UNIVERSIDADE FEDERAL DE JUIZ DE FORA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Bruno Caetano Vidigal

ANÁLISE DE CORRESPONDÊNCIA PARA AVALIAÇÃO DE INDICADORES SOCIOECONÔMICOS, DEMOGRÁFICOS E DE MORTALIDADE

JUIZ DE FORA
2010

Bruno Caetano Vidigal

ANÁLISE DE CORRESPONDÊNCIA PARA AVALIAÇÃO DE INDICADORES SOCIOECONÔMICOS, DEMOGRÁFICOS E DE MORTALIDADE

Monografia apresentada ao curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para a obtenção do grau de Bacharel em Estatística.

Orientador: Ronaldo Rocha Bastos

PhD em Urban and Regional Planning-Liverpool University

JUIZ DE FORA
2010

Bruno Caetano Vidigal

ANÁLISE DE CORRESPONDÊNCIA PARA AVALIAÇÃO DE INDICADORES SOCIOECONÔMICOS, DEMOGRÁFICOS E DE MORTALIDADE

Monografia apresentada ao curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para a obtenção do grau de Bacharel em Estatística.

Aprovada em 06 de dezembro de 2010

BANCA EXAMINADORA

_____

Ronaldo Rocha Bastos (orientador)

PhD em Urban and Regional Planning - Liverpool University

_____

Clécio da Silva Ferreira

Doutor em Estatística pela Universidade de São Paulo

_____

Marcel de Toledo Vieira

Ph.D. em Estatística pela University of Southampton

Vidigal, Bruno – Juiz de Fora, 2010
Análise de Correspondência para avaliação de indicadores socioeconômicos, demográficos e de mortalidade / Bruno Vidigal

58.p

Monografia – Universidade Federal de Juiz de Fora e Instituto de Ciências Exatas

Orientador: Ronaldo Rocha Bastos

À minha mãe

A vida é feito andar de bicicleta, se parar você cai!

Gabriel, o pensador

AGRADECIMENTOS

Agradeço primeiramente à DEUS por ter me guiado nesses anos.

À minha mãe, pela minha formação como pessoa e por todo incentivo, dedicação, empenho, carinho e amor ao longo de toda minha vida.

Aos queridos tios, Fernando e Paloma, que me deram literalmente casa, comida e roupa lavada nesses últimos seis anos. Agradeço eternamente a vocês pela oportunidade de estudo e por toda confiança, atenção e carinho. Muito obrigado.

Ao professor Ronaldo, por ter acreditado em meu trabalho e por todo conhecimento passado, além da paciência e sempre boa vontade em ajudar.

Ao professor Marcel, por toda educação e clareza nas aulas e ao professor Clécio por sempre estar disposto em ajudar e é claro, pelas caronas.

Agradeço também aos professores Joaquim e Camila, e principalmente ao André Hallack, professor do Departamento de Matemática, pelo show nas aulas de Cálculo e a todos os outros professores que contribuíram de alguma forma para minha formação.

Aos meus avós Chico, Luzia e Carmelita, ao meu pai, à tia Loló, e à tia Rosinha, pelo imenso carinho e valiosos conselhos.

À toda família, que sempre preocupou-se comigo e em especial aos meus primos-irmãos.

Aos meus queridos amigos de Rio Pomba Dominique e Helisson, pela verdadeira amizade e companheirismo.

À minha querida amiga e parceira de missa Laura, que me aturou nesses anos e que sempre me fez feliz.

Ao Samuel, pelas aventureiras caronas no Trovão Azul e pela agradável companhia e risadas dadas na cantina do ICE.

Ao Iago, fiel combatente de curso, muito obrigado pela amizade e incentivo. Ao Luís, outro grande companheiro que se fez presente nessa jornada, obrigado por ter lhe conhecido e pela motivação para malhar. Ao Roberto, grande amigo de faculdade, obrigado pela companhia e hospitalidade. O mundo é nosso!

E aos demais amigos e colegas de faculdade Carol, Priscila , Thiago e Victor.

Enfim, um muito obrigado a todos que me ajudaram nessa longa caminhada!

RESUMO

Esse trabalho apresenta um estudo sobre as associações, similaridades ou dissimilaridades de alguns indicadores socioeconômicos, demográficos e de mortalidade para todos os estados e regiões brasileiras a partir da técnica estatística multivariada denominada Análise de Correspondência (AC). Esses indicadores foram retirados do IDB-2009 (Indicadores e Dados Básicos-2009), e também do site do Instituto de Pesquisa Econômica Aplicada (IPEA). Inicialmente, calculamos algumas estatísticas descritivas para os indicadores referentes aos estados brasileiros e fizemos gráficos *box-plot* para cada variável (indicador). Para a aplicação aqui apresentada, já que a Análise de Correspondência (AC) é estritamente desenvolvida para trabalhar com variáveis categóricas ou categorizadas, e os indicadores estudados são de natureza numérica, utilizamos dois métodos a fim de agrupá-los em categorias ordinais quanto aos estados e regiões - método das k-Médias e dos percentis. Apresentamos em nossa análise o método das k-Médias pois este possibilitou uma maior explicação da variância do sistema ao aplicarmos a AC. Contudo, o resultado utilizando o método dos percentis também foi bastante satisfatório e similar ao primeiro. Recodificamos os valores da matriz obtida pelo método das k-Médias através da correção chamada *doubling* e obtivemos resultados similares àqueles quando não fizemos tal correção, embora a visualização das similaridades e associações tenha ficado mais direta. Desta forma, pudemos analisar a posição relativa dos estados, suas similaridades em termos dos indicadores e quais indicadores são similares para os estados brasileiros, assim como a associação entre os estados e determinadas categorias dos indicadores. A utilização de AC foi escolhida em função da possibilidade de analisar tais similaridades e associações de forma gráfica, o que caracteriza esta técnica. Todos os cálculos e gráficos foram obtidos através do software livre R.

Palavras-chave: Análise de Correspondência; Análise Multivariada; Indicadores Sociais.

ABSTRACT

We present a study of the associations, similarities or dissimilarities among selected socioeconomic, demographic and mortality indicators for all states and regions in Brazil, through the adoption of a multivariate statistical technique called Correspondence Analysis (CA). The indicators were gathered from IDB-2009 (Indicators and Basic Data-2009) and also from IPEA (Applied Economic Research Institute). Some descriptive statistics for all indicators used in the analysis are presented, along with corresponding box-plots. As CA was strictly developed for categorical data, and the indicators we selected were all numerical variables, two grouping methods were used in order to categorize such variables: k-means and percentiles. The analysis of the categorized variables through the k-means method presented a solution which best explained the overall variance in the CA solution and was therefore adopted. However, the results obtained through the percentiles method presented similar results. The scaling correction known as "doubling" was then applied to the k-means generated data and another CA solution was obtained. These results were similar to the previous one, but the visual interpretation was more straightforward. We could thus analyze the relative position of states, their similarities in terms of the indicators adopted and their association with such indicators. AC was selected for its unique feature of presenting the solution in a clear, easy-to-interpret graphic form. All calculations were made in the open-source programming code R.

Key-words: Correspondence Analysis; Multivariate Analysis; Social Indicators.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

Indicadores demográficos, socioeconômicos e de mortalidade, entre outros, são muito importantes para conhecer a estrutura de determinado lugar ou região e ajudar na decisão de investir ou não em políticas públicas e o quanto investir.

Esse trabalho apresenta um estudo sobre as associações, similaridades ou dissimilaridades de alguns indicadores socioeconômicos, demográficos e de mortalidade para todos os estados e regiões brasileiras a partir da técnica estatística multivariada denominada Análise de Correspondência (AC).

Nesse trabalho iremos estudar uma aplicação da AC em um caso particular, onde temos uma estrutura de tabela de classificação cruzada e não uma tabela de contingência, como podemos ver na maioria dos artigos e trabalhos publicados.

Assim, nossos elementos de linha são compostos pelos estados e regiões brasileiras e as colunas, pelos indicadores. As células, ao invés de representarem freqüências observadas, são na verdade, a categoria (ordinal) assumida pelo elemento de linha para determinado indicador.

Esses indicadores foram retirados do IDB (Indicadores e Dados Básicos-2009) e também do site do Instituto de Pesquisa Econômica Aplicada (IPEA).

Desta forma, podemos analisar a posição relativa dos estados, suas similaridades em termos dos indicadores e quais indicadores são similares para os estados brasileiros, assim como a associação entre os estados e determinadas categorias dos indicadores. A utilização da AC foi escolhida em função da possibilidade de analisar tais similaridades e associações de forma gráfica, o que caracteriza esta técnica.

2 OBJETIVOS

Esse trabalho visa aplicar a técnica exploratória denominada Análise de Correspondência a um conjunto de dados dispostos em forma de tabela cruzada, onde as células representam o valor da categoria assumida por tal estado ou região frente a cada um dos 15 indicadores demográficos, socioeconômicos e de mortalidade, com o intuito de encontrarmos associações, similaridades ou dissimilaridades entre os elementos de linha e coluna.

Faremos uma comparação dos resultados da AC para o caso em que adotamos uma recodificação dos dados chamada *doubling* que objetiva gerar pólos positivo e negativo a cada uma das variáveis do estudo com o intuito de obter totais marginais de linha iguais na tabela cruzada.

Além do gráfico da AC, também conhecido por mapa, iremos interpretar as contribuições dos pontos de linha e coluna, assim como as contribuições relativas das dimensões para as inércias desses pontos.

Como outro objetivo, destacamos a divulgação da flexibilidade da técnica de AC, cuja utilização não é restrita a tabelas de contingência, fato nem sempre conhecido e divulgado.

3 INDICADORES

Usamos 7 indicadores socioeconômicos (Trabalho Infantil-2008, PIB (Produto Interno Bruto) *per capita* - 2007, Taxa de Desemprego em pessoas de 16 e mais anos-2008, Percentual da população com Renda Inferior a 1/2 Salário Mínimo - 2008, Proporção de Analfabetos em maiores de 15 anos - 2008, Proporção da População de 15 anos ou mais com mais de 8 anos de estudo - 2008, Índice de Gini - 2009) e 7 indicadores demográficos, todos referentes ao ano de 2008, com exceção do IDH-2000 (Esperança de vida ao nascer e aos 60 anos de idade, Índice de Envelhecimento, Razão de Dependência de Idosos, IDH, Área e População) a fim de explorar suas associações com os estados e regiões brasileiras.

Usamos também um indicador de mortalidade: o indicador de Mortalidade Infantil do ano de 2008.

