Presidente da República
Luiz Inácio Lula da Silva

Ministro do Planejamento, Orçamento e Gestão
Paulo Bernardo Silva

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Eduardo Pereira Nunes

Diretor-Executivo
Sérgio da Costa Côrtes

ÓRGÃOS ESPECÍFICOS SINGULARES

Diretoria de Pesquisas
Wasmália Socorro Barata Bivar

Diretoria de Geociências
Luiz Paulo Souto Fortes

Diretoria de Informática
Luiz Fernando Pinto Mariano

Centro de Documentação e Disseminação de Informações
David Wu Tai

Escola Nacional de Ciências Estatísticas
Sérgio da Costa Côrtes (interino)

Ministério do Planejamento, Orçamento e Gestão
Instituto Brasileiro de Geografia e Estatística - IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 68 número 229 julho/dezembro 2007

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 68, n. 229, p. 1-97, jul./dez. 2007

Instituto Brasileiro de Geografia e Estatística - IBGE
Av. Franklin Roosevelt, 166 - Centro - 20021-120 - Rio de Janeiro - RJ - Brasil

Revista Brasileira de Estatística, ISSN 0034-7175

Órgão oficial do IBGE e da Associação Brasileira de Estatística – ABE.

Publicação semestral que se destina a promover e ampliar o uso de métodos estatísticos (quantitativos) na área das ciências econômicas e sociais, através de divulgação de artigos inéditos.

Temas abordando aspectos do desenvolvimento metodológico serão aceitos, desde que relevantes para os órgãos produtores de estatísticas.

Os originais para publicação deverão ser submetidos em três vias (que não serão devolvidas) para:

Francisco Louzada-Neto
Editor responsável – RBEs – IBGE.
Av. República do Chile, 500 – Centro
20031-170 – Rio de Janeiro, RJ.

Os artigos submetidos às RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.

A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Editoração
Helem Ortega da Silva - Coordenação de Métodos e Qualidade - DPE/COMEQ/IBGE

Impressão
Gráfica Digital/Centro de Documentação e Disseminação de Informações - CDDI/IBGE, em 2004.

Capa
Renato J. Aguiar – Coordenação de *Marketing*/CDDI/IBGE

Ilustração da Capa
Marcos Balster – Coordenação de *Marketing*/CDDI/IBGE

Revista brasileira de estatística/IBGE, - v.1, n.1 (jan./mar.1940), - Rio de Janeiro:IBGE, 1940-

v.

Trimestral (1940-1986), semestral (1987-).

Continuação de: Revista de economia e estatística.

Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v. 50 (1980-1989).

Co-edição com a Associação Brasileira de Estatística a partir do v.58.

ISSN 0034-7175 = Revista brasileira de estatística.

I. Estatística – Periódicos. I. IBGE. II. Associação Brasileira de Estatística.

IBGE. CDDI. Div. de Biblioteca e Acervos Especiais CDU 31 (05)
RJ-IBGE/88-05 (rev.98) PERIÓDICO

Impresso no Brasil/*Printed in Brazil*

Nota do Editor

Neste segundo volume da RBEs do ano de 2007 temos quatro artigos interessantes. O primeiro artigo, de autoria de Maria Eugénia Ferrão, Kaizô Iwakami Beltrão e Denis Paulo dos Santos, avalia o impacto do regime de organização do ensino (promoção automática) sobre o desempenho dos alunos da quarta série da Região Sudeste do Brasil. O segundo artigo, de autoria de Simone Maffini Cerezer, Sílvia Regina Costa Lopes e Valdério Anselmo Reisen, compara a estimação do parâmetro de diferenciação de processos de longa dependência, ARFIMA(p, d, q), quando $d \in (0,0;0,5)$ utilizando os métodos exato e aproximado para simulação do processo. O terceiro artigo, de autoria de Maria Ivanilde Araújo e Basilio de Bragança Pereira, propõe procedimentos Bayesianos alternativos de comparação de sistemas de equações de regressão separadas (não encaixadas). O quarto artigo, de autoria de José André de Moura Brito, Rosemary Vallejo de Azevedo e Flávio Marcelo Tavares Montenegro, discute um novo algoritmo para o problema de estratificação com base em algoritmos genéticos.

Aproveito a oportunidade para agradecer a colaboração de todos os Editores Associados, revisores do periódico, autores e à equipe do IBGE.

Uma excelente leitura.

Francisco Louzada-Neto
Editor Responsável

Sumário

Algoritmos genéticos aplicados ao problema de estratificação

José André de Moura Brito[*]
Rosemary Vallejo de Azevedo[*]
Flávio Marcelo Tavares Montenegro[*]

Resumo

O presente trabalho trata de um novo algoritmo proposto para o problema de estratificação. Uma vez definida uma população de tamanho N, uma amostra de tamanho n e um número fixado L de estratos, deve-se determinar quais observações da população estão associadas a cada um dos estratos de forma a minimizar a soma das variâncias dos estimadores em cada estrato. Para a resolução deste problema, é proposta uma heurística baseada em Algoritmos Genéticos. Um conjunto de resultados computacionais, relacionados a populações assimétricas, é apresentado e usado para comparações de desempenho com outros métodos da literatura.

Palavras-chave: Estratificação, amostra, algoritmos genéticos.

[*] Endereço para correspondência: Instituto Brasileiro de Geografia e Estatística – IBGE - Av.República do Chile, nº 500, 10º andar, CEP: 21031-170 – Rio de Janeiro – RJ. *e-mails*: flavio.montenegro@ibge.gov.br, jose.m.brito@ibge.gov.br e rosemary.azevedo@ibge.gov.br.

1. Introdução

O levantamento de informações por amostragem se constitui em uma ferramenta estatística de especial importância, pois permite a obtenção de estimativas de parâmetros populacionais desconhecidos por meio da observação de apenas uma parte (amostra) dos elementos do universo de estudo (população).

Os elementos de uma população são as unidades de observação e análise, determinadas pelos objetivos do levantamento. Do ponto de vista matemático, a população é definida como um conjunto de elementos que possuem pelo menos uma característica em comum. Na prática, compreende um agregado de elementos, que deve ser definido em termos de sua localização no espaço e no tempo.

De acordo com as características da população em estudo, das restrições de orçamento e do grau de precisão das informações que se deseje obter a partir da amostra, podemos considerar diferentes esquemas de amostragem (Cochran, 1977 e Bolfarine e Bussab, 2005).

Dentre tais esquemas, o presente trabalho concentra-se, em particular, no estudo da amostragem estratificada, sendo proposta uma nova metodologia de resolução para o problema de delimitação dos estratos. Este problema consiste em separar uma população de tamanho N em um conjunto de L subpopulações, chamadas de estratos. A construção desses estratos é efetuada levando-se em conta uma variável de estratificação X, também chamada de medida de tamanho, cujas observações são conhecidas para todas as unidades da população.

Assim, o estrato 1 será definido pelas observações de X que sejam menores ou iguais a um valor b_1, o estrato h $(h=2...,L-1)$ será definido pelas observações de X que estejam no intervalo (b_{h-1},b_h), onde $b_1<b_2<.....<b_h<.....<b_{L-1}$ são os limites desses estratos, e o estrato L pelas observações maiores do que b_{L-1}. O problema de estratificação a ser tratado consiste em encontrar limites de forma a minimizar a variância de um estimador de total de uma variável de estudo Y (que seja correlacionada com a variável X) ou a própria variável X, considerando um tamanho total fixo de amostra. Este estudo trata simultaneamente do problema da delimitação e

da alocação ótima de Neyman na repartição do tamanho de amostra entre os estratos.

Com a finalidade de fornecer um método de resolução alternativo aos já existentes na literatura referente ao problema, será proposto neste trabalho um procedimento heurístico baseado em algoritmos genéticos.

A apresentação deste trabalho é dividida da seguinte forma: Na seção 2, são apresentados de forma concisa os principais conceitos de amostragem estratificada, com a finalidade de auxiliar o entendimento do problema de estratificação e da nova metodologia proposta. Na seção 3, apresenta-se o problema de estratificação e uma breve descrição das metodologias existentes na literatura. Na seção 4, é apresentada uma nova proposta de metodologia para a resolução do problema. Como conclusão do trabalho, apresenta-se um conjunto de resultados computacionais obtidos pelo algoritmo proposto, a partir de dados reais, e uma comparação destes com os resultados obtidos por outras metodologias existentes na literatura.

2. Conceitos básicos sobre amostragem estratificada

Na amostragem estratificada, uma população U com N unidades é particionada em L subpopulações com $N_1, N_2, ..., N_h, ..., N_L$ unidades, respectivamente, chamadas de estratos. Essas subpopulações não se superpõem e, juntas, abrangem a totalidade da população, de tal modo que

$$N_1 + N_2 + ... + N_h + ... + N_L = N \tag{2.1}$$

Depois de definidos os estratos, a partir do conhecimento de uma ou mais características da população, seleciona-se, de maneira independente, uma amostra em cada um deles, com tamanhos denotados por $n_1, n_2, ..., n_h, ..., n_L$, respectivamente. Os parâmetros são estimados para cada estrato e as estimativas são agregadas para o conjunto da população.

A seguir, apresenta-se a notação básica associada à amostragem estratificada:

N - número de observações na população;

N_h - número de observações da população no h- ésimo estrato;

n - tamanho da amostra;

n_h - tamanho da amostra no h-ésimo estrato;

Y_{hi} - valor da variável de estudo Y considerando o h-ésimo estrato e a i-ésima unidade;

$\overline{Y}_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h}$ - valor médio de Y na população do *h-ésimo* estrato;

$Y = \sum_{h=1}^{L} N_h \overline{Y}_h$ - total de Y na população considerando todos os estratos; e

$S_{yh}^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \overline{Y}_h)^2}{N_h - 1}$ - medida da variância de Y na população do *h-ésimo* estrato.

Um plano amostral estratificado requer a solução de dois tipos de problemas simultaneamente. O primeiro está relacionado com a construção dos estratos e o outro com a escolha dos métodos de seleção e alocação das amostras nos estratos Inicialmente tem que ser escolhido que tipo de estratificação é mais adequado aos objetivos da pesquisa: se a estratificação estatística ou a estratificação natural, nesta os estratos formam domínios naturais de interesse. A escolha da(s) variável(eis) para a estratificação, do número de estratos a serem formados e como estabelecer pontos de corte entre os estratos de forma eficiente são demandas que vêm em seguida. Quando a estratificação é estatística, números de estratos maiores que 6 não produzem ganhos significativos de precisão para as estimativas (ver Cochran,1977, seção 5A.8).

3. O Problema de estratificação

Considere uma população de pesquisa identificada pelo conjunto finito e ordenado U de rótulos, representando todas as unidades elementares de interesse, denotado por

$U=\{1,2,...,i,...,N\}$, e sejam: Y a variável de interesse na pesquisa, a partir da qual serão calculadas estimativas, e X a variável de tamanho (Azevedo, 2004) usada para a estratificação.

Seja $Y_U=\{y_1,y_2,...,y_N\}$ um vetor populacional associado à variável Y e $X_U=\{x_1,x_2,...,x_N\}$ o vetor populacional gerado pela variável auxiliar X, tal que, sem perda de generalidade, se supõe que $x_1\leq x_2\leq...\leq x_N$.

Suponha, ainda, que a população seja dividida em L conjuntos disjuntos e exaustivos, denotados por $U_1,U_2,...,U_h,...,U_L$. Assim, a união destes L conjuntos ou estratos corresponde à população completa, isto é, $U_1\cup U_2\cup...U_L=U$. Denota-se por N_h, conforme a seção 2, o total de unidades da população em cada estrato h e por n_h o número de unidades a serem selecionadas na amostra dentro de cada estrato h, $h=1,2,...,L$.

Algumas dessas definições, incluindo o tamanho acumulado da amostra nos estratos, estão esquematizadas na Tabela 1, apresentada a seguir.

$$M_h=\sum_{j=1}^{h}N_j\,,\quad h=1,2,...,L \qquad (3.1)$$

Tabela 1 – Visualização do processo de estratificação.

Estrato	Tamanho da População	Tamanho Acumulado	Rótulos no Estrato	Dados da Variável Auxiliar no Estrato
1	N_1	$M_1=N_1$	$U_1=\{1,2,...,M_1\}$	$\{x_1,...,x_{M_1}\}$
2	N_2	$M_2=N_1+N_2$	$U_2=\{M_1+1,...,M_2\}$	$\{x_{M_1+1},...,x_{M_2}\}$
...	...	...	...	...
h	N_h	$M_h=M_{h-1}+N_h$	$U_h=\{M_{h-1}+1,...,M_h\}$	$\{x_{M_{h-1}+1},...,x_{M_h}\}$
...	...	...	...	...
L	N_L	$M_L=M_{L-1}+N_L$	$U_L=\{M_{L-1}+1,...,M_L\}$	$\{x_{M_{L-1}+1},...,x_{M_L}\}$

Assim, a amostra é identificada a partir dos conjuntos U_h, $h=1,2,...,L$, que contém as unidades que estão em cada estrato e são tais que:

$$U_1=\{i:x_i\leq b_1\}=\{1,2,...,M_1\}\,,$$

$$U_h=\{i:b_{h-1}<x_i\leq b_h\}=\{M_{h-1}+1,...,M_h\}\,,\ h=2,3,...,L-1$$

e $U_L = \{i : b_{L-1} < x_i\} = \{M_{L-1}+1, ..., M_L\}$,

onde $b_1 < b_2 < ... < b_h < ... < b_{L-1}$ denotam os pontos de corte que delimitam os estratos na população. Uma vez determinados os valores $b_1 < b_2 < ... < b_h < ... < b_{L-1}$, é retirada uma amostra de dentro de cada um dos estratos.

Deve-se ainda destacar que as observações (componentes) repetidas, associadas ao vetor populacional X_U, devem, obrigatoriamente, pertencer a um mesmo estrato.

A partir dos tamanhos de amostra n_h e de população N_h, definidos em cada um dos estratos e considerando uma certa variável de interesse Y, pode-se então expressar o estimador de total $\hat{Y}$ (Bolfarine e Bussab, 2005):

$$\hat{Y} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{k=1}^{n_h} y_{hk} \tag{3.2}$$

sendo

Y_{hk} - valor da variável de estudo Y considerando o *h*-ésimo estrato e a *k*-ésima unidade da amostra;

O problema de estratificação consiste, portanto, em determinar os limites $b_1 < b_2 < ... < b_h < ... < b_{L-1}$ de forma a minimizar a variância de $\hat{Y}$ (ou da própria variável de tamanho X).

$$v(\hat{Y}) = \sum_{h=1}^{L} N_h^2 . \frac{S_{yh}^2}{n_h} . (1 - \frac{n_h}{N_h}) \tag{3.3}$$

Observe-se que N_h e S_{yh}^2 são definidas em função dos limites dos estratos, enquanto o total n é fixado previamente. Os valores dos tamanhos de amostra n_h, em cada um dos estratos, podem ser definidos considerando a expressão abaixo:

$$n_h = \frac{n . N_h . S_{yh}}{\sum_{g=1}^{L} N_g . S_{yg}} . \tag{3.4}$$

A expressão (3.4) está associada à chamada alocação de Neyman (Cochran, 1977 e Bolfarine e Bussab, 2005).

Sigman e Monsour (1995) comentam que a alocação de Neyman é apenas aproximadamente ótima. A razão disto é que n_h deve ser um número inteiro e, na alocação de Neyman, isto normalmente não ocorre. Em termos práticos, procura-se arredondar (para cima ou para baixo) os valores n_h para valores inteiros de forma que $\sum_{h=1}^{L} n_h$ não ultrapasse o tamanho de amostra de n. Considerando tal questão, no trabalho de Brito (2005) é proposta uma formulação de programação inteira, que efetua a alocação de Neyman e define os tamanhos de amostra alocados a cada um dos estratos como valores inteiros.

Convém observar que obter um mínimo global (Resende e Sousa, 2004) para a variância expressa acima, considerando alocação de Neyman, é um problema de difícil resolução tanto analítica quanto computacional, pois S_{yh}^2 é uma função não linear dos valores $b_1,\ldots, b_{L-1}$ e o número de possibilidades diferentes de escolha desses valores (para $N \geq 4$ observações distintas, $L>1$, e ao menos duas observações em cada estrato para evitar indeterminação em (2.4)) é, no mínimo, igual ao número de combinações de $(\lfloor N/2 \rfloor - 1)$ tomados $(L\text{-}1)$ a $(L\text{-}1)$: $C_{L-1}^{\lfloor N/2 \rfloor - 1}$.

Como conseqüência dessa dificuldade, vários métodos aproximados têm sido sugeridos. O problema de estratificação foi estudado inicialmente por Dalenius (1952) sob a hipótese de que os fatores de correção para população finita eram desprezíveis, ou seja: $(1-\frac{n_h}{N_h}) \approx 1, \forall h$. Sob essa hipótese, o problema se resumia a encontrar limites $b_1 < b_2 < \ldots. < b_h . < \ldots. < b_{L-1}$ que minimizassem a variância aproximada do estimador de total, considerando a alocação de Neyman.

Um outro método bastante conhecido é o da regra de Dalenius-Hodges (1959). Este método consiste em aproximar a distribuição dos dados da variável de estratificação X na população por um histograma com muitas classes, o que implica em adotar a hipótese de que a variável de estratificação é uniformemente distribuída

(Cochran, 1977) dentro de cada classe. Com isto, o problema tem solução simples com a aplicação da Regra da Distribuição Cumulativa da Raiz da Freqüência, ou regra de Dalenius-Hodges, cuja descrição pode ser encontrada em Cochran (1977 cap. 5).

Em estudos mais recentes, Lavallée e Hidiroglou (1988) empregaram um plano amostral usando estratificação ótima para uma população com variável de tamanho assimétrica. O método procura encontrar os limites de estratificação fixando o número de estratos e o nível de precisão desejado para estimar o total da característica de interesse, de tal modo que o tamanho total da amostra n seja mínimo. No estrato delimitado para as maiores unidades da população, denominado estrato certo, as unidades amostrais são incluídas com certeza na amostra. Nos demais estratos, denominados estratos amostrados, as unidades são selecionadas por amostragem aleatória simples sem reposição. Os autores usaram alocação potência estudado por Bankier (1988) para a distribuição do tamanho total da amostra entre os estratos amostrados. Este tipo de alocação distribui a amostra nos estratos amostrados proporcionalmente à potência da soma da variável de estratificação em cada estrato, e busca uma uniformidade dos coeficientes de variação para as estimativas de totais da variável de estratificação dentro de cada estrato.

No método desenvolvido por Hedlin (1998, 2000), é considerada a delimitação dos estratos de forma tal que a variância do estimador de total da variável de estudo, seja mínima, levando em conta um tamanho total fixo da amostra n, assumindo uma estratificação com mais de dois estratos, associando simultaneamente o emprego de um estrato certo e alocação de Neyman na distribuição da amostra nos estratos a serem amostrados.

Gunning e Horgan (2004) desenvolveram um algoritmo simples e de fácil implementação, que utiliza a fórmula do termo geral de uma progressão geométrica para construir os limites dos estratos, considerando populações assimétricas e a hipótese de que os coeficientes de variação associados a cada um dos estratos são iguais.

4. Algoritmo para o problema de estratificação

Conforme já observado, as metodologias existentes na literatura normalmente fornecem apenas soluções viáveis, não necessariamente ótimas, quando se leva em conta a minimização da variância, seja qual for o método de alocação aplicado ao problema de estratificação. Como contribuição à busca por soluções de melhor qualidade ou até mesmo ótimas, será apresentada, nesta seção, uma proposta alternativa para a resolução do problema de estratificação com alocação de Neyman, considerando que seja fixado o número de estratos e o tamanho de amostra.

Para a resolução deste problema, será proposto um algoritmo heurístico (Viana, 1998), baseado em algoritmos genéticos (Resende e Sousa, 2004 e Holland, 1975). Inicialmente, para aplicação do algoritmo, serão consideradas algumas transformações nos dados de entrada fornecidos. Tendo em vista que as N observações X_U estão em ordem crescente, é possível agrupá-las considerando-se apenas seus valores distintos. Dessa forma, obtêm-se K valores distintos de X_U, reunidos em um conjunto $Q=\{q_1,q_2,\ldots,q_K\}$, que são os possíveis limites para a estratificação da população.

Considerando, por exemplo, $N=9$, $L=3$ e $X=(2,4,4,8,10,10,10,15,15)$, obtém-se $Q=(2,4,8,10,15)=(q_1,q_2,q_3,q_4,q_5)\Rightarrow|Q|=5=K$.

Definindo $U_1=\{i\,|\,x_i\le q_2, x_i\in X_U\}$, $U_2=\{i\,|\,q_2<x_i\le q_4, x_i\in X_U\}$ e $U_3=\{i\,|\,q_4<x_i, x_i\in X_U\}$, tem-se a Tabela 2, a seguir:

Tabela 2 – Ilustração do exemplo

Estrato	Tamanho da População	Tamanho Acumulado	Rótulos no Estrato	Dados de X_U
1	$N_1=3$	$M_1=3$	$U_1=\{1,2,3\}$	$\{x_1,x_2,x_3\}$
2	$N_2=4$	$M_2=3+4$	$U_2=\{4,5,6,7\}$	$\{x_4,\ldots,x_7\}$
3	$N_3=2$	$M_3=3+4+2$	$U_3=\{8,9\}$	$\{x_8,x_9\}$

Analogamente, para L estratos e K limites , deve-se escolher (L-1) limites q_k a partir de Q. Considerando uma população de tamanho N, que será dividida em L estratos, e os valores ordenados de X_U, define-se o conjunto Q. A solução do problema consistirá, então, em determinar quais limites q_k serão selecionados do conjunto Q, de forma a obter a menor variância possível considerando as expressões (3.3) e (3.4).

4.1. Algoritmos genéticos

Os *algoritmos genéticos* (AG), introduzidos por HOLLAND (1975), baseiam-se na teoria de Darwin de evolução natural das espécies. Termos como *cromossomos*, *genes*, *alelos*, *genótipo* e *fenótipo*, comuns no estudo da Genética, têm correspondentes nesse modelo computacional especialmente proposto para simular processos evolutivos.

A idéia básica parte do princípio de que, de forma similar à teoria biológica da evolução das espécies, os "melhores" indivíduos sobrevivem e geram descendentes com suas características hereditárias. Estes elementos, que comporão as novas gerações, tendem a ter a mesma aparência, ou *fenótipo*, que seus antecessores.

Assim, de forma análoga, um algoritmo genético parte de uma população de indivíduos gerados aleatoriamente (conjunto de soluções iniciais de um problema), realiza a avaliação de cada um (calculo da função objetivo), seleciona os "melhores" (escolha daqueles cuja função objetivo tenha os maiores valores, se for um problema de maximização, ou menores, no caso de minimização) e promove manipulações genéticas, como cruzamento e mutação (correspondente a perturbações das soluções escolhidas) a fim de criar uma nova população, a partir da qual reinicia-se o processo iterativamente. Este procedimento adaptativo pode ser usado para resolver qualquer problema de otimização.

