UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

PARTIÇÃO DE GRUPOS E VALIDAÇÃO DA ANÁLISE DE AGRUPAMENTO PARA EQUIPAMENTOS DE FISCALIZAÇÃO ELETRÔNICA DE TRÂNSITO

CURITIBA
JUNHO / 2008

ADRIANO SCHON MAXIMILIANO

MARCOS TADEU ANDRADE CORDEIRO

PARTIÇÃO DE GRUPOS E VALIDAÇÃO DA ANÁLISE DE AGRUPAMENTO PARA EQUIPAMENTOS DE FISCALIZAÇÃO ELETRÔNICA DE TRÂNSITO

Trabalho de Conclusão de Curso apresentado à disciplina de Laboratório de Estatística II do Curso de Estatística do Departamento de Estatística do Setor de Ciências Exatas da Universidade Federal do Paraná, sob orientação do Prof$^{o.}$ Dr. Bruno Grimaldo Martinho Churata

CURITIBA

JUNHO / 2008

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

Este trabalho apresenta uma análise multivariada de agrupamento (clusters) aplicada a equipamentos de fiscalização eletrônica de trânsito. Um banco de dados com características destes equipamentos foi utilizado, sendo este constituído de variáveis categóricas e numéricas. Foi empregado o algoritmo de agrupamento não hierárquico PAM, que é capaz de trabalhar com a matriz de dissimilaridades. A matriz de similaridade foi calculada usando o coeficiente de Gower, que trabalha com variáveis categóricas e numéricas simultaneamente, e posteriormente transformada em uma matriz de dissimilaridade através de uma relação apropriada. Para obter o número de *cluster* na população, foi usado o gráfico da silhueta. No intuito de avaliar as técnicas utilizadas, PAM e gráfico da silhueta, foi gerada uma população simulada, onde se aplicou a técnica proposta. A população simulada foi utilizada para avaliar uma forma proposta de encontrar o número de clusters ao qual população foi dividida. A análise de agrupamento, aplicada aos equipamentos, obteve 21 *clusters*.

Palavras-Chave: Métodos de Particionamento, Gower, PAM, Gráfico da Silhueta.

1. INTRODUÇÃO

O trânsito no Brasil é uma das principais causas de mortes, segundo a Organização Mundial de Saúde (OMS), sendo que os acidentes mais graves geralmente são associados á imprudência no trânsito. Na tentativa de diminuir os números desse problema em 23 de setembro de 1997 foi sancionada a lei n° 9.503 que instituiu o Código de Trânsito Brasileiro (CTB), o qual é considerado internacionalmente como um dos mais evoluídos do mundo. O novo CTB regulamentou a utilização de instrumentos de fiscalização eletrônica de trânsito, entendendo que estes equipamentos são necessários na aplicação da lei com um maior rigor. Um dos primeiros instrumentos deste tipo utilizado foi o Redutor Eletrônico de Velocidade (R.E.V.), conhecido popularmente como “Lombada Eletrônica”. Para que este tipo de equipamento tenha um funcionamento de qualidade, os departamentos de trânsito contratam empresas que oferecem a tecnologia e manutenção destes equipamentos.

Depois de dez anos estes equipamentos receberam novas funcionalidades auxiliando os departamentos de trânsito no gerenciamento de trafego, o que nos últimos anos fez o volume de dados gerados por estes equipamentos aumentarem muito, além disto, estes equipamentos tornaram-se muito singulares, pois cada equipamento é customizado, muitas vezes, de acordo com as necessidades de cada cliente.

A metodologia aplicada foi a análise multivariada de agrupamento (*clusters*) por particionamento utilizando uma medida de distância para dados misturados (numéricos e categóricos), além de apresentar uma metodologia para avaliar a qualidade do agrupamento, permitindo encontrar o número natural de grupos (*clusters*) realmente presente em que os equipamentos serão alocados.

1.1 OBJETIVO

Realizar uma análise de agrupamento (cluster), com dados de equipamentos de fiscalização eletrônica de trânsito, contendo variáveis categóricas e numéricas, utilizando métodos de agrupamento não hierárquico. A metodologia utilizada também foi avaliada através de simulação.

1.3 JUSTIFICATIVA

Devido à heterogeneidade destes equipamentos, muitas comparações entre eles tornam-se cada dia mais difícil, dificultando assim a melhoria destes equipamentos, pois não se consegue verificar se um componente instalado em um determinado equipamento é melhor do que o instalado em outro.

A heterogeneidade também causa transtornos ao tentar-se modelar o volume de informação gerada por estes equipamentos. Por isto neste trabalho foca-se na separação destes equipamentos em grupos, tentando assim separar em grupos homogêneos uma população heterogênea de uma maneira eficiente.

1.4 ESTRUTURA DO TRABALHO

Além desta introdução, esta monografia dispõe no segundo capítulo de uma revisão de literatura, onde são descritos basicamente métodos de agrupamento, medidas de parecença e o gráfico da silhueta. No terceiro capítulo estão descritos os materiais utilizados e métodos aplicados. No quarto capítulo são apresentados e discutidos os resultados obtidos. Ao final, tem-se a conclusão, as referências e um apêndice.

2. REVISÃO DE LITERATURA

2.1. ANÁLISE DE AGRUPAMENTOS (CLUSTER)

A análise de agrupamento (*clusters*) busca padrões em um conjunto de dados agrupando-os em *clusters* de forma automática. Cada *cluster* é composto por indivíduos parecidos, porém não há similaridade entre *clusters* diferentes, em outras palavras tem-se homogeneidade dentro dos *clusters* e heterogeneidade entre eles. Ao se fazer uma análise de *clusters* tem-se a esperança de encontrar o agrupamento natural dos objetos, para que os *clusters* formados façam sentido ao pesquisador.

2.2. DISTÂNCIAS E COEFICIENTES DE SIMILARIDADES

Para criar os *clusters,* é necessário calcular alguma medida de parecença, estas medidas podem ser apenas distâncias ou uma correlação entres as observações ou ainda uma ponderação destas medidas.

2.3. MEDIDAS DE DISTÂNCIA (DISSIMILARIDADES)

Na maioria dos casos a medida de parecença é feita pela medida de distância entre as observações e é aconselhável utilizar distâncias verdadeiras, ou seja, que estas medidas satisfaçam as seguintes propriedades:

I) $d(P,Q) = d(Q,P)$;
II) $d(P,Q) > 0$ se $P \neq Q$;
III) $d(P,Q) = 0$ se $P = Q$;
IV) $d(P,Q) \leq d(P,R) + d(R,Q)$;

Onde;

R é um ponto intermediário;
I é a Simetria;
II e III é a Positividade;
IV é a desigualdade triangular.

Por outro lado alguns algoritmos de agrupamentos aceitam distâncias que não satisfação todas estas propriedades, por exemplo, que não satisfaça a desigualdade triangular.

Quando os objetos da análise não podem ser representados por medidas p-dimensionais então estes são comparados em relação à ausência ou presença de determinada característica e logicamente itens semelhantes estão mais próximos que os não semelhantes (CHAVES NETO, 2007).

Existem inúmeras medidas de distâncias que poderão ser utilizadas, para agrupar itens, porém as mais usuais estão dispostas na TABELA 1.

TABELA 1 – MEDIDAS DE DISTÂNCIA (DISSIMILARIDADE)

Nome	Expressão	Explicação
Distância Euclidiana	$d(\underline{x},\underline{y}) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$	É a mais comum, simplesmente é a distância geométrica (intuitiva) em um espaço p-dimensional.
Distância Euclidiana Quadrática	$d(\underline{x},\underline{y}) = \sum_{j=1}^{p}(x_j - y_j)^2$	É similar a distância euclidiana, porém atribui um peso maior a distancias entre objetos mais distantes.
Distância de Mahalanobis (Estatística)	$d(\underline{x},\underline{y}) = \sqrt{(\underline{x}-\underline{y})^T S^{-1}(\underline{x}-\underline{y})}$	S^{-1} é a inversa da matriz de covariância. Contudo sem conhecimento dos grupos distintos está matriz não pode ser calculada, portanto é preferível utilizar a distância euclidiana.
Manhattan (City-block)	$d(\underline{x},\underline{y}) = \left[\sum_{i=1}^{p}\lvert x_i - y_i\rvert\right]$	A distância de Manhattan é a distância média entre as dimensões. Na maioria dos casos a distância encontrada é similar a distância Euclidiana, entretanto, o efeito de grandes diferenças são suavizados
Chebchev	$d(\underline{x},\underline{y}) = \max(\lvert x_i - y_i\rvert)$,i = 1,...,p	A distância de Chebchev é apropriada quando o objetivo é definir dois elementos como diferentes, onde apenas uma das dimensões difere.
Minkowski	$d(\underline{x},\underline{y}) = \left[\sum_{i=1}^{p}\lvert x_i - y_i\rvert^m\right]^{\frac{1}{m}}$	A distância de Minkowski é uma generalização das demais distâncias, pois as distâncias são basicamente normas de vetores. Por exemplo, para m = 2 temos a expressão da distância euclidiana.

