CADERNOS DO IME – Série Estatística

Universidade do Estado do Rio de Janeiro - UERJ
ISSN impresso 1413-9022 / ISSN on-line 2317-4536 - v. 42, p. 31 - 43, 2017
DOI: 10.12957/cadest.2017.30347

DESEMPENHO DAS ESCOLAS PÚBLICAS E PRIVADAS DA REGIÃO DO VALE DO PARAÍBA: UMA APLICAÇÃO DA TÉCNICA DE AGRUPAMENTOS *KMEANS* COM BASE NAS VARIÁVEIS DO ENEM 2015

Roberto Campos Leoni
Academia Militar das Agulhas Negras
rcleoni@yahoo.com.br

Nilo Antonio de Souza Sampaio
Universidade do Estado do Rio de Janeiro
nilo.samp@terra.com.br

Resumo

Este artigo avalia o desempenho dos alunos de escolas públicas e privadas da região sul fluminense no Exame Nacional do Ensino Médio (Enem) de 2015 por meio da técnica não hierárquica de agrupamentos kmeans. As escolas são classificadas de acordo com o desempenho dos alunos em função das variáveis indicadoras de proficiência média em ciências da natureza e suas tecnologias, ciências humanas e suas tecnologias, linguagens, códigos e suas tecnologias, matemática e suas tecnologias, redação, formação docente, taxas de rendimento escolar de aprovação e participação percentual de alunos no Enem, todas divulgadas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Os resultados indicam a prevalência de dois grupos que são aqui caracterizados em função do porte da escola, município, dependência administrativa e indicador socioeconômico.

Palavras-chave*: Kmeans, Análise de Agrupamento, Enem.*

1. Introdução

O Exame Nacional do Ensino Médio (Enem) foi criado em 1998 com o objetivo de avaliar o conhecimento adquirido pelos alunos que estavam concluindo o segundo grau. O exame propõe mensurar modalidades estruturais da inteligência mediante uma concepção construtivista com foco na resolução de problemas.

Atualmente o Enem vai muito além de suas expectativas iniciais, pois é a porta de entrada para uma faculdade pública ou privada e serve para certificação de competências de jovens e adultos (Encceja) ou de exames de certificação de competência ou de avaliação de jovens e adultos realizados pelos sistemas estaduais de ensino.

O Enem é pré-requisito para o estudante se inscrever no sistema informatizado do Ministério da Educação (SiSU). Neste sistema, instituições públicas de ensino superior oferecem vagas a candidatos participantes do Enem. Outros sistemas também empregam o resultado do Enem, sendo eles: o programa Universidade para todos (ProUni) e o fundo de financiamento estudantil (Fies).

O resultado do Enem auxilia gestores educacionais a tomar decisões sobre o aprendizado dos estudantes e auxiliam no desenvolvimento de estratégias em favor da qualidade da educação. As médias apresentadas por área do conhecimento e redação são usadas pela comunidade escolar para auxiliar na análise dos desafios a serem enfrentados.

Além dos resultados agregados de proficiências por área de conhecimento e redação, a análise dos resultados do Enem por escola considera também informações contextuais, tais como: o indicador de nível socioeconômico, permanência na escola, o indicador de formação docente da escola e as taxas de rendimento escolar do ensino médio (INEP, 2015).

O Enem é um bom instrumento construtivista para fins educacionais. Gomes e Borges (2009) realizaram um estudo exploratório e constataram que a prova do Enem pode ser tomada como um teste de inteligência. É uma avaliação que ativa fundamentalmente processos cognitivos complexos e sofisticados de resolução de problemas.

O Enem não tem a finalidade de avaliar escolas, mas sim o desempenho individual dos alunos. Castro (2009) afirma que a comparação das escolas públicas com as particulares, no caso do Enem, provoca enorme polêmica entre os especialistas em avaliação, pois trata-se de uma comparação frágil, do ponto de vista metodológico, que

não considera os fatores socioeconômicos associados ao desempenho individual dos alunos.

Existe um longo debate sobre as importâncias relativas dos papeis das famílias e das escolas no aprendizado dos alunos. Em geral, argumenta-se que a família tem um papel primordial no aprendizado dos alunos, mas que a escola pode também adicionar conhecimento e aprendizagem nesse processo. Por razões óbvias, as escolas particulares recebem os alunos que podem pagar, em geral oriundos de famílias de maior escolaridade e com acesso a bens culturais.

