COMPARAÇÃO ENTRE RESULTADOS DA COMPOSIÇÃO PROBABILÍSTICA DE PREFERÊNCIAS E DA FORMAÇÃO DE AGRUPAMENTOS PELO MÉTODO DAS K-MÉDIAS

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

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

Resumo

Neste trabalho o método CPP-TRI, que emprega a composição probabilística de preferências para classificar alternativas em categorias ordenadas, é usado para avaliar o agrupamento gerado pelo método das k-médias. A validade do uso deste método depende, na situação estudada, da ocorrência de relações custo/benefício próprias nos diferentes agrupamentos. Considerou-se a aplicação à segmentação dos municípios do Estado do Rio de Janeiro visando a avaliar a relação entre recursos e resultados na oferta de creches. As variáveis consideradas são, de um lado, a proporção de crianças de dois a cinco anos atendidas em creches de educação infantil 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 com base nessas três variáveis. Após a segmentação, 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.

Área principal: ADM - Apoio à Decisão Multicritério, AdP – PO na Administração Pública, MP – Métodos Probabilísticos.

Abstract

In this work, CPP-TRI, which employs the probabilistic composition of preferences to sort alternatives into ordered categories, is used to evaluate the clustering by the k-means method. The validity of using this method in the case studied depends on the occurrence of proper cost/benefit ratios in the different clusters. The case is of segmentation of the municipalities of the State of Rio de Janeiro to evaluate the relationship between resources and results on offering child daycare. The variables considered are, on one hand, proportion of children aged two to five years in day care centers and, on the other hand, municipal government 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 by taking into account these three variables. There was agreement between the grouping performed by the k-means and CPP-TRI based on cost/benefit ratios.

KEYWORDS: probabilistic preferences composition, k-means, early childhood education, efficiency, evaluation of municipal management.

1. Introdução

Neste trabalho o método CPP-TRI, que emprega a composição probabilística de preferências para classificar alternativas em categorias ordenadas, é usado para avaliar os agrupamentos gerados pelo uso da técnica das k-médias.

Para a avaliação da eficiência de unidades em que se identificam variáveis de dois tipos caracterizados como de custos e benefícios, a comparação de desempenhos pode ser mais informativa se é feita entre unidades separadas em subgrupos homogêneos. A técnica das k-médias pode ser usada para identificar tais agrupamentos homogêneos. A validade desse uso depende da ocorrência de relações custo/benefício próprias nos agrupamentos.

Considerou-se aqui a aplicação dessa técnica à segmentação dos municípios do estado do Rio de Janeiro visando avaliar a relação entre recursos e resultados na oferta de creches. A variável considerada como resultado foi a proporção de crianças de dois a cinco anos de idade atendidas em creches de educação infantil. Por outro lado, foram consideradas como recursos a renda per capita do município e a despesa pública municipal per capita com educação e cultura. A técnica das k-médias foi empregada para segmentar esses municípios em estudos anteriores das relações entre essas três variáveis.

Aqui são comparados os resultados dessa segmentação com os resultados da aplicação da CPP-TRI para a classificação com base nas relações custo/benefício. Verifica-se considerável concordância entre as classificações dos municípios pelas k-médias e pela CPP-TRI. Considerando os agrupamentos gerados pelas k-médias ordenados segundo a mediana dos valores da proporção de crianças atendidas em creches e a classificação pela CPP-TRI de acordo com as razões insumo/produto, a classificação de 67 dos 90 municípios ficou entre os extremos superior e inferior admitidos pela CPP-TRI com taxa de flexibilização da precisão em 50%.

As seções seguintes resumem, respectivamente, a técnica das k-médias e a CPP-TRI. A seção 4 apresenta o problema estudado. A seção 5 apresenta o resultado da aplicação da técnica das k-médias. A seção 6 descreve a apresentação da composição probabilística e compara os resultados. A seção 7 apresenta os comentários finais.