Todos esses indicadores, com exceção de área, índice de Gini e IDH foram retirados do IDB (Indicadores e Dados Básicos-2009), que é um produto da Rede Interagencial para a Saúde – RIPSA. O IDB é composto por um conjunto de mais de 100 indicadores, divididos em 6 grupos: demográficos, socioeconômicos, mortalidade, morbidade e fatores de risco, recursos e cobertura. São desdobrados segundo as unidades da federação, suas capitais e regiões metropolitanas, podendo ser categorizadas também por faixa etária, sexo ou outras características, de acordo com o indicador.

A RIPSA viabiliza parcerias entre entidades representativas dos segmentos técnicos e científicos nacionais envolvidos na produção, análise e disseminação de dados, objetivando sistematizar informações úteis ao conhecimento e à compreensão da realidade sanitária brasileira e de suas tendências.

Formalizada por Portaria Ministerial e Termo de Cooperação com a Opas/OMS, a RIPSA promove a organização e a manutenção de uma base de indicadores relativos ao

estado de saúde da população e aos aspectos sociais, econômicos e organizacionais que influenciam e determinam a situação de saúde. Os produtos da Rede resultam de um processo de construção coletiva, no qual as instituições parceiras contribuem com a própria expertise, por meio de seus profissionais e bases técnico-científicas.

Compõem a RIPSA cerca de 30 entidades representativas dos segmentos técnicos e científicos nacionais envolvidos na produção e análise de dados (produtores de informações estrito senso, gestores do sistema de saúde e unidades de ciência e tecnologia), que se associaram para aperfeiçoar informações de interesse comum.

Quanto aos indicadores, a taxa de Mortalidade Infantil (TMI) tem sido utilizada internacionalmente como um dos principais indicadores da qualidade de vida da população (LAURENTI et al.,1985). Ela refere-se à mortalidade ocorrida no primeiro ano de vida e mede o risco que os nascidos vivos com menos de um ano têm de morrer antes de completar um ano de idade.

O Trabalho Infantil refere-se ao percentual da população residente de 10 a 15 anos de idade que se encontra trabalhando ou procurando trabalho na semana de referência, em determinado espaço geográfico, no ano considerado (DATASUS, 2010).

O Percentual de população com renda domiciliar mensal per capita de até meio salário mínimo, em determinado espaço geográfico, no ano considerado, expressa a proporção da população considerada em situação de pobreza ou de extrema pobreza, de acordo com a renda domiciliar mensal per capita. Esse indicador possibilita discriminar as regiões brasileiras em condições precárias de sobrevivência, que requerem maior atenção de políticas públicas de geração de renda, saúde, educação etc (DATASUS, 2010).

A proporção de analfabetos em pessoas com 15 e mais anos de idade mede o percentual da população adulta que não sabe ler e escrever pelo menos um bilhete simples, no idioma que conhecem, na população total residente da mesma faixa etária, em determinado

espaço geográfico, no ano considerado. Esse indicador é amplamente utilizado para dimensionar a situação de desenvolvimento socioeconômico de um grupo social em seu aspecto educacional. Contribui também para a análise das condições de vida e saúde da população, utilizando esse indicador como *proxy* da condição econômico-social da população. A atenção à saúde das crianças é influenciada positivamente pela alfabetização da população adulta, sobretudo das mães (DATASUS, 2010).

A proporção da população de 15 anos e mais com mais de 8 anos de estudo expressa o nível de instrução dessa população, e assim como o indicador de analfabetismo descrito acima, é utilizada como *proxy* da situação socioeconômica da população (DATASUS, 2010).

O coeficiente de GINI mede o grau de desigualdade existente na distribuição da renda domiciliar per capita. Seu valor varia de 0 a 1, assumindo o valor 0 quando não há desigualdade alguma (todas as pessoas da região recebendo exatamente a mesma renda) e o valor 1 quando há o mais extremo grau de desigualdade possível (um único indivíduo concentrando toda a renda e todos os demais da região sem renda alguma) (IPEA, 2010).

A Taxa de Desemprego é o percentual da população economicamente ativa que se encontra sem trabalho na semana de referência, em determinado espaço geográfico, no ano considerado. É calculado como a razão entre o número de residentes de 16 e mais anos de idade que se encontram desocupados e procurando trabalho, na semana de referência, sobre o número de residentes economicamente ativos (PEA) desta faixa etária. Essa razão é multiplicada por 100 (DATASUS, 2010).

O PIB *per capita* é o valor médio agregado por indivíduo, em moeda corrente e a preços de mercado, dos bens e serviços finais produzidos em determinado espaço geográfico, no ano considerado.Ele indica o nível de produção econômica em um território, em relação ao seu contingente populacional. Valores muito baixos assinalam, em geral, a existência de

segmentos sociais com precárias condições de vida. O PIB *per capita* é construído pelo Sistema de Contas Nacionais e calculado como a razão do valor do PIB em moeda corrente, a preços de mercado pela população total residente (DATASUS, 2010).

Quanto aos indicadores demográficos, a Esperança de Vida ao Nascer é o número médio de anos de vida esperados para um recém-nascido, mantido o padrão de mortalidade existente na população residente, em determinado espaço geográfico, no ano considerado. É calculada pela metodologia denominada tábuas de vida. O aumento desse indicador sugere melhoria das condições de vida e saúde da população (DATASUS, 2010).

A Esperança de Vida aos 60 anos é calculada de forma similar à Esperança de Vida ao Nascer, sendo uma importante informação para os setores de saúde, previdência e assistência social. Esses dois indicadores se limitam pelo fato de haver imprecisões relacionadas a falhas na declaração de idades nos levantamentos estatísticos ou à metodologia empregada para elaborar estimativas e projeções populacionais na base de dados utilizados para o cálculo do indicador (DATASUS, 2010).

O Índice de Envelhecimento é o número de pessoas de 60 e mais anos de idade, para cada 100 pessoas menores de 15 anos de idade, na população residente em determinado espaço geográfico, no ano considerado. Valores elevados desse índice indicam que a transição demográfica encontra-se em estágio avançado. É importante para acompanhar a evolução do ritmo de envelhecimento da população, comparativamente entre áreas geográficas e grupos sociais. Também é limitado devido as mesmas condições citados na Esperança de Vida ao Nascer e aos 60 anos (DATASUS, 2010).

A Razão de Dependência de Idosos é calculada como a razão entre os que possuem 60 e mais anos de idade (economicamente dependentes) e o segmento etário potencialmente produtivo (entre 15 e 59 anos de idade). Esse indicador mede a participação relativa do contingente populacional potencialmente inativo que deveria ser sustentado pela

parcela da população potencialmente produtiva. Valores elevados indicam que a população em idade produtiva deve sustentar uma grande proporção de dependentes, o que significa consideráveis encargos assistenciais para a sociedade (DATASUS, 2010).

A população é o número de indivíduos residentes e sua estrutura relativa, em determinado espaço geográfico, no ano considerado e a área representa a dimensão de cada estado e região (DATASUS, 2010).

No Anexo C encontra-se o link de todos indicadores utilizados nesse estudo.

4 MÉTODO DAS K-MÉDIAS

O método das k-Médias é um dos métodos de agrupamento mais utilizados e consiste simplesmente em alocar os elementos amostrais àquele cluster cujo centróide (vetor de médias amostral) é o mais próximo do vetor de valores observados para tal elemento (MINGOTI, 2005).

O método possui quatro passos que são:

1. Escolhe-se k centróides, chamados de “sementes”, para iniciar o processo;

2. Através da distância Euclidiana, cada elemento é comparado com cada centróide inicial (semente). Daí, o elemento é alocado ao grupo cuja distância é a menor;

3. Após o passo 2, calcula-se os valores dos centróides para cada novo grupo formado e então, repete-se o passo 2;

4. Os passos 2 e 3 são repetidos até que nenhuma realocação de elementos seja necessária.

Existem formas distintas de se determinar as sementes iniciais para implementar o método das k-Médias. Em nosso trabalho, utilizamos o método de escolha prefixada, onde escolhemos os percentis 10, 30, 50, 70 e 90 (mediana de cada uma das cinco categorias do método dos percentis) de cada indicador a fim de categorizá-lo quanto aos estados e regiões como nossas “sementes”. Existem outros métodos de escolha como o de técnicas aglomerativas, escolha aleatória, escolha via variável aleatória, valores discrepantes e até dos k primeiros valores do banco de dados.

5 DECOMPOSIÇÃO EM VALORES SINGULARES

A decomposição em valores singulares é uma extensão da teoria de diagonalização de matrizes simétricas nxn para matrizes mxn arbitrárias (ANTON, 2006)

Existem dois caminhos para percorrer na procura de outros tipos de fatoração de uma matriz quadrada A arbitrária. Podemos procurar fatorações da forma

$$A = PJP^{-1}$$

em que P é invertível mas não necessariamente ortogonal, ou então da forma

$$A = U\sum V^T ,$$

onde $\sum$ é uma matriz diagonal composta de valores singulares organizados em ordem decrescente em sua diagonal principal, U é a matriz que contém os vetores singulares à esquerda de A e V é a matriz que contém os vetores singulares de A. Neste trabalho iremos trabalhar com o segundo caminho.

5.1 Teorema

Se A é uma matriz nxn de posto k, então A pode ser fatorada como

$$A = U\sum V^T$$

onde U e V são matrizes ortogonais nxn e $\sum$ é uma matriz diagonal nxn cuja diagonal principal tem k entradas positivas e $n-k$ entradas nulas. Então,

(a) $V = [v_1\ v_2 ... v_n]$ diagonaliza ortogonalmente $A^T A$.

(b) As entradas não nulas de $\sum$ são

$$\sigma_1 = \sqrt{\lambda_1}\ ,\ \sigma_2 = \sqrt{\lambda_2}\ ,...,\ \sigma_k = \sqrt{\lambda_k}$$

onde $\lambda_1,\ \lambda_2,...,\ \lambda_k$ são autovalores não-nulos de $A^T A$ associados aos vetores-coluna de V.

(c) Os vetores-coluna de V são ordenados de tal forma que $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_k > 0$.

(d) $u_i = \dfrac{Av_i}{\|Av_i\|} = \dfrac{1}{\sigma_i} Av_i \ \ (i = 1, 2, ..., k)$.

(e) $\{u_1, u_2, ..., u_k\}$ é uma base ortonormal de $col(A)$ (subespaço de R^n gerado pelos vetores - coluna de A)

(f) $\{u_1, u_2, ..., u_k, u_{k+1}, ..., u_n\}$ é uma extensão de $\{u_1, u_2, ..., u_k\}$ a uma base ortonormal de R^n.

