CADERNOS DO IME – Série Estatística

Universidade do Estado do Rio de Janeiro - UERJ
Rio de Janeiro - RJ - Brasil
ISSN impresso 1413-9022 / ISSN on-line 2317-4536 - v.34, p.17 - 31, 2013

APLICAÇÃO DA COMPOSIÇÃO PROBABILÍSTICA E DO MÉTODO DAS K-MÉDIAS À CLASSIFICAÇÃO DE MUNICÍPIOS QUANTO À OFERTA DE CRECHES

Annibal Parracho Sant'Anna
Universidade Federal Fluminense
annibal.parracho@gmail.com

Flavia Faria
Instituto Federal de Educação, Ciência e Tecnologia Fluminense
flavia_faria@uol.com.br

Helder Gomes Costa
Universidade Federal Fluminense
helder.uff@gmail.com

Resumo

Neste trabalho o método CPP-TRI, que emprega a composição probabilística de preferências para classificar em categorias ordenadas, e o método de identificação de conglomerados das k-médias são usados para classificar os municípios fluminenses quanto ao cumprimento da disposição constitucional de disponibilidade de creches para a população infantil. Consideramos a segmentação dos municípios do Estado do Rio de Janeiro visando a avaliar a relação entre recursos e produtos na oferta de creches. As variáveis consideradas são, de um lado, a proporção de crianças de dois a cinco anos atendidas em creches e, de outro lado, a despesa pública municipal per capita com educação e cultura e a renda per capita do município. O método das k-médias foi empregado para segmentar esses municípios em estudos anteriores com base nessas três variáveis. Verificou-se concordância entre o agrupamento realizado pelas k-médias e pela CPP-TRI com base nas relações custo/benefício.

***Palavras-chave**: Composição Probabilística de Preferências, k-médias, Educação Infantil, Eficiência, Avaliação de Gestão Municipal.*

1. Introdução

Dentre os problemas de decisão destaca-se o de *sorting* ou classificação ordenada, conforme denominado em Costa *et al*. (2007). Sant'Anna *et al*. (2013) propôs o CPP-TRI, que se baseia no uso de composições probabilísticas como método de reduzir a subjetividade ao se tratar a classificação de objetos em categorias ordenadas entre si.

Observa-se que ao se empregar variáveis conflitantes, do tipo caracterizadas como custos e benefícios, à avaliação da eficiência de unidades, a comparação de desempenhos pode ser mais informativa se as unidades forem organizadas em subgrupos homogêneos.

Este trabalho propõe o uso da técnica k-médias para identificar agrupamentos homogêneos, como etapa prévia a modelagem de problemas de classificação pelo CPP-TRI. Como pano de fundo para descrição desta proposta, toma-se o problema de da classificação dos municípios do Estado do Rio de Janeiro em classes ordenadas quanto à oferta de creches.

2. Agrupamentos pelas k-médias

Conforme Samoilenko e Osei-Bryson (2008), Johnson e Wichern (1998) e Witten e Frank (2005), a k-médias é uma técnica não-hierárquica de agrupamento que visa particionar um conjunto de elementos numa coleção de k agrupamentos, onde cada elemento é alocado no agrupamento de cujo centróide se encontra mais próximo, por meio das seguintes etapas:

a) Escolha de k centroides, que serão inicialmente os centros dos k grupos.
b) Alocar cada elemento no agrupamento de cujo centróide esse elemento está mais próximo, em termos de distância Euclideana.
c) Determinar um novo centroide para cada grupo, com coordenadas dadas pelas médias das coordenadas das observações alocadas no grupo.
d) Realocação de cada elemento no grupo cujo centróide esteja mais próximo.
e) Repetição dos passos (c), (d) e (e), até que nenhum elemento mude de agrupamento.