Além do trabalho pioneiro de Holland (1975), Goldberg (1989) e outros consolidaram o uso desta técnica. Em resumo, os algoritmos genéticos têm as seguintes características (Viana, 1998):

- são robustos, isto é, dependem muito pouco das soluções iniciais e do problema a ser resolvido;
- trabalham unicamente com o valor da função objetivo (sem a necessidade de manipulações explícitas de restrições formuladas no problema);
- utilizam somente regras probabilísticas (não determinísticas); e
- são de uso geral.

4.2 Algoritmo genético aplicado ao problema de estratificação

A presente seção descreve os principais componentes do *algoritmo genético* implementado para o problema de estratificação definido na seção 3.

Inicialmente, para a aplicação do AG, deve-se escolher a forma de representação dos cromossomos (soluções). Em geral, para tal representação, utilizam-se vetores. Neste trabalho, os cromossomos são representados da seguinte forma: dado o conjunto $Q = \{q_1, q_2, ..., q_K\}$ com K limites, que são os possíveis pontos de corte para definição dos estratos, e definido o número L de estratos, uma solução para o problema de estratificação corresponderá a um vetor com $L-1$ posições que contém, em seqüência crescente, $L-1$ possíveis pontos de corte extraídos de Q. Por exemplo, se $L=4$ e $K=20$, pode-se ter o seguinte vetor solução:

$S_i =$

q_4	q_{11}	q_{16}

Ou seja, os quatro estratos são definidos da seguinte forma:

$U_1 = \{i \mid x_i \leq q_4, x_i \in X_U\}$, $U_2 = \{i \mid q_4 < x_i \leq q_{11}, x_i \in X_U\}$, $U_3 = \{i \mid q_{11} < x_i \leq q_{16}, x_i \in X_U\}$ e $U_4 = \{i \mid q_{16} < x_i, x_i \in X_U\}$.

Como população inicial, são gerados m vetores (soluções), isto é, m configurações $S = \{S_1, S_2, ..., S_m\}$ que correspondem a um pequeno subconjunto do

espaço total de configurações. Cada solução define uma possível estratificação para o problema. Cada solução S_i é construída, selecionando, aleatoriamente, $L-1$ limites entre os K limites possíveis que estão no conjunto Q. E em seguida, os $L-1$ valores de S_i são ordenados crescentemente.

Uma vez definidos os pontos de corte em cada solução, ficam também definidos automaticamente os valores N_h e S_{yh}^2 e, assim, considerando um tamanho de amostra n e a expressão $n_h = \frac{n.N_h.S_{yh}}{\sum_{h=1}^{L} N_h.S_{yh}}$, é possível calcular os valores da função objetivo f_i, que corresponderá à variância $v(\hat{Y}) = \sum_{h=1}^{L} N_h^2.\frac{S_{yh}^2}{n_h}.(1-\frac{n_h}{N_h})$.

Obs.: Todas as soluções que compõem S têm pontos de corte que produzem estratos com pelo menos duas observações $(N_h \geq 2)$ e com tamanho de amostra maior ou igual a dois $(n_h \geq 2)$.

Após o cálculo da função objetivo, são aplicados operadores genéticos na seguinte ordem:

i) Reprodução dos melhores indivíduos (seleção) – Com base nos valores de f_i, as melhores configurações, ou seja, aquelas com menor valor de f_i, são selecionadas e "clonadas" em substituição às piores. Para realizar esta operação, é utilizada uma "roleta" viciada (Viana, 1998). A roleta é mecanismo de escolha aleatória que é acionada *m* vezes (uma vez para cada elemento da nova população). A cada vez que a roleta é acionada, a probabilidade de que uma dada configuração seja copiada para a nova população é inversamente proporcional ao valor de sua correspondente função objetivo. Dessa maneira, quanto menor for o valor da função objetivo relacionada a uma configuração, maior o número de cópias esperadas dessa configuração na nova população.

ii) Cruzamento – É a operação que possibilita a recombinação das estruturas genéticas da nova população gerada no passo anterior. Permite uma diversificação da população no espaço de soluções ao gerar configurações diferentes. O operador de cruzamento escolhe aleatoriamente duas soluções, S_i e S_j, e troca partes de seu padrão genético. Considere o exemplo no esquema abaixo, com $L=6$ e $K=40$. No caso de duas soluções S_i e S_j, é possível trocar até $L-2$ pontos de corte destas soluções.

No exemplo abaixo, foi efetuada a troca de um trecho com dois pontos de corte entre as soluções S_i e S_j.

Após o cruzamento, ordenamos os pontos de corte em cada solução S_i e S_j, de forma que $q_{i_1} < q_{i_2} < \ldots < q_{i_{L-1}}$ e $q_{j_1} < q_{j_2} < \ldots < q_{j_{L-1}}$. Em conseqüência, obtemos duas novas soluções S_i' e S_j'. Neste processo, verifica-se, também, se as restrições da observação acima são cumpridas.

O cruzamento pode ser feito tomando-se, dois a dois, os elementos da população inteira. Todavia, é mais comum atribuir um percentual *PX*, na faixa de 25% (0.25) a 75% (0.75) da população.

iii) Mutação – Corresponde a uma pequena perturbação, realizada em algumas configurações S_i, que tem por objetivo tentar regenerar valores entre q_1 e q_K que porventura tenham sido eliminados na reprodução ou no cruzamento. Tal operação é efetuada em um pequeno percentual dos elementos que compõem a configuração S_i (ver esquema a seguir).

No caso do problema de estratificação, a operação de mutação consistirá em selecionar um ponto de corte q_k que esteja na posição r de uma solução S_i e, em seguida, sortear um novo ponto de corte para substituir q_k. Este novo ponto de corte

estará compreendido entre $q_{k'+2}$ e $q_{k''-2}$, sendo k' o índice associado ao ponto de corte que está na posição $r-1$ de S_i e k'' o índice associado ao ponto de corte que está na posição $r+1$ de S_i. No exemplo abaixo, se selecionamos q_{17} (posição $r=2$), o novo ponto de corte possível para a 2ª posição de S_i estará compreendido entre q_{11} e q_{20} (k'=9,k"22)

Para que uma determinada população não sofra muitas mutações, esta operação é processada para um pequeno percentual (*PM*) de seus elementos (em torno de 1% dos gens). Ou seja, se $m=100$ e $PM=0.01$ (1%), teremos o número total de mutações igual a 1.

Após todos estes procedimentos terem sido executados, repetem-se, iterativamente, os passos i, ii e iii, que correspondem a uma geração ou iteração. O critério de parada do algoritmo é a repetição destes passos um certo número de vezes (*MAXGEN*). A saída do algoritmo é representada pela última população gerada, da qual a melhor configuração (pontos de corte) é a solução heurística do problema.

Alguns autores adotam um critério de tempo máximo de execução do algoritmo em substituição ao número máximo de gerações (*MAXGEN*). Além disso, é interessante incluir, no critério de parada, um parâmetro que represente o número máximo (*NMAX*) de gerações consecutivas durante as quais a solução pode permanecer inalterada.

5. Resultados Computacionais

Nesta seção são apresentados alguns resultados obtidos a partir do algoritmo genético proposto neste trabalho e dos algoritmos propostos nos trabalhos de Hedlin (1998, 2000) e Gunning e Horgan (2004).

O algoritmo apresentado na seção 4 foi implementado utilizando a linguagem R e está disponível na função chamada *gene_estrata* na biblioteca *OptStrata* e pode ser solicitado por *e-mail* a um dos autores.

A função tem os seguintes parâmetros de entrada: o nome do arquivo de entrada, o número de estratos, o tamanho de amostra, o tamanho da população (número de soluções viáveis), o número máximo de gerações, o número de iterações sem melhora, a probabilidade de cruzamento, a probabilidade de mutação, o índice da coluna relativa à variável de estratificação ou tamanho X e o índice da coluna relativa à variável de estudo Y.

A Tabela 3, abaixo, contém os valores dos parâmetros que foram considerados para o algoritmo genético proposto neste trabalho.

Tabela 3 – Valor dos parâmetros utilizados no algoritmo genético.

Parâmetro	Valor
Tamanho da População	50
Número Máximo de Gerações	200
Número Máximo de Iterações sem Melhora	10
Probabilidade de Cruzamento	0.50
Probabilidade de Mutação	0.01

A função retorna os seguintes valores: o tempo de processamento do algoritmo, o menor valor de X, o maior valor de X, a média de X, o coeficiente de assimetria da população, os pontos de corte dos estratos, a variância total, a variância por estrato, o coeficiente de variação total, o coeficiente de variação por estrato, o número de elementos da população em cada estrato, o tamanho de amostra em cada estrato considerando a alocação de Neyman, a medida de homogeneidade R^2 e o total associado à variável Y.

O arquivo de entrada de dados deve ser do tipo texto, com as observações das n colunas separadas por ";". Os valores em cada uma das colunas correspondem às observações de X e/ou Y que serão utilizadas nos cálculos.

5.1. Comparação entre os algoritmos

Esta seção contém um conjunto de resultados obtidos pelo AG proposto e pelos algoritmos de Hedlin (1998, 2000) e Gunning e Horgan (2004) considerando as três populações descritas a seguir.

A primeira população está associada a um conjunto de estabelecimentos agropecuários produtores de café, com efetivo de pés de café acima de 1000 *ha*, do Estado do Paraná, pesquisados no Censo Agropecuário 1995-1996 (Azevedo, 2004). As duas outras populações foram obtidas a partir dos dados da PAM (Produção Agrícola Municipal 1990-2004), uma pesquisa produzida pelo IBGE. Os dados desta última pesquisa podem ser obtidos na Internet acessando o banco de dados SIDRA (http://www.sidra.ibge.gov.br/bda/agric/).

Nas três populações, a variável X foi usada tanto para a estratificação quanto para o cálculo da variância (ou seja, usou-se $X = Y$).

No caso da pesquisa de produção de café, foram consideradas duas possíveis variáveis X de estratificação: o efetivo (população 1^A) e o número de pés colhidos (população 1^B). Tais variáveis correspondem, respectivamente, à quantidade de pés de café novos mais à quantidade de pés de café em idade produtiva e à quantidade de pés de café que deram origem à colheita.

Com relação à pesquisa de Produção Agrícola Municipal - PAM, foram escolhidas as Unidades da Federação de São Paulo e Minas Gerais. Nestas unidades foi selecionada, como variável X de estratificação, área total colhida por município, considerando a cultura de milho (populações 2 e 3, associadas respectivamente a SP e MG).

Na Tabela 4, são informadas algumas características das distribuições correspondentes às observações da variável X considerada em cada população, quais sejam, o total populacional, o coeficiente de assimetria e os valores de mínimo, de máximo, da média e da variância populacional.

Tabela 4 – Informações gerais sobre as populações.

População	N	Coeficiente de Assimetria	Mínimo	Máximo	Média	Variância
1^A (Efetivo)	20472	19.7	1000	1321200	9170.0	929399973.0
1^B (Pés colhidos)	20472	28.8	0	1098500	4945.0	424483609.0
2 (Milho-SP)	586	4.9	3	30385	1832.1	11016631.8
3 (Milho-MG)	845	7.5	1	46000	1561.4	10145561.0

As Tabelas 5, 6 e 7 contêm os resultados obtidos pelos três algoritmos (Hedlin, Gunning e Horgan e AG) na estratificação das populações 1^A , 2 e 3. São apresentados, para cada população, os coeficientes de variação, os pontos de corte b_h, o tamanho da população nos estratos (N_h) e os tamanhos de amostra nos estratos (n_h) obtidos por cada algoritmo. Considerou-se, para cada população, o tamanho da amostra igual a 100 e o número de estratos variando entre 3 e 5.

Pode-se verificar, através dessas tabelas, que o algoritmo genético encontrou, em todos os casos, coeficientes de variação menores ou ao menos iguais àqueles obtidos pelos algoritmos de Hedlin e de Gunning e Horgan, o que indica que as soluções encontradas pelo AG corresponderam, em geral, a estratos mais homogêneos segundo a variável X.

Tabela 5 – Pontos de corte nos estratos com $L=3$ e $n=100$.

População	Algoritmo	CV		Estratos 1	2	3
1[A]	Genético	6.7	b_h	13500	98800	
			N_h	17963	2326	183
			n_h	41	31	28
	Gunning e	7.0	b_h	10973	120405	
	Horgan		N_h	17100	3239	133
			n_h	32	47	21
	Hedlin	9.1	b_h	23000	385010	
			N_h	19233	1215	24
			n_h	42	34	24
2	Genético	3.2	b_h	1333	4800	
			N_h	383	153	50
			n_h	26	24	50
	Gunning e	7.7	b_h	65	1404	
	Horgan		N_h	35	352	199
			n_h	1	12	87
	Hedlin	3.2	b_h	1300	4475	
			N_h	383	153	50
			n_h	26	24	50
3	Genético	4.1	b_h	1350	6700	
			N_h	588	226	31
			n_h	28	41	31
	Gunning e	9.3	b_h	36	1284	
	Horgan		N_h	27	553	265
			n_h	1	12	87
	Hedlin	7.5	b_h	1400	5500	
			N_h	598	204	43
			n_h	28	29	43

Tabela 6 – Pontos de corte nos estratos com $L=4$ e $n=100$.

População	Algoritmo	CV		Estratos 1	2	3	4
1[A]	Genético	4.9	b_h	6700	27000	143000	
			N_h	14297	5137	934	104
			n_h	23	26	26	25
	Gunning e	5.1	b_h	6029	36348	219142	
	Horgan		N_h	14013	5765	628	66
			n_h	21	39	24	16
	Hedlin	6.0	b_h	9350	50000	430000	
			N_h	16323	3732	399	18
			n_h	28	29	25	18
2	Genético	2.2	b_h	682	1880	4200	
			N_h	265	164	101	56
			n_h	14	14	16	56
	Gunning e	5.1	b_h	30	302	3029	
	Horgan		N_h	15	138	341	92
			n_h	1	2	30	67
	Hedlin	2.2	b_h	760	2051	4475	
			N_h	285	164	87	50
			n_h	17	18	15	50
			n_h	1	1	37	61
3	Genético	2.5	b_h	763	2200	8200	
			N_h	476	201	147	21
			n_h	20	16	43	21
	Gunning e	6.6	b_h	15	214	3141	
	Horgan		N_h	11	186	543	105
			n_h	1	1	37	61
	Hedlin	4.1	b_h	700	2220	5200	
			N_h	476	207	115	47
			n_h	19	16	18	47

Tabela 7 - Pontos de corte nos estratos com $L=5$ e $n=100$.

População	Algoritmo	CV		Estratos				
				1	2	3	4	5
1[A]	Genético	3.9	b_h	5650	16500	52000	199888	
			N_h	12994	5600	1468	334	76
			n_h	22	21	18	15	24
	Gunning e Horgan	4.1	b_h	4209	17717	74573	313888	
			N_h	10895	7816	1497	230	34
			n_h	14	32	25	18	11
	Hedlin	4.4	b_h	6000	20000	86000	145000	
			N_h	14012	5060	1188	196	16
			n_h	23	20	21	20	16
2	Genético	1.5	b_h	550	1388	2900	9322	
			N_h	232	153	104	78	19
			n_h	12	11	14	44	19
	Gunning e Horgan	3.4	b_h	19	120	759	4803	
			N_h	5	65	214	253	49
			n_h	1	1	6	43	49
	Hedlin	1.6	b_h	500	1200	2420	4212	
			N_h	225	150	94	63	54
			n_h	13	11	12	10	54
3	Genético	1.6	4	410	1170	2700	9300	
			N_h	323	239	142	123	18
			n_h	9	12	14	47	18
	Gunning e Horgan	4.7	b_h	9	73	628	5373	
			N_h	7	56	375	360	47
			n_h	1	1	7	47	44
	Hedlin	2.7	b_h	510	1450	3000	5800	
			N_h	388	215	134	69	39
			n_h	15	14	18	14	39

Na Tabela 8, temos o número de estratos, os tamanhos de amostra e os coeficientes de variação obtidos a partir da aplicação do algoritmo de Hedlin e do algoritmo genético, considerando a população 1[B] (o algoritmo de Gunning e Horgan não foi usado neste experimento, pois não é aplicado quando há valores $X=0$ na população, Tabela 4). O tamanho de amostra variou entre 734 e 186, e o número de estratos entre 3 e 8.

Cabe observar que a adoção dos mesmos tamanhos de amostra, bem como do número de estratos apresentados na Tabela 8, teve por finalidade possibilitar a comparação com os resultados apresentados na dissertação de Azevedo (2004). Neste trabalho, na implementação do algoritmo desenvolvido por Hedlin, considerou-se a

variável de estratificação efetivo para a obtenção dos pontos de corte, mediante o tamanho total da amostra estabelecido e conseqüentemente, os tamanhos de amostra alocados em cada um dos estratos.

A partir destes pontos de corte e dos tamanhos de amostra, foram calculados os coeficientes de variação para a variável efetivo e para um conjunto de sete variáveis também consideradas na pesquisa do café, em particular, para a variável pés colhidos.

De igual modo, aplicou-se o algoritmo genético (AG) considerando a variável de estratificação efetiva e os mesmos tamanhos de amostra e número de estratos utilizados no trabalho de Azevedo (2004). Em seguida, utilizando os pontos de corte e os tamanhos de amostra produzidos pelo AG, para esta variável, foram calculados os coeficientes de variação da variável pés colhidos.

Pode-se verificar, nessa tabela, que o AG novamente encontrou as soluções mais homogêneas, considerando-se o critério de menor coeficiente de variação relativo ao total estimado da variável X.

Tabela 8 – Coeficientes de variação - população 1[8].

Algoritmo	Estratos	Tamanhos de Amostra	Coeficientes de Variação
Genético	3	734	3.5
	4	439	2.9
	5	327	3.8
	6	242	3.8
	7	198	3.7
	8	186	3.2
Hedlin	3	734	3.6
	4	439	4.4
	5	327	4.8
	6	242	5.6
	7	198	6.2
	8	186	6.2

Para se ter uma avaliação quantitativa da qualidade das soluções obtidas, efetuou-se o cálculo da razão (eficiência relativa) entre as variâncias da estimativa obtidas mediante a aplicação dos algoritmos de Hedlin e de Gunning e Horgan e aquelas obtidas a partir

do algoritmo genético. Nas equações, a seguir, temos as expressões que definem a eficiência relativa em função do novo algoritmo:

$$eff_{Gunning,Genetico} = \frac{\hat{V}_{Gunning}(X)}{\hat{V}_{Genetico}(X)}, \quad eff_{Hedlin,Genetico} = \frac{\hat{V}_{Hedlin}(X)}{\hat{V}_{Genetico}(X)}.$$

Os resultados obtidos pela aplicação destas expressões são apresentados nas Tabelas 9, 10 e 11 e foram obtidos a partir das informações das Tabelas 5, 6, 7 e 8. A partir destas tabelas, é possível verificar que o algoritmo genético foi, em geral, mais eficiente do que os outros dois algoritmos, destacando-se principalmente quando comparado com o algoritmo de Gunning e Horgan.

Tabela 9 – Eficiência do algoritmo Gunning e Horgan em relação ao algoritmo genético

Estratos	População		
	1[A]	2	3
3	1.1	5.1	5.8
4	1.1	5.4	7.0
5	1.1	5.1	8.6

Tabela 10 – Eficiência do algoritmo Hedlin em relação ao algoritmo genético

Estratos	População		
	1[A]	2	3
3	1.8	1.0	3.3
4	1.5	1.0	2.7
5	1.3	1.1	2.8

Tabela 11 – Eficiência do algoritmo Hedlin em relação ao algoritmo genético

Estratos	População 1[B]
3	1.1
4	2.3
5	1.6
6	2.2
7	2.8
8	3.8

5.2. Considerações Finais

A partir dos resultados da seção anterior, pode-se verificar que o algoritmo genético forneceu soluções de qualidade igual ou superior àquelas obtidas pelos algoritmos de Hedlin e de Gunning e Horgan, mostrando ser ao menos promissora a abordagem heurística utilizada neste trabalho.

Os autores acreditam, entretanto, que novos desenvolvimentos do AG apresentado podem ser feitos, considerando algumas modificações nos operadores de reprodução, mutação e cruzamento (Glover e Kochenberger, 2003) ou na própria representação dos cromossomos.

Além disso, outras técnicas heurísticas, tais como: VNS (*Variable Neighborhood Search*) e GRASP (*Greedy Randomized Adaptive Search Procedure*) (Glover e Kochenberger, 2003), também podem ser aplicadas ao problema, resultando em soluções de qualidade ainda melhor (coeficientes de variação menores). Observamos, ainda, que a proposta de aplicação de algoritmos genéticos (AG) para este particular problema é inovadora, não tendo sido encontrado, ao longo de nossa pesquisa bibliográfica, nenhum outro trabalho que utiliza-se o A.G.

Entretanto, no trabalho de Day C.D. (2006) é feita uma proposta de um algoritmo genético aplicado ao problema de alocação ótima multivariada (Bethel, 1989) em amostras estratificadas. Neste problema, considera-se a necessidade de produzir estimativas com nível precisão razoável (fixo) para um conjunto de variáveis de interesse, mediante a uma boa alocação do tamanho de amostra em cada um dos estratos, produzindo o menor custo de alocação possível.

No problema de alocação, define-se, previamente, o número de estratos, o tamanho de amostra, os limites nos estratos e as variâncias associadas (e coeficientes de variação) a cada uma das variáveis de interesse.

O algoritmo genético proposto para a resolução de tal problema tem a população formada por um conjunto de m cromossomos, cada um, com L genes (sendo L o número de estratos). A cada uma das L posições dos cromossomos é atribuído um

tamanho de amostra n_h (no estrato) de forma que $\sum_{h=1}^{L} n_h = n$. A partir destes cromossomos, o algoritmo avalia uma função objetivo, que agrega, basicamente, os tamanhos de amostra e os coeficientes de variação associados a cada uma das variáveis de interesse.

A seguir, são apresentadas de forma resumida as principais características dos dois algoritmos considerando sua estrutura de representação, as operações e a execução.