2. Agrupamentos pelas k-médias

As técnicas de agrupamentos, segundo Johnson e Wichern (1998), Witten e Frank (2005) e Po *et al.* (2009), se constituem num ramo da Análise Estatística Multivariada e um aprendizado não supervisionado sobre reconhecimento de padrões, uma vez que busca, basicamente, uma estrutura de agrupamentos "naturais" para um conjunto de dados, com base em similaridades ou distâncias (dissimilaridades).

Mooi e Sarstedt (2011) corroboram com esta ideia ao afirmar que a análise de agrupamentos é um método conveniente para a identificação de grupos homogêneos de objetos chamados *clusters* (grupos). Os objetos (ou casos ou observações) em um grupo específico compartilham muitas características, mas os objetos que não pertencem a esse grupo são muito diferentes.

Por outro lado, conforme Samoilenko e Osei-Bryson (2008), tais técnicas de agrupamentos são técnicas de mineração de dados, que envolvem a partição de um conjunto de objetos (ou variáveis) em subconjuntos mutuamente excludentes, tais que a similaridade entre as observações dentro de cada grupo é elevada, enquanto a similaridade entre as observações dos diferentes grupos é baixa. Ainda segundo Samoilenko e Osei-Bryson (2008), existem vários algoritmos disponíveis para se realizar agrupamentos, os quais, por sua vez, podem ser categorizados de muitas formas, tais como: hierárquico ou particional (não-hierárquico); determinístico ou probabilístico; *hard* ou *fuzzy*.

As técnicas de agrupamentos não hierárquicas ou particionais, como, por exemplo, k-mediana e k-médias, são destinadas a agrupar itens dentro de uma coleção de k grupos. O número de grupos, k, pode ser especificado antecipadamente ou determinado durante o processo de agrupamento (Johnson e Wichern, 1998).

Dentro desse contexto, o método particional ou não-hierárquico k-médias se configura o mais apropriado para ser utilizado no presente trabalho, tendo em vista que o objetivo do mesmo

é a identificação de grupos homogêneos "naturais".

Assim, conforme reportado em Johnson e Wichern (1998) e em Witten e Frank (2005) a técnica k-médias é uma abordagem 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 centroide se encontra mais próximo. Em suma, pode-se descrever o algoritmo dessa técnica por meio de três etapas:

1- Partição dos elementos em k agrupamentos iniciais;

2- Alocação de cada elemento no agrupamento de cujo centroide (média) esse elemento está mais próximo. Então, recalculam-se os centroides para os agrupamentos que perderam um item e que receberam um novo item.

3- Repete-se a etapa 2 até que nenhum elemento mude de agrupamento.

Cabe ressaltar que, antes de iniciar a partição de todos os elementos em k grupos definidos na etapa 1, deve-se especificar os k iniciais centroides e, então, proceder à etapa 2.

Uma questão importante é como decidir sobre o número de grupos, isto é, o valor de k que se deve considerar para a aplicação, no caso, da técnica k-médias. Pesquisas têm sugerido vários procedimentos para a determinação do número de agrupamentos de um conjunto de dados, tais como: o critério de agrupamento cúbico (*cubic clustering criterion* – CCC), desenvolvido por Sarle (1983); a estatística pseudo-F (critério de razão de variância – *variance ratio criterion* - *VRC*), que teve origem em Calinski e Harabasz (1974) e estatística pseudo t^2 , desenvolvida por Hotelling (1931). Segundo Mooi e Sarstedt (2011) e Milligan e Cooper (1985), o procedimento mais notável é o critério de razão de variância (VRC), o qual, para uma situação com n objetos e k grupos, é dado pela seguinte fórmula:

$$VRC_k = \left(SS_B / (k-1)\right) / \left(SS_W / (n-k)\right) \qquad [1]$$

onde SS_B é a soma dos quadrados entre os grupos e SS_W é a soma dos quadrados dentro dos grupos. O critério, na verdade, é o F-valor de uma Análise de Variância um fator (*one-way ANOVA*), com k representando os níveis de fator.