Há vários procedimentos para a definição do número de agrupamentos, avaliando a qualidade da classificação. Segundo Milligan e Cooper (1985) e Mooi e Sarstedt (2011), o mais eficiente é o razão de variância (VRC), para n objetos e k grupos:

$$VRC_k = \left(SS_B / (k-1)\right) / \left(SS_W / (n-k)\right) \quad (1)$$

onde SS_B é a soma dos quadrados entre os grupos e SS_W a soma dos quadrados dentro dos grupos. O número adequado de grupos é o valor de k que minimiza ω_k, com:

$$\omega_k = \left(VRC_{k+1} - VRC_k\right) - \left(VRC_k - VRC_{k-1}\right) \quad (2)$$

3. Composição Probabilística de Preferências aplicada a Classificação

CPP-TRI (SANT'ANNA *et al*., 2012) é um método de classificação multicritério baseado em relações de sobreclassificação que prescinde da atribuição de pesos aos critérios. Este método tem a mesma estrutura do ELECTRE-TRI-nC (ALMEIDA *et al*., 2011), distinguindo-se por substituir o uso de patamares de indecisão pela adição de perturbações aleatórias para levar em conta a imprecisão nas avaliações. Assim os valores segundo cada critério, tanto para as avaliações das alternativas quanto para os perfis característicos das classes são tratados como parâmetros de locação de distribuições de probabilidades e as comparações são realizadas entre tais distribuições. Isto permite que a credibilidade das relações e a composição dos critérios sejam baseadas em probabilidade conjuntas em vez de em médias ponderadas.

As medidas exatas iniciais são tratadas como parâmetros de locação das distribuições. Seguindo os princípios da modelagem econométrica clássica, assume-se, além de distribuições normais, idêntica distribuição e independência entre perturbações provocando imprecisão em diferentes medidas. Caso se disponha de informação aconselhando outras distribuições, tal informação pode ser usada sem alterações substanciais nos cálculos.

Uma vez representada a avaliação segundo cada critério por uma distribuição de probabilidade, é possível calcular a probabilidade de cada alternativa ter uma avaliação acima ou abaixo dos perfis de cada classe, ainda segundo cada critério. Dessas probabilidades de sobreclassificação segundo cada critério é, então, possível derivar classificações globais sem atribuir pesos aos critérios, usando as abordagens já consideradas para composição probabilística de preferências por Sant'Anna (2002).

Caso seja possível e desejável atribuir pesos aos critérios, em vez de probabilidades conjuntas, pode-se usá-los tratando as probabilidades de preferência segundo cada critério como probabilidades condicionais e os pesos como probabilidades marginais dos critérios.

Formalmente, considera-se o problema de classificar uma alternativa *A* com avaliação a_l segundo o *l*-ésimo critério, para *l* variando de 1 a *m*, com *m* denotando o número de critérios, em uma dentre *k* classes ordenadas, cada uma identificada por um certo número *n* de perfis de referência, constituídos por avaliações de alternativas previamente construídas. Denote-se por C_{ijl} a avaliação pelo *l*-ésimo critério que aparece no *j*-ésimo perfil da *i*-ésima classe. Os perfis são definidos de modo que as classes estão efetivamente ordenadas em ordem crescente, essas diferenças constituem uma sucessão decrescente. Isto é, se, para $i_1<i_2$, j_1 e j_2 quaisquer, para algum *l*, observa-se $C_{i1j1l} > C_{i2j2l}$, então os perfis precisam ser revistos.

As medidas exatas a_l e C_{ijl} são usadas como médias de distribuições de variáveis aleatórias X_l e Y_{ijl}. A_{il}^+ e A_{il}^- representam, respectivamente, a probabilidade de a alternativa *A* apresentar valor respectivamente acima e abaixo dos valores informados para o critério *l* nos perfis da classe i. Assumindo independência,

$$A_{il}^+ = \prod_j P[X_l > Y_{ijl}] \text{ e } A_{il}^- = \prod_j P[X_l < Y_{ijl}]. \quad (3)$$

Para as credibilidades A_i^+ e A_i^- de a alternativa *A* estar, respectivamente, em classe acima ou abaixo da classe *i*-ésima serão usados, respectivamente, os produtos dos A_{il}^+ e dos A_{il}^-, para *l* variando ao longo dos critérios.

O procedimento de classificação é baseado na comparação das diferenças $A_i^+ - A_i^-$. A regra de classificação é simples: a alternativa A pertence à classe i para qual essa diferença é mais próxima de zero.

Um algoritmo para aplicar essa regra pode ser desenvolvido com duas etapas. Primeiro, se identifica o menor valor de $_i$ para o qual é negativa a diferença $A_i^+ - A_i^-$. Se para este valor de i, a classe *i*-ésima é a primeira classe, a alternativa é classificada nesta classe. Caso contrário, comparamos os valores absolutos das diferenças $A_i^+ - A_i^-$ para tal classe e para a que a precede e classificamos a alternativa naquela em que esse valor seja menor.