Tabela 12 – Principais características dos algoritmos

Algoritmo Genético para Estratificação	Algoritmo Genético para Alocação
Cada cromossomo contém $L-1$ posições (genes).	Cada cromossomo contém L posições (genes).
Cada cromossomo contém os pontos de corte para definição dos estratos.	Cada cromossomo contém os tamanhos de amostra que serão alocados aos estratos.
A função objetivo a ser minimizada é variância do estimador de total.	A função objetivo a ser minimizada agrega os tamanhos de amostra e os coeficientes de variação.
No processo seleção dos cromossomos, é utilizado o método da roleta.	No processo seleção dos cromossomos, é utilizado o método da roleta.
O algoritmo converge para soluções de boa qualidade, necessitando de poucas iterações (gerações).	O algoritmo converge para soluções de boa qualidade necessitando de poucas iterações (gerações).
O operador de mutação é aplicado em um gene e produz a mudança de um ponto de corte.	O operador de mutação é aplicado em um gene e produz a mudança no tamanho de amostra alocado a um dos estratos.
O operador de cruzamento é aplicado em apenas um segmento dos cromossomos, alterando até $L-2$ pontos de corte.	O operador de cruzamento é aplicado em um segmento dos cromossomos. É selecionado um gene na posição k $(1 \le k < L)$, e, em seguida, todos os tamanhos de amostra alocados da posição $(k+1)$ até L são recalculados.

Pela Tabela 12, é possível observar que os dois algoritmos têm apenas algumas similaridades, fato naturalmente esperado, tendo em vista que foram desenvolvidos para problemas com características diferentes.

Todavia, em trabalhos futuros, pode-se tentar uma combinação destes dois algoritmos, implementando um algoritmo genético com duas fases, sendo 1ª fase responsável pela definição dos pontos de corte e a 2ª fase responsável pela alocação dos tamanhos de amostra nos estratos.

Referências bibliográficas

Azevedo, R. V. (2004). Estudo Comparativo de Métodos de Estratificação Ótima de Populações Assimétricas. Dissertação de Mestrado. IBGE/ENCE.

Brito, J. A.M. (2005). Uma Formulação de Programação Inteira para o Problema de Alocação Ótima em Amostras Estratificadas. *Anais do XXXVI Simpósio de Pesquisa Operacional* – SOBRAPO. Gramado, RS.

Bankier, M.D. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, v. 42, p. 174-177.

Bethel, J. (1989). Sample Allocation in Multivariate Surveys. Survey Methodology, 15, 47-57.

Bolfarine, H. e Bussab, Wilton O. (2005). Elementos de Amostragem. ABE – Projeto Fisher. Editora Edgard Blücher,

Cochran, Willian G. (1977). Sampling Techniques. Third Edition – New York, John Wiley.

Dalenius, T. (1952). The Problem of Optimum Stratification in Special Type of Design. Skandinavisk Aktuarietidskrift, 35, 61-70.

Dalenius, T. and Hodges, J. L. Jr. (1959). Minimum variance stratification. Skandinavisk Aktuarietidskrift, 54, 88-101.

Day C.D. (2006). Application of an Evolutionary Algorithm to Multivariate Optimal Allocation in Stratified Sample Designs. American Statistical Assocations Conference. Internal Revenue Service, Statistics of Income Division.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*, Boston, MA: Addison-Wesley.

Glover F. and Kochenberger G.A. (2003). *Handbook of Metaheuristics*. Kluwer Academic Publishers.

Gunning P. and Horgan, Jane M. (2004). A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. Survey Methodology, 30 (2), 159-166.

Hedlin, D. (1998). On the stratification of highly skewed populations. R&D Report. *Statistics Sweden*, Sweden.

Hedlin, D. (2000). A Procedure for Stratification by an Extended Ekman Rule. Journal of Official Statistics, 16 (1), 15-29.

Holland, J. H. (1975). Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor.

Lavallée P. and Hidiroglou M. A. (1988). On the Stratification of Skewed Populations. Survey Methodology (Statistics Canada), 14 (1), 33-43.

Resende, M. G. C. e Sousa, J. P. (2004). *Metaheuristics : Computer Decision-Making*. Kluwer Academic Publishers.

Viana, G. V. R. (1998). Meta-Heurísticas e Programação Paralela em Otimização Combinatória. Editora UFC.

Sigman, R. e Monsour, R. S. (1995). Selecting Samples from List Frames of Business. In: COX, B. G. et al. (Ed.). Business Survey Methods, New York: Wiley, 133-152.

Abstract

This paper reports a new algorithm to solve the stratification problem. Given a population of size N, a sample size n and a number of strata L, the stratification problem is to determine which observations of the population must be associated to what stratum in order to minimize the sum of the variances of the estimators in each stratum. To solve this problem, we propose a heuristic method based in genetic algorithms. A set of computational results is presented and used in performance comparisons with other methods in the literature.

Keywords: Stratification, sampling and genetic algorithms.

Fatores de Bayes para discriminação de modelos separados de regressão multivariada usando *priores* impróprias

Maria Ivanilde Araujo[*]
Basílio de Bragança Pereira[**]

Resumo

Para discriminar entre modelos alternativos, em inferência Bayesiana, fatores de Bayes alternativos foram desenvolvidos para superar as dificuldades com o uso de *priores* impróprias para os parâmetros dos modelos. Neste artigo, estendemos estes resultados para sistemas de equações de regressão separadas (não encaixadas) e apresentamos duas aplicações.

Palavras-chave: Discriminação, fator de Bayes, fator de Bayes intrínseco, fator de Bayes fracionário, fator de Bayes *a posteriori*, regressão multivariada.

[*] Endereços para correspondência: Departamento de Estatística – UFAM – *e-mail*: miaraujo@ufam.edu.br.
[**] Faculdade de Medicina, COPPE e HUCFF – UFRJ – Caixa Postal: 68507, CEP.: 21945-970 - Rio – RJ, Brasil – *e-mail*: basilio@ufrj.br.

1. Introdução

Em qualquer área do conhecimento científico, pesquisadores são freqüentemente levados a escolher entre modelos estatísticos alternativos. Surgem, então, as seguintes questões:

- Existe alguma evidência significativa de que os modelos forneçam diferentes ajustes para os dados?
- Assumindo que um dos modelos é verdadeiro, qual é a evidência fornecida pelos dados de qual é o modelo verdadeiro?
- Se um modelo representa a hipótese atualmente mantida, existe alguma evidência de não adequação deste modelo na direção de outro?

A teoria de teste de hipóteses de Neyman-Pearson aplica-se aos modelos pertencentes à mesma família de distribuições. Alternativamente, procedimentos especiais foram desenvolvidos por Cox (1961, 1962) no caso dos modelos pertencerem a famílias separadas ou não-encaixadas, no sentido de que um membro arbitrário de uma família não pode ser obtido como limite dos membros de outra. Cox (1961) sugeriu os seguintes enfoques alternativos: modificação da razão de Neyman-Pearson, mistura exponencial e linear dos modelos, verossimilhança, decisão Bayesiana, e fator de Bayes assintótico, ver Araújo *et al.* (2005) para detalhes destes desenvolvimentos.

Uma grande variedade de artigos discutindo famílias de hipóteses separadas surgiu desde o trabalho fundamental de Cox (1961, 1962). Para *reviews* e referências, Pereira (1977, 1981, 2005), Gourieroux e Monfort (1994), McAller (1995), e Pesaran e Weeks (2001).

A análise Bayesiana também encontra dificuldades ao usar os fatores de Bayes para discriminar modelos separados. Primeiro, *a priori* para um modelo e *a priori* para os seus parâmetros devem ser coerentes com *as priores* para o outro modelo. Se os espaços de parâmetros possuem dimensões diferentes e não existe relação simples entre os parâmetros, o problema não é simples. Segundo, quando a informação *a priori* é fraca e *priores* impróprias são usadas também, existe dificuldades e paradoxos com o

uso do fator de Bayes que torna-se indefinido. Fatores de Bayes alternativos são necessários.

A seção 2 discute algumas propostas alternativas para superar as dificuldades com o fator de Bayes. A seção 3 estende estes resultados para modelos separados de regressão multivariada. Dois exemplos são apresentados na seção 4.

2. Fatores de Bayes alternativos

A sugestão de Cox (1961) utiliza o procedimento de inferência Bayesiana, razão *a posteriori* para H_0 *versus* H_1, para o problema geral de famílias separadas de hipóteses.

Seja $y=(y_1,\ldots,y_n)$ uma variável aleatória de uma distribuição desconhecida. A hipótese nula H_0 e a hipótese alternativa H_1, respectivamente, especificam as densidades paramétricas $f_0(y|a_0)$ e $f_1(y|a_1)$ para o vetor aleatório y, onde a_0 e a_1 são vetores de parâmetros desconhecidos.

A razão *a posteriori* para H_0 *versus* H_1 é

$$\frac{\pi_0}{\pi_1}\frac{q_0(y)}{q_1(y)}=\frac{\pi_0}{\pi_1}B_{01}(y), \tag{1}$$

onde π_j é a probabilidade *a priori* de H_j e

$$q_j(y)=\int f_j(y|\alpha_j)\pi_j(\alpha_j)d\alpha_j \tag{2}$$

denota a distribuição preditiva com probabilidade *a priori* $\pi_j(a_j)$ para os parâmetros sob $H_j(j=0,1)$. O fator de Bayes, $B_{01}(y)$, representa o peso da evidência nos dados favorecendo H_0 sobre H_1. Cox também fornece uma expressão geral considerando a função perda na razão *a posteriori*, e descreve a aproximação para grandes amostras para distribuição do fator de Bayes.

Esta abordagem possui duas grandes limitações. Primeiro, o conhecimento *a priori* dado por π_0 e $\pi_0(\alpha_0)$ deve ser coerente com aquele dado por π_1 e $\pi_1(\alpha_1)$. Se os espaços de parâmetros têm diferentes dimensões, por exemplo, não existe relações simples entre os parâmetros. Segundo, se a informação *a priori* é fraca e *priores* impróprias são usadas, o fator de Bayes usual não está bem definido (AITKIN, 1991; O`HAGAN, 1995). Por exemplo, caso se use *priores* impróprias para os parâmetros de cada modelo proporcional a constantes, C_0 e C_1 e *priores* iguais a *1/2* para cada modelo, o fator de Bayes usual depende da razão C_0/C_1 e, portanto, não está especificado.

Os fatores de Bayes alternativos permitem o uso de *priores* impróprias para os parâmetros e também podem ser interpretados usando a regra de Jeffreys (KASS e RAFTERY, 1995). Estas regras serão utilizadas nos exemplos deste artigo.

Tabela 1 - Regra de Jeffreys para os fatores de Bayes

$2\ln B_{01}(y)$	$B_{01}(y)$	Evidência contra H_1
0 a 2	1 a 3	Fraca
2 a 6	3 a 20	Positiva
6 a 10	20 a 150	Forte
> 10	> 150	Muito Forte

Na seqüência, discutem-se os principais fatores de Bayes alternativos que podem ser aplicados no lugar do fator de Bayes usual.

2.1. Fator de Bayes *a posteriori*

Aitkin (1991) propôs o fator de Bayes *a posteriori* que compara as médias *a posteriori* das funções de verossimilhança sob H_0 e H_1. Mais formalmente, a densidade *a posteriori* sob $H_j (j=0,1)$ é

$$\pi_j(\alpha_j|y) \equiv \frac{f_j(y|\alpha_j)\pi_j(\alpha_j)}{\int f_j(y|\alpha_j)\pi_j(\alpha_j)d\alpha_j}, \qquad (3)$$

o que resulta em

$$q_j^P(y) = \int f_j(y|\alpha_j)\pi_j(\alpha_j)d\alpha_j \qquad (4)$$

como a média *a posteriori* da função de verossimilhança sob $H_j(j=0,1)$. Se $\pi_j(\alpha_j)=C_j h_j(\alpha_j)$ em (3), então as constantes C_j cancelam-se. O fator de *Bayes a posteriori* corresponde então à razão das médias *a posteriori*, isto é, $B_{01}^{\ P}(y)=q_0^{\ P}(y)/q_1^{\ P}(y)$.

2.2. Fator de Bayes parcial

O`Hagan (1995) derivou uma alternativa ao fator de Bayes *a posteriori* de Aitkin (1991) para evitar problemas com *priores* impróprias. Considere a partição $y=(y_1,y_2)$ da amostra. Da subamostra y_1, obtém-se densidades *a posterioris* próprias $\pi_0(\alpha_0|y_1)$ e $\pi_1(\alpha_1|y_1)$ para usá-las como *priores* para a subamostra y_2. O fator de Bayes parcial $B_{01}(y_2|y_1)$ é então

$$B_{01}(y_2|y_1) \equiv \frac{\int f_0(y_2|y_1,\alpha_0)\pi_0(\alpha_0|y_1)\,d\alpha_0}{\int f_1(y_2|y_1,\alpha_1)\pi_1(\alpha_1|y_1)\,d\alpha_1} = \frac{q_0(y)/q_0(y_1)}{q_1(y)/q_1(y_1)} = \frac{B_{01}(y)}{B_{01}(y_1)}. \qquad (5)$$

A idéia é que *as priores* impróprias afetam $B_{01}(y)$ e $B_{01}(y_1)$ da mesma maneira e assim seus efeitos cancelam-se.

2.3. Fator de Bayes intrínseco

Berger e Pericchi (1996) definiram que a amostra de treinamento y_1 na amostra particionada $y = (y_1, y_2)$ é mínima se *as posterioris* para α_0 e α_1 são próprias e não existe subconjunto de y_1 que leve a uma *posteriori* própria. Existem usualmente muitas, digamos R, partições caracterizando uma amostra de treinamento minimal. A idéia do fator de Bayes intrínseco, $B_{01}^{I}(y)$, é tomar a média ou a mediana dos fatores de Bayes parciais $\{B_{01}(y_{2r}|y_{1r}); r = 1, ..., R\}$ obtidos de todas as R amostras de treinamento minimais. Se R é muito grande, pode-se selecionar uma amostra aleatória de todas as possíveis

amostras de treinamento. O fator de Bayes intrínseco é para a média aritmética e para a média geométrica, respectivamente, assim definido:

$$B_{01}^{IA}(y) = \frac{1}{R}\sum_{r=1}^{R} B_{01}(y_2 r|y_1 r) \tag{6}$$

$$B_{01}^{IG}(y) = \left[\prod_{r=1}^{R} B_{01}(y_2 r|y_1 r)\right]^{1/R}, \tag{7}$$

enquanto o fator de Bayes intrínseco para a mediana é

$$B_{01}^{IM}(y) = \text{med}\{B_{01}(y_2 r|y_1 r); r = 1,\ldots R\}. \tag{8}$$

A versão geométrica é preferível porque é a única versão na qual a desejável propriedade de reverter os índices resulta em $B_{0l}(y) = 1/B_{0l}^{IG}(y)$. Para as outras versões, tem-se que impor isto como definição (BERGER e PERICCHI, 1996b).

2.4. Fator de Bayes fracionário

O'Hagan (1995) propôs uma modificação do fator de Bayes que também utiliza amostras de treinamento. Seja $b = n_{yl}/n$ denotando a fração de treinamento, onde n_{yl} é o tamanho da amostra de treinamento y_l. O fator de Bayes fracionário é então $B_{0l}^{[b]}(y) = q_0^{[b]}(y)/q_l^{[b]}(y)$, com

$$q_j^{[b]}(y) = \frac{\int f_j(y|\alpha_j)\pi_j(\alpha_j)\,d\alpha_j}{\int [f_j(y|\alpha_j)]^b \pi_j(\alpha_j)\,d\alpha_j} \tag{9}$$

para $j = 0, 1$. Se $\pi_j(\alpha_j) = C_j h_j(\alpha_j)$, a constante cancela-se em $q_j^{[b]}(y)$. Para entender o fator de Bayes fracionário, observe que, para n_{yl}, suficientemente grande, a verossimilhança

para a amostra completa é aproximadamente igual a verossimilhança para a amostra de treinamento elevada a b. Portanto, consistente, dado que b tende a zero quando n cresce.

3. Regressão linear multivariada

3.1. Algumas definições

Considere dois modelos de regressão multivariados H_0: $\mathbf{Y} = \mathbf{XB}_0 + \mathbf{U}_0$ e H_1: $\mathbf{Y} = \mathbf{ZB}_1 + \mathbf{U}_1$, onde $\mathbf{Y}$ é uma matriz $n \times m$ de respostas, $\mathbf{X}$ e $\mathbf{Z}$ são matrizes de regressores $n \times p$ e $n \times q$, respectivamente, e $\mathbf{B}_0$ e $\mathbf{B}_1$ são matrizes de parâmetros $p \times m$ e $q \times m$, respectivamente. Os termos de erros $\mathbf{U}_0$ e $\mathbf{U}_1$ possuem linhas independentes e identicamente distribuídas como vetores aleatórios normais com média zero e matrizes de covariância Σ_0 e Σ_1, respectivamente. Também assumimos que $\mathbf{X}$ e $\mathbf{Z}$ são de postos completos com $n \geq m + p$ e $n \geq m + q$. Segue-se então que $\mathbf{U}_0 \sim \mathrm{N}(0, I_n \otimes \Sigma_0)$ e $\mathbf{U}_1 \sim \mathrm{N}(0, I_n \otimes \Sigma_1)$, enquanto $\mathbf{Y} \sim \mathrm{N}(\mathbf{XB}_0, I_n \otimes \Sigma_0)$ sob H_0 e $\mathbf{Y} \sim \mathrm{N}(\mathbf{ZB}_1, I_n \otimes \Sigma_1)$ sob H_1, onde $\otimes$ é o produto de Kronecker (ver RAO e RAO ,1998).

As matrizes de regressores $\mathbf{X}$ e $\mathbf{Z}$ são fixas e não encaixadas no sentido de que não é possível obter as colunas de $\mathbf{X}$ a partir das colunas de $\mathbf{Z}$ e vice-versa. Assumiremos mais adiante que as matrizes $\Sigma_{X'X} = lim_{n\to\infty}(1/n)X'X$ e $\Sigma_{Z'Z} = lim_{n\to\infty}(1/n)Z'Z$ são não singulares, e que $\Sigma_{X'Z} = lim_{n\to\infty}(1/n)X'Z$ é uma matriz não nula. As suposições acima asseguram que o estimador de máxima verossimilhança $\hat{\beta}_0 = (X'X)^{-1}X'Y$ e $\hat{\beta}_1 = (Z'Z)^{-1}Z'Y$ são consistentes sob H_0 e H_1, respectivamente.

3.2. Fator de Bayes

Nesta seção, estendem-se os resultados de Aitkin (1991), O`Hagan (1995) e Berger e Pericchi (1996) para o contexto de regressão linear multivariada. A *razão a posteriori* para H_0 contra H_1 é $(\pi_0/\pi_1)B_{01}$. Suponha o uso de *priores* impróprias para os parâmetros tais que $\pi_0(\alpha_0)$ e $\pi_1(\alpha_1)$ sejam respectivamente proporcional as constantes K_0 e K_1. O fator de Bayes B_{01} é, então, proporcional a K_0/K_1 não estando definido. Para o modelo de regressão multivariada *a priori* difusa de Jeffreys é dada por

$$\pi_j(\alpha_j)=\pi_j(B_j)\pi_j(\Sigma_j)=K_j\left|\Sigma_j\right|^{-\frac{m+1}{2}},j=0,1 \tag{10}$$

dando a seguinte distribuição preditiva sob a hipótese nula

$$q_0(Y)=\pi^{\frac{m(2n-2p-m+1)}{4}}K_0\left|X'X\right|^{-m/2}\left|S_0\right|^{-\frac{n-p}{2}}\prod_{s=1}^{m}\Gamma\left(\frac{n-p-s+1}{2}\right), \tag{11}$$

onde $S_0\equiv(Y-X\hat{B}_0)'(Y-X\hat{B}_0)$. Uma expressão similar ocorre para o modelo alternativo $Y=ZB_1+U_1$. O fator de Bayes resultante é, então,

$$B_{01}(Y)=\pi^{m(p-q)/2}\frac{K_0}{K_1}\left(\frac{\left|Z'Z\right|}{\left|X'X\right|}\right)^{m/2}\frac{\left|S_1\right|^{(n-q)/2}}{\left|S_0\right|^{(n-p)/2}}\prod_{s=1}^{m}\frac{\Gamma\left(\frac{n-p-s-1}{2}\right)}{\Gamma\left(\frac{n-q-s-1}{2}\right)}, \tag{12}$$

onde $S_1\equiv(Y-Z\hat{B}_1)'(Y-Z\hat{B}_1)$. Vê-se de (12) que o fator de Bayes não está bem definido uma vez que depende da razão desconhecida K_0/K_1.

De (11) e (12), é possível obter o fator de Bayes alternativo. Por exemplo, o fator de Bayes *a posteriori*, $B_{01}(y)\,B_{01}^{P}(Y)$, de Aitkin (1991) resulta da razão entre

$$q_0^P(Y) = (2\sqrt{\pi})^{-mn} |S_0|^{-n/2} \prod_{s=1}^{m} \frac{\Gamma\left(\frac{2n-p-s+1}{2}\right)}{\Gamma\left(\frac{n-p-s+1}{2}\right)}$$

e

$$q_1^P(Y) = (2\sqrt{\pi})^{-mn} |S_1|^{-n/2} \prod_{s=1}^{m} \frac{\Gamma\left(\frac{2n-q-s+1}{2}\right)}{\Gamma\left(\frac{n-q-s+1}{2}\right)}.$$

resultando em

$$B_{01}^P(Y) = \left(\frac{|S_1|}{|S_0|}\right)^{n/2} \prod_{s=1}^{m} \frac{\Gamma\left(\frac{2n-p-s+1}{2}\right)\Gamma\left(\frac{n-q-s+1}{2}\right)}{\Gamma\left(\frac{2n-q-s+1}{2}\right)\Gamma\left(\frac{n-p-s+1}{2}\right)}. \tag{13}$$

O fator de Bayes intrínseco de BERGER e PERICCHI (1996) é, então,

$$B_{01}^{IA}(Y) = \frac{1}{R}\sum_{r=1}^{R} \frac{B_{01}(Y)}{B_{01}(Y_{(r)})} = B_{01}(Y)\frac{1}{R}\sum_{r=1}^{R} B_{10}(Y_{(r)})$$

onde $Y_{(r)}$ é a amostra de treinamento minimal com matrizes $X_{(r)}$ e $Z_{(r)}$ sob H_0 e H_1, respectivamente. Por definição, $Y_{(r)}$ é uma matriz tal que ambos $X_{(r)}'X_{(r)}$ e $Z_{(r)}'Z_{(r)}$ são não singulares. Ela tem dimensão $\bar{n} \times m$, onde $\bar{n} = \lceil (m+1)/2 \rceil + \max(p,q)$ e $\lceil \cdot \rceil$ retorna o menor inteiro maior do que seu argumento. De (12), segue-se que

$$\begin{aligned} B_{01}^{IA}(Y) = & \left(\frac{|Z'Z|}{|X'X|}\right)^{m/2} \frac{|S_1|^{\frac{n-q}{2}}}{|S_0|^{\frac{n-p}{2}}} \prod_{s=1}^{m} \frac{\Gamma\left(\frac{n-p-s+1}{2}\right)\Gamma\left(\frac{\bar{n}-q-s+1}{2}\right)}{\Gamma\left(\frac{n-q-s+1}{2}\right)\Gamma\left(\frac{\bar{n}-p-s+1}{2}\right)} \\ & \times \frac{1}{R}\sum_{r=1}^{R} \left(\frac{|X_{(r)}'X_{(r)}|}{|Z_{(r)}'Z_{(r)}|}\right)^{m/2} \frac{|S_{0(r)}|^{(\bar{n}-p)/2}}{|S_{1(r)}|^{(\bar{n}-q)/2}}, \end{aligned} \tag{14}$$

onde $S_{j(r)}$ é análogo a S_j para a r-ésima amostra de treinamento ($j=0,1$).

