REGIÕES HOMOGÊNIAS DO MUNICIPIO DE BELÉM – UMA APLICAÇÃO DAS TÉCNICAS DE AGRUPAMENTO

Francisco do Nascimento Felix.*[1]
Sérgio Castro Gomes.*[2]

RESUMO: O presente trabalho apresenta uma aplicação das Técnicas Multivariadas, em particular, a Análise de Conglomerados, através do método das K-médias, objetivando a obtenção de 3 regiões homogêneas(conglomerados de bairros) a partir do conjunto de variáveis sócio-econômicas selecionadas no Censo demográfico de 1991 e decorrentes das tabulações especiais realizada pela Fundação Instituto Brasileiro de Geografia e Estatística – F.IBGE. As variáveis escolhidas segundo os bairros que integram o Município de Belém, referem-se à situação de moradia no bairro; a infra-estrutura sanitária dos domicílios; proporção da população escolarizada; o contigente populacional e a renda média dos chefes do domicílio.

Após escolha das variáveis, foi elaborado a matriz de dados, necessária à geração das estatísticas descritivas, a padronização da escala das variáveis para posterior aplicação da técnica.

INTRODUÇÃO

A partir dos resultados apresentados, no Censo Demográfico de 1991, para o Estado do Pará(em 1996), e dos dados disponíveis a nível de Município de Belém e dos bairros que o compõe, torna-se possível a elaboração de estudos mais detalhados sobre a situação sócio - econômica da população residente nos bairros do Município de Belém.

Os dados apresentados em tabulação especial desenvolvida pelo F.IBGE, aos municípios que compõe as regiões metropolitanas do País demonstram a preocupação desse instituto em produzir informações sócio - econômicas que reflitam a real situação dessas regiões , bem como a desigualdade existente entre elas.

A desagregação dos resultados censitários por município possibilita a geração de estudos, como o aqui apresentado, em que o objetivo principal é o de criar grupos de bairros homogêneos com relação às variáveis selecionadas. Este trabalho segue as idéias de Bussab e Dini (1985), que apresentaram resultados para o Estado de São Paulo. O processo de formação dos grupos será desenvolvido utilizando-se técnicas de análise multivariada, em particular a análise de conglomerados pelo método das K-médias.

[1]* Bacharel em Estatistica – UFPA. Especialista em Estatistica- UFPA. Departamento de Matemática e Estatistica. Universidade da Amazônia. Belém – Pa. Assessor Estatistico da Pesquisa de Emprego e Desemprego na Região Metropolitana de Belém – SINE/PA / IDESP./DIEESE/SUDAM/FADESP/F.SEADE-SP.
[2]*Licenciado Pleno em Matemática – UFPA. Especialista em Aplicações Estatisticas – UFPA. Especialista em Estatistica- UFPA. Chefe do Departamento de Matemática e Estatistica. Universidade da Amazônia. Belém – Pa. Assessor Estatistico do Instituto de Previdência e Assistência dos Servidores Públicos do Estado do Pará - IPASEP

DESENVOLVIMENTO

A região de interesse é formada pelos 20 bairros que compõe o município de Belém, que são: B. Campos, Canudos, C. Velha, Condor, Cremação, Guamá, Jurunas, Marambaia, Marco, Fátima, Nazaré, Pedreira, Reduto, Sacramenta, São Braz, Souza, Telégrafo, Terra Firme, Comércio, Umarizal. Para os quais se dispõe das seguintes informações:

a) Percentual de domicílios particulares permanentes com localização casa (PLO).
b) Percentual de domicílios particulares permanentes com abastecimento de água com canalização interna (PAA).
c) Percentual de domicílios particulares permanentes com instalação sanitária só no domicílio (PIS).
d) Percentual de domicílios particulares permanentes em condição de ocupação (PCO).
e) Percentual de domicílios particulares permanentes com lixo coletado (PLC).
f) Número médio de dormitórios por domicílio (MDD).
g) Número médio de banheiros por domicílio (MBD).
h) Número médio de pessoas por domicílio (MPD).
i) Renda média nominal dos chefes (RMC).
j) População (POP).
k) Percentual de pessoas alfabetizadas (PPA)

A Tabela 1 a seguir apresenta o conjunto de observações das 11 variáveis para os 20 bairros de Belém.

