	0
Japo	8	11	1	4	3	3	0	1

Titre du tableau MARCAS X CRITERIOS
1Suite2Enregistrement3Impression4Enregistrement et Impression5Retour

A próxima tela a aparecer é a Tela 5, apresentando a tabela de contingência concluindo-se a etapa de entrada de dados. Na tela 5 escolheremos a opção 1 SUITE para passar à demonstração dos resultados.

UFRGS
SISTEMAS DE BIBLIOTECAS
BIBLIOTECA SETORIAL DE MATEMÁTICA

Tela 7

AXES FACTORIELS ET CONTRIBUTIONS

67 % de la variance totale est expliquée par les axes. Dont:

AXE 1 : 42 %	AXE 2 : 25 %
Volkswagen 34 %+	Ford 76 %-
Renault 27 %-	Japonnaise 7 %+
Citroen 14 %-	Fiat 6 %+
Fiat 13 %-	Renault 3 %-
Japonnaise 5 %+	Talbot 2 %+
Opel 3 %+	Volkswagen 2 %+
Peugeot 2 %+	Peugeot 1 %+
CONSUMO 34 %-	SEGURANCA 45 %-
CONFORTO 24 %+	ATENDIMENTO 29 %-
PRECO 19 %-	CONFORTO 15 %+
VELOCIDADE 9 %+	PRECO 7 %+
SEGURANCA 7 %+	DISTRIBUIDOR 3 %+
ATENDIMENTO 5 %-	CONSUMO 1 %+

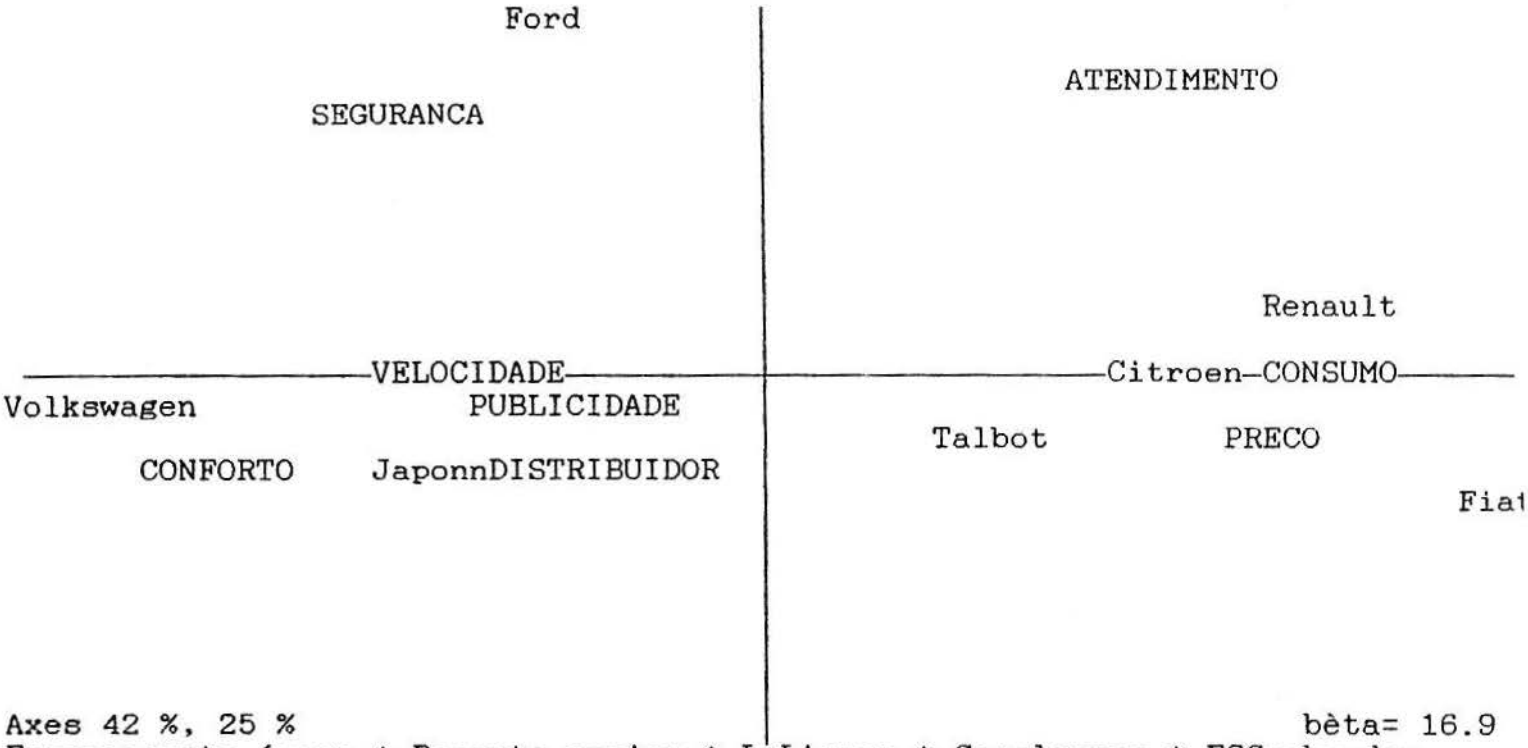
Dépendance significative. Béta= 16.9
1Représentation graphique 2Edition contributions 3Retour