Para o fator de Bayes geométrico correspondente, simplesmente substitua a média aritmética pela média geométrica.

Finalmente, o fator de Bayes fracionário de O´Hagan (1995) resulta da razão entre

$$q_0^{[b]}(Y) = \pi^{mn(1-b)/2} b^{mnb/2} |S_0|^{-n(1-b)/2} \prod_{s=1}^{m} \frac{\Gamma\left(\frac{n-p-s+1}{2}\right)}{\Gamma\left(\frac{nb-p-s+1}{2}\right)}$$

e

$$q_1^{[b]}(Y) = \pi^{mn(1-b)/2} b^{mnb/2} |S_1|^{-n(1-b)/2} \prod_{s=1}^{m} \frac{\Gamma\left(\frac{n-q-s+1}{2}\right)}{\Gamma\left(\frac{nb-q-s+1}{2}\right)}. \quad (15)$$

O que implica em

$$B_{01}^{[b])}(Y) = \left(\frac{|S_1|}{|S_0|}\right)^{n(1-b)/2} \prod_{s=1}^{m} \frac{\Gamma\left(\frac{n-p-s+1}{2}\right)\Gamma\left(\frac{nb-q-s+1}{2}\right)}{\Gamma\left(\frac{n-q-s+1}{2}\right)\Gamma\left(\frac{nb-p-s+1}{2}\right)}. \quad (16)$$

4. Exemplos de Aplicação

Em ambos os exemplos, utilizou-se mínimos quadrados ordinários, tendo em vista que nos sistemas de equações de cada exemplo temos as mesmas variáveis exógenas em ambas as equações, e neste caso, o estimador de mínimos quadrados coincide com o estimador Seemingly Unrelated Regresssion - SUR (Harvey, 1981, cap. 2). Os resultados foram obtidos usando o *software* S-Plus.

Exemplo 1: Componentes químicos da folha de tabaco

Bedrick e Tsai (1994) apresentaram dados sobre os componentes químicos de uma amostra de 25 folhas de tabaco. Ajustaremos todos os modelos de regressão tendo um termo de intercepto, duas variáveis respostas: Y_1 = taxa de cigarros queimados em polegadas por 1 000 segundos e Y_2 = % de açúcar (na folha); e cinco possíveis

preditores: Z_1 = % de nitrogênio, Z_2 = % de cloro, Z_3 = % de potássio, Z_4 = % de fósforo e Z_5 = % de magnésio.

Os seguintes modelos são comparados:

$M_{\{123\}}$: $[Y_1, Y_2] = [Z_1, Z_2, Z_3]\, B_{\{123\}} + U_{\{123\}}$

$M_{\{125\}}$: $[Y_1, Y_2] = [Z_1, Z_2, Z_5]\, B_{\{125\}} + U_{\{125\}}$

$M_{\{1235\}}$: $[Y_1, Y_2] = [Z_1, Z_2, Z_3, Z_5]\, B_{\{1235\}} + U_{\{1235\}}$

$M_{\{1245\}}$: $[Y_1, Y_2] = [Z_1, Z_2, Z_4, Z_5]\, B_{\{1245\}} + U_{\{1245\}}$

As estimativas dos parâmetros dos modelos são:

Tabela 2 - Estimativas dos parâmetros das regressões

(a) Modelo $M_{\{123\}}$

B_{123}	*Taxa Cigarro*	*Açúcar*
$B_{123.0}$	1.2183	32.9504
$B_{123.1}$	0.1255	-8.9980
$B_{123.2}$	0.0289	0.8652
$B_{123.3}$	0.0395	0.2380

(b) Modelo $M_{\{125\}}$

B_{125}	*Taxa Cigarro*	*Açúcar*
$B_{125.0}$	1.4348	33.2570
$B_{125.1}$	0.2855	-9.1087
$B_{125.2}$	0.0509	0.8608
$B_{125.3}$	-0.5487	0.4862

(a) Modelo $M_{\{123\}}$

$B_{\{1235\}}$	*Taxa Cigarro*	*Açúcar*
$B_{\{1235\}.0}$	1.7103	31.8964
$B_{\{1235\}.1}$	0.3683	-9.5180
$B_{\{1235\}.2}$	0.0663	0.7851
$B_{\{1235\}.3}$	-0.1153	0.5697
$B_{\{1235\}.4}$	-0.7951	1.7030

(b) Modelo $M_{\{125\}}$

$B_{\{1245\}}$	*Taxa Cigarro*	*Açúcar*
$B_{\{1245\}.0}$	1.8167	32.5450
$B_{\{1245\}.1}$	0.2552	-9.0524
$B_{\{1245\}.2}$	0.0511	0.8605
$B_{\{1245\}.3}$	-0.7877	1.4689
$B_{\{1245\}.4}$	-0.4771	0.3528

Os fatores de Bayes para comparar os modelos são:

Tabela 3 - Logaritmo dos fatores de Bayes

Modelos	ln(FBI)	ln(FBP)	ln(FBF)
$M_{\{123\}} \times M_{\{125\}}$	0.1596	-1.0627	-0.7651
$M_{\{125\}} \times M_{\{1235\}}$	3.0665	13.2497	2.6623
$M_{\{125\}} \times M_{\{1245\}}$	3.5095	13.3850	2.7597
$M_{\{123\}} \times M_{\{1235\}}$	1.0569	12.1869	1.8971
$M_{\{123\}} \times M_{\{1245\}}$	6.8640	12.3223	1.9945

Claramente os três fatores de Bayes destacam dois modelos com três preditores: $M_{\{123\}}$ e $M_{\{125\}}$. Mas não existe razão óbvia para se preferir um destes dois modelos. A Tabela 4, abaixo, é apresentada em Bedrick e Tsai (1994). Podemos ver que eles chegaram à mesma conclusão sobre os modelos $M_{\{123\}}$ e $M_{\{125\}}$ usando outros procedimentos, neste caso o modelo escolhido é o que apresenta valor mínimo para cada um dos critérios. Valores mínimos para cada critério são identificados por*.

Tabela 4 - Critérios para seleção de modelos

Modelo	AIC	AICc	BIC	HQ	PRESS
$M_{\{123\}}$	-35.6	-27.0	-72.2	-81.8	45.5
$M_{\{125\}}$	-35.8	-27.2*	-72.4*	-82.0	44.6
$M_{\{1245\}}$	-36.5*	-24.3	-70.7	-82.1*	42.4*

Exemplo 2: inflação brasileira

A inflação brasileira no período do Pós-guerra foi discutida por Barbosa (1983) através dos modelos: *Monetarismo e Estruturalismo* referindo-se as duas principais correntes que influenciaram o pensamento sobre inflação desde a década de 1950. Ele diz que "Os Monetaristas atribuem o crescimento exagerado da oferta de moeda como principal razão por trás do processo inflacionário. Os estruturalistas afirmam que a inflação é gerada dentro do sistema econômico através de mudanças de preços relativos resultantes do crescimento econômico, e que tem sua origem na política monetária, que é passiva e acomoda a variação da renda nominal da economia".

O ponto de vista dos monetaristas é geralmente a curto prazo e favorece um rápido controle da inflação. O modelo estruturalista emergiu de esforços de pesquisas

como a abordagem dominante na década de 1990 combinando atributos clássicos e Keynesianos.

Não houve tentativas de melhorar um ou outro modelo, incluindo ou excluindo certas variáveis. Somente foram usadas aqui as variáveis definidas. Barbosa (1983) apresenta alguma variação na forma destes modelos, trabalhando, todavia, com equações separadas para discriminar entre os dois modelos.

Os dados usados são de Barbosa (1983), referindo-se aos anos 1948 a 1980.

O índice de inflação usado foi o Índice Geral de Preços ao Consumidor - IGP.

Da seção 3.1, $n=32$, $m=2$, $p=4$ (modelo monetarista) e $q=5$ (modelo estruturalista).

Os modelos podem ser escritos da seguinte forma:

a) Monetarismo:

Neste modelo a premissa básica é que crescimento monetário causa inflação e é explicada pelas equações (ver BARBOSA, 1983):

$$\begin{bmatrix} p_t & h_t \end{bmatrix} = \begin{bmatrix} p_{t-1} & h_{t-1} & \mu_{t-1} & D\overline{y}_t \end{bmatrix} \begin{bmatrix} \alpha_1 - \beta\alpha_2 & 1 \\ -\beta\alpha_1 & \alpha_1 \\ \beta & -1 \\ \beta\alpha_1 & \alpha_1 \end{bmatrix} \phi + \varepsilon$$

t varia de 1949 a 1980;

$$\phi = 1/[\alpha_1 + \beta(1-\alpha_2)]$$
$$D\overline{y}_t = Dy_t + h_t - h_{t-1}$$

com

p_t = taxa de inflação ao ano t;

h_t = nível de capacidade ociosa no ano t;

μ_t = taxa de expansão monetária no ano t (oferta de moeda);

$D\overline{y}_t$ = taxa de crescimento do produto potencial; e

Dy_t = taxa de crescimento do produto real total (incluindo agricultura, indústria e comércio).

b) Estruturalismo (CEPAL, anos 1950, anos 1960):

Neste modelo a inflação é causada pela escassez de produtos e conflito entre agentes.

$$\left[\begin{array}{cc} p_t & h_t \end{array}\right] = \left[\begin{array}{ccccc} p_{t-1} & h_{t-1} & S_{m,t} & DZ_t & 1 \end{array}\right] \left[\begin{array}{cc} \beta_{11} - \gamma_{12}\beta_{21} & \gamma_{12}\beta_{11} + \beta_{21} \\ \beta_{12} + \gamma_{12} & \gamma_{12}\beta_{12} + 1 \\ \beta_{13} & \beta_{13}\gamma_{21} \\ \gamma_{12}\beta_{23} & \beta_{23} \\ \beta_{10} + \gamma_{12}\beta_{20} & \beta_{20} + \beta_{10} \end{array}\right] \psi + \varepsilon$$

onde,

$$\psi = 1/[1 - \gamma_{12}\gamma_{21}]$$

com

p_t e h_t os mesmos do modelo monetarista;

S_{mt} = salário mínimo no ano ;

DZ_t = deficit do governo no ano t; e

DZ_t pode ser qualquer variável exógena, ou um grupo delas, que afeta o nível da demanda, tais como: as despesas reais do governo, sua estrutura tributária, taxa real de câmbio, distribuição de renda, etc.

Para estes modelos, foi calculado o fator de Bayes intrínseco, fracionário e *a posteriori*.

Os parâmetros de cada modelo podem ser determinados através da matriz $\hat{\mathbf{B}}$.

Tabela 5 - Estimativas dos parâmetros do modelo monetarista

	p_t	h_t
p_{t-1}	0.3232	0.1429
h_{t-1}	-0.2886	0.7570
μ_t	0.8674	-0.1526
$D\bar{y}_t$	-1.5073	0.2955

Tabela 6 - Estimativa dos parâmetros do modelo estruturalista

	p_t	h_t
p_{t-1}	0.8262	0.0928
h_{t-1}	-0.0716	0.7728
$S_{m,t}$	0.0716	-0.0158
DZ_t	0.0043	-0.0003
1	-0.3974	-1.1388

Fator de bayes intrínseco

Aqui é necessário o cálculo das amostras de treinamento, ao todo eram 26 amostras uma vez que tínhamos séries temporais, e o mínimo $\overline{n}$ de tal maneira que $Z_{(1)}'Z_{(1)}$ seja não singular é $\overline{n}=7$.

De (14),

$$B^{IA}_{mon\times est}(Y) = \frac{|\, Z'_{est}Z_{est}\,||\,S_{est}\,|^{13,5}}{|\, Z'_{mon}Z_{mon}\,||\,S_{mon}\,|^{14}}\frac{\Gamma(14)\Gamma(2)\Gamma(13,5)\Gamma(1,5)}{\Gamma(13,5)\Gamma(2,5)\Gamma(13)\Gamma(2)} \times\frac{1}{24}\sum_{l=1}^{24}\frac{|\, Z'_{mon}(l)Z_{mon}(l)\,||\,S_{mon}(l)\,|^{2,5}}{|\, Z'_{est}(l)Z_{est}(l)\,||\,S_{est}(l)\,|^{2}}$$

com $2\log B^{IA}_{mon\times est}(Y)=27{,}172$.

Fator de bayes fracionário

De (16),

$$B^{[b]}_{mon\times est}(Y)=\left(\frac{|\,S_{est}\,|}{|\,S_{mon}\,|}\right)^{11,5}\frac{\Gamma(14)\Gamma(2)\Gamma(13,5)\Gamma(1,5)}{\Gamma(13,5)\Gamma(2,5)\Gamma(13)\Gamma(2)}$$

com $2\log B^{[b]}_{mon\times est}(Y)=7{,}121$.

Fator de Bayes *a posteriori*

De (13},

$$B^{P}_{mon\times est}(Y) = \left(\frac{|\,S_{est}\,|}{|\,S_{mon}\,|}\right)^{16} \frac{\Gamma(30)\Gamma(13,5)\Gamma(29,5)\Gamma(13)}{\Gamma(29,5)\Gamma(14)\Gamma(29)\Gamma(13,5)}$$

com $2\log B^{P}_{mon\times est}(Y) = 5{,}028$.

Para os modelos descritos, os três fatores de Bayes escolhem o modelo monetarista como melhor modelo para explicar a inflação e a capacidade ociosa brasileira no período do Pós-guerra. Este foi o modelo também escolhido por Barbosa (1983) ao utilizar o critério BIC (mínimo entre -254,2 relativo ao modelo monetarista contra -250,3 relativo ao modelo estruturalista). Entretanto, o uso de critérios clássicos por Barbosa (1983), derivados da proposta de mistura exponencial de modelos (Cox 1961), não permitiu escolher um modelo, pois os testes t rejeitaram ambos os modelos. Para detalhes deste teste ver Pesaran e Weeks (2001) ou Pereira (2005).

5. Conclusão

Em Araujo e Pereira (2007), vários resultados usando simulações foram analisados para os três fatores de Bayes alternativos, bem como outros exemplos univariados, isto é, sem covariáveis. Para os resultados das simulações naquele artigo e para os dados reais aqui analisados, tem-se que:

- Aparte da inconsistência, como apresentado na discussão de Aitkin (1991), o Fator de Bayes *a posteriori* não deveria ser recomendado também em vista dos problemas computacionais encontrados para o seu cálculo, tais como: instabilidade, necessidade de mais precisão computacional especialmente quando n cresce.
- Fator de Bayes intrínseco e o fracionário mostraram possuir comportamento similar. O Fator de Bayes fracionário parece ser mais prático (não exige amostras de treinamento) e requer menor esforço computacional.

- Quando ambos os modelos ajustam os dados igualmente bem, todos os Fatores de Bayes alternativos parecem escolher o modelo mais simples.

Referências bibliográficas

Aitkin, M. (1991). Posterior Bayes factors (with discussion). *J. R. Statist. Soc. B*, n. 53, p. 111-142.

Araujo, M. I.; Pereira, B. de B. (2007). A comparison among Bayes factors for separate models: some simulation results. *Communications in Statistics – Simulation and Computation*, 36, p. 297-309.

Araujo, M. I.; Pereira, B. de B,Cleroux ,R., Fernandes, M., Lazraq, A. (2005). Separated families of models: Sir David Cox contributions and recent developments. *Student* ,V. 5, n. 9, p. 251-258.

Bedrick, E. J.; Tsai, C-L. (1994). Model selection for multivariate Regression in small samples. *Biometrics*, n. 50, p. 226-231.

Barbosa, F. H. (1983*). A Inflação Brasileira no Pós-Guerra: Monetarismo versus Estruturalismo*. Rio de Janeiro: IPEA/INPES.

Berger, J. O.; Pericchi, L. R. (1996). The intrinsic Bayes factor for linear model. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, (Eds.*). Bayesian Statistics 5,* Oxford: University Press, p. 25-44.

Berger, J. O.; Pericchi, L. R. (1996b). The intrinsic Bayes factor for model selection and prediction. *J. American Statist. Association*, 91, p. 109-122.

Cox, D. R. (1961). Test of separated families of hypothesis. *Proc. 4th Berkeley Simposium*, 1, p. 105-123.

Cox, D. R. (1962). Further results on tests of separated families of hypothesis. *J. R. Statist Soc.* B, n. 24, p. 406-423.

Gourieroux, C., Monfort, A. (1994). Testing non nested hypothesis. In: Engle R., Mcfadden, D. L., (Eds.), *Handbook of Econometrics.* Vol. IV. Elservier. p. 2585-2637.

Harvey, A. C. (1981). *The Econometric Analysis of Time Series*. Reprinted 1982. Oxford: Philip Allan Publishers Limited.

Kass, R. E., Rafttery, A. E. (1995). Bayes Factor. *J. American Satist. Association*, 90, p. 773-795.

McAller, M. (1995). The significance of testing empirical non nested models. *Journal of Econometrics* 65: 149-171.

O'Hagan, A. (1995). Fractional Bayes factor for model comparison (with discussion). *J. R. Statist. Soc.* B, n. 57, p. 99-138.

Pereira, B. de B. (1977). Discriminating among separate models: a bibliography. *International Statistical Reviews,* 45, p. 163-172.

Pereira, B. de B. (1981). Discriminating among separate models: an additional bibliography. In: Katti, S. K., ed. On the Preliminary Test for CEAS Model versus the Thompson Models for Predicting Soybean Production. Technical report 125. Department of Statistics, University of Missouri-Columbia. *International Statistical Information,* 62 n. 2, p. 3.

Pereira, B. de B. (2005). Separate families of hypotheses. In: Armitage, P., Colton, T., (Eds.) *Encyclopedia of Biostatistics*, 2 Ed., Vol. 7, New York: John Wiley, pp. 4881-4886.

Pesaran, M. H., Weeks, M. (2001). Non nested hypothesis testing: an overview. In: Badi, H. Baltagi, B. H., (Eds.), *Companion to Theoretical Econometrics*. Oxford: Basil Blackwell.

Rao, C.R. Rao, M.B. (1998) *Matix Álgebra and Its Applications to Statistics and Econometrics*. World Scientific Publishing.

Abstract

To discriminate among alternative models, in Bayesian inference, alternative Bayes factors were developed to overcome the difficulties with the use of improper prior for the parameters of the models. Here we extend these results for systems of separated (non nested) regression equations and present two applications.

Palavras-chave: Discrimination, bayes factor, Intrinsic bayes factor, fractional bayes factor, posterior bayes factor, multivariate regression.

Comparação da estimação do parâmetro d em modelos ARFIMA(p, d, q) para dois métodos de simulação do processo

Simone Maffini Cerezer[*]
Sílvia Regina Costa Lopes[**]
Valdério Anselmo Reisen[***]

Resumo

Neste trabalho comparamos a estimação do parâmetro de diferenciação de processos de longa dependência, ARFIMA(p, d, q), quando $d \in (0,0;0,5)$ utilizando os métodos exato e aproximado para simulação do processo, ambos sugeridos por Hosking (1984). Para estimação do parâmetro de longa dependência foram considerados os estimadores obtidos através do método de regressão, baseado nas funções periodograma e periodograma suavizado. O comportamento destes estimadores é analisado através da média, variância e do erro quadrático médio.

Palavras-chave: Processos de longa dependência, métodos de simulação exato e aproximado, estimação.

[*] Endereços para correspondências: Universidade Federal do Rio Grande do Sul - UFRGS, *e-mail*: scerezer@uri.com.br.
[**] Universidade Federal do Rio Grande do Sul - UFRGS.
[***] Universidade Federal do Espírito Santo - UFES.

1. Introdução

Uma das preocupações fundamentais na análise de séries temporais é o grau de relação entre observações feitas em tempos distantes. Esta característica em uma série temporal é chamada *longa dependência* e uma das formas de medi-la é através da função de autocorrelação entre observações nos tempos t e $t+k$. Um dos modelos que exibe esta característica é o *processo auto-regressivo de médias móveis com integração fracionária*, denotado por ARFIMA(p,d,q), com $d=\in(-,5;0,5)$, região na qual é estacionário e invertível. Estes processos foram introduzidos por Granger e Joyeux (1980) e Hosking (1981).

O fenômeno de longa dependência é conhecido há muito tempo. Beran (1994, pág. 34) cita a Bíblia quando diz "Sete anos de grande abundância são chegados pelo País do Egito, mas sete anos de fome seguirão", que pode ser interpretado como a longa dependência dos efeitos do Rio Nilo. A relação é feita porque as enchentes fertilizam o solo e a safra é relativamente abundante nestes anos. Mas, existem muitos outros exemplos de processos com característica de longa dependência na natureza. Hosking (1984), além de investigar os dados do Rio Nilo (onde $\hat{d} \cong 0,4$), encontra longa dependência na média anual de temperaturas no centro da Inglaterra de 1659 a 1976 (onde $\hat{d} \cong 0,3$).

Os processos ARFIMA(p,d,q), com d fracionário, que em seguida discutiremos, generalizam os processos ARMA(p, q) e ARIMA(p,d,q) com d inteiro, introduzidos por Box *et al.* (1995), apresentando funções de autocorrelação que não decaem para zero de forma geométrica e evidenciando funções de densidade espectral ilimitadas para freqüências próximas de zero. O espectro de um modelo ARMA(p,q) é finito e estritamente maior que zero na freqüência zero.

A geração de séries temporais para a obtenção de amostras de um processo ARFIMA(p,d,q) é um problema complexo que envolve métodos computacionalmente intensivos. Este é um problema grande de síntese onde muitos dos métodos conhecidos exigem grande quantidade de memória computacional, além de serem computacionalmente intensivos. Na literatura existem pelo menos três destes métodos,

propostos por Mcleod e Hipel (1978), Hosking (1982 e 1984) e Davies e Harte (1987).

Os métodos exato e aproximado foram propostos por Hosking (1982 e 1984). O método exato utiliza a estrutura em função da autocorrelação do processo, dada pela equação (4.6), enquanto o método aproximado utiliza a representação auto-regressiva e média móvel do processo, equação (4.8). Desta forma, os resultados na estimativa do parâmetro d para o processo ARFIMA$(0, d, 0)$ coincidem para os métodos de geração, exato e aproximado, do processo. O método proposto por Mcleod e Hipel (1978), de inverter a matriz de correlação do processo, é equivalente ao método exato proposto por Hosking (1982 e 1984). Já o método proposto por Davies e Harte (1987) é baseado na idéia de encobrimento circulante, bastante conhecida nas áreas de engenharia e física. Neste trabalho, escolhemos o algoritmo de Hosking (1982 e 1984) por ser computacionalmente eficiente.