Para, finalmente, determinar o número adequado de grupos, calcula-se ω_k para cada solução de grupo, como se segue:

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

A partir daí, escolhe-se o número de grupos k que minimiza o valor de ω_k. Devido ao termo VRC_{k-1}, o número mínimo de grupos que podem ser selecionados é três, o que é uma clara desvantagem desse critério, o que limita a sua aplicação na prática. No entanto, essa limitação não representa um problema para o presente trabalho, como pode ser visto na aplicação e análise do critério.

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 completamente 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.

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.

O método se aplica ao caso aqui estudado com um único perfil para cada classe. Uma vez substituídas as medidas exatas a_k e C_{ijk} por distribuições de variáveis aleatórias X_k e Y_{ijk} centradas nessas medidas, podemos calcular probabilidades de sobreclassificação. Denotemos por A_{ik}^+ e A_{ik}^- a probabilidade de a alternativa A apresentar valor respectivamente acima e abaixo dos valores informados para o critério k nos perfis da classe i. Por independência,

$Aik^+ = \prod_j P[Xk > Yijk]$

e [3]

$Aik^- = \prod_j P[Xk < Yijk]$.

Para as credibilidades A_i^+ e A_i^- de a alternativa A estar acima ou abaixo da classe i serão usados, respectivamente, os produtos dos Aik^+ e dos Aik^-. Para k variando ao longo de todos os critérios.

O procedimento de classificação é baseado na comparação das diferenças $A_i^+ - A_i^-$. Se os perfis estão definidos de modo que as classes estão efetivamente ordenadas em ordem crescente, essas diferenças constituem uma sucessão decrescente. Se isto não acontece, os perfis precisam ser revistos. 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, a classificação otimista 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%.

4. Caracterização da área em estudo

O território brasileiro, de acordo com a Constituição da República Federativa do Brasil de 5 de outubro de 1988, é dividido em um Distrito Federal e vinte e seis estados-membros. Dentre estes estados está o Rio de Janeiro que se situa na Região Sudeste, limitando-se com Minas Gerais, ao norte e noroeste; São Paulo, a oeste; Espírito Santo, a nordeste; Oceano

Atlântico, ao sul, leste, sudeste e sudoeste.

O Rio de Janeiro se divide em 92 municípios, constituindo, pelo artigo 343 da sua Constituição, "... unidades territoriais que integram a organização político-administrativa da República Federativa do Brasil, dotados de autonomia política, administrativa e financeira, nos termos assegurados pela Constituição da República, por esta Constituição [a do Estado] e pela respectiva Lei Orgânica". Cada município possui a sua sede numa cidade, a qual lhe dá nome. Para fins administrativos, a maior parte dos municípios fluminenses divide-se em distritos, que também podem ser subdivididos.

Além da divisão político-administrativa, o Estado do Rio de Janeiro também possui uma divisão regional, apoiada na Lei n° 1.227/87, que aprovou o Plano de Desenvolvimento Econômico e Social 1988/1991. Desde então, foram feitas algumas alterações tanto na denominação quanto na composição dessas Regiões. Ao todo se têm oito Regiões de Governo: Metropolitana, Noroeste Fluminense, Norte Fluminense, Baixadas Litorâneas, Serrana, Centro-Sul Fluminense, Médio Paraíba e Costa Verde, como mostra a Figura 1.

