Foguel, Miguel Nathan; Veloso, Fernando

Working Paper

Agrupando unidades de sistemas de serviços públicos

Texto para Discussão, No. 2051

Provided in Cooperation with:
Institute of Applied Economic Research (ipea), Brasília

Suggested Citation: Foguel, Miguel Nathan; Veloso, Fernando (2015) : Agrupando unidades de sistemas de serviços públicos, Texto para Discussão, No. 2051, Instituto de Pesquisa Econômica Aplicada (IPEA), Brasília

This Version is available at:
https://hdl.handle.net/10419/121536

2051
TEXTO PARA DISCUSSÃO

ipea

AGRUPANDO UNIDADES DE SISTEMAS DE SERVIÇOS PÚBLICOS

Miguel Nathan Foguel
Fernando Veloso

ipea

2051
TEXTO PARA DISCUSSÃO
Brasília, março de 2015

AGRUPANDO UNIDADES DE SISTEMAS DE SERVIÇOS PÚBLICOS[1]

Miguel Nathan Foguel[2]
Fernando Veloso[3]

1. Gostaríamos de agradecer a Sergio Guimarães Ferreira e Maína Celidonio de Campos pelos comentários e à equipe do Instituto Pereira Passos pela organização das bases de dados no nível das Coordenadorias Regionais de Educação (CREs).
2. Pesquisador da Diretoria de Estudos e Políticas Sociais (Disoc) do Ipea.
3. Pesquisador do Instituto Brasileiro de Economia (Ibre) da Fundação Getulio Vargas (FGV).

Governo Federal

Secretaria de Assuntos Estratégicos da Presidência da República
Ministro Roberto Mangabeira Unger

ipea **Instituto de Pesquisa Econômica Aplicada**

Fundação pública vinculada à Secretaria de Assuntos Estratégicos da Presidência da República, o Ipea fornece suporte técnico e institucional às ações governamentais – possibilitando a formulação de inúmeras políticas públicas e programas de desenvolvimento brasileiro – e disponibiliza, para a sociedade, pesquisas e estudos realizados por seus técnicos.

Presidente
Sergei Suarez Dillon Soares

Diretor de Desenvolvimento Institucional
Luiz Cezar Loureiro de Azeredo

Diretor de Estudos e Políticas do Estado, das Instituições e da Democracia
Daniel Ricardo de Castro Cerqueira

Diretor de Estudos e Políticas Macroeconômicas
Cláudio Hamilton Matos dos Santos

Diretor de Estudos e Políticas Regionais, Urbanas e Ambientais
Rogério Boueri Miranda

Diretora de Estudos e Políticas Setoriais de Inovação, Regulação e Infraestrutura
Fernanda De Negri

Diretor de Estudos e Políticas Sociais, Substituto
Carlos Henrique Leite Corseuil

Diretor de Estudos e Relações Econômicas e Políticas Internacionais
Renato Coelho Baumann das Neves

Chefe de Gabinete
Ruy Silva Pessoa

Assessor-chefe de Imprensa e Comunicação
João Cláudio Garcia Rodrigues Lima

Ouvidoria: http://www.ipea.gov.br/ouvidoria
URL: http://www.ipea.gov.br

Texto para Discussão

Publicação cujo objetivo é divulgar resultados de estudos direta ou indiretamente desenvolvidos pelo Ipea, os quais, por sua relevância, levam informações para profissionais especializados e estabelecem um espaço para sugestões.

Texto para discussão / Instituto de Pesquisa Econômica Aplicada.- Brasília : Rio de Janeiro : Ipea , 1990-

ISSN 1415-4765

1.Brasil. 2.Aspectos Econômicos. 3.Aspectos Sociais. I. Instituto de Pesquisa Econômica Aplicada.

CDD 330.908

JEL: H11; I28; C81.

SUMÁRIO

SINOPSE

Neste trabalho, são analisados alguns métodos que podem ser usados para realizar agrupamentos de unidade que compõem algum sistema de provisão de serviços públicos (por exemplo, escolas, postos de saúde e delegacias). Os métodos aqui analisados são: *i*) o baseado em alguma *partição espacial* das unidades (por exemplo, agrupamento por região administrativa de uma cidade); *ii*) o que se denominou *indicador ordenado*, no qual as unidades do sistema são agrupadas segundo alguma partição da distribuição (por exemplo, décimos) da média ponderada das variáveis de interesse das unidades; e *iii*) o de *clusters*, que utiliza uma medida de distância entre essas múltiplas variáveis para agrupar as unidades do sistema de forma a que haja uma elevada homogeneidade das unidades dentro dos grupos e, ao mesmo tempo, uma alta heterogeneidade entre os grupos. Um exercício empírico utilizando informações do sistema de escolas públicas municipais do Rio de Janeiro é apresentado no texto para fins de comparação dos métodos em foco.

Palavras-chave: *clusters*; provisão de serviços públicos; organização de dados.

ABSTRACT

In this study we analyze some methods that can be used to perform grouping of unit that make up a public service delivery system (e.g., schools, health centers and police stations). The methods discussed here are: *i*) one based on some spatial partition of the units (for example, grouping by administrative region of a city); *ii*) what we call ordered indicator, in which the system units are grouped according to some partition of the distribution (e.g., deciles) of the weighted average of the variables of interest of the units; and *iii*) clusters that use a distance measure across these multiple variables to group the units of the system so that there is a high homogeneity of the units within the groups and, at the same time, a high heterogeneity among the groups. An empirical exercise using information from the municipal public school system of Rio de Janeiro is presented in the text for comparison of the methods in focus.

Keywords: clusters; public service provision; data structure.

1 INTRODUÇÃO

Monitorar o desempenho de unidades provedoras de serviços públicos como escolas, postos de saúde e delegacias é uma atividade fundamental para gestão desses serviços. Entre outras utilidades, o monitoramento permite a detecção de diversos tipos de problema dentro do sistema, o acompanhamento da conclusão das metas de cada unidade, bem como o fornecimento de informações para a avaliação dos efeitos potenciais de políticas e programas públicos.