Nos artigos que tratam da estimação do parâmetro d em processos ARFIMA (p, d, q) que utilizam o método aproximado para gerar amostras deste processo (Reisen, 1994; Lopes *et al.*, 2004 e Hurvich e Ray 1995), os autores afirmam que não existem prejuízos para as estimativas de d ao utilizar o método aproximado para gerar as amostras. A grande vantagem do uso do método aproximado é o tempo computacional ser inferior àquele quando o método exato é utilizado. Portanto, existe uma lacuna, na literatura, de uma análise mais criteriosa, tendo em vista que o método aproximado é preferido.

Nosso objetivo, neste trabalho, é comparar os resultados das estimativas do parâmetro d, em processos ARFIMA(p,d,q), quando utilizamos os métodos exato e aproximado, para gerar amostras do processo. Para efeito de comparação, os procedimentos utilizados para a estimação do parâmetro de diferenciação foram os semiparamétricos. Nesta classe, cujos estimadores são obtidos através do método de regressão, estamos considerando aqui aqueles que utilizam as funções periodograma e periodograma suavizado. Ainda na classe dos semiparamétricos, ressaltamos também os procedimentos propostos por Robinson (1994, 1995), Velasco (1999) e Hurvich e Ray (1995). O artigo Lopes *et al.* (2004) reporta um estudo extenso comparando

diversos procedimentos de estimação para o parâmetro d, no caso não-estacionário. Ressaltamos também a classe dos estimadores de máxima verossimilhança, ditos paramétricos, propostos por Sowell (1992) e Fox e Taqqu (1986).

O artigo é apresentado na seguinte forma: os processos ARFIMA(p, d, q), e algumas de suas propriedades, são apresentados na seção 2. Na seção 3 são descritos os dois procedimentos de estimação do parâmetro d, ambos semiparamétricos e baseados no método de regressão. A seção 4 apresenta a simulação dos processos ARFIMA(p, d, q), baseada nos métodos exato e aproximado. Os resultados são apresentados na seção 5 e as conclusões do artigo estão na seção 6.

2. Processo ARFIMA(p, d, q)

O processo $\{Y_t\}_{t\in\mathbf{Z}}$ é um *processo auto-regressivo de médias móveis com integração fracionária*, denotado por ARFIMA(p, d, q), se este é uma solução da equação de diferenças

$$\phi(B)(1-B)^d Y_t = \theta(B)_{\varepsilon_t}, \text{ para todo } t\in\mathbf{Z} \qquad (2.1)$$

onde $d\in(-0{,}5;0{,}5)$ é o parâmetro ou grau de diferenciação, B é o operador da forma $B^j X_t = X_{t-j}$ (também chamado operador defasagem) e $\phi(B)$ e $\theta(B)$ são polinômios de ordem p e q (inteiros), respectivamente, dados por $\phi(B)=1-\phi_1(B)-\ldots-\phi_p B^p$ e $\theta(B)=1-\theta_1(B)-\ldots-\theta_p B^p$, onde ϕ_i, $1\le i\le p$, e θ_j, $1\le j\le q$ são constantes reais. O termo $(1-B)^d$ é definido pela expansão binomial

$$(1-B)^d=\sum_{k=o}^{\infty}\binom{d}{k}(-B)^k=1-dB-\frac{d}{2!}(1-d)B^2-\frac{d}{3!}(1-d)(2-d)B^3-\ldots .$$

O processo $\{\varepsilon_t\}_{t\in\mathbf{Z}}$ é chamado ruído branco, com média constante $E(\varepsilon_t)\equiv\mu_\varepsilon$, usualmente assumida ser zero, variância constante $Var(\varepsilon_t)\equiv\sigma_\varepsilon^2$ e função de

autocovariância $\gamma_\varepsilon(k) \equiv Cov(\varepsilon_t, \varepsilon_{t+k}) = 0$, para todo $k \neq 0$.

Se os polinômios auto-regressivo $\phi(B)$ e de médias móveis $\theta(B)$ têm todas as suas raízes fora do círculo unitário e não possuem raízes em comum, diremos que o processo $\{Y_t\}_{t\in\mathbb{Z}}$ é um ARFIMA(p, d, q) quando o processo $U_t = (1 - B)^d Y_t$ for um ARMA(p, q) estacionário e invertível. Para assegurar a invertibilidade, assume-se que $d > -0{,}5$. Demonstra-se que um ARFIMA(p, d, q) é estacionário e invertível quando $d \in (-0{,}5; 0{,}5)$.

Observamos que, quando $d = 0$, $\{Y_t\}_{t\in\mathbb{Z}}$ é um processo auto-regressivo de médias móveis, denotado por ARMA(p, q).

2.1. Propriedades do processo ARFIMA(p, d, q)

A função de densidade espectral de um processo $\{Y_t\}_{t\in\mathbb{Z}}$ ARFIMA(p, d, q), com $d \in (-0{,}5; 0{,}5)$, pode ser obtida considerando $(1 - B)^d Y_t \equiv U_t$ na expressão (2.1). Desta forma, o processo $\{U_t\}_{t\in\mathbb{Z}}$, dado por $\phi(B)U_t = \theta(B)\varepsilon_t$, é um ARMA($p$, q) com função densidade espectral $f_U(\cdot)$ definida por

$$f_U(\omega) = \frac{\sigma_\varepsilon^2}{2\pi}\left|\frac{\theta(e^{-i\omega})}{\phi(e^{-i\omega})}\right|^2 \text{, para todo } \omega \in [-\pi, \pi]. \qquad (2.2)$$

Considere $X_t = \frac{\phi(B)}{\theta(B)} Y_t$. Então, $\{X_t\}_{t\in\mathbb{Z}}$ é um processo ARFIMA(0, d, 0) com função densidade espectral $f_X(\cdot)$ dada por

$$f_X(\omega) = \left[2 sen\left(\frac{\omega}{2}\right)\right]^{-2d} \text{, para todo } \omega \in [-\pi, \pi]. \qquad (2.3)$$

Portanto, a função densidade espectral do processo $\{Y_t\}_{t\in\mathbb{Z}}$ é dada por

Nota-se imediatamente que a função densidade espectral na vizinhança de zero é dominada pelo termo introduzido pela diferenciação fracionária, isto é,

$$f_Y(\omega) \sim C\omega^{-2d}, \text{ quando } \omega = 0, \tag{2.5}$$

onde $C > 0$ é uma constante.

Sendo assim, os processos ARFIMA(p,d,q) estacionários têm um pólo na freqüência zero quando $d > 0$ e um zero nessa mesma freqüência quando $d < 0$. O que é notável é que essa divergência nas baixas freqüências, também chamada persistência, é obtida com um processo estacionário (o que era impossível para os modelos ARMA e ARIMA) e que a raiz na freqüência zero pode ser obtida por processos ARFIMA invertíveis (também impossível para os modelos ARMA e ARIMA).

Enquanto a função densidade espectral dos processos ARFIMA(p,d,q) é fácil de ser obtida, a sua função de auto-correlação é de cálculo mais complexo.

No entanto, pela transformação do espectro, verifica-se que ela decai hiperbolicamente com a seguinte relação assintótica

$$\rho y(k) \sim Ck^{2d-1}, \text{ quando } k \to \infty, \tag{2.6}$$

onde $\rho y(k) = \frac{\gamma y(k)}{\gamma y(0)}$, para todo $k \in \mathbb{N}$, $\gamma y(\cdot)$ a função de autocovariância do processo e C é uma outra constante maior do que zero.

As condições (2.5) e (2.6) são equivalentes e definem a chamada *memória longa* ou *persistência*. Por vezes distinguem-se os casos de processos com $d > 0$, chamados persistentes, dos com $d < 0$ chamados antipersistentes ou de memória intermediária.

3. Estimação do parâmetro em modelos ARFIMA(p, d, q)

Apresentamos dois procedimentos de estimação, para o parâmetro d dos processos ARFIMA(p, d, q), baseados no método de regressão que utiliza a função

periodograma e a função periodograma suavizado. Estas duas funções são estimadores da função densidade espectral para processos estacionários. Outros estimadores do parâmetro d são amplamente conhecidos na literatura. Entre eles os estimadores de máxima verossimilhança propostos por Sowell (1992) e Fox e Taqqu (1986) e os estimadores semiparamétricos propostos por Robinson (1994, 1995) e Velasco (1999).

3.1. Análise espectral de processos estacionários

Seja $\{Y_t\}_{t\in\mathbb{Z}}$ um processo estocástico estacionário com função de autocovariância $\hat{\gamma}y(\cdot)$ absolutamente convergente, isto é, $\sum_{k\in\mathbb{Z}}|\gamma y(k)|<\infty$. A função densidade espectral de $\{Y_t\}_{t\in\mathbb{Z}}$ é dada por

$$fy(\omega)=\frac{1}{2\omega}\sum_{k=-\infty}^{\infty}\gamma y(k)e^{-i\omega k}=\frac{1}{2\pi}\left[\gamma y(0)+2\sum_{k=1}^{\infty}\gamma y(k)\cos(\omega k)\right], \qquad (3.1)$$

para todo $\omega\in[-\pi,\pi]$.

A equação (3.1) está relacionada com a inversa da transformada de FOURIER (BROCKWELL e DAVIS, 1991), dada por

$$\gamma y(k)=\int_{-\pi}^{\pi}fy(\omega)e^{i\omega k}d\omega\,. \qquad (3.2)$$

Para uma série temporal $\{Y_t\}_{t=1}^{n}$ obtida de um processo $\{Y_t\}_{t\in\mathbb{Z}}$, a função periodograma é definida por

$$fI(\omega)=\frac{1}{2\omega}\left[\hat{\gamma}y(0)+2\sum_{k=1}^{\infty}\hat{\gamma}y(k)\cos(\omega k)\right],\omega\in[-\pi,\pi], \qquad (3.3)$$

onde $\hat{\gamma}y(\cdot)$ é a função de autocovarância amostral do processo $\{Y_t\}_{t\in\mathbb{Z}}$.

A função periodograma $I(\cdot)$, dada pela expressão (3.3), é um estimador não viciado e não consistente da função densidade espectral dada pela expressão (3.1) (BROCKWELL e DAVIS, 1991).

3.2. Estimador de *d* usando a função periodograma

Considere o conjunto de freqüências harmônicas $\omega_j = \frac{2\pi j}{n}$, para $0 \leq j \leq \left[\frac{n}{2}\right]$, onde n é o tamanho da amostra e $[x]$ indica a parte inteira de x. Tomando o logaritmo da função densidade espectral $fy(\cdot)$ e adicionando $lnI(\omega_j)$ e $ln\, f_U(0)$ em ambos os membros da expressão (2.4), temos

$$lnI(\omega_j) = lnf_U(0) - dln\left[2sen\left(\frac{\omega_j}{2}\right)\right]^2 + ln\left[\frac{fU(\omega_j)}{fU(0)}\right] + ln\left[\frac{I(\omega_j)}{fy(\omega_j)}\right], \qquad (3.4)$$

onde $I(\cdot)$ é a função periodograma dada pela expressão (3.3). Quando ω_j está próximo de zero, por exemplo, $\omega_j \leq \omega_l$, onde ω_l é pequeno, então o penúltimo termo da expressão (3.4) é negligível quando comparado com os demais termos da direita nesta mesma expressão (Brockwell e Davis, 1991 e Geweke e Porter-Hudak, 1983).

O estimador de d, pelo método de regressão de mínimos quadrados, é dado por

$$\hat{d}_p = -\frac{\sum_{j=1}^{g(n)} (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{g(n)} (x_j - \bar{x})^2}, \qquad (3.5)$$

onde $E(\hat{d}_p) = d$, $Var(\hat{dp}) = \frac{\pi^2}{6\sum_{j=1}^{g(n)} (x_j - \bar{x})^2}$, $\bar{x}$ é a média de $x_j = ln\left[2sen\left(\frac{\omega_j}{2}\right)\right]^2$ e $\bar{y}$ é a média de $y_j = ln\, l(\omega_j)$, para todo $j = 1,2,\cdots,g(n)$ e $g(n) = n^{\alpha}$, $0 < \alpha < 1$, determina o número de regressores (Geweke e Porter-Hudak, 1983). Neste trabalho consideramos $\alpha \in \{0,5;0,6;0,7\}$.

3.3. Estimador de d usando a função periodograma suavizado

O estimador de d, usando a função periodograma suavizado, foi proposto por REISEN (1994). O objetivo, neste método de estimação, é usar um estimador consistente para a função densidade espectral. Para isto, REISEN (1994) utiliza a função periodograma suavizado $f_s(\cdot)$, que possui uma expressão similar a (3.3), com um núcleo dado pela janela espectral de Parzen, dada pela seguinte forma

$$f_s(\omega)=\frac{1}{2\pi}\sum_{k=-(n-1)}^{n-1}\lambda(k)\hat{\gamma}y(k)\cos(\omega k),\ \text{para todo}\ \omega\in[-\pi,\pi], \tag{3.6}$$

onde $\lambda(\cdot)$ é a janela de Parzen que é dada por

$$\lambda(k)=\begin{cases}1-6\left(\frac{k}{m}\right)^2+6\left(\frac{|k|}{m}\right)^3, & \text{se } |k|\le\frac{m}{2}\\ 2\left(1-\frac{|k|}{m}\right)^3, & \text{se } \frac{m}{2}\le|k|\le m\\ 0, & \text{caso contrário,}\end{cases}$$

onde m é o *ponto de truncamento* da janela de Parzen. Este ponto de truncamento é função do tamanho amostral n sendo escolhido da forma $m=n^{\beta}$, para algum $\beta\in(0,1)$ fixo. Neste trabalho consideramos $\beta=0{,}9$ (este valor foi escolhido a partir de estudos apresentados por Reisen, 1994). O estimador de d é obtido substituindo a função densidade espectral na expressão (3.4) pela função $f_s(\cdot)$ com a janela de Parzen. Reisen (1994) mostra que este estimador tem expressão igual aquela dada por (3.5) onde agora $y_j=ln(f_s(\omega_j))$ em vez de $y_j=ln(I(\omega_j))$. Denotaremos este novo estimador por $\hat{d}_{sp}$, onde $g(n)$ é escolhido como no estimador $\hat{d}_p$.

4. Simulação de modelos ARFIMA(p, d, q)

Nosso objetivo, nesta seção, é a comparação entre os métodos de simulação exato e aproximado usados para gerar amostras de modelos ARFIMA(p, d, q), ambos propostos por Hosking (1982, 1984).

Inicialmente, Hosking (1982, 1984) sugere gerar amostras $\{X_t\}_{t=0}^{n-1}$ de tamanho n de um processo estacionário $\{X_t\}_{t\in\mathbb{Z}}$ ARFIMA(0, d, 0), quando o ruído branco tem distribuição normal, através do algoritmo que possui as seguintes etapas:

Primeira Etapa: Gerar uma variável aleatória X_0 com distribuição $N(0,\sigma_0^2)$, onde $\sigma_0^2 = \frac{\Gamma(1-2d)}{[\Gamma(1-d)]^2}$ é a variância do processo $\{X_t\}_{t\in\mathbb{Z}}$, sendo $\Gamma(\cdot)$ a função gama. Definir $N_0 = 0$ e $D_0 = 1$;

Segunda Etapa: Para $t = 1,2,\ldots,n-1$, calcular $\phi_{t,j}$, para $1 \le j \le t$, recursivamente através das equações

$$N_t = \rho x(t) - \sum_{j=1}^{t-1} \phi_{t-1,j}\, \rho x(t-j), \tag{4.1}$$

$$D_t = D_{t-1} - \frac{N_{t-1}^2}{D_{t-1}}, \tag{4.2}$$

$$\phi_{t,t} = \frac{N_t}{D_t} \tag{4.3}$$

e $\phi_{t,j} = \phi_{t-1,j} - \phi_{t,t}\, \phi_{t-1,t-j}$, para $j \in \{1,2,\ldots,n-1\}$. (4.4)

Terceira Etapa: Calcular $m_t = \sum_{j=1}^{t} \phi_{t,j} X_{t-j}$ para todo $t \in \{1,2,\ldots,n-1\}$, e $v_t = (1-\phi_{t,t}^2) v_{t-1}$, onde $v_0 = \sigma_0^2$, para todo $t \in \{1,2,\ldots,n-1\}$.

Quarta Etapa: Gerar a variável X_t a partir de uma distribuição $N(m_t,v_t)$. Para o processo $\{X_t\}_{t\in\mathbb{Z}}$ ARFIMA(0, *d*, 0), o algoritmo pode ser simplificado, substituindo-se o cálculo dos coeficientes $\phi_{t,j}$ pela função de autocorrelação parcial de ordem k dada por

$$\phi_{k,k}=\frac{d}{k-d}, \text{para } k\in\mathbb{N}. \tag{4.5}$$

Fazendo algumas manipulações algébricas podemos perceber que as expressões (4.1), (4.2) e (4.3) são equivalentes a expressão (4.5) quando as ordens p e q são iguais a zero. Segundo Hosking (1982, 1984), reduzimos o tempo de computação pela metade para a execução do algoritmo que gera amostras do modelo ARFIMA(0,*d*,0) se substituirmos, na segunda etapa, as expressões (4.1), (4.2) e (4.3) pela expressão (4.5).

Para gerar um processo ARFIMA(*p, d, q*) pelo método de simulação exato utilizamos a estrutura de correlação do processo onde

$$\rho y(k)=\frac{\gamma y(k)}{\gamma y(0)} \text{ com } \gamma y(k)=\sum_{j\in\mathbb{Z}}\gamma U(j)\gamma X(j+k), \tag{4.6}$$

com $\rho Y(\cdot)$ e $\gamma Y(\cdot)$ as funções de autocorrelação e autocovarância, respectivamente, do processo $\{Y_t\}_{t\in\mathbb{Z}}$ ARFIMA(*p,d,q*); $\gamma U(\cdot)$ a função de autocovariância do processo $\{U_t\}_{t\in\mathbb{Z}}$ ARMA(*p, q*); $\gamma X(\cdot)$ a função de autocovariância do processo $\{X_t\}_{t\in\mathbb{Z}}$ ARFIMA(*0, d, 0*). A soma infinita em (4.6) deve ser truncada em $j=k$ tal que $\gamma Y(l)$, para $l>k$, seja negligível. Para maiores detalhes Hosking (1984). Neste trabalho o valor de k considerado foi igual ao tamanho amostral da série (Cerezer, 1999, para maiores detalhes). Para gerar um processo $\{Y_t\}_{t\in\mathbb{Z}}$ ARFIMA(*p, d, q*), dado pela expressão (2.1), pelo método aproximado, deve-se decompor este processo em

$$\phi(B)Y_t=\theta(B)X_t \text{ e } \Delta^d X_t=\varepsilon_t. \tag{4.7}$$

O processo X_t é um ARFIMA(0, d, 0) e pode ser simulado como descrito anteriormente. O processo Y_t pode ser simulado usando a forma recursiva

$$Y_t = \sum_{j=1}^{p} \phi_j Y_{t-j} + X_t - \sum_{j=1}^{q} \theta_j X_{t-j} . \qquad (4.8)$$

A recursão para Y_t envolve $(n+q)$ valores de X_t que são necessários para gerar n valores de Y_t. Quando o modelo possui forma auto-regressiva a equação (4.8) também envolve valores Y_t que não estão disponíveis para iniciar o processo. Nesta situação, HOSKING (1984) sugere gerar Y_t, $-r \le t \le n$, r é um número inteiro, considerando $Y_t = 0$ para $t \le -r$ e eliminando $Y_t - r \le t \le 0$. Então, a amostra resultante é $Y_1, Y_2, \ldots, Y_n$. Ele também sugere que r pode ser escolhido de tal forma que satisfaça $|\eta|^r \le \xi$, onde η é a raiz da equação (em z), $1 - \sum_{j=1}^{p} \phi_{jz^{-j}} = 0$, e ξ é um número pequeno (por exemplo, $\xi = 0{,}01$).

5. Resultados numéricos

Nesta seção, apresentamos, através de simulações realizadas usando sub-rotinas do FORTRAN (IMSL), o comportamento dos procedimentos de estimação de d obtido do método de regressão baseado nas funções periodograma e periodograma suavizado utilizando os métodos exato e aproximado para gerar processos ARFIMA(p, d, q). Os estimadores obtidos por esses métodos, $\hat{d}_p$ e $\hat{d}_{sp}$, são comparados com base na média, variância e erro quadrático médio para diferentes casos do processo ARFIMA. O efeito do tamanho amostral também é analisado. Nas expressões de $\hat{d}_p$ e $\hat{d}_{sp}$ usamos a função $g(n) = n^{\alpha}$, com $\alpha \in \{0{,}5; 0{,}6; 0{,}7\}$. No estimador $\hat{d}_{sp}$ consideramos o ponto de truncamento $m = n^{\beta}$, com $\beta = 0{,}9$. Para cada estimador $(i = p, sp)$ as tabelas apresentam a média nas estimativas do parâmetro $d(\hat{d}_i)$, a variância $(\mathrm{var}(\hat{d}_i))$ e o erro quadrático médio $(eqm(\hat{d}_i))$ referentes a 1 000 replicações do processo. Para cada valor de ϕ e θ

as tabelas apresentam duas linhas onde na primeira estão os resultados referentes ao método de geração exato, enquanto na segunda estão os resultados do aproximado.

A Tabela 5.1 apresenta resultados das simulações para o modelo ARFIMA (*1, d, 0*) quando são realizadas 1 000 replicações do processo com número de regressores dado pela função $g(n)=n^{\alpha}$, com $\alpha=0{,}5$.