Tabela 1

Variáveis por bairros do Município de Belém

BAIRRO	PLO	PAA	PIS	PCO	PLC	MDD	MBD	MPD	RMC	POP	PPA
B. CAMPOS	0,4801	0,9892	0,9625	0,7068	0,9873	2,48	2,24	4,40	494682	19633	0,8948
CANUDOS	0,9743	0,9215	0,9069	0,8073	0,9709	2,30	1,14	5,00	137022	12924	0,8404
C. VELHA	0,8883	0,9951	0,9493	0,6285	0,9860	2,48	1,61	5,00	260291	12388	0,8764
CONDOR	0,9396	0,8197	0,7253	0,7818	0,9478	1,97	0,81	5,00	91724	41083	0,7497
CREMAÇÃO	0,8197	0,9360	0,8605	0,7626	0,9417	2,18	1,25	5,00	192931	32180	0,8301
GUAMÁ	0,9609	0,7975	0,7475	0,8285	0,8471	1,94	0,86	5,00	106314	90252	0,7334
JURUNAS	0,9247	0,6966	0,6713	0,7526	0,9452	2,06	0,88	5,00	121653	63559	0,7527
MARAMBAIA	0,9281	0,8975	0,8818	0,8011	0,9007	2,17	1,18	5,00	149544	64341	0,8151
MARCO	0,8516	0,9226	0,8721	0,7542	0,9489	2,21	1,35	5,00	202401	67571	0,8437
FÁTIMA	0,9652	0,8565	0,7149	0,7757	0,8717	2,05	0,89	5,00	119858	13884	0,8037
NAZARÉ	0,2995	0,9980	0,9945	0,6814	0,9963	2,62	2,51	4,00	576236	18776	0,9206
PEDREIRA	0,8777	0,8587	0,7680	0,7742	0,8750	2,02	1,04	5,00	136940	68789	0,8130
REDUTO	0,4345	0,9913	0,9785	0,6903	0,9948	2,56	2,35	4,00	490217	7328	0,9043
SACRAMENTA	0,9750	0,8716	0,7419	0,7770	0,6991	1,92	0,85	5,00	100650	26666	0,7811
SÃO BRAZ	0,7193	0,9771	0,9639	0,7438	0,9866	2,46	1,75	5,00	332925	22124	0,8908
SOUZA	0,8608	0,8803	0,8524	0,7735	0,8539	2,06	1,14	5,00	151747	64299	0,8019
TELÉGRAFO	0,9581	0,8248	0,7456	0,7988	0,8352	2,05	0,95	5,00	105201	44454	0,7793
TERRA FIRME	0,9802	0,6356	0,7595	0,8939	0,6413	1,79	0,73	5,00	75959	59158	0,6919
COMÉRCIO	0,4960	1,0000	0,9953	0,5296	0,9993	2,19	1,92	4,00	442883	5720	0,9042
UMARIZAL	0,7260	0,9880	0,9695	0,7787	0,9872	2,41	1,76	5,00	345271	31190	0,8925

Fonte: FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA.

ESCALA DAS VARIÁVEIS

Um aspecto importante a ser considerado é a homogeneidade entre as variáveis. Ao se agrupar observações é necessário combinar todas as variáveis em um único índice de similaridade, de forma que a contribuição de cada variável depende tanto de sua escala de mensuração como das escalas das demais variáveis. Como as variáveis são medidas em escalas diferentes, com variação diferente, e visando reduzir o efeito dessas diferenças, decidiu-se utilizar as variáveis padronizadas, ou seja,

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

onde:

x_{ij} é o valor para o i-ésimo bairro da j-ésima variável em estudo;

$\bar{x}_j$ é a média aritmética da j-ésima variável em estudo;

s_j é o desvio padrão da j-ésima variável em estudo.

ANÁLISE DE CONGLOMERADOS

As técnicas de análise de conglomerados (Johnson e Wichern, 1992; Morrison, 1976 e Bussab, Miazaki e Andrade, 1990) objetivam dividir um conjunto de observações em um número de grupos homogêneos, segundo algum critério de homogeneidade. Várias são as opções de medidas de homogeneidade e de técnicas para obtenção dos conglomerados. Neste trabalho, usou-se a distância euclidiana como medida de dissimilaridade, e o processo das K-médias como técnica de agrupamento.

Formalmente, tem-se um conjunto de N observações, identificadas pelos seus índices $I = \{1,2,...N\}$.