Fonte: Os autores.

2.4. COEFICIENTES DE SIMILARIDADES

Os coeficientes de similaridades surgiram para tratar diferenciadamente distâncias entre indivíduos quando a variável de comparação é categórica, por exemplo, a variável é dada pela presença (1) ou ausência (0) de uma determinada característica, sendo assim, os pares possíveis de comparações são (1,0), (0,1), (1,1) e (0,0), nota-se que as distâncias entre (1,0) e (0,1) são iguais e as distâncias dos empates (1,1) e (0,0) também, porém julgando apenas os empates (1,1) e (0,0) percebe-se que suas distâncias não devem ser iguais. Para evidenciar analisa-se o seguinte exemplo:

Se 1 significa "*lê grego antigo*" e 0 significa "*não lê grego antigo*", é óbvio que o empate (1,1) mostra que estes indivíduos são mais parecidos que os empates (0,0), (CHAVES NETO, 2007).

Para resolver este problema surgiram os coeficientes de similaridades, que atribuem pesos maiores para as distâncias dos empates (1,1) e pesos menores ou até mesmo desconsideram as distâncias dos empates (0,0), para introduzir esta metodologia são apresentados resultados de coincidência e divergência dos objetos h e i na TABELA 2:

TABELA 2 – TABELA DE CONTINGÊNCIA PARA O CÁLCULO DOS COEFICIENTES.

		Item I		Totais
		1	0	
Item h	1	a	b	a + b
	0	c	d	c + d
Totais		a + c	b + d	p = a + b + c + d

Fonte: Adaptação de Johnson & Wichern (1999).

Onde;

a = Freqüência de igualdades 1-1;

b = Freqüência de desigualdades 1-0;

c = Freqüência de desigualdades 0-1;

d = Freqüência de igualdades 0-0;

As informações da TABELA 2 possibilitam a construção de muitos coeficientes. Vários deles estão propostos na literatura e alguns estão apresentados na TABELA 3.

TABELA 3 – COEFICIENTES USUAIS DE SIMILARIDADE

Nome	Expressão	Explicação	Variação
Coincidência Simples	$\frac{a+d}{p}$	Pesos iguais para 1-1 e 0-0	$(0,1)$
Sokal e Sneath	$\frac{2(a+d)}{2(a+d)+b+c}$	Peso Duplo 1-1 e 0-0	$(0,1)$
Rogers e Tanimoto	$\frac{a+d}{a+2(b+c)+d}$	Duplo peso para pares não coincidentes	$(0,1)$
Russel e Rao	$\frac{a}{p}$	Nenhum 0-0 no numerador	$(0,1)$
Jaccard	$\frac{a}{a+b+c}$	As coincidências 0-0 são tratadas como irrelevantes	$(0,1)$
Sorenson	$\frac{2a}{2a+b+c}$	0-0 é irrelevante e duplo peso para 1-1	$(0,1)$
Distancia Binária de Sokal	$\sqrt{\frac{b+c}{p}}$	Única medida de dissimilaridade	$(0,1)$
Ochiai	$\frac{a}{\sqrt{(a+b)(a+c)}}$	Concordâncias positivas sobre adaptação da média geométrica de discordâncias	$(0,1)$
Baroni-Urbani-Buser	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$	Coincidência positiva sobre adaptação da média geométrica de concordância positiva e negativa	$(0,1)$
Haman	$\frac{(a+d)-(b+c)}{p}$	Proporção de coincidências menos a proporção de discordâncias	$(-1,1)$
Yule	$\frac{ad-bc}{ad+bc}$	Proporção de ad menos a de bc.	$(-1,1)$
Φ	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	Produto de momento de correlação aplicado a variáveis binárias	$(-1,1)$
Ochiai II	$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	Proporção de coincidências em relação à média geométrica total modificada	$(0,1)$

Fonte: Adaptação de Ferreira (1996).

COEFICIENTE DE GOWER

Outra medida de parecença foi proposta por Gower (1971). Esta medida trata-se de um coeficiente geral de similaridade que permite a combinação simultânea de variáveis categóricas e numéricas. Este coeficiente é aplicável aos mais diversos tipos de variáveis: binárias, numéricas, nominais (multiníveis) e ordinais (multiníveis). Este índice é calculado pela expressão (2.1):

$$S_{ij} = \frac{1}{p}\sum_{i=1}^{p} S_i \qquad (2.1)$$

Para dados categóricos (qualitativos), temos:

- S_{ij} = similaridade entre o objeto i e o objeto j.
- S_i = 1, quando há uma concordância entre os dados.
- S_i = 0, quando não há uma concordância dos dados.
- p = número de variáveis.

A semelhança entre as duas amostras é a média dos valores de similaridade para as *p* variáveis.

As variáveis quantitativas (numéricas) são tratadas de maneira diferente. Inicialmente é calculado o desvio entre os valores de duas amostras, $|y_{i1}-y_{i2}|$, este valor é então dividido pelo desvio máximo ($\left|\max(x_{.k}) - \min(x_{.k})\right|$) que pode ser calculado para esta variável, utilizando a amostra disponível. Assim, a proporção é de fato uma distância padrão e subtraindo-a de um encontra-se uma semelhança, este calculo é feito pela expressão (2.2):

$$S_i = 1 - \left[\left|\frac{y_{ik} - y_{jk}}{\max(x_{.k}) - \min(x_{.k})}\right|\right] \qquad (2.2)$$

Onde:

- y_{ik} = valor da *k-ésima* variável para o objeto i.
- y_{jk} = valor da *k-ésima* variável para o objeto j.
- $\max(x_{.k})$ = valor máximo da *k-ésima* variável.
- $\min(x_{.k})$ = valor mínimo da *k-ésima* variável.

Ao coeficiente de Gower foi introduzido um elemento w_i (Delta de Kronecker) de flexibilidade. A comparação não é feita para as variáveis em que existe uma falta de informação para uma das duas amostras. O w_i funciona como uma espécie de “função indicadora”, informando a presença ou ausência de determinada informação, que assume valor 0 (zero) se não há nenhuma informação que pertence a variável *i* de uma das duas amostras, e o valor de 1 (um) quando a informação este presente para ambas as amostras. Neste caso, o coeficiente toma a forma (2.3) mostrada abaixo:

$$S_{ij} = \frac{\sum_{k=1}^{p} W_k S_k}{\sum_{k=1}^{p} W_k} \qquad (2.3)$$

Onde:

- w_k = 1, quando se tem os valores da k-ésima variável para ambas amostras;
- w_k = 0, quando não se tem os valores da k-ésima variável para quaisquer das duas amostras.
- S_{ij} = similaridade entre os objetos i e j.
- p = número total de variáveis.

Segundo CUADRAS (1989) [2] para converter uma medida de similaridade em uma distância, utilizando (2.3), temos que:

$$\delta_{ij} = 1 - s_{ij} \qquad (2.4)$$

Onde:

δ_{ij} = distância entre *i* e *j*

s_{ij} = similaridade entre *i* e *j*

Na literatura existem muitas outras distâncias, porém as mais usuais estão contidas neste tópico. Cabe ao pesquisador decidir de acordo com seus dados qual é a melhor distância a ser utilizada.

2.5. AGRUPAMENTO HIERÁRQUICO

Os métodos que utilizam um agrupamento hierárquico são divididos em dois, aglomerativos e divisivos.

Os métodos aglomerativos começam inicialmente com tantos *clusters* quanto indivíduos, contendo apenas um individuo, une-se um *cluster* a outro formando outros *clusters*. Esta união entre *clusters* acontece ao relaxar-se o grau de parecença entre os indivíduos de um e de outro *cluster*, isto ocorre sucessivamente até que exista apenas um *cluster*.

Quando se utiliza os métodos divisivos, é tomado o sentido contrário ao método aglomerativo, ou seja, inicialmente se tem um único *cluster* contendo todos os objetos e este *cluster* se subdivide em *sub-clusters* contendo itens mais parecidos dentro destes e uma grande diferença entre os objetos de *clusters* distintos, isto ocorre até que se tenha tantos *clusters* quanto objetos.

Neste trabalho apresenta-se apenas o método hierárquico alglomerativo. A explicação de um destes é suficiente para o entendimento do outro.

A forma de representar os métodos hierárquicos são gráficos denominados dendogramas. Os dendogramas apresentam de forma visual os *clusters* e as fusões que ocorrem até a formação de um único *cluster*.

O procedimento a ser utilizado quando se trabalha com o método alglomerativo hierárquico é o seguinte:

1. Iniciando com o mesmo numero de *clusters* quanto indivíduos calcula-se a matriz de similaridade $D\ nxn$.

$$D = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \vdots & \vdots & \dots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{nn} \end{bmatrix}_{n \times n}$$

2. Na matriz D acha-se o par de *clusters* mais próximos e juntando-os em um novo *cluster*.

3. Calcula-se a distância entre os outros *clusters* e o novo *cluster* formado, ligando o novo grupo de acordo com algum método de ligação.
4. Repete-se os passos 2 e 3 até (n-1) vezes, observando-se a identidade dos clusters que estão sendo formados e as distâncias entre estes clusters.