Por exemplo, veja como é feita a decomposição em valores singulares da matriz

$$A = \begin{pmatrix} \sqrt{3} & 2 \\ 0 & \sqrt{3} \end{pmatrix}$$

Primeiramente devemos encontrar os autovalores da matriz

$$A^T A = \begin{pmatrix} \sqrt{3} & 0 \\ 2 & \sqrt{3} \end{pmatrix} \begin{pmatrix} \sqrt{3} & 2 \\ 0 & \sqrt{3} \end{pmatrix} = \begin{pmatrix} 3 & 2\sqrt{3} \\ 2\sqrt{3} & 7 \end{pmatrix}$$

O polinômio característico de $A^T A$ é

$$\lambda^2 - 10\lambda + 9 = (\lambda - 9)(\lambda - 1)$$

onde os autovalores de $A^T A$ são $\lambda_1 = 9$ e $\lambda_2 = 1$ e os valores singulares de A são

$$\sigma_1 = \sqrt{\lambda_1} = \sqrt{9} = 3\,,\ \sigma_2 = \sqrt{\lambda_2} = \sqrt{1} = 1$$

Os autovetores unitários de $A^T A$ associados aos autovalores $\lambda_1 = 9$ e $\lambda_2 = 1$ são

$$v_1 = \begin{bmatrix} \frac{1}{2} \\ \frac{\sqrt{3}}{2} \end{bmatrix} e\ v_2 = \begin{bmatrix} -\frac{\sqrt{3}}{2} \\ \frac{1}{2} \end{bmatrix}$$ respectivamente.

Assim,

$$u_1 = \frac{1}{\sigma_1} A v_1 = \frac{1}{3}\begin{pmatrix} \sqrt{3} & 2 \\ 0 & \sqrt{3} \end{pmatrix}\begin{pmatrix} \frac{1}{2} \\ \frac{\sqrt{3}}{2} \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{3}}{2} \\ \frac{1}{2} \end{pmatrix},$$

$$u_2 = \frac{1}{\sigma_2} A v_2 = 1\begin{pmatrix} \sqrt{3} & 2 \\ 0 & \sqrt{3} \end{pmatrix}\begin{pmatrix} -\frac{\sqrt{3}}{2} \\ \frac{1}{2} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} \\ \frac{\sqrt{3}}{2} \end{pmatrix}$$

e portanto

$$U = [u_1\ u_2] = \begin{pmatrix} \sqrt{3} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix} \text{ e } V = [v_1\ v_2] = \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}$$

A decomposição em valores singulares de A é

$$\begin{pmatrix} \sqrt{3} & 2 \\ 0 & \sqrt{3} \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{-\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}$$

$$A \quad = \quad U \quad \sum \quad V^T$$

A decomposição em valores singulares de matrizes não quadradas se dá de forma similar ao método para matrizes quadradas.

6 ANÁLISE DE CORRESPONDÊNCIA

A Análise de Correspondência (AC) é uma das diversas técnicas de análise multivariada desenvolvida para o estudo da relação entre variáveis qualitativas em tabelas de contingência (BEH, 2004). Ela permite a visualização gráfica das linhas e colunas como pontos em espaços vetoriais de dimensões reduzidas em um novo sistema de eixos ortogonais.

A AC tem se tornado muito comum devido a sua fácil implementação com o aumento dos recursos computacionais. Necessita apenas de uma tabela com números positivos que representam freqüências observadas de objetos ou indivíduos classificados por uma categoria de linha e uma categoria de coluna. Tais categorias devem ser mutuamente exclusivas e exaustivas, ou seja, um indivíduo ou objeto não pode ser classificado em mais de uma categoria de uma mesma variável e devem existir categorias suficientes para esse ser classificado (GREENACRE, 1984).

AC também pode ser estendida para a situação em que a tabela representa apenas o cruzamento de duas variáveis categóricas, onde a variável de coluna normalmente representa uma série de indicadores medidos na mesma escala ou normalizados (GOUVÊA, 1990) e as linhas são categorias de uma variável nominal.

6.1 Análise de Correspondência Simples

A seguir apresentaremos conceitos da Análise de Correspondência Simples (ACS), além de um exemplo de aplicação a uma tabela de contingência de dimensão $3x6$.

A tabela de contingência de duas entradas coloca em correspondência duas variáveis categóricas, onde, por exemplo, *I* é o conjunto das classes sociais e *J*, o conjunto das

rádios. Os elementos que constituem estes conjuntos são denotados por *i*, com *i* =1,..., *i*, ... *I* e *j*, com *j* = 1,..., *J*, respectivamente.

Tabela 6.1 - Representação de uma tabela de contingência

	B						
A	1	2	...	j	...	J	Coluna marginal
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1J}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2J}	$n_{2.}$
...	...	...	...	...	...	...	
i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iJ}	$n_{i.}$
...	...	...	...	...	...	...	
I	n_{I1}	n_{I2}	...	n_{Ij}	...	n_{IJ}	$n_{I.}$
Linha marginal	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.J}$	$n_{..}$

Na tabela acima, denotamos:

- n_{ij}, a interseção da linha i com a coluna j, é o número de ocorrências de i e j, conjuntamente, sendo $i \in I$ e $j \in J$.
- $n_{.j}$ é um dos elementos da linha marginal ou, o j-ésimo termo da linha marginal, que é a soma do número de observações na j-ésima coluna, isto é:

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

Analogamente, $n_{i.}$ é o i-ésimo termo da coluna marginal, que é a soma do número de observações da linha i

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

- $n_{..}$ é o total principal, a soma de todas as observações , portanto equivale à soma da linha marginal ou da coluna marginal.

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

A partir desses conceitos iniciais, segue abaixo a Tabela 6.2, referente ao exemplo das classes sociais versus rádio preferida na cidade de Juiz de Fora. Esses dados são de uma Pesquisa de Opinião encomendada ao Departamento de Estatística da Universidade Federal de Juiz de Fora no ano 2000.

Tabela 6.1 - Tabela de Contingência entre Classe Social e Rádio Preferida acrescida das linhas e colunas marginais

Classe Social	Rádio Preferida						
	Solar	Globo	Itatiaia	Cidade	Pio XI	Outras	$n_{i.}$
A	48	10	9	6	12	14	99
B	119	24	20	8	10	16	197
C	68	5	3	4	1	2	83
$n_{.j}$	235	39	32	18	23	32	$n_{..} = 379$

Fonte: Pesquisa de Opinião DE/ICE/UFJF (2000)

A figura 1 a seguir mostra o mapa de correspondência das variáveis Classe Social e Rádio Preferida. Veja como a saída gráfica da ACS nos permite fazer algumas interpretações sobre as proximidades entre as categorias das linhas e colunas, como por exemplo, que existe uma associação entre a classe social C e a rádio Solar, bem como a associação entre a classe A e as rádios Cidade, Pio XII e Outras.

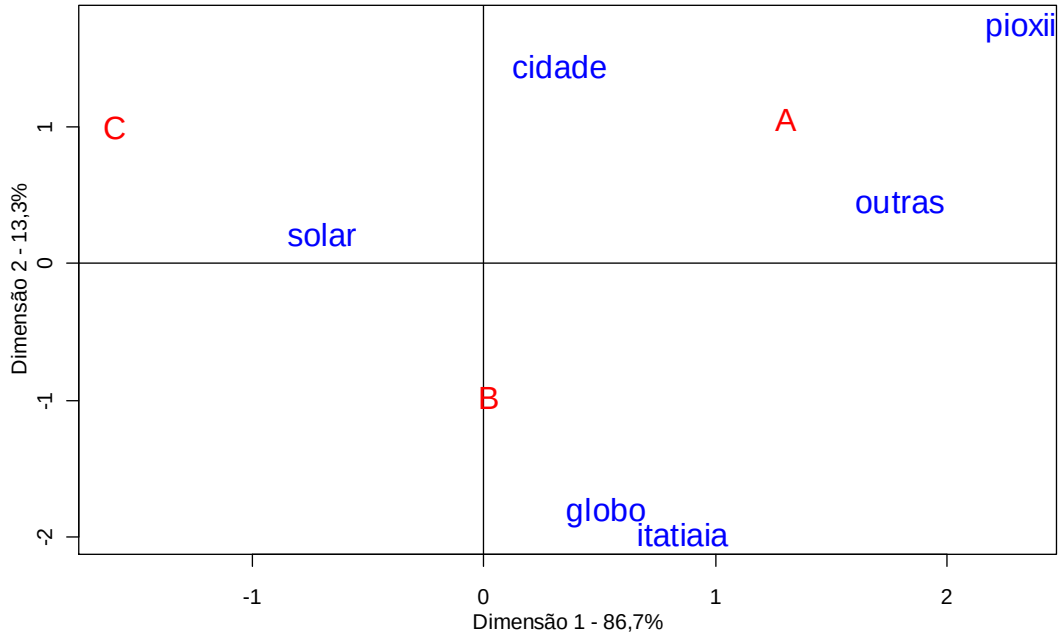

Figura 6.1 – Análise de Correspondência entre classe social e preferência de rádio

6.2 Matriz de Correspondência

A matriz de correspondência é calculada como:

$$p_{ij} = \frac{n_{ij}}{n..}, \; i=1,2,...,I \text{ e } j=1,2,...,J.$$

Encontra-se abaixo a matriz de correspondência para o exemplo das Classes Sociais e Rádio Preferida.

$$P_{(3X6)} = \frac{1}{n..} N_{(3x6)} = \begin{bmatrix} 0{,}126 & 0{,}026 & 0{,}023 & 0{,}015 & 0{,}031 & 0{,}036 \\ 0{,}313 & 0{,}063 & 0{,}052 & 0{,}021 & 0{,}026 & 0{,}042 \\ 0{,}179 & 0{,}013 & 0{,}007 & 0{,}010 & 0{,}002 & 0{,}005 \end{bmatrix}$$

6.2.1 Massa de um elemento

A massa de um elemento $i \in I$ é definida como a divisão do total da *i*-ésima linha pelo total geral sendo denotada por r_i.

$$r_i = \frac{n_{i.}}{n..}$$

A massa de um elemento $j \in J$ é definida como:

$$c_j = \frac{n_{.j}}{n..}$$

Assim, o vetor de massas das linhas e colunas é a proporção de cada elemento da coluna marginal e linha marginal em relação ao total geral. Para o exemplo das Classes Sociais e Rádio Preferida, os vetores massas das linhas e colunas são

$$r^T_{(3X1)} = [0,261\ 0,519\ 0,218]$$

$$c^T_{(6X1)} = [0,620\ 0,102\ 0,084\ 0,047\ 0,060\ 0,084]$$

onde r^T é o vetor transposto de massa das linhas e c^T o vetor transposto de massa das colunas.

6.3 Variáveis suplementares

A AC permite incluir uma ou mais variáveis como suplementares em sua análise. Isso significa que elas figuram na solução gráfica, mas não fazem parte da solução - têm, portanto massa igual a zero - ou seja, não colaboram com os totais de linha, coluna e geral.