O foco na qualidade tem levado o governo federal, através do Ministério da Educação (MEC), a tentar mensurar a qualidade da educação, desenvolvendo métricas e divulgando *rankings* das instituições de ensino. O objetivo é fornecer informações sobre a qualidade das instituições para os diferentes agentes, as quais podem auxiliar pais e estudantes na escolha de onde estudar. Evidências sugerem que os impactos da divulgação são relevantes. Os dirigentes das escolas mal colocadas no *ranking* são questionados pelos pais dos alunos, que exigem explicações e estratégias para melhorar a posição das escolas. A procura pelas escolas mais bem classificadas parece também ser maior (ANDRADE, 2011).

Do exposto, formulou-se a seguinte questão de pesquisa: tomando-se como parâmetro os indicadores de proficiência média por escola, indicadores contextuais e de participação de alunos no Enem, verificam-se padrões de desempenho similares ou dissimilares suficientemente significativos para permitir afirmar a existência de agrupamentos naturais dentre os tipos de escolas? Para alcançar a resposta à presente questão de pesquisa, procurou-se empregar uma das técnicas mais utilizadas por pesquisadores em diversos campos do conhecimento, a técnica de agrupamento denominada *kmeans* (FÁVERO *et al.*, 2009; HAIR *et al*., 2009).

O restante do artigo está assim estruturado: a seção 2 apresenta o método, as técnicas e as etapas da pesquisa. A seção 3 os resultados obtidos com base no algoritmo de agrupamento e a caracterização dos grupos formados. Na seção 4, são apresentadas algumas considerações finais e sugestões para pesquisa.

2. Métodos e técnicas de pesquisa

A análise de agrupamentos é um grupo de técnicas multivariadas cuja finalidade principal é agregar objetos (respondentes, produtos ou outras entidades) com base nas

características que eles possuem (HAIR *et al*., 2009). Esta pesquisa aplicou o algoritmo de agrupamento *kmeans*, pois é extremamente rápido e, portanto, pode ser facilmente aplicado em grandes conjuntos de dados. É uma ferramenta exploratória de análise de dados que agrupa objetos em grupos ou *clusters* semelhantes. Além disso, só exige a especificação do número de grupos (LEWIS, 2017).

Empregou-se a análise de agrupamentos *kmeans* com o objetivo de classificar escolas públicas e privadas em grupos com desempenhos similares por meio das variáveis indicadoras do Enem que são descritas brevemente a seguir (INEP, 2015):

i) **Indicador de proficiência média por escolas:** para o Enem apresenta proficiências médias, por unidade escolar, para cada uma das Áreas do Conhecimento e para Redação.
   - proficiência média em ciências da natureza e suas tecnologias (MEDIA_CN);
   - proficiência média em ciências humanas e suas tecnologias (MEDIA_CH);
   - proficiência média em linguagens, códigos e suas tecnologias (MEDIA_LC);
   - proficiência média em matemática e suas tecnologias (MEDIA_MT);
   - proficiência média em redação (MEDIA_RED).

ii) **Indicadores contextuais**: para o Enem apresentam informações que permitem uma melhor compreensão da realidade de cada escola e uma análise mais adequada de seus resultados de proficiência, uma vez que esses estão associados às características e contexto das escolas e seus alunos. O Indicador de Adequação da Formação Docente analisa a formação dos docentes que lecionam no ensino médio na escola. Apresenta o percentual de disciplinas que são ministradas por professores com formação superior de Licenciatura (ou Bacharelado com complementação pedagógica) na mesma disciplina que leciona. As taxas de rendimento escolar correspondem às taxas de aprovação, reprovação e abandono baseadas em informações sobre o movimento e rendimento escolar dos alunos, registrados no Censo Escolar. A soma das três taxas resulta 100%.
   - indicador de formação docente (IND_FORM_DOCENTE);
   - taxas de rendimento escolar de aprovação (TAXA_APROVACAO).

iii) **Indicador para taxas de participação:** corresponde à razão entre o número total de estudantes do ensino médio regular da escola, de acordo com o Censo Escolar 2015, que tenham realizado as quatro provas objetivas e a redação do Enem 2015, com

proficiências superiores a zero em todas as provas objetivas, e o número total de alunos do ensino médio regular, declarado pela unidade escolar ao Censo Escolar 2015. Os alunos matriculados em mais de uma turma na mesma escola foram considerados apenas uma vez para o cálculo do número total de alunos do ensino médio regular dessa unidade escolar.