Tabela 5.1 - Estimação do parâmetro *d* para o modelo ARFIMA(1,*d*,0), com $d=0{,}1$ e $\phi \in \{-0{,}3;-0{,}1;0{,}1;0{,}3\}$, para diferentes tamanhos amostrais

n	ϕ	$\hat{d}_p$	$\hat{d}_{sp}$	$var(\hat{d}_p)$	$var(\hat{d}_{sp})$	$eqm(\hat{d}_p)$	$eqm(\hat{d}_{sp})$
100	-0,3	0,0783	0,0059	0,0851	0,0426	0,0855	0,0514
		0,0933	0,0203	0,0842	0,0432	0,0842	0,0495
	-0,1	0,0951	0,0217	0,0835	0,0433	0,0834	0,0494
		0,1020	0,0831	0,0281	0,0435	0,0830	0,0486
	0,1	0,1278	0,0527	0,0821	0,0438	0,0828	0,0460
		0,1180	0,0436	0,0830	0,0437	0,0842	0,0468
	0,3	0,1934	0,1171	0,0808	0,0442	0,0894	0,0445
		0,1519	0,0761	0,0815	0,0440	0,0841	0,0445
350	-0,3	0,0982	0,0533	0,0357	0,0220	0,0357	0,0241
		0,1014	0,0566	0,0358	0,0220	0,0357	0,0239
	-0,1	0,1017	0,0568	0,0358	0,0220	0,0358	0,0239
		0,1032	0,0583	0,0357	0,0220	0,0357	0,0237
	0,1	0,1091	0,0641	0,0353	0,0220	0,0354	0,0233
		0,1069	0,0619	0,0354	0,0220	0,0354	0,0235
	0,3	0,1258	0,0804	0,0350	0,0221	0,0356	0,0224
		0,1153	0,0701	0,0352	0,0220	0,0354	0,0229
500	-0,3	0,1044	0,0645	0,0276	0,0172	0,0276	0,0184
		0,1070	0,0668	0,0278	0,0172	0,0278	0,0183
	-0,1	0,1073	0,0669	0,0277	0,0173	0,0278	0,0183
		0,1083	0,0679	0,0277	0,0173	0,0277	0,0183
	0,1	0,1124	0,0719	0,0276	0,0173	0,0277	0,0181
		0,1109	0,0704	0,0276	0,0173	0,0277	0,0181
	0,3	0,1242	0,0833	0,0275	0,0173	0,0281	0,0176
		0,1168	0,0761	0,0276	0,0173	0,0278	0,0178

Analisando os resultados apresentados na Tabela 5.1 observamos que o estimador $\hat{d}_p$ produziu resultados mais satisfatórios na média das estimativas do parâmetro d quando comparamos com os resultados obtidos através do estimador $\hat{d}_{sp}$, para os dois métodos de geração do processo ARFIMA(*1, d, 0*) com d = 0, 1. Porém, a variância e o

erro quadrático médio do estimador $\hat{d}_{sp}$ são menores do que para o estimador $\hat{d}_p$. Ressaltamos que o tamanho amostral da série não exerceu influência significativa nos resultados médios das estimativas do parâmetro d, mas com o aumento do tamanho amostral a variância e o erro quadrático médio para os estimadores $\hat{d}_p$ e $\hat{d}_{sp}$ reduziram, para os dois métodos de geração do processo, como era esperado.

Na Tabela 5.2, apresentamos os resultados das simulações para o modelo ARFIMA(*0,d,1*), quando são realizadas 1 000 replicações desse processo com tamanho amostral 500 e com número diferente de regressores.

Tabela 5.2 - Estimação do parâmetro d para o modelo ARFIMA(0,d,1), com $d = 0,2$, $\theta \in \{-0,3; -0,1; 0,1; 0,3\}$ tamanho amostral 500

α	θ	$\hat{d}_p$	$\hat{d}_{sp}$	$var(\hat{d}_p)$	$var(\hat{d}_{sp})$	$eqm(\hat{d}_p)$	$eqm(\hat{d}_{sp})$
0,5	-0,3	0,2171	0,1720	0,0276	0,0178	0,0279	0,0186
		0,2147	0,1699	0,0277	0,0178	0,0279	0,0187
	-0,1	0,2145	0,1698	0,0276	0,0178	0,0278	0,0187
		0,2134	0,1687	0,0277	0,0178	0,0278	0,0187
	0,1	0,2092	0,1648	0,0277	0,0178	0,0278	0,0190
		0,2108	0,1663	0,0277	0,0178	0,0278	0,0189
	0,3	0,1966	0,1530	0,0277	0,0177	0,0277	0,0199
		0,2047	0,1645	0,0277	0,0177	0,0277	0,0193
0,6	-0,3	0,2237	0,1970	0,0136	0,0087	0,0142	0,0087
		0,2170	0,1904	0,0136	0,0087	0,0142	0,0087
	-0,1	0,2164	0,1899	0,0136	0,0086	0,0139	0,0088
		0,2133	0,1868	0,0135	0,0085	0,0138	0,0089
	0,1	0,2010	0,1747	0,0136	0,0087	0,0136	0,0094
		0,2057	0,1792	0,0136	0,0087	0,0137	0,0091
	0,3	0,1666	0,1409	0,0137	0,0087	0,0148	0,0122
		0,1886	0,1624	0,0136	0,0087	0,0138	0,0101
0,7	-0,3	0,2508	0,2356	0,0063	0,0042	0,0088	0,0055
		0,2291	0,2139	0,0063	0,0042	0,0071	0,0044
	-0,1	0,2270	0,2118	0,0063	0,0042	0,0070	0,0044
		0,2173	0,2020	0,0063	0,0042	0,0060	0,0042
	0,1	0,1803	0,1650	0,0063	0,0042	0,0067	0,0055
		0,1940	0,1786	0,0063	0,0042	0,0063	0,0047
	0,3	0,0887	0,0734	0,0064	0,0042	0,0188	0,0203
		0,1483	0,1329	0,0063	0,0042	0,0090	0,0087

Comparando os resultados na estimativa do parâmetro d para séries geradas pelos métodos exato e aproximado, com o mesmo tamanho amostral (n = 500), mas com números diferentes de regressores (Tabela 5.2), observamos que à medida que aumenta o número de regressores ocorre uma redução considerável nos valores da variância e do erro quadrático médio dos estimadores $\hat{d}_p$ e $\hat{d}_{sp}$. No entanto, o vício para os estimadores $\hat{d}_p$ e $\hat{d}_{sp}$, em geral, é maior para valores de $\theta<0$ e menor para valores de $\theta>0$, para os dois métodos de geração do processo ARFIMA(0, d, 1).

Na Tabela 5.3, são apresentados os resultados das simulações para o modelo ARFIMA(1, d, 1), quando são realizadas 1 000 replicações desse processo com tamanho amostral 500 e número de regressores dado pela $g(n)=n^{\alpha}$, com $\alpha=0{,}5$.

Tabela 5.3 - Estimação do parâmetro d para o modelo ARFIMA(1,d,1), com $d=0{,}2$, para diferentes valores de ϕ e θ e tamanho amostral 500

ϕ	θ	$\hat{d}_p$	$\hat{d}_{sp}$	$var(\hat{d}_p)$	$var(\hat{d}_{sp})$	$eqm(\hat{d}_p)$	$eqm(\hat{d}_{sp})$
0,1	0,2	0,2076	0,1632	0,0277	0,0178	0,0278	0,0191
		0,2100	0,1655	0,0277	0,0178	0,0278	0,0189
0,1	0,3	0,2001	0,1561	0,0278	0,0177	0,0277	0,0196
		0,2063	0,1620	0,0277	0,0177	0,0277	0,0192
0,2	0,1	0,2171	0,1722	0,0277	0,0178	0,0279	0,0185
		0,2147	0,1700	0,0277	0,0178	0,0279	0,0187
0,2	0,4	0,1930	0,1492	0,0277	0,0177	0,0277	0,0203
		0,2028	0,1585	0,0277	0,0177	0,0277	0,0194
0,3	0,2	0,2196	0,1748	0,0277	0,0178	0,0280	0,0184
		0,2159	0,1713	0,0277	0,0178	0,0279	0,0186
0,3	0,4	0,2005	0,1563	0,0277	0,0177	0,0277	0,0196
		0,2063	0,1620	0,0277	0,0178	0,0277	0,0192
0,4	0,1	0,2359	0,1906	0,0276	0,0178	0,0289	0,0179
		0,2242	0,1792	0,0277	0,0178	0,0283	0,0182
0,4	0,3	0,2241	0,1791	0,0277	0,0178	0,0282	0,0182
		0,2182	0,1734	0,0277	0,0178	0,0280	0,0185

Para os resultados da estimativa do parâmetro d apresentados na Tabela 5.3, observamos que o estimador $\hat{d}_{sp}$ apresenta menor variância e menor erro quadrático médio nos dois métodos de geração do processo ARFIMA(1, d, 1), com d = 0, 2, quando

comparado com $\hat{d}_p$. Ressaltamos que os estimadores $\hat{d}_{sp}$ e $\hat{d}_p$ apresentam resultados similares na estimação do parâmetro d para os métodos, exato e aproximado, de geração do processo.

6. Conclusões e trabalhos futuros

Neste trabalho analisamos o comportamento dos estimadores do parâmetro d de longa dependência de processos estacionários ARFIMA(1, d, 0), ARFIMA (0, d, 1) e ARFIMA(1, d, 1), quando $d \in (0,0;0,5)$, através do método de regressão utilizando as funções periodograma e periodograma suavizado, pelos métodos, exato e aproximado, de geração destes processos. Os resultados das estimativas do parâmetro d no modelo ARFIMA(1, d, 0), obtidas por $\hat{d}_p$ e $\hat{d}_{sp}$ para o procedimento exato, são inferiores às estimativas de d pelos mesmos estimadores quando o procedimento utilizado para gerar o processo é o aproximado, quando ϕ é menor que zero. No entanto, quando ϕ é maior que zero, os resultados das estimativas do parâmetro d, obtidas por $\hat{d}_p$ e $\hat{d}_{sp}$ para o procedimento exato, são superiores às estimativas de d pelos mesmos estimadores, quando o procedimento utilizado para gerar o processo é o aproximado.

Percebemos que para os modelos ARFIMA(1, d, 0), ARFIMA(0, d, 1) e ARFIMA (1, d, 1) os resultados de $\hat{d}_p$ são muito próximos para os métodos, exato e aproximado, de geração do processo. O mesmo acontece para os resultados de $\hat{d}_{sp}$. No entanto, o estimador $\hat{d}_p$ fornece melhores estimativas do que o $\hat{d}_{sp}$, mas com maior variância e erro quadrático médio, independentemente do método de geração do processo.

Analisando o tamanho amostral da série, observamos que o mesmo não exerceu influência significativa nas estimativas do parâmetro d, porém com o aumento do tamanho amostral a variância e o erro quadrático médio para os estimadores $\hat{d}_p$ e $\hat{d}_{sp}$ reduziram, para os métodos, exato e aproximado, de geração do processo, o que era esperado. Ressaltamos também que os valores da variância e do erro quadrático médio

dos estimadores $\hat{d}_p$ e $\hat{d}_{sp}$ são reduzidos consideravelmente com o aumento do número de regressores, para os dois métodos de geração do processo. Em contrapartida, em geral, os vícios aumentam.

Portanto, concluímos que o método aproximado pode continuar sendo utilizado para gerar modelos ARFIMA(p, d, q) sem prejuízos nas estimativas do parâmetro d, quando os procedimentos nas estimativas deste parâmetro são baseados nas funções periodograma e periodograma suavizado, é necessário um estudo mais aprofundado para se verificar se o mesmo ocorre quando p e q superam 1 e quando são utilizados outros estimadores do parâmetro d.

Referência bibliográfica

Beran, J. (1994). *Statistics for Long-Memory Processes*. New York: Chapman & Hall.

Box, G.E.P., Jenkins, G.M. e Reinsel, G.C. (1995). *Time Series Analysis: Forecasting and Control*. New Jersey: Prentice Hall.

Brockwell, P.J. e Davis, R.A. (1991). *Time Series: Theory and Methods*. New York: Springer-Verlag.

Cerezer, S.M. (1999). *Estimação do Parâmetro d em Modelos ARFIMA (p,d,q) utilizando o Método de Simulação Exata*. Dissertação de Mestrado. Instituto de Matemática da UFRGS, Porto Alegre.

Davies, R. e Harte, D. (1987). Tests for Hurst effect. *Biometrika*, Vol. 74, 95-101.

Fox, R. e Taqqu, M.S. (1986). Large-Sample Properties of Parameter Estimates for Strongly Dependent Stationary Gaussian Time Series. *Annals of Statistics*, Vol. 14(2), 517-532.

Geweke, J. e Porter-Hudak, S. (1983). The Estimation and Application of Long Memory Time Series Model. *Journal of Time Series Analysis*, Vol. 4(4), 221-238.

Granger, C.W.J. e Joyeux, R. (1980). An Introduction to Long-Memory Time Series Models and Fractional Differencing. *Journal of Time Series Analysis*, Vol. 1(1), 15-29.

Hosking, J. (1981). Fractional Differencing. *Biometrika*, Vol. 68(1), 165-167.

Hosking, J. (1982). Some models of persistence in time series analysis: theory and practice. *North Holland Publishing Company*, 641-653.

Hosking, J. (1984). Modelling Persistence in Hydrological Time Series using Fractional Differencing. *Water Resources Research*, Vol. 20(12), 1898-1908.

Hurvich, C.M. e Ray, B.K. (1995). Estimation of the memory parameter for nonstationary or noninvertible fractionally integrated processes. *Journal of Time Series Analysis*, Vol. 16, 017-042.

Lopes, S.R.C., Olbermann, B.P. e Reisen, V.A. (2004). A Comparison of Estimation Methods in Non-Stationary ARFIMA Processes. *Journal of Statistical Computation & Simulation*, Vol. 74(5), 339-347.

McLeod, B. e Hipel, K. (1978). Preservation of the rescaled adjusted range, I. A reassessment of the Hurst phenomenon. *Water Resources Research*, Vol. 14, 491-518.

Reisen, V.A. (1994). Estimation of the Fractional Difference Parameter in 15 the ARIMA(p, d, q) model using the Smoothed Periodogram. *Journal of Time Series Analysis*, Vol. 15(3), 335-350.

Robinson, P.M. (1994). Rates of convergence and optimal spectral bandwidth for long range dependence. *Probability Theory and Related Fields*, Vol. 99, 443-473.

Robinson, P.M. (1995). Log-periodogram regression of time series with long range dependence. *Annals of Statistics*, Vol. 23(3), 1630-1661.

Sowell, F. (1992). Maximum Likelihood estimation of stationary univariate fractionally integrated time series models. *Journal of Econometrics*, Vol. 53, 165-188.

Velasco, C. (1999). Gaussian Semiparametric Estimation of Non-stationary Time Series. *Journal of Time Series Analysis*, Vol. 20(1), 87-127.

Abstract

In this work we compare the estimation of the differencing parameter in long dependence processes, ARFIMA(p, d, q), when $d\hat{I}(0,0;0,5)$ by using the exact and approximated simulation methods, both suggested by Hosking (1984). We investigate the estimators of d through the regression method based on the periodogram and on the smoothed periodogram functions. The performance of these estimators is analyzed by their mean, variance and their mean squared error values.

Keywords: Long dependence processes, exact and approximated simulation methods, estimation.

Agradecimentos

S.R.C. Lopes foi parcialmente financiada pelo CNPq-Brasil, pelo Pronex Probabilidade e Processos Estocásticos (Convênio MCT/CNPq/FAPERJ - Edital 2003), pelo Edital Universal Modelos com Dependência de Longo Alcance: Análise Probabilística e Inferência (CNPq-Nº. 476781/2004-3) e também pela Fundação de Amparo à Pesquisa no Estado do Rio Grande do Sul-FAPERGS). V.A. Reisen foi parcialmente financiado pelo CNPq-Brasil.

O impacto da política de não-repetência na proficiência dos alunos da quarta série: um estudo sobre o Sudeste brasileiro

Maria Eugénia Ferrão[*]
Kaizô Iwakami Beltrão[**]
Denis Paulo dos Santos[***]

[*] Endereço para correspondências: Rua André Cavalcanti, 106; Rio de Janeiro 20231-050 - Escola Nacional de Ciências Estatísticas (ENCE/IBGE), *e-mail*: mariabarbosa@ibge.gov.br; Departamento de Matemática da Universidade da Beira Interior /Portugal (UBI), *e-mail*: meferrao@noe.ubi.pt.
[**] Escola Nacional de Ciências Estatísticas (ENCE/IBGE), *e-mail*: kaizo@ibge.gov.br.
[***] Escola Nacional de Ciências Estatísticas (ENCE/IBGE), *e-mail*: denis.santos@ibge.gov.br

Resumo

Este trabalho investiga o impacto do regime de organização do ensino (promoção automática) sobre o desempenho dos alunos da quarta série da Região Sudeste do Brasil.

A proporção de alunos com defasagem idade-série no Brasil é de 44% no ensino fundamental e de 55% no ensino médio. A defasagem pode ser decorrente de três fenômenos: entrada tardia na escola, repetência e abandono precoce com posterior reingresso no sistema educacional. A repetência é um fenômeno implícito ao regime de organização do ensino em séries – ensino seriado e para atenuar os elevados índices de repetência têm sido adotadas políticas de não-repetência, particularmente na forma de reorganização do ensino. Elas têm sido designadas como "promoção automática", "avaliação continuada" ou ensino organizado em ciclos, fases ou etapas. Na Região Sudeste, a promoção automática encontra-se mais fortemente disseminada nos Estados de Minas Gerais e São Paulo, onde mais de 50% das escolas adotam esse regime de organização.

Considerando-se que o problema da cobertura da rede já parece em vias de solução, o grande desafio do sistema educacional transfere-se então para a correção da defasagem idade-série sem perda da qualidade na educação provida à população.

A nossa questão de pesquisa é: a promoção automática, ou políticas equivalentes de não-repetência, corrigem a defasagem idade-série sem perda de qualidade na educacão?

Aplicaram-se modelos de regressão multinível aos dados do Sistema Nacional de Avaliação da Educação Básica - SAEB (algumas das variáveis são provenientes do Censo Educacional) referentes à Região Sudeste. Modelos específicos foram ajustados para São Paulo e Minas Gerais. Os resultados sugerem que o impacto dessa política sobre o desempenho dos alunos depende do estado em que tem sido implementada. Nas escolas da rede pública, não se encontraram evidências de que o impacto seja substancial, mesmo quando estatisticamente significativo.

Palavras-chave: Avaliação educacional, defasagem idade-série, promoção automática, modelo multinível.

1. Introdução

1.1. Contexto

Em 1999, o número de alunos matriculados no Ensino Fundamental foi superior a 36 milhões, dos quais 91% estavam matriculados em escolas públicas, conforme pode ser observado na Tabela 1. Esta Tabela mostra a distribuição dos alunos por tipo de administração das escolas – pública (federal, estadual, ou municipal) e particular. A proporção de alunos que estudam em escolas sob administração estadual é de 46% e sob administração municipal, 45%.

Tabela 1- Distribuição dos alunos do Ensino Fundamental por tipo de administração das escolas, Brasil, 1999 (números em milhares de alunos)

Administração		Número de alunos	Distribuição %
Pública	Federal	28,6	0,08
	Estadual	16 589,5	46,01
	Municipal	16 164,4	44,82
Particular		3 277,3	9,09
Total		36 059,8	100,00

Fonte: MEC/INEP/SEEC.

A taxa de escolarização, na faixa etária de 7 a 14 anos, era de 83,8% em 1989 e de 95,7% em 1999 (Fonte: IBGE, PNAD[1]). O percentual de crianças entre 7 e 9 anos fora da escola era de 29,3% em 1981, 15,0% em 1989 e 3,8% em 1999. Estas taxas mostram a progressiva melhoria que tem ocorrido no sistema educacional em termos de cobertura.

Os gráficos das Figuras 1.1 e 1.2 mostram, como função do tempo, a taxa de matrícula para cada Grande Região e o percentual de crianças fora da escola para faixas etárias selecionadas. A região com a menor taxa de matrículas foi o Nordeste, que até 1996 era a única com taxa abaixo da média nacional, fazendo-se uma ressalva da ausência de amostragem na população rural da Região Norte, o que pode inflar os

[1] Pesquisa Nacional de Amostra de Domicílios - PNAD é uma pesquisa nacional domiciliar cuja amostragem cobre áreas urbana e rural de todos os estados brasileiros, com exceção das áreas rurais da Região Norte. A PNAD é realizada todos os anos, com exceção daquelas em que se realizam os censos. A partir de 2004, passou a abranger também as áreas rurais de todas as Unidades da Federação da Região Norte.

resultados para tal região. A partir dessa data, a Região Norte também começou a apresentar taxas abaixo da média nacional. O Sudeste é a região com maiores taxas de matrícula em todo o período. O Sul e o Centro-Oeste, que apresentavam uma diferença em relação ao Sudeste de cerca de 3,5% em 1989, reduziram tal diferença para 0,3% em 1999. Os ganhos no período, apresentados pelas regiões menos favorecidas, foram maiores do que os das regiões mais afluentes: o Nordeste aumentou a sua cobertura em 17,6%, e o Sudeste em somente 8,1%.

Concomitantemente, observa-se que o percentual de crianças fora da escola vem caindo regularmente (Figura 1.2). Para os grupos etários de 7 a 14 anos, esse percentual cruzou a linha dos 5% antes de 1999. Um estudo mais detalhado (Beltrão e Ferrão, 2002) mostra que a percentagem de alunos fora da escola aumenta com a idade. Verifica-se que o grupo etário 5-6 (educação infantil) também vem apresentando melhorias substanciais e, vale notar que, estar no sistema educacional no grupo etário 5 e 6 anos é um bom prenúncio de entrada no ensino fundamental com a idade recomendada. A maior proporção de alunos que freqüentam a educação infantil registra-se no Nordeste.