Verifica-se então, na Tela 7, os primeiros resultados da análise de correspondência, tais como: a taxa de inércia e a contribuição absoluta (para melhor visualização dos coeficientes aconselha-se a impressão destes resultados, opção 2). No programa SPHINX, onde está escrito contribuição relativa, leia-se contribuição absoluta (o programa não fornece

as contribuições relativas). A última coluna da contribuição absoluta é a média ponderada das contribuições absolutas dos eixos .

Verifica-se também, na Tela 7, um coeficiente de associação denominado Beta, que determina se é significativo ou não a dependência entre as linhas e as colunas da tabela de contingência. O valor de Beta, para ter significado o resultado, deve ser maior ou igual a três. No exemplo Beta = 16.9, portanto a dependência entre as linhas e as colunas da tabela de contingência é significativa. Passaremos para a representação gráfica das linhas e das colunas da tabela de contingência (opção 1 REPRESENTATION GRAPHIQUE, na Tela 7).

Tela 8



Na Tela 8, temos o gráfico e várias outras opções. Para imprimir digite P e para abandonar a seção digite ESC. Retornaremos então a Tela 2 e digitaremos o número 3 para abandonar o programa SPHINX.

4.4. Interpretação dos Resultados

A interpretação do gráfico representando as nuvens de pontos nos planos de projeção formados pelos primeiros eixos fatoriais dois a dois é o principal e mais difícil objetivo da análise de correspondência, por isso, contamos com a ajuda de alguns coeficientes: taxa de inércia, contribuição relativa e contribuição absoluta.

No exemplo, do Automóvel temos que 67% da variância total das variáveis é explicada pelos dois primeiros eixos, sendo que o primeiro eixo extraído pela análise, explica 42% e o segundo eixo explica 25% da variância total das variáveis. Então, vamos considerar, somente estes dois primeiros eixos (fatores).

A tabela 4.2 mostra as contribuições absolutas dos pontos-linha e pontos-coluna para estes fatores.

Contributions relatives)