- participação percentual de alunos (TAXA_PARTICIPACAO).

Delimitou-se como população alvo na presente avaliação o conjunto de escolas públicas e privadas da região sul fluminense, cujos municípios participantes (Angra dos Reis, Barra do Piraí, Barra Mansa, Itatiaia, Paraty, Piraí, Porto Real, Resende, Três Rios e Volta Redonda) foram os que apresentaram os melhores resultados nos indicadores PIB per capta de 2014 e IDH de 2010 (IBGE, 2017).

Etapas empregadas na pesquisa:

1) Coleta; exploração e preparação dos dados:
   Preparação da base de dados; verificação de observações discrepantes, padronização dos dados com o escore Z e verificação da representatividade dos dados empregados na análise;
2) Escolha do modelo de agrupamento:
   Formação dos agrupamentos; escolha do algoritmo de agrupamento, determinação do intervalo de número de grupos aceitáveis e processamento do agrupamento;
3) Avaliar o desempenho do modelo de agrupamento:
   Avaliação, interpretação e validação dos resultados.

Todos os cálculos e procedimentos computacionais foram realizados com o software livre *R* (R CORE TEAM, 2017).

3. Resultados e discussão dos resultados

A base de dados é composta por 106 escolas públicas e privadas da região sul fluminense. Algumas características descritivas das escolas são apresentadas na Tabela 1. Todas as variáveis se destacam em relação ao elevado desvio padrão, indicando heterogeneidade nos resultados. Prosseguir-se-á, nas próximas etapas, em busca de grupos de escolas com desempenho homogêneo dentro dos grupos e desempenho heterogêneo entre os grupos.

Tabela 1: Estatísticas descritivas das variáveis indicadoras.

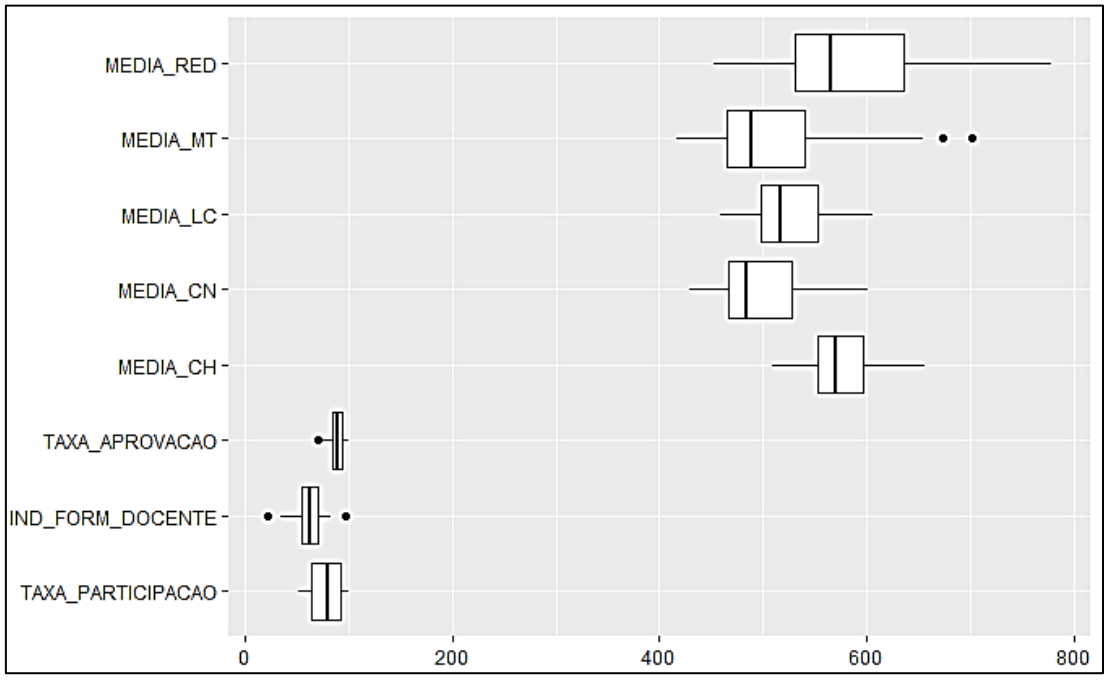
Variáveis	média	desvio-padrão	mediana	mínimo	máximo
TAXA_PARTICIPACAO	78.56	15.11	79.41	51.06	100.00
IND_FORM_DOCENTE	62.14	12.00	62.95	21.30	97.10
TAXA_APROVACAO	88.56	6.80	89.20	68.60	100.00
MEDIA_CH	578.53	36.72	570.83	508.42	657.18
MEDIA_CN	498.61	45.41	485.04	428.58	609.16
MEDIA_LC	527.58	36.11	518.14	458.71	606.49
MEDIA_MT	509.54	64.22	488.48	416.95	702.15
MEDIA_RED	588.87	76.52	566.74	452.50	811.35