Um subconjunto qualquer, contendo m_k índices será denotado por

$$Ik = \{J_1(k), J_2(k),..., J_{mk}(k)\}.$$

Deseja-se obter uma partição de I em K subconjuntos, que tornem mínima uma determinada medida de homogeneidade. Deve-se lembrar que uma partição ε de I é uma coleção de subconjuntos $I_1, I_2, ..., I_K$ (K ≤ N), tal que:

(i) $Ii \neq \phi$, i = 1, 2, ...,K (cada subconjunto é não vazio);

(ii) dois subconjuntos não tem pontos em comum $Ii \cap Ij = \phi$, i ≠ j;

(iii) a reunião dos subconjuntos reproduz o conjunto todo:

$I_1 \cup I_2 \cup ... \cup I_K = I$.

Associado a cada observação $J_j(k)$, há um vetor de dados Y, que será indicado por

$$Yj(k) = \{Y_{1j}(k), Y_{2j}(k),..., Y_{pj}(k)\}$$

e para todas as observações do subgrupo I_k, pode-se encontrar as médias das variáveis de interesse, representadas pelo vetor médio

$$\bar{Y}_k = \{\bar{Y}_1(k), \bar{Y}_2(k),..., \bar{Y}_p(k)\},$$

também chamado de centro do grupo, onde

$$\bar{Y}_i(k) = \frac{1}{m_k}\sum_{j=1}^{m_k} Y_{ij}(k) \quad , i = 1, 2,..., p.$$

Pode-se encontrar a distância do j-ésimo elemento do grupo I_k, ao centro do grupo, que será indicada por

$$dj^2(k) = \sum_{i=1}^{p}\left(Y_{ij}(k) - \bar{Y}_i(k)\right)^2 .$$

Quanto menor for esta distância, mais próxima do centro estará a observação. Desse modo, pode-se definir uma medida de homogeneidade para o grupo I_k como

$$d^2(k) = \sum_{j=1}^{m_k} d_j^2(k) .$$

Analogamente, quanto menor for $d^2(k)$, mais próximas do centro estarão as m_k observações e, portanto, mais parecidos entre si serão estes pontos. Finalmente, para cada partição ε do conjunto I, é possível encontrar a medida

global de homogeneidade induzida por essa partição, ou seja,

$$d^2(\varepsilon) = \sum_{k=1}^{K} d^2(k) .$$

A partição ε que produz o menor valor de $d^2(\varepsilon)$ é a partição que agrupa os elementos mais parecidos entre si. Entretanto, por ser impraticável explorar todas as possíveis partições de N elementos em K grupos, para encontrar a que produza o menor valor de $d^2(\varepsilon)$, procura-se, por processo iterativo, encontrar uma solução que, embora não sendo a melhor, esteja bem próxima desta. Este é o algoritmo das K-médias. Inicia-se por uma partição arbitrária de K grupos e através de mudanças sucessivas de elementos, de um grupo para outro, tenta-se encontrar a melhor partição.

Uma questão delicada e difícil de resolver em análise de conglomerados é a fixação do número final K de grupos desejados. O procedimento mais comum é utilizar vários valores de K, e por algum critério "ótimo", selecionar o mais conveniente.

O critério adotado será o baseado no princípio de análise de variância, o qual objetiva medir a diminuição dos resíduos quadráticos ao se passar de uma partição com J a outra com K grupos, com J < K. Assim, na partição ε_J, com J grupos, a soma de quadrados dos resíduos devida à i-ésima variável é dada por

$$SQR_i(\varepsilon_J) = \sum_{j=1}^{J} \sum_{l=1}^{mj} \left(Y_{ijl} - \bar{Y}_{ij}\right)^2 .$$

Uma medida para avaliar a diminuição na soma de quadrados dos resíduos devida à i-ésima variável, ao passar de J para K grupos, é dada pela estatística

$$F_i(J/K) = \frac{SQR_i(\varepsilon_J) - SQR_i(\varepsilon_K)}{SQR_i(\varepsilon_K)} \cdot \frac{(N-K-1)}{(K-J)} ,$$

que sob determinadas condições, tem distribuição conhecida. A generalização para mais de uma variável pode ser construída através da soma das médias correspondentes às p variáveis, isto é,

$$SQR(\varepsilon_J) = \sum_{i=1}^{p} SQR_i(\varepsilon_J)$$

e

$$F(J/K) = \frac{SQR(\varepsilon_J) - SQR(\varepsilon_K)}{SQR(\varepsilon_K)} \cdot \frac{(N-K-1)}{(K-J)}$$

a qual pode ser usada como medida descritiva da diminuição relativa dos resíduos quadráticos. Grandes valores da estatística F indicam diminuição significante nestes resíduos, quando se passa de J para K grupos.

