ESTUDO E APLICAÇÕES DO MÉTODO DE AGRUPAMENTO BASEADO NO MODELO

PATRÍCIA BATISTA LEAL

Orientadora: Prof.ª Dra. Viviana Giampaoli
Área de concentração: Estatística Aplicada

Dissertação submetida como requerimento parcial para obtenção do grau de Mestre em Estatística pela Universidade Federal de Pernambuco

Recife, 2004

Universidade Federal de Pernambuco
Mestrado em Estatística

27 de Fevereiro de 2004
(data)

Nós recomendamos que a dissertação de mestrado de autoria de

Patrícia Batista Leal

Intitulada

Estudo e aplicações do método de agrupamento Clusterwise em modelos

seja aceita como cumprimento parcial dos requerimentos para o grau de Mestre em Estatística.

Coordenador da Pós-Graduação em Estatística

Prof. Francisco Cribari Neto
Coordenador Substituto
em Exercício - PPGE

Banca Examinadora:

Vivian Cristina Gonçalves
orientador

Maria Cristina Falcão Raposo

Ana Katarina Campêlo

Este documento será anexado à versão final da dissertação.

Aos meus pais, Ivonete e Antônio.
As minhas irmãs, Rosana e Cristiane.

Agradecimentos

A Deus, por me conceder a graça de concluir mais uma etapa da minha vida.

A minha família que sempre me incentivou, em especial a minha mãe.

Agradeço, de forma especial, a minha orientadora Viviana Giampaoli pela segurança, disponibilidade, paciência e incentivo.

A professora Cristina, por estar sempre disposta a me ajudar, por sua alegria.

A Lindomberg, por seu amor, por acreditar em mim, pelo seu incentivo e compreensão.

A Gilmar, meu amigo, meu irmão, pela amizade incondicional.

A Michelli, minha querida amiga, pela amizade, pelo apoio, por saber que posso contar com ela sempre.

A Horácio, pela amizade e pelas vezes que sempre se dispôs a me ajudar quando precisei.

A Silvia, minha irmã de coração, pela amizade, pelo companheirismo e pelos momentos felizes que dividimos.

A Gilson, por sua amizade, por saber transmitir alegria e calma.

A Moisés, que nos meus momentos de dificuldade sempre aparecia disposto a me ajudar.

A minha turma de mestrado, que se tornou uma família: Sílvia, Gilson, Moisés, Bartolomeu, Felipe, Tatiene, Tarciana, Raydonal, João Marcelo e Keila. Juntos dividimos nossas saudades, angustias, apreensões, nosso sucesso, nossas experiências e principalmente nossas alegria.

Aos meus colegas de mestrado Michelli, Diana, Carla, Patrícia Leone, Amanda, Heráclito, Sandra Pinheiro, Tatiane, Andreia, Sandra Rêgo, Lenaldo, Cherubino, Fernando, Sílvia, Gecy, André, Júnior e Cristina Morais.

Aos professores do Programa de Mestrado em Estatística, pela competência e atenção.

A Valéria Bittencourt, pela competência, organização, atenção e amizade.

Aos funcionários do departamento de estatística, Adriana, Cícero e Antônio, pela atenção.

Aos colegas de graduação, Lindomberg, Michelli, Horácio, Diana e Juliana.

Aos professores do Departamento de Matemática e Estatística da Universidade Federal do Estado da Paraíba, em especial aos professores Francisco Morais, Luiz Mendes e Vandik Estevam.

A Michelli, Horácio, Diana, Lenaldo, Renata, Avelita e Kátia, pela convivência pacífica e harmoniosa.

A CAPES, pelo apoio financeiro.

Resumo

Análise de agrupamento é um termo genérico que abrange vários métodos numéricos para estudar dados multivariados com o intuito de classificar as observações em grupos homogêneos.

Neste trabalho se apresenta uma síntese dos métodos de agrupamento existentes e se realiza uma descrição detalhada de um método particular de agrupamento hierárquico aglomerativo baseado em modelos que considera um critério bayesiano para a determinação do melhor modelo e dos possíveis grupos existentes.

Duas aplicações desta metodologia foram realizadas. A primeira utilizando o banco de dados do Instituto de Pesquisa Econômica e Aplicada que contém informações sobre finanças públicas. A segunda com o banco de dados do Sistema Único de Saúde relacionadas a área de saúde.Na realização das duas foram utilizados os municípios de Pernambuco visando uma classificação dos mesmos verificando os grupos mais homogêneos, os municípios com valores extremos para cada variável. É apresentada também uma discussão dos resultados obtidos.

Abstract

Cluster analysis are a generic term for a variety of numerical methods to study multivariate data with the purpose of classifying in especific groups homogeneous observations. In this work we present the synthesis of the existing methodologies. A particular method is detailed: model-based clustering for the determination of the best model and the possible existing groups is utilized an Bayesian criterion. Two different applications of this methodology are carried with data of the Instituto de Pesquisa Econômica Aplicada and of Sistema Único de Saúde. The first database contains information on public finances and health of the boroughs of Pernambuco.

Índice

Capítulo 1

Introdução

O estudo da classificação de objetos contidos em grandes conjuntos de dados reunidos em grupos, segundo alguma(s) característica(s) de interesse, é caracterizado como análise de agrupamento. Esta necessidade existe há muito tempo em várias áreas do conhecimento. No século XVIII, Linnaeus e Sauvages realizaram uma extensa classificação de animais, plantas, minerais, e enfermidades (ver Holman, 1985). Em ciências sociais, por exemplo, existe a necessidade de classificar as pessoas com respeito às suas preferências e seu comportamento. Em marketing, identificar segmentos do mercado, isto é, grupos de costumes com necessidades semelhantes. Mais exemplos poderiam ser dados em geografia (grupos de regiões), medicina (tipos específicos de câncer), química (classificação de componentes), história (grupos de achados arqueológicos), genética, processamento de imagens e muito mais. Como podemos perceber a análise de agrupamento é um assunto com muitas aplicações práticas e sua abordagem não se limita apenas a periódicos estatísticos mas também a periódicos nas áreas já citadas. Além disso, a análise de agrupamento pode ser usada não apenas para identificar uma estrutura já presente nos dados, mas também para impor uma estrutura mais ou menos homogênea num conjunto de dados que deve ser dividido, de maneira correta, em grupos.

O objetivo desta dissertação é trabalhar com o agrupamento hierárquico baseado em modelos, que fundamenta-se na suposição que as observações são geradas de uma mistura de probabilidades subjacentes. Foi utilizado o pacote MCLUST que foi escrito em `Fortran` e possui interface para o `S-Plus` e também para o `R`, ver Fraley e Raftery (2002a). O MCLUST é um pacote para agrupamento baseado em modelos e implementa algoritmos de agrupamento hierárquico para modelos de mistura Gaussiana. Nele são incluídas funções que combinam agrupamento hierárquico, o algoritmo EM e o Critério de Informação Bayesiano (BIC) em uma estratégia tanto para agrupamento quanto para a análise discriminante. O MCLUST pode ser obtido em um dos endereços abaixo:

`http://www.stat.washington.edu/mclust`

ou

`http://lib.stat.cmu.edu/S/mclust`.

A plataforma computacional utilizada é o pacote estatístico `R`. O programa `R` é uma versão gratuita do programa `S-PLUS` e é baseado na linguagem de programação `S`; ver Cribari-Neto & Zarkos (1999) e Venables & Ripley (2002). Uma vantagem dessa plataforma é sua flexibilidade em permitir a criação de novas funções e a possibilidade de modificações de muitas funções internas. Este programa pode ser obtido em versões `Windows`, `Linux`, `Unix` e `Macintosh` da página `http://www.r-project.org`.

O presente trabalho está organizado em mais 3 capítulos além desta introdução. No capítulo 2 é citada a classificação das variáveis e as suas correspondentes medidas de similaridade e dissimilaridade. Também é apresentada uma síntese dos métodos de agrupamento existentes. No capítulo 3 é feita uma revisão das densidades de mistura finita, o algoritmo EM e o agrupamento hierárquico baseado nos modelos. No capítulo 4 encontra-se uma aplicação deste método objetivando agrupar os municípios do estado de Pernambuco usando os bancos de dados do Instituto de Pesquisa Econômica e Aplicada (IPEA) e do Sistema Único de Saúde (SUS).

Capítulo 2

Conceitos Básicos de Análise de Agrupamento

O resultado de uma análise de agrupamento deve ser um conjunto de grupos que podem ser consistentemente descritos através de suas características, atributos e outras propriedades. Estas características são descritas pelas variáveis utilizadas no estudo. Assim, um dos fatores que mais influencia os resultados neste tipo de análise é a escolha das variáveis. Aquelas variáveis que assumem praticamente o mesmo valor para todos os objetos são pouco discriminatórias, e sua inclusão contribuiria muito pouco para a determinação da estrutura do agrupamento. Em contrapartida, a inclusão de variáveis com grande poder de discriminação porém irrelevantes ao problema pode mascarar os grupos e levar a resultados equivocados. Além disso é desejável que unidades ou objetos de estudo sejam comparáveis segundo o significado de cada uma das variáveis. Em geral, o número de variáveis medidas é grande dificultando a análise. Deve-se então procurar diminuir seu número de forma que sua seleção leve em consideração tanto a sua relevância como seu poder de discriminação face ao problema em estudo. Primeiro é necessário identificar quais são os tipos das variáveis que estão sendo avaliadas para que se possa escolher uma medida adequada para definir a distância entre os objetos. A posteriori se escolherá o método de agrupamento apropriado para cada tipo de variável. Neste capítulo apresentaremos uma classificação das variáveis, considerando-as em dois tipos: *quantitativas* e *qualitativas* e as medidas de similaridade e dissimilaridade associadas. Apresentaremos também uma síntese dos métodos de agrupamentos existentes.

Variáveis Quantitativas - São aquelas que podem ser expressas em termos numéricos. Em geral são as resultantes de medições, enumerações ou contagens. Podem ser classificadas em *contínuas* e *discretas*.

Variáveis Discretas - Quando só podem assumir determinados valores num certo intervalo, podendo ser associadas ao conjunto dos números inteiros, ou seja, seus possíveis valores formam um conjunto finito ou enumerável. Em geral, estes números inteiros são resultantes de um processo de contagem, como por exemplo: o número de bactérias, a quantidade de pacientes atendidos diariamente num hospital, etc.

Variáveis Contínuas - São aquelas que podem assumir qualquer valor num certo intervalo de medida, podendo ser associada ao conjunto dos números reais, ou seja, seus valores possíveis formam um conjunto não enumerável. Entre outras, enquadram-se nesta categoria

as medidas de tempo, comprimento, espessura, área, volume, peso e velocidade. As variáveis contínuas podem pertencer a dois tipos de escala:

a) *Intervalar* - São medidas contínuas em uma escala linear. Neste nível o zero da escala é relativo. Como exemplo temos as escalas termo métricas e o fuso horário.

b) *Razão* - Essa escala é muito parecida com a intervalar exceto pelo valor do zero visto que neste caso o zero é um valor absoluto.

Variáveis Qualitativas - Quando o resultado da observação é apresentado na forma de qualidade ou atributo. Nesses casos, as variáveis podem ser classificadas em *nominais* ou *ordinais*.

Variáveis Nominais - Quando puderem ser reunidas em categorias ou espécies com idênticos atributos. Aqui se incluem, por exemplo, os agrupamentos por sexo, área de estudo, desempenho, cor, raça, nacionalidade e religião. Podem ser de dois tipos: binomiais (ou dicotômicas) e policotômicas, a saber:

a) *Binárias* - Possuem apenas dois resultados possíveis, que podem ser codificados como 0 ou 1, e podem ser caracterizadas de duas formas: simétricas e assimétricas. No primeiro tipo as categorias possuem a mesma importância e portanto recebem o mesmo peso. Por exemplo, a variável sexo. No segundo tipo os resultados tem pesos de importância diferentes. Por convenção o código para o resultado mais importante é 1 e o 0 para o resultado menos significativo. Normalmente 1 representa a presença de algum atributo já o 0 sua ausência.

b) *Policotômicas* - Possuem mais que dois resultados possíveis, por exemplo religião. Variáveis ordinais são do tipo policotômicas.

Variáveis Ordinais - São variáveis que têm níveis ordenados, por exemplo, classe social (baixa, média, alta), grau de instrução (ensino: fundamental, médio, superior). Estas variáveis tem suas categorias ordenadas, porém a distância absoluta entre cada uma das categorias é desconhecida.

As técnicas existentes na análise de agrupamento utilizam algum tipo de critério para medir a distância entre dois objetos ou quantificar o quanto eles são parecidos. Esta distância pode ser classificada em duas categorias: medidas de *similaridade* ou *dissimilaridade*, denotadas por $s(i, j)$ e $d(i, j)$, respectivamente. Na primeira distância temos que, quanto maior o seu valor mais parecidos são os objetos. Já na segunda, quanto maior o valor observado menos parecidos são os objetos. Da medida de similaridade é possível construir uma de dissimilaridade e vice-versa. A seguir são apresentadas algumas dessas medidas.

2.1 Medidas de Similaridade e Dissimilaridade para Variáveis Quantitativas

O conjunto de dados pode ser organizado numa matriz $n \times p$, onde as linhas correspondem aos objetos e as colunas correspondem as variáveis. Quando a l-ésima medida do i-ésimo objeto é denotada por x_{il} (onde $i = 1, \ldots, n$ e $l = 1, \ldots, p$) a matriz é denotada da seguinte forma

$$\begin{array}{c} \\ \text{objetos} \end{array} \begin{array}{c} \text{variáveis} \\ \begin{bmatrix} x_{11} & \cdots & x_{1l} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{il} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nl} & \cdots & x_{np} \end{bmatrix} \end{array}.$$

Um fato importante a ser considerado, em algumas situações é que mudando a unidade de medida pode-se levar à estruturas de agrupamentos muito diferentes. Para evitar este tipo de problema deve-se padronizar os dados originais convertendo-os em variáveis adimensionais

$$z_{il} = \frac{x_{il} - m_l}{s_l},$$

onde

$$m_l = \frac{1}{n}(x_{1l} + x_{2l} + \cdots + x_{nl}),$$

e

$$s_l = \frac{1}{n}\{\mid x_{1l} - m_l \mid + \mid x_{2l} - m_l \mid + \cdots + \mid x_{nl} - m_l \mid\}.$$

A vantagem de se usar s_l no denominador é que o valor do z_{il} não será muito afetado com a existência de valores extremos.

Por construção o z_{il} tem valor médio igual a zero e seu desvio médio absoluto é igual a 1. Quando se utiliza a padronização os dados originais são desprezados e passa-se a trabalhar com uma nova matriz de dados, a saber:

$$\begin{array}{c} \\ \text{objetos} \end{array} \begin{array}{c} \text{variáveis} \\ \begin{bmatrix} z_{11} & \cdots & z_{1l} & \cdots & z_{1p} \\ \vdots & & \vdots & & \vdots \\ z_{i1} & \cdots & z_{il} & \cdots & z_{ip} \\ \vdots & & \vdots & & \vdots \\ z_{n1} & \cdots & z_{nl} & \cdots & z_{np} \end{bmatrix} \end{array}.$$

As métricas que são calculadas na análise de agrupamento satisfazem os seguintes requerimentos matemáticos de uma função de distância:

(1) $d(i,j) \geq 0$;

(2) $d(i,i) = 0$;

(3) $d(i,j) = d(j,i)$;

(4) $d(i,j) \leq d(i,h) + d(h,j)$;

onde i, j, e h são os índices indicadores dos objetos. A condição (1) diz que as distâncias são números não negativos; a (2) que a distância de um objeto à ele mesmo é zero, a (3) é a propriedade de simetria da função distância e, a condição (4) é chamada Desigualdade Triangular. Note que $d(i,j) = 0$ não implica necessariamente que $i = j$, o que pode ocorrer, por exemplo, é que dois elementos diferentes possuem a mesma medida para a variável em estudo.

2.1.1 Para Variáveis de Escala Intervalar

A medida mais conhecida para indicar a proximidade entre os objetos i e j é a *distância Euclidiana*:

$$d(i,j) = \left[\sum_{l=1}^{p} (x_{il} - x_{jl})^2 \right]^{1/2}.$$

Outra métrica bem conhecida é a *distância de Manhattan*, definida por

$$d(i,j) = \sum_{l=1}^{p} \mid x_{il} - x_{jl} \mid .$$

Uma generalização da distância Euclidiana e da métrica Manhattan é a *distância de Minkowski* dada por

$$d(i,j) = \left[\sum_{l=1}^{p} \mid x_{il} - x_{jl} \mid^q \right]^{1/q},$$

onde q é um número real igual ou maior que 1. Também é chamada métrica L_q e possui a métrica Euclidiana ($q = 2$) e Manhattan ($q = 1$) como casos especiais.