Os motivos para considerar pontos como suplementares estão no fato deles distorcerem a análise (ao distorcerem a configuração do gráfico) ou quando essas observações representam informações adicionais sobre os indivíduos (objetos).

6.4 Perfil

O perfil de uma linha (coluna) é um vetor formado pela divisão de cada termo n_{ij} pelo total $n_{i.}$ ($n_{.j}$). Os perfis de linha podem ser definidos como:

$$r^{i}{}_{J} = \left\{ r^{i}{}_{J} = \frac{n_{ij}}{n_{i.}} \;\middle|\; j \in J \right\} \forall_{j} = 1,...,J \text{ se } \sum_{j=1}^{J} r^{i}{}_{j} = 1.$$

Assim, o perfil de uma linha *i*, por exemplo, é a descrição da relação entre esta linha com todos os elementos de *J*.

Para as colunas,

$$c^{j}{}_{I} = \left\{ c^{j}{}_{i} = \frac{n_{ij}}{n_{.j}} \;\middle|\; i \in I \right\}$$

é o perfil da coluna *j*, se $c^{j}{}_{I}$ é a participação relativa de *i* na coluna *j*, $\forall_{i} = 1,...,I$ se $\sum_{i=1}^{I} c^{j}{}_{i} = 1$.

Além do perfil da linha e coluna, pode-se determinar também o perfil das marginais de linha e coluna, tal que *c* é o perfil da coluna marginal e *r* o perfil da linha marginal, definidos por:

$$\text{Perfil da linha marginal: } r_{J} = \left\{ r_{j} = \frac{n_{.j}}{n_{..}} \mid j \in J \right\}$$

$$\text{Perfil da coluna marginal: } c_{I} = \left\{ c_{i} = \frac{n_{i.}}{n_{..}} \mid i \in I \right\}$$

Assim, o perfil da linha marginal é igual ao vetor massas de coluna e o perfil da coluna marginal é igual ao vetor massas de linha.

No propósito de enriquecer a ideia apresentada acima, segue a matriz dos perfis de linha e coluna para o exemplo das Classes Sociais e Rádio Preferida.

$$R_{(3x6)} = D_r{}^{-1}{}_{(3x3)} P = \begin{bmatrix} 0{,}484 & 0{,}101 & 0{,}090 & 0{,}060 & 0{,}121 & 0{,}141 \\ 0{,}604 & 0{,}121 & 0{,}101 & 0{,}040 & 0{,}050 & 0{,}081 \\ 0{,}819 & 0{,}060 & 0{,}036 & 0{,}048 & 0{,}012 & 0{,}024 \end{bmatrix} \text{ e}$$

$$C_{(6x3)} = D_c{}^{-1}{}_{(6x6)} P^T{}_{(3X6)} = \begin{bmatrix} 0{,}204 & 0{,}506 & 0{,}289 \\ 0{,}256 & 0{,}615 & 0{,}128 \\ 0{,}281 & 0{,}625 & 0{,}093 \\ 0{,}333 & 0{,}444 & 0{,}222 \\ 0{,}521 & 0{,}434 & 0{,}043 \\ 0{,}437 & 0{,}500 & 0{,}062 \end{bmatrix}, \text{ onde}$$

$$D_{r(3x3)} = diag(r) = \begin{bmatrix} 0{,}2612 & 0{,}000 & 0{,}000 \\ 0{,}000 & 0{,}5197 & 0{,}000 \\ 0{,}000 & 0{,}000 & 0{,}2189 \end{bmatrix}$$

é a matriz diagonal do perfil da coluna marginal e

$$D_{c(6x6)} = diag(c) = \begin{bmatrix} 0{,}620 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 \\ 0{,}000 & 0{,}102 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 \\ 0{,}000 & 0{,}000 & 0{,}084 & 0{,}000 & 0{,}000 & 0{,}000 \\ 0{,}000 & 0{,}000 & 0{,}000 & 0{,}047 & 0{,}000 & 0{,}000 \\ 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}060 & 0{,}000 \\ 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}084 \end{bmatrix},$$

matriz diagonal do perfil da linha marginal.

6.5 Nuvem de pontos

Uma nuvem de pontos é a representação dos perfis em espaços vetoriais de n dimensões, sendo $n = \min(I-1, J-1)$ (BENÉCRI, 1992).

A nuvem dos perfis de *I* (N_I), definida como o conjunto dos perfis de cada linha *i*, cada um associado à sua massa, estará representada em um espaço de dimensão igual ao número de elementos de *J-1*, já que há uma dependência linear entre as coordenadas, ou seja, os elementos de cada perfil somam 1. Raciocínio análogo para a nuvem dos perfis de *J* (N_J).

Pensando no exemplo das Classes Sociais e Rádio Preferida, é como se cada perfil de linha pudesse ser representado em um espaço de 6 dimensões e cada perfil de coluna, em um espaço tridimensional. Como cada perfil soma 1, há uma dependência linear entre as coordenadas que significa que os perfis de linha estão contidos em um espaço penta-dimensional e os perfis de coluna, em um espaço bidimensional.

6.6 Centro de gravidade, Perfil Médio ou Centróide

A ideia de centro de gravidade é a generalização espacial da noção de média ponderada em que cada ponto é representado de forma proporcional à sua massa. Dessa forma, o centro de gravidade da nuvem de pontos formada pelos perfis do conjunto *I* (N_I) é o perfil da linha marginal r_J, e o centro de gravidade da nuvem de pontos N_J é o perfil da coluna marginal c_I. Como o centro de gravidade é uma espécie de média ponderada pelas massas, ele fica sujeito a maior influência daqueles pontos com maior massa - isto pode fazer com que o resultado da AC seja muito influenciado por esses pontos. Por isso, é conveniente, em determinadas situações, considerá-los como elementos suplementares.

6.7 Distância

A distância entre dois perfis de linha (i e i') é dada por:

$$d^2(r_J^i, {}_J^{i'}) = \sum_{j \in J} \frac{(r_j^i - r_j^{i'})^2}{r_j}$$

e entre dois perfis de coluna (j e j')

$$d^2(c_I^j, {}_I^{j'}) = \sum_{j \in J} \frac{(c_i^j - c_i^{j'})^2}{c_i}$$

Esta métrica é conhecida como métrica χ^2 por apresentar uma relação entre a distância a ela associada (distância χ^2) e a estatística χ^2 de Pearson (GREENACRE, 1984, p.31-33; BENZÉCRI, 1992, 54-5; BEH, 2004, 261-2).

6.8 Inércia

A inércia de uma nuvem de pontos em relação ao seu centro de gravidade é uma medida de variação total, que objetiva captar a dispersão dos pontos da nuvem em torno do centro de gravidade. A inércia total pode ser calculada como o quociente da estatística Qui-Quadrado de Pearson da tabela em relação ao total da mesma

$$I = \frac{\chi^2}{n}$$

Uma nuvem será tão mais concentrada em torno do centro de gravidade quanto menor for a inércia (variância total do sistema).

Em notação matricial, a inércia total é dada por:

$$I = traço[D_r^{-1}(P - rc^T)D_c^{-1}(P - rc^T)^T]$$

sendo D_r e D_c as matrizes diagonais do perfis da coluna e linha marginal, respectivamente.

Em nosso exemplo, a inércia total é dada por

$$I = traço \begin{bmatrix} 0,034 & -0,004 & -0,029 \\ -0,002 & 0,005 & -0,002 \\ -0,035 & -0,006 & 0,041 \end{bmatrix}$$

$$I = 0,034 + 0,005 + 0,041 = 0,081$$

A inércia dos perfis de linha é igual à dos perfis de coluna e ambas iguais a inércia total.

Do mesmo modo, a inércia de uma nuvem de pontos em relação a um eixo E_1 é dada por:

$$I_1 = \sum_{i=1}^{n} p_i d^2(i, E_1)$$

em que $d^2(i, E_1)$ é a distância ao quadrado do ponto *i* até o eixo.

E a inércia projetada sobre o eixo:

$$I_2 = \sum_{i=1}^{n} p_i d^2(p_J, x_i)$$

em que x_i é a projeção do ponto i sobre E_1. Assim, essa associação se dá de tal forma que

$$I = I_1 + I_2$$

O que interessa é que a inércia em relação ao eixo seja a menor possível, ou seja, que a inércia projetada seja a maior. Já que I é constante, para que I_1 seja mínima é preciso que I_2 seja máxima.

6.9 Subespaço ótimo

O objetivo da AC é encontrar o subespaço ótimo de menor dimensão que melhor se ajusta aos pontos da nuvem (BENZÉCRI, 1992). Daí, esse subespaço ótimo é formado pelos eixos fatoriais que passam pelo centro de gravidade e que minimizam as distâncias dos pontos até ele. Primeiramente encontra-se um eixo que passa pelo centro de gravidade e que minimiza a distância perpendicular dos pontos a reta (primeira dimensão). Logo após, traça-se uma segunda reta ortogonal a primeira que também passa pelo centro de gravidade e que minimize as distâncias dos pontos em relação a essa reta. Esse processo ocorrerá até se encontrar min (*I-1*, *J-1*) dimensões, pois os perfis somam 1 (dependência linear).

6.10 Coordenadas dos perfis no novo sistema de eixo fatorial

Uma forma de se obter as coordenadas dos perfis de linha e coluna no novo sistema de eixo fatorial é através da decomposição em valores singulares da matriz $A = UD_{\mu}V^{T}$, onde U contém os vetores singulares a esquerda de A, V os valores singulares a direita de A e D_{μ} é uma matriz diagonal de números positivos, valores singulares, em ordem decrescente. Dessa forma as coordenadas principais dos perfis de linha e coluna são dadas por $F = D_r^{-1/2}UD_{\mu}$ e $G = D_c^{-1/2}VD_{\mu}$, respectivamente.