Esta regra pode ser exposta formalmente da seguinte forma, denotando por C(A) a classe em que a alternativa A vem a ser classificada.

Partimos da classificação provisória $CP(A) = \min\{i: A_i^+ - A_i^- < 0\}$.

Se CP(A)=1, a alternativa pertence à classe 1.

Se CP(A) >1, se $A_{CP(A)}^- - A_{CP(A)}^+ < A_{CP(A)-1}^+ - A_{CP(A)-1}^-$, então C(A) = CP(A).

Caso contrário, C(A) = CP(A)-1.

Para oferecer informação sobre a incerteza na classificação final, classificações alternativas resultantes da aplicação de planos de corte menos exigentes para as credibilidades de localização acima ou abaixo dos perfis da classe são produzidas. Esses planos de corte são identificados por percentuais aplicados para reduzir a exigência de um ou do outro lado da classe. Assim, uma classificação benevolente para a alternativa *A* com plano de corte determinado pelo percentual *c* a colocará na classe $Cc(A)^+$ para a qual seja mínimo o valor absoluto da diferença $A_i^+ - cA_i^-$. Do mesmo modo, a classificação pessimista $Cc(A)^-$ será a classe para a qual seja mínimo o valor absoluto da diferença $cA_i^+ - A_i^-$.

Os valores de $Cc(A)^+$ e $Cc(A)^-$ podem ser obtidos por procedimentos de classificação ascendente e descendente desenvolvidos de forma análoga à descrita na seção anterior, que corresponde aos planos de corte com o percentual de 100%.

Este método se aplicará adiante ao caso aqui estudado, com um único perfil para cada classe.

4. Caracterização da área em estudo

Foi considerado o conjunto formado por 90 municípios do Estado do Rio de Janeiro. Ressalta-se que há ausência de dados mais recentes para as variáveis utilizadas na modelagem. A escolha das variáveis foi baseada no trabalho desenvolvido por Faria (2005): gastos *per capita* com educação e cultura (GEDUC) e o valor do rendimento médio mensal dos responsáveis pelos domicílios particulares permanentes (RENDA), como *inputs* do modelo em questão. As despesas consideradas aqui se referem ao ano de 2000 e foram obtidas na Secretaria do Tesouro Nacional do Ministério da Fazenda. O *input* RENDA, obtido no Censo Demográfico de 2000 e calculado em nível municipal, deve ser considerado como uma variável exógena, ambiental ou não discricionária (LINS e MEZA 2000, EMROUZNEJAD 2001), introduzida no modelo para levar em conta os efeitos que um padrão mais elevado de renda pode ter sobre o *output*, independentemente do nível de gasto público alocado.

Já a variável considerada como *output* foi definida como proporção de crianças de 2 a 5 anos matriculadas em creches ou em escolas de educação infantil (PPCRECH). Este indicador também foi obtido do Censo Demográfico de 2000, realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE). O Quadro 1 mostra as variáveis em análise e a Tabela 1 mostra os valores dessas variáveis por município.

Quadro 1 – Indicadores selecionados para o estudo

Indicador	Definição	Variável	Fonte
Gastos com Educação e Cultura (GEDUC)	Gastos anuais municipais *per capita* em educação e cultura, calculados como a razão dos gastos informados na rubrica, Educação e Cultura em 2000, pelo total de residentes no município em 2000.	*Input*	STN – Secretaria do Tesouro Nacional
RENDA	Valor do rendimento médio mensal dos responsáveis pelos domicílios particulares permanentes.	*Input*	CENSO 2000
Proporção de crianças de 2 a 5 anos matriculadas em creches ou em escolas de educação infantil (PPCRECH)	Proporção de crianças de 2 a 5 anos matriculadas em creches ou escolas de educação infantil.	*Output*	CENSO 2000

Fonte: Faria (2005)