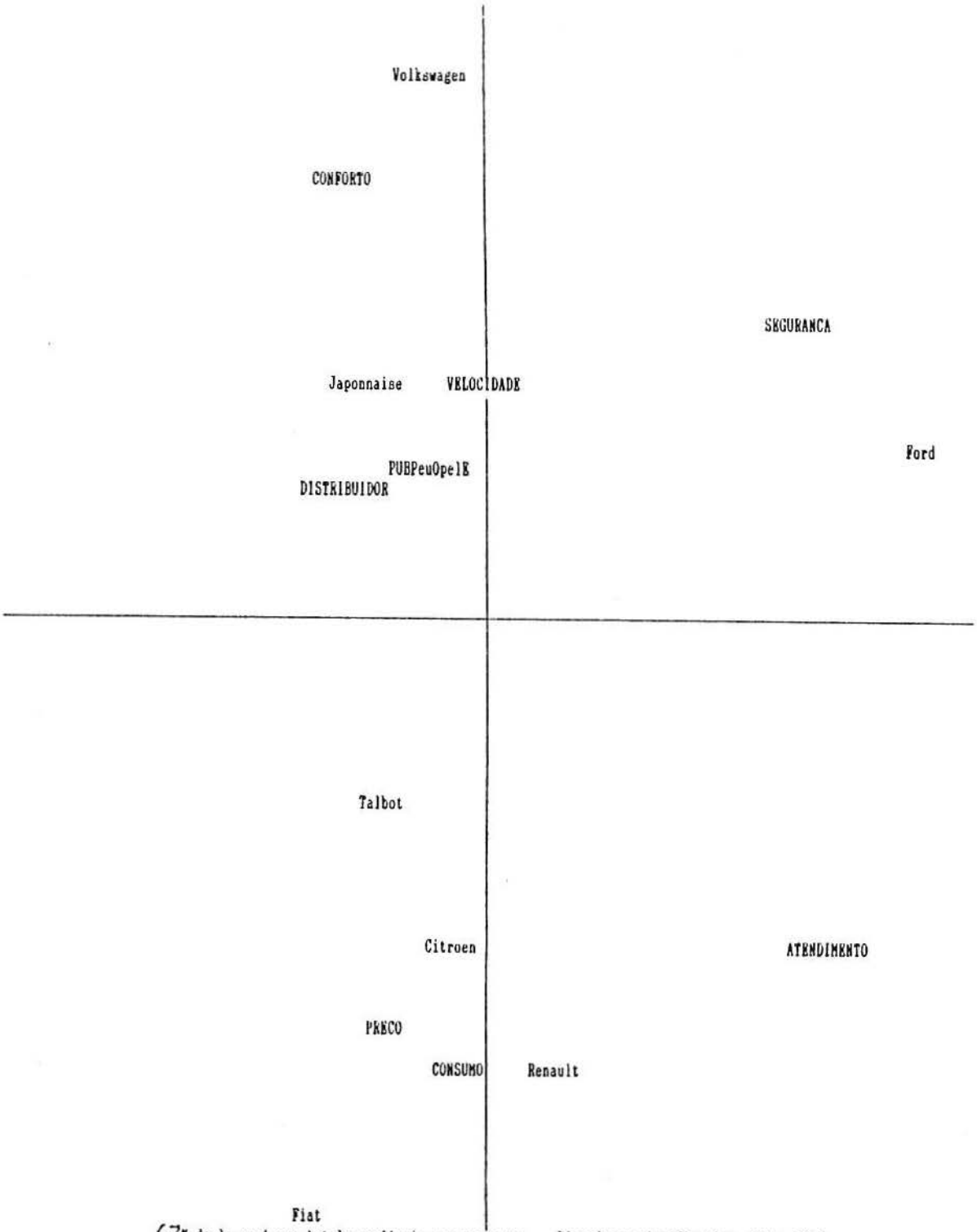
	AxeN° 1 (42%)	AxeN° 2 (25%)	CONTRIB ABSOLUES
Renault	27%-	3%-	12%
Peugeot	2%+	1%+	1%
Citroen	14%-	1%+	6%
Talbot	1%-	2%+	0%
Ford	2%+	76%-	20%
Fiat	13%-	6%+	6%
Volkswagen	34%+	2%+	14%
Opel	3%+	1%+	1%
Japonnaise	5%+	7%+	4%
VELOCIDADE	9%+	0%+	4%
CONFORTO	24%+	15%+	14%
SEGURANCA	7%+	45%-	14%
CONSUMO	34%-	1%+	14%
PRECO	19%-	7%+	10%
PUBLICIDADE	1%+	1%+	1%
DISTRIBUIDOR	0%+	3%+	1%
ATENDIMENTO	5%-	29%-	9%

Pourcentages arrondis à l'entier
UFRGS Dprt Stat. PORTO ALEGRE

Tabela 4.2. - Contribuições Absolutas dos Eixos

O primeiro fator é muito mais importante que o segundo para explicação da dispersão das variáveis. Somente para a marca Ford e os critérios segurança e atendimento é que o fator 2 se destaca mais que o fator 1.