A matriz A possui como valor característico

$$a_{ij} = \frac{pij - p_{i.}p_{.j}}{\sqrt{p_{i.}p_{.j}}}$$

e pode ser calculada também a partir da multiplicação de matrizes a seguir

$$A = D_r^{-1/2}(P - rc^T)D_c^{-1/2}$$

A matriz A de nosso exemplo é dada por

$$A = \begin{bmatrix} -0{,}087 & -0{,}003 & 0{,}011 & 0{,}030 & 0{,}125 & 0{,}100 \\ -0{,}014 & 0{,}042 & 0{,}042 & -0{,}022 & -0{,}029 & -0{,}007 \\ 0{,}118 & -0{,}062 & -0{,}077 & 0{,}001 & -0{,}092 & -0{,}097 \end{bmatrix}$$

e sua decomposição em valores singulares fornece

$$U = \begin{bmatrix} -0{,}665 & 0{,}544 & 0{,}511 \\ -0{,}012 & -0{,}692 & 0{,}720 \\ 0{,}746 & 0{,}473 & 0{,}467 \end{bmatrix}$$

$$D_\mu = \begin{bmatrix} 0{,}266 & 0 & 0 \\ 0 & 0{,}104 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ e}$$

$$V = \begin{bmatrix} 0{,}552 & 0{,}176 & 0{,}711 \\ -0{,}169 & -0{,}579 & 0{,}515 \\ -0{,}248 & -0{,}574 & 0{,}133 \\ -0{,}071 & 0{,}318 & 0{,}237 \\ -0{,}571 & 0{,}428 & 0{,}387 \\ -0{,}522 & 0{,}134 & 0{,}065 \end{bmatrix}$$

A partir desses resultados, encontramos as coordenadas principais dos perfis de linha e coluna em relação a todos eixos principais, de forma que cada coluna representa as coordenadas de uma dimensão (GREENACRE, 1984).

$$F = \begin{bmatrix} \text{-0,346} & \text{0,111} & 0 \\ \text{-0,004} & \text{-0,100} & 0 \\ \text{0,424} & \text{0,105} & 0 \end{bmatrix} \text{e}$$

$$G = \begin{bmatrix} \text{0,186} & \text{0,023} & 0 \\ \text{-0,140} & \text{-0,188} & 0 \\ \text{-0,227} & \text{-0,206} & 0 \\ \text{- 0,087} & \text{0,152} & 0 \\ \text{-0,617} & \text{0,181} & 0 \\ \text{-0,478} & \text{0,048} & 0 \end{bmatrix}$$

6.11 Decomposição da inércia total

Temos que inércia total pode ser obtida através da soma dos quadrados dos valores singulares (diagonal de D_μ) de A.

$$I = traço[D_\mu^2] = \lambda_1 + \lambda_2 + ... + \lambda_\alpha$$

Cada valor singular está associado a um eixo fatorial, de maneira que o primeiro valor singular (λ_1) ao quadrado se refere à parcela da inércia total captada pelo primeiro eixo principal de inércia. Assim a parcela da inércia total captada pelo eixo α é λ_α. Diante disto determina-se a participação relativa de cada eixo fatorial, que é dado por:

$$\tau_\alpha = \frac{\lambda_\alpha}{I}$$

Para ilustrar, a participação relativa do primeiro eixo fatorial no exemplo das rádios é

$$\tau_{\alpha} = \frac{(0,266)^2}{0,0817} = 0,867$$

A participação do eixo fatorial α na inércia do ponto *i* é definida como $f_i(F_{\alpha}(i))^2$, sendo f_i a massa de *i* e $F_{\alpha}(i)$ a coordenada do ponto *i* no eixo α. Podemos então calcular a contribuição relativa desse mesmo eixo pela razão:

$$COR_{\alpha}(i) = \frac{f_i(F_{\alpha}(i))^2}{I_{f_J}(f_J^i, f_i)}$$

Sendo $I_{f_J}(f_J^i, f_i)$ a inércia do ponto *i* em relação ao centro de gravidade. Esta medida nos permite conhecer aqueles fatores que explicam a posição de um ponto em relação ao centro da nuvem.

Para o exemplo apresentado, o valor da COR para a classe social A na primeira dimensão é

$$COR_1(1) = \frac{0,03136968}{0,03459918} = 0,906$$

Ou seja, isso significa que 90,6% da inércia desse ponto de linha é explicado pela primeira dimensão.

Além da COR, pode-se determinar também QLT, que é a qualidade da representação de um ponto em um subespaço, calculada pela soma das contribuições relativas dos eixos considerados à inércia do ponto. Quanto mais próximo de 1, melhor será a representação do ponto *i*.

Em nosso exemplo, a qualidade da representação será máxima, ou seja, 1, pois existem apenas duas soluções nesse sistema.

Definimos também a contribuição relativa de cada elemento *i* para a inércia do eixo α, dada por CTR.

$$CTR_{\alpha}(i) = \frac{f_i (F_{\alpha}(i))^2}{\lambda_{\alpha}}$$

O numerador dessa divisão nada mais é do que a contribuição absoluta do elemento *i* para λ_{α}. Para o caso do primeiro perfil para a inércia da primeira dimensão, CTR corresponde à 0,443.

A contribuição relativa do elemento *i* para a inércia total é dada por

$$INR(i) = \frac{f_i d^2(f_J, f^i{}_J)}{in(I)}$$

Considerando o perfil $i = 1$ do exemplo,

$$INR(i=1) = \frac{0,261[-0,346^2 + 0,111^2 + 0]}{0,08176792} = 0,423, \text{ ou seja,}$$

42,3% da inércia dos pontos de linha são explicados pelo perfil 1.

7 RESULTADOS E ANÁLISE

Utilizamos dois métodos para categorizar os estados e regiões quanto aos indicadores. O primeiro envolve a ideia dos percentis e o quadro abaixo mostra como se deu tal categorização.

Quadro 7.1 – Categorização dos indicadores pelo método dos percentis	
Categoria 1	Valores menores ou iguais ao percentil 20
Categoria 2	Valores maiores que o percentil 20 e menores que ou iguais ao percentil 40
Categoria 3	Valores maiores que o percentil 40 e menores que ou iguais ao percentil 60
Categoria 4	Valores maiores que o percentil 60 e menores que ou iguais ao percentil 80
Categoria 5	Valores maiores que o percentil 80

A segunda forma de categorização se deu pelo método das k-Médias, onde usamos como valores iniciais os percentis 10, 30, 50, 70 e 90 de cada indicador para classificá-los quanto aos estados e regiões. Nesse método calculamos apenas uma matriz de percentis, não fazendo distinção de estados e regiões.

Preferimos apresentar os resultados da AC pelo segundo método, visto que conseguimos uma maior explicação da inércia pela dimensão 1 do que comparado ao método de classificação por percentis, apesar das inércias e coordenadas serem muito próximas.

7.1 Estatísticas Descritivas

Calculamos algumas estatísticas descritivas como média, mediana, coeficiente de variação (cv) e valores de mínimo e máximo para todos indicadores referentes apenas aos estados brasileiros.

Observamos que população e área possuem uma alta dispersão, fazendo com que seus respectivos coeficientes de variação ultrapassem 1-aliás, esses indicadores possuem média maior que mediana, caracterizando uma assimetria positiva. Esperança de vida ao nascer e aos 60 anos possuem as menores variações em torno da média assumindo os valores 0,032 e 0,053 respectivamente.

Essas estatísticas ficam mais fáceis de serem entendidas visualizando os gráficos box-plots de cada indicador. Esses gráficos estão disponíveis no Anexo A.

Tabela 7.1 – Estatísticas descritivas referente aos estados brasileiros

Indicadores	Média	Mediana	CV	Mínimo	Máximo
População (pop)	7.022.696,81	3453648	1,19	412783	41011635
Área	315.230,63	224118	1,19	5801,9	1570946,8
Esperança de vida ao nascer (Espn)	72,08	71,82	0,032	67,24	75,82
Esperança de vida aos 60 anos (Esp.60)	20,80	20,94	0,053	18,3	22,58
Índice de Envelhecimento (Envl)	31,32	32,1	0,367	12,1	56,1
Razão de Dependência de Idosos (Dpds)	13,5	14,1	0,232	7,4	19,4
PIB *per capita*	11.863,06	9458,86	0,626	4611,34	41061,89
Proporção da população com mais de 8 anos de estudo (EF)	52,33	51,46	0,171	37,07	72,53
Proporção de analfabetos em maiores de 15 anos (Anlf)	12,16	9,31	0,555	4,02	25,74
Taxa de desemprego (Dsmp)	6,87	6,22	0,319	4,05	13,67
Trabalho Infantil (Trb.inf)	10,47	11,1	0,387	2,36	17,36
Mortalidade Infantil (Mrt.inf)	21,59	20,85	0,343	11,09	41,16
Proporção da população com renda inferior a ½ salário mínimo (SM)	36,3	39,36	0,392	12,72	59,48
Índice Gini	0,53	0,52	0,070	0,459	0,62
IDH	0,73	0,73	0,078	0,63	0,84

7.2 Resultados e Análise da AC

Como já foi dito, a proposta desse trabalho é averiguar similaridades dos estados e regiões em relação a alguns indicadores demográficos, socioeconômicos e de mortalidade.

Para efeito de análise, consideramos os indicadores de população e área como pontos suplementares de coluna, assim como as 5 regiões brasileiras como pontos suplementares de linha. Isso se deu pelo fato de tanto as regiões quanto a área e população serem informações adicionais aos nossos dados.

Nessa solução de AC teremos os estados e regiões associados (próximos) aos maiores níveis de classificação dado pelo método das k-Médias para cada indicador. AC foi implementada através da função *ca*, da biblioteca *ca* no software livre R (NENADIC e GREENACRE, 2007).

Temos que lembrar que esses dados não são de uma tabela de contingência, onde as células representam freqüências observadas para cada categoria de uma determinada variável e sim de uma tabela cruzada.

Quanto ao número de dimensões, essa solução possui $\min(I-1, J-1)$ que no caso são 12, pois consideramos 2 elementos de coluna como suplementares entre os 15 existentes.

Quanto à solução, a dimensão ou fator 1 é responsável por explicar 63,8% de toda variância entre os indicadores e os estados e regiões, ao passo que a segunda dimensão explica 16,9%. Como esses dois eixos são responsáveis por 80,7% de toda a variância, não iremos analisar os demais, visto que não agregam informações significativas para a compreensão do resultado.

A seguir encontra-se o gráfico da ACS desse estudo para as dimensões 1 e 2. A partir dele observamos que a primeira dimensão separa as regiões Sul, Sudeste e Centro-Oeste no lado direito do gráfico das regiões Norte e Nordeste – lado esquerdo do gráfico.

A dimensão 1 conseguiu discriminar os estados em relação às suas respectivas regiões, com exceção do estado do Amapá (AP) que ficou localizado no lado contrário à sua região, junto ao Centro-Oeste, mas com uma inércia baixa nessa primeira dimensão.

Quanto à dimensão 2, os estados de Alagoas (AL) e Sergipe (SE) figuram no quadrante referente à região Norte e o estado do Tocantins (TO), no quadrante da região Nordeste com uma inércia baixa em relação a esse fator. Mato Grosso do Sul (MS) figura junto ao quadrante dos estados do Sul e Sudeste, também com uma inércia baixa nesse eixo.

Os demais estados estão dispostos no mesmo lado do gráfico ou até mesmo no mesmo quadrante de suas regiões, como ocorreu com os estados do Sul e Sudeste, localizados no quadrante inferior à direita.