2.6. LIGAÇÕES

Para ligar um grupo a outro no intuito de formar os dendogramas através do agrupamento aglomerativos hierárquico, são utilizadas as ligações e algumas destas ligações estão dispostas na TABELA 4.

TABELA 4 – MÉTODOS DE LIGAÇÃO.

Nome	Expressão	Explicação
Ligação Simples (Vizinho mais próximo)	$d_{(AB)C} = \min\{d_{AC}, d_{BC}\}$	Esta ligação é feita através da menor distância entre os objetos dos grupos a serem ligados.
Ligação Completa (Vizinho mais distante)	$d_{(AB)C} = \max\{d_{AC}, d_{BC}\}$	A ligação dos grupos é feita pela maior distância entre os elementos de dois grupos a serem unidos, isto garante que todos os elementos de um grupo estão dentro de alguma distância máxima
Ligação Média	$d_{(AB)C} = \frac{\sum_i \sum_k d_{ik}}{n_{(AB)} n_C}$	Esta ligação é feita pela distância média entre dois pares de objeto.
Método de Ward	$ESS_k = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$	Este método busca minimizar a soma dos quadrados de quaisquer dois hipotéticos *clusters* que podem ser formados em cada passo.

Fonte: os autores.

Existem outros métodos de ligação, mas não serão abrangidos neste trabalho por serem menos utilizados.

2.7. DENDOGRAMA

O dendograma é a representação gráfica resultante da aplicação do método hierárquico. É através deste gráfico que se pode encontrar quais são e em quantos agrupamentos o conjunto de dados está dividido. Um exemplo é o gráfico apresentado na FIGURA 1.

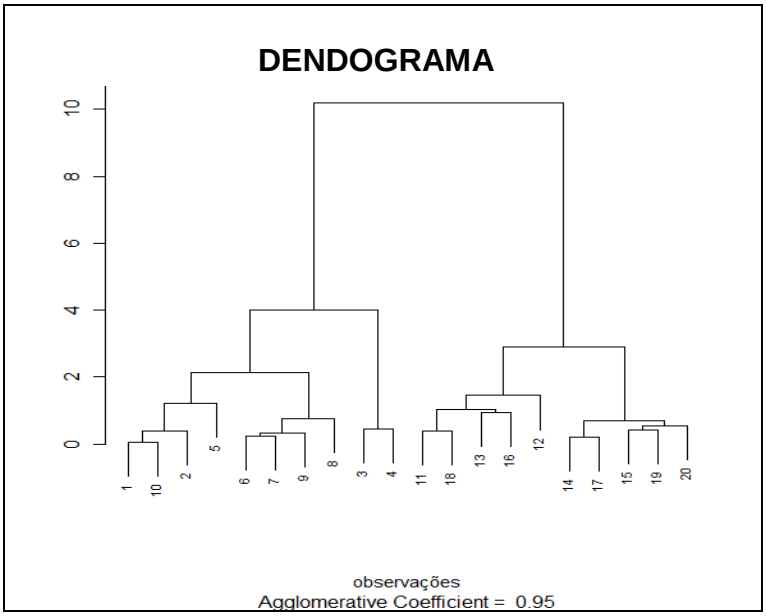

FIGURA 1 - EXEMPLO DE DENDOGRAMA.

Fonte: os autores.

No eixo vertical está representado em que distância cada grupo está unido e no eixo horizontal estão contidos os itens do banco de dados.

2.8. AGRUPAMENTO NÃO HIERARQUICO (PARTICIONAL)

Os métodos *não hierárquicos*, também conhecidos como *particionais*, são métodos baseados na minimização de uma função custo onde os objetos são agrupados em K grupos escolhidos *a priori.* Cada item é agrupado na classe em que essa função custo é minimizada (VALE, 2005).

As principais vantagens do método *particional* em relação ao agrupamento *hierárquico* são duas*:*

1. Um item pode mudar de agrupamento com a evolução do algoritmo.
2. Há a possibilidade de se operar um maior conjunto de dados, pois estes métodos são mais rápidos de serem processados do que os métodos *hierárquicos*, devido a não necessidade de guardar na memória a matriz de *similaridade* completa durante o processamento.

Uma das principais desvantagens do método *particional* em relação ao método *hierárquico* é o fato de se ter que escolher um número K de *clusters* inicialmente, o que pode levar à conclusões erradas de acordo com o número de *clusters* escolhido. Isto ocorre porque o algoritmo poderá impor aos dados uma estrutura, em vez de encontrar a estrutura intrínseca aos dados.

2.8.1. K-médias

O método das k-médias consiste em agrupar os itens em k *clusters* mutuamente exclusivos, sendo que para encontrar estes *clusters* o algoritmo utiliza um processo iterativo com o objetivo de minimizar a soma das distâncias de cada item em relação ao centróide de cada *cluster,* assim como em outros métodos *particionais*. A principal diferença entre este método e outros *particionais* é que o centróide de cada *cluster* é dado pela média dos mesmos.

Para agrupar itens em *clusters*, utilizando o método das k-médias, aplica-se o procedimento descrito a seguir:

1. Repartir os itens em k grupos iniciais, aleatoriamente;
2. Encontrar as k médias de cada grupo ;
3. Para cada item determinar o grupo mais próximo usando uma medida de *dissimilaridade* (distância);
4. Calcular a média de cada partição. Se houver mudança na média das partições, voltar ao passo 2;
5. Fim do processo com as médias dos k grupos definidas.

Como muitos dos problemas de minimização, a solução encontrada pelo método das k-médias, geralmente depende do ponto de partida, pois o algoritmo encontra um mínimo local (VALE, 2005).

Este método é prático e computacionalmente eficiente, porém é suscetível a ruídos e *outliers,* também não é indicado a agrupamentos não convexos e não trabalha com dados categóricos.

2.8.2. PAM

O algoritmo PAM (*Partitioning Around Medoids*), como os demais métodos de agrupamento por *particionamento*, minimiza uma função custo em relação a um determinado vetor contendo k centróides, porém neste algoritmo estes centróides são objetos denominados *medóides*. Os *medóides* são objetos representativos de cada agrupamento e contêm os padrões onde a dissimilaridade média dos itens pertencentes a um dado agrupamento é mínima (VALE, 2005).

Este algoritmo é dividido em duas fases:

Fase 1: Construção

Essa fase é relativa à construção dos objetos *medóides* e são construídos através de k seleções de objetos representativos. O primeiro *medóide* é o item onde a soma das *dissimilaridades* entre todos os itens é mínima, os *medóides* seguintes são selecionados de forma a minimizar a função objetivo (2.5) o máximo possível, está função é dada por:

$$F_o = \sum_{i=1}^{n} d(x_i, m(x_i)) \tag{2.5}$$

Onde n é o total de itens no conjunto de dados, x_i é o *i – ésimo* item do conjunto de dados, $m(x_i)$ é o *medóide* mais próximo ao objeto x_i e $d(x_i, m(x_i))$ é dissimilaridade entre x_i e $m(x_i)$.

Os passos para encontrar os medóides são os seguintes (VALE, 2005):

1. Considere um item x_i que não tenha sido selecionado ainda;
2. Considere um item x_j não selecionado e calcule a diferença entre a sua dissimilaridade em relação ao último objeto selecionado (D_j) com a dissimilaridade do objeto x_i selecionado no passo anterior ($d(x_j, x_i)$);
3. Se a diferença for positiva. Calcule:

$$C_{ij} = \max(D_j - d(x_i, x_j), 0) \tag{2.6}$$

4. Calcule o total obtido por selecionar o item x_i:

$$\text{Total} = \sum_{j=1} C_{ij} \tag{2.7}$$

5. É selecionado o item x_i que maximize (2.7).

Fase 2: Troca

Nesta fase, tenta-se melhorar o conjunto de *medóides* trocando os itens entre eles. Dessa forma, se houver uma minimização da função objetivo, então mantém-se a troca, caso contrário ela é desfeita.

O resultado final é medido pela *Distância Média Final* (2.8):

$$DMF = \frac{1}{n}\sum_{i=1}^{n} d(x_i, m(x_i)) \qquad (2.8)$$

Onde n é o total de itens no conjunto de dados, x_i é o i–ésimo item do conjunto de dados, $m(x_i)$ é o medóide mais próximo ao objeto x_i e $d(x_i, m(x_i))$ é dissimilaridade entre x_i e $m(x_i)$.

2.9. GRÁFICO DA SILHUETA

O gráfico da silhueta é uma técnica que foi proposta por Rousseeuw, em 1986 [9] para avaliar particionamentos. Cada objeto (observação) é representado por um valor $s(x_i)$ chamado de *silhueta*, que é baseado na comparação da "consistência" e na "separação" de cada cluster. O agrupando inteiro é exibido combinando as silhuetas em um único gráfico, permitindo assim uma avaliação da qualidade relativa dos agrupamentos e uma avaliação da configuração dos dados. A largura média da silhueta fornece uma avaliação de validação do agrupamento, e poderia ser usada para selecionar um número "adequado" de agrupamentos.