Figura 1 – Mapa do Estado do Rio de Janeiro dividido em municípios e em Regiões de Governo
Fonte: http://mapasblog.blogspot.com.br/2011/11/mapas-do-estado-do-rio-de-janeiro.html

O conjunto estudado, já utilizado em Faria (2005), é formado por 90 municípios do Estado do Rio de Janeiro. O fato de se trabalhar apenas com os dados de 90 dos 92 municípios do Estado do Rio de Janeiro se deve à falta de dados referentes ao *input* denominado gastos com educação e cultura para os municípios de Mesquita e Seropédica em 2000. No caso de Mesquita, de fato, não seria possível obter estes dados, uma vez que a sua criação se deu em 25/09/1999 e a sua instalação em 01/01/2001. Entretanto, o mesmo não se pode dizer a respeito de Seropédica, o qual foi criado em 12/10/1995 e foi instalado em 01/01/1997. Assim, optou-se por retirar estes dois municípios da análise.

Faz-se importante ressaltar o fato de que, como os dados mais atualizados não estão disponíveis para todas as variáveis utilizadas na modelagem, apenas para parte delas, não foi possível realizar uma análise mais recente dos mesmos.

A escolha das variáveis foi baseada no trabalho desenvolvido por Faria (2005). Assim, optou-se por trabalhar com as despesas por função, no caso, 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.

Cabe destacar que, apesar do objetivo deste estudo estar voltado para a área de educação, foi preciso considerar também as despesas municipais na área da cultura, uma vez que no referido ano, tais despesas eram apresentadas em conjunto, não sendo possível especificar os gastos dos municípios fluminenses apenas com educação.

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 (PROPCRECH). Este indicador também foi obtido do Censo Demográfico de 2000, realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e, assim como os demais, são os mesmos abordados em Faria (2005) e reportados em Faria *et al.* (2008), no que diz respeito à área de educação e cultura.

O Quadro 1 mostra as variáveis em análise e a Tabela 1 mostra os valores dessas variáveis para os municípios estudados.

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 (PROPCRECH)	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 e Renda per capita e Crianças em Creches

	Municípios	**GEDUC (R$)**	**RENDA (R$)**	**PROPCRECH (%)**
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	Armação 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

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

O emprego da técnica k-médias se inicia com a determinação do número de grupos e a escolha dos centros para a classificação inicial. Para a escolha dos centros iniciais foram escolhidos pontos igualmente espaçados segundo o *output* PROPCRECH, uma vez que se busca avaliar o comportamento dos municípios quanto à oferta deste serviço social.

Para determinar o número ótimo de grupos que devem ser formados com os 90 municípios em estudo, realizaram-se algumas simulações com diferentes valores de $k = 2,...,8$ e aplicou-se a ANOVA um fator para cada um desses valores de k, obtendo-se os F-valores apresentados na Tabela 2.

Tabela 2 – F-valores obtidos da aplicação da ANOVA um fator

K	F-valor
2	233,32
3	234,24
4	356,83
5	422,79
6	453,67
7	93,11
8	549,93

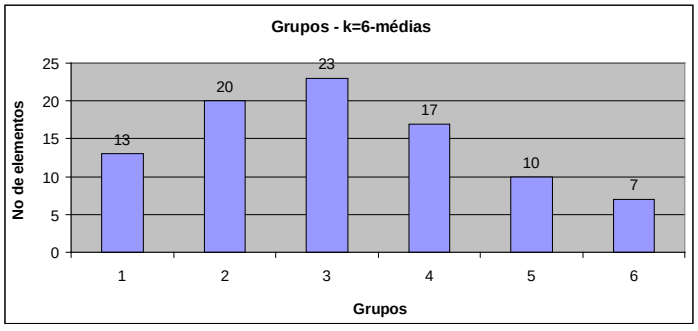
A partir destes resultados, calculou-se o valor de ω_k para cada solução de grupo, como mostra a Tabela 3.

Tabela 3 –Valores de ω_k para cada solução de grupo

ω_k	Valor
ω_3	121,67
ω_4	-56,63
ω_5	-35,08
ω_6	-391,45
ω_7	817,39

Como o menor valor de ω_k é $\omega_6 = -391{,}45$, então, seguindo Milligan e Cooper (1985) e Mooi e Sarstedt (2011), se conclui que $k = 6$ grupos é a melhor partição para o conjunto de dados em análise, a qual está representada no Gráfico 1.