É usual que os gestores desses serviços façam comparações do desempenho das unidades que compõem tais sistemas. Normalmente, estas comparações são realizadas por meio de dois expedientes. O primeiro é utilizar algum parâmetro global do sistema, tipicamente a média ou a mediana de desempenho de todas as unidades. O segundo é comparar unidades que pertençam a uma mesma partição espacial, tais como bairros, regiões administrativas ou mesmo municípios. No entanto, embora úteis, essas comparações têm suas limitações, já que pode haver elevada heterogeneidade de características entre as unidades do sistema. As escolas de uma grande cidade, por exemplo, tendem a ser bastante diferentes com respeito ao nível socioeconômico dos seus alunos – fato que pode afetar significativamente tanto o nível quanto as variações observadas no desempenho das escolas ao longo do tempo. Nessas situações, é desejável agrupar unidades que sejam semelhantes entre si, de modo a permitir comparações que levem em consideração as diferenças de condições das unidades do sistema.

Vários são os métodos disponíveis de agrupamento de unidades. Um comumente utilizado é baseado num indicador composto pela média ponderada de um conjunto de variáveis e que, após ser ordenado entre as unidades do sistema, é repartido em dois ou mais grupos contíguos das unidades.[1] O principal problema desse tipo de método é encontrar bons ponderadores para as variáveis que compõem o indicador. Outro método é o de agrupamento por *clusters*. Neste método, as unidades são agrupadas com base em alguma métrica que combine as diversas variáveis que caracterizam as unidades. Os *clusters* formados procuram agrupar unidades similares com base nessa métrica. Na próxima seção, serão apresentados em mais detalhes os procedimentos tipicamente utilizados na metodologia dos *clusters*.

1. Este método é utilizado pelo Departamento de Educação de Nova Iorque para agrupar as escolas da cidade (New York, 2011).

Qualquer que seja o método empregado, um ponto importante é definir quais são as características das unidades a serem utilizadas na análise. Embora a escolha dependa da situação em questão, quando se quer analisar o efeito de ações dos gestores sobre alguma dimensão das unidades (por exemplo, o desempenho dos alunos num sistema escolar), as variáveis escolhidas devem ser exógenas às unidades em análise, ou seja, as variáveis devem captar as condições externas enfrentadas por cada unidade e que podem afetar a dimensão de interesse. Isso implica que não devem entrar variáveis associadas às decisões e ações dos gestores do sistema, pois elas são potencialmente endógenas, no sentido de que podem refletir um esforço dos gestores para modificar o desempenho das unidades do sistema. No exemplo das escolas, anteriormente mencionado, a escolha recairia sobre variáveis que captem as condições socioeconômicas dos alunos, mas não, por exemplo, sobre o volume de recursos recebidos pela escola ou o número de professores que nela trabalham.

Além desta introdução, este estudo possui três seções. Na próxima, discutem-se algumas das principais metodologias de agrupamento disponíveis na literatura. A terceira seção está dedicada à apresentação dos dados que foram empregados no exemplo empírico deste trabalho, que é baseado no sistema de escolas municipais de ensino fundamental da cidade do Rio de Janeiro. A última seção contém os resultados da aplicação dos métodos utilizados para esses dados.

2 METODOLOGIA

Neste estudo, serão utilizados três tipos de agrupamentos das unidades que compõem um sistema de provisão de serviços públicos. O primeiro é baseado em alguma partição espacial das unidades, por exemplo, bairros ou áreas de planejamento/administração de uma cidade ou estado. Denominar-se-á este método de *partição espacial*. No exemplo empírico das escolas da cidade do Rio de Janeiro, será utilizada a partição dessas escolas – denominada Coordenadoria Regional de Educação (CRE) –, que é realizada pela própria prefeitura da cidade.

O segundo tipo de agrupamento é baseado na ordenação das unidades do sistema segundo um indicador computado pela média ponderada de um conjunto de variáveis pré-escolhidas para compor este indicador. Ao transformar múltiplas variáveis em um

único indicador, esse método tem a vantagem de sintetizar, em um único escalar, todas as informações que foram selecionadas sobre as unidades do sistema. Sua principal desvantagem é que ele depende do esquema de ponderação das variáveis, que normalmente é definido *ad hoc*. Mais formalmente, seja X_{jk} uma das $k = 1,\ldots,K$ características pré-escolhidas para compor o indicador de interesse das $j = 1,\ldots,J$ unidades que compõem o sistema. Dado um vetor de pesos $W' = (W_1, W_2, \ldots, W_K)$, o indicador para cada escola j é calculado por:

$$X_j = W_1 X_{j1} + W_2 X_{j2} + \cdots + W_K X_{jK} = \sum_{k=1}^{K} W_k X_{jk}. \quad (1)$$

Para cada vetor de pesos W escolhido, o indicador assumirá um valor distinto. Tipicamente, o indicador é calculado com base na média aritmética simples das variáveis, ou seja, cada variável recebe o mesmo peso $1/K$:

$$X_j = \left(\frac{1}{K}\right) X_{j1} + \left(\frac{1}{K}\right) X_{j2} + \cdots + \left(\frac{1}{K}\right) X_{jK} = \frac{1}{K}\sum_{k=1}^{K} X_{jk}. \quad (2)$$

No exemplo empírico, utilizar-se-á a média aritmética simples das características das escolas municipais do Rio de Janeiro.

Tendo calculado o valor do indicador para o conjunto das unidades, $\{X_j\}_{j=1}^{J}$, faz-se um ordenamento (por exemplo, ascendente) desse conjunto de valores para se agrupar as unidades de acordo com algum critério de partição. No caso aqui analisado, cada escola foi agrupada de acordo com o décimo da distribuição do indicador ao qual ela pertence. Naturalmente, há inúmeros outros critérios que podem ser empregados, por exemplo, o utilizado pela cidade de Nova Iorque, que agrupa cada escola do sistema com as vinte escolas acima e as vinte escolas abaixo dela em termos do valor do indicador escolhido (New York, 2011, p. 3-4). Qualquer que seja o caso, denominar-se-á este como o método de *indicador ordenado*.