MARCAS X CRITERIOS



67% de la variance totale expliquée par ces axes Dépendance significative. Béta= 16.9
Verticalement: Axe N° 1 (42 %)
Horizontalement: Axe N° 2 (25 %)
UFRGS Dept Stat. PORTO ALEGRE

Observando, no gráfico, as proximidades dos pontos que representam as variáveis, percebe-se uma associação entre marcas e critérios. A marca Japonnaise, Peugeot e Opel estão associadas com velocidade, publicidade e distribuidor. Volkswagen está associada com conforto, Ford com segurança ,Citroen, Renaut, Fiat e Talbot estão associadas com preço e consumo.

5. APLICAÇÕES NUMÉRICAS

Os dados numéricos, da aplicação 1, aos quais utilizamos a técnica de análise de correspondência, foram cedidos pela professora Sídia M. C. Jacques do Departamento de Estatística da UFRGS e os da aplicação 2 foram cedidos por Nilton C. M. de Araújo, os quais foram utilizados em sua Tese de Mestrado em Economia, da Faculdade de Ciências Econômicas, Centro de Estudo e Pesquisas Econômicas- IEPE, UFRGS (1990), sob o título: "Fatores Locacionais da Agroindústria Alimentar do Rio Grande do Sul.

5.1. Aplicação 1

Na genética, o GM é o sistema genético relacionado com as imunoglobulinas existindo vários aplótipos no sistema que são herdados como alelos (AG, AXG, ABST e OUTROS). Estes alelos possuem pequenas diferenças nas várias populações.

O objetivo deste trabalho é verificar, nas tribos indígenas norte-americanas, a existência de associação entre grupos linguísticos e alelos. Os grupos linguísticos pesquisados

UFRGS
SISTEMAS DE BIBLIOTECAS
BIBLIOTECA SETORIAL DE MATEMÁTICA

foram: Atabascos, Yuit, Inuit, Almosan-Keresiouan, Penutian, Uto-Azteca e Chibcha.

SID(Contributions relatives)

	AxeN° 1 (57%)	AxeN° 2 (27%)	CONTRIB ABSOLUES
Alm-Ker	1%+	5%+	1%
Penutian	3%-	6%-	4%
Uto-Azt	0%-	44%+	12%
Chibcha	48%-	18%-	33%
Yuit	14%+	13%-	11%
Inuit	23%+	5%-	14%
Athab	10%+	9%-	8%
AG	1%+	17%+	4%
AXG	55%-	32%-	40%
ABST	45%+	47%-	38%
OUTROS	0%-	4%-	1%