Sendo A o agrupamento ao qual o objeto x_i pertence, a dissimilaridade (ou similaridade) média $a(x_i)$ do objeto x_i em relação aos outros objetos de *A* é dada por (2.9):

$$a(x_i) = \frac{1}{N_A - 1}\sum_{j \in A, j \neq i} d(x_i, x_j) \qquad (2.9)$$

Onde N_A representa o total de itens contidos no agrupamento e $d(x_i, x_j)$, é a dissimilaridade (ou similaridade) entre o item x_i e x_j.

Considere-se qualquer agrupamento *C* diferente de *A*. A dissimilaridade (ou similaridade) média do objeto x_i para todos os objetos *C* será dada por (2.10):

$$d(x_i, C) = \frac{1}{N_C}\sum_{j \in C} d(x_i, x_j) \qquad (2.10)$$

Onde N_C representa o total de objetos contido no cluster *C* e $d(x_i, x_j)$, é a similaridade entre o item x_i e x_j.

A menor distância de dissimilaridade entre x_i a um agrupamento A é dada por (2.11):

$$b(x_i) = \min\{d(x_i, C),\ \forall C \neq A\} \tag{2.11}$$

Considera-se como *B* o agrupamento *C* que contém a menor distância dada acima. Esse agrupamento chamado de vizinho do objeto x_i e é o segundo melhor agrupamento possível para esse objeto.

O valor da silhueta do objeto x_i, é calculada por (2.12):

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}} \tag{2.12}$$

O valor de $s(x_i)$ está entre -1 e 1 e pode ser interpretado da seguinte forma:

$S(x_i) \approx 1$ ⇨ objeto x_i está bem classificado no *cluster A*

$S(x_i) \approx 0$ ⇨ objeto x_i está entre os *clusters A e B*

$S(x_i) \approx -1$ ⇨ objeto x_i mal classificado no *cluster A* e mais próximo do *cluster B.*

O gráfico do *cluster A* é dado pelo gráfico dos valores da silhueta de todos os objetos pertencentes ao *cluster A* em ordem decrescente. Quanto mais próximo de 1, melhor é a qualidade do agrupamento.

Os valores da silhueta média podem ser interpretados como se segue na TABELA 5:

TABELA 5 – VALORES DA SILHUETA MÉDIA.

$S(x_i)$	**Descrição**
0,71 - 1,00	Uma estrutura forte foi encontrada.
0,51 - 0,70	Uma estrutura razoável foi encontrada.
0,26 - 0,50	A estrutura é fraca e pode ser superficial. É aconselhável o uso de outros métodos para esses dados.
≤ 0,25	Nenhuma estrutura substancial foi encontrada.

Fonte: Adaptação de Vale (2006).

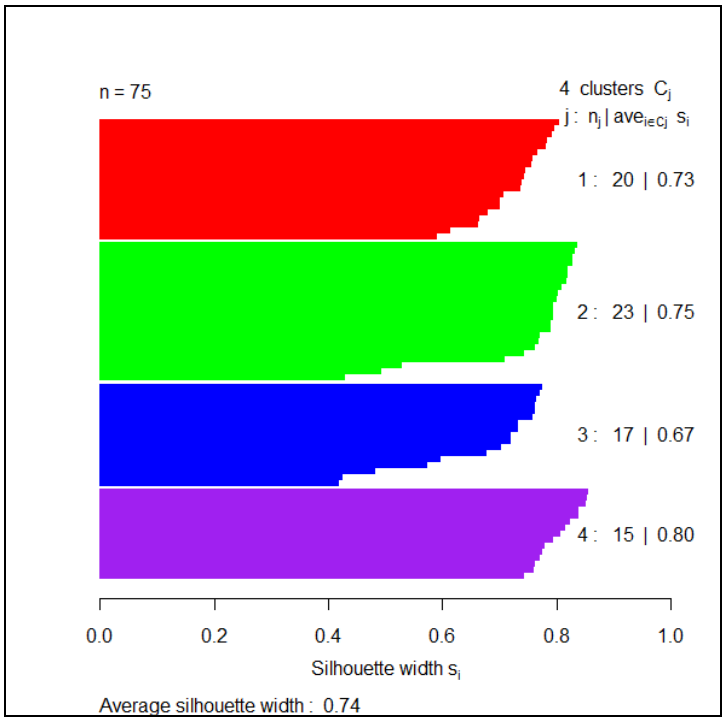

FIGURA 2: GRÁFICO DA SILHUETA.

Fonte: Ruspini (1970).

O eixo vertical representa as n observações enquanto o eixo horizontal representa o valor da silhueta para cada observação. Na figura 2, temos um gráfico da silhueta, para uma população que foi subdividida em quatro *clusters*.

3. MATERIAL E MÉTODO

3.1 SIMULAÇÃO

Para avaliar a metodologia utilizada, simulamos um conjunto de dados composto por 5 (cinco) sub-populações normais bivariadas com tamanho 1000, totalizando 5000 indivíduos. Os vetores de médias utilizados foram, $\mu'_{1=}$ [0,3 0,3], $\mu'_{2=}$ [0,5 0,5], $\mu'_{3=}$ [0,65 0,5], $\mu'_{4=}$ [0,7 0,3] e $\mu'_{5=}$ [0,4 0,5] e a matriz de variância-covariância $\sum$ é:

$$\Sigma = \begin{bmatrix} 0.003 & 0 \\ 0 & 0.004 \end{bmatrix}$$

Após a simulação foi feita uma análise de *clusters* pelos métodos não hierárquicos das *K-médias* e *PAM* apresentados nas seções 2.8.1 e 2.8.2 respectivamente. Para exibir os resultados, foram feitos gráficos das silhuetas variando o número de *clusters* iniciais k de 3 a 6 para os dois métodos utilizados. Também foi gerado um gráfico contendo as silhuetas médias encontradas para os métodos *PAM* e o K-médias variando o número de *clusters* inicial k entre 2 e 51, este gráfico é uma proposta para simplificar a visualização de quantos *clusters* deverão ser utilizado inicialmente na análise.

3.2 DADOS REAIS

Para agrupar os equipamentos em k *clusters* foi utilizado um banco de dados composto por 969 observações (equipamentos), contendo 33 variáveis descritas na TABELA 6.

TABELA 6 – DESCRIÇÃO DAS VARIÁVEIS DO BANCO DE DADOS.

Nome	**Tipo**	**Descrição**
T00_ID_contrato	Categórica	Código do contrato
T00_Tipo_Contrato	Categórica	Tipo de contrato
T00_Regiao	Categórica	Região do país
T00_Frota_automovel	Numérica	Frota de carros
T00_Frota_Bonde	Numérica	Frota de bonde
T00_Frota_Caminhao	Numérica	Frota de Caminhao
T00_Frota_Caminhao_Trator	Numérica	Frota de Caminhao_Trator
T00_Frota_Caminhonete	Numérica	Frota de Caminhonete
T00_Frota_Caminhoneta	Numérica	Frota de Caminhoneta
T00_Frota_Chassi_plataforma	Numérica	Frota de Chassi_plataforma
T00_Frota_Ciclomotor	Numérica	Frota de Ciclomotor
T00_Frota_Micro_Onibus	Numérica	Frota de Micro_Onibus
T00_Frota_Motocicleta	Numérica	Frota de Motocicleta
T00_Frota_Motoneta	Numérica	Frota de Motoneta
T00_Frota_Onibus	Numérica	Frota de Onibus
T00_Frota_Quadriciclo	Numérica	Frota de Quadriciclo
T00_Frota_Reboque	Numérica	Frota de Reboque
T00_Frota_Semi_Reboque	Numérica	Frota de Semi_Reboque
T00_Frota_Side_Car	Numérica	Frota de Side_Car
T00_Frota_Outros	Numérica	Frota de Outros Veículos
T00_Frota_Trator_Esteira	Numérica	Frota de Trator_Esteira
T00_Frota_Trator_Rodas	Numérica	Frota de Trator_Rodas
T00_Frota_Triciclo	Numérica	Frota de Triciclo
T00_Frota_Utilitarios	Numérica	Frota de Utilitarios
T00_Idade_Contrato	Numérica	Idade do contrato
T00_Area_km2	Numérica	Área em Km² do contrato
T00_Habitantes	Numérica	Numero de Habitantes
T00_Velocidade_Regulamentada	Categórica	Velocidade regulamentada
T00_Modelo	Categórica	Modelo do equipamento
T00_Horas_operacional	Numérica	Horas diárias de operação
T00_Captura_Moto	Binaria	Monitora a velocidade de motos
T00_Safe	Binaria	Reconhecimento de caractere
T00_Fluxo_Veículos	Numerica	Fluxo de Veículos no ponto

Fonte: os autores.

Para preservar a empresa que cedeu os dados as variáveis categóricas foram substituídas por códigos com exceção da variável T00_região.