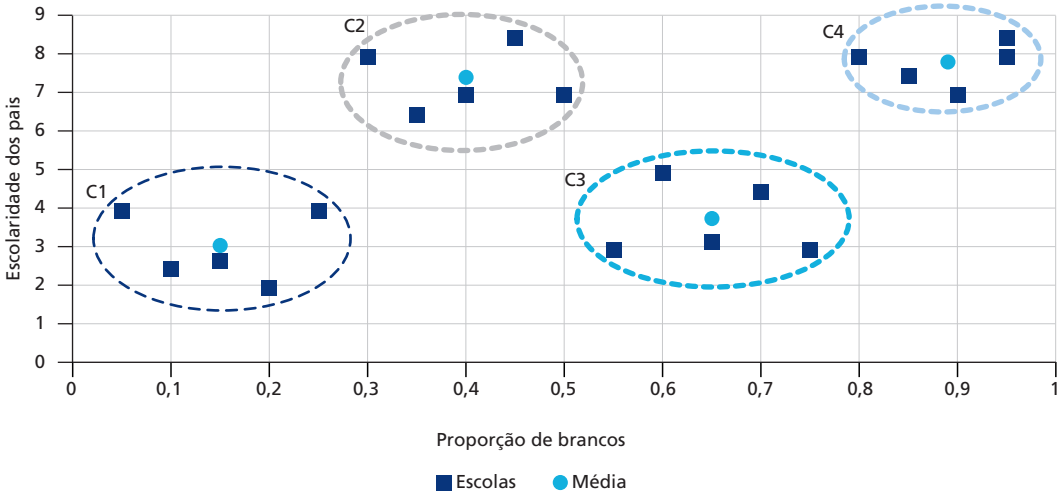
O terceiro tipo que se empregará será o agrupamento de *clusters*.[2] Distintamente do procedimento anterior, o método de *clusters* não sintetiza o conjunto de variáveis em um único indicador. Distintamente, ele mantém a multidimensionalidade dessas

2. A nossa exposição do método de *clusters* será naturalmente sucinta. Tratamentos detalhados do método podem ser encontrados em diversos livros de estatística, por exemplo, Hair *et al*. (2009).

variáveis e utiliza uma medida de distância para combiná-las de forma a encontrar os grupos de unidades que sejam as mais similares entre si. O que método busca é que haja uma elevada homogeneidade das unidades dentro dos grupos e, ao mesmo tempo, uma alta heterogeneidade entre os grupos. Ao não condensar as características das unidades num indicador, o método realiza um agrupamento de unidades e não de variáveis.

A fim de tornar mais clara a distinção entre o método de *cluster* e o do indicador ordenado, considere-se uma situação hipotética na qual um gestor quer agrupar um conjunto de 20 escolas com base em apenas duas características dessas escolas: a escolaridade média dos pais dos estudantes de cada escola e proporção dos estudantes que são de cor branca. O gráfico 1 plota o valor destas duas variáveis para as 20 escolas, cada qual correspondendo a um triângulo no gráfico. De uma simples observação da disposição dos triângulos no espaço das duas variáveis é possível ver que existem quatro *clusters* naturais das 20 escolas, os quais estão marcados pelas circunferências no gráfico. Note que os *clusters* são formados pela proximidade entre as escolas segundo o par de variáveis que as está caracterizando.

GRÁFICO 1
Exemplo hipotético de *clusters* de escolas



Elaboração dos autores com base num exemplo hipotético.

O método do indicador ordenado calcula inicialmente a média ponderada das variáveis para cada escola e posteriormente as ordena de acordo com o valor dessa média. A tabela 1 apresenta as informações que permitem demarcar quatro grupos formados por cada método. As colunas 3 e 4 contêm os valores das duas variáveis para cada escola tais

como apresentados no gráfico 1. A última coluna exibe o valor do indicador sintético, o qual foi obtido pela média aritmética dessas duas variáveis para cada escola. A primeira e a quinta colunas informam o agrupamento a que pertence cada escola respectivamente pelos métodos de *clusters* e do indicador ordenado. A segunda coluna informa o número da escola de acordo com o método de *cluster*, ao passo que a sexta coluna rearranja esse mesmo número com base no método do indicador sintético. Como se pode ver, os dois métodos chegam a resultados distintos em relação à partição das vinte escolas em grupos.

TABELA 1
Exemplo hipotético de agrupamentos de escolas pelos métodos de *cluster* e do indicador ordenado

Clusters				Indicador ordenado		
Grupos	Escola	Proporção de brancos	Escolaridade dos pais	Grupos	Escola	Média aritmética
C1	1	0,05	4,0	IO1	4	1,10
C1	2	0,10	2,5	IO1	2	1,30
C1	3	0,15	2,7	IO1	3	1,43
C1	4	0,20	2,0	IO1	11	1,78
C1	5	0,25	4,0	IO1	15	1,88
C2	6	0,30	8,0	IO2	13	1,93
C2	7	0,35	6,5	IO2	1	2,03
C2	8	0,40	7,0	IO2	5	2,13
C2	9	0,45	8,5	IO2	14	2,60
C2	10	0,50	7,0	IO2	12	2,80
C3	11	0,55	3,0	I O3	7	3,43
C3	12	0,60	5,0	IO3	8	3,70
C3	13	0,65	3,2	IO3	10	3,75
C3	14	0,70	4,5	IO3	18	3,95
C3	15	0,75	3,0	IO3	6	4,15
C4	16	0,80	8,0	IO4	17	4,18
C4	17	0,85	7,5	IO4	16	4,40
C4	18	0,90	7,0	IO4	9	4,48
C4	19	0,95	8,5	IO4	20	4,48
C4	20	0,95	8,0	IO4	19	4,73

Elaboração dos autores com base num exemplo hipotético.

Nesse exemplo fictício, a configuração dos grupos pelo método de *clusters* foi bastante direta, pois as escolas estavam naturalmente arranjadas em forma de *clusters* isolados. Entretanto, na maior parte dos casos, esse tipo de padrão não é fácil de visualizar, seja porque o número de variáveis utilizadas para formar os *clusters* é

elevado, seja porque as unidades de interesse se encontram arranjadas de forma mais homogênea. Assim, é preciso fazer uso de uma metodologia mais estruturada para demarcar os *clusters*.

Existem duas questões básicas no método de *clusters*: *i*) qual a medida de similaridade entre as unidades do sistema que será utilizada para construir os *clusters*; e *ii*) como formar os grupos, ou seja, como agrupar as unidades de forma a maximizar simultaneamente a similaridade das unidades *intraclusters* e as diferenças *entreclusters*.