Pourcentages arrondis à l'entier
UFRGS Dprt Stat. PORTO ALEGRE

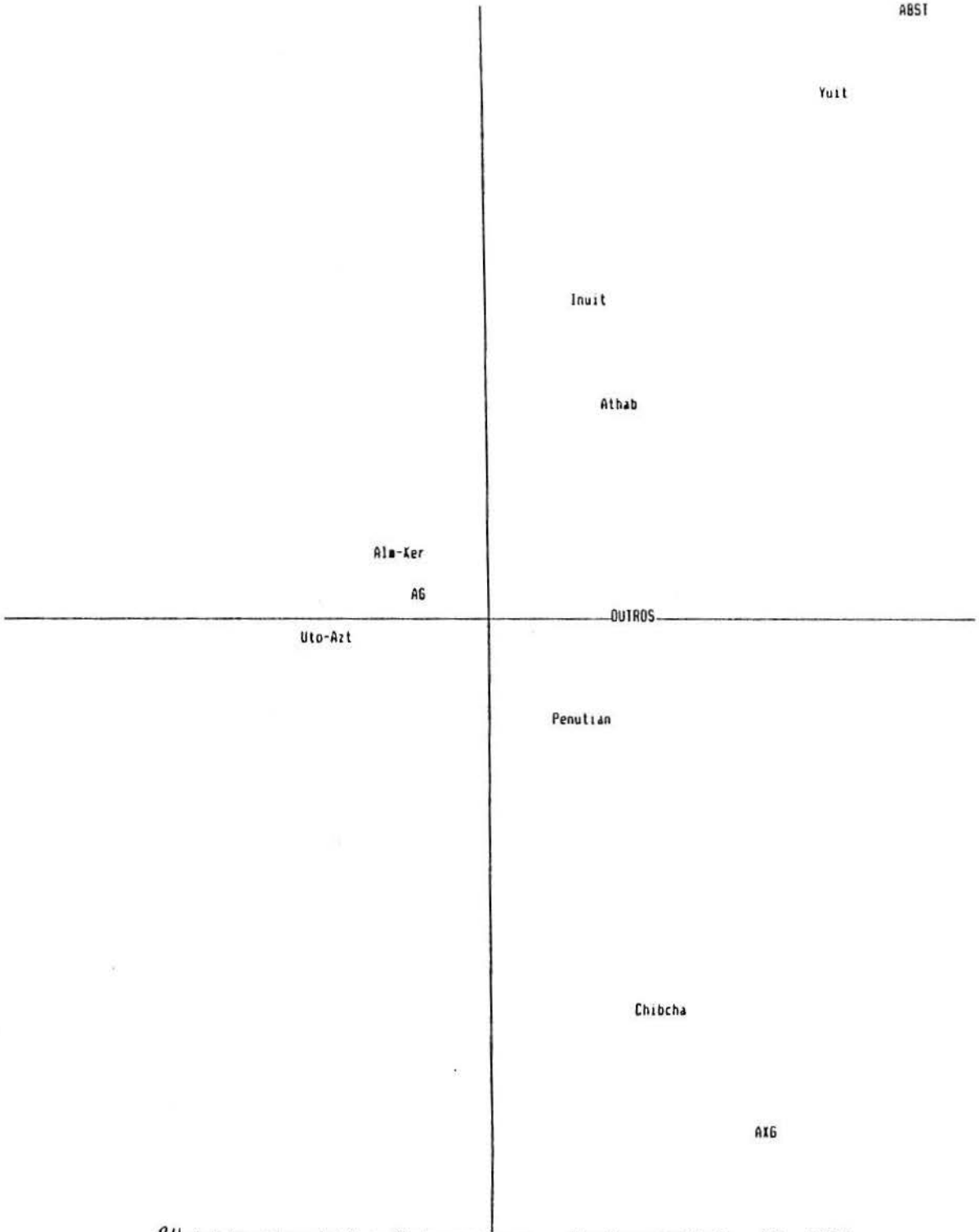
Tabela 5.1. - Contribuições Absolutas dos Eixos

A tabela 5.1. mostra que o primeiro fator, extraído pela análise, explica 57% da variância total das variáveis e o segundo fator explica 27% desta mesma variâcia. Assim, os dois primeiros fatores explicam 84% da variância total.

Analisando as contribuições absolutas, tabela 5.1., verifica-se que as variáveis que mais contribuem na construção do

fator 1 são Chibcha, Inuit, AXG e ABST. Em relação ao fator 2, as variáveis que mais se destacaram são Uto-Azt, AXG e ABST.

TRONCO LING. x ALELOS



84 % de la variance totale expliquée par ces axes Dépendance significative. Bêta= 1103.5
Verticalement: Axe N° 1 (57 %)
Horizontalement: Axe N° 2 (27 %)
UFRGS Dprt Stat. PORTO ALEGRE

Observando o gráfico, verifica-se algumas associações entre grupos linguísticos e alelos. Nota-se claramente a associação entre o alelo ABST e o grupo linguístico Yuit, também associação entre o alelo AG e os grupos linguísticos Alm-Ker e Uto-Azt e entre o alelo AXG e o grupo linguístico Chibcha.

5.2. Aplicação 2

Os dados desta pesquisa, tratam das localizações atuais das empresas e de situações envolvendo relocalizações ou abertura de filiais.

As 204 empresas pesquisadas, foram divididas em 4 classes de tamanho, pelo critério de número de empregados: "Pequenas" (PQ - 10 a 44 empregados), "Médias-Pequenas" (MP - 45 a 155 empregados), "Médias-Grandes" (MG - 156 a 484 empregados) e "Grandes" (GR - 485 ou mais empregados).