Existem situações onde se faz necessário trabalhar com a *distância Euclidiana Ponderada*, a saber

$$d(i,j) = \left[\sum_{l=1}^{p} \omega_l (x_{il} - x_{jl})^2 \right]^{1/2},$$

onde ω_l é o peso que cada variável recebe de acordo com sua importância relativa ao problema.

2.1.2 Para Variáveis de Razão Intervalar

Uma das maneiras de se calcular uma medida de dissimilaridade para este tipo de variável é usar uma transformação logaritmica, e tratá-la como sendo uma variável de escala intervalar. Notemos que isto é factível se não existem observações nulas.

Existe uma quantidade considerável de medidas de similaridade e dissimilaridade para variáveis quantitativas que surgem a partir do objetivo de moldar situações especiais de interesse do pesquisador. Em Cormack (1971), encontra-se uma revisão dessas medidas e, as principais delas são mostradas no Quadro 1.

Quadro 1. Principais medidas de similaridades e dissimilaridades para variáveis quantitativas.

Nome	Expressão
Distância de Minkowsky[1]	$d(i,j)\left[\sum_{l=1}^{p}\omega_l \mid x_{il}-x_{jl} \mid^l\right]^{1/l}$
Coeficiente de Similaridade de Cattel	$s(i,j)=\frac{2(p-\frac{2}{3})-d^2}{2(p-\frac{2}{3})+d^2}$
Coeficiente de Canberra	$d(i,j)=\frac{1}{p}\sum_{l=1}^{p}\frac{\lvert x_{il}-x_{jl}\rvert}{x_{il}+x_{jl}}$
Coeficiente de Bray-Curtis	$d(i,j)=\frac{\Sigma\lvert x_{il}-x_{jl}\rvert}{p\Sigma(x_{il}+x_{jl})}$
Coeficiente de Sokal e Sneath	$d(i,j)=\left\{\frac{1}{p}\sum\left(\frac{(x_{il}-x_{jl})}{(x_{il}+x_{jl})}\right)^2\right\}^{1/2}$

[1] os ω_k's representam as ponderações para as variáveis
p representa o número de variáveis
d a dimensão dos dados

2.2 Medidas de Similaridade e Dissimilaridade para Variáveis Qualitativas

É comum se usar variáveis qualitativas na busca de identificar objetos semelhantes, surge daí a necessidade de medidas que definam o grau de proximidade ou de afastamento entre os objetos segundo variáveis desse tipo.

Na matriz de dados os valores observados destas variáveis são codificados por zero ou um; Como já foi dito anteriormente o valor 1 representa a presença de algum atributo, enquanto 0 a sua ausência. Algumas vezes estas variáveis são tratadas como se fossem de escala intervalar, ou seja, se aplicando a distância Euclidiana ou distância de Manhattan. Outra possibilidade é calcular uma matriz de dissimilaridade (ou similaridade) de dados binários e então simplesmente aplicá-los num algoritmo de agrupamento que opera com esta matriz. Suponhamos que consideremos dois objetos i, j e

		Objeto j		
		0	1	
Objeto i	0	a	b	a+b
	1	c	d	c+d
		a+c	b+d	p

onde a e d representam o número de variáveis que assumiram o valor 0 e 1 respectivamente para os dois objetos, b é o número de variáveis onde as características estudadas se apresentam no objeto j e não se apresentam no objeto i, c é o número de variáveis onde as características estudadas se apresentam no objeto i e não se apresentam no objeto j. Obviamente p é o número total de variáveis. Quando ocorre dados faltantes troca-se o p pelo número de variáveis que são avaliadas para i e j simultaneamente. A partir desta notação definiremos as medidas de similaridade.

Medida de Coincidência Simples

Esta é uma das medidas de similaridades mais utilizadas, a saber

$$s(i,j) = \frac{a+d}{a+b+c+d},$$

onde $0 \leq s(i,j) \leq 1$. Logo, quanto mais próximo de 1 o resultado se encontrar maior similaridade entre os objetos.

Medida de Concordância Positiva

Um dos coeficientes mais usados quando se deseja medir a similaridade baseando-se apenas na presença de determinada característica e não na ausência é o coeficiente de Jaccard

$$s(i,j) = \frac{a}{a+b+c}.$$

Tentando ressaltar propriedades específicas criou-se uma série de coeficientes, que são derivados dos anteriores. Esses coeficientes, apresentados no trabalho de Romesburg (1984), estão descritos na Quadro 2.

Quadro 2. Coeficientes de semelhança para variáveis binárias.

Nome	Expressão	Intervalo de Variação
Distância Binária de Sokal	$\left(\frac{b+c}{a+b+c+d}\right)^{1/2}$	[0,1]
Rogers e Tanimoto	$\frac{a+d}{a+2(b+c)+d}$	[0,1]
Sokal e Sneath	$\frac{2(a+d)}{2(a+d)+b+c}$	[0,1]
Russel e Rao	$\frac{a}{a+b+c+d}$	[0,1]
Sorenson	$\frac{2a}{2a+b+c}$	[0,1]
Ochiai	$\frac{a}{[(a+b)(a+c)]^{1/2}}$	[0,1]
Russel e Rao	$\frac{a}{a+b+c+d}$	[0,1]
Baroni-Urbani-Buser	$\frac{a+(ad)^{1/2}}{a+b+c+(ad)^{1/2}}$	[0,1]
Haman	$\frac{(a+d)-(b+c)}{a+b+c+d}$	[-1,1]
Yule	$\frac{ad-bc}{ad+bc}$	[-1,1]

2.3 Métodos de Agrupamento

A partir da definição de distâncias entre os objetos pode-se então proceder o agrupamento dos mesmos. A análise de agrupamento pode ser realizada a partir de uma variedade de técnicas e algoritmos matemáticos. Estes algoritmos, de maneira geral, podem ser classificados em dois tipos: algoritmos de *partição* e algoritmos *hierárquicos* (Banfield e Raftery, 1993). Uma síntese destes métodos é apresentada a seguir.

2.3.1 Métodos de Partição

Os algoritmos de partição descrevem um método em que o conjunto de dados é particionado em um número predeterminado de grupos, k. Verificar todas as possíveis partições é na maioria das vezes inviável, portanto o processo tende a investigar algumas delas visando encontrar aquela que torne ótimo um critério de adequacidade da partição ou aquela que seja quase ótima.

Os algoritmos de partição diferem um do outro pela escolha de um ou mais dos seguintes procedimentos:

(a) Método de iniciar os agrupamentos;

(b) Método de designar objetos aos agrupamentos iniciais;

(c) Método de redesignar um ou mais objetos já alocados para outros agrupamentos.

Método das k-Médias

O algoritmo das k-médias alterna entre calcular os centróides baseados nos atuais membros dos grupos e designar as observações aos grupos baseados nos novos centróides. O centróide de cada grupo é a média das coordenadas do vetor de observações de seus membros e as observações são designadas para os grupos cujo centróide está mais próximo, baseando-se num critério de mínimos quadrados. O uso de mínimos quadrados para este método de k-médias é adequado visto que é menos resistente a valores extremos que o método baseado nos medoides PAM (partitioning around medoids), que está baseado na soma das dissimilaridades no lugar da soma das distâncias euclidianas ao quadrado. Para mais detalhes ver Kaufman e Rousseeuw (1990).

O método das k-médias é um dos mais conhecidos métodos de partição . Numa versão mais simples é composto dos seguintes passos:

1. Os dados são separados em k grupos iniciais.

2. Designar cada objeto ao grupo cujo centro está mais próximo dele. A distância Euclidiana é utilizada para realizar este cálculo estando as observações padronizadas ou não.

3. Repetir o passo 2 até não ocorrer mais deslocamentos das observações.

Uma forma diferente de utilizá-lo é especificando k centros iniciais dos k grupos prede-

terminados e então realizar o passo 2.

2.3.2 Métodos Hierárquicos

As técnicas hierárquicas podem ser classificadas em *aglomerativas* ou *divisivas*. Métodos hierárquicos aglomerativos iniciam considerando que cada um dos n objetos representam um grupo e segue unindo os mais similares, formando assim novos grupos, até que todos os objetos façam parte de um mesmo grupo. Já os métodos divisivos partem de um único grupo que vai sendo dividido sucessivamente até que o número de subgrupos formados coincida com o número de objetos. O que caracteriza estes processos é que a reunião de dois agrupamentos numa certa etapa produzem um dos agrupamentos da etapa superior, caracterizando o processo hierárquico. Os processos aglomerativos são mais utilizados do que métodos divisivos. Classificações hierárquicas produzidas tanto pelos métodos divisivos ou aglomerativos podem ser representadas por um diagrama bidimensional conhecido como *dendograma*. A seguir são apresentados alguns algoritmos hierárquicos.

Método do Centróide

Este processo é o mais direto dentre os algoritmos hierárquicos. A distância entre os grupos neste método é definida pela distância entre os seus centros. A união entre os grupos é realizada por aqueles que possuem a menor distância entre si. A maior dificuldade desta técnica é ter que recalcular os centros dos grupos a cada união realizada.

Método da Ligação Simples ou do Vizinho Mais Próximo

Neste método a distância entre dois grupos é definida pelos dois elementos mais parecidos, isto é, entre as medidas de similaridade e dissimilaridade entre os objetos de um grupo e do outro é escolhida a de maior valor como a medida entre os dois grupos. Portanto, a distância entre os grupos X e Y é definida como

$$d(X, Y) = \min\{d(i, j) : i \in X \text{ e } j \in Y\}$$

no caso de dissimilaridades, e

$$s(X, Y) = \max\{s(i, j) : i \in X \text{ e } j \in Y\}$$

no caso de similaridades.

Método da Ligação Completa ou do Vizinho Mais Distante

De forma contrária ao método anterior a similaridade e dissimilaridade entre dois grupos é definida pelos objetos de cada grupo que menos se parecem, ou seja, no caso de dissimilaridades a distância entre os grupos X e Y é calculada da seguinte forma

$$d(X, Y) = \max\{d(i, j) : i \in X \text{ e } j \in Y\}$$

já a similaridade

$$s(X, Y) = \min\{s(i, j) : i \in X \text{ e } j \in Y\}.$$

Convém ressaltar que a união entre os grupos ainda será feita com os mais parecidos, isto é, aqueles que possuirem a menor distância.

Método de Ward

Ward (1963) propôs um procedimento de agrupamento que busca formar partições, $P_n, \ldots, P_1$, até o ponto que minimize a perda de informação associada com cada agrupamento. Ward define esta perda como a soma do quadrado dos erros.

Por fim vale a pena destacar que das técnicas hierárquicas s métodos divisivos tem a mesma potencialidade que os aglomerativos, porém caso se pretenda obter um grande número de grupos é preferível se utilizar os métodos divisivos já que nos aglomerativos isto só é obtido depois de um grande número de uniões de pequenos grupos.

Os algoritmos aglomerativos são geralmente classificados em *monotéicos* e *politéicos*. No monotéico a divisão de um grupo em dois é realizada baseando-se apenas em uma variável. Denominado como MONA (Monothetic Analysis) e é descrito no capítulo 7 de Kaufman e Rousseaw (1990). Os politéicos consideram p variáveis simultaneamente para fazer a partição. Podemos citar como exemplo o algoritmo DIANA também descrito em Kaufman e Rousseaw (1990).

No próximo capítulo apresentamos um método hierárquico particular: o chamado *método de agrupamento baseado em modelos*.

Capítulo 3

Método de Agrupamento Baseado no Modelo

Como já foi mencionado nos capítulos anteriores, na análise de agrupamento existem vários métodos para a formação dos grupos das observações , dentre eles aqueles envolvendo o agrupamento hierárquico. Uma das propostas deste tipo é o agrupamento baseado no modelo (model-based clustering) que está fundamentado na suposição que as observações são geradas de uma mistura de probabilidades subjacentes (ver por exemplo, Bock 1996, 1998a, 1998b).

Modelos de mistura finita têm sido estudados no contexto de agrupamento. Neles cada componente da distribuição de probabilidade corresponde a um grupo. Os problemas de determinar o número de grupos e da escolha do agrupamento apropriado podem ser resolvidos com a escolha de um modelo estatístico. Apresentaremos os métodos de agrupamento hierárquico baseados na verossimilhança e no algoritmo EM para a estimação por máxima verossimilhança de mistura de distribuições normais multivariadas, sendo estas propostas complementares. O agrupamento hierárquico baseado em modelos produz razoáveis partições mesmo quando é iniciado sem nenhuma informação sobre os grupos. A inicialização do algoritmo EM é um ponto crítico. Então é possível iniciar a iteração do algoritmo EM com as partições obtidas a partir do agrupamento basedo nos modelos, isto é, na verossimilhança da classificação .

Neste capítulo faremos uma revisão das densidades de mistura finita, apresentaremos o algoritmo EM para modelos de mistura e o agrupamento hierárquico baseado nos modelos.

3.1 Densidades de Mistura Finita

Sejam $y_1, \ldots, y_n$ observações independentes multivariadas. A verossimilhança para um modelo de mistura com G componentes é

$$L_{\mathrm{M}}(\theta_1, \ldots, \theta_G; \tau_1 \ \ldots, \tau_G \mid y) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f(y_i \mid \theta_k), \tag{3.1}$$

onde f_k e θ_k são as densidades e os parâmetros, respectivamente, da k-ésima componente na mistura , e τ_k é a probabilidade de uma observação pertencer a k-ésima componente, $\tau_k \geq 0$ e $\sum_{k=1}^{G} \tau_k = 1$, com $y = (y_1, \ldots, y_n)^T$.

Misturas finitas fornecem modelos adequados para análise de agrupamento se assumimos que cada grupo de observações tem origem em populações, cada uma com uma distribuição de probabilidade diferente. Esta pode pertencer a mesma família de distribuições mas diferir nos valores dos parâmetros. Um caso particular muito utilizado é quando f_k é a densidade normal multivariada ϕ_k, parametrizada pelo vetor de média μ_k e a matriz de covariância Σ_k:

$$\phi_k(y_i \mid \mu_k, \Sigma_k) \equiv \frac{\exp\{-\frac{1}{2}(y_i - \mu_k)^T \Sigma_k^{-1}(y_i - \mu_k)\}}{\sqrt{\mid 2\pi\Sigma_k \mid}}.$$

Dados gerados por misturas de densidades normais multivariadas têm como característica grupos centrados nas médias μ_k. Notemos que a densidade aumenta para os pontos próximos da média. As superfícies correspondentes a densidades constantes são de forma elipsoidal. As características geométricas dos grupos: a forma, o volume e a orientação , são determinados pela matriz de covariância Σ_k que também pode ser parametrizada impondo restrições ao longo dos grupos. Podemos citar como exemplos: $\Sigma_k = \lambda I$, onde a distribuição de todos os grupos é esférica e possuem o mesmo tamanho; $\Sigma_k = \Sigma$ em que a matriz de covariância é constante ao longo dos grupos, todos possuem a mesma geometria mas não necessariamente são esféricos (Friedman e Rubin, 1967); e Σ_k sem restrição onde cada grupo pode ter uma geometria diferente (Scott e Symons, 1971).

Banfield e Raftery (1993) desenvolveram critérios mais gerais para agrupamentos que permitem variar algumas características da distribuição do grupo (orientação, tamanho e forma) entre os grupos, enquanto força outras a serem a mesma. O que eles propuseram foi uma reparametrização da matriz de covariância Σ_k em termos da decomposição dos seus autovalores da forma

$$\Sigma_k = D_k \Lambda_k D_k^T, \tag{3.2}$$

onde D_k é a matriz ortogonal de autovetores e Λ_k é uma matriz diagonal com os autovalores de Σ_k na diagonal. A orientação das principais componentes de Σ_k é determinada por D_k, enquanto Λ_k especifica o tamanho e forma do contorno da densidade. Descrevemos $\Lambda_k = \lambda_k A_k$, onde λ_k é o primeiro autovalor de Σ_k, $A_k = \text{diag}\{\alpha_{1k}, \ldots, \alpha_{pk}\}$, e $1 = \alpha_{1k} \geq \alpha_{2k} \geq \cdots \geq \alpha_{pk} > 0$. Assim D_k determina a orientação do k-ésimo grupo, λ_k seu tamanho, e A_k sua forma. Pelo tamanho medimos o volume ocupado pelo grupo no p-espaço em vez do número de elementos contidos nele. Se os α_{jk}'s são de magnitudes semelhantes, então o k-ésimo grupo tenderá a ser hiperesférico, enquanto se $\alpha_{2k} << 1$, ele estará concentrado sobre uma linha e se $\alpha_{3k} << 1$ ele estará concentrado sobre um plano 2-dimensional no p-espaço e assim sucessivamente.