Gráfico 1 – Grupos formados pela aplicação do k-médias

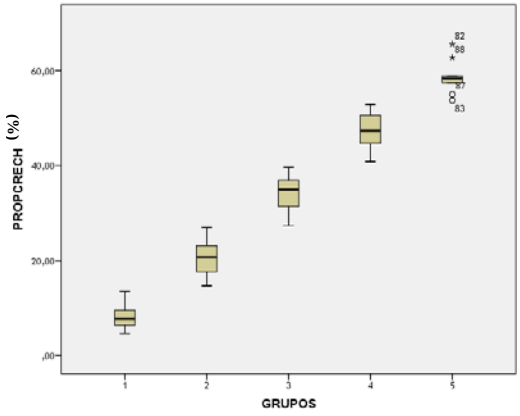


Os agrupamentos formados se encontram parcialmente de acordo com os parâmetros limitantes da técnica k-médias estipulados por Samoilenko e Osei-Bryson (2008). Isto se deve ao fato de que, para k = 2, 3, 4 e 5, o número de elementos de cada um dos grupos formados atendeu ao limite de veto ($k_{máx}$ = 10% do total dos elementos). 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) dentre os seis grupos formados foi constituído por menos de 10% do total dos

municípios, como mostra o Gráfico 1.

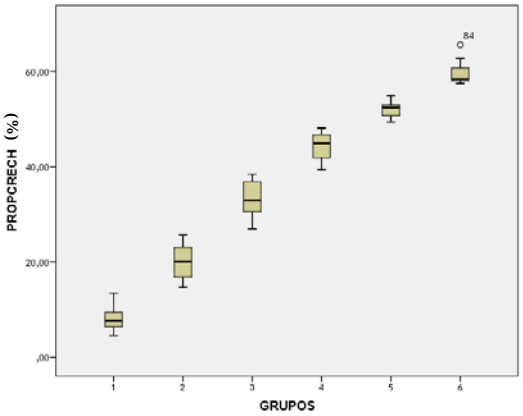
Todavia, apesar dos agrupamentos formados para $k = 6$ atenderem parcialmente a esses parâmetros limitantes da técnica k-médias estipulados por Samoilenko e Osei-Bryson (2008), esta partição se mostra mais robusta estatisticamente, como comprovaram os testes baseados na aplicação da ANOVA. Isto também pode ser notado nos gráficos 2, 3 e 4, que mostram a redução de *outliers* quando se aplica o k-médias a este conjunto de dados para $k = 6$.

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



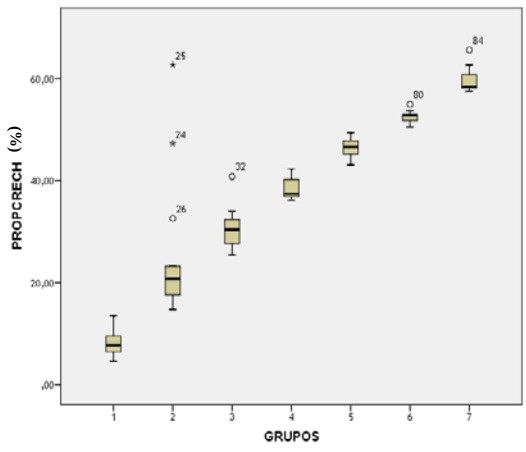
Os *outliers* dos grupos formados pelo k-médias com $k = 5$ se encontram no grupo 5 e correspondem aos seguintes municípios: Aperibé, Cambuci, Natividade e Porciúncula. Já nos grupos formados pelo k-médias com $k = 6$, apenas um *outlier*, o município de Aperibé, foi destacado neste gráfico, também pertencente ao último grupo, no caso grupo 6.