Fonte: o próprio autor.

Verificou-se a existência de dados multivariados discrepantes com a técnica denominada *Blocked Adaptive Computationally-Efficient Outlier Nominators* (BILLOR; HADI; VELLEMAN, 2000). Três escolas se comportaram como discrepantes e, por essa razão, foram excluídas da análise para não influenciar a formação dos grupos.

O algoritmo empregado, *kmeans*, usa a distância euclidiana ao quadrado para alocar os objetos nos grupos. Isso requer que os dados tenham aproximadamente a mesma escala. A Figura 1 ilustra a necessidade de padronização dos dados. Muito claramente, as faixas diferem.

Figura 1 - Box-plot das variáveis indicadoras.



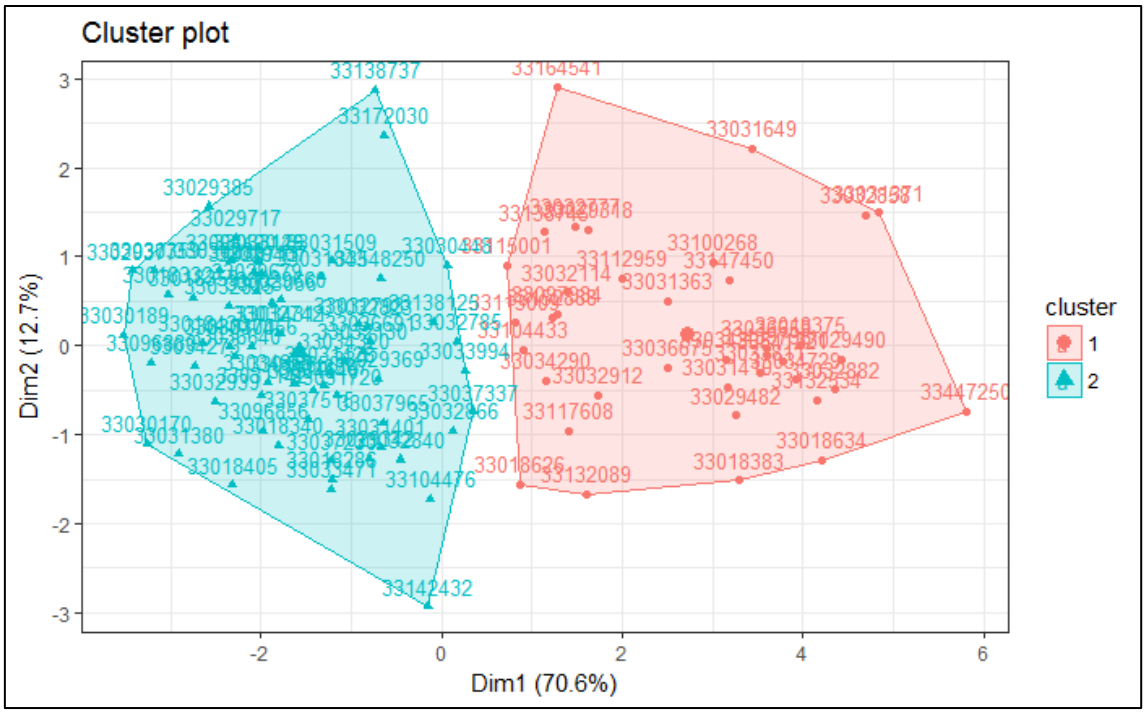
Fonte: o próprio autor.