Tabela 1. Gastos em Educação e Cultura, Renda per capita e Crianças em Creches

	Municípios	GEDUC (R$)	RENDA (R$)	PPCRECH (%)
1	Aperibé	307,14	435	65,6
2	Angra dos Reis	262,83	732	14,8
3	Araruama	137,92	624	17,6
4	Areal	353,16	540	32,8
5	Armacão de Búzios	426,64	764	20,9
6	Arraial do Cabo	131,57	686	36,9
7	Barra do Piraí	81,58	653	40,8
8	Barra Mansa	190,05	671	18,6
9	Belford Roxo	68,62	461	7,6
10	Bom Jardim	166,07	508	36,3
11	Bom Jesus do Itabapoana	135,37	548	39,6
12	Cabo Frio	141,45	737	20,7
13	Cachoeiras de Macacu	135,78	534	29,4
14	Cambuci	249,87	402	53,7
15	Campos dos Goytacazes	135,78	588	33
16	Cantagalo	256,75	563	52,8
17	Carapebus	586,03	455	52,9
18	Cardoso Moreira	286,02	345	42,3
19	Carmo	234,28	522	45,9
20	Casimiro de Abreu	240,12	640	39,4
21	Comendador Levy Gasparian	335	466	41,3
22	Conceição de Macabu	133,27	477	45
23	Cordeiro	112,97	641	49,4
24	Duas Barras	224,94	460	58,3
25	Duque de Caxias	98,94	539	6,2
26	Engenheiro Paulo de Frontin	142,07	488	38,4
27	Guapimirim	119,66	566	14,7
28	Iguaba Grande	237,51	765	18,5
29	Itaboraí	127,75	483	9,8
30	Itaguaí	164,77	597	25,7
31	Italva	205,64	422	32,1
32	Itaocara	157,5	484	51,8
33	Itaperuna	132,64	608	36,9
34	Itatiaia	428,59	779	46,7
35	Japeri	120,85	397	9,5
36	Laje do Muriaé	278,78	390	52
37	Macaé	266,89	928	48
38	Macuco	251,91	500	13,5
39	Magé	78,62	498	4,6
40	Mangaratiba	318,72	802	57,9
41	Maricá	123,3	752	23

42	Mendes	114,11	571	44,4
43	Miguel Pereira	184,53	784	30,4
44	Miracema	102,63	494	58,4
45	Natividade	264,24	453	55
46	Nilópolis	56,6	702	8,9
47	Niterói	133,89	1741	23,2
48	Nova Friburgo	146,18	753	29,5
49	Nova Iguaçu	75,51	560	6,4
50	Paracambi	135,93	548	30,7
51	Paraíba do Sul	106,61	552	43,1
52	Parati	137,02	725	16,1
53	Paty do Alferes	178,03	480	22,8
54	Petrópolis	163,33	894	19,4
55	Pinheiral	167,47	598	22
56	Piraí	430,81	588	47,3
57	Porciúncula	142,14	474	62,7
58	Porto Real	200,29	516	32,6
59	Quatis	209,77	558	25,4
60	Queimados	107,7	483	5,3
61	Quissamã	793,5	421	46,5
62	Resende	161,11	899	23,1
63	Rio Bonito	10621,97	599	36,2
64	Rio Claro	168,28	484	41,9
65	Rio das Flores	281,63	428	50,7
66	Rio das Ostras	352,78	812	36,8
67	Rio de Janeiro	139,82	1354	18,9
68	Santa Maria Madalena	315,77	423	52,8
69	Santo Antônio de Pádua	149,39	527	45,4
70	São Fidélis	117,19	425	37
71	São Francisco de Itabapoana	0,12	335	50,5
72	São Gonçalo	39,58	614	8,2
73	São João da Barra	277,26	421	57,5
74	São João de Meriti	50,55	547	6,5
75	São José de Ubá	202,67	426	27
76	São José do Vale do Rio Preto	135,96	481	27,9
77	São Pedro da Aldeia	102,05	677	15,4
78	São Sebastião do Alto	313,24	355	58,8
79	Sapucaia	204,12	502	36,1
80	Saquarema	117,23	618	27,4
81	Silva Jardim	187,87	451	34
82	Sumidouro	233,32	484	7,7
83	Tanguá	164,87	417	11,7
84	Teresópolis	151,73	811	15,2
85	Trajano de Morais	0,28	401	48,1
86	Três Rios	53,19	599	37,8
87	Valença	106,65	601	47,5
88	Varre-Sai	308,45	388	32,1
89	Vassouras	114,74	615	36,9
90	Volta Redonda	238,14	834	23,2