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



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. Os municípios correspondentes a estes *outliers* são: Piraí, Porciúncula, Porto Real, Barra do Piraí, Natividade, Aperibé. Tal fato corrobora para a escolha do $k = 6$ grupos para a aplicação da técnica k-médias.

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



6. Validação pela CPP-TRI

A CPP-TRI foi aplicada a partir da caracterização de seis categorias de municípios por perfis de referência dados pelas médias dos vetores de valores das razões *input/output* nos seis grupos determinados pela aplicação da técnica das k-médias. Para calcular a média das razões entre despesas municipais per capita com educação e a proporção de crianças atendidas por creches, no grupo 4, foi excluído 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 4.

Tabela 4. 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

O grupo de menor tamanho pela CPP-TRI foi o grupo 4 com 9 municípios. No grupo 5 ficaram 10 e no grupo 1 ficaram 11. Os grupos maiores foram os grupos 2, 3 e 6 com respectivamente, 24, 19 e 17 municípios. Nota-se que embora os tamanhos dos grupos sejam, em alguns casos, bastante diferentes dos obtidos pelas k-médias, o total de municípios no conjunto dos três primeiros e no conjunto dos três últimos grupos é o mesmo nas duas classificações.

A Tabela 5 apresenta para cada município as classificações inferior, central e superior pela CPP-TRI precedida pela classificação inicial pelas k-médias. As classificações superior e inferior são produzidas com a flexibilização para um ponto de corte de 50%.

Tabela 5. 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

Na Tabela 5, contamos 66 municípios em que o valor da classificação inicial, gerada pelo k-médias, está entre os limites da CPP-TRI e 24 municípios com classificações discordantes. Destes 24, apenas o município de 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 pode ser considerado alto, dada a ausência de correlação encontrada nos municípios do Estado do Rio de Janeiro para as três variáveis e a alta proximidade entre as classificações benevolente e exigente pela CPP-TRI.

7. Conclusão

O uso da composição probabilística de preferências para a classificação dos municípios a partir de uma classificação pelas k-médias foi realizado com sucesso, conduzindo a considerável concordância entre as classificações benevolente e exigente.

Com a concordância entre os resultados das duas 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.

O estudo realizado permitiu 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. e Roy, B. (2010), Electre Tri-C: A multiple criteria sorting method based on characteristic reference actions. *European Journal of Operational Research*, 204, 3, 565-580,

Calinski, T. e Harabasz, J. (1974), A Dendrite Method for Cluster Analysis. *Communications in Statistics*, 3, 1-27.

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), Escola Nacional de Ciências Estatísticas, Rio de Janeiro, Brasil, 2005.

Faria, F. P., Jannuzzi, P. M. e Silva, S. J. (2008), 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.

Hotelling, H. (1931), The generalization of Student's ratio. *Annals of Mathematical Statistics*, 2, 3, 360-378.

Johnson, R. A. e Wichern, D. W. *Applied Multivariate Statistical Analysis*. Prentice Hall, New Jersey, 1998.

Milligan, G. W. e Cooper, M. C. (1985), An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*, 50,159-179.

Mooi, E. e Sarstedt, M. *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics.* Springer, Heidelberg, 2011.

Po, R., Guh, Y. e Yang, M. (2009), A new clustering approach using data envelopment analysis. *European Journal of Operational Research*, 199, 276-284.

Samoilenko, S. e Osei-Bryson, K. (2008), 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.

Sant'Anna, A. P. (2002), Aleatorização e composição de medidas de preferência. *Pesquisa Operacional*, 22, 1, 87-103.

Sant'Anna, A. P., Costa, H. G. e Pereira, V. (2012), CPP-TRI: um método de classificação ordenada baseado em composição probabilística, *Relatórios de Pesquisa em Engenharia de Produção*, 12, 8, 104-117.

Sarle, W. S. *Cubic Clustering Criterion*. SAS Technical Report A-108, Cary, NC: SAS Institute Inc, 1983.

Witten, I. H. e Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, San Francisco, 2005.