A análise foi realizada com o método não hierárquico *PAM,* pois muitas destas variáveis são qualitativas impossibilitando a utilização do método das k-médias. Pelo mesmo motivo a medida de parecença utilizada foi o coeficiente de Gower. No intuito de verificar o número de *clusters* ao qual a população deverá ser dividida utilizou-se um método iterativo, onde, se elaborou

o particionamento da população variando o número de *clusters* k de 2 a 51, e para verificar a qualidade dos ajustes utilizou-se de um gráfico da silhueta média em relação ao número de *clusters* a qual foi gerada.

O programa estatístico utilizado, tanto para a geração e análise dos dados simulados e para a análise dos dados reais, foi o *software* R. A versão utilizada foi a 2.7.0. No *software* R, foram ainda utilizados dois pacotes: *cluster* e *MASS*.

4. RESULTADOS E DISCUSSÃO

4.1. RESULTADOS DA SIMULAÇÃO

Neste tópico, demonstra-se como o gráfico da silhueta média pode ser utilizado para encontrar o número de *clusters* ideal.

A população simulada está representada na FIGURA 3.

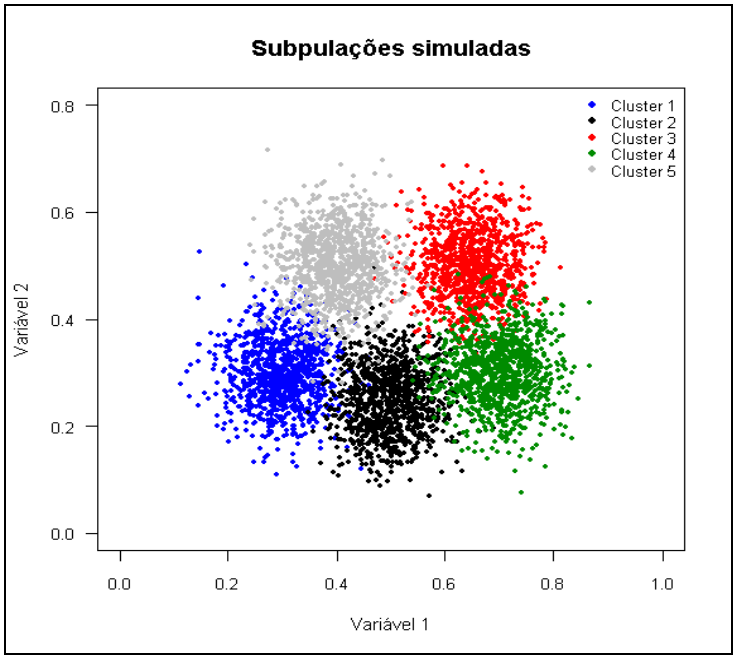

FIGURA 3: DIAGRAMA DE DISPERSÃO DOS CLUSTERS SIMULADOS.

Fonte: os autores.

Após esta simulação, a população foi dividida em clusters utilizando dois algoritmos de particionamento bastante difundidos: k-médias *e PAM*. A qualidade dos agrupamentos é mostrada pelos gráficos das silhuetas contidas na FIGURA 4 e na FIGURA 5.

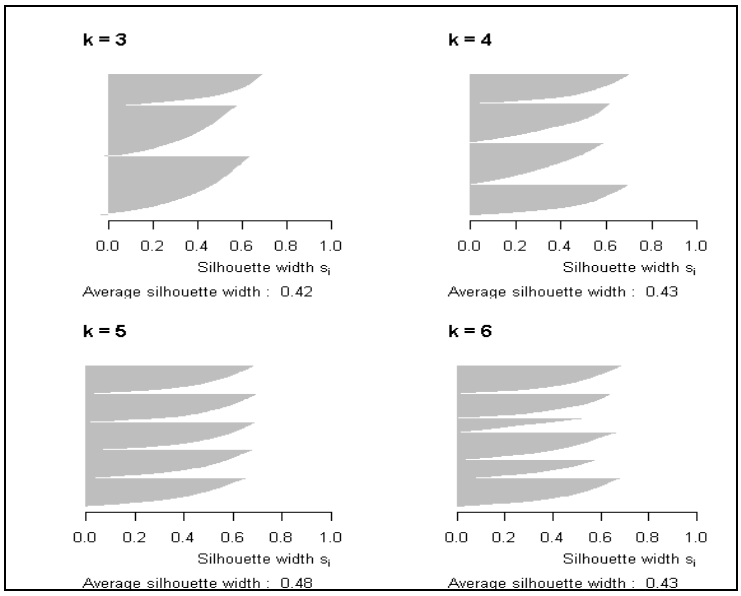

FIGURA 4 – GRÁFICOS DA SILHUETA PARA O ALGORITMO K-MÉDIAS, COM K VARIANDO DE 3 A 6.

Fonte: os autores.

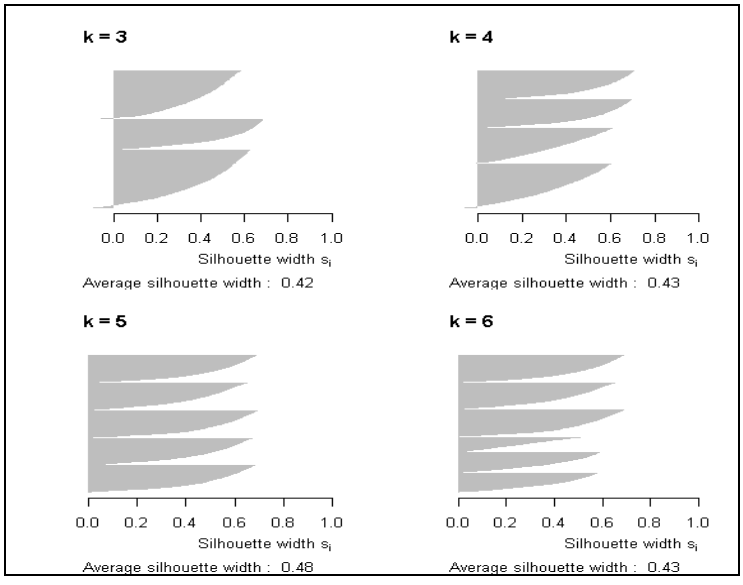

FIGURA 5 – GRÁFICOS DA SILHUETA PARA O ALGORITMO PAM, COM K VARIANDO DE 3 A 6.

Fonte: os autores.

Nas FIGURAS 4 e 5 estão apresentados os gráficos das silhuetas variando os valores para *k* de 3 a 6, onde se verifica que a melhor qualidade de agrupamento é quando k = 5, tanto para o método *PAM* quanto o método das k médias. Conforme já se esperava, pois o número de sub-populações era conhecido *a priori*.

Para tornar a utilização das silhuetas mais eficiente gerou-se um gráfico contendo o valor da silhueta média calculada para k variando de 2 a 51, tanto para o *PAM* quanto para o método das k médias (FIGURA 6).

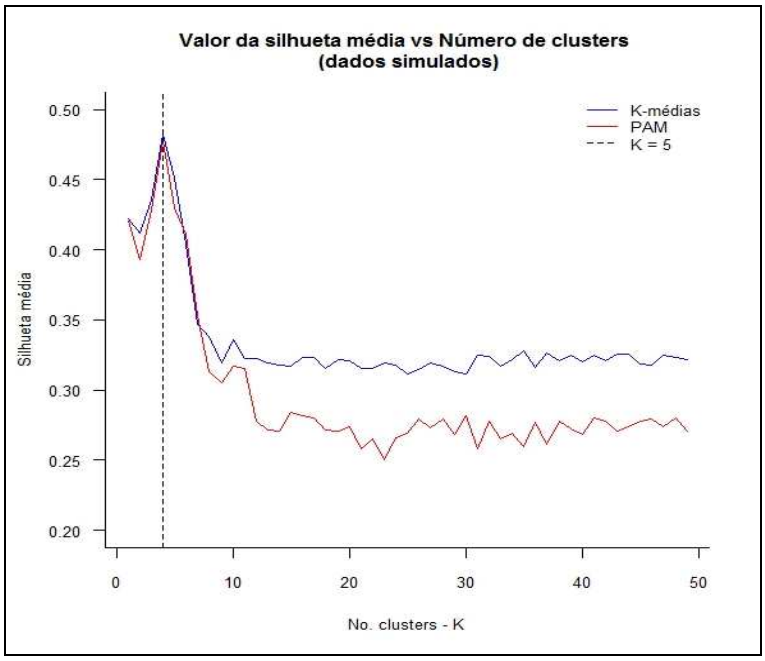

FIGURA 6: VALOR DA SILHUETA MÉDIA VS NÚMERO K DE CLUSTERS, PARA O K-MÉDIAS E PAM.

Fonte: os autores.

Na FIGURA 6, tem-se no eixo y (das ordenadas) o valor da silhueta média e no eixo x (das abscissas) a quantidade de clusters k que foi utilizada.

Todas as análises feitas neste item foram feitas com a utilização da matriz de distância euclidiana.