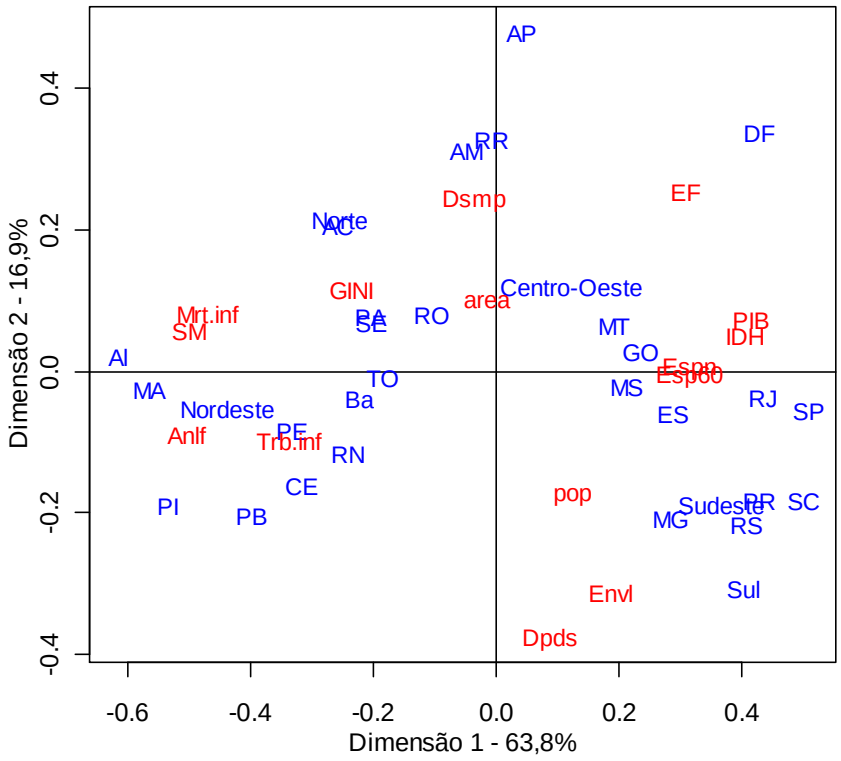

Figura 7.1 – Análise de Correspondência em indicadores socioeconômicos, demográficos e de mortalidade

A partir do gráfico acima, principal saída da AC, observamos que os estados do Sul, Sudeste e Centro-Oeste estão associados aos maiores valores de PIB, IDH, Taxa de Dependência de Idosos, Taxa de Envelhecimento, Proporção da população com mais de 8 anos de estudo (EF), Esperança de Vida ao Nascer e aos 60 anos ao passo que os estados do Norte e Nordeste associam-se aos maiores níveis dos indicadores de Mortalidade e Trabalho Infantil, Analfabetismo, Índice Gini - que mensura desigualdade, Proporção de pessoas com renda inferior a ½ salário mínimo, desemprego e área.

Como o objetivo da AC é reduzir pontos vetoriais localizados em espaços de muitas dimensões a um espaço de dimensão inferior, torna-se importante conhecer a contribuição de cada dimensão para a inércia total.

Dessa forma temos a participação relativa de cada dimensão na inércia de um ponto *i* (*j*) conhecida como COR. Ela permite conhecer as dimensões que explicam a posição do ponto relativo ao centro da nuvem.

A tabela 4 do anexo B mostra os pontos de linha que são melhores representados pela dimensão 1. Fica mais fácil entender a contribuição da dimensão para a inércia do ponto quando olhamos no gráfico a localização dos mesmos pois estão dispostos nas extremidades desta dimensão 1. Assim, os estados/regiões com maiores valores associados a COR (1) são: AL (0,969), MA (0,961), SP (0,888), Nordeste (0,867) e SC (0,833).

A mesma ideia é válida para os pontos de coluna (indicadores), onde os maiores valores estão associados a Analfabetismo (0,927), Percentual de população com renda domiciliar mensal *per capita* de até ½ salário mínimo (0,924), Mortalidade Infantil (0,916), IDH (0,873) e PIB (0,856).

Quanto à dimensão 2, os pontos de linha com as melhores representações são, em ordem decrescente, AP (0,738), AM (0,666), RR (0,614), DF (0,329) e MG (0,304). Já os pontos de coluna que são melhores explicados são a Taxa de Dependência de Idosos (0,757),

Taxa de Envelhecimento (0,651), Desemprego (0,387), Proporção da população com mais de 8 anos de estudo (0,338) e população (0,194).

Se pensarmos de forma inversa, os menores valores de inércia do ponto em relação aos fator 1 (COR 1) terão uma tendência de serem assumidos por aqueles pontos que obtiveram maiores valores na outra dimensão, como por exemplo, RR (0,000), AP (0,006), AM (0,017), Centro-Oeste (0,073) e RO (0,115).Quanto aos pontos de coluna, os menores valores de COR (1) estão associados a área, desemprego, dependência de idosos, população e taxa de envelhecimento.

Em relação aos menores valores de COR (2) – agora na segunda dimensão, temos TO, AL, MA, RJ e MS. Ainda sobre os menores valores de COR nessa dimensão, quanto aos pontos de coluna, são esperança de vida ao nascer, esperança de vida aos 60 anos, Proporção com menos de ½ salário mínimo, IDH e analfabetismo.

Uma análise descritiva desses resultados afirma que Roraima, Amapá e Amazonas que estão associados aos menores valores de COR na dimensão 1, possuem os maiores valores de COR na dimensão 2**.** O raciocínio é análogo para os pontos de coluna.

A qualidade da representação de um ponto em um subespaço (QLT) é obtida pela soma das contribuições relativas dos eixos considerados à inércia do ponto. Assim, quanto mais próximo de 1 estiver, melhor será a representação do ponto. A representação de um ponto de linha é dado por

$$QLT = COR_1(i) + COR_2(i)$$

Daí, podemos notar que RJ, AL, PI, SC , PE e MA compõem os estados com os menores valores de QLT. Podemos pensar de forma similar para os pontos de coluna.

Podemos pensar também na contribuição relativa do ponto para a inércia da dimensão (CTR).

É interessante lembrar que, como estamos considerando pontos suplementares na solução, estes não contribuem para a inércia da dimensão pois não participam da solução (massa igual a zero).

Nas dimensões 1 e 2, as maiores inércias foram assumidas por AL e AP, e as menores contribuições, por AP e RR no fator 1, e por MA, MS e TO, na segunda dimensão.

Podemos observar que o perfil de linha com maior contribuição para a inércia da dimensão 2 é o Amapá (AP) e, ao mesmo tempo, esse perfil possui junto com Roraima (RR) a menor contribuição para a inércia da dimensão 1, e isso fica bastante visível quando olhamos o gráfico e vemos esses pontos muito próximos do primeiro eixo e um tanto distante do eixo dois.

Para os pontos de coluna é possível realizar a mesma interpretação.

Por último, analisaremos a contribuição relativa do elemento *i* para a inércia total do sistema (*INR*), onde a maior contribuição é dada pelo DF que encontra-se distante tanto do eixo 1 quanto do 2 e a menor, por SE e MS . O raciocínio é análogo para os perfis de coluna.

7.3 Codificação *Doubling*

Depois de categorizarmos todos os indicadores do estudo quanto aos estados segundo o método de agrupamento das k-Médias, realizamos uma nova recodificação desses indicadores denominada *doubling* (GREENACRE, 2010), uma das diferentes formas de recodificação que podem ser utilizadas em AC.

Em classificações, rankings e comparações pareadas, cada variável gera duas variáveis recodificadas que podem ser pensadas como pólos positivo e negativo (GREENACRE, 2010).

Por exemplo, o estado de Minas Gerais (MG) foi classificado na categoria 4 do indicador IDH numa escala de 1 a 5, onde 1 concentra os estados com os menores valores do indicador, e 5, os estados associados aos maiores valores de tal indicador. Podemos recodificar esse valor em 3 e 1 respectivamente, onde quatro representa 3 pontos de classificação acima da origem da escala e 1 ponto de classificação abaixo do extremo da mesma escala de classificação.

A ideia é simplesmente medir a distância do ponto de classificação aos extremos da escala, obtendo assim dois pólos (positivo e negativo). Essa correção faria com que os totais marginais ficassem iguais a fim de fazer sentido a comparação de perfis.

Em casos em que temos todas variáveis crescendo em um mesmo sentido numa escala de 1 a 5, por exemplo, faria-se necessário o uso de alguma codificação pois poderíamos nos deparar com dois perfis que essencialmente são diferentes mas que para a Análise de Correspondência seriam idênticos.

O método de codificação *doubling* impede que isso ocorra fazendo uma correção na matriz de dados, deixando os totais marginais iguais a fim de fazer sentido a comparação desses perfis.

Esse método se resume à criação de dois extremos da escala que são complementares.

7.4 Análise dos resultados via codificação *doubling*

Aplicamos o método de codificação proposto acima e obtivemos a nova matriz de dimensão $32x30$ e os seguintes resultados acerca da solução de Análise de Correspondência. Agora podemos comparar perfis de linha sem nos preocupar com algum tipo de erro metodológico, já que todos os objetos de nossa aplicação terão a mesma massa.

Em relação à inércia explicada pelas dimensões ou fatores, o eixo 1 é responsável por concentrar 61,2% de toda a variabilidade dos dados e o segundo eixo, 17,5%. A soma acumulada dos três primeiros fatores explicam 87,6% de toda inércia.

Assim como foi feito na análise 1 de AC, consideramos pontos suplementares de linha as regiões brasileiras e quanto aos pontos de coluna (indicadores), consideramos novamente a área e a população, agora com seus respectivos pólos positivo e negativo.

Como já foi dito no capítulo 6, variáveis suplementares possuem massas iguais a zero, ou seja, apenas figuram no gráfico. Daí a AC, através do método de decomposição em valores singulares, trabalhou com a matriz de dimensão $27x26$ e determinaria 25 dimensões, porém, como há 13 dependências lineares devido aos pólos positivo e negativo somarem o mesmo valor (4) para cada um dos indicadores, o resultado encontrado foi 13 valores singulares, cada um associado a uma dimensão.

A seguir encontra-se o gráfico dessa solução, onde observamos que os pólos de cada indicador estão dispostos em sentidos opostos e se cruzam no centróide. Veja a ilustração para os indicadores Taxa de Envelhecimento (Envl) e Taxa de Analfabetismo (Anlf). Isso se dá devido a dependência linear existente entre os pontos de coluna da matriz.

É importante comentar que foi necessário invertermos os labels dos indicadores para que assim, associássemos o pólo positivo, denotado pelo símbolo $+$ aos maiores níveis de classificação e o pólo negativo $(-)$, aos menores níveis.