3.2 O Algoritmo EM para Modelos de Mistura

O algoritmo EM foi originalmente proposto por Dempster, Laird e Rubin (1977) como um método geral para se obter a estimação de máxima verossimilhança dos parâmetros em problemas com dados imcompletos. Seu uso para a estimação em modelos de mistura foi estudado em detalhes por McLachlan e Basford (1988) e Roeder e Walker (1984).

Os dados podem ser considerados como n observações multivariadas x_i, recuperadas por (y_i, z_i) onde y_i é observado e z_i é a parte não observada. Sendo as x_i independentes e identicamente distribuidas (iid) com função de densidade f com parâmetro θ, assim temos que a verossimilhança para dados completos é

$$L_c(x_i \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta). \tag{3.3}$$

Além disso, se a probabilidade de uma variável em particular ser não observada depender apenas dos dados observados y e não de z, obtemos a verossimilhança dos dados observados, $L_o(y_i \mid \theta)$, integrando z sob (3.3),

$$L_o(y \mid \theta) = \int L_c(x \mid \theta) dz. \tag{3.4}$$

O estimador de máxima verossimilhança (EMV) de θ baseado nos dados observados maximiza (3.4).

No algoritmo EM para modelos de mistura, os dados completos são considerados como sendo $x_i = (y_i, z_i)$, onde $z_i = (z_{i1}, \ldots, z_{iG})$ é a parte não observada dos dados

$$z_{ik} = \begin{cases} 1, & \text{se } x_i \text{ pertence ao grupo } k, \\ \\ 0, & \text{caso contrário.} \end{cases} \tag{3.5}$$

Os z_i são considerados iid segundo uma distribuição multinomial extraída de alguma das G categorias com probabilidades $z_1, \ldots, z_G$ e, a densidade de uma observação y_i dado z_i é dada por $\prod_{k=1}^{G} f_k(y_i \mid \theta_k)^{z_{ik}}$. O logaritmo da verossimilhança dos dados completos é:

$$l(\theta_k, \tau_k, z_{ik} \mid x) = \sum_{i=1}^{n} \sum_{k=1}^{G} z_{ik} \log[\tau_k f_k(y_i \mid \theta_k)]. \tag{3.6}$$

O termo EM foi utilizado porque cada iteração deste algoritmo consiste de dois passos: o passo E e o passo M. O passo E, da esperança, calcula $\hat{z}_{ik} = E(z_{ik} \mid y_i, \theta_1, \ldots, \theta_G)$ que é a esperança condicional dada a observação y_i e o vetor de parâmetros de y_i pertencer ao k-ésimo grupo que para os modelos de mistura é calculada da seguinte forma

$$\hat{z}_{ik} \leftarrow \frac{\hat{\tau}_k f_k(y_i \mid \hat{\theta}_k)}{\Sigma_{j=1}^{G} \hat{\tau}_j f_j(y_i \mid \hat{\theta}_j)}.$$

O passo M, da maximização, onde são determinados os parâmetros que maximizam a log-verosimilhança (3.6) em termos de τ_k e θ_k com os valores fixos de z_{ik} dados pelos valores obtidos no passo E, ($\hat{z}_{ik}$ calculada no passo E anterior). Os valores z^*_{ik} de $\hat{z}_{ik}$ no máximo de (3.1) é a probabilidade condicional estimada de que a i-ésima observação pertença ao grupo k. A máxima verossimilhança da classificação da i-ésima observação é tal que $(1 - \max_k z^*_{ik})$ é uma medida da incerteza da classificação (Bensmail et. al, 1997). A parte não observada dos dados envolve valores que são faltantes devido a falta de resposta e/ou quantidades que são introduzidas para reformular o problema para EM. Sob certas condições de regularidade, EM pode ser demonstrado convergir para um máximo local da verossimilhança dos dados observados (exemplos podem ser verificados em Dempster, Laird e Rubin, 1977; Boyles, 1983; Wu, 1983; McLachlan e Krishnan, 1997, Seild et al., 2000; Karlis e Xekalaki, 2003; Biernacki et. al, 2003). Embora essas condições nem sempre sejam asseguradas na prática, o algoritmo EM tem sido muito usado para estimação da verossimilhança máxima para modelos de mistura, com bons resultados, (ver por exemplo, Celeux e Govaert, 1993).

Entre as vantagens deste algoritmo podemos citar:

• Facilmente implementado já que se baseia em resultados de dados completos.

• Numericamente estável, a sucessão de iterações converge, quase sempre, para um máximo local do logaritmo da função de verossimilhança.

As principais críticas que podem ser destacadas ao algoritmo são :

• O algoritmo pode ter problemas de convergência quando a covariância associada com uma ou mais componentes é singular. Também pode falhar ou dar resultados imprecisos se um ou mais grupos contêm poucas observações (isto acontece se existem demasiadas componentes na mistura) ou se as observações dos grupos estão concentradas num subespaço linear de dimensão menor que os dados.

• O passo E pode ser analiticamente intratável em alguns problemas. Nestas situações pode ser utilizada uma aproximação via Monte Carlo.

3.3 Seleção do Modelo

Na análise de agrupamento existem duas questões importantes a serem consideradas: a seleção do método de agrupamento e a determinação do número de grupos. Na modelagem de mistura estas questões reduzem-se a seleção de modelos já que cada combinação de um número de grupos e um método de agrupamento equivalem a um modelo estatístico. Utilizando um modelo mais complexo, um número menor de grupos poderá ser suficiente, no entanto escolhendo um modelo mais simples existirá talvez a necessidade de um número maior de grupos para ajustar os dados adequadamente. Uma das vantagens desta técnica é o

uso do fator de Bayes para a escolha do melhor modelo (ver Kass e Raftery, 1995; Giampaoli e Singer, 2003).

Consideremos vários modelos, $M_1, \ldots, M_K$, com probabilidades a priori $p(M_k)$, $k = 1, \ldots, K$, respectivamente, então pelo teorema de Bayes a probabilidade a posteriori do modelo M_k dado que foram observados os dados D, $P(M_k \mid D)$, é proporcional a probabilidade do modelo M_k multiplicada pela probabilidade a priori do modelo, a saber

$$p(M_k \mid D) \propto p(D \mid M_k)p(M_k), \tag{3.7}$$

onde $p(D \mid M_k)$ é a chamada verossimilhança integrada do modelo M_k e é obtida por

$$p(D \mid M_k) = \int p(D \mid \theta_k, M_k)p(\theta_k \mid M_k)d\theta_k, \tag{3.8}$$

sendo $p(\theta_k \mid M_k)$ a chamada distribuição a priori de θ_k, vetor paramétrico do modelo M_k. Uma proposta bayesiana para selecionar o modelo é escolher aquele com maior probabilidade a posteriori, logo se as probabilidades a priori de cada modelo $P(M_k)$ são as mesmas, então por (3.7), basta considerar as verossimilhanças integradas $P(D \mid M_k)$. Para se comparar dois modelos, como por exemplo M_1 e M_2, o fator de Bayes é definido como a razão de duas verossimilhanças integradas

$$B_{12} = \frac{p(D \mid M_1)}{p(D \mid M_2)}.$$

Se $B_{12} > 1$ o modelo M_1 é escolhido em detrimento ao M_2, caso contrário é escolhido o modelo M_2. No entanto, a maior dificuldade para se usar o fator de Bayes é o cálculo das integrais envolvidas em (3.8). Na maioria dos casos estas são realizadas através de aproximações numéricas. Fraley e Raftery (2002b) supõem que os modelos são igualmente prováveis e propõem uma aproximação baseada no critério de informação bayesiana (Bayesian Information Critério ou BIC) proposto por Schwarz (1978):

$$2\log p(D \mid M_k) \approx 2\log p(D \mid \hat{\theta}_k, M_k) - v_k \log(n) = BIC_k,$$

onde v_k é o número de parâmetros independentes a serem estimados no modelo M_k e $\hat{\theta}_k$ é o estimador de máxima verossimilhança (EMV) de θ_k. O valor alto para o BIC de um determinado modelo indica forte evidência a favor deste. Geralmente diferenças maiores que 10 entre os BIC's de dois modelos é considerada uma forte evidência a favor de um deles. Como os modelos de misturas finitas não satisfazem condições de regularidade, esta aproximação não é válida para esta situação. Porém, várias aplicações de agrupamentos baseados nos modelos que consideram o critério BIC para a seleção do modelo apresentam bons resultados (ver, Fraley e Raftery, 2002b, Stanford e Raftery, 2000).

3.4 Agrupamentos Hierárquicos Baseados nos Modelos

Sejam $x_1, \ldots, x_n$ as observações e $f_k(x_i \mid \theta_k)$ a densidade da observação x_i da k-ésima componente que tem θ_k como seu parâmetro e G o número de componentes da mistura. Uma das maneiras de se formular o modelo para a composição dos grupos é através da escolha de θ e τ que maximizem a chamada *verossimilhança de classificação*

$$L(\theta, \gamma) = \prod_{i=1}^{n} f_{\gamma_i}(x_i \mid \theta_{\gamma_i}), \tag{3.9}$$

onde γ_i são indicadores da única classificação de cada observação , se $\gamma_i = k$ então x_i pertence a k-ésima componente. Na verossimilhança (3.1) cada componente da mistura está ponderada pela probabilidade de que cada observação pertença a cada componente. A presença do indicador γ_i na verossimilhança de classificação (3.9) faz com que não seja possível se obter a maximização exata.

Quando a $f_k(x; \theta)$ é uma densidade normal multivariada (μ_k, Σ_k) conforme McLachlan et. al, 2003 e Biernacki et. al, 2003, a verossimilhança de classificação assume a seguinte forma

$$L(\theta, \gamma) = \text{const} \prod_{k=1}^{G} \prod_{i \in E_k} \mid \Sigma_k \mid^{-1/2} \exp\{\frac{-1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)\} \tag{3.10}$$

onde $E_k = \{i : \gamma_i = k\}$. O estimador de máxima verossimilhança (EMV) de μ_k é $\overline{x}_k = n_k^{-1} \Sigma_{i \in E_k} x_i$, onde n_k é o número de elementos em E_k. Substituindo μ_k pelo seu EMV em (3.10) e calculando a log-verossimilhança obtemos

$$l(\theta, \gamma) = \text{const} - \frac{1}{2} \Sigma_{k=1}^{G} \{\text{tr}(W_k \Sigma_k^{-1}) + n_k \log \mid \Sigma_k \mid\}, \tag{3.11}$$

onde W_k é a matriz produto cruzado da amostra para o k-ésimo grupo, a saber

$$W_k = \Sigma_{i \in E_k} (x_i - \overline{x}_k)(x_i - \overline{x}_k)^T.$$

Note que W_k / n_k é o EMV de Σ_k.

Banfield e Raftery (1993) citam o seguinte:

• Se $\Sigma_j = \alpha^2 I$ $(k = 1, \ldots, G)$ então a log-verossimilhança em (3.10) é maximizada pela escolha do γ que minimiza $\text{tr}(W)$, onde $W = \Sigma_{k=1}^{G} W_k$.

• Se $\Sigma_k = \Sigma$ $(k = 1, \ldots, G)$ então a log-verossimilhança em (3.10) é maximizada pela escolha do γ que minimiza $\det(W)$.

• Se Σ_k não é restrita então a log-verossimilhança em (3.10) é maximizada pela escolha do γ que minimiza $\prod_{k=1}^{G} [\det(W_k / n_k)]^{n_k}$.

A diferença entre o procedimento da verossimilhança de classificação e a formulação de mistura finita é que o último assume que os x_i vem de uma distribuição de mistura, cujos parâmetros são estimados e então os membros dos grupos são determinados pelos valores máximos das probabilidades a posteriori estimadas. Em contrapartida, na proposta de verossimilhança de classificação assume-se que os x_i são originários de uma distribuição única $f_{\gamma_i}(x_i, \theta_{\gamma_i})$ que é determinada pelo parâmetro desconhecido γ_i. A associação de grupos é então estimada diretamente pelos γ_i que maximizam a verossimilhança de classificação.

3.5 Combinando Aglomeração Hierárquica, EM, e Fator de Bayes

No agrupamento aglomerativo, cada etapa onde os grupos são unidos corresponde a um único número de grupos e uma única partição dos dados. Uma determinada partição pode ser transformada em uma variável indicadora do tipo (3.5), que pode ser então usada na probabilidade condicional no passo M do algoritmo EM para estimação dos parâmetros, inicializando uma iteração no EM. Combinando isto com a seleção do modelo via BIC resulta na estratégia de agrupamento detalhada abaixo:

• Determinar um número máximo de grupos (M), e um conjunto de modelos de mistura a serem analisados.

• Realizar a aglomeração hierárquica visando maximizar aproximadamente a verossilhança de classificação para cada modelo e obter a classificação correspondente para todos os possíveis números de grupos até M.

• Implementar o algoritmo EM para cada modelo de mistura e cada número de grupos $2, \ldots, M$, iniciando com a classificação e aglomeração hierárquica.

• Calcular o BIC de um determinado caso de cada modelo e para o modelo de mistura com os parâmetros ótimos de EM para $2, \ldots, M$, grupos.

A escolha do melhor modelo associado a um número de grupos será direcionada para o modelo cujo BIC obteve valor máximo. Um conjunto de modelos apropriados para a realização do agrupamento dos dados em geral, na prática é aquele formado pela parametrização de misturas normais multivariadas através da decomposição dos autovalores como em (3.2). Com estes modelos o cálculo pode ser reduzido fazendo aglomeração hierárquica apenas para um dos modelos, por exemplo quando a matriz de covariância não possui restrição, usando as partições resultantes como valores iniciais do EM com alguma outra parametrização (ver Fraley e Raftery, 2000b).

3.6 Modelando as pertubações

A estratégia para análise de agrupamento baseada no modelo descrita até agora não

é diretamente aplicada a dados definidos como perturbações. No entanto o modelo pode ser modificado de forma que o algoritmo EM trabalhe de maneira satisfatória com uma identificação inicial do que é uma observação e uma pertubação. A pertubação é considerada como proveniente de um processo de razão constante de Poisson, resultando assim a seguinte mistura de verossimilhança

$$\tilde{L}_{\mathrm{M}}(\theta_1, \ldots, \theta_G; \tau_1 \ \ldots, \tau_G \mid y) = \prod_{i=1}^{n} \left[\frac{\tau_0}{V} + \sum_{k=1}^{G} \tau_k f(y_i \mid \theta) \right],$$

onde V é o hipervolume da região dos dados, $\tau_k \geq 0$, e $\sum_{k=0}^{G} \tau_k = 1$. Alternativamente pontos extremos isolados podem ser tratados por amostragem iterativa (ver, Fayyad e Smith, 1996) na qual pontos com baixa probabilidade são retiradas do grupo, e o processo é reiterado até que todas as observações restantes tenham alta probabilidade. Outra alternativa é trabalhar com as pertubações modelando as misturas via distribuição t (ver Peel and McLachlan, 2000).

Quando nos deparamos com um conjunto de dados contendo uma grande quantidade de pertubações devemos modificar o método de agrupamento baseado no modelo da seguinte forma:

• Obter inicialmente uma classificação de cada observação como sendo uma pertubação ou não. Um método possível para isto inclui o método Voronoï (Allard e Fraley, 1997) e um método do vizinho mais próximo (Byers e Raftery, 1998).

• Aplicar um método de agrupamento hierárquico para os dados sem as perturbações.

• Aplicar o algoritmo EM baseado no modelo Gaussiano com a inclusão de termos definidos como pertubações para todo o conjunto de dados. Valores iniciais para z_{ik} são formados pelo aumento das variáveis indicadoras do passo do agrupamento hierárquico com uma linha de zeros para cada observação inicialmente avaliada como sendo pertubações , e uma coluna de variáveis indicadoras dando o resultado do passo em que foram retiradas as perturbações . Um exemplo pode ser visto em Fraley e Raftery, (2000b).

3.7 Vantagens e limitações

Os métodos de agrupando baseados em modelos de mistura normais multivariadas descritos por Fraley e Raftery (2000b) tem sido usado com sucesso em várias aplicações tais como a detecção de minas e falhas sísmicas (Dasgupta e Raftery, 1998), identificação de falhas de imagens (Campbell et. al, 1997), e a classificação de dados astronômicos (Mukherjee et. al, 1998). Contudo, o uso destes métodos sem qualquer modificação poderá ser limitado para dados não normais, de alta dimensão ou para um grande conjunto de dados.