4.2. DADOS REAIS

Neste tópico foi realizada a análise de cluster para os dados reais, conforme a proposta contida no item 3.1.

4.2.1. DEFININDO O NÚMERO DE CLUSTERS

Como visto na revisão de literatura os métodos não hierárquicos necessitam que o número de *clusters k* seja informado *a priori*. Para isto aplicou-se a metodologia proposta no item 3.2, que é análoga a elaborada no item 4.1 para os dados simulados, com a diferença de que agora a matriz de dissimilaridade utiliza o coeficiente de Gower. Outro ponto importante é de que neste caso é aplicado apenas o algoritmo *PAM*, tanto o coeficiente quanto o algoritmo foram escolhidos por utilizar variáveis categóricas em conjunto com variáveis numéricas. O gráfico gerado para as silhuetas médias em relação ao número de clusters está apresentado na FIGURA 7.

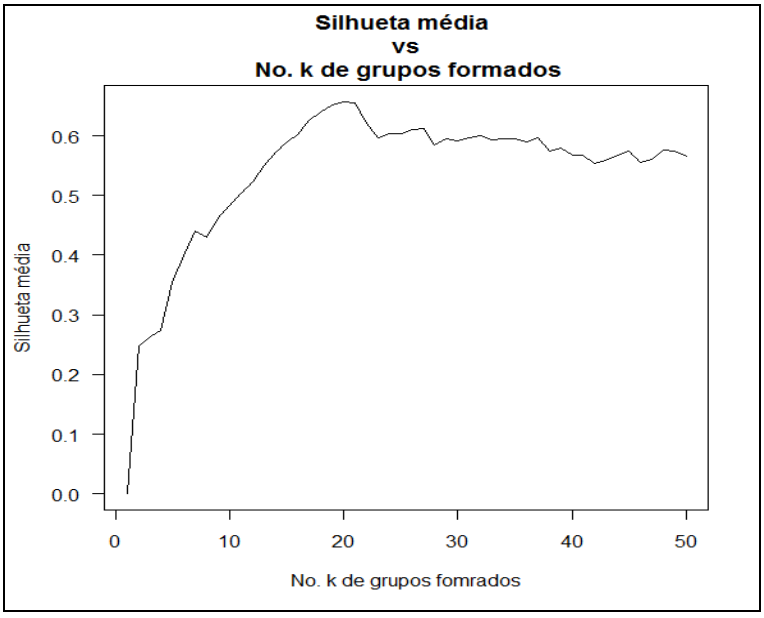

FIGURA 7 – GRÁFICO DA SILHUETA MÉDIA

Fonte: os autores.

Com base na FIGURA 7 tem-se que o número de *clusters* é k = 21, pois representa o maior valor para a silhueta média, e segundo a TABELA 5 uma estrutura razoável foi encontrada.

4.2.2. RESULTADOS DA ANÁLISE DE CLUSTER

Utilizando a quantidade de *clusters* k = 21, conforme encontrado no item anterior elaborou – se a TABELA 7 contendo algumas características da qualidade de ajuste do agrupamento.

TABELA 7 – TABELA DE VERIFICAÇÃO DA QUALIDADE DO AGRUPAMENTO

Núm. do *Cluster*	Quantidade de equip.	Valor médio da silhueta
1	53	0,6615
2	111	0,7794
3	22	0,8498
4	53	0,6638
5	59	0,1529
6	39	0,3741
7	21	0,9298
8	25	0,6898
9	79	0,7587
10	25	0,5906
11	44	0,6698
12	35	0,6760
13	36	0,5238
14	41	0,3411
15	29	0,6626
16	22	0,6741
17	38	0,6568
18	49	0,8045
19	110	0,8178
20	21	0,4426
21	57	0,7084
Mínimo	21	0,1529
Média	46	0,6394
Máximo	111	0,9298

Fonte: os autores.

Pela TABELA 7, tem-se que a quantidade de equipamentos por agrupamentos, variou de 21 (*cluster* 20) a 111 (*cluster* 2). Observa-se também, que 46 é o número médio de equipamentos por *cluster* formado.

Ainda na TABELA 7, o valor médio da silhueta, para cada agrupamento, variou de 0,1529 (cluster 5) a 0,9298 (cluster 7). A média das médias dos valores médios da silhueta foi de 0,6394.

A FIGURA 8 apresenta silhuetas, com k = 21 *clusters* formados, complementado a TABELA 7.

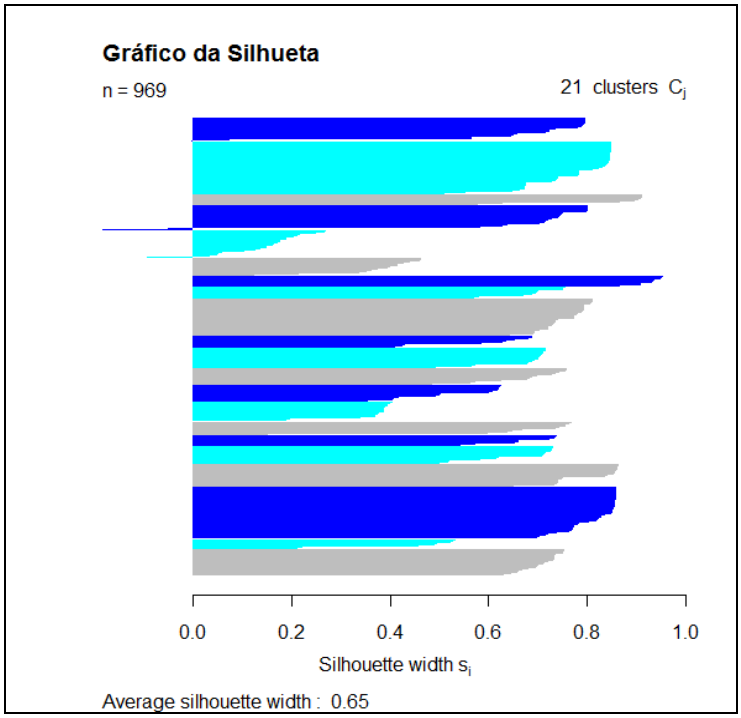

FIGURA 8 – GRÁFICO DA SILHUETA DOS 21 CLUSTERS FORMADOS.

Fonte: os autores.

4.2.3. PERFIL DOS CLUSTERS FORMADOS

Após a obtenção dos grupos, interessa-se saber qual é o perfil destes clusters.

***Cluster* 1**

O c*luster* 1 é formado por 53 equipamentos, localizados na região sudeste, sendo estes pertencentes à contratos do tipo 3. A velocidade máxima permitida na via onde estão instalados estes dispositivos, varia de 40 a 60 Km/h. Os modelos de equipamentos predominantes são do tipo A e R e 90% destes aparelhos têm a capacidade de fazer aferição da velocidade de motocicletas.

***Cluster* 2**

O c*luster* 2 é formado por 111 equipamentos, foi o maior grupo gerado. Estes equipamentos estão todos localizados na região sudeste, sendo estes pertencentes à contratos do tipo 1. A velocidade máxima permitida na via onde estão instalados estes dispositivos é predominantemente 50 Km/h. Os aparelhos são em sua maioria dos modelos A e R e aproximadamente 30% deles têm a capacidade de aferir a velocidade de motocicletas.

***Cluster* 3**

O c*luster* 3 é formado por 22 equipamentos, localizados na região nordeste, e pertencentes à contratos do tipo 3. A velocidade máxima permitida na via onde estão instalados estes dispositivos, é predominantemente 60 Km/h. Outras características destes equipamentos são que todos são habilitados a fazer a aferição de motocicletas, a maioria são do modelo do tipo A e funcionam por 15 horas diárias.

***Cluster* 4**

O c*luster* 4 é formado por 53 equipamentos, sendo que 90% destes estão na região centro-oeste, e contratos tipo 3. A velocidade máxima permitida na via onde estão instalados estes aparelhos, é predominantemente 30 Km/h, ou seja, vias de baixa velocidade de circulação. Outras características deste agrupamento são: 98% têm a capacidade de fazer a aferição em motocicletas, são em sua maioria da marca R e marca I.

***Cluster* 5**

O c*luster* 5 é formado por 59 equipamentos, sendo que 98% estão na região sudeste e alocados em contratos tipo 3. A velocidade máxima permitida na via onde estão instalados estes aparelhos, varia de 30 a 50 Km/h. Os modelos de equipamentos pertencentes ao *cluster* 5, mostrou-se bastante heterogêneo, sendo os modelos F e P os predominantes (juntas, representam aproximadamente 45% dos equipamentos). Quanto a aferição de motocicletas, aproximadamente 35% dos equipamentos do agrupamento, dispõe de tal funcionalidade.

***Cluster* 6**

O *cluster* 6 é formado por 39 equipamentos, sendo que estes estão localizados na região nordeste e em contratos do tipo 3. A velocidade máxima permitida na via onde estão instalados estes aparelhos, na sua maioria, varia de 40 a 50 Km/h. Outras características deste grupo são que o modelo F é o mais freqüente e apenas 12% dos equipamentos têm a funcionalidade de aferir motocicletas.