Muitos índices têm sido propostos na literatura para determinar o número ótimo de agrupamentos (TIBSHIRANI; WALTHER; HASTIE, 2001; CHARRAD *et al*., 2014).

Três métodos foram usados para validar o número de grupos, sendo eles: *wss*, *silhouette* e *gap_statistic*. Os três métodos apresentaram respostas convergentes, indicando a formação de 2 (dois) grupos para a base de dados do Enem disponibilizada pelo INEP.

A Figura 2 ilustra a formação dos dois grupos com base no algoritmo *kmeans*. O grupo 1 resultou em 38 escolas e o grupo 2 resultou em 65 escolas. Ressalta-se que o algoritmo foi replicado mil vezes para que fosse avaliada a estabilidade do procedimento de agrupamento e embasasse, de maneira consistente, a alocação das escolas nos grupos.

Figura 2 - Grupos resultantes da aplicação do algoritmo *kmeans*.

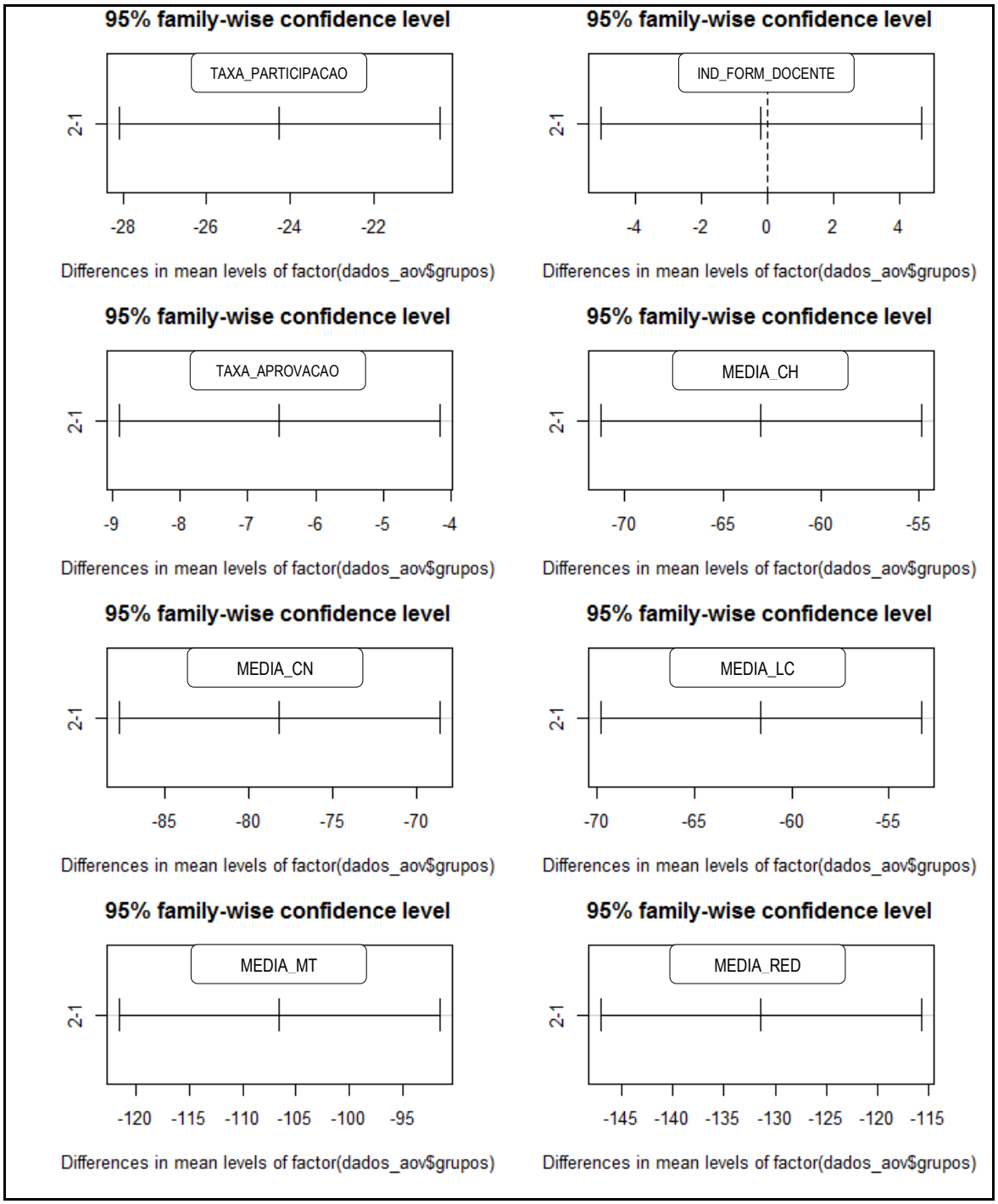


Fonte: o próprio autor.

Verificou-se por meio do teste *F* da análise de variância de um fator se os valores de cada uma das oito vaiáveis indicadoras na análise são estatisticamente diferentes entre os dois grupos.