Figura 1.1 - Taxa de escolarização de crianças de 7 a 14 anos de idade, por Grandes Regiões

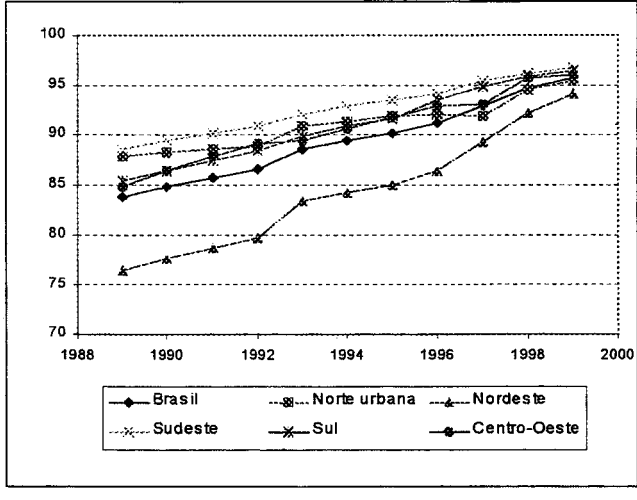
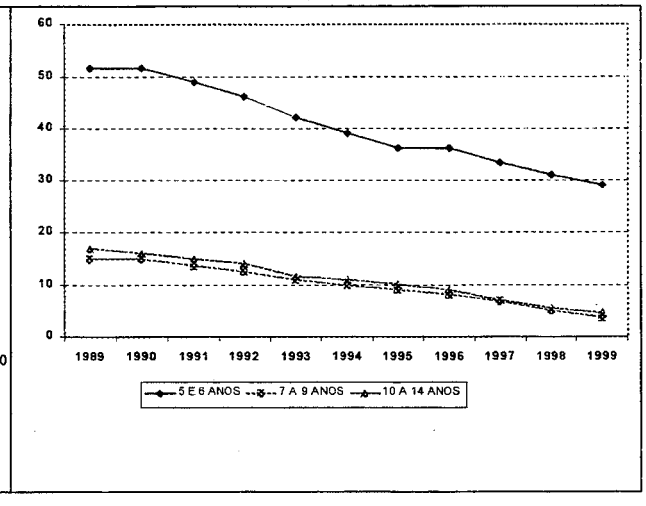


Figura 1.2 - Percentual de crianças de 7 a 14 anos fora da escola



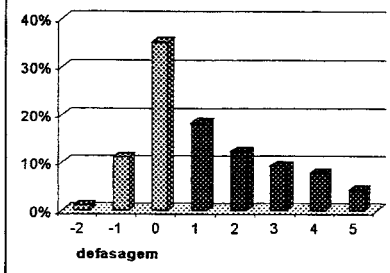
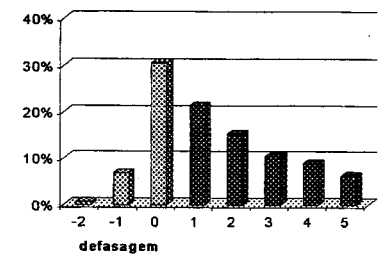
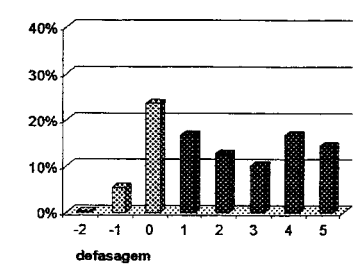
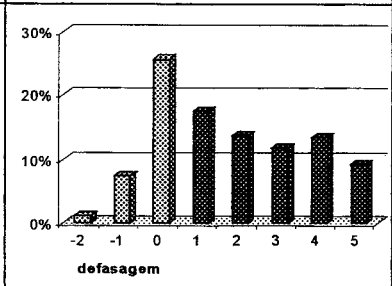
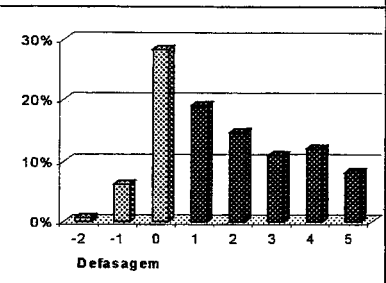
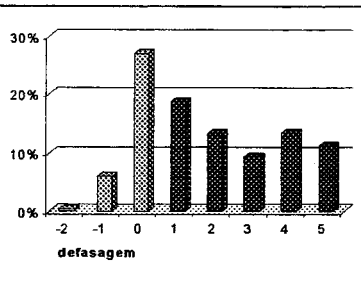
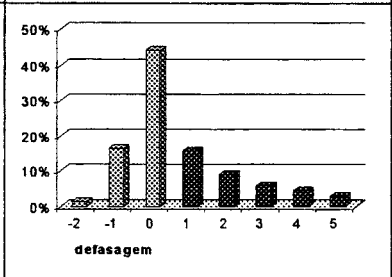
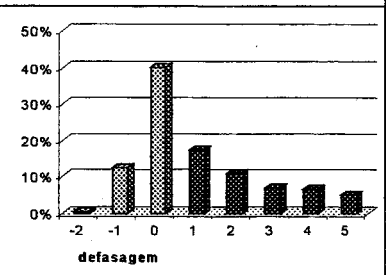
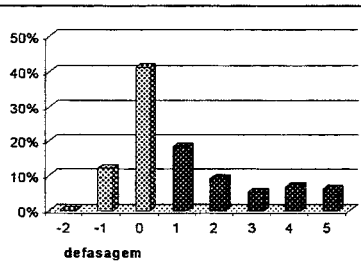
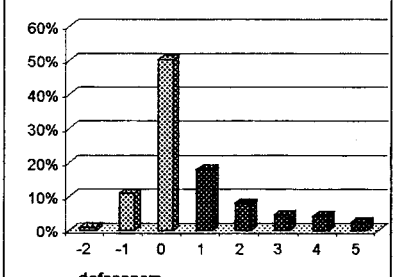
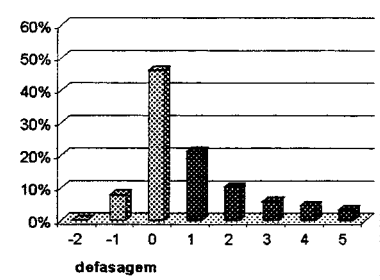
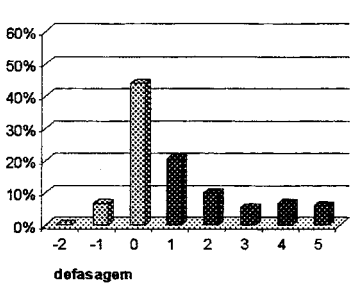
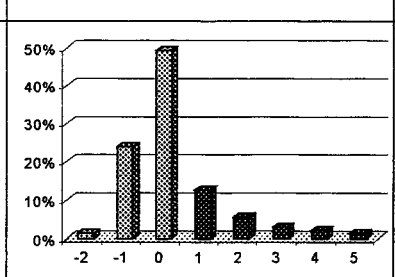
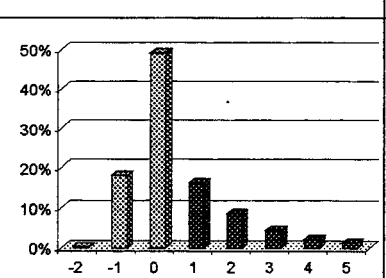
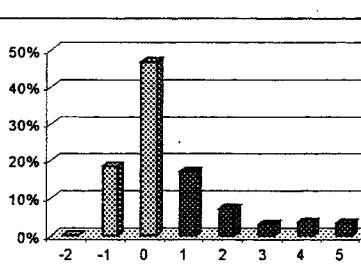
Fonte: PNAD/IBGE, diversos anos. Os valores correspondentes aos anos de 1991 e 1994 foram obtidos por interpolação logística.

Apesar da consistente melhoria nos indicadores educacionais, no ensino fundamental (EF) os alunos com defasagem idade-série[2] ainda constituem um contingente de 44% e, no ensino médio (EM), esse número sobe para de 55% (Fonte: INEP/MEC, 1999).

A Figura 1.3 mostra a distribuição dos estudantes com defasagem escolar de acordo com a região e séries selecionadas[3]. O Nordeste é a região que apresenta a maior proporção de alunos com defasagem. Apesar de a comparação não ser estrita[4], a Região Norte ocupa um segundo lugar bem próximo, com redução da diferença nas séries mais altas. Na 3ª série do EM, o Norte apresenta, relativamente ao Nordeste, percentual mais elevado de alunos nesta situação. O Centro-Oeste e o Sudeste apresentam proporções muito semelhantes de estudantes com defasagem idade-série. A diferença é de 0,2% a favor do Sudeste, na 4ª série, do EF e de 2,9% na 3ª série do EM. O Sul apresenta os melhores números. Em todas as regiões, a proporção de alunos com defasagem cresce com a série.

[2] Alunos com idade acima da adequada para a série que freqüentam.
[3] Fonte: SAEB/1999. Séries finais de cada ciclo: quarta e oitava do ensino fundamental e terceira do ensino médio..
[4] Visto que os dados da Região Norte só incluem área urbana, o que já foi anteriormente assinalado.

Figura 1.3 - Distribuição dos alunos por defasagem idade-série (em anos)– análise por região para os graus selecionados

Região	4ª	8ª	11ª
Norte			
Nordeste			
Centro-Oeste			
Sudeste			
Sul			

Existem três causas para o fenômeno da defasagem idade-série: entrada tardia no sistema educacional, repetência e reingresso após evasão (Teixeira de Freitas (1947), Schiefelbein (1975), Schiefelbein & Wolf (1993), e Klein & Ribeiro (1991)).

A entrada tardia na escola vem sendo reduzida por programas especiais de incentivo promovidos pelos governos federal, estadual e municipal.

A repetência e a evasão escolar são fenômenos correlacionados, visto que a maior parte dos alunos que abandonam a escola fazem-no por desalento, após sucessivos fracassos. Os malefícios da repetência têm repercussões individual e coletiva. As individuais resumem-se ao estímulo à evasão e fomento do autoconceito negativo, enquanto as coletivas dizem respeito ao congestionamento do sistema e desperdício de recursos (Almeida Júnior, 1957; Leite, 1959; Mello *et al.*, 1991; Klein e Ribeiro, 1991).

O maior desafio para o sistema educacional brasileiro, atualmente, é corrigir a defasagem idade-série sem perda da qualidade na educação. A solução vem sendo implementada através de dois tipos de experiência: regime de organização alternativo ao ensino seriado[5] e as classes de aceleração. A reorganização progressiva das escolas no que diz respeito à avaliação/aprovação dos alunos segue uma tendência de implementação de políticas de não-repetência e, para o seu êxito, precisa ser acompanhada não só da reformulação dos currículos, mas também da capacitação de professores (até mesmo para lidar com turmas homoetária e heterogênea quanto ao desenvolvimento cognitivo), etc. Essas medidas estão exaustivamente enunciadas em Almeida Júnior (1957), Leite(1959), Silva e Davis(1993), Mainardes (2001), entre outros.

1.2. Décadas importantes: os anos 1950, 1970 e 1990

A discussão sobre a promoção automática e a repetência não é recente. No final dos anos 1950, ela foi bastante acalentada por peritos internacionais e apontada como a solução para os sistemas educacionais dos países do terceiro mundo. Politicamente

[5] Experiências designadas por promoção automática, progressão continuada, avaliação por ciclos, etapas ou fases, entre outras.

falando, a solução foi abraçada pelo Presidente Juscelino Kubitschek, que logo a incorporou ao seu discurso.

Nos anos 1970, implantaram-se experiências pioneiras de promoção automática em São Paulo (1968), Santa Catarina (1970) e Rio de Janeiro (1979). Infelizmente, não se tomaram as medidas complementares que garantissem o seu sucesso, e a experiência falhou.

Ao longo dos anos 1980, nos Estados de São Paulo (1984), Minas Gerais (1985) e Paraná (1988), houve uma proposta coerente e positiva de promoção automática com o surgimento do Ciclo Básico de Alfabetização. Este consistiu em "eliminar a reprovação no final da primeira série, ampliando o período de alfabetização e assegurando a continuidade desse processo; mudar o enfoque da avaliação, que deveria centrar-se no processo de aprendizagem, indicando o progresso do aluno e dando informações sobre as necessidades de reforço e atendimento de dificuldades específicas; (...); capacitar os professores que atuavam na proposta; alterar a concepção e a prática de alfabetização" (Mainardes, 2001, p.44). Existem ainda, entretanto, alguns professores, pesquisadores e autoridades da educação que têm dúvidas quanto aos resultados desse sistema de promoção. Importa pois monitorar os resultados da implementação da promoção automática para evitar possíveis implicações negativas apontadas anteriormente neste documento e também enunciadas por aquele autor, tais como:

a) A promoção automática pode atenuar as taxas de repetência e atender a interesses economicistas sem qualquer preocupação com a aprendizagem; deve haver a garantia de que a promoção formal corresponde efetivamente à promoção real;
b) As mudanças na administração e a falta de sustentação desse tipo de programas podem contribuir para desacreditar o sistema; e
c) A falta de trabalho coletivo na escola bem como a falta de estratégia de supervisão, apoio docente e projetos pedagógicos consistentes, podem tornar impraticável a consolidação da promoção automática.

1.3. O objetivo da pesquisa e a organização do artigo

Este trabalho tem por finalidade investigar o impacto das políticas de não-repetência no desempenho acadêmico do aluno e, em particular, avaliar alguma evidência de que alunos com defasagem idade-série, em escolas com promoção automática, têm desempenho acadêmico diferenciado dos demais. Para tal foram aplicados modelos de regressão multinível aos dados do Sistema Nacional de Avaliação da Educação Básica - SAEB.

O documento está organizado da seguinte forma: na seção 2 apresentamos uma breve descrição do SAEB, bem como a análise exploratória das covariáveis utilizadas no modelo. A especificação do modelo multinível está apresentada na seção 3. A seção 4 contém o resultado do ajuste dos modelos, e as conclusões, finalmente, aparecem na seção 5.

2. Avaliação do sistema educacional brasileiro e dados

2.1. Descrição do sistema de avaliação e variáveis selecionadas

O estudo foi realizado pela aplicação de modelos de regressão multinível aos dados do Sistema Nacional de Avaliação da Educação Básica - SAEB coletados em 1999, juntamente com algumas variáveis extraídas do Censo Escolar do mesmo ano. O SAEB é um levantamento amostral, em ampla escala, realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais - INEP/MEC. Esse levantamento tem representatividade estadual e nacional, com base em uma amostra de alunos das 4ª e 8ª séries do ensino fundamental e da 3ª série do ensino médio (que doravante designaremos por 11ª série). O plano amostral do SAEB se dá em dois estágios. No primeiro, selecionam-se as escolas[6] e, dentro de cada escola, uma ou duas turmas, dependendo do tamanho da escola (Bussab *et al.*, 2000). Em cada turma, alocam-se sistematicamente os alunos aos

[6] Não estão incluídas na amostra nem escolas com turmas multisseriadas nem sob administração federal. Só foram consideradas escolas rurais no subuniverso da 4ª série, na Região Nordeste, e nos Estados de Minas Gerais e do Mato Grosso do Sul.

testes das diferentes disciplinas[7]. O plano amostral está integralmente descrito nos relatórios técnicos Andrade *et al.*(1999), Bussab *et al.*(1999), Silva *et al.* (2000), e resumido na seção 2 de Ferrão, Beltrão e Fernandes (2002).

Este estudo restringe-se à Região Sudeste do Brasil (Estados de São Paulo, Rio de Janeiro, Espírito Santo e Minas Gerais). Utilizam-se os dados referentes aos alunos da 4ª série de suas respectivas escolas, estando envolvidos 16 066 alunos de 514 escolas. Nos dados da subamostra de Minas Gerais, há 195 escolas, e nos de São Paulo, 88. O percentual de alunos em turmas com promoção automática é de 35,6% (1 862 alunos) em Minas Gerais e de 62,2% (2 276 alunos) em São Paulo[8].

A variável de interesse é a proficiência em Matemática, Ciências, Língua Portuguesa, História e Geografia. A estimativa de proficiência está baseada em modelos da teoria de resposta ao item. Essa metodologia torna possível a classificação dos alunos das 4ª, 8ª e 11ª séries em uma só escala. A escala de proficiência varia de 0 a 500 pontos.

As variáveis explicativas usadas nos modelos especificados adiante são mensuradas em dois níveis: Nível 1(alunos) e Nível 2 (escola/turma/professor). As variáveis do Nível 1 são:

• "raça/cor" é uma variável categórica nominal para mulato/pardo, amarelo, indígena, negro e branco. É codificada como um conjunto de variáveis indicadoras tendo como nível de referência o branco;

• "defasagem idade-série" é uma das variáveis explicativas de interesse. A defasagem foi computada como a diferença da idade do aluno e a idade adequada para série que cursa (sete anos completos até julho do ano da matrícula[9]). A idade foi computada utilizando-se a data de nascimento. Além dessa, foi criada uma variável indicando defasagem negativa (alunos matriculados na série com idade abaixo da recomendada). Os gráficos do anexo mostram a distribuição de freqüência dessa

---

[7] Na 4ª e 8ª séries do Ensino Fundamental são testadas as disciplinas de Matemática, Ciências, Geografia, História e Português. Na 3ª série do Ensino Fundamental as disciplinas de Física e Química substituem Ciências.
[8] Segundo o Censo Escolar 1999, a distribuição das escolas por regime de organização de ensino indica números diferentes: 45% de escolas em Minas Gerais estão organizadas somente em ciclos, e em São Paulo esse número atinge 70%.
[9] No Censo Educacional, computa-se a defasagem idade-série considerando-se a idade do aluno em 31 de dezembro.

variável nas áreas de estudo para algumas desagregações de interesse: nível socioeconômico e existência de promoção automática; e

• "nível socioeconômico da família do aluno" - classificação do *status* socioeconômico da família calculado a partir de variáveis primárias, tais como a educação dos pais e a posse de bens *(freezer,* refrigerador, máquina de lavar roupa, automóvel, etc.) de acordo com o índice sintético proposto pela Associação Brasileira dos Estudos de Mercado (Jannuzzi, 2001; pág. 99-100). A escala[10] varia de A a E, indo das classes mais afluentes para as menos favorecidas; nome da variável, "NSE".

As variáveis de Nível 2 são:

• "nível socioeconômico médio da escola" é uma variável contextual criada a partir da média do nível socioeconômico dos alunos da escola; nome da variável, "NSE-escola";

• o regime de organização do ensino é uma das variáveis de interesse e é do tipo categórica nominal, operacionalizada através de variáveis indicadoras - "promoção automática" para as escolas com esse regime, "misto"[11] para as escolas com classes em regime seriado e classes em regime de promoção automática; o regime seriado foi considerado o nível de referência. Essa variável foi retirada do Censo Escolar 1999; e

• "pública" é uma variável indicadora para as escolas sob esse tipo de administração.

2.2. Análise exploratória das variáveis envolvidas

Nesta seção, apresentar-se-á uma breve análise exploratória das variáveis utilizadas na modelagem.

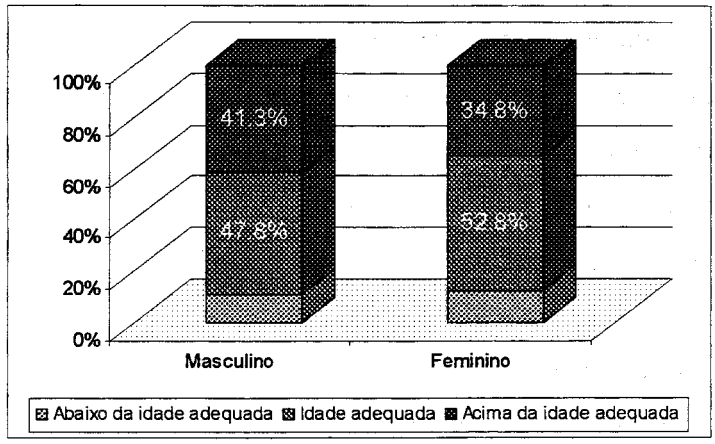
Para se obter maior facilidade na análise gráfica, a variável defasagem idade-série vem categorizada em três níveis: aluno abaixo da idade adequada, na idade adequada e acima da idade adequada. A Figura 2.1 mostra a distribuição dos alunos por esses três

[10] Na seção 4, essa variável foi recodificada por uma variável ordinal que vai de 1, muito pobre, a 7, muito rico.

[11] O questionário de turma do SAEB 1999 não inclui a informação relativa ao regime de promoção. Assim, essa informação foi retirada do Censo Escolar para cada umas das escolas da amostra. Nos casos em que a escola declara a existência de ambos os regimes de promoção (aqui designado por regime misto de promoção), torna-se impossível determinar qual o regime a que turma amostrada está sujeita (na edição 2001 do SAEB essa informação foi coletada no questionário relativo à turma).

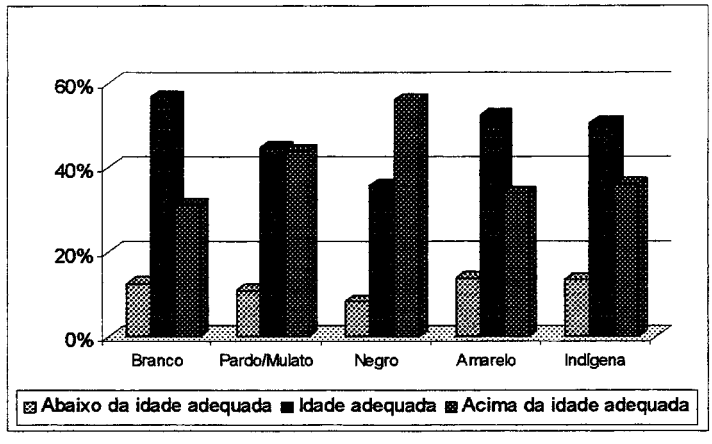
níveis segundo o sexo. Pode-se observar que os meninos apresentam maior percentual na idade acima da adequada.

Figura 2.1- Defasagem por sexo



A Figura 2.2 mostra a relação entre a defasagem idade-série e raça/cor autodeclarada. Há maior percentagem de alunos negros acima da idade adequada do que em qualquer outro grupo. Assim, 56% dos alunos negros estão nessa condição enquanto os alunos brancos atingem o total de 31%.

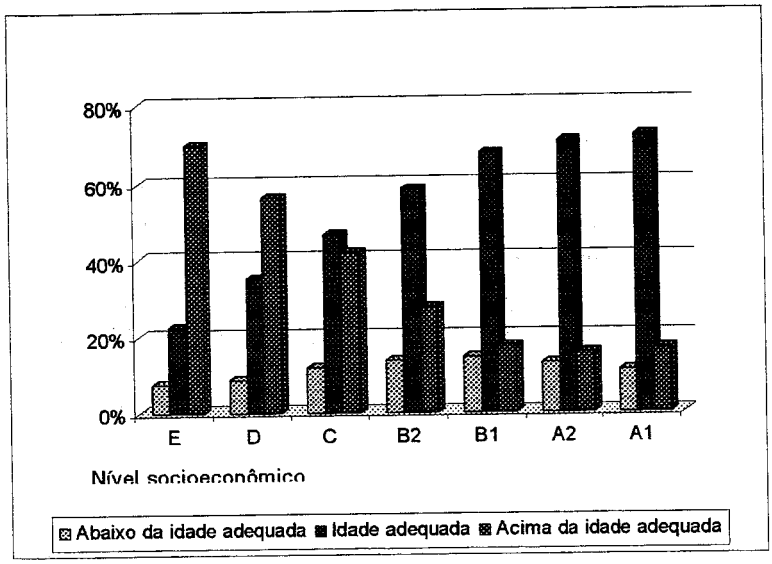
Figura 2.2 – Defasagem por raça/cor



A Figura 2.3 mostra que existe correlação negativa entre a defasagem idade-série e o nível socioeconômico do aluno. As classes socioeconômicas menos favorecidas apresentam maior proporção de alunos de idade mais elevada. O anexo apresenta números equivalentes (Figuras A1 e A2) para São Paulo e Minas Gerais, bem como se

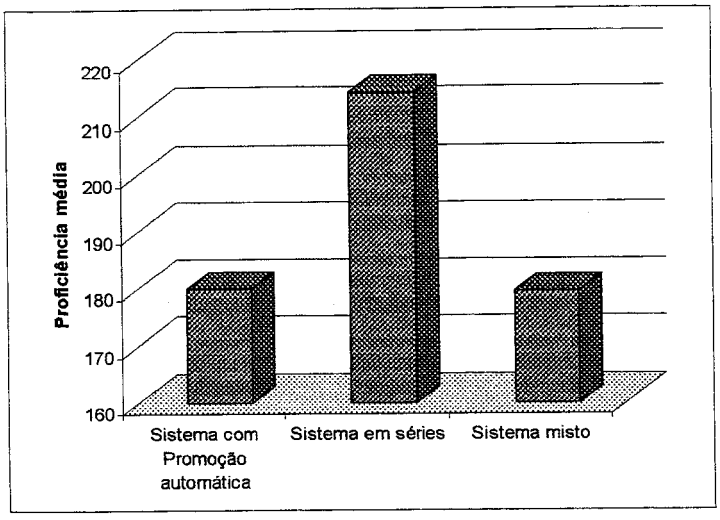
apresenta a relação entre a proporção de alunos a estudar em escolas com sistema de promoção automática e o nível socioeconômico dos alunos (A3 e A4).