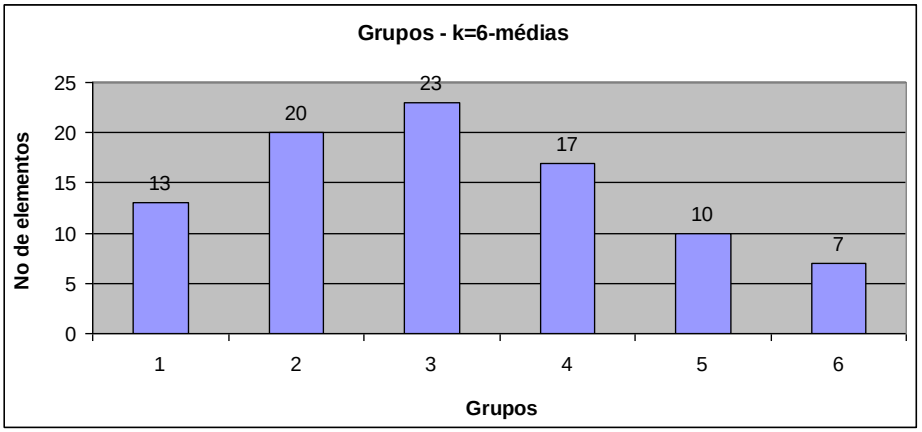
Fonte: Autores, 2013

5. Aplicação da técnica de agrupamentos k-médias para determinar os grupos

Os centros iniciais foram escolhidos como pontos igualmente espaçados segundo o *output* PPCRECH, uma vez que se busca avaliar o comportamento dos municípios quanto à esta oferta. Para determinar o número ótimo de grupos, realizaram-se simulações com diferentes valores de $k = 2,...,8$, sendo calculado o valor de ω_k para cada solução de grupo, concluindo que $k = 6$ grupos é o mais adequado para o conjunto de dados em análise.

Conforme pode ser observado do Gráfico 1, os agrupamentos formados se encontram parcialmente de acordo com os parâmetros limitantes da técnica k-médias estipulados por Samoilenko e Osei-Bryson (2008), que estabelece que todos os grupos devem ter uma quantidade ($k_{\min}$) superior ou igual a 10% do número total de elementos a serem categorizados. Isto se deve ao fato de que, para $k = 2,3,4$ e 5, o número de elementos de cada um dos grupos formados atendeu ao limite de veto). Porém, os grupos formados pela técnica k-médias para $k = 6$ atenderam parcialmente a este limite de veto, uma vez que um único grupo (grupo 6) foi constituído por menos de 10% do total dos municípios, como mostra o Gráfico 1.

Gráfico 1 – Grupos formados pela aplicação do k-médias



Fonte: Autores, 2013

Apesar dos agrupamentos para $k = 6$ atenderem parcialmente as restrições da k-médias descritas em Samoilenko e Osei-Bryson (2008), esta partição se mostrou mais robusta estatisticamente.. Isto também é notado nos gráficos 2, 3 e 4, que mostram a redução de *outliers* quando se aplica o *k*-médias com $k = 6$. Os *outliers* dos grupos formados pelo k-médias com $k = 5$ se encontram no grupo 5 e correspondem aos seguintes municípios: Aperibé, Cambuci, Natividade e Porciúncula. Nos grupos formados pelo *k*-médias com $k = 6$, apenas um *outlier*, o município de Aperibé, foi destacado neste gráfico, também pertencente ao grupo 6. Já no caso em que $k = 7$, têm-se, ao todo, seis *outliers*, sendo três pertencentes ao grupo 2, enquanto os demais pertencem, cada um, aos grupos 3, 6 e 7. Neste caso, os *outliers* são: Piraí, Porciúncula, Porto Real, Barra do Piraí, Natividade, Aperibé. Isto corrobora para a escolha do $k = 6$.