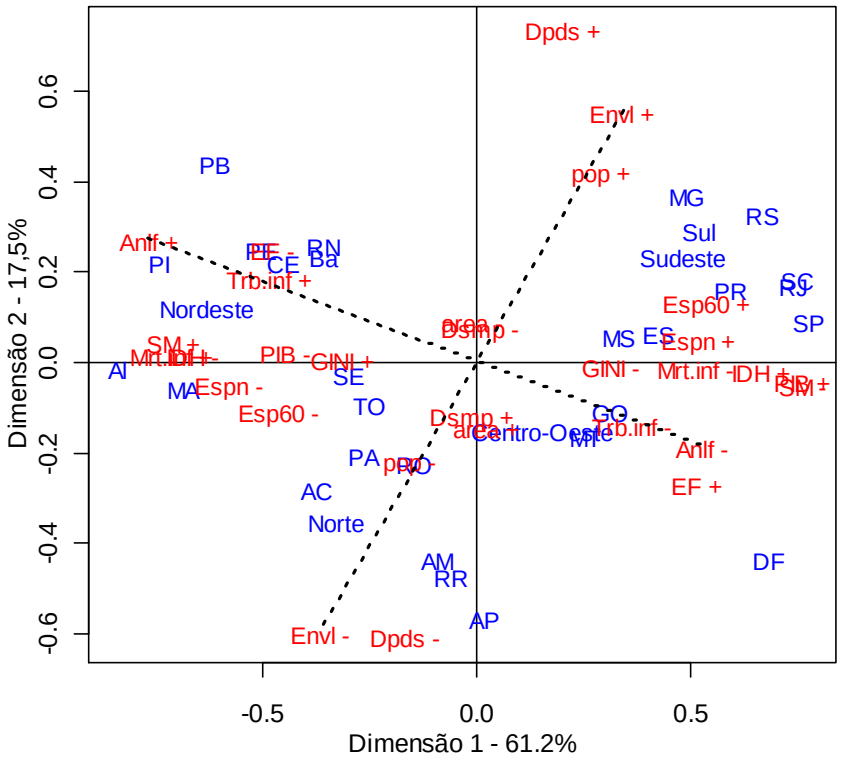

Figura 7.2 – Análise de Correspondência utilizando codificação *doubling* em indicadores socioeconômicos, demográficos e de mortalidade

Observamos que o gráfico acima apresenta a mesma informação daquele construído sem a aplicação da codificação *doubling*.

Os pólos positivos dos indicadores IDH, EF, Esp.60, Espn associam-se aos estados do Sul, Sudeste Centro-Oeste, ao passo que os estados do Norte e Nordeste estão plotados próximos aos pólos negativos desses indicadores e aos positivos dos indicadores de Mortalidade Infantil (Mrt.inf), Trabalho Infantil (Trb.inf), Índice Gini, Analfabetismo em maiores de 15 anos (Anlf) e porcentagem da população com menos de meio salário mínimo (SM).

Em relação as análises das contribuições de COR, QLT, CTR e INR, observamos que os resultados da AC também são muito próximos aos obtidos sem a codificação *doubling*.

Isso indica que tanto faz usar ou não essa correção nesses dados, pois como já havíamos comentado antes, temos variáveis crescendo em duas direções, fazendo com que seja improvável classificarmos dois perfis como idênticos ou parecidos sabendo que são diferentes por natureza.

8 CONCLUSÃO

A partir desse trabalho observamos que a Análise de Correspondência Simples pode ser utilizada tanto para tabelas de contingência quanto para tabelas cruzadas, como foi o caso.

Utilizamos duas técnicas a fim de categorizar todos os 15 indicadores em 5 categorias e preferimos adotar àquele que utilizou-se do método das k-Médias ao método dos percentis pois obtivemos uma maior explicação dos eixos 1 e 2 ao aplicar a técnica de Análise de Correspondência.

Após essa etapa, aplicamos AC usando duas metodologias. A primeira consiste na simples obtenção dos resultados a partir da matriz fornecida pelo método das k-Médias, e a segunda decorre de uma correção metodológica na matriz dos dados conhecida como *doubling*, de tal forma que faça sentido a comparação de perfis.

Quanto à comparação desses dois resultados para AC, estes foram similares, o que nos faz concluir que, para esse caso particular, os dois métodos são eficazes.

Em relação à análise gráfica da AC, observamos as proximidades, similaridades ou dissimilaridades entre os elementos estudados e concluímos nos dois gráficos que as regiões Norte e Nordeste, incluindo seus respectivos estados, estão associados aos indicadores de mortalidade e trabalho infantil, desemprego, índice Gini, analfabetismo e percentual da população com renda inferior a ½ salário mínimo, caracterizando assim, estados com maiores necessidades de políticas públicas do governo federal com foco no desenvolvimento, geração de renda, educação e saúde.

Já os estados do Sul e Sudeste, junto com os do Centro-Oeste, estão associados a indicadores que suscitam a ideia de que sua população possui comparativamente boa

qualidade de vida – esperança de vida ao nascer e aos 60 anos, IDH, PIB, porcentagem da população com 8 ou mais anos de estudo, taxa de envelhecimento e dependência de idosos.

Assim, fica evidenciado nesse estudo a disparidade dos estados do Sul, Sudeste e Centro-Oeste em relação aos estados do Norte e Nordeste.

É claro concluirmos também o poder da Análise de Correspondência em detectar e discriminar os estados e regiões em relação aos indicadores ou os indicadores em relação aos estados e regiões. A regionalização do país, portanto, reflete a ação múltipla dos indicadores utilizados nesse estudo, dentre outros fatores.

ANEXO A – Box-plots para os 15 indicadores referentes aos 27 estados brasileiros

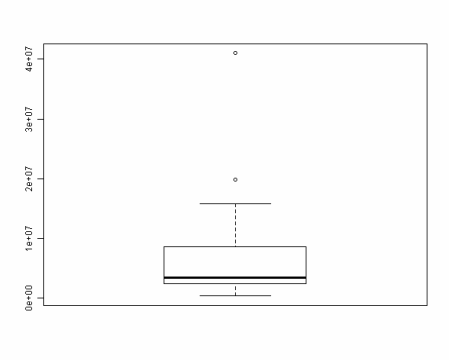

Figura A.1 – Box-plot do indicador de população

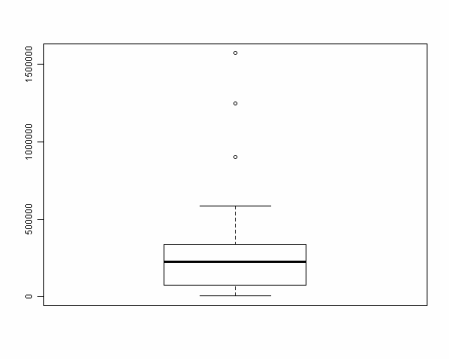

Figura A.2 – Box-plot do indicador de área

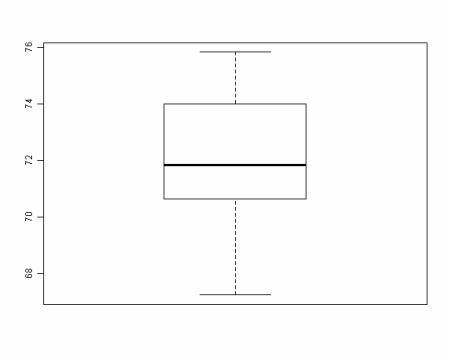

Figura A.3 – Box-plot do indicador de Esperança de vida ao nascer

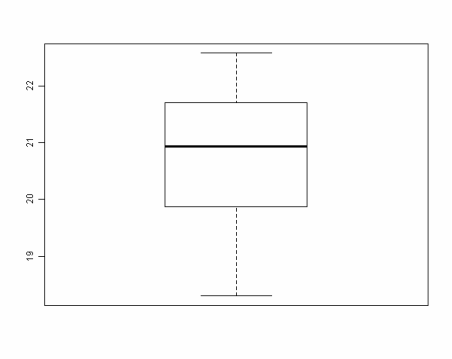

Figura A.4 – Box-plot do indicador de Esperança de vida aos 60 anos

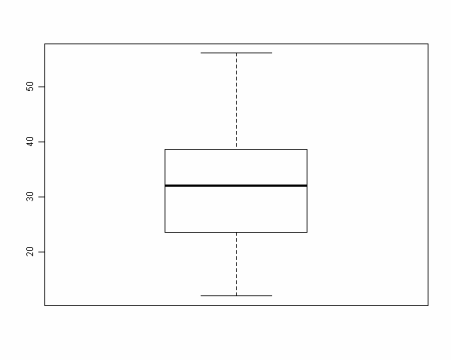

Figura A.5 – Box-plot do indicador de Taxa de Envelhecimento

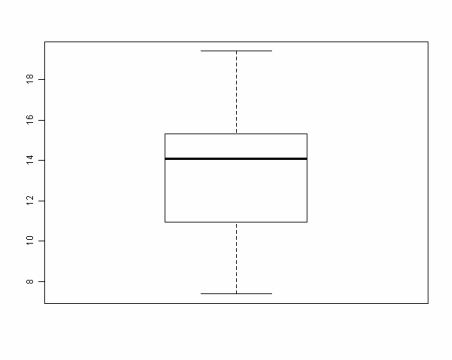

Figura A.6 – Box-plot do indicador de Taxa de Dependência de Idosos

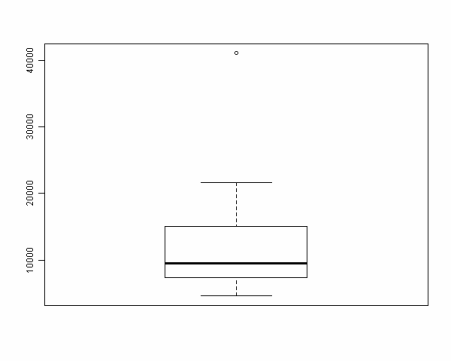

Figura A.7 – Box-plot do indicador de PIB

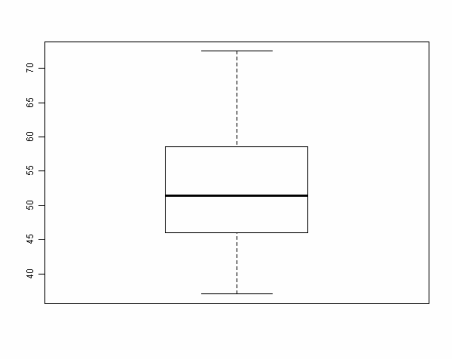

Figura A.8 – Box-plot do indicador do Percentual de pessoas de 15 anos e mais com mais de 8 anos de estudo