A questão da medida similaridade em análises de *cluster* é tipicamente baseada na noção de distância entre dois pontos ou objetos quaisquer, ou seja, quanto menor a distância entre eles, maior a similaridade.[3] Por exemplo, quanto mais próximos estiverem dois triângulos quaisquer no gráfico 1, mais similares eles serão em termos das variáveis que os caracterizam na figura. Essa noção é fácil de entender e visualizar quando se está num plano cartesiano ou até tridimensional. Todavia, ao se passar para um número maior de dimensões, torna-se imperativo fazer uso de medidas formais de distância.

Sem dúvida, a medida de distância mais utilizada é a euclidiana. Sejam dois pontos quaisquer, X e Y, com dimensão K, que corresponde ao número de variáveis pré-selecionadas para construir os *clusters*. A distância euclidiana é calculada por:

$$d(X,Y) = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \cdots + (X_K - Y_K)^2}\,. \tag{3}$$

Se K=2, então a distância Euclidiana passa a ser simplesmente o comprimento do segmento de reta que liga dois pontos no plano cartesiano. No gráfico 1, ela corresponde ao tamanho do segmento de reta que liga dois triângulos quaisquer na figura. Em alguns casos, utiliza-se uma variante da distância euclidiana, que é seu valor ao quadrado.[4]

3. Embora menos utilizadas em análises de *cluster*, há duas outras formas de se medir similaridade entre pares de unidades. A primeira é o coeficiente de correlação, que é calculado para o vetor de variáveis pré-estabelecidas para cada par de unidades do sistema. A segunda é alguma medida de associação entre as unidades em casos em que as características destas unidades não são medidas de forma métrica, por exemplo, respostas qualitativas dadas por respondentes de uma pesquisa.
4. Outras medidas de distância também são empregadas, por exemplo, a soma dos valores absolutos das diferenças das variáveis (conhecida como *city-block*) ou essa mesma soma elevada a algum expoente escolhido pelo analista.

A segunda questão a ser tratada é o método para a formação dos *clusters*. Em linhas gerais, há duas grandes abordagens, a *hierárquica* e a de *partição*.[5] A primeira é baseada na conformação sequencial de uma "árvore" de *clusters*, que pode ser realizada a partir de dois procedimentos. No primeiro, denominado *aglomerativo*, cada unidade representa inicialmente um *cluster* e, sequencialmente, as unidades vão se aglomerando para formar os *clusters* até que se chegue a um único *cluster* final com todas as unidades. Em cada passo da sequência, um *cluster* é aglomerado a outro de acordo com a proximidade existente entre eles. Dada a medida de distância escolhida (por exemplo, a euclidiana), o agrupamento dos *clusters* a cada passo é realizado por meio de algum método de aglomeração. Existem diversos métodos de aglomeração, os mais comuns sendo os de ligação (*linkage*).

1) Simples (*single*): um novo *cluster* é formado a cada passo de acordo com a distância mínima existente entre duas unidades pertencentes a *clusters* distintos, ou seja, de acordo a distância do vizinho mais próximo.
2) Completo (*complete*): similar ao método anterior, porém aglomera os *clusters* de acordo com a distância máxima existente entre duas unidades pertencentes a *clusters* distintos.
3) Média: distintamente dos dois procedimentos anteriores – que se baseiam em distâncias extremas –, esse método calcula a média de distância entre as unidades de um mesmo *cluster* e agrupa os novos com base na menor distância dessas médias.
4) Centroide: calcula-se o centroide de cada *cluster* (ou seja, o vetor de médias das características das unidades do *cluster*) e aglomeram-se com base na menor distância entre os centroides.
5) Ward: este método, proposto originalmente por Ward (1963), agrupa os *clusters* buscando a cada passo a aglomeração que produz o menor incremento na soma das variâncias dentro dos *clusters*.

O outro procedimento hierárquico, denominado *divisivo*, pode ser entendido como o reverso do aglomerativo, isto é, inicia-se com o *cluster* formado por todas as unidades e a cada passo os *clusters* mais dissimilares são particionados até que se chegue à situação em que cada um seja formado por uma unidade do sistema. Devido a sua maior complexidade operacional e semelhança com o método aglomerativo, este procedimento não está disponível na maior parte dos *softwares* estatísticos.

5. Há uma terceira, da qual não trataremos, que é baseada na *densidade* com que as unidades estão distribuídas no "espaço". Este método é particularmente interessante para demarcar *clusters* com formatados não convexos.

Qualquer que seja o procedimento hierárquico utilizado, esta abordagem não informa qual o número de *clusters* "ideal", ou seja, não há uma regra de parada no processo de construção (ou desconstrução) da "árvore" de *clusters*. Em boa medida, isto se deve ao fato de que não existem critérios (estatísticos) gerais que informem qual a melhor solução final para o número de *clusters* presentes na estrutura de dados. Apesar disso, algumas medidas e procedimentos têm sido propostos na literatura especializada para orientar a escolha. Dois exemplos utilizados pelo *software* Stata são: o índice pseudo-F proposto por Calinski e Harabasz (1974) e o índice de Duda, Hart e Stork (2001).

Além do método hierárquico, a análise de *clusters* também utiliza com frequência a abordagem de partição. Neste método, escolhe-se inicialmente um número fixo de *clusters* e as unidades vão se arranjando através de um procedimento iterativo que tipicamente busca reduzir a dispersão das unidades *intraclusters*. Dois são os procedimentos mais empregados nesta abordagem, o *kmeans* e o *kmedians*. No primeiro, tendo pré-estabelecido o número de *clusters* (digamos *L*), escolhe-se uma semente (*seed*) inicial de valores – tipicamente designados de forma aleatória – para *L* vetores de características das unidades. No primeiro passo, as unidades são agrupadas de acordo com a menor distância de cada uma em relação aos *L* vetores escolhidos.[6] No passo seguinte, calculam-se os centroides dos *L* clusters formados (i.e., a média dos vetores de características de todas as unidades de cada *cluster*), e cada unidade é realocada ao grupo cujo centroide é o mais próximo. Esse processo de recálculo dos centroides e rearranjo das unidades em *clusters* é repetido até que não haja mais troca de unidades entre eles (ou, equivalentemente, até que os centroides não mudem de valor). O procedimento de *kmedians* é semelhante, sendo a única diferença o cálculo da mediana – e não da média – dos vetores de características das unidades de cada *cluster*.