Em todos os testes o *p*-valor associado a estatística *F* é significativo (p-valor <0,01) com exceção da variável indicadora índice de formação docente (IND_FORM_DOCENTE) que não difere entre os dois grupos. Dessa análise, observou-se que o grupo 1 apresenta resultados superiores em relação ao grupo 2, como pode ser observado na Figura 3.

Figura 3 - Diferença de médias entre os grupos.



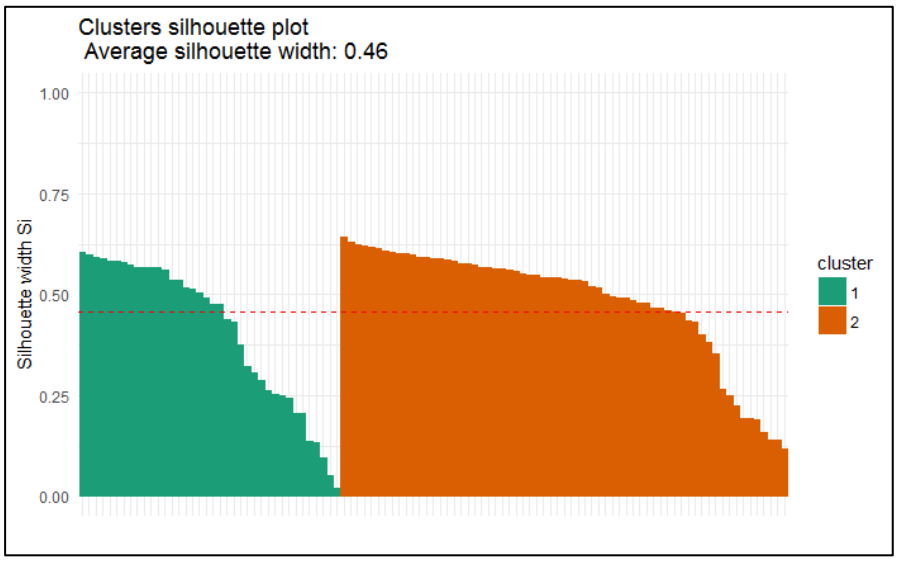
Fonte: o próprio autor.

A análise de silhueta mede o quão bem uma observação é agrupada e estima a distância média entre agrupamentos. O gráfico de silhueta exibe uma medida de quão próximo cada ponto de um grupo é para pontos nos grupos vizinhos.

A largura da silhueta (*Si*) pode ser interpretada da seguinte forma: Observações com um *Si* grande (quase 1) são muito bem agrupadas. Um *Si* pequeno (em torno de 0) significa que a observação está entre dois grupos. Observações com um *Si* negativo são, provavelmente, colocadas no grupo errado (BROCK *et al.*, 2008; LEWIS, 2017).

Observa-se na Figura 4 que os elementos de cada grupo estão, em média, bem agrupados e não há *Si* negativos.