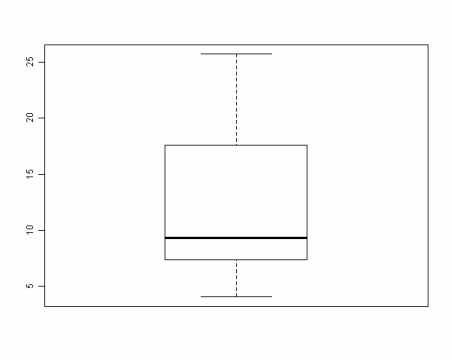

Figura A.9 – Box –plot do indicador de Analfabetismo em pessoas de 15 anos e mais

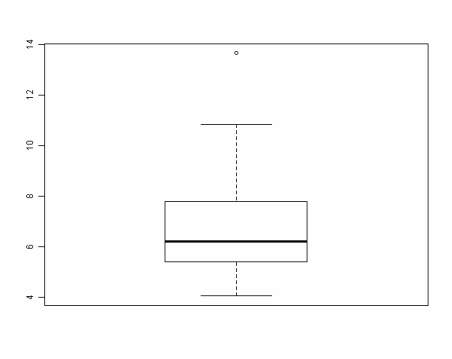

Figura A.10 – Box-plot do indicador de Taxa de Desemprego

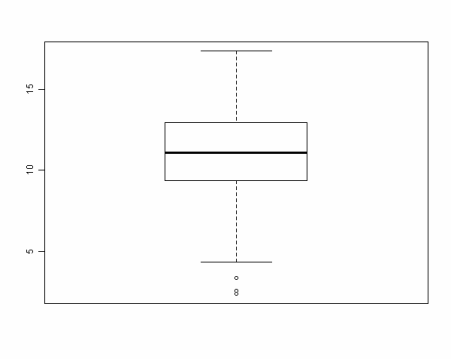

Figura A.10 – Box-plot do indicador de Trabalho Infantil

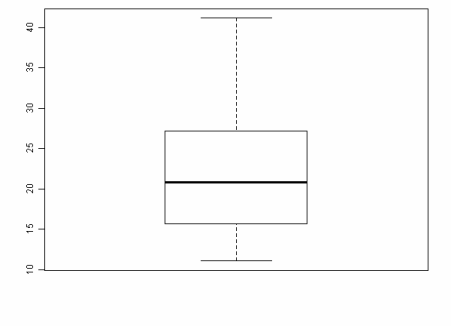

Figura A.12 – Box-plot do indicador de Mortalidade Infantil

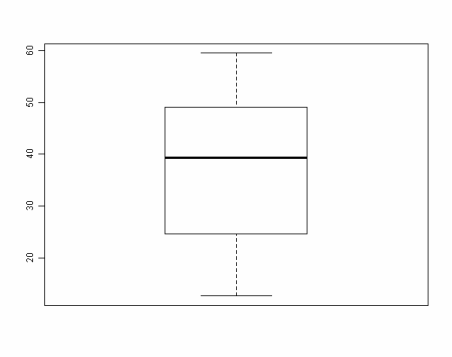

Figura A.13 – Box-plot do indicador da proporção da população com renda inferior a ½ salário mínimo

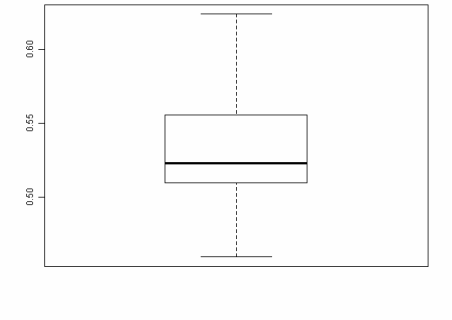

Figura A.14 – Box-plot do indicador de Índice de Gini

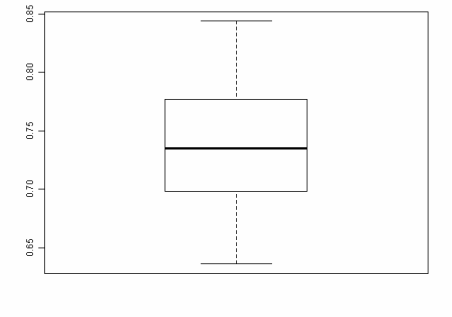

Figura A.15 – Box-plot do indicador de IDH

ANEXO B

Tabelas referentes às inércias dos pontos de linha e coluna

Tabela da decomposição da inércia total dos pontos de linhas (valores multiplicados por 1000)

Estados/Regiões	QLT	INR	COR [1]	CTR [1]	COR [2]	CTR[2]
AC	837	32	403	20	265	51
AL	972	73	969	111	1	1
AM	686	25	17	1	666	99
AP	927	54	6	0	738	234
BA	689	19	640	19	16	2
CE	872	28	696	31	176	30
DF	855	76	526	63	329	149
ES	830	21	790	26	30	4
GO	834	15	704	16	13	1
MA	964	54	961	81	2	0
MG	877	31	573	28	304	57
MS	830	12	805	15	6	0
MT	844	16	433	11	50	5
PA	625	18	383	11	59	6
PB	943	53	649	54	166	52
PE	964	37	644	37	38	8
PI	970	67	823	87	101	40
PR	892	46	757	55	131	36
RJ	978	54	799	67	5	2
RN	944	20	640	20	138	16
RO	701	16	115	3	69	7
RR	629	26	0	0	614	93
RS	951	49	735	57	205	60
SC	967	61	833	80	105	38
SE	789	12	698	14	90	7

SP	940	60	888	84	10	3
TO	787	23	266	10	0	0
Centro-Oeste	159	-	73	-	81	-
Nordeste	894	-	867	-	11	-
Norte	612	-	320	-	224	-
Sudeste	558	-	339	-	89	-
Sul	542	-	340	-	202	-

Tabela da decomposição da inércia total dos pontos de coluna (valores multiplicados por 1000)

Indicadores	QLT	INR	COR[1]	CTR	COR[2]	CTR [2]
População	342	-	96	-	194	-
Área	231	-	1	-	40	-
Esperança ao nascer	947	58	756	69	0	0
Esperança aos 60 anos	733	56	713	63	0	0
Índice de Envelhecimento	942	65	237	24	651	250
Razão de Dependência de Idosos	954	76	42	5	757	339
PIB	901	72	856	97	28	12
Proporção da população de 15 anos e mais com mais de 8 anos de estudo	838	82	493	63	338	164
Analfabetismo	954	104	927	151	27	16
Taxa de desemprego	871	56	9	1	387	128
Trabalho infantil	921	82	540	70	44	21
Mortalidade Infantil	946	93	916	134	29	16
Percentual da população com renda domiciliar inferior a ½ salário mínino	938	121	924	175	14	10
Índice de Gini	592	54	453	38	112	36
IDH	903	81	873	111	16	8

ANEXO C – Informação sobre a origem dos dados

Fonte e base de dados		
Indicadores	Fonte/Ano	Base de dados
População	IBGE/PNAD/2007	http://tabnet.datasus.gov.br/cgi/idb2009/matriz.htm
Área	IBGE/2007	http://tabnet.datasus.gov.br/cgi/idb2009/matriz.htm
Esperança ao nascer	IBGE/Projeções demográficas preliminares/2008	http://tabnet.datasus.gov.br/cgi/idb2009/matriz.htm
Esperança aos 60 anos	IBGE/2008	http://tabnet.datasus.gov.br/cgi/idb2009/matriz.htm
Índice de Envelhecimento	IBGE/Pesquisa projeções e estimativas demográficas (2001-2008)	http://tabnet.datasus.gov.br/cgi/idb2009/matriz.htm
Razão de Dependência de Idosos	IBGE/2008	http://tabnet.datasus.gov.br/cgi/idb2009/matriz.htm
PIB	IBGE/2007	http://tabnet.datasus.gov.br/cgi/idb2009/matriz.htm
Proporção da população de 15 anos e mais com mais de 8 anos de estudo	IBGE/2008	http://tabnet.datasus.gov.br/cgi/idb2009/matriz.htm
Taxa de Analfabetismo	PNAD/2008	http://tabnet.datasus.gov.br/cgi/idb2009/matriz.htm
Taxa de desemprego	IBGE/Pesquisa Nacional por Amostra de Domicílios – PNAD/2008	http://tabnet.datasus.gov.br/cgi/idb2009/matriz.htm
Trabalho infantil	IBGE/Pesquisa Nacional por Amostra de Domicílios – PNAD/2008	http://tabnet.datasus.gov.br/cgi/idb2009/matriz.htm
Mortalidade Infantil	MS/SVS - Sistema de Informações sobre Mortalidade – SIM (2008)	http://tabnet.datasus.gov.br/cgi/idb2009/matriz.htm
Percentual da população com renda domiciliar inferior a ½ salário mínino	IBGE/Pesquisa Nacional por Amostra de Domicílios – PNAD/2008	http://tabnet.datasus.gov.br/cgi/idb2009/matriz.htm
Índice de Gini	IPEA/2009	www.ipea.gov.br
IDH	IPEA/2000	www.ipea.gov.br

REFERÊNCIAS BIBLIOGRÁFICAS

[1] ANTON,H e Busby, R. (2006). Álgebra Linear Contemporânea. Porto Alegre: Bookman.

[2] BEH, E.J Simple Correspondence analysis: a bibliographic review. International Statistical. Review, v.72,n.2, p 257-284.

[3] BENZÉCRI, J.P. (1992). Correspondence Analysis Handbook. New York: Marcel Dekker.

[4] BERRINGTON, A. (2002). Exploring relationships between entry into parenthood and gender role attitudes: evidence from the British Household Panel Study . In Lesthaeghe R. ed. Meaning and Choice: Value Orientations and Life Course Decisions. Brussels: NIDI.

[5] DATASUS. Disponível em < www.datasus.gov.br >. Acesso em setembro de 2010.

[6] DEPARTAMENTO DE ESTATÍSTICA. Pesquisa de Opinião-Contrato de Prestação de Serviços,2000.

[7] GREENACRE,M. (2007). Correspondence Analysis in Practice, second edition: Boca Raton: Chapman & Hall/CRC.

[8] GREENACRE,M. (2010). Correspondence Analysis. Focus Article. *Computational Statistics*, vol. 2, september/october 2010, 613-619.

[9] GOUVÊA, V.H. (1990). ANÁLISES DE CORESPONDÊNCIAS. Notas de Aula. ENCE/IBGE.

[10] IPEADATA. Disponível em: < www.ipeadata.gov.br > . Acesso em setembro de 2010.

[11] MINGOTI, S. (2005). Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: Editora UFMG.

[12] NENADIC, O. e GREENACRE, M. (2007). Corerspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package. Journal of Statistical Software, vol. 20. issue 3. http://www.jstatsoft.org/

[13] R Development Core Team. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, 2009. Disponível em http://www.R-project.org.