Um dos principais problemas da abordagem de partição é que o arranjo final dos *clusters* pode ser bastante dependente do ponto de partida, isto é, da semente inicial de valores.[7] Não há uma solução totalmente robusta para esse problema, mas realizar

6. É comum realizar uma análise para a presença de *oultiers* na metodologia de *clusters* (principalmente a de *kmeans*), uma vez que esse tipo de unidade pode acabar distorcendo bastante o resultado final. Em alguns casos, no entanto, não se quer eliminar essas unidades, seja porque elas são de interesse em si (por exemplo, unidades extremas ou "excêntricas"), seja porque não se pode deixá-las de fora, como no exemplo das escolas.

7. Esse é um problema comum a vários problemas de otimização, em que não se tem garantia de que a solução final corresponde ao ótimo global. Na realidade, a abordagem de participação de *kmeans* pode ser entendida como um problema de minimização da soma dos erros quadrados, já que este método calcula o desvio de cada ponto ao centroide mais próximo e depois soma esses desvios ao quadrado. A solução final é a que oferece a menor soma desses erros ao quadrado.

a inicialização com diferentes sementes aleatoriamente escolhidas é uma das mais empregadas. Outro problema dessa abordagem é que o número de *clusters* é estabelecido *ad hoc*. Normalmente, esse problema é resolvido realizando-se todo o procedimento para diferentes números de *clusters* e comparando-se as respectivas soluções finais por meio de algum indicador.

3 DADOS E ESPECIFICAÇÕES

O exemplo empírico deste trabalho é baseado nas 982 escolas de ensino fundamental do município do Rio de Janeiro em 2011. Para medir o que se chama de condições externas às escolas na introdução deste estudo, serão utilizadas informações sobre duas características dos alunos: *i*) a proporção de alunos brancos na escola; e *ii*) a escolaridade média dos pais deles – especificamente, a média da escolaridade da mãe e do pai de todos os alunos desta escola. Naturalmente, outras variáveis poderiam entrar no exercício, mas optou-se pela parcimônia. De todo modo, como o interesse é relacionar os agrupamentos de escolas a serem formados com o desempenho dos alunos em provas padronizadas, essas duas variáveis são apropriadas, pois sabe-se que elas tendem a afetar (positivamente) esse desempenho. Isso se deve ao fato de que, em média, alunos brancos possuem condições econômicas melhores que os não brancos; alunos com pais mais escolarizados se desempenham melhor na escola que os com pais menos escolarizados.[8]

As informações sobre cor dos alunos e escolarização dos pais foram obtidas pelos dados disponíveis na Prova Brasil. Nem todas as escolas do município possuíam essas informações em 2011. Nesses casos, buscaram-se as informações dessas escolas na Prova Brasil de 2009 e, caso essas também não estivessem disponíveis naquele ano, atribuiu-se a média do bairro onde se encontravam estas escolas no ano de 2011. Assim mesmo, cinco escolas do município do Rio de Janeiro ficaram sem informações para as duas variáveis.

As informações do desempenho dos alunos em provas padronizadas de português e matemática foram obtidas pela Prova Rio para as 3ª, 4ª, 7ª e 8ª séries das escolas. No exercício, não se fez distinção entre as escolas que possuíam somente o primeiro ciclo

8. Vale assinalar que, com base numa regressão, essas duas variáveis mostram valor preditivo positivo e estatisticamente significativo sobre o desempenho dos alunos nas provas padronizadas de português e matemática da Prova Rio (resultados não mostrados).

ou o segundo ciclo (ou ambos), portanto as médias foram computadas a partir das notas de ambas as matérias para todas as séries que as escolas possuíam. Em 2011, havia quinze escolas do município sem informação para a Prova Rio. Esse conjunto de escolas não foi utilizado em nenhuma etapa do exercício. A amostra final ficou, portanto, com 962 (= 982 - 15 - 5) escolas.

Construíram-se os agrupamentos das escolas utilizando os três métodos descritos na seção anterior. Para o de partição espacial usou-se a divisão das dez Coordenadorias Regionais de Educação (CREs) existentes em 2011. Naturalmente, como esse método é baseado somente numa divisão do espaço do município, não houve uso das informações das duas características das escolas para criar os agrupamentos.

Para os outros dois métodos, realizou-se inicialmente uma padronização usual (*z-score*) das variáveis, ou seja, para cada variável subtraiu-se seu valor da média geral de todas as escolas e dividiu-se o resultado pelo desvio-padrão geral. Esse é um procedimento comumente empregado em análises de *cluster* e tem como benefício principal eliminar vieses decorrentes do uso de diferentes escalas para medir cada variável (que é o caso aqui estudado). A padronização utilizada transforma as variáveis brutas em novas variáveis com média zero e desvio-padrão unitário.

No método do indicador ordenado, calcula-se a média aritmética simples no nível da escola das variáveis padronizadas proporção de alunos brancos e escolaridade dos pais. Após ordenar o indicador, realiza-se uma partição das escolas baseada nos décimos da distribuição desse indicador, ou seja, as unidades foram agrupadas em dez grupos com (aproximadamente) o mesmo número de escolas em cada grupo. Outras formas de divisão poderiam ter sido utilizadas – por exemplo, empregando o procedimento de Nova Iorque (seção 2) –, mas optou-se pela partição em dez grupos por este ser o número de CREs.

Utilizaram-se as duas abordagens usuais de *cluster*: de partição e hierárquica. No primeiro método, estabeleceu-se, também, em dez o número de *clusters* a serem formados e aplicou-se o procedimento de *kmeans* com medida euclidiana de distância. Como se trata apenas de um exercício ilustrativo dos resultados, utilizou-se apenas um valor inicial para a semente de partida do procedimento (*seed*). No método hierárquico, aplicou-se o método aglomerativo de Ward com distância euclidiana ao quadrado. A operacionalização deste método foi feita com diferentes regras de parada para o

número de *clusters* (especificamente, 2 a 15 *clusters*). Apesar de o número de *clusters* "ideal" ter sido menor que dez, segundo o índice pseudo-F de Calinski and Harabasz (1974), optou-se por apresentar os resultados com 10 *clusters* para manter a consistência com os demais métodos.