ANÁLISE DESCRITIVA

A Tabela 2 apresenta o valor mínimo, o valor máximo, a média, os três quartis, o desvio padrão e o coeficiente de variação das 11 variáveis (Morettin e Bussab, 1987 e Toledo e Ovalle, 1985).

i) **Percentual de Domicílios Particulares Permanentes com Localização Casa.**

Essa variável apresentou o valor mínimo de 0,2995, referente a uma baixa percentagem de casas no bairro de Nazaré. Esse valor pode ser explicado pelo fato do bairro apresentar uma grande concentração de prédios de apartamento.

ii) **Percentual de Domicílios Particulares Permanentes com Condição de Ocupação.**

Esta variável apresentou o valor mínimo de 0,5773 referente ao bairro da Terra Firme e o valor máximo de 0,8954 referente ao bairro do Comércio. O primeiro pode ser explicado pelo fato do bairro apresentar um grande número de invasões, onde as pessoas consideram-se proprietárias de seus domicílios.

O segundo é um bairro com grande número de estabelecimentos comerciais.

iii) Percentual de Domicílios Particulares Permanentes com Lixo Coletado.

O bairro da Terra Firme foi o que apresentou o menor percentual de domicílios com lixo coletado (0,6413). Esse bairro tem regiões de difícil acesso.

observa-se também que alguns bairros apresentam um número médio de banheiros por domicílio menor que um, o que indica que nesses bairros existe mais de um domicílio cujos moradores utilizam o mesmo banheiro.

Com relação às variáveis que medem percentual, a única que apresenta valor máximo igual a 1 é o Percentual de domicílios particulares permanentes com abastecimento de água com canalização interna (PAA). Isto significa que existe pelo menos um bairro onde todos os domicílios têm esse serviço. Para os outros percentuais nenhum bairro é atendido em 100% dos casos.

A característica que menos aparece nos domicílios é a condição de ocupação, que apresenta os menores percentuais.

Tabela 2

Medidas descritivas das variáveis

Variáveis	Valor Mínimo	Valor Máximo	Media	Primeiro Quartil	Mediana	Terceiro Quartil	Desvio padrão	Coef. de Variação
PLO	0,2995	0,9802	0,8030	0,7227	0,8830	0,9595	0,2096	26,10%
PAA	0,6356	1,0000	0,8929	0,8407	0,9095	0,9886	0,1022	11,45%
PCO	0,5773	0,8954	0,7545	0,7253	0,7738	0,7903	0,0705	9,34%
PLC	0,6413	0,9993	0,9108	0,8628	0,9465	0,9869	0,0993	10,90%
PIS	0,6713	0,9953	0,8531	0,7466	0,8663	0,9632	0,1092	12,80%
MBD	0,73	2,51	1,36	0,89	1,16	1,36	0,55	40,44%
MDD	1,79	2,62	2,20	2,04	2,17	2,44	0,24	10,91%
MPD	3,71	5,26	4,77	4,56	4,90	5,03	0,41	8,60%
POP	5.720	90.252	38.316	16.330	31.685	63.929	25.241	65,88%
PPA	0,6819	0,9206	0,8254	0,7802	0,8226	0,8917	0,0663	8,03%
RMC	74.959	576.236	231.722	113.086	150.646	339.098	157.959	68,17%

iv) Número Médio de Pessoas por Domicílio.

O bairro do Comércio é um dos mais antigos e por isso a grande maioria das famílias ali residentes são compostas por pessoas idosas que moram sozinhas; com isso, nesse bairro, o número médio de pessoas por domicílio é o mais baixo (3,71).

A Tabela 2, mostra ainda, que a maioria das variáveis tem baixo coeficiente de variação, com exceção das variáveis População (POP), Rendimento Médio Nominal dos Chefes (RMC) e Número Médio de Banheiros por Domicílio (MBD). Com relação a esta última variável,

ANÁLISE ESTATÍSTICA

Utilizando-se o método das K-médias contido no programa "SPSS for Windows 6.0", atribuiu-se o valor J=3. Desse modo, obtivemos, para o Município de Belém, 3 grupos homogêneos formados pelos bairros apresentados na Tabela 3.