Gráfico 2 – Boxplot dos grupos formados pelo *k*- médias com $k = 5$

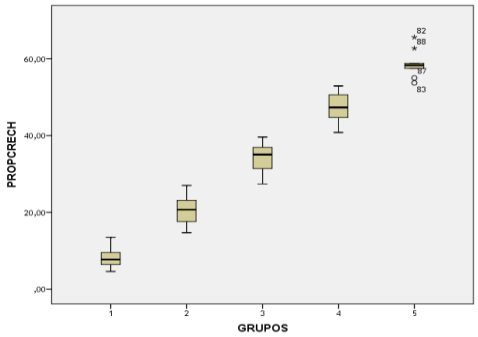


Gráfico 3 – Boxplot dos grupos formados pelo k-médias com $k = 6$

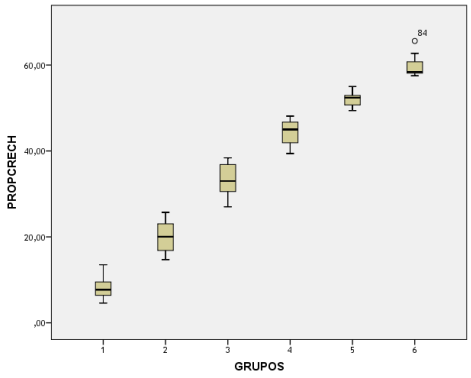
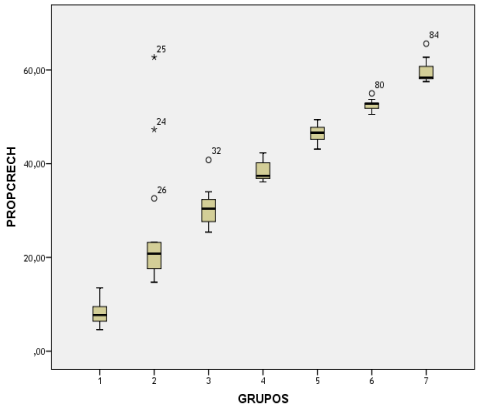


Gráfico 4 – Boxplot dos grupos formados pelo k-médias com $k = 7$



6. Classificação pela CPP-TRI

Na aplicação da CPP-TRI os perfis de referência foram dados pelas médias dos vetores de valores das razões *input/output* nos seis grupos determinados pela aplicação da *k*-médias. No caso do grupo 4, excluiu-se o *outlier* Rio Bonito, cujo valor das

despesas com educação é mais de 30 vezes superior ao de qualquer outro município do grupo. Os perfis de referência são apresentados na Tabela 2.

Tabela 2. Perfis de Referência

Grupo	Educação/Creches	Renda/Creches
1	4,1	8,3
2	4,8	8,8
3	5,2	12,9
4	5,5	17
5	9,2	40,5
6	14	69,2

Fonte: Autores, 2013

O grupo de menor tamanho pela CPP-TRI é o grupo 4 com 9 municípios. No grupo 5 ficaram 10 e no grupo 1 ficaram 11. Os maiores são os grupos 2, 3 e 6 com 24, 19 e 17 municípios, respectivamente. Embora os tamanhos dos grupos sejam, em alguns casos, muito diferentes dos obtidos pela *k*-médias, o total de municípios no conjuntos dos três primeiros e dos três últimos grupos é igual nas duas classificações. A Tabela 3 apresenta as classificações inferior, central e superior pela CPP-TRI precedida da classificação pela k-médias. As classificações superior e inferior são obtidas da flexibilização para um ponto de corte de 50%.

Tabela 3. Classificações dos Municípios

Município	k-médias	CPPinferior	CPPcentral	CPPsuperior
Aperibé	6	3	5	5
Duas Barras	6	4	6	6
Mangaratiba	6	3	3	3
Miracema	6	6	6	6
Porciúncula	6	6	6	6
São João da Barra	6	3	5	5
São Sebastião do Alto	6	3	5	5
Cambuci	5	3	5	5
Cantagalo	5	3	4	5
Carapebus	5	2	2	3
Cordeiro	5	5	6	6
Itaocara	5	4	6	6
Laje do Muriaé	5	3	5	5
Natividade	5	3	5	5
Rio das Flores	5	3	4	4
Santa Maria Madalena	5	3	4	4
São Francisco de Itabapoana	5	6	6	6
Barra do Piraí	4	4	6	6
Bom Jesus do Itabapoana	4	4	6	6
Cardoso Moreira	4	3	3	3
Carmo	4	3	4	5
Casimiro de Abreu	4	3	3	3
Comendador Levy Gasparian	4	2	3	3
Conceição de Macabu	4	4	6	6
Itatiaia	4	2	3	3