4 RESULTADOS

A tabela 2 apresenta o perfil dos agrupamentos formados por cada um dos métodos. A segunda e a terceira colunas mostram respectivamente o total e a proporção das escolas em cada grupo por método. Como se pode ver, cada método aloca as escolas de forma distinta, o que era esperado. Como serve para fins de administração e coordenação do sistema educacional do município do Rio de Janeiro, o sistema de CREs segue uma lógica de proximidade espacial entre as escolas e, pelo que se vê na tabela, não há grande discrepância na proporção de escolas em cada CRE. No caso do método do indicador ordenado, não há nenhuma discrepância, mas isso se deve ao fato de que se impôs, para cada grupo, 10% do total de escolas da amostra final. Já nos métodos de *cluster* de partição e hierárquico, vê-se bastante variação em seus tamanhos, com alguns contendo menos de 5% das escolas e outros mais de 20%.

As colunas 4 e 5 mostram a média para cada método e agrupamento das variáveis *proporção de alunos brancos na escola* e a *escolaridade média dos pais desses alunos*. Como mostra a antepenúltima linha referente a cada método, o desvio-padrão de cada característica é menor no arranjo das CREs, seguido do método do indicador ordenado. O método de partição apresenta maior dispersão para a proporção de brancos, ao passo que a escolarização dos pais é mais heterogênea no método hierárquico. A última coluna, que apresenta a média aritmética das duas características em termos padronizados – ou seja, a variável usada no método do indicador ordenado –, mostra um ordenamento de maior dispersão para o método das CREs, seguido pelo indicador ordenado, *cluster* hierárquico e *cluster* de partição. Nesse sentido, olhando apenas para essa medida de dissimilaridade entre grupos, o método de partição é o que apresenta melhor desempenho no que diz respeito à capacidade de criar dez agrupamentos distintos. As penúltima e última linhas de cada método mostram o teste *F* e seu corresponde *P*-valor para testar se as características médias são iguais entre todos os agrupamentos formados por cada método. Essas estatísticas revelam que todos os métodos foram capazes de gerar grupos diferentes nessas características.

Além de analisar se os agrupamentos formados por cada método são distintos nessas características, pode-se também testar se há diferenças em variáveis que não foram empregadas para construir os grupos e que são afetadas de alguma forma por essas características. Se os grupos forem diferentes nessas variáveis, há evidências de que são capazes de aglutinar as unidades do sistema em termos de variáveis de "resultado", isto é, de variáveis que são preditas pelas características utilizadas para formar os grupos. Como mencionado anteriormente, utilizar-se-á aqui o desempenho dos alunos nas provas de português e matemática na Prova Rio de 2011, que, como se viu, é uma variável afetada positivamente pelas duas características que construíram os agrupamentos. A tabela 3 contém os escores médios, mínimos e máximos nessas provas dos agrupamentos formados. Como se pode observar na tabela, em geral, todos os métodos geram grupos dispersos em termos do desempenho dos alunos nas provas. Ademais, os testes *F* rejeitam a hipótese de que os grupos possuem escores médios iguais – a única exceção talvez seja o caso das CREs, cuja estatística de teste só é significativa a 5%.

TABELA 2
Características dos grupos por método de agrupamento

Método/agrupamento	Número de escolas	Proporção de escolas	Percentual de alunos brancos	Escolaridade média dos pais	Indicador médio das características
CREs					
1	42	4%	32,0%	9,2	-0,05
2	99	10%	32,0%	9,0	-0,09
3	90	9%	31,0%	9,8	0,15
4	130	14%	31,0%	9,3	-0,06
5	95	10%	31,0%	10,3	0,28
6	66	7%	28,0%	9,7	-0,11
7	109	11%	32,0%	9,2	-0,05
8	136	14%	28,0%	10,3	0,13
9	96	10%	29,0%	10,2	0,13
10	99	10%	26,0%	9,2	-0,40
Desvio-padrão	27,5	3,0%	2,1%	0,5	0,19
Estatística *F*			7,0196	17,4625	6,4076
P-Valor			0,0000	0,0000	0,0000
Indicador ordenado					
1	96	10%	20,0%	7,9	-1,31
2	96	10%	23,0%	8,6	-0,79
3	96	10%	27,0%	8,8	-0,49
4	96	10%	29,0%	9,0	-0,29
5	97	10%	29,0%	9,6	-0,11

(Continua)

(Continuação)

Método/agrupamento	Número de escolas	Proporção de escolas	Percentual de alunos brancos	Escolaridade média dos pais	Indicador médio das características
6	96	10%	30,0%	9,9	0,09
7	96	10%	31,0%	10,1	0,26
8	96	10%	33,0%	10,4	0,47
9	96	10%	35,0%	10,8	0,74
10	97	10%	42,0%	11,3	1,41
Desvio-padrão	0,4	0,0%	6,1%	1,0	0,78
Estatística *F*			129,5875	143,8662	1.937,1500
P-Valor			0,0000	0,0000	0,0000
Cluster de partição					
1	99	10%	22,0%	10,4	-0,21
2	84	9%	17,0%	8,9	-1,12
3	97	10%	37,0%	9,4	0,38
4	69	7%	27,0%	11,6	0,60
5	84	9%	36,0%	7,9	-0,30
6	186	19%	31,0%	10,4	0,36
7	84	9%	25,0%	7,6	-1,08
8	161	17%	28,0%	9,1	-0,32
9	27	3%	49,0%	10,5	1,52
10	71	7%	40,0%	11,4	1,33
Desvio-padrão	45,8	4,7%	9,4%	1,4	0,90
Estatística *F*			514,9677	580,1937	678,0667
P-Valor			0,0000	0,0000	0,0000
Cluster hierárquico					
1	121	13%	23,0%	10,7	-0,02
2	199	21%	32,0%	10,5	0,47
3	25	3%	30,0%	12,1	0,95
4	76	8%	44,0%	11,3	1,49
5	149	15%	28,0%	9,4	-0,20
6	85	9%	39,0%	9,3	0,44
7	129	13%	30,0%	8,1	-0,61
8	49	5%	38,0%	7,6	-0,27
9	69	7%	16,0%	9,4	-0,97
10	60	6%	20,0%	7,9	-1,27
Desvio-padrão	52,6	5,5%	8,8%	1,5	0,85
Estatística *F*			477,5077	501,2321	565,5308
P-Valor			0,0000	0,0000	0,0000
Total	**962**	**100%**	**30,0%**	**9,6**	**0,0**

Elaboração dos autores com base nos dados da Prova Brasil (MEC).