3.8 Dados não gaussianos

Misturas finitas com componentes normais multivariadas tem sido muito utilizadas para modelos de dados multivariados contínuos. Uma das vantagens destas misturas é sua conveniência computacional. Eles podem ser facilmente ajustados diretamente pela máxima verossimilhança ou via algoritmo EM. No entanto, em algumas situações o uso de componentes normais pode ser inadequado já que podem tornar o modelo de mistura inviável. Visando solucionar este problema Peel e McLachlan (2000) consideraram como uma alternativa adequada misturas de distribuições t multivariadas.

Uma componente não-Gaussiana pode ser frequentemente aproximada por várias Gaussianas (ver, Dasgusta e Raftery, 1998; e Fraley e Raftery, 1998). Suponha que um componente está concentrado sobre uma curva não-linear, isto pode ser possível realizando uma aproximação linear por partes que pode ser representada por vários grupos gaussianos, cada um concentrado sobre um subespaço linear. Uma proposta para solucionar o caso quando os grupos estão concentrados em torno de curvas não lineares em vez de linhas é modelar as curvas usando o conceito de *curvas principais* (Hastie e Stuetzle, 1989). O agrupamento sobre curvas principais foi proposto e desenvolvido por Banfield e Raftery (1992) e Stanford e Raftery (2000). Banfield e Raftery (1993) também sugeriram alguns critérios para algumas situações não -Gaussianas.

Capítulo 4

Aplicação

Para aplicar a técnica de agrupamento hierárquico baseada no modelo, utilizando o pacote MCLUST já citado, foram utilizadas variáveis de tipo contínuas onde não ocorreram falta de informações das variáveis, em nenhum dos elementos a serem agrupados.

Na presente aplicação objetiva-se agrupar os municípios de Pernambuco, e, para isso foram obtidos dados oficiais do IPEA e do SUS do ano de 2001. O IPEADATA é uma base de dados macroeconômicos sobre o Brasil organizada, pelo Instituto de Pesquisa Econômica Aplicada (IPEA). Contém mais de 5000 séries sendo 2500 de uso público com acesso gratuito na Internet abrangendo os seguintes temas: Balanço de Pagamentos, Câmbio, Comércio Exterior, Consumo e Vendas, Contas Nacionais, Economia Internacional, Emprego, Finanças Públicas, Indicadores Sociais, Juros, Moeda e Crédito, População, Preços, Produção , Salário e Renda. O Departamento de Informática do Sistema Único de Saúde (SUS), DATASUS, tem como missão prover os órgãos do SUS de sistemas de informação e suporte de informática necessários ao processo de planejamento, operação e controle do SUS, através da manutenção de bases de dados nacionais, apoio e consultoria na implantação de sistemas e coordenação das atividades de informática inerentes ao funcionamento integrado dos mesmos. Suas principais linhas de atuação são: manutenção das bases nacionais do Sistema de Informações de Saúde; desenvolvimento e disseminação de sistemas de informação de saúde; desenvolvimento, seleção e disseminação de tecnologias de informática para a saúde, adequadas ao país; consultoria para a elaboração de sistemas do planejamento, controle e operação do SUS; suporte técnico para informatização dos sistemas de interesse do SUS, em todos os níveis; normatização de procedimentos, softwares e de ambientes de informática para o SUS; apoio à capacitação das secretarias estaduais e municipais de saúde para a absorção dos sistemas de informações no seu nível de competência; incentivo à formação de uma rede para intercâmbio e disseminação de informações de interesse do SUS via Internet, BBS e outras formas complementares. No banco do IPEA no ano de 2001 faltam informações para 31 dos municípios do estado de Pernambuco, dos quais 17,no banco do SUS, também não são fornecidas informações , portanto estes municípios não foram considerados e estão listados no Apêndice C. O município do Recife não foi considerado para podermos ter uma melhor classificação já que este município apresenta valores das variáveis bem distintos dos demais levando-os a ficarem em um mesmo grupo. Estes dados estão disponíveis na internet nos

endereços: `www.ipeadata.gov.br` e `www.datasus.gov.br` respectivamente. No apêndice A e B estão as saídas para o agrupamento realizado com o banco de dados do IPEA e do SUS, respectivamente,

Nos Quadros 1 e 2 encontram-se descritas as variáveis que foram retiradas do arquivo de dados do IPEA e do SUS, respectivamente.

Quadro 1. Variáveis referentes ao banco de dados do IPEA - 2001.

Código	Variável
V1	Desp. Municipais por Função de Saúde e Saneamento - Anual - R$
V2	Desp. Municipais por Função de Transporte - Anual - R$
V3	Desp. Municipais por Função Legislativa - Anual - R$
V4	Investimento Municipal - Anual - R$
V5	Impostos Municipais Total - Anual - R$
V6	IPTU - Anual - R$
V7	ISS - Anual - R$
V8	Outras Receitas Correntes Municipais - Anual - R$
V9	Transferências para os Municípios Referentes ICMS - Anual - R$
V10	Transferência para os Municípios referentes ao IPVA - Anual - R$
V11	Transferências Correntes de Tributos Estaduais Para os Munípios - Anual - R$
V12	Receita Corrente Municipal - Anual - R$
V13	Outros Impostos Municipais - Anual - R$
V14	Outras Transferências Correntes para os Municípios - Anual - R$
V15	Receita Municipal com Transferências Correntes Total - Anual - R$
V16	Total das Despesas Municipais por Função - Anual - R$
V17	Receita Orçamentária Municipal - Anual - R$
V18	Receita Tributária Municipal - Anual - R$
V19	Receita Municipal de Capital - Anual - R$
V20	Taxas Municipais Total - Anual - R$
V21	Transferência para os Municípios referentes ao IPVA - Anual - R$
V22	População

Quadro 2. Variáveis referentes ao banco de dados do SUS. - 2001

Código	Variável
V23	Despesas Total com Saúde - Anual - R$
V24	Despesa de Recursos Próprios - Anual - R$
V25	Receita de Impostos e Transferências Constitucionais e Legais - Anual - R$
V26	Transferências SUS - Anual - R$
V27	Despesa Pessoal - Anual - R$

A técnica de agrupamento apresentada no capítulo anterior foi aplicada para os dois bancos separadamente, visando comparar assim as diferenças obtidas nos agrupamentos. Para tanto foram feitos agrupamentos para diferentes casos: considerando as variáveis na

sua forma original (dados brutos) usando a população como uma variável (DBCP); utilizando o quantitativo monetário percapita (QMP) e utilizando os dados originais sem a população como variável (DBSP). Nos apêndices A e B estão as saídas do pacote MCLUST para agrupamente dos três casos considerados, para os bancos do IPEA e do SUS, respectivamente.

Os modelos considerados nos agrupamentos se referem à distribuição dos dados levando em consideração a parametrização da matriz de covariância, seu volume, forma e a orientação da distribuição para os grupos. No Quadro 3 estão todos os modelos que foram considerados nesta aplicação. O número máximo de grupos a ser testado foi 9, já que a medida que o número de grupos aumentava o BIC apresentava um decréscimo para os modelos em questão.

Quadro 3. Modelos considerados para agrupar os municípios.

Código	Distribuição	Volume	Forma	Orientação
EII	Esférica	igual	igual	NA
VII	Esférica	variável	igual	NA
EEI	Diagonal	igual	igual	eixos de coordenadas
VEI	Diagonal	variável	igual	eixos de coordenadas
EVI	Diagonal	igual	variável	eixos de coordenadas
VVI	Diagonal	variável	variável	eixos de coordenadas
EEE	Elipsoidal	igual	igual	igual
VVV	Elipsoidal	variável	variável	variável
EEV	Elipsoidal	igual	igual	variável
VEV	Elipsoidal	variável	igual	igual

4.1 Resultado do agrupamento com os dados do IPEA

A princípio foi realizada uma análise exploratória dos dados visando identificar dados faltantes, observações discrepantes bem como algumas medidas de locação. O Quadro 4 apresenta um resumo das variáveis de interesse deste banco. Podemos destacar uma grande dispersão relativa nas variáveis principalmente em V6 e V20, cujos coeficientes de variação são respectivamente 440% e 410%. Dentre as 22 variáveis, 9 apresentam dados faltantes. Percebemos também assimetria positiva na distribuição de todas as variáveis já que os valores das médias são maiores do que das medianas.

Para o agrupamento foram escolhidas as variáveis que não possuiam dados faltantes por ser esta uma restrição do método. Para o caso, se considerou como melhor modelo aquele que apresentou um valor máximo de BIC. No Quadro 5 é apresentado qual é o melhor modelo com seu respectivo BIC, a incerteza associada a cada modelo, o número de grupos resultantes e o número de elementos de cada grupo. Deste quadro podemos observar que para o caso em que se considerou os dados brutos com a população como variável (DBCP), o melhor modelo

foi VEV, se constituiram 4 grupos não havendo uma concentração dos municípios em um determinado grupo. Ao retirar a população (DBSP), o agrupamento se manteve semelhante no número de grupos e no modelo, VEV, no entanto ocorreu uma concentração de 66% dos municípios no grupo 1. Com os valores percapita (QMP), os municípios formaram 6 grupos sendo o modelo agora o EEE, concentrando os municípios no grupo 1. Dentre os três agrupamentos o que apresentou a menor incerteza foi aquele com os dados brutos sem a população e foi este o considerado como melhor.

Os grupos resultantes deste melhor agrupamento são apresentados de forma ilustrativa no Mapa 1, onde 0 significa que o município não foi considerado no agrupamento. Neste mapa podemos constatar que os agrupamentos não dependem das regiões geográficas. Algumas medidas descritivas de cada uma das variáveis nos 4 grupos são apresentadas nos Quadros 6, 7 e 8. Nos Gráficos de 1 à 12 estão ilustrados os diagramas box-plot de cada variável, em cada grupo. Os municípios com valores extremos nestes gráficos estão a seguir em ordem crescente de valores.

Variável V1 - grupo 2: Cabo de Santo Agostinho, Petrolina, Camaragibe; grupo 3: Santa Cruz do Capibaribe.

Variável V4 - grupo 1: Belém de Maria, Lagoa de Ouro, Rio Formoso, Betânia, Jatobá, grupo 2: Petrolina.

Variável V5 - grupo 1: Afogados da Ingazeira, grupo 2: Petrolina, Ipojuca, Caruaru, Olinda, grupo 4: Salgueiro.

Variável V7 - grupo 1: Lagoa Grande, grupo 2: Petrolina, Olinda, grupo 3: Santa Maria da Boa Vista. Salgueiro.

Variável V9 - grupo 1 Amaraji, Itacuruba, Jatobá, Joaquim Nabuco, Lagoa do Itaenga, Rio Formoso, grupo 2: Cabo de Santo Agostinho, grupo 4: Vicência.

Variável V11 - grupo 1: Amaraji, Itacuruba, Jatobá, Lagoa do Itaenga, Maraial, Rio Formoso, grupo 2: Paulista, Cabo de Santo Agostinho grupo 4: Vicência.

Variável V12 - grupo 1: Rio Formoso, grupo 2: Petrolina, Caruaru, Paulista, Cabo de Santo Agostinho.

Variável V14 - grupo 1: Rio Formoso, grupo 2: Caruaru, Paulista, Petrolina, Cabo de Santo Agostinho, Camaragibe. grupo 3: Santa Maria da Boa Vista.

Variável V15 - grupo 1: Rio Formoso, grupo 2: Petrolina, Paulista, Cabo de Santo Agostinho, grupo 4: Chã Grande.

Variável V16 - grupo 1: Rio Formoso, grupo 2: Caruaru, Paulista, Petrolina, Cabo de Santo Agostinho.

Variável V17 - grupo 1: Rio Formoso, grupo 2: Caruaru, Petrolina, Olinda, Paulista, Cabo de Santo Agostinho.

Variável V18 - grupo 1: Sertânia, grupo 2: Paulista, Caruaru, Olinda, grupo 4: Salgueiro.

Os resultados apresentados nos Quadros 6, 7 e 8 revelam que ao realizar o agrupamento os grupos gerados apresentam, como era esperado, uma menor dispersão em todas as variáveis. Através dos diagramas box-plot percebemos que o grupo 1 agrega 101 municípios com os menores valores para todas as variáveis, e que existem alguns valores extremos especialmente nas variáveis V4, V7, V9 e V11. Estas apresentam também os maiores coeficientes de variação, 62%, 62%, 136% e 130%, respectivamente.

Os grupos 3 e 4 são bastante semelhantes quanto a grandeza das variáveis, com o grupo 3 apresentando valores um pouco maiores para quase todas as variáveis. A presença de valores extremos não é muito frequente nestes grupos. Os diagramas box-plot referidos revelaram uma certa homogeneidade dos valores de cada uma das variáveis nos grupos 1, 3 e 4 constatada pelo baixo valor dos coeficientes de variação. No entanto, os resultados para o grupo 2 demonstram uma grande heterogeneidade entre os 28 municípios para todas as variáveis, com valores médios superiores ao restante dos grupos. Vale a pena destacar que neste grupo 2 os municípios que possuem valores extremos em quase todas as variáveis são: Cabo de Santo Agostinho, Camaragibe, Caruaru, Olinda, Paulista e Petrolina, ainda contribuindo para a maior assimetria positiva de todas as variáveis identificamos: Agrestina, Araripina, Belo Jardim, Bezerros, Floresta, Garanhuns, Goiana, Gravatá, Iati, Igarassu, Ipojuca, Itapissuma, João Alfredo, Moreno, Nazaré da Mata, Ouricuri, Palmares, Pesqueira, Petrolândia, Petrolina, Pombos, Serra Talhada e Vitória de Santo Antão.

Quadro 4. Estatísticas básicas do banco de dados do IPEA para as variáveis de interesse.

Variáveis	Medidas						
	MIN	$\overline{x}$	Md	MAX	D.P.	C.V. (%)	NA's[1]
V1	61738,44	2290692,09	1534969,73	27149473,93	3301601,21	144	0
V2	100,31	184034,25	76977,56	4101641,23	448174,88	244	19
V3	9284,06	453109,07	272912,74	4655272,67	648107,85	143	10
V4	117392,50	940381,30	572591,40	10539035,80	1251281,80	130	0
V5	11874,00	479037,16	86794,52	16879165,12	1699933,56	355	0
V6	680,95	148385,03	11065,43	6722104,04	652292,61	440	3
V7	8898,13	299038,82	59916,66	9126363,72	969749,16	296	0
V8	203,00	594618,08	162748,96	14584624,62	1504578,33	253	1
V9	53805,14	1963574,58	462722,00	32712585,34	4269084,21	217	0
V10	2220,75	143222,42	28304,72	4019718,11	430649,67	301	1
V11	56233,38	2105860,91	48771,00	33157863,36	4528445,40	215	0
V12	2309146,00	10854135,00	6758938,00	84056667,00	13381740,00	123	0
V13	57,30	35689,13	5435,50	1030697,36	114583,78	321	5
V14	328269,70	3164961,90	2300852,10	25720043,90	3564208,90	113	0
V15	2168221,00	9582754,00	6373751,00	64243472,00	10358728,00	108	0
V16	1985388,00	10210274,00	6496946,00	77490057,00	12061009,00	118	0
V17	2309146,00	11085109,00	6989454,00	84522332,00	13713404,00	124	0
V18	12396,00	680648,98	109952,47	25620800,51	2486094,63	365	0
V19	146,10	289665,00	122415,20	5875747,00	668195,00	231	31
V20	23,00	204282,17	23873,92	8741635,39	837522,53	410	2
V21	2220,75	143222,42	28304,72	4019718,11	430649,67	301	1
V22	3734	34182	20219	372014	47920,67	140	0

[1] Quantidade de dados faltantes.

Quadro 5. Resumo dos agrupamentos feitos com o banco de dados do IPEA.

Variáveis	Grupos	Elementos por grupos	Melhor modelo	Incerteza	BIC
DBCP	4	28-12-52-61	VEV	0,20	−55084,10
DBSP	4	101-28-12-12	VEV	0,18	−52286,60
QMP	6	123-7-11-4-4-4	EEE	0,49	−15045,07

Quadro 6. Estatísticas básicas para as variáveis V1, V4, V5 e V7 dos quatro grupos gerados no agrupamento feito com os dados brutos sem a população.