As outras variáveis utilizadas, com uma classificação de 0 (nenhuma importância) a 3 (muito importante), foram:

x04 - Existência de instituições de treinamento de mão-de-obra

x05 - Proximidade dos principais compradores

x06 - Proximidade de firmas industriais fornecedoras

x07 - Proximidade de zonas produtoras agrícolas

x08 - Proximidade de firmas não industriais fornecedoras

x09 - Proximidade de porto fluvial ou marítmo

x10 - Acesso à administração central ou a outras fábricas da empresa

x12 - Fácil acesso ferroviário

x14 - Aproveitamento de incentivos governamentais

x16 - Fácil acesso a serviços urbanos especializados

x17 - Fácil acesso a serviços de manutenção e assistência técnica

x19 - Proximidade de locais com grande disponibilidade de água

x21 - Energia para uso próprio ou outras fontes energéticas

x22 - Possuía prédio próprio e/ou terreno no local

x23 - Compra de terreno barato e espaçoso

x24 - Próximo da residência do empresário e/ou vínculo familiar

x25 - Região com tradição na produção do setor.

O objetivo deste trabalho é verificar a existência de associação entre o tamanho da empresa e fatores locacionais.

Observamos, na tabela 5.2., que a maioria dos pontos estão bem representados no plano formado pelos dois primeiros eixos (fatores). O primeiro eixo, explica 56% da variância total das variáveis e o segundo eixo explica 26% desta mesma variância. Assim os dois primeiros eixos explicam 82% da variância total.

O primeiro eixo representa o porte da empresa, ou seja, separa as empresas Pequenas e Médias-Pequenas de Médias-Grandes e Grandes.

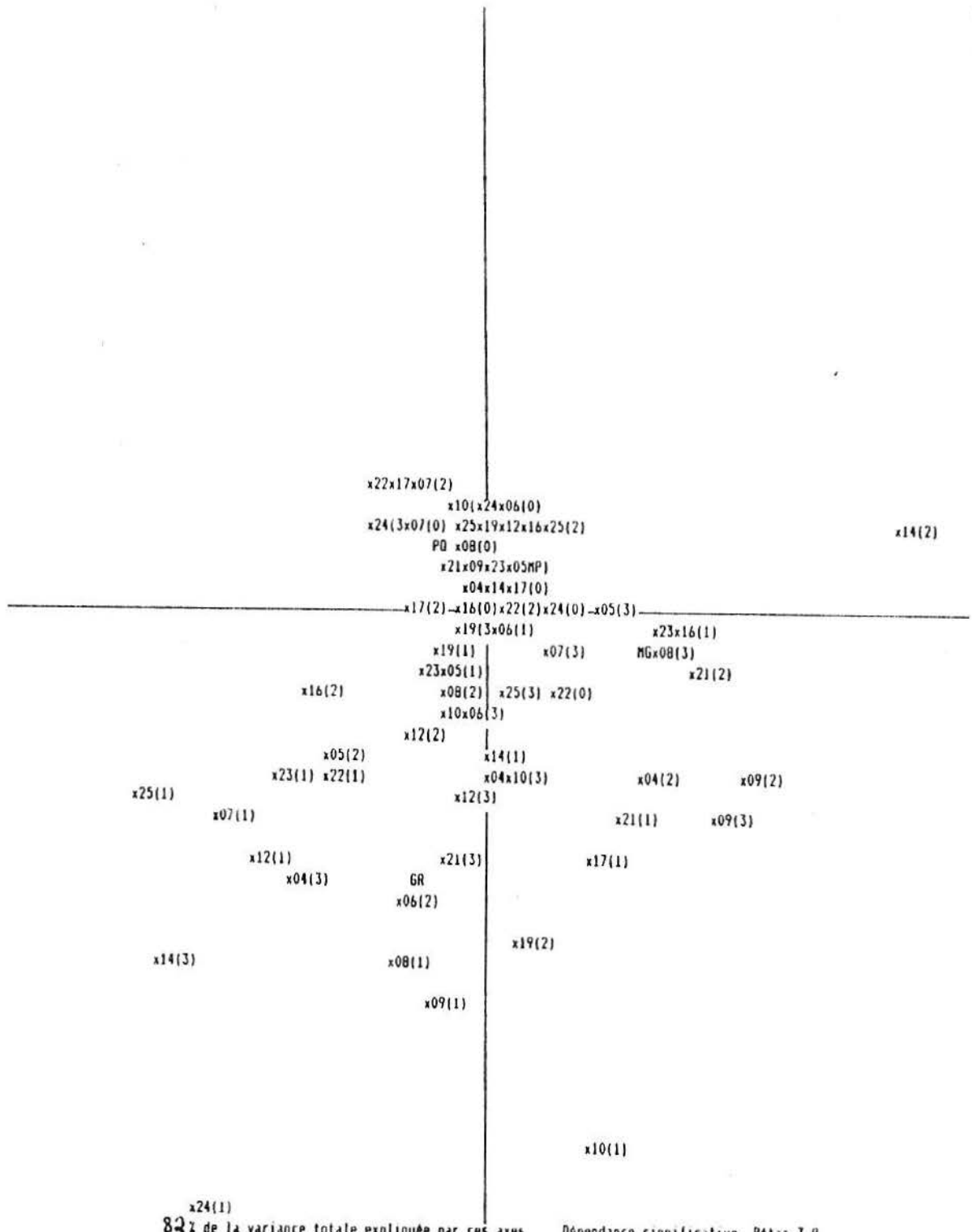
MILT(Contributions relatives)