TABELA 3
Desempenho dos alunos por método e agrupamento

Método/agrupamento	Escore médio em português e matemática	Escore mínimo em português e matemática	Escore máximo em português e matemática
CREs			
1	187,1	140,6	246,6
2	196,0	144,1	254,1
3	186,0	125,7	254,5
4	185,2	138,9	263,7
5	189,6	142,6	253,2
6	182,7	135,9	240,4
7	190,9	141,1	251,9
8	185,0	138,5	250,9
9	188,1	145,7	249,6
10	183,5	139,0	231,9
Desvio-padrão	4,0	5,5	8,6
Estatística *F*	2,2367		
P-valor	0,0180		
Indicador ordenado			
1	179,9	125,7	232,6
2	183,9	144,0	247,4
3	181,4	140,6	241,8
4	184,5	134,4	238,9
5	188,1	138,5	251,9
6	189,2	149,3	249,7
7	189,4	138,9	248,8
8	186,9	141,6	247,8
9	196,7	157,8	254,5
10	194,6	154,3	263,7
Desvio-padrão	5,4	9,4	8,6
Estatística *F*	4,1484		
P-valor	0,0000		
Cluster de partição			
1	182,7	134,4	238,9
2	181,2	125,7	232,6
3	188,9	141,6	254,1
4	183,4	154,7	235,8
5	179,8	140,6	249,7
6	194,6	138,9	254,5
7	185,2	135,9	246,8
8	187,2	138,5	245,3

(Continua)

(Continuação)

Método/agrupamento	Escore médio em português e matemática	Escore mínimo em português e matemática	Escore máximo em português e matemática
9	198,7	154,3	252,4
10	193,2	154,9	263,7
Desvio-padrão	6,3	9,8	9,6
Estatística *F*	4,5674		
P-valor	0,0000		
Cluster hierárquico			
1	181,3	134,4	238,9
2	194,6	138,9	254,5
3	186,3	157,8	231,3
4	196,6	154,9	263,7
5	188,4	138,5	245,3
6	188,1	141,6	254,1
7	186,2	140,6	247,4
8	178,1	144,1	249,7
9	181,2	125,7	232,6
10	179,6	135,9	230,0
Desvio-padrão	6,2	9,4	11,3
Estatística *F*	5,3884		
P-valor	0,0000		
Total	**187,5**	**125,7**	**263,7**

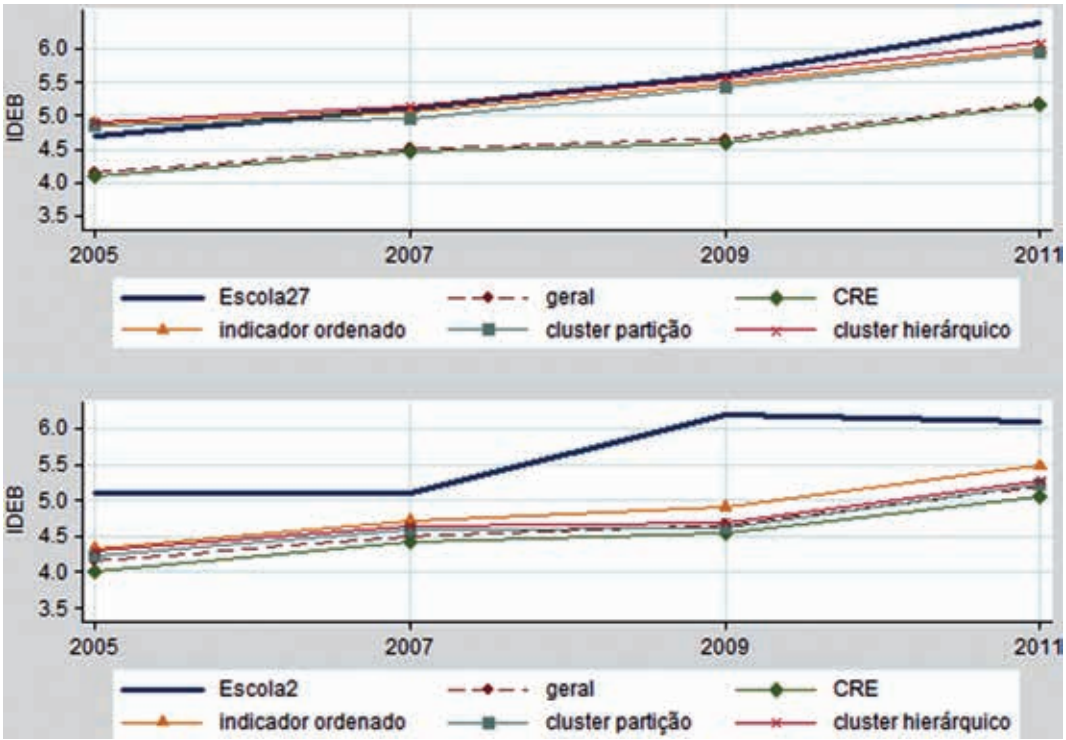
Elaboração dos autores com base nos dados da Prova Rio (Prefeitura do Rio de Janeiro).

Um importante uso de métodos de agrupamentos é o monitoramento do desempenho das unidades do sistema. No exemplo deste trabalho, os gestores podem querer acompanhar o desempenho de cada escola municipal carioca, por exemplo, para analisar o progresso alcançado após a introdução de alguma política ou programa. Em geral, esse acompanhamento é realizado por meio de comparações da escola individual com a média ou algum outro indicador geral do sistema. No entanto, nem sempre esse procedimento é o mais indicado, já que as escolas tendem a ser heterogêneas em características que influenciam seus desempenhos. Uma das utilidades da construção de agrupamentos internamente semelhantes nessas características é permitir comparações de uma escola individual com outras escolas que tenham maior grau de similaridade a ela.

Os gráficos 2 e 3 ilustram essa situação para comparações da evolução temporal respectivamente do índice de desenvolvimento da educação básica (Ideb) e da taxa de aprovação de duas escolas do município do Rio de Janeiro aleatoriamente escolhidas.