Macaé	4	3	3	3
Mendes	4	4	6	6
Paraíba do Sul	4	5	6	6
Piraí	4	2	3	3
Quissamã	4	1	2	2
Rio Claro	4	3	5	5
Santo Antônio de Pádua	4	4	6	6
Trajano de Moraes	4	6	6	6
Valença	4	5	6	6
Areal	3	2	2	3
Arraial do Cabo	3	3	4	5
Bom Jardim	3	3	4	5
Cachoeiras de Macacu	3	3	3	4
Campos dos Goytacazes	3	3	4	5
Engenheiro Paulo de Frontin	3	3	5	5
Italva	3	3	3	4
Itaperuna	3	3	5	5
Miguel Pereira	3	2	3	3
Nova Friburgo	3	3	3	3
Paracambi	3	3	4	5
Porto Real	3	3	3	3
Rio Bonito	3	1	1	1
Rio das Ostras	3	2	2	3
São Fidélis	3	4	6	6
São José de Ubá	3	2	3	3
São José do Vale do Rio Preto	3	3	4	4
Sapucaia	3	3	3	3
Saquarema	3	3	4	4
Silva Jardim	3	3	3	4
Três Rios	3	5	6	6
Varre-Sai	3	2	2	3
Vassouras	3	3	6	6
Angra dos Reis	2	1	1	1
Araruama	2	2	2	3
Armação de Búzios	2	1	1	1
Barra Mansa	2	2	2	2
Cabo Frio	2	2	2	3
Guapimirim	2	2	2	3
Iguaba Grande	2	1	2	2
Itaguaí	2	2	3	3
Maricá	2	2	3	3
Niterói	2	1	2	2
Parati	2	2	2	2
Paty do Alferes	2	2	3	3
Petrópolis	2	2	2	2
Pinheiral	2	2	2	3
Quatis	2	2	3	3
Resende	2	2	2	3
Rio de Janeiro	2	1	2	2
São Pedro da Aldeia	2	2	2	3
Teresópolis	2	1	2	2
Volta Redonda	2	2	2	2
Belford Roxo	1	1	2	2
Duque de Caxias	1	1	1	1
Itaboraí	1	1	1	2
Japeri	1	1	2	2
Macuco	1	1	1	1
Magé	1	1	1	1
Nilópolis	1	1	2	2
Nova Iguaçu	1	1	1	1
Queimados	1	1	1	1
São Gonçalo	1	2	2	2
São João de Meriti	1	1	1	2
Sumidouro	1	1	1	1
Tanguá	1	1	2	2

Fonte: Autores, 2013

Na Tabela 3, há 66 municípios em que o valor da classificação inicial está entre os limites da CPP-TRI e 24 municípios com classificações discordantes. Destes 24, apenas Mangaratiba, classificado no grupo 1 pelas *k*-médias e no grupo 4 pela CPP-TRI, apresenta um afastamento maior que 2 entre as duas classificações. Este nível de concordância foi considerado alto, dada a ausência de correlação encontrada entre os municípios para as três variáveis e a alta proximidade entre as classificações benevolente e exigente pela CPP-TRI. Uma segunda aplicação da CPP-TRI foi realizada, desta vez construindo-se perfis de referência por um critério baseado no princípio de igualar o espaçamento entre os perfis. As coordenadas dos perfis de referência são quartis igualmente espaçados nas distribuições das avaliações segundo cada critério das observações; Para construir perfis representativos das 6 classes, foram usados afastamentos de 1/6. Para as classes extremas foram usados os quartis de 1/12 e 11/12. Assim, para a classe no extremo inferior da população de 90 municípios foram usadas médias aritméticas entre os valores das posições 7 e 8 e para a classe no extremo superior médias aritméticas entre os das posições 82 e 83. As classes intermediárias foram formadas mediante afastamentos de 15 em 15 posições, isto é, com médias aritméticas entre os valores das posições 22,5, 37,5, 52,5 e 67,5. Os perfis representativos obtidos foram os da Tabela 4.