***Cluster* 7**

O c*luster* 7 é formado por 21 equipamentos, situados na região sudeste e em contratos do tipo 2. A velocidade máxima permitida na via onde

estão instalados estes equipamentos é predominantemente 60 Km/h. O modelo dos equipamentos, mais freqüente é o modelo R. Dos equipamentos neste *cluster*, aproximadamente 30% têm a funcionalidade de aferir a velocidade de motocicletas.

***Cluster* 8**

O c*luster* 8 é formado por 25 equipamentos, situados na região sul e em contratos do tipo 2. A velocidade máxima permitida na via onde estão instalados estes equipamentos varia de 40 a 60 Km/h. O modelo dos equipamentos predominantes é o modelo R. Neste agrupamento, aproximadamente 30% dos equipamentos têm a funcionalidade de aferir a velocidade de motocicletas.

***Cluster* 9**

O c*luster* 9 é formado por 79 equipamentos, situados na região sudeste e encontram-se em contratos tipo 2. A velocidade máxima permitida na via onde estão instalados estes equipamentos, são em sua maioria, de 50 e 60 Km/h. Os modelos dos equipamentos são bastante heterogêneos, sendo os mais freqüentes os do tipo F, I, R e H. Neste agrupamento, apenas 12% dos aparelhos têm a funcionalidade de aferir a velocidade de motocicletas.

***Cluster* 10**

O c*luster* 10 é formado por 25 equipamentos, situados na região sul e em contratos tipo 2. A velocidade máxima permitida na via onde estão instalados estes equipamentos é de 60 Km/h. As marcas dos equipamentos são bastante heterogêneas, sendo as mais freqüentes, as marcas S e E. Quase que a totalidade dos aparelhos deste agrupamento, não tem a funcionalidade de aferir a velocidade de motocicletas.

***Cluster* 11**

O c*luster* 11 é formado por 44 equipamentos, situados na região centro-oeste e contratos tipo 2. A velocidade máxima permitida na via onde estão instalados estes equipamentos é de 40 Km/h. Os modelos F e H correspondem a 86% dos modelos de equipamentos presentes no agrupamento. Quase que a totalidade dos equipamentos (97%) não têm a funcionalidade de aferir a velocidade de motocicletas.

***Cluster* 12**

O c*luster* 12 é formado por 35 equipamentos, situados na região sudeste e contratos tipo 1. A velocidade máxima permitida na via onde estão instalados estes equipamentos varia predominantemente de 40 a 60 Km/h, no entanto, foi alocado a este agrupamento os dois equipamentos instalados em vias com limite de 80 Km/h. Os modelos F e P, com uma incidência de 72%, são os modelos mais presentes no agrupamento. Grande parte dos equipamentos (aproximadamente 72%), têm a funcionalidade de aferir a velocidade de motocicletas.

***Cluster* 13**

O c*luster* 13 é formado por 36 equipamentos, situados na região nordeste e contratos tipo 2. A velocidade máxima permitida na via onde estão instalados estes equipamentos varia de 50 a 60 Km/h. Os modelos F e H correspondem a praticamente todos os equipamentos do agrupamento. Nenhum dos equipamentos do agrupamento tem a funcionalidade de aferir velocidade de motocicletas. Quanto ao tempo de operação diária, 66% dos equipamentos operam menos de 24 horas diárias.

***Cluster* 14**

O c*luster* 14 é formado por 41 equipamentos, sendo que 38 equipamento estão na região centro-oeste e 3 no sudeste, e contratos tipo 2. A velocidade máxima permitida na via onde estão instalados estes equipamentos são em sua maioria (93%) de 40 Km/h. Os modelos F e I correspondem a maioria dos equipamentos alocados no agrupamento. Nenhum dos equipamentos do agrupamento tem a funcionalidade de aferir a velocidade de motocicletas.

***Cluster* 15**

O c*luster* 15 é formado por 29 equipamentos, sendo que estes estão situados na região centro-oeste e 93% deles são regidos por contratos tipo 1. A velocidade máxima permitida na via onde estão instalados estes equipamentos, varia de 40 a 60 Km/h. Os modelos F e H correspondem a grande maioria dos equipamentos presentes neste agrupamento. Praticamente nenhum dos equipamentos, apenas 4%, têm a funcionalidade de aferir a velocidade de motocicletas.

***Cluster* 16**

O c*luster* 16 é formado por 22 equipamentos, sendo que estes estão instalados na região sul e contratos tipo 2. A velocidade máxima permitida na via onde estão instalados estes equipamentos, varia de 40 a 50 Km/h. O modelo E e o modelo I são os mais freqüentes neste agrupamento. Nenhum dos equipamentos presentes no *cluster*, tem a funcionalidade de aferir a velocidade de motocicletas.

***Cluster* 17**

O c*luster* 17 é formado por 38 equipamentos, sendo que estes estão situados no nordeste e com contratos tipo 1. A velocidade máxima permitida na via onde estão instalados estes equipamentos, são em sua maioria de 40 Km/h,

no entanto, alguns estão instalados em vias limitadas a 60 Km/h. O tempo de funcionamento diário destes equipamentos é de 18 horas. Pouco menos da metade, 37%, apresentam a funcionalidade de aferir a velocidade de motocicletas.

Cluster 18

O c*luster* 18 é formado por 49 equipamentos, sendo que estes estão situados na região sul do Brasil e contratos tipo 1. A velocidade máxima permitida na via onde estão instalados estes equipamentos, varia de 40 a 50 Km/h. Outras características deste agrupamento: são todos modelos do tipo P e apenas 17% apresentam a funcionalidade de aferir a velocidade de motocicletas.

Cluster 19

O c*luster* 19 é o 2º. maior agrupamento formado. Os equipamentos estão localizados no sudeste do Brasil, e são regidos pelo contrato tipo 3. A velocidade máxima permitida na via onde estão instalados estes equipamentos, varia de 40 a 50 Km/h. Os modelos P, F e C abrangem aproximadamente 90% dos equipamentos presentes neste agrupamento. A funcionalidade de aferir a velocidade de motocicletas não é uma característica deste agrupamento (menos de 1% de incidência).

Cluster 20

O c*luster* 20 é formado por 21 equipamentos, sendo que estes estão situados na região nordeste e, são regidos por contratos tipo 1. A velocidade máxima permitida na via onde estão instalados estes equipamentos, predominantemente é de 40 Km/h neste agrupamento. Os modelos dos equipamentos são bastante heterogêneos, sendo que o modelo F corresponde sozinho a 50% dos aparelhos. A funcionalidade de aferir a velocidade de motocicletas ocorre em 30% dos equipamentos do agrupamento.

Cluster 21

O c*luster* 21 é formado por 57 equipamentos, sendo que estes estão situados na região centro-oeste e contratos tipo 1. A velocidade máxima permitida na via onde estão instalados estes equipamentos, predominantemente é de 40 Km/h. Os modelos dos equipamentos são em sua maioria do tipo F, E e I. Os equipamentos não são habilitados a aferir a velocidade de motocicletas. O tempo de operação diária destes aparelhos é de 19 horas.

5. CONCLUSÃO

Os resultados deste trabalho mostraram que a metodologia proposta é eficiente, pois de acordo com a simulação proposta no item 4.1 e com seus resultados apresentados no item 5.1 podemos perceber que o algoritmo das k-médias e o *PAM* tiveram resultados próximos para os valores de k avaliados, outra conclusão retirada dos dados simulados, é que utilizar um processo iterativo para escolher um numero k de *clusters* inicial traz resultados muito satisfatórios.

Em relação aos dados reais, a metodologia empregada obteve um bom agrupamento para os equipamentos analisados, de acordo com a análise técnica de profissionais que trabalham com estes equipamentos. Os *clusters* encontrados vêm ao encontro dos sentimentos dos mesmos.

O conhecimento dos clusters viabiliza futuras modelagens, utilizadas para determinar o volume de informações geradas, bem como possibilita a seleção, de forma mais homogênea, de equipamentos para analisar o desempenho de determinadas peças instaladas.

6. REFERENCIAS BIBLIOGRAFICAS

1. CHAVES NETO, A. **Notas de Aula - Análise Multivariada II.** Curitiba: [s.n.], 2007.

2. CUADRAS, C. M. **Distâncias Estadísticas**. Vol 30, Núm. 119, pág. 300. Departamento d' Estadística, Universidad de Barcelona.

3. FERREIRA, D.F. **Análise Multivariada.** Lavras: 1996. Disponível em < http://www.dex.ufla.br/~danielff/dex522.pdf >. Acesso em: 05/04/2008.

4. GOWER, J. C. (1971). **A general coefficient of similarity and some of its properties**. Biometrics, 77, 623-637.

5. JOHNSON, R. A.; WICHERN, D. W. **Aplied multivariate statistical analysis**. *4th*, ed. Prentice-Hall, New Jersey, 1999.

6. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. (2005). **Cluster Analysis Basics and Extensions**; unpublished

7. R Development Core Team (2008). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

8. RENCHER, A. C. **Methods of Multivariate Analysis**. Second Edition. J. Wiley & Sons. INC. Publication.

9. ROUSSEEUW, P.J. (1987). **Silhouettes: A graphical and to the interpretation and validation of cluster analysis**. J. Comput. Appl. Math., 20, 53-65.

10. RUSPINI, H.R.(1970) **Numerical method for fuzzi clustering**. Information Sciences 2, 310-350.

11. SALAZAR, M. E. R.; HERNÁNDEZ, S.A.; NUÑEZ, E. B. **Coeficientes de Asociación**. P y V. Madri, Espanha: 2001.

12. VALE, M. N. **Agrupamentos de dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos.** Rio de Janeiro, 2005. 120 f. Dissertação (Mestrado) – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

13. Venables, W. N. & Ripley, B. D. (2002) **Modern Applied Statistics with S**. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

APÊNDICE - Comandos do R [7] utilizados na análise.

```
dados <- read.table("dadostcc.txt",h=T,sep=";",dec=",")

###################
require(cluster)

### Dados Reais

## Criando a matriz de dissimilaridade
matriz.dissim <- daisy(dados[,2:34], metric = "gower", stand = FALSE)

##### Código usando para obter o número de
##### clusters pelo gráfico da silhueta:

m = numeric(50)
for(i in 2:50){
k = i
result.agrup.pam <- pam(matriz.dissim,k, diss = TRUE)
result.silhueta <- silhouette(result.agrup.pam)
summary(result.silhueta)$si.summary
m[i] <- summary(result.silhueta)$avg.wid}

#### Com base na simulação, o número de cluster que
#### apresentou um maior valor da silhueta média, foi
#### a análise que considerou um tamanho de cluster igual
#### a k = 21.

k = 21

#### PARTICIONANDO A AMOSTRA DE EQUIPAMENTOS (PAM)

# Particiona a amostra de equipamentos, em 21 clusters
# utilizando o algoritmo não-hierárquico PAM
#
result.agrup.pam <- pam(matriz.dissim,k, diss = TRUE)

#### GRÁFICO DA SILHUETA

# Após a divisão da população nos 21 clusters (usando o PAM)...
# ...Calcula o valor da silhueta, para cada objeto (equipamento).
#
result.silhueta <- silhouette(result.agrup.pam)

## Gráfico da silhueta, para um k = 21 clusters.
plot(silhouette(result.agrup.pam),
main=paste("Gráfico da Silhueta ","",sep=""),
do.clus.stat=F,col=c(rep(c(4,5,8),7)),border=0)

# Resumo das análises
summary(result.silhueta)

###############################################
############ Clusters formados:
########

#cluster 1;cluster1 = subset(dados2,result.agrup.pam$clust==1)
summary(cluster1)
#cluster 2;cluster2 = subset(dados2,result.agrup.pam$clust==2)
summary(cluster2)
#cluster 3;cluster3 = subset(dados2,result.agrup.pam$clust==3)
summary(cluster3)
#cluster 4;cluster4 = subset(dados2,result.agrup.pam$clust==4)
summary(cluster4)
#cluster 5;cluster5 = subset(dados2,result.agrup.pam$clust==5)
summary(cluster5)
#cluster 6;cluster6 = subset(dados2,result.agrup.pam$clust==6)
```

```
summary(cluster6)
#cluster 7;cluster7 = subset(dados2,result.agrup.pam$clust==7)
summary(cluster7)
#cluster 8;cluster8 = subset(dados2,result.agrup.pam$clust==8)
summary(cluster8)
#cluster 9;cluster9 = subset(dados2,result.agrup.pam$clust==9)
summary(cluster9)
#cluster 10;cluster10 = subset(dados2,result.agrup.pam$clust==10)
summary(cluster10)
#cluster 11;cluster11 = subset(dados2,result.agrup.pam$clust==11)
summary(cluster11)
#cluster 12;cluster12 = subset(dados2,result.agrup.pam$clust==12)
summary(cluster12)
#cluster 13;cluster13 = subset(dados2,result.agrup.pam$clust==13)
summary(cluster13)
#cluster 14;cluster14 = subset(dados2,result.agrup.pam$clust==14)
summary(cluster14)
#cluster 15;cluster15 = subset(dados2,result.agrup.pam$clust==15)
summary(cluster15)
#cluster 16;cluster16 = subset(dados2,result.agrup.pam$clust==16)
summary(cluster16)
#cluster 17;cluster17 = subset(dados2,result.agrup.pam$clust==17)
summary(cluster17)
#cluster 18;cluster18 = subset(dados2,result.agrup.pam$clust==18)
summary(cluster18)
#cluster 19;cluster19 = subset(dados2,result.agrup.pam$clust==19)
summary(cluster19)
#cluster 20;cluster20 = subset(dados2,result.agrup.pam$clust==20)
summary(cluster20)
#cluster 21;cluster21 = subset(dados2,result.agrup.pam$clust==21)
summary(cluster21)

#################
##########Simulados
##################

require(MASS)
## Criando matrizes de correlação (consideraremos
## nesta simulação, subpopulação com matrizes de
## variância-covariância iguais e, variáveis inde-
## pendentes.

matriz.var <- c(.003,0,0,.004)

sigma1 <- matrix(matriz.var,nr=2)
sigma2 <- matrix(matriz.var,nr=2)
sigma3 <- matrix(matriz.var,nr=2)
sigma4 <- matrix(matriz.var,nr=2)
sigma5 <- matrix(matriz.var,nr=2)

## Gerando as 5 subpopulações...
mvr1 <- mvrnorm(1000, c(.3,.3), sigma1)
mvr2 <- mvrnorm(1000, c(.5,.25), sigma2)
mvr3 <- mvrnorm(1000, c(.65,.5), sigma3)
mvr4 <- mvrnorm(1000, c(.7,.3), sigma4)
mvr5 <- mvrnorm(1000, c(.4,.5), sigma5)

dat <- rbind(mvr1,mvr2,mvr3,mvr4,mvr5)

### Diagrama de dispersão para as subpopulações geradas
plot(mvr1, pch=19, cex=.6, col='blue',xlab="Variável 1",
ylab="Variável 2",xlim=c(0,1),ylim=c(0,.8),
main="Subpulações simuladas",las=1,cex.axis=.8,cex.lab=.9)

points(mvr2, pch=19, cex=.6, col='black')
points(mvr3, pch=19, cex=.6, col='red')
points(mvr4, pch=19, cex=.6, col='green4')
points(mvr5, pch=19, cex=.6, col='gray')
```

```
legend("topright", c("Cluster 1", "Cluster 2", "Cluster 3",
"Cluster 4","Cluster 5"),col = c("blue","black","red","green4","gray"),
pch=c(19,19,19,19,19),box.lty=0,cex=.6)

##########
### Simulação para o método das K-médias
########

#########################################################
## Rodando o k-médias para um "k" variando de 3 a 6
#########################################################

lista.temp <- list()
for(i in 1:4){
lista.temp[[i]] <- kmeans(dat, i+2)}

### Criando o gráfico de silhueta

par(mfrow=c(2,2)) # divide a janela gráfica em 3 linhas e 3 colunas
start1 = Sys.time() #inicia um contador de tempo (hora atual sistema)
for(i in 1:9){
si2 <- silhouette(lista.temp[[i]]$cluster, dist(dat, "euclidian"))
plot(si2, do.n.k=FALSE,do.clus.stat=FALSE,cex.names=0.6,
main=paste("k = ",i+2,sep=""),adj=1)}

### Simulação para o PAM

## Rodando o PAM para um "k" variando de 3 a 6
lista.temp <- list()
for(i in 1:4){
lista.temp[[i]] <- pam(dat, i+2)}

## Criando o gráfico da silhueta
par(mfrow=c(2,2))

for(i in 1:4){
si2 <- silhouette(lista.temp[[i]]$cluster, dist(dat, "euclidian"))
plot(si2, do.n.k=FALSE,do.clus.stat=FALSE,cex.names=0.6,
main=paste("k = ",i+2,sep=""),adj=1)}

m1 = numeric(49)
for(i in 2:50){
temp <- kmeans(dat,i)
si2 <- silhouette(temp$cluster, dist(dat, "euclidian"))
m1[i-1] <- summary(si2)$avg.wid}

m2 = numeric(49)

for(i in 2:50){
temp <- pam(dat,i)
si2 <- silhouette(temp$cluster, dist(dat, "euclidian"))
m2[i-1] <- summary(si2)$avg.wid}

plot(m1,type="l",xlab="No. clusters - K",ylab="Silhueta média",
cex.axis=.8,col="blue",las=1,cex.lab=.8,bty="l",ylim=c(0.2,0.5))

lines(m2,col="red")
abline(v=4,lty=2)

legend("topright", c("K-médias", "PAM", "K = 5"),
col = c("blue","red","black"),lty = c(1, 1, 2),box.lty=0,cex=.8)
title("Valor da silhueta média vs Número de clusters \n (dados
simulados)",cex.main=1)
```