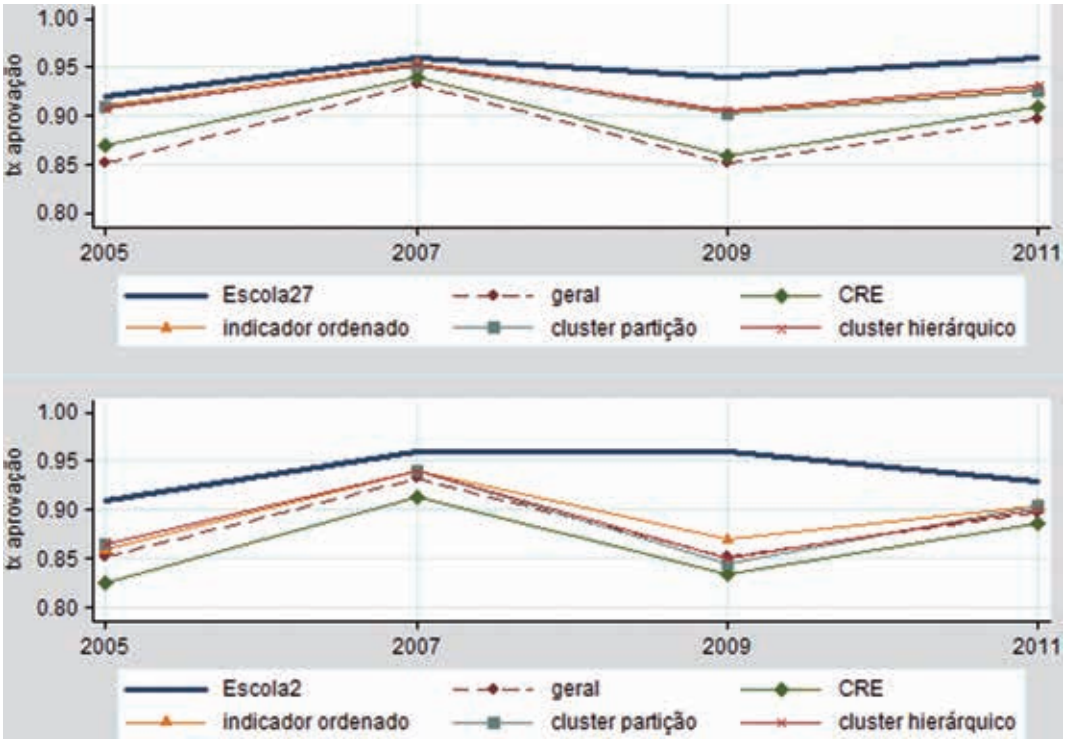
A primeira escola apresenta tanto níveis quanto trajetórias nas duas variáveis que desde o início do período são parecidos com os das escolas dos agrupamentos aos quais ela pertence nos métodos do indicador ordenado, *cluster* de partição e *cluster* hierárquico. Note-se que os níveis de desempenho dessa escola são distintos do apresentado pela média geral e pela CRE da qual ele faz parte. Já a segunda escola tem tanto níveis quanto trajetórias diferentes dos seus agrupamentos em cada o método. Por exemplo, enquanto a taxa de aprovação dessa escola permanece constante entre 2007 e 2009 e depois cai no biênio subsequente, vê-se uma queda desse indicador no primeiro intervalo e um aumento no segundo para todos os grupos aos quais ela está associada em cada método. No caso do Ideb, as diferenças também são marcadas com dois períodos de estagnação para escola em análise e com um aumento continuado ao longo de todo o período para os grupos. Essas evidências mostram que, apesar de similar a esses grupos – em termos das suas características –, essa escola teve uma evolução de desempenho diferenciada. Essa pode ser uma informação importante não só para os gestores gerais do sistema, mas também para os dirigentes da escola.

GRÁFICO 2
Evolução do Ideb de escolas individuais e dos seus agrupamentos (2005-2011)



Elaboração dos autores com base nos dados da IDEB (MEC).
Obs.: Imagem cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais disponibilizados pelos autores para publicação (nota do Editorial).

GRÁFICO 3
Evolução da taxa de aprovação de escolas individuais e dos seus agrupamentos (2005-2011)
(Em %)



Elaboração dos autores com base nos dados da IDEB (MEC).
Obs.: Imagem cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais disponibilizados pelos autores para publicação (nota do Editorial).

REFERÊNCIAS

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics**, n. 3, p. 1-27, 1974.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification and scene analysis**. 2nd ed. New York: Wiley, 2001.

HAIR, J. F. *et al*. **Multivariate data analysis**. 7th ed. New Jersey: Prentice Hall, 2009.

NEW YORK. Department of Education. **Educator guide**: the New York City progress report. New York: Department of Education, 2011.

WARD, J. H. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, n. 58, p. 236-244, 1963.

Ipea – Instituto de Pesquisa Econômica Aplicada

EDITORIAL

Coordenação
Cláudio Passos de Oliveira

Supervisão
Everson da Silva Moura
Reginaldo da Silva Domingos

Revisão
Ângela Pereira da Silva de Oliveira
Clícia Silveira Rodrigues
Idalina Barbara de Castro
Leonardo Moreira Vallejo
Marcelo Araujo de Sales Aguiar
Marco Aurélio Dias Pires
Olavo Mesquita de Carvalho
Regina Marta de Aguiar
Bárbara Seixas Arreguy Pimentel (estagiária)
Laryssa Vitória Santana (estagiária)
Manuella Sâmella Borges Muniz (estagiária)
Thayles Moura dos Santos (estagiária)
Thércio Lima Menezes (estagiário)

Editoração
Bernar José Vieira
Cristiano Ferreira de Araújo
Daniella Silva Nogueira
Danilo Leite de Macedo Tavares
Diego André Souza Santos
Jeovah Herculano Szervinsk Junior
Leonardo Hideki Higa

Capa
Luís Cláudio Cardoso da Silva

Projeto Gráfico
Renato Rodrigues Bueno

The manuscripts in languages other than Portuguese published herein have not been proofread.

Livraria do Ipea

SBS – Quadra 1 - Bloco J - Ed. BNDES, Térreo.
70076-900 – Brasília – DF
Fone: (61) 3315-5336

Correio eletrônico: livraria@ipea.gov.br

Composto em adobe garamond pro 12/16 (texto)
Frutiger 67 bold condensed (títulos, gráficos e tabelas)
Brasília-DF

Missão do Ipea
Aprimorar as políticas públicas essenciais ao desenvolvimento brasileiro por meio da produção e disseminação de conhecimentos e da assessoria ao Estado nas suas decisões estratégicas.

ISSN 1415-4765

ipea Instituto de Pesquisa Econômica Aplicada

Secretaria de Assuntos Estratégicos

GOVERNO FEDERAL BRASIL PÁTRIA EDUCADORA