Variáveis	Medidas	GRUPOS			
		1	2	3	4
V1	MIN	61738,44	1895395,00	2055919,60	1752494,00
	$\overline{x}$	1178084,01	6127572,00	2670821,30	2322293,50
	Md.	1082855,72	3469824,00	2607071,8	2361601,60
	MAX	2557790,28	27149474,00	3962421,50	3062664,90
	D.P.	530338,10	6337313,00	502439,90	490141,20
	C.V.(%)	45	103	20	21
V4	MIN	117392,50	150255,20	499731,50	393588,70
	$\overline{x}$	520507,40	2365875,90	1142163,30	946384,70
	Md.	422596,20	1716617,10	966319,60	917112,70
	MAX	1547356,00	10539035,80	2535403,00	1673352,40
	D.P.	321243,90	2320206,50	705548,00	387227,80
	C.V.(%)	62	98	62	41
V5	MIN	11874,00	56763,94	99000,88	28323,51
	$\overline{x}$	70303,90	2161354,30	306960,16	165879,18
	Md.	59976,49	484633,13	290717,20	156252,09
	MAX	216350,80	16879165,12	539357,00	410293,77
	D.P.	40245,29	3553780,69	175341,96	96304,13
	C.V.(%)	57	164	57	58
V7	MIN	11874,00	17221,37	50515,50	8898,13
	$\overline{x}$	53494,76	1316496,11	182325,90	108347,28
	Md.	46234,00	415106,00	161358,20	129596,00
	MAX	176578,40	9126363,72	464745,40	161800,18
	D.P.	33326,45	1990570,15	112265,90	55363,61
	C.V.(%)	62	151	62	51

MIN = mínimo valor da variável
MAX = máximo valor da variável
$\overline{x}$ = média
Md = mediana
D.P. = Desvio Padrão
C.V.(%) = Coeficiente de Variação

Quadro 7. Estatísticas básicas para as variáveis V9, V11, V12 e V14 dos quatro grupos gerados no agrupamento feito com os dados brutos sem a população.

Variáveis	Medidas	GRUPOS			
		1	2	3	4
V9	MIN	53805,14	255454,70	989695,60	235184,30
	$\overline{x}$	518052,19	7470018,50	2370967,10	874293,00
	Md.	253384,00	5359958,90	2466888,80	763779,60
	MAX	4214883,23	32712585,30	3592514,00	2699302,20
	D.P.	706625,07	7770987,40	888852,80	663874,80
	C.V.(%)	136	104	37	76
V11	MIN	56233,38	266394,60	1088430,90	268835,00
	$\overline{x}$	545421,86	8032495,10	2555200,40	961403,70
	Md.	285876,61	5627116,00	2643970,50	857423,20
	MAX	4261448,92	33157863,40	4000138,00	2774718,90
	D.P.	710088,64	8175970,30	900778,50	691573,70
	C.V.(%)	130	101	35	72
V12	MIN	2309146,00	6477675,00	11859454,60	9485208,00
	$\overline{x}$	5614184,00	28871229,00	12984599,90	10786701,00
	Md.	5336812,00	17651915,00	13117764,70	10424094,00
	MAX	1112089,00	84056667,00	14532415,00	12743742,00
	D.P.	1849458,00	23523400,00	780995,50	1071953,00
	C.V.(%)	33	81	6	10
V14	MIN	328269,70	581350,60	2339589,00	2435391,30
	$\overline{x}$	1806790,10	7296419,10	3998999,00	4122136,40
	Md.	1775021,20	5472468,80	3736965,00	4331817,10
	MAX	4012242,70	25720043,90	7100816,00	5124993,90
	D.P.	805756,00	6550243,70	1246424,00	791895,10
	C.V.(%)	46	90	31	19

Quadro 8. Estatísticas básicas para as variáveis V15, V16, V17 e V18, dos quatro grupos gerados no agrupamento feito com os dados brutos sem a população.

Variáveis	Medidas	GRUPOS			
		1	2	3	4
V15	MIN	2168221,00	5921654,00	10283881,50	8525766,50
	$\overline{x}$	5315340,00	23789924,00	12070202,80	9862644,30
	Md.	5106857,00	15804656,00	12238374,20	9940203,00
	MAX	10769602,00	64243472,00	13553396,00	10697133,70
	D.P.	1757668,00	17662328,00	896029,30	552155,80
	C.V.(%)	33	74	7	6
V16	MIN	1985388,00	7437533,00	11782654,20	9037860,00
	$\overline{x}$	5421486,00	26178236,00	2763614,80	10703993,00
	Md.	5351160,00	16544266,00	12587807,80	10575369,00
	MAX	10787213,00	77490057,00	14000749,00	12069495,00
	D.P.	1899886,00	21252997,00	742720,30	1036181,00
	C.V.(%)	35	81	6	10
V17	MIN	2309146,00	6477675,00	11859455,00	9522307,00
	$\overline{x}$	5748216,00	29528918,00	13114836,00	10938680,00
	Md.	5487164,00	17971452,00	13267954,00	10473488,00
	MAX	11191283,00	84522332,00	14659055,00	13023108,00
	D.P.	1861184,00	24163563,00	776707,00	1194828,00
	C.V.(%)	32	82	6	11
V18	MIN	12396,00	56763,94	111138,20	107504,90
	$\overline{x}$	92231,21	3081889,34	479532,30	231387,60
	Md.	77741,76	821515,08	531402,30	180652,50
	MAX	322639,22	25620800,51	798600,20	590362,30
	D.P.	56487,22	5231346,47	252460,70	137701,30
	C.V.(%)	61	170	53	60

Gráfico 1. BoxPlot de V1 (Despesas Municipais por Função de Saúde e Saneamento), por grupo.

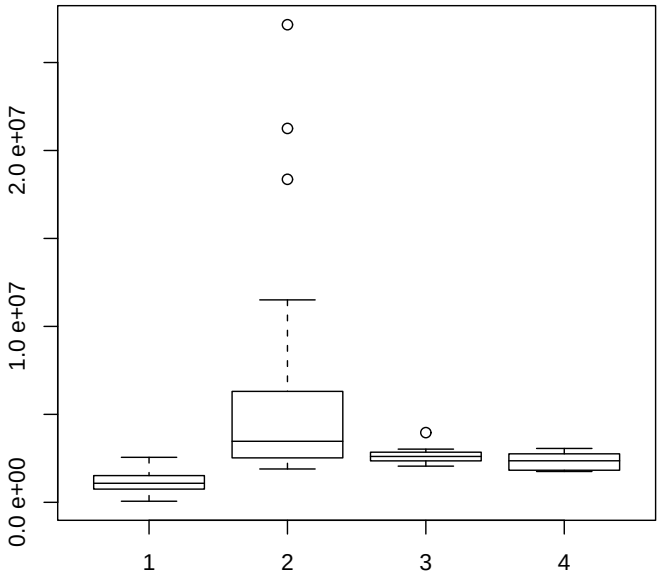


Gráfico 2. BoxPlot de V4 (Investimento Municipal), por grupo.

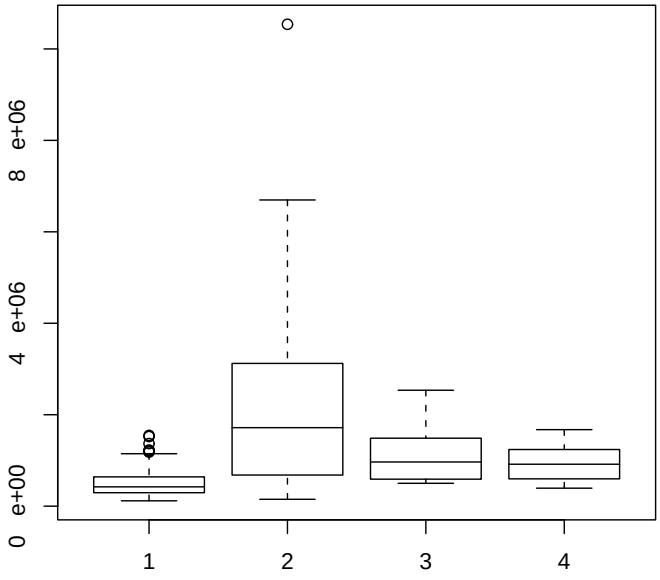


Gráfico 3. BoxPlot de V5 (Impostos Municipais), por grupo.

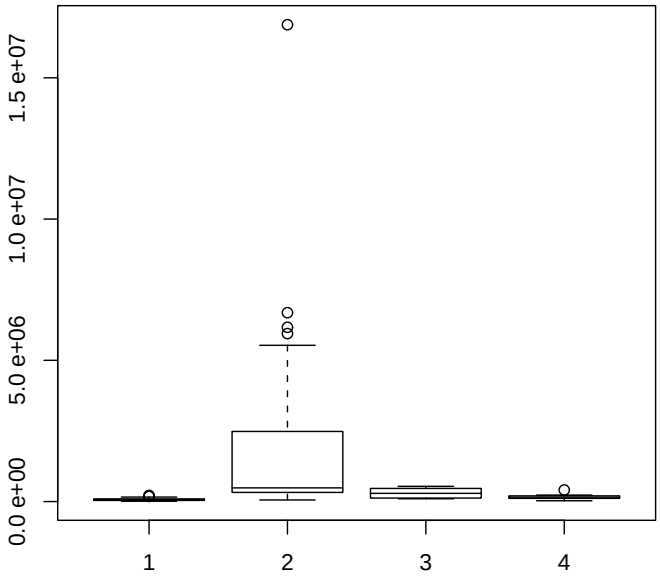


Gráfico 4. BoxPlot de V7 (ISS), por grupo.

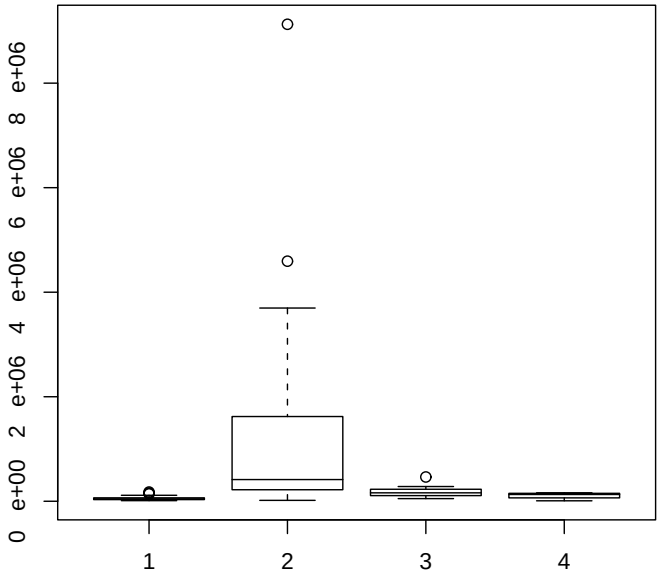


Gráfico 5. BoxPlot de V9 (Transferências para os Municípios Referentes ICMS), por grupo.

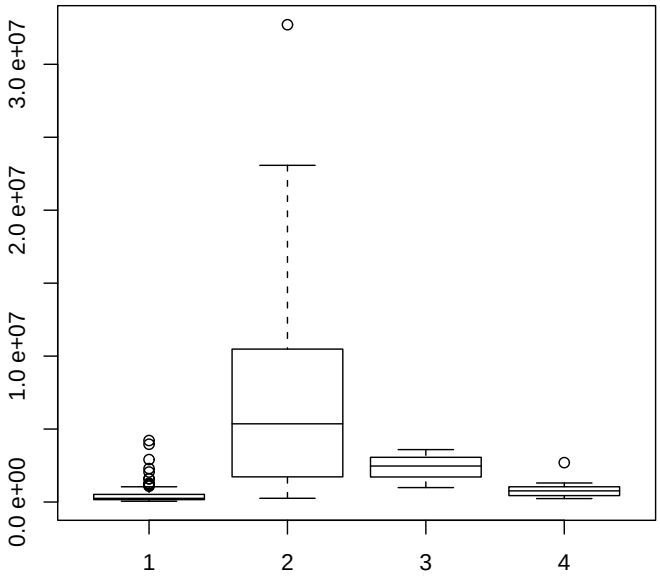


Gráfico 6. BoxPlot de V11 (Transferências Correntes de Tributos Estaduais para os municípios), por grupo.

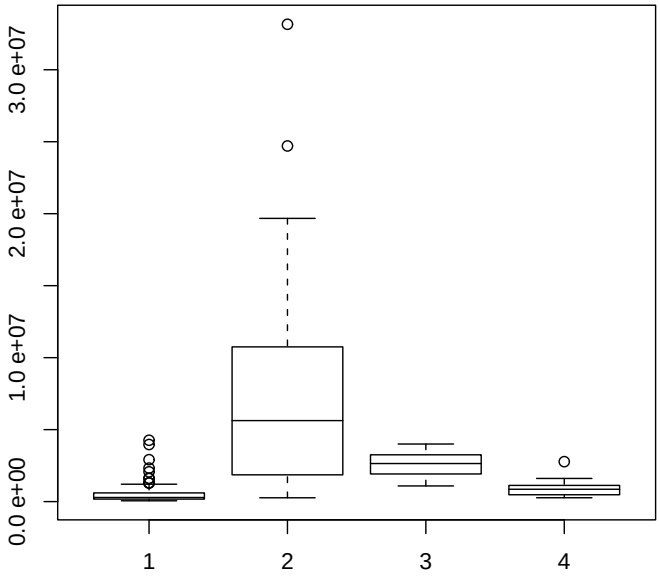


Gráfico 7. BoxPlot de V12 (Receitas Correntes Municipais), por grupo.

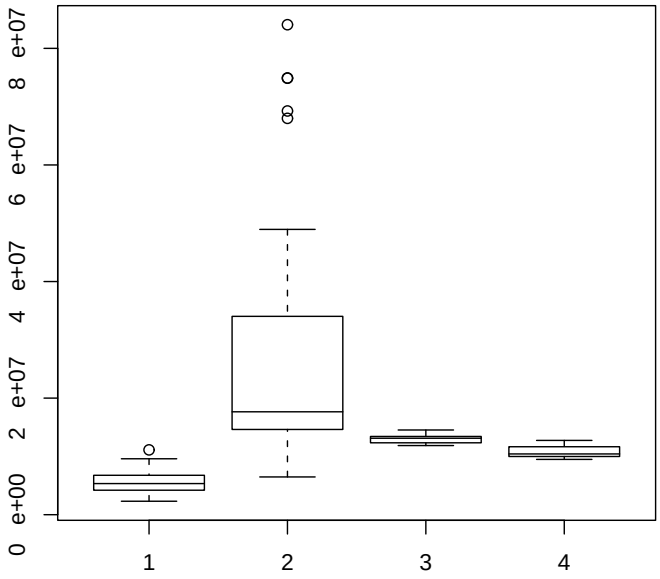


Gráfico 8. BoxPlot de V14 (Outras Transferências Correntes para os municípios), por grupo.

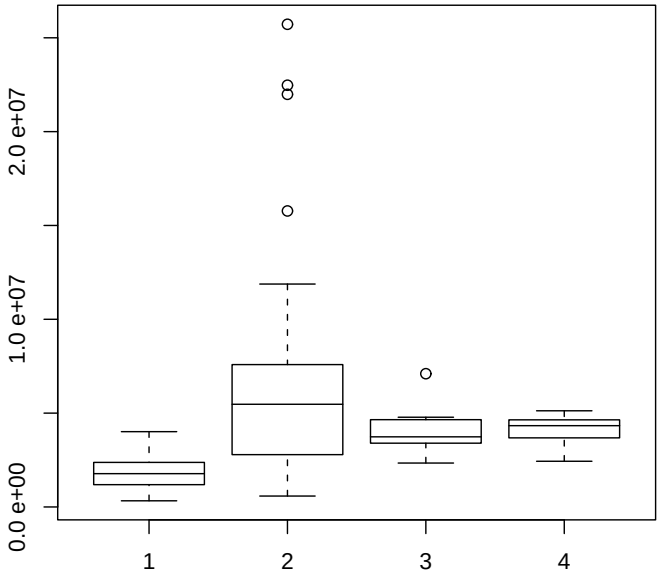


Gráfico 9. BoxPlot de V15 (Receita Municipal com Transferências Correntes), por grupo.

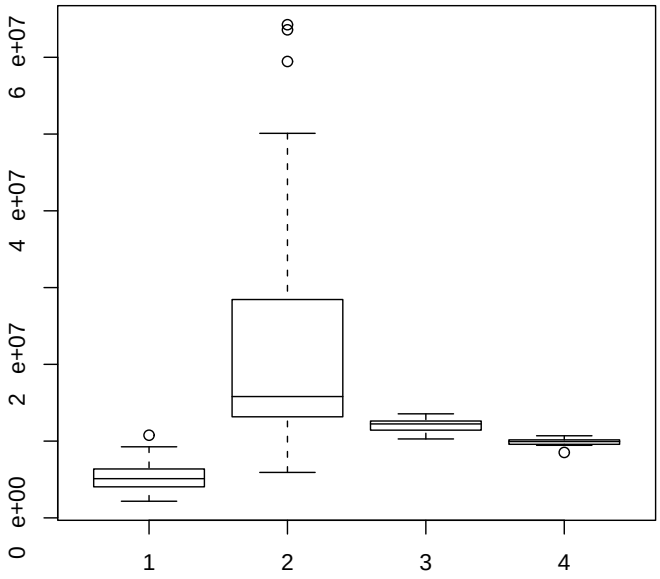


Gráfico 10. BoxPlot de V16 (Totais de Despesas Municipais por Função), por grupo.