	AxeN° 1 (56%)	AxeN° 2 (26%)	CONTRIB ABSOLUES
x04(0)	0%+	0%+	0%
x04(1)	1%-	0%-	1%
x04(2)	0%-	1%-	0%
x04(3)	1%-	1%+	1%
x05(0)	1%+	0%-	0%
x05(1)	0%-	0%+	0%
x05(2)	2%-	3%+	2%
x05(3)	0%+	3%-	1%
x06(0)	3%+	0%-	2%
x06(1)	0%-	0%-	0%
x06(2)	7%-	1%+	4%
x06(3)	1%-	0%+	0%
x07(0)	1%+	2%+	1%
x07(1)	1%-	5%+	2%
x07(2)	1%+	0%+	0%
x07(3)	1%-	4%-	1%
x08(0)	1%+	0%+	1%
x08(1)	6%-	1%+	3%
x08(2)	0%-	0%+	0%
x08(3)	0%-	3%-	1%
x09(0)	1%+	0%+	0%
x09(1)	4%-	0%+	2%
x09(2)	0%-	3%-	1%
x09(3)	1%-	4%-	2%
x10(0)	3%+	0%+	2%
x10(1)	11%-	1%-	6%
x10(2)	0%-	0%+	0%
x10(3)	2%-	0%-	1%
x12(0)	2%+	0%-	1%
x12(1)	2%-	3%+	2%
x12(2)	1%-	0%+	0%
x12(3)	3%-	0%+	2%
x14(0)	0%+	0%-	0%
x14(1)	0%-	0%-	0%
x14(2)	0%+	3%-	1%
x14(3)	3%-	5%+	3%
x16(0)	0%-	0%+	0%
x16(1)	0%-	5%-	1%
x16(2)	0%-	4%+	1%
x16(3)	1%+	0%-	0%
x17(0)	0%+	0%-	0%
x17(1)	3%-	2%-	2%
x17(2)	0%+	1%+	0%
x17(3)	1%+	1%+	1%
x19(0)	2%+	0%+	1%
x19(1)	0%-	0%+	0%
x19(2)	7%-	0%-	4%
x19(3)	0%-	0%+	0%
x21(0)	1%+	0%+	0%
x21(1)	2%-	2%-	1%
x21(2)	0%-	2%-	1%
x21(3)	2%-	0%+	1%
x22(0)	1%-	4%-	2%
x22(1)	1%-	1%+	0%
x22(2)	0%+	0%-	0%
x22(3)	3%+	4%+	3%
x23(0)	0%+	0%+	0%
x23(1)	1%-	2%+	1%
x23(2)	0%-	5%-	1%
x23(3)	0%-	1%+	0%
x24(0)	0%-	4%-	1%
x24(1)	8%-	4%+	6%
x24(2)	0%+	0%-	0%
x24(3)	1%+	4%+	2%
x25(0)	1%+	0%+	1%
x25(1)	1%-	6%+	2%
x25(2)	0%+	0%-	0%
x25(3)	1%-	1%-	1%
PQ	13%+	16%+	12%
MP	7%+	1%-	4%
MG	3%-	73%-	21%
GR	77%-	10%+	46%

Pourcentages arrondis à l'entier

UFRGS Dprt Stat. PORTO ALEGRE

GRAFICO 1



82 % de la variance totale expliquée par ces axes Dépendance significative. Béta= 3.9

Verticalement: Axe N° 1 (56 %)

Horizontalement: Axe N° 2 (26 %)

UFRGS Dprt Stat. PORTO ALEGRE

UFRGS
SISTEMAS DE BIBLIOTECAS
BIBLIOTECA SETORIAL DE MATEMÁTICA

Observando no gráfico, a proximidade dos pontos que representam as variáveis, chegamos à conclusão de que existe uma associação entre o tamanho da empresa e os fatores locacionais, no sentido de que existe uma maior valorização das empresas de porte maior.