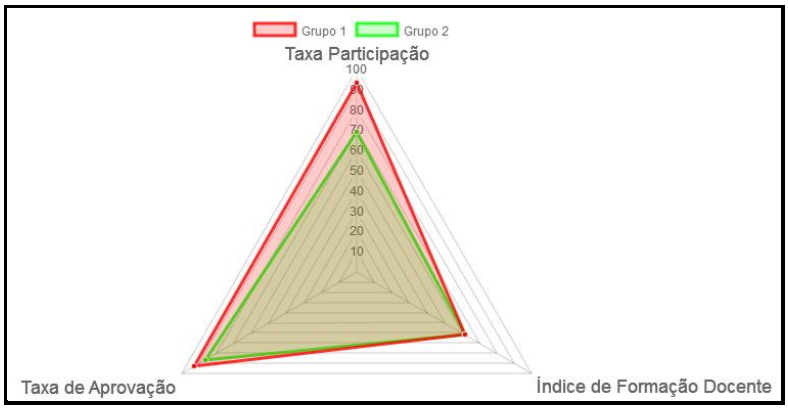
Figura 4 - Gráfico da Silhueta.



Fonte: o próprio autor.

Os grupos 1 e 2 são caracterizados de acordo com as Figuras 5 a 7. Observa-se que o grupo 1 apresenta níveis superiores em relação aos indicadores taxa de participação e taxa de reprovação. Contudo, o índice de formação docente é equivalente em ambos os grupos (ver Figura 5).

Figura 5 - Gráfico radar dos indicadores contextuais e taxas de participação.

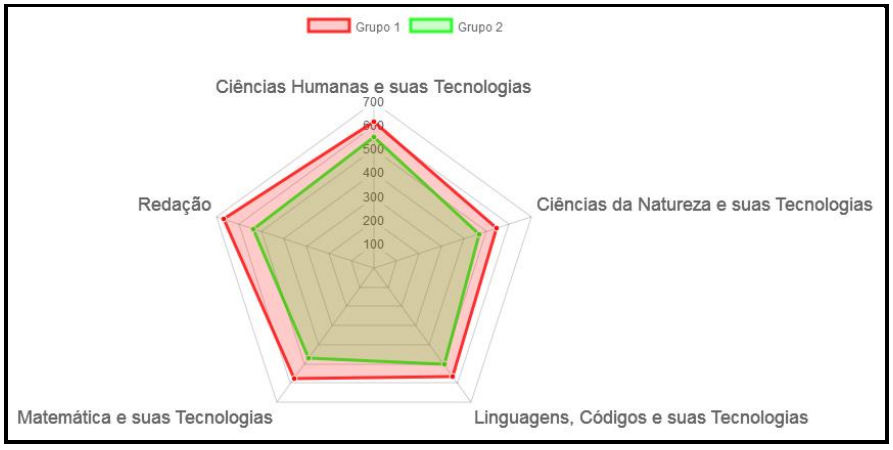


Fonte: o próprio autor.

A Figura 6 ilustra o indicador de proficiência média por escolas. É fácil observar a vantagem do Grupo 1 em relação ao Grupos 2 em todas as áreas do conhecimento.

Na Figura 7, retêm-se que o Grupo 1 é basicamente caracterizado por escolas em sua maioria privadas, com indicador socioeconômico alto ou muito alto e porte não maior que 60 alunos. A distribuição das escolas por Municípios está muito próxima da quantidade de escolas públicas e privadas em cada região.

Figura 6 - Gráfico radar do indicador de proficiência média por escolas.



Fonte: o próprio autor.

Figura 7 - Caracterização dos grupos por porte, município, dependência administrativa e indicador socioeconômico

	Grupo			Grupo	
	1	**2**		**1**	**2**
PORTE			DEPENDÊNCIA ADMINISTRATIVA		
De 1 a 30 alunos	20	11	Estadual	1	54
De 31 a 60 alunos	11	17	Municipal	1	4
De 61 a 90 alunos	4	12	Privada	36	7
Maior que 90 alunos	3	25			
MUNICÍPIO			INDICADOR SOCIOECONÔMICO		
Angra dos Reis	6	7	Médio	0	13
Barra do Piraí	4	6	Médio alto	3	39
Barra Mansa	4	14	Alto	17	11
Itatiaia	0	2	Muito alto	17	0
Paraty	1	3			
Piraí	0	1			
Porto Real	1	1			
Resende	7	7			
Três Rios	6	10			
Volta Redonda	9	14			

Fonte: o próprio autor.