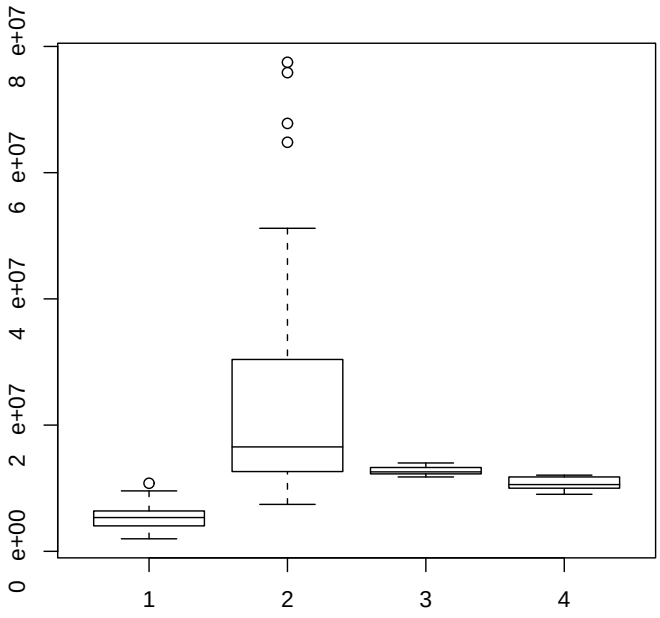


Gráfico 11. BoxPlot de V17 (Receita Orçamentária Municipal), por grupo.

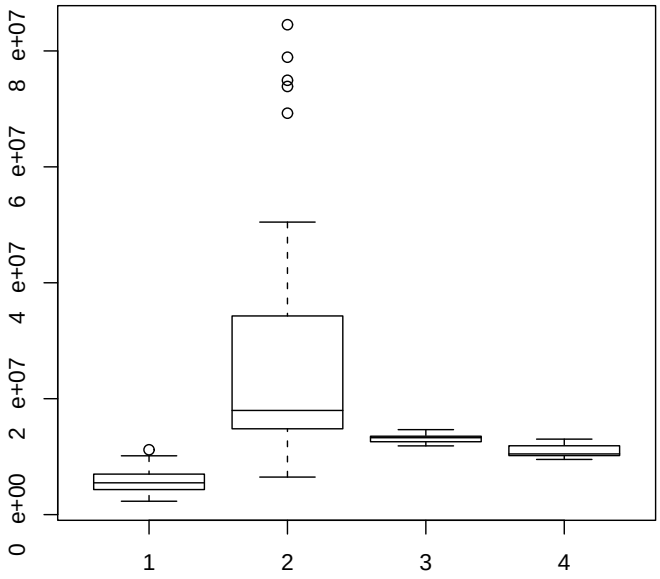
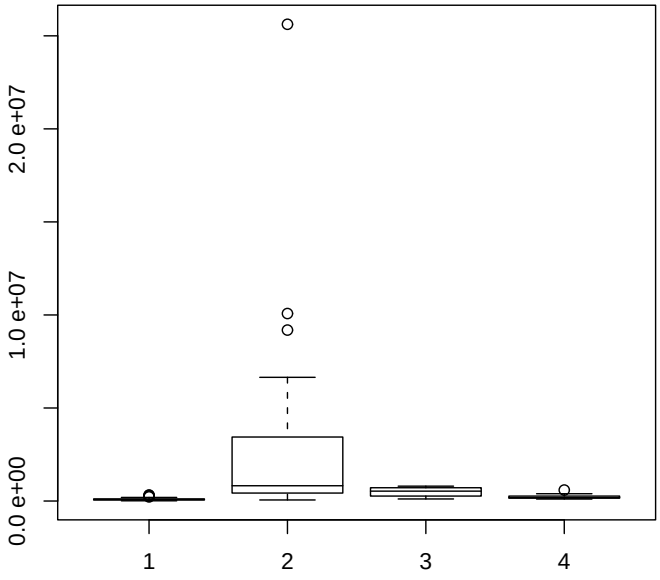


Gráfico 12. BoxPlot de V18 (Receita Tributária Municipal), por grupo.



4.2 Resultado do Agrupamento com o Banco de Dados do SUS

Realizada a análise exploratória para este banco, ver Quadro 9, verificamos que nenhuma das variáveis apresenta dados faltantes, no entanto para este banco também existe uma grande dispersão relativa nas variáveis, especialmente em V23 e V26, cujos coeficientes de variação são 153% e 181% respectivamente.

Na maioria dos agrupamentos realizados observou-se que a inclusão da variável 24 (Despesa de Recursos Próprios) acarreta o agrupameno de todos os municípios em um só grupo. Por este motivo esta variável foi retirada e todos os agrupamentos refeitos. Considerando as 3 situações já mencionadas: DBCP, DBSP e QMP, foram testados todos os modelos existentes no Quadro 3, já referido variando a quantidade de grupos de 1 a 9. Para cada situação, os melhores modelos foram aqueles que obtiveram maior BIC. O resumo destes resultados é encontrado no Quadro 10. Para DBCP os dados foram agregados em 7 grupos, com VEI como melhor modelo, já retirando a população os grupos foram reduzidos a cinco tendo como modelo VEE, este número de grupos se mantem com os dados percapita, no entanto o melhor modelo agora é o EEE. Comparando os valores de incerteza verifica-se que a menor incerteza refere-se a situação em que foram utilizados os dados brutos sem população, tal como ocorreu no agrupamento com os dados do IPEA. Este agrupamento em especial é apresentado de forma ilustrativa no Mapa 2. Também é apresentada no Quadro 11 e nos Gráficos 13 a 16, uma descrição das variáveis dos cinco grupos resultantes deste agrupamento.

Notemos que nos grupos 1 e 2 se concentram a maioria dos municípios. O grupo 3 é o que contém apenas três municípios: Cabo de Santo Agostinho, Caruaru e Paulista. No grupo 4 encontram-se Bezerros, Camaragibe, Ipojuca, Olinda e Petrolina. E no grupo 5: Garanhuns, Goiana, Igarassu e Vitória de Santo Antão . Analisando os box-plot verifica-se que no grupo 1 se concentraram as cidades que possuem os menores valores para todas as variáveis, apresentando pouca dispersão em todas as variáveis visto que o coeficiente de variação é relativamente baixo, atingindo no máximo 55% para a variável V27. Nos gráficos 13, 14, 15 e 16 são mostrados os municípios com valores extremos nas variáveis. Na variável V23 no grupo 2 se observam valores extremos em Petrolândia, Pesqueira, Belo Jardim. Na variável V25 não existe nenhum. Na V26 no grupo 2 temos Belo Jardim, Gravatá, Moreno, Pesqueira. Na V26 também no grupo 2 Belo Jardim e Pesqueira. Por último na V27 temos valores extremos nos grupos 2 com Belo Jardim e Pesqueira, e no grupo 4 com Olinda.

Os grupos 1, 2, e 5 , do ponto de vista da grandeza absoluta das variáveis podem ser considerados ordenados e com maior homogeneidade que os outros dois grupos. No grupo 4 é notada uma dispersão maior das variáveis. Este grupo quase não apresenta valores extremos,

enquanto o grupo 2 apresenta “outliers” em quase todas as variáveis.

Quadro 9. Estatísticas básicas do banco de dados do SUS para as variáveis de interesse.

Variáveis	Medidas						
	MIN	$\overline{x}$	Md	MAX	D.P.	C.V.(%)	NA's
V23	374432,50	2454921,70	1606725,50	27061697,60	3766902,80	153	0
V24	-863488,50	1019651,70	677026,50	9668601,00	1283994,60	126	0
V25	1851422,00	7176914,00	4280536,00	58778965,00	9222992,00	128	0
V26	126750,60	1435270,00	822010,20	20791694,80	2601470,70	181	0
V27	19942,20	1014424,10	609100,00	9907205,20	1386923,90	137	0

Quadro 10. Resumo dos agrupamentos feitos com o banco de dados do DATASUS.

Variáveis	Grupos	Elementos por grupo	Melhor Modelo	Incerteza	BIC
DBCP	7	33-41-12-14-13-34-6	VEI	0,48	-21401,74
DBSP	5	83-58-3-5-4	VEE	0,45	-18218,64
QMP	5	24-14-112-2-1	EEE	0,57	-5554,88

Quadro 11. Estatísticas básicas para as variáveis V23, V25, V26 e V27, dos cinco grupos gerados no agrupamento feito com os dados brutos sem a população .

Variáveis	Medidas	GRUPOS				
		1	2	3	4	5
V23	MIN	374432,50	889164,30	11854257,00	5185099,00	3239346,00
	$\overline{x}$	1088290,10	2439917,60	14353090,00	16921611,00	4023100,00
	Md.	964137,40	2402400,00	15193892,00	20136225,00	3934632,00
	MAX	2122563,00	5338360,70	16011122,00	27061698,00	4983787,00
	D.V	433576,10	842784,60	2202292,00	10648344,00	746114,00
	C.V.(%)	40	34	15	63	19
V25	MIN	1851422,00	2971203,00	43411967,00	7904743,00	19418695,40
	$\overline{x}$	3462793,00	7132513,00	50659385,00	33232209,00	19707775,20
	Md.	3359050,00	6360685,00	49787224,00	33250604,00	19689772,20
	MAX	5905157,00	14455033,00	58778965,00	56842386,00	20032861,00
	D.P.	1046496,00	2870301,00	7720534,00	18782105,00	285151,40
	C.V.(%)	30	40	15	57	1
V26	MIN	144742,50	126750,60	7244652,00	2733264,00	1542384,50
	$\overline{x}$	568417,90	1345009,40	8606083,00	11712061,00	2507129,90
	Md.	468910,10	1308360,50	9028973,00	14143734,00	2635597,80
	MAX	1295459,30	3027239,30	9544624,00	20791695,00	3214939,30
	D.P.	290176,80	630975,70	1206895,00	7825231,00	700899,10
	C.V.(%)	51	47	14	67	28
V27	MIN	19942,20	138982,10	3135593,00	1991020,00	2061062,00
	$\overline{x}$	503281,20	1028864,80	6619307,00	4776660,00	2504794,50
	Md.	429913,70	901430,10	6815123,00	4196088,00	2358190,90
	MAX	1131755,80	3081784,60	9907205,00	9869278,00	3241734,10
	D.P.	277966,40	698764,40	3390050,00	3116378,00	510931,40
	C.V.(%)	55	68	51	65	20

Gráfico 13. BoxPlot de V23 (Despesas Total de Saúde), por grupo.

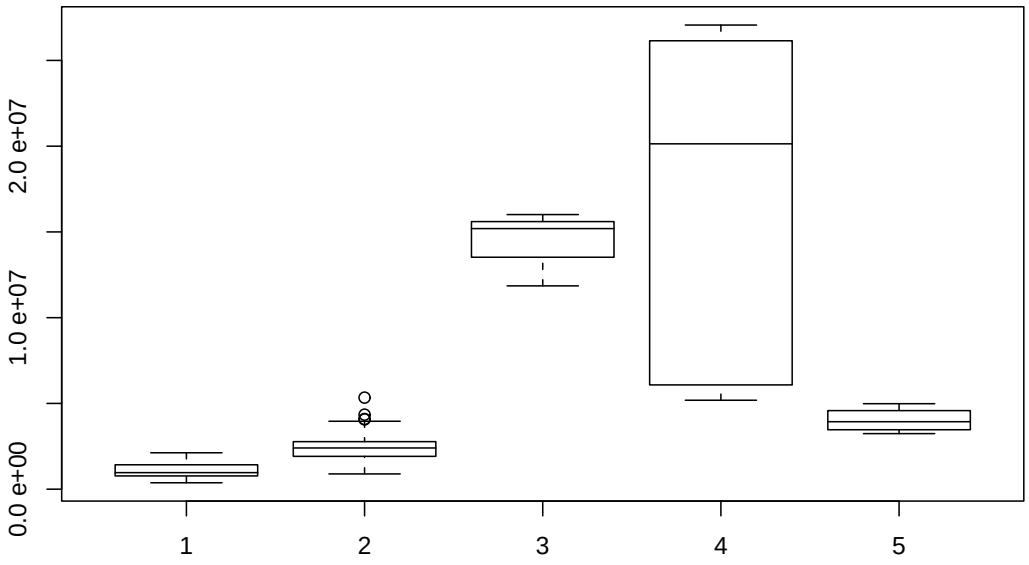


Gráfico 14. BoxPlot de V25 (Receita de Impostos e Transferências Constitucionais e Legais), por grupo.

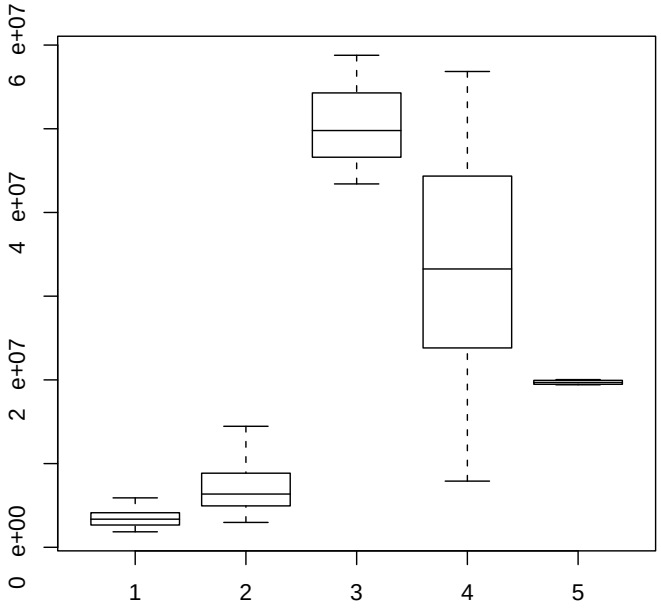


Gráfico 15. BoxPlot de V26 (Transferências SUS), por grupo.

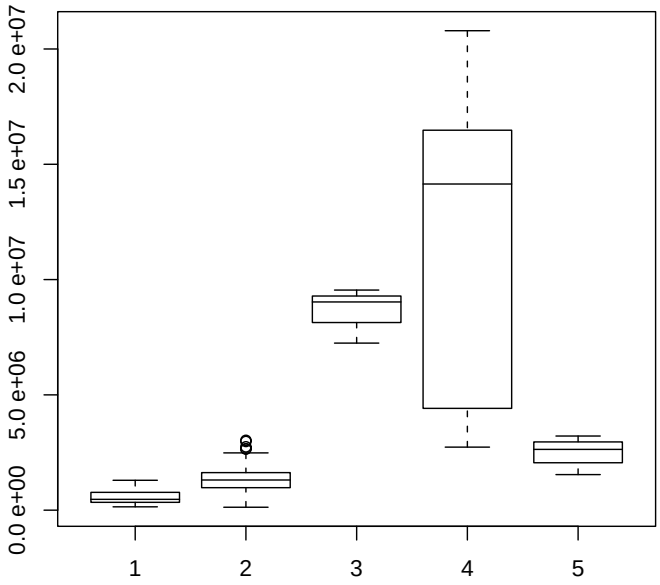
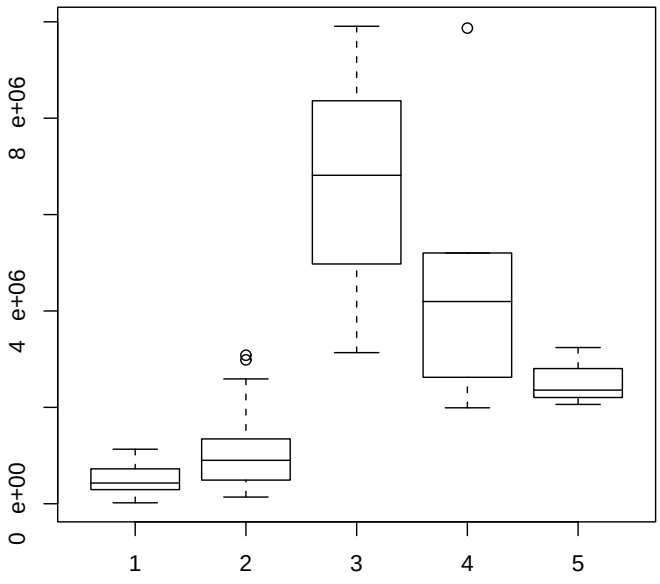


Gráfico 16. BoxPlot de V27 (Despesa Pessoal), por grupo.