Nota-se que os principais fatores locacionais, associados às empresas de grande porte são: x04- Existência de instituições de treinamento de mão-de-obra, x06- Proximidade de firmas industriais fornecedoras, x10- Acesso à administração central ou a outras fábricas da empresa, x12- Fácil acesso ferroviário, x14- Aproveitamento de incentivos governamentais, x19- Proximidade de locais com grande disponibilidade de água e x21- Energia para uso próprio ou outras fontes energéticas. Os principais fatores locacionais associados à empresas Médias-Grandes são: X06- Proximidade de firmas industriais fornecedoras, x08- Proximidade de firmas não industriais fornecedoras, x09- Proximidade de porto fluvial ou marítmo, x10- Acesso à administração central ou a outras fábricas da empresa, x19- Proximidade de locais com grande disponibilidade de água e x25- Região com tradição na produção do setor. As empresas Pequenas e Médias-Pequenas estão associadas com os fatores locacionais x22- Possuía prédio próprio e/ou terreno no local e x24- Próximo da residência do empresário e/ou vínculo familiar.

6. REFERÊNCIAS BIBLIOGRÁFICAS

BENZÉCRI, J. P. (1973). **L'analyse des données: T.2, L'analyse des correspondances**. Paris: Dunod.

BENZÉCRI, J. P. (1977a). Histoire et préhistoire de l'analyse des données: l'analyse des correspondances . **Les Cahiers de L'analyse des Données**, 2, 9-53.

BENZÉCRI, J. P. (1977b). Sur l'analyse des tableaux binaires associés à une correspondance multiple . **Les Cahiers de L'analyse des Données**, 2, 55-71.

CANTON, A. W. P. (1982). Análise de Correspondência. **Atas do 1º Encontro de Doscentes de Estatística da Região Sul**, Porto Alegre, UFRGS, 40-50.

FERNANDEZ, P., YOHAI, V. e KLEIN, R. (1980). **Analisis de datos multivariados**. Rio de Janeiro, IMPA (monografia).

GIFI, A. (1981). **Nonlinear multivariate analysis**. Leiden, The Netherlands: University of Leiden, Afdeling Datatheorie.

GREENACRE, M. J. (1984). **Theory and applications of correspondence analysis**. London: Academic Press.

HILL, M. O. (1973). Correspondence analysis: a neglected multivariate method. **Appl Stat 23**, 340-354.

LEBART, L., MORINEAU, A. & TABARD, N. (1977). **Techniques de la description statistique**. Paris: Dunod.

LEBART, L., MORINEAU, A., & WARWICK, K. M. (1984). **Multivariate descriptive analysis: correspondence analysis and related techniques for large matrics**. New York: Wiley-Intercience.

MOSER, E. B. (1989). Exploring contingency tables with correspondence analysis. **CAMBIOS**, Vol. 5., nº 3, 183-189.

MURTAGH, F., HECK, A. (1987). **Multivariate data analysis**. Dordrecht, The Netherlands: D.Reidel Publishing Company.

NISHIATO, S. (1980). **Analysis of categorical data: dual scaling and its applications**. Toronto: University of Toronto Press.

SOUZA, A. M. R. (1982). **Análise de correspondência** . Tese de Mestrado. IME-USP.

SCHIEVER, B. F. (1983). Scaling of order dependent categorical variables with correspondence analysis. **International Statistical Rewiew, 51**, 225-238.

UFRGS
SISTEMAS DE BIBLIOTECAS
BIBLIOTECA SETORIAL DE MATEMÁTICA

TEMENHAUS, M. . YONG, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for qualifying categorical multivariate data. **Psychometrika**, vol. 50, nº 1, 91-119.