Tabela 3

Grupo de Bairros do Município de Belém

Grupos	Bairros
A	Batista Campos, Nazaré, Reduto, São Braz, Comércio, Umarizal e Cidade Velha
B	Canudos, Cremação, Marambaia, Marco, Pedreira e Souza
C	Condor, Guamá, Jurunas, Fátima, Sacramenta, Telégrafo e Terra Firme

Para melhor entendimento das características dessas regiões homogêneas foi elaborada a Tabela 4, que contém as médias dos 20 bairros, as médias de cada grupo e o valor das estatísticas F , que é uma medida da diminuição dos resíduos, ao se passar de 1 para 3 grupos.

Assim, observou-se que as variáveis menos atuantes na separação dos bairros foram a População (POP), o Percentual de Domicílios Particulares Permanentes com Lixo Coletado (PLC) e o Percentual de Domicílios Particulares Permanentes em Condição de Ocupação (PCO).

Tabela 4
Média das variáveis dos Grupos do Município de Belém

Variáveis	Média				Estatística F
	Total	Grupo A	Grupo B	Grupo C	
PLO	0,8030	0,5259	0,8858	0,9577	17,69
PAA	0,8929	0,9906	0,9160	0,7860	25,02
PCO	0,7545	0,6964	0,7545	0,8014	10,03
PLC	0,9108	0,9919	0,9253	0,8268	8,69
PIS	0,8531	0,9774	0,8701	0,7294	96,80
MBD	1,3605	2,0883	1,2443	0,8529	512,22
MDD	2,1960	2,4533	2,2029	1,9686	32,71
MPD	4,7735	4,2533	4,8657	5,1271	15,17
POP	38.316	17.462	46.070	48.437	6,08
PPA	0,8254	0,9012	0,8314	0,7545	50,44
RMC	231.722	447.036	175.839	103.051	42,04

Os grupos de bairros do Município de Belém apresentam os seguintes perfis:

- Grupo A: formado por bairros considerados de classe alta.
- Grupo B: formado por bairros considerados de classe média.
- Grupo C: formado por bairros considerados de classe baixa.

CONCLUSÕES

O propósito deste trabalho foi construir regiões homogêneas no Município de Belém, segundo algumas variáveis demográficas, possibilitando um estudo das condições sócio - econômicas dos bairros. Os grupos obtidos, quando analisados sob o ponto de vista qualitativo e geográfico, mostraram resultados interessantes. Por exemplo, é bem conhecido o caráter de desenvolvimento sócio-econômico e habitacional do grupo **A**, ou seja , sabe-se que é um grupo composto por bairros centrais onde o sistema de saneamento, educação e saúde são os melhores. Nessa região, reside a maioria da população considerada de classe alta. O grupo **B** é composto por bairros com características intermediárias; são considerados de classe média. Já no Grupo **C**, encontram-se os bairros considerados mais pobres, localizados na periferia do município, onde a maioria dos espaços foram invadidos e habitados, em parte, por pessoas provenientes de outros municípios do Estado.

BIBLIOGRAFIA CONSULTADA

BUSSAB, Wilton O ., DINI, Nádia P. **Pesquisa em Desemprego. SEADE/DIEESE:** Regiões Homogêneas da Grande São Paulo. Rev . Fundação SEAD. São Paulo: São Paulo em Perspectiva, set./dez. 1985.

BUSSAB, Wilton de Oliveira, MIAZAKI, Édina Shizue, ANDRADE, Dalton Francisco de. **Introdução à análise de agrupamento**. Simpósio Nacional de Probabilidade e Estatística da Associação Brasileira de Estatística - ABE. IME, 9. São Paulo: USP, 1990.

FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Censo Demográfico 1991. Pará: Tabulações Especiais. Rio de Janeiro, 1993

JOHNSON, Richard A., WICHERN, Dean W. **Applied Multivariate Statistical Analysis. Prentice Hall**. New Jersey, 1992.

MORENTTIN, Pedro A., BUSSAB, W. O. **Estatística Básica.** 4. ed. São Paulo: Atual, 1987.

MORRISON, D F. **Multivariate statistical methods.** New York: Mc Graw Hill Book .,1976.

TOLEDO, Geraldo Luciano, OVALLE, Ivo. **Estatística Básica**. 2. ed. São Paulo: Atlas, 1985