4.3 Conclusões Gerais dos Agrupamentos

Para ter presente o tamanho dos municípios em termos da quantidade de habitantes a ordem decrescente dos 10 maiores municípios estudados em relação a população é: Olinda (372014), Paulista (268282), Caruaru (258176), Petrolina (225199), Cabo de Santo Agostinho (156004), Camaragibe (132215), Garanhuns (119336), Vitória de Santo Antão (118894), Igarassu (83424) e Goiana (71940). Considerando a variável despesas municipais por função legislativa em ordem decrescente se observaram os municípios de: Paulista (4655272,67), Cabo de Santo Agostinho (4421681,98), Petrolina (3013918,79), Caruaru (2831987,38) e Olinda (2398801,76).

Ao comparar os dois agrupamentos verificamos que os grupos 3, 4 e 5 formados com o banco de dados do SUS estão contidos no grupo 2 para o banco de dados do IPEA. Este grupo possui os maiores coeficientes de variação para quase todas as variáveis. Comportamento semelhante apresenta o grupo 4 dos dados do SUS. O que é interessante perceber é que os valores extremos do grupo 2 do IPEA fazem parte dos grupos 3 e 4, já descritos, do SUS. O grupo 4 do SUS apresenta valor extremo apenas para o município de Olinda na variável V27 (Despesa Pessoal).

No gráfico 16 observamos que o grupo 3 apresenta na variável V27 coeficiente de variação superior aos demais. Isto se deve porque os municípios que fazem parte deste grupo: Cabo de Santo Agostinho, Caruaru e Paulista possuem Despesa de Pessoal bem distintas.

Cabo de Santo Agostinho é um município que apresenta valores extremos em 8 variáveis analizadas com o banco de IPEA, apresentando os maiores valores extremos para as variáveis V9 (Transferências correntes de tributos estaduais para os municípios), V12 (Receitas correntes municipais), V14 (Outras transferências correntes para os municípios), V15 (Receita municipal com transferências correntes), V16 (Total de despesas municipais por função), V17 (Receita orçamentária municipal) e sendo o menor valor extremo dos valores observados para V1 (Despesas municipais por saúde e saneamento). Note que este município encontra-se no grupo 3 para as variáveis consideradas com o banco de dados do SUS, que é o grupo com a maior Despesa com Pessoal.

Camaragibe destaca-se por ser o município com maior valor de V1 (Despesa municipal por função de saúde e saneamento).

Caruaru é um município com valores extremos para V5 (Impostos municipais), V12 (Receitas correntes municipais), V14 (Outras receitas correntes para os municípios), V16 (Total de despesas municipais), V17 (Receita orçamentária municipal).

Petrolina é o único município com valor extremo na Variável Investimentos municipais, apresenta-se também com valor extremo em V1 (Despesas municipais por funções de saúde

e saneamento), V5 (Impostos municipais), V7 (ISS), V12 (Receita corrente municipal), V14 (Outras transferências correntes para os municípios), V15 e V16.

Paulista apresenta valor extremo para as variáveis V11, V12, V14, V15, V16, V17, V18 e forma parte do Grupo 3 para as variáveis do SUS.

Olinda é o maior valor extremo dos valores observados de V5 (Impostos municipais) e V7 (ISS), e V18 (Receita tributária municipal) e de V27 (Despesas com Pessoal), sendo equivalente a Caruaru.

Notemos que Ipojuca e Bezerros municípios menores em termos de quantidade de habitantes (ordem 18 e 19, respectivamente) aparecem no mesmo grupo (grupo 4) que municípios maiores como Olinda.

Belo Jardim e Pesqueira (municípios na ordem 14 e 20 em termos de número de habitantes) e aparecem com valores extremos na variável V23 (Despesa total de saúde), V26 (Transferência do SUS) e V27 (Despesa com Pessoal).

De uma forma geral, analisando comparativamente os resultados dos agrupamentos pode-se destacar os seguintes pontos:

• A incerteza calculada com o banco de dados do IPEA é sempre menor do que aquela calculada com os dados do SUS muito provavelmente devido ao fato de que o banco de dados do IPEA contem mais variáveis do que o do SUS.

• Tanto para o banco do IPEA quanto para o banco do SUS o agrupamento realizado com as variáveis monetárias per capita (QMP), reproduziram a maior incerteza muito provavelmente pelo fato de que ao diminuir a variância de todas as variâveis os municípios ficaram "mais parecidos" ficando mais difícil a classificação.

• Para os agrupamentos realizados com DBCP e DBSP a incerteza com os dados do IPEA é cerca da metade da incerteza calculada com os dados do SUS.

Apêndice A

Saídas referentes ao banco de dados do IPEA.

• Dados brutos com a população como variável.

```
> dados<-matrix(scan("C:/Usuarios/PatriciaLeal/grupos1/dados_ipea.txt", 0), ncol=13,
               byrow=TRUE)
Read 1989 items
> bic<-EMclust(dados)
> bic

 BIC:
        EII       VII       EEI       VEI       EVI       VVI       EEE
1 -68570.78 -68570.78 -65028.92 -65028.92 -65028.92 -65028.92 -57840.20
2 -65761.11 -64076.50 -62763.48 -60193.18 -63334.20 -59734.40 -57786.49
3 -64766.82 -62214.64 -62383.99 -58479.13 -62235.23 -58219.40 -57701.11
4 -64062.85 -61263.33 -62454.06 -57830.99 -62361.55 -57654.25 -57899.13
5 -63979.71 -60723.81 -61941.64 -57539.70 -61901.57 -57386.58 -57968.60
6 -63960.51 -60506.01 -61982.95 -57290.59 -62035.98 -57139.55 -58014.02
7 -64000.47 -60090.23 -62052.54 -57214.11 -62183.15 -57149.25 -58085.22
8 -64068.84 -60081.47 -62119.48 -57139.50 -62362.75 -57028.27 -57684.40
9 -64140.81 -59983.33 -62192.62 -57209.66 -62498.55 -57011.03 -57686.96
        EEV       VEV      VVV
1 -57840.20 -57840.20 -57840.2
2 -58245.97 -58005.72       NA
3 -57226.12 -55230.39       NA
4 -57418.16 -55084.10       NA
5 -57678.94 -55522.59       NA
6 -58032.77 -55778.58       NA
7 -58280.74 -56580.38       NA
8 -58267.04 -56555.56       NA
9 -58929.86 -56895.02       NA

> sum<-summary(bic,dados)
> sum

classification table:

 1  2  3  4
28 12 52 61

uncertainty (quartiles):
          0%          25%          50%          75%         100%
0.000000e+00 2.131939e-11 1.517217e-08 4.854766e-05 1.993063e-01

best BIC values:
    VEV,4     VEV,3     VVI,9
-55084.10 -55230.39 -57011.03

best model: ellipsoidal, equal shape
```

```
> sum$classification
  [1] 4 3 1 4 4 3 2 4 4 3 1 2 3 4 1 4 1 4 2 4 2 3 4 3 1 1 4 4 4 3 1 4 4 3 3 3 1 3
      2 4 4 3 4 4 4 4 4 4 4 3 3 3 1 3 1 4 1 3 1 4 3
 [61] 1 3 3 3 1 4 4 4 1 4 3 3 4 4 4 3 4 4 3 4 4 3 4 4 2 4 3 4 3 4 1 4 1 3 3 4 1 3
      4 3 3 4 2 1 4 1 1 1 3 1 4 4 3 4 4 3 3 1 3 3
[121] 3 3 2 3 1 3 2 4 4 4 4 3 1 3 4 2 3 2 4 4 3 1 4 3 3 4 3 3 4 3 2 1 4
```

• Dados divididos pela população .

```
> dadosd<-dados/dados[,13]
> dadosd<-dadosd[,-13]
> bic<-EMclust(dadosd)
> bic

 BIC:
        EII       VII       EEI       VEI       EVI       VVI       EEE
1 -22013.26 -22013.26 -19511.78 -19511.78 -19511.78 -19511.78 -15207.51
2 -20372.47 -19733.35 -18756.33 -17571.87 -18685.51 -17513.64 -15101.87
3 -19363.20 -18866.17 -18522.69 -17063.54 -18030.75 -17001.38 -15072.70
4 -18773.11 -18409.22 -18496.78 -16815.63 -17875.01 -16763.91 -15058.35
5 -18496.48 -18095.95 -18366.04 -16720.62 -17823.65 -16634.88 -15064.54
6 -18483.55 -18066.76 -18235.69 -16575.05 -17788.92 -16534.75 -15045.07
7 -18131.88 -17724.08 -18281.84 -16594.99 -17819.95 -16460.37 -15047.92
8 -18093.90 -17680.93 -18347.49 -16407.39 -17914.02 -16334.77 -15048.05
9 -18051.99 -17658.91 -18342.51 -16288.68 -17938.03 -16341.94 -15094.51
        EEV       VEV       VVV
1 -15207.51 -15207.51 -15207.51
2 -15123.27 -15046.51        NA
3 -15349.83 -15092.62        NA
4 -15463.83 -15084.62        NA
5 -15718.47 -15555.85        NA
6 -15941.66 -15787.67        NA
7 -15933.80 -15754.57        NA
8 -16285.74 -15926.90        NA
9 -16564.14 -16170.43        NA

> sum<-summary(bic,dadosd)
> sum

classification table:

  1   2   3   4   5   6
123   7  11   4   4   4

uncertainty (quartiles):
          0%          25%          50%          75%         100%
0.000000e+00 1.877472e-06 1.317952e-05 3.518385e-04 4.867641e-01

best BIC values:
    EEE,6     EEE,7     VEV,2
-15045.07 -15047.92 -15046.51

best model: elliposidal, equal variance
```

```
> sum$classification
  [1] 1 1 1 1 1 1 1 1 1 1 1 2 2 1 2 1 1 1 1 1 1 1 1 3 1 1 1 1 2 1 1 1 1 1 1 1
      1 1 2 1 1 1 1 1 3 1 1 1 1 3 1 3 1 1 2 1 4 1
 [61] 1 5 1 1 6 6 1 1 6 1 1 1 3 1 1 1 1 1 1 4 1 3 1 1 1 1 1 1 1 3 1 5 1 5 1 1 1
      1 5 1 1 1 1 1 1 6 3 1 3 1 1 1 1 1 1 4 1 1 1
[121] 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1 4 1 1 1 1 1 1 1 1 1 1 2 1 1 3 1
```

• Dados brutos sem população .

```
> dados<-dados[,-13]
> bic<-EMclust(dados)
> bic

 BIC:
        EII       VII       EEI       VEI       EVI       VVI       EEE
1 -63443.45 -63443.45 -61287.82 -61287.82 -61287.82 -61287.82 -54804.30
2 -60854.66 -59309.42 -59179.11 -56753.52 -59661.70 -56346.54 -54748.06
3 -59948.58 -57598.79 -58838.21 -55189.82 -58381.77 -54929.90 -54514.80
4 -59303.48 -56753.40 -58481.23 -54610.55 -58192.15 -54450.72 -54730.96
5 -59233.77 -56286.34 -58475.36 -54335.17 -57752.77 -54292.51 -54638.42
6 -59219.92 -55865.84 -58540.52 -54139.93 -57871.57 -54042.50 -54699.48
7 -59284.83 -55680.60 -58582.27 -54064.40 -58712.78 -54086.33 -54478.65
8 -59324.86 -55542.45 -58647.23 -53976.25 -58838.68 -54083.71 -54880.27
9 -59389.56 -55494.04 -58713.59 -53991.63 -58247.23 -54005.57 -54523.94
        EEV       VEV      VVV
1 -54804.30 -54804.30 -54804.3
2 -55097.21 -54946.47       NA
3 -53903.20 -55083.34       NA
4 -54340.95 -52286.60       NA
5 -54509.18 -52414.95       NA
6 -54712.86 -52596.94       NA
7 -54769.54 -52845.96       NA
8 -55055.19 -53090.37       NA
9 -55085.98 -53258.90       NA

> sum<-summary(bic,dados)
> sum

classification table:

  1   2   3   4
101  28  12  12

uncertainty (quartiles):
          0%          25%          50%          75%         100%
0.000000e+00 5.644374e-13 1.242073e-11 8.278280e-10 1.849429e-01

best BIC values:
    VEV,4     VEV,5     EEV,3
-52286.60 -52414.95 -53903.20

best model: ellipsoidal, equal shape
```

```
> sum$classification
  [1] 1 1 2 3 4 1 3 1 1 1 2 3 1 1 2 1 2 4 4 4 3 1 4 1 4 2 1 1 1 1 2 1 1 1 1 1 2 1
      3 1 4 1 1 1 1 1 1 1 1 1 1 2 1 2 1 2 1 2 2 1
 [61] 2 1 1 1 2 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 2 2 2 1 1 2 2 1
      4 1 1 1 3 2 1 2 2 2 1 2 1 1 1 1 1 1 1 4 1 1
[121] 1 1 3 1 3 1 4 4 1 1 1 1 2 1 1 3 1 3 1 1 1 3 1 1 1 1 1 1 1 1 4 2 1
```

Apêndice B

Saídas referentes ao banco de dados do SUS.

• Dados brutos com a população como variável.

```
> dados<-matrix(scan("C:/Usuarios/PatriciaLeal/grupos1/datasus
               /dados_datasus.txt", 0), ncol=6, byrow=TRUE)
Read 918 items
> dados<-dados[,-3]
> bic<-EMclust(dados)
> bic

 BIC:
        EII       VII       EEI       VEI       EVI       VVI       EEE
1 -25684.11 -25684.11 -23907.66 -23907.66 -23907.66 -23907.66 -22529.85
2 -24635.11 -23872.75 -22915.19 -22254.16 -22908.67 -22267.09 -22374.95
3 -24402.05 -23319.91 -22771.45 -21718.08 -22776.54 -21735.75 -22316.09
4 -24290.53 -22993.06 -22754.57 -21528.75 -22800.15 -21590.58 -22091.49
5 -24319.84 -22951.25 -22776.17 -21446.38 -22844.42 -21496.04 -22099.96
6 -24339.25 -22974.76 -22806.34 -21457.33 -22899.50 -21529.98 -22130.16
7 -24369.45 -22873.44 -22836.53 -21401.74 -22954.94 -21546.71 -22160.34
8 -24399.63 -22887.82 -22866.72 -21414.26 -23010.14 -21567.11 -22190.53
9 -24429.84 -22737.76 -22896.92 -21426.14 -23065.36 -21564.41 -22167.80
        EEV       VEV       VVV
1 -22529.85 -22529.85 -22529.85
2 -22255.37 -22185.94        NA
3 -22319.84 -22195.92        NA
4 -22344.83 -22234.27        NA
5 -22137.57 -22284.25        NA
6 -22011.38 -22317.31        NA
7 -22043.19 -22327.79        NA
8 -22199.58 -22378.31        NA
9 -22235.79 -22407.41        NA

> sum<-summary(bic,dados)
> sum

classification table:

 1  2  3  4  5  6  7
33 41 12 14 13 34  6

uncertainty (quartiles):
         0%         25%         50%         75%        100%
0.000000000 0.000150767 0.002772467 0.029394500 0.480085993

best BIC values:
    VEI,7     VEI,8     VVI,5
-21401.74 -21414.26 -21496.04

best model: diagonal, equal shape
```

```
> sum$classification
  [1] 2 6 2 3 2 1 3 2 2 1 4 4 1 6 5 6 5 2 3 2 3 1 5 1 4 7 2 6 2 1 7 6 2 6 6 1
      7 6 4 1 2 2 6 2 2 2 6 2 2 1 2 4 6 5 2 5 1 5 2 2
 [61] 5 1 1 1 5 6 3 6 4 6 6 6 6 2 6 1 1 1 1 2 2 6 3 2 4 2 6 6 6 6 5 3 7 2 6 4
      4 1 2 1 2 2 4 7 6 5 5 7 1 2 1 2 1 2 2 1 1 4 6 6
[121] 1 1 5 1 3 1 3 3 2 6 2 6 4 2 2 3 1 4 6 2 1 4 2 6 6 6 6 1 1 2 3 5 1
>
```