4. Considerações finais e recomendações

Apresentou-se, neste artigo, o desempenho dos alunos em função dos indicadores de proficiência média por escolas, indicadores contextuais e indicador para taxas de participação por escolas. O método empregado revelou a formação de dois grupos, validados através de técnicas estatísticas, em função desses indicadores. Procurou-se caracterizar o perfil das escolas e alunos em cada grupo utilizando os indicadores de porte, município, dependência administrativa e indicador socioeconômico. Verificando-se, portanto, padrões de desempenho similares suficientemente significativos para permitir afirmar a existência de agrupamentos naturais dentre os tipos de escolas.

Algoritmos de agrupamento semelhantes ao empregado no presente artigo podem ajudar na formação de grupos de escolas particulares que possuam desempenho similar ao de algumas escolas públicas, que em geral, não apresentam bom desempenho no Enem. Recomenda-se como sugestão para trabalho futuro identificar os fatores críticos de sucesso determinantes para o bom desempenho das escolas no Enem.

Referências

ANDRADE, E. D. C. Rankings em Educação: Tipos, Problemas, Informações e mudanças - Análise dos Principais Rankings Oficiais Brasileiros. **Estudos Econômicos (São Paulo)**, v. 41, n. 2, p. 323–343, 2011.

BILLOR, N.; HADI, A. S.; VELLEMAN, P. F. BACON: Blocked Adaptive Computationally efficient Outlier Nominators. **Computational Statistics & Data Analysis**, v. 34, n. 3, p. 279–298, set. 2000.

BROCK, G.; PIHUR, V.; DATTA, S.; DATTA, S. clValid: An R Package for Cluster Validation. **Journal of Statistical Software**, v. 25, n. 4, p. 1–22, 2008.

CASTRO, M. H. G. DE. Sistemas de Avaliação da Educação no Brasil. **São Paulo Perspec.**, v. 23, n. 1, p. 5–18, 2009.

CHARRAD, M.; GHAZZALI, N.; BOITEAU, V.; NIKNAFS, A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. **Journal of Statistical Software**, v. 61, n. 6, 2014.

FÁVERO, L. P.; BELFIORE, P.; SILVA, F. L. da; CHAN, B. L. **Modelagem Multivariada para Tomada de Decisões**. Rio de Janeiro - RJ: Elsevier, 2009.

GOMES, C. M. A.; BORGES, O. O ENEM é uma Avaliação Educacional Construtivista? Um Estudo de Validade de Construto. **Estudos em Avaliação Educacional**, v. 42, p. 73–88, 2009.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise Multivariada de Dados**. 6. ed. Porto Alegre: Bookman, 2009.

IBGE. **BRASIL**. Disponível em: <http://cidades.ibge.gov.br/xtras/uf.php?lang=&coduf=33&search=rio-de-janeiro>. Acesso em: 28 maio. 2017.

INEP. **Nota Explicativa Enem 2015 por Escola**. Disponível em:

<http://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2015/nota_explicativa_enem2015_por_escola.pdf>. Acesso em: 28 maio. 2017.

LEWIS, N. D. **Machine Learning made easy with R: An Intuitive Step by Step Blueprint for Beginners**. 1. ed. -: Paperback, 2017.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**.Vienna, Austria, 2017. . Disponível em: <https://www.r-project.org/>. Acesso em: 28 maio. 2017.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the Number of Clusters in a Data Set via the Gap Statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 63, n. 2, p. 411–423, 2001.

PERFORMANCE OF THE PUBLIC AND PRIVATE SCHOOLS OF THE PARAÍBA VALLEY REGION: AN APPLICATION OF THE KMEANS CLUSTERING TECHNIQUE BASED ON ENEM 2015 VARIABLES

Abstract

This article evaluates the performance of students from public and private schools in the southern region of Rio de Janeiro in the National High School Examination (Enem) of 2015. The non-hierarchical K-means clustering technique was used. The schools are classified according to the indicator variables of average proficiency in natural sciences and their technologies, human sciences and their technologies, languages, codes and their technologies, mathematics and its technologies, writing, teacher training, school achievement rates, and percent participation of students in Enem, all of which are reported by the National Institute of Educational Studies and Research Anísio Teixeira (INEP). The results indicate the prevalence of two groups that are characterized by size of school, municipality, administrative dependence, and socioeconomic indicator.

Key-words: *Kmeans, Cluster Analysis, Enem.*