Tabela 4. Perfis Equidistantes

Grupo	Educação/Creches	Renda/Creches
1	2	8
2	4	11
3	5	15
4	7	19
5	9	36
6	17	74

Fonte: Autores, 2013

O número de municípios cuja classificação pelas k-médias fica fora da classificação intervalar pela CPP é igual a 14. Observam-se 69 classificações pontuais coincidentes nas duas classificações e 21 divergentes. Nenhuma diferença entre as duas classificações é maior que 1, isto é, quando dois municípios não estão na mesma classe estão em classes vizinhas. Há interseção entre todas as classificações intervalares.

7. Conclusão

O uso da composição probabilística de preferências para a classificação dos municípios seja a partir de uma classificação pelas *k*-médias seja a partir de perfis equidistantes conduziu a considerável concordância entre as classificações benevolente e exigente. Observou-se, também, a concordância destas classificações com a oriunda da aplicação isolada das k-médias.

Com a concordância entre os resultados das diferentes formas de classificação, foi possível estabelecer a validade da segmentação dos municípios em grupos homogêneos para a avaliação das relações entre os insumos e produtos considerados.

Com isto, torna-se possível identificar características das variáveis analisadas que podem ser aproveitadas em outras avaliações de desempenho na administração pública municipal.

Referências

ALMEIDA-DIAS, J.; FIGUEIRA, J. R. ; ROY, B. Electre Tri-C: A multiple criteria sorting method based on characteristic reference actions. **European Journal of Operational Research**, 204, 3, 565-580, 2010.

COSTA, H. G.; MANSUR, A. F. U.; FREITAS, A. L. P.; DE CARVALHO, R. A. ELECTRE TRI aplicado a avaliação da satisfação de consumidores. **Produção**, v.17, n.2, p.230-245. 2007.

FARIA, F. P. **Gastos Sociais e Condições de Vida nos municípios fluminenses: Uma avaliação através da Análise Envoltória de Dados.** Dissertação (Mestrado em Estudos Populacionais e Pesquisas Sociais), Rio de Janeiro: Escola Nacional de Ciências Estatísticas, 2005.

FARIA, F. P., JANNUZZI, P. M. e SILVA, S. J. Eficiência dos gastos municipais em saúde e educação: uma investigação através da análise envoltória no estado do Rio de Janeiro. **Revista de Administração Pública**, 42, 1, 155-177, 2008.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. New Jersey: Prentice Hall, 1998.

MILLIGAN, G. W.; COOPER, M. C. An Examination of Procedures for Determining the Number of Clusters in a Data Set. **Psychometrika**, 50,159-179, 1985.

MOOI, E.; SARSTEDT, M. **A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistic.** Heidelberg: Springer, 2011.

SAMOILENKO, S.; OSEI-BRYSON, K. Increasing the discriminatory power of DEA in the presence of the sample heterogeneity with cluster analysis and decision trees. **Expert Systems with Applications**, 34, 1568-1581, 2008.

SANT'ANNA, A. P. Aleatorização e composição de medidas de preferência. **Pesquisa Operacional**, 22, 1, 87-103, 2002.

SANT'ANNA, A. P.; COSTA, H. G.; PEREIRA, V. CPP-TRI: um método de classificação ordenada baseado em composição probabilística. **Relatórios de Pesquisa em Engenharia de Produção**, 12, 8, 104-117, 2012.

WITTEN, I. H.; FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**. San Francisco: Elsevier, 2005.

APPLICATION OF PROBABILISTIC COMPOSTION AND OF K-MEANS TO THE CLASSIFICATION OF MUNICIPALITIES FOR CHILDS DAY CARE SUPPLY

Abstract

In this work the CPP - TRI method, which employs the probabilistic composition of preferences to classify into ordered categories, and the method of identification of clusters of k -means are used to classify municipalities with respect to the fulfillment of the constitutional provision for the availability of childcare child population. The segmentation of the municipalities of the State of Rio de Janeiro in order to evaluate the relationship between resources and products on childs daycare supply is considered. The variables considered are , by one side, the proportion of children aged two to five years in day care centers and, on the other side, the public spending per capita on education and culture and per capita income of the municipality. The method of k - means was employed to segment these municipalities in previous studies based on these three variables. There was agreement between the grouping performed by k -means and CPP - TRI based on cost / benefit ratios.

Key-words: *Probabilistic Composition of Preferences, k-means, Early Childhood Education, Efficiency, Evaluation of Municipal Management.*