• Dados sem a população .

```
> dadospop<-dados[,-2]
> bic<-EMclust(dadospop)
> bic

 BIC:
        EII       VII       EEI       VEI       EVI       VVI       EEE
1 -20598.61 -20598.61 -18831.02 -18831.02 -18831.02 -18831.02 -18118.32
2 -19770.19 -19156.43 -18105.93 -17570.00 -18097.78 -17581.00 -17899.37
3 -19485.86 -18731.86 -18024.08 -17192.58 -17921.98 -17216.01 -17777.47
4 -19362.82 -18686.00 -17846.08 -17053.70 -17939.26 -17214.74 -17772.85
5 -19217.10 -18671.41 -17791.05        NA -17869.84 -17190.07 -17798.04
6 -19214.89 -18431.00 -17801.10        NA -17911.49 -17108.47 -17804.84
7 -19239.98 -18398.21 -17825.35        NA -17937.56 -17212.25 -17803.07
8 -19248.30 -18397.72 -17850.35 -17083.94 -17982.03 -17173.03 -17828.26
9 -19273.93 -18413.97 -17867.88 -17072.97 -18027.27 -17198.56 -17853.41
        EEV       VEV       VVV
1 -18118.32 -18118.32 -18118.32
2 -17963.08 -17151.77        NA
3 -17750.45 -17736.04        NA
4 -17754.45 -17183.92        NA
5 -17525.45 -17010.92        NA
6 -17452.06 -17191.82        NA
7 -17360.39 -17197.35        NA
8 -17378.67 -17226.29        NA
9 -17425.44 -17248.46        NA

> sum<-summary(bic,dadospop)
> sum

classification table:

 1  2  3  4  5
83 58  3  5  4

uncertainty (quartiles):
          0%          25%          50%          75%         100%
0.0000000000 0.0003670508 0.0025044388 0.0198050353 0.4505122688

best BIC values:
    VEV,5     VEV,6     VEI,6
-18218.64 -18246.17 -18251.34

best model: ellipsoidal, equal shape
```

```
> sum$classification
  [1] 2 2 2 2 2 1 2 1 1 1 2 2 1 1 2 1 4 2 2 2 2 1 2 1 2 3 2 1 2 1 4 1 1 2 1 1
      3 1 2 1 2 1 1 1 1 2 1 1 1 1 2 2 1 5 1 5 1 2 2 2
 [61] 5 1 1 1 4 2 2 1 2 1 1 1 1 1 2 1 1 1 1 2 2 1 2 1 2 1 1 1 1 2 2 2 4 1 1 2
      2 1 1 1 2 1 2 3 1 2 2 4 1 1 1 1 1 1 2 1 1 2 1 1
[121] 1 1 2 1 2 1 2 2 2 1 2 1 2 1 1 2 1 2 1 2 1 2 1 1 1 1 1 1 1 2 2 5 1
>
```

• Dados divididos pela população .

```
>  dadosd<-dados/dados[,1]
>  dadosd<-dados[,-1]
>  bic<-EMclust(dadosd)
> bic

 BIC:
       EII       VII       EEI       VEI       EVI       VVI       EEE
1 -20684.84 -20684.84 -20166.55 -20166.55 -20166.55 -20166.55 -19108.50
2 -19858.01 -19271.37 -19365.89 -18821.61 -19355.10 -18832.50 -18900.93
3 -19760.02 -18850.39 -19199.51 -18437.25 -19257.98 -18454.73 -18742.84
4 -19490.89 -18808.21 -19012.30 -18456.15 -19246.50 -18448.80 -18689.82
5 -19308.93 -18807.55 -19000.86 -18312.89 -19127.25 -18474.60 -18682.70
6 -19304.89 -18570.59 -18988.86 -18251.34 -19133.61 -18406.95 -18696.02
7 -19321.12 -18576.82 -19011.83 -18283.33 -19152.64 -18436.95 -18716.06
8 -19343.36 -18550.40 -19024.51 -18303.68 -19191.27 -18449.50 -18715.24
9 -19367.88 -18552.67 -19049.71 -18312.83 -19244.24 -18417.59 -18711.96
       EEV       VEV       VVV
1 -19108.50 -19108.50 -19108.50
2 -18787.85 -18344.30 -18328.35
3 -18700.13 -18377.88        NA
4 -18588.52 -18386.25        NA
5 -18591.63 -18218.64        NA
6 -18591.76 -18246.17        NA
7 -18778.44 -18282.93        NA
8 -18593.21 -18330.49        NA
9 -18602.77 -18373.81        NA

>  sum<-summary(bic,dadosd)
>  sum

classification table:

  1   2   3   4   5
 24  14 112   2   1

uncertainty (quartiles):
          0%          25%          50%          75%         100%
0.000000e+00 3.206902e-05 4.005178e-04 5.704239e-03 5.746276e-01

best BIC values:
    EEE,5     EEE,9     VVV,2
-5554.879 -5580.390 -5589.363
```

```
best model: elliposidal, equal variance

> sum$classification
  [1] 3 1 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 1 3 3 3 1 3 3 3 2 1 3 3 3 4 3 3 1 3 1
      3 3 1 2 1 3 2 3 3 3 3 3 3 3 1 2 3 3 3 3 1 3 2 3
 [61] 3 3 3 2 2 5 3 3 2 3 3 3 2 3 1 1 3 3 3 3 4 3 1 3 3 3 3 1 1 1 3 3 3 3 3 3
      3 3 3 1 1 3 3 3 3 3 2 1 3 3 3 3 3 3 2 3 1 3 3 3
[121] 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 2 3 3 1 1 3 3 3 3 2 3 3 3 1 3 3 3
```

Apêndice C

• Municípios que não fizeram parte do agrupamento.

Abreu e Lima
Araçoiaba
Barreiros
Belém de São Francisco
Brejinho
Calumbi
Camutanga
Carpina
Cedro
Condado
Escada
Fernando de Noronha
Gameleira
Ibirajuba
Ilha de Itamaracá
Ipubi
Itaíba
Jaboatão dos Guararapes
Manari
Recife
Ribeirão
Santa Terezinha
São Benedito do Sul
São José da Coroa Grande
São José do Egito
São Loureno da Mata
Tabira
Tacaimbó
Tamandaré
Terra Nova
Tracunhaém
Trindade

• Municípios que fizeram parte do agrupamento e seus respectivos grupos para o banco do IPEA e do SUS.

IPEA	SUS	Municípios
2	2	Afogados da Ingazeira
2	2	Afrânio
2	2	Agrestina
2	2	Água Preta
2	2	Águas Belas
4	1	Alagoinha
3	2	Aliana
4	1	Altinho
2	1	Amaraji
4	1	Angelim
1	2	Araripina
1	2	Arcoverde

4	1	Barra de Guabiraba
5	1	Belém de Maria
1	2	Belo Jardim
2	1	Betânia
6	4	Bezerros
2	2	Bodocó
3	2	Bom Conselho
2	2	Bom Jardim
2	2	Bonito
4	1	Brejão
2	2	Brejo da Madre de Deus
4	1	Buenos Aires
3	2	Buíque
1	3	Cabo de Santo Agostinho
2	2	Cabrobó
4	1	Cachoeirinha
4	2	Caetés
4	1	Calçado
1	4	Camaragibe
2	1	Camocim de São Félix
2	1	Canhotinho
2	2	Capoeiras
4	1	Carnaíba
4	1	Carnaubeira da Penha
1	3	Caruaru
4	1	Casinhas
2	2	Catende
4	1	Chã de Alegria
2	2	Chã Grande
2	1	Correntes
4	1	Cortês
4	1	Cumaru
2	1	Cupira
2	2	Custódia
2	1	Dormentes
4	1	Exu
4	1	Feira Nova
4	1	Ferreiros
4	2	Flores
6	2	Floresta
4	1	Frei Miguelinho
1	5	Garanhuns
4	1	Glória do Goitá
1	5	Goiana
4	1	Granito
1	2	Gravatá
1	2	Iati
4	2	Ibimirim
1	5	Igarassu
2	1	Iguaraci
4	1	Inajá
4	1	Ingazeira
1	4	Ipojuca

4	2	Itacuruba
4	2	Itambé
2	1	Itapetim
3	2	Itapissuma
4	1	Itaquitinga
4	1	Jaqueira
4	1	Jataúba
5	1	Jatobá
4	1	João Alfredo
4	2	Joaquim Nabuco
4	1	Jucati
4	1	Jupi
2	1	Jurema
4	1	Lagoa do Carro
2	2	Lagoa do Itaenga
2	2	Lagoa do Ouro
4	1	Lagoa dos Gatos
5	2	Lagoa Grande
2	1	Lajedo
1	2	Limoeiro
2	1	Macaparana
4	1	Machados
4	1	Maraial
4	1	Mirandiba
4	2	Moreilândia
6	2	Moreno
2	2	Nazaré da Mata
1	4	Olinda
4	1	Orobó
2	1	Orocó
6	2	Ouricuri
1	2	Palmares
4	1	Palmeirina
2	1	Panelas
2	1	Paranatama
4	2	Parnamirim
2	1	Passira
1	2	Paudalho
1	3	Paulista
4	1	Pedra
6	2	Pesqueira
1	2	Petrolândia
1	4	Petrolina
4	1	Poção
1	1	Pombos
4	1	Primavera
4	1	Quipapá
4	1	Quixaba
2	1	Riacho das Almas
2	2	Rio Formoso
4	1	Sairé
2	1	Salgadinho
1	2	Salgueiro

4	1	Saloá
4	1	Sanharó
4	1	Santa Cruz
4	1	Santa Cruz da Baixa Verde
1	2	Santa Cruz do Capibaribe
4	1	Santa Filomena
1	2	Santa Maria da Boa Vista
4	1	Santa Maria do Cambucá
2	2	São Bento do Una
2	2	São Caitano
4	2	São João
2	1	São Joaquim do Monte
4	2	São José do Belmonte
4	1	São Vicente Ferrer
1	2	Serra Talhada
5	1	Serrita
2	1	Sertânia
2	2	Sirinhaém
2	1	Solidão
3	2	Surubim
2	1	Tacaratu
4	2	Taquaritinga do Norte
4	1	Terezinha
6	2	Timbaúba
2	1	Toritama
4	1	Triunfo
4	1	Tupanatinga
4	1	Tuparetama
4	1	Venturosa
4	1	Verdejante
4	1	Vertente do Lério
4	2	Vertentes
2	2	Vicência
1	5	Vitória de Santo Antão
4	1	Xexéu

REFERÊNCIAS

[1] Allard, D. and Fraley, C.(1997). Nonparametric maximum likelihood estimation of features in spatial point processes using Voronoï tessellation. *Journal of the American Statistical Association,* **92**, 1485-1493. (disponível como relatório técnico no. 293R em *http://www.stat.washington.edu/www/research/reports*).

[2] Banfield, J. D. and A. E. Raftery (1992). Ice floe identification in satellite images using mathematical morphology and clustering about principle curves. *Journal of the American Statistical Association*, **87**, 7-16.

[3] Banfield, J.D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**:803-821.

[4] Bensmail, H. and G. Celeux, A. E. Raftery, and C.P. Robert (1997). Inference in model-based cluster analysis. *Statistics and Computing*, **7**, 1-10.

[5] Biernacki, C., Celeux, G., Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixtures. Computational Statistics & Data Analysis, **41**

[6] Bock, H. H. (1996). Probabilistic models in cluster analysis. *Computational Statistics and Data Analysis*, **23**, 5-28.

[7] Bock, H. H. (1998a). Probabilistic approaches in cluster analysis. *Bulletin of the International Statistical Institute*, **57**, 603-606.

[8] Bock, H. H. (1998b). Probabilistic aspects in classification. In C. Hayashi, K. Yajima, H. H. Bock, N. Oshumi, Y. Tanaka, and Y. Baba (Eds.), *Data science, classification and related methods,* pp. 3-21. Springer-Verlag.

[9] Boyles, R. A. (1983). On the convergence of the EM algorithm. *Journal of the Royal Statistical Society,* Series B, **45**, 47-50.

[10] Byers, S. D. and A. E. Raftery (1998). Nearest neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association* **93**, 577-584.

[11] Campbell, J. G., C. Fraley, F. Murtagh, and A. E. Raftery (1997). Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters*, **18**, 1539-1548.

[12] Celeux, G. and Govaert (1993). Comparison of the mixture and the classification maximum likelihood in cluster amalysis. *Journal of Statistical Computation and Simulation*, **47**, 127-146.

[13] Cormack, R. M. (1971). A Review of Classifications. JRSS, A, 134, 321-367.

[14] Cribari-Neto, F. & Zarkos, S.G. (1999). `R:` yet another econometric programming environment. *Journal of Applied Econometrics,* 14, 319–329.

[15] Dasgupta, A. and A. E. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* **93**, 294-474.

[16] Dempster, A.P.,N.M. Laird, and D.B.Rubin (1977). Maximum likelihood for incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Series B, **39**, 1-38

[17] Fayyad, U. and Smyth (1996). From massive data sets to science catalogs: applications

and challenges. In J. Kettenring and D. Pregibon (Eds.), *Statistics and Massive Data Sets: Report to the Committee on Applied and Theoretical Statistics.* National Research Council.

[18] Fraley, C. and A. E. Raftery (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, **41**, 578-588.

[19] Fraley, C. and Raftery, A. E., (2002a). MCLUST: Software for Model-Based Clustering, Density Estimation and Discriminant Analysis. Technical Report 415, Department of Statistics, University of Washington.

[20] Fraley, C. and Raftery, A. E., (2002b). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association,* **97**, 611-631.

[21] Friedman, H.P. and J.Rubin (1967). One some invariant criteria for grouping data. *Journal of the American Statistical Association*, **62**, 1159-1178.

[22] Giampaoli, V. and Singer, J. (2004). Bayes factor for comparing the mean diastolic pressure of hypertense individuals. A ser publicado no *Journal of Data Science.*

[23] Hastie, T. and W. Stuetzle (1989). Principal curves. *Journal of the American Statistical Association*, **84**, 502-516.

[24] Holman, E. W. (1985), Evolutionary and psychological effects in pre-evolutionary classifications, *J. Classification*, **2**, 29-39.

[25] Karlis, D., Xekalaki, E. (2003), Choosing initial values for the EM algorithm for finite mixtures. Special issue on mixtures. *Computational Statistics & Data Analysis,* **41**

[26] Kass, R. E. and Raftery, A. E. (1995). Bayes Factor. *Journal of the American Statistical Association*, **90** ; 773-795.

[27] Kaufman. L. and Rousseeuw P.J. (1990). Finding Groups in Data. An Introduction to Cluster Analysis. Wiley-Interscience. New York.

[28] McLachlan, G. J. and K. E. Basford (1988). *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker

[29] McLachlan, G. J. and T. Krishnan (1997). The EM Algorithm and Extensions. Wiley.

[30] McLachlan, G. J., Peel, D., Bean, R. W. (2003). Modelling high-dimensional data by mixtures of factor analyzers. Special issue on mixtures. *Comput. Statist. Data Anal.*, **41**

[31] Mukherjee, S., E. D. Feigelson, G. J. Babu, F. Murtagh, C. Fraley, and A. E. Raftery (1998). Three types of gamma ray bursts. *The Astrophysical Journal*, **508**, 314-327.

[32] Peel, D. and McLachlan, G. J. (1999). User's Guinde to EMMIX: Version 1.3 *http://www.maths.uq.edu.au/ gim/emmix/emmix.html.*

[33] Peel, D. and G. J. McLachlan (2000). Robust mixture modeling using the t-distribution. *Statistics and Computing* (to appear).

[34] Roeder, R. A., and Walker, H. F. (1984), "Mixture Densities Maximum Likelihood and the EM Algorithm", *SIAM* Review, *26*, 195-239.

[35] Romesburg, H. C. (1984). Cluster Analysis for Researches. *Lifetime Learning Publications* California.

[36] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals os Statistics*, **6**, 461-464.

[37] Scott, A.J. and M.J.Symons (1971). Clustering methods based on likelihood ratio criteria. *Biometrics.* **27**, 387-397.

[38] Seild, W., Mosler, K., Alker, M. (2000). A cautionary note on likelihood ratio tests in mixture models. *Ann. Inst. Statist. Math.* **52**, 481-487.9

[39] Stanford, D. and A. E. Raftery (2000). Principal curve clustering with noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 601-609.

[40] Venables, W.N. & Ripley, B.D. (2002). *Modern Applied Statistics with* `S`, $4^{\underline{a}}$ ed. New York: Springer-Verlag.

[41] Ward, J. H. (1963). Hierarchical groupings to optimize an objective function. *Journal of American Statistical Association*, **58**, 234-244.

[42] Wu, C. F. J. (1983). On convergence properties of the EM algorithm. *The Annals of Statistics* **11**, 95-103.