Figura 2.3 – Defasagem por nível socioeconômico



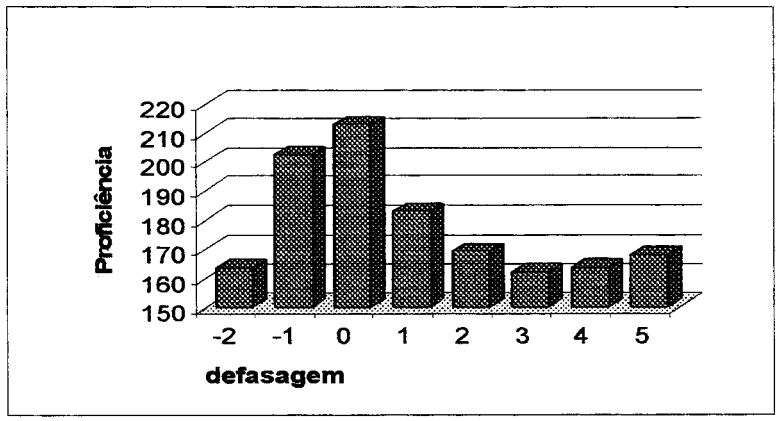
A Figura 2.4 mostra a média da proficiência por regime de organização do ensino e nela se pode observar que a média da proficiência dos alunos que freqüentam escolas com promoção seriada é superior em mais de 30 pontos do que a proficiência média dos que freqüentam escolas com promoção automática.

Figura 2.4 – Proficiência por regime de organização do ensino



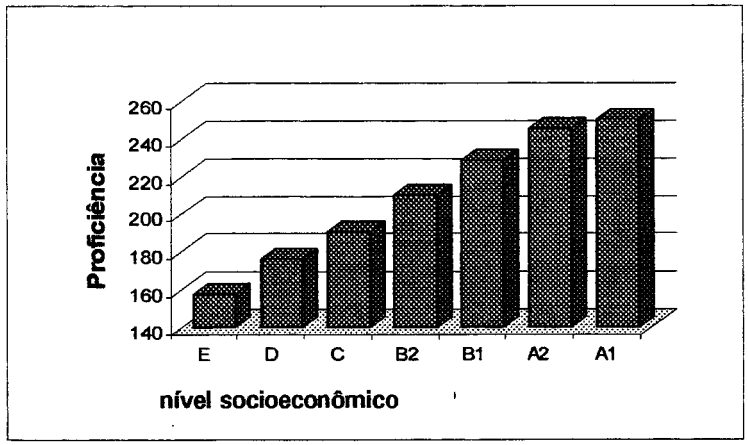
A Figura 2.5 mostra a relação entre proficiência e defasagem idade-série. Podemos verificar que os alunos abaixo da idade adequada e os que estão acima dela têm a sua proficiência reduzida, quando comparada com a dos alunos em idade adequada. FERRÃO e BELTRÃO (2001) estudaram o impacto da idade sobre o desempenho do aluno. Os autores mostram que a defasagem idade-série apresenta um efeito quadrático[12] sobre a proficiência e que existe um efeito escolar estatisticamente significativo da defasagem idade-série sobre a proficiência. Isto significa que, em algumas escolas, o desempenho dos estudantes com idade acima da adequada é mais penalizado do que em outras.

Figura 2.5 - Proficiência por defasagem idade-série



A Figura 2.6 mostra a relação entre a proficiência e o nível socioeconômico, sendo evidente a associação entre estas variáveis.

Figura 2.6 - Proficiência por nível socioeconômico



[12] Para os alunos com defasagem idade-série positiva.

3. Modelo multinível

3.1. Especificação do modelo

Os dados de avaliação educacional apresentam estrutura hierárquica, pois os alunos encontram-se agrupados em turmas, as turmas em escolas, as escolas em municípios, e assim por diante. A mensuração pode ser realizada para captar as características de alunos, turmas e/ou professores, e escolas. O modelo de regressão multinível[13] incorpora naturalmente a estrutura hierárquica ou de agrupamento dos dados e, por conseguinte, da população em estudo.

Na modelagem de dados educacionais, a aplicação de modelos de regressão multinível tem duas vantagens em face dos modelos de regressão clássica. Na presença de correlação intraclasse, a estimação dos parâmetros do modelo via regressão clássica produz estimativas do erro padrão muito pequenas. As estimativas produzidas pelos modelos de regressão multinível são geralmente mais conservadoras. Por outro lado, ao decompor a variância do erro segundo os níveis hierárquicos, o modelo de regressão multinível permite ao analista a melhor compreensão e/ou explicação do processo que está a modelar. Torna-se mais simples, por exemplo, estudar a capacidade explicativa de variáveis intra-escolares em face das extra-escolares, ou de variáveis passíveis de intervenção direta. Usualmente os pesquisadores estão interessados em saber como é que a estrutura organizativa da escola (no nosso caso, o regime de organização do ensino) influencia o desempenho acadêmico dos alunos, ou saber como é que as características do professor (experiência, estilo pedagógico, etc.) têm impacto na aprendizagem. Exemplos clássicos deste tipo de análises são Cronbach *and* Webb(1975), Burstein *et al.*(1978) e Longford (1986).

Além de permitir a correta análise de contexto (com eventual efeito de interação do grupo nos indivíduos, isto é, interação entre as variáveis da escola e dos alunos), o Modelo trata as escolas como uma amostra extraída da população de todas as escolas, com determinada distribuição de probabilidade subjacente.

[13] Também conhecidos como modelos lineares hierárquicos ou modelos de coeficientes aleatórios.

O modelo de dois níveis[14] considera "aluno" como a unidade do nível 1 e "escola" como a unidade do nível 2. Os alunos são identificados pelo índice i, e as escolas, pelo índice k. O índice k varia de 1 a K e o índice i varia de 1 a n_k, onde n_k representa o número de alunos da escola k.

A variável resposta é *proficiência*$_{ik}$, que representa o desempenho escolar do aluno i na escola k, e as variáveis explicativas são, por exemplo, a defasagem idade-série do aluno ik, *defasagem*$_{ik}$, e a variável indicadora de regime de organização do ensino em promoção automática, *prom_aut*$_k$. A primeira é mensurada no nível 1, aluno, e a segunda no nível 2, escola.

O modelo de dois níveis para a *proficiência* do i^{ésimo} aluno na k^{ésima} escola apresenta-se em seguida:

$$proficiencia_{ik} = \beta_{ok} + \beta_{1k} defasagem_{ik} + e_{ik} \quad (1)$$

$$\beta_{0k} = \beta_{00} + \beta_{01}\text{prom\_aut}_k + u_{ok} \quad (2)$$

$$\beta_{1k} = \beta_{10} + \beta_{11}\text{prom\_aut}_k + u_{1k} \quad (3)$$

$$e_{ik} \sim N(0, \sigma_e^2)$$

$$u_{0k} \sim N(0, \sigma_{u0}^2)$$

$$u_{1k} \sim N(0, \sigma_{u1}^2)$$

$$\text{cov}(e_{ik}, u_{ok}) = \text{cov}(e_{ik}, u_{1k}) = 0$$

$$\text{cov}(u_{0k}, u_{1k}) = \sigma_{u01}$$

Observe-se que tanto o intercepto como o coeficiente de inclinação são parâmetros aleatórios, variando de escola para escola (têm associado o índice k). O erro de nível 1 é e_{it}, e u_{0k} e u_{1k} são os erros de nível 2, associados ao intercepto e à inclinação, respectivamente. Pressupõem-se que o erros seguem distribuição normal com média *0* e variâncias σ^2_e, σ^2_{uo}, σ^2_{u1}. A covariância entre o intercepto e o coeficiente de inclinação é dada por σ_{uo1}. O parâmetro de variância σ_e^2 representa a variabilidade intra-escolar, enquanto σ_{u0}^2 e σ_{u1}^2 representam variabilidade entre escolas.

[14] Poderíamos considerar um nível intermédio na hierarquia para representar o efeito da turma. No entanto, devido às características do plano amostral do SAEB (dados usados), que contempla poucas turmas por escola, essa abordagem torna-se inviável.

Pressupõe-se que os erros de nível 1 e 2 sejam não-correlacionados entre si. A equação (4) decorre da substituição de (2) e (3) em (1). A primeira linha do lado direito da equação (4) é a componente preditiva ou parte determinística do modelo, e a segunda linha é a parte aleatória ou estocástica. O modelo é dado por

$$\begin{aligned} proficiencia_{ik} &= \beta_{00} + \beta_{10} defasagem_{ik} + \beta_{01} prom\_aut_k + \beta_{11} prom\_aut_k \times defasagem_{ik} + \\ &+ u_{0k} + u_{1k} defasagem_{ik} + e_{ik} \end{aligned} \quad (4)$$

onde β_{00} representa a média global da proficiência controlada por "defasagem" e "promoção automática"; β_{10}, β_{01} são os efeitos principais das variáveis "defasagem" e "promoção automática" na proficiência. β_{10} representa, em média, o efeito marginal na proficiência do aluno ik devido a cada ano de defasagem idade-série; β_{10} representa, em média, o efeito marginal na proficiência do aluno ik em virtude de o regime de organização do ensino na escola que ele frequenta ser "promoção automática"; β_{11} é o coeficiente associado ao termo de interação entre "defasagem" e "promoção automática" e representa o efeito marginal, por cada ano de "defasagem" idade-série, na proficiência do aluno que estuda numa escola com "promoção automática". O referido termo é um dos que permitem verificar a existência do efeito do grupo no indivíduo. No caso, do efeito duma variável que diz respeito a toda a escola (poderíamos classificá-la como uma variável de gestão/administração escolar) no desempenho individual.

O termo $u_{1k}defasagem_{ik}$ representa a ação contextual da escola, no aluno ik com *defasagem* escolar, que não é captada pelas variáveis incluídas na componente determinística do modelo.

Os parâmetros fixos - β_{00}, β_{10}, β_{01}, β_{11}, e aleatórios - σ_e^2, σ_{u0}^2, σ_{u1}^2, σ_{u01} são desconhecidos e estimados a partir dos dados. Dependendo do programa computacional utilizado na estimação destes modelos (MlwiN, HLM, ou outros), o procedimento de estimação pode ser de mínimos quadrados generalizados iterativos ou de máxima verossimilhança. Quando a variável resposta segue distribuição normal, as estimativas produzidas são equivalentes (Goldstein, 1986).

Quando os modelos de regressão multinível são aplicados a dados mensurados em unidades amostrais com diferentes probabilidades de seleção, os procedimentos de

estimação acima mencionados podem conduzir a estimativas do erro padrão viesadas, a menos que a informação subjacente à probabilidade de seleção seja incorporada como covariáveis do modelo Pfeffermann *et al.*(1998). Estes autores propõem duas abordagens para a correção do viés: o método designado por *probability-weighted IGLS* (mínimos quadrados generalizados iterativos ponderado pela probabilidade) que segue os princípios de estimação de pseudo-verossimilhança, e o escalonamento dos pesos com duas formas possíveis. O MLwiN implementa o segundo método de escalonamento.

À semelhança do que usamos no trabalho empírico, nesta seção apresentamos o modelo com apenas dois níveis, mas a sua extensão para três ou mais níveis é direta. Podem encontrar-se maiores detalhes sobre modelos de regressão multinível em BRYK & Raudenbusch (1992), Longford (1993), Goldstein (1995) ou Kreft e de Leeuw(1998).

3.2. Especificação do modelo para os dados do SAEB

Esta subseção descreve cada um dos termos incluídos nos modelos cujos resultados são apresentados e comentados neste artigo.

3.2.1. Intercepto

Como descrito anteriormente, dentro de cada turma, os alunos são alocados a um teste de uma das seguintes disciplinas: Matemática, Ciências, Geografia, História ou Português. Os modelos captam as diferenças de escala inerentes a cada disciplina através do ajuste em separado dos interceptos. Os parâmetros aleatórios são considerados nos níveis 1 e 2 do modelo, conforme apresentado na subseção anterior.

3.2.2. Controle socioeconômico e raça/cor

Geralmente, os alunos de estratos sociais mais baixos apresentam proficiência mais baixa e maior defasagem idade-série. Para controlar esses efeitos, incluímos um indicador do nível socioeconômico da família do aluno e um do contexto, além da informação de "raça/cor" do aluno.

Ferrão *et al.* (2002) mostram que o modelo nulo de dois níveis para a Região Sudeste apresenta coeficiente de correlação intra-escola de 35%. Após o controle pelo *status* socioeconômico, o valor cai para 12,1%. Essas estimativas tornam evidente o hiato socioeconômico entre as famílias do Sudeste. Os modelos apresentados e discutidos naquele documento sugerem que o desempenho dos alunos negros é menor que o dos demais grupos, o que é confirmado pelos resultados das Tabelas 4.1 e 4.2. Dado que os negros, no Brasil, são em média mais pobres e de menor escolaridade, os pesquisadores admitem a hipótese de que a magnitude e a significância da estimativa associada à raça/cor se devem à inadequação do controle da variável socioeconômica utilizada. No entanto, ainda está em curso trabalho adicional sobre o assunto, nomeadamente a investigação da existência de discriminação por raça/cor em sala de aula.

3.2.3. Efeito da idade do aluno

O impacto da defasagem idade-série do aluno sobre a proficiência é modelado da forma proposta anteriormente pelos autores[15]. Isto significa que o modelo contém um polinômio de 2º grau para a variável defasagem idade-série com um termo aleatório associado ao coeficiente linear. Adicionalmente, inclui uma função indicadora que assinala defasagem idade-série negativa.

3.2.4. Variáveis da escola e interação

Além da variável contextual socioeconômica da escola, outras variáveis deste nível são consideradas nos modelos: o sistema de promoção (variáveis indicadoras do regime de promoção - automático e misto, com nível de referência regime seriado) e tipo de administração (público[16] *versus* privado).

---

[15] Ver FERRÃO e BELTRÃO (2001).

[16] O efeito da administração municipal não é estatisticamente diferente do efeito da estadual. Na amostra da Região Sudeste, há 200 escolas sob administração municipal e 114 sob administração estadual.

O modelo inclui os termos de interação entre o sistema de promoção (automática) e a escola pública, bem como entre a defasagem idade-série e o sistema de promoção. Através desses termos de interação, pretende-se testar se o impacto do sistema de promoção na proficiência dos alunos depende ou não do tipo de administração da escola. Com o termo de interação entre defasagem idade-série e sistema de promoção, procura-se verificar se os alunos com defasagem que estudam em escolas com sistema de promoção automática têm ou não o seu desempenho reduzido.

Foram testadas outras interações, como entre o NSE do aluno e o sistema de promoção que testaria se alunos de classes sociais menos favorecidas teriam sua proficiência impactada pelo regime de promoção de forma diferente dos alunos de classes mais afluentes. Essas estimativas, porém, não se apresentaram estatisticamente significativas.

4. Resultados

Todos os cálculos foram realizados utilizando-se o MlwiN 1.1 (RASBASH *et al.* 2000) e o procedimento de estimação utilizado foi o IGLS. O plano amostral não foi considerado nas estimativas que se apresentam[17]. Os modelos foram ajustados para a Região Sudeste (resultados na Tabela 4.1) e em separado para os Estados de São Paulo (resultados na Tabela 4.2) e Minas Gerais (resultados na Tabela 4.3). As Tabelas 4.1, 4.2 e 4.3 apresentam as estimativas pontuais e respectivos erros padrão para os parâmetros fixo e aleatório. O modelo 1 contém apenas variáveis explicativas associadas aos alunos (nível 1). Os modelos 2 e 3 contêm todas as variáveis explicativas descritas anteriormente, à exceção da variável contextual socioeconômica que não está incluída no modelo 2.

A Tabela 4.1 mostra os parâmetros estimados e o erro padrão correspondente para os modelos propostos. Em todos os modelos, as estimativas dos parâmetros da defasagem idade-série confirmam os resultados obtidos em Barbosa e Beltrão (2001).

[17] Foi feita a estimação considerando o plano amostral, através do 2º procedimento de escalonamento dos pesos (PFEFFERMAN *et al.* 1998) e os resultados obtidos confirmam as conclusões substantivas deste artigo.

Os modelos ajustados sugerem que, na Região Sudeste, os alunos do sistema de promoção automática apresentam, em média, uma proficiência mais baixa (nos modelos 2 e 3). Nas escolas públicas, o sistema de promoção não apresenta efeito estatisticamente significativo, já que o termo da interação (sistema de promoção automática e administração pública) é de mesma magnitude e de sinal contrário ao efeito principal (sistema de promoção automática).

Ao comparar os resultados dos modelos 2 e 3, nota-se uma correlação entre o tipo de administração da escola e a variável do contexto socioeconômico da mesma (NSE-escola). Quando se introduz esta última variável no modelo 3, a estimativa do efeito da primeira decresce, ainda que o mesmo continue estatisticamente significativo. O coeficiente de determinação, condicionado às variáveis de alunos (nível 1) foi calculado para avaliar a capacidade explicativa do modelo devido à inclusão das variáveis relativas à escola (nível 2). Observou-se uma redução de 57% da variância de nível 2.

Os resultados do impacto da promoção automática em São Paulo e Minas Gerais são diferentes, principalmente devido à interferência da variável contextual socioeconômica naquele primeiro estado. Comparando os resultados dos modelos 2 e 3 na Tabela 4.2, observamos que, quando o modelo inclui a variável contextual socioeconômica (modelo 3), todas as estimativas relativas ao sistema de promoção automática e ao tipo de administração da escola tornam-se estatisticamente não significativas.

Tabela 4.1- Estimativas dos modelos ajustados para a Região Sudeste do Brasil

	Modelo 1 Estimativa (e.p.)	Modelo 2 Estimativa (e.p.)	Modelo 3 Estimativa (e.p.)
Parâmetros Fixos			
Ciências	205,0 (1,5)	231,6 (1,7)	214,1 (2,0)
Geografia	222,5 (1,5)	249,1 (1,7)	231,7 (2,0)
História	208,5 (1,5)	235,1 (1,7)	217,7 (2,0)
Português	199,4 (1,5)	225,9 (1,7)	208,4 (2,0)
Matemática	209,1 (1,5)	235,6 (1,7)	218,2 (2,0)
NSE	3,9 (0,4)	2,8 (0,4)	1,6 (0,4)
Abaixo da idade adequada	-5,5 (1,2)	-6,0 (1,1)	-5,7 (1,1)
Defasagem idade-série	-14,8 (0,9)	-13,3 (0,9)	-12,8 (0,9)
Defasagem idade-série ^2	1,9 (0,2)	1,9 (0,2)	1,8 (0,2)
Pardo/Branco	-2,0 (0,8)	-2,0 (0,8)	-1,8 (0,8)
Negro/Branco	-15,4 (1,4)	-15,0 (1,4)	-14,8 (1,4)
Amarelo/Branco	-2,4 (1,7)	-2,4 (1,7)	-2,4 (1,7)
Indígena/Branco	1,4 (2,0)	1,3 (2,0)	1,5 (2,0)
Promoção Automática	---	-27,7 (6,8)	-15,6 (5,8)
Promoção Mista	---	3,3 (2,7)	-3,8 (2,4)
Escola Pública	---	-44,0 (2,6)	-11,5 (3,3)
NSE-escola	---	---	16,1 (1,2)
Promoção Automática x Escola Pública	---	28,2 (7,3)	10,7 (6,3)
Defasagem idade-série x Promoção Automática	---	-2,2 (0,8)	-2,0 (0,8)
Parâmetros Aleatórios			
Nível 2 – Escola			
Intercepto	733,8 (53,3)	316,2 (26,6)	204,7 (19,2)
Defasagem	16,3 (3,8)	13,5 (3,5)	13,8 (3,5)
Intercepto x Defasagem	-58,9 (11,7)	-21,0 (7,9)	-12,6 (6,7)
Nível 1 – Aluno			
Intercepto	1637,8 (20,0)	1632,5 (20,0)	1632,2 (20,0)

Tabela 4.2 - Estimativas dos modelos ajustados para São Paulo

	Modelo 1 Estimativa (e.p.)	Modelo 2 Estimativa (e.p.)	Modelo 3 Estimativa (e.p.)
Parâmetros Fixos			
Ciências	208,5 (3,5)	235,6 (3,9)	206,9 (5,8)
Geografia	227,5 (3,5)	254,8 (3,9)	226,1 (5,8)
História	214,8 (3,5)	241,8 (3,9)	213,1 (5,8)
Português	207,4 (3,5)	234,4 (3,9)	205,7 (5,8)
Matemática	215,2 (3,5)	242,3 (3,9)	213,6 (5,8)
NSE	3,9 (0,8)	3,1 (0,8)	2,1 (0,8)
Abaixo da idade adequada	-6,5 (2,3)	-6,9 (2,3)	-6,9 (2,3)
Defasagem idade-série	-18,4 (2,3)	-17,3 (2,6)	-16,5 (2,7)
Defasagem idade-série ^2	2,3 (0,5)	2,2 (0,5)	2,2 (0,5)
Pardo/Branco	-4,7 (1,8)	-4,1 (1,8)	-3,7 (1,8)
Negro/Branco	-21,5 (3,0)	-21,0 (3,0)	-20,9 (3,0)
Amarelo/Branco	-7,0 (3,7)	-6,9 (3,7)	-6,9 (3,7)
Indígena/Branco	-4,3 (5,0)	-4,4 (5,0)	-3,8 (5,0)
Promoção Automática	---	-31,4 (7,1)	-8,8 (7,0)
Promoção Mista	---	-11,4 (12,1)	-12,0 (9,9)
Escola Pública	---	-35,8 (9,4)	-0,5 (9,7)
NSE-escola	---	---	18,5 (3,0)
Promoção Automática x Escola Pública	---	20,8 (11,6)	-3,2 (10,4)
Defasagem idade-série x Promoção Automática	---	-1,1 (2,3)	-1,6 (2,3)
Parâmetros Aleatórios			
Nível 2 – Escola			
Intercepto	712,5 (119,3)	277,7 (53,2)	176,8 (37,4)
Defasagem	27,3 (11,9)	24,3 (11,1)	25,1 (11,2)
Intercepto x Defasagem	-32,4 (29,9)	-29,7 (19,6)	-25,3 (16,6)
Nível 1 – Aluno			
Intercepto	1749,1 (44,9)	1749,9 (45,1)	1749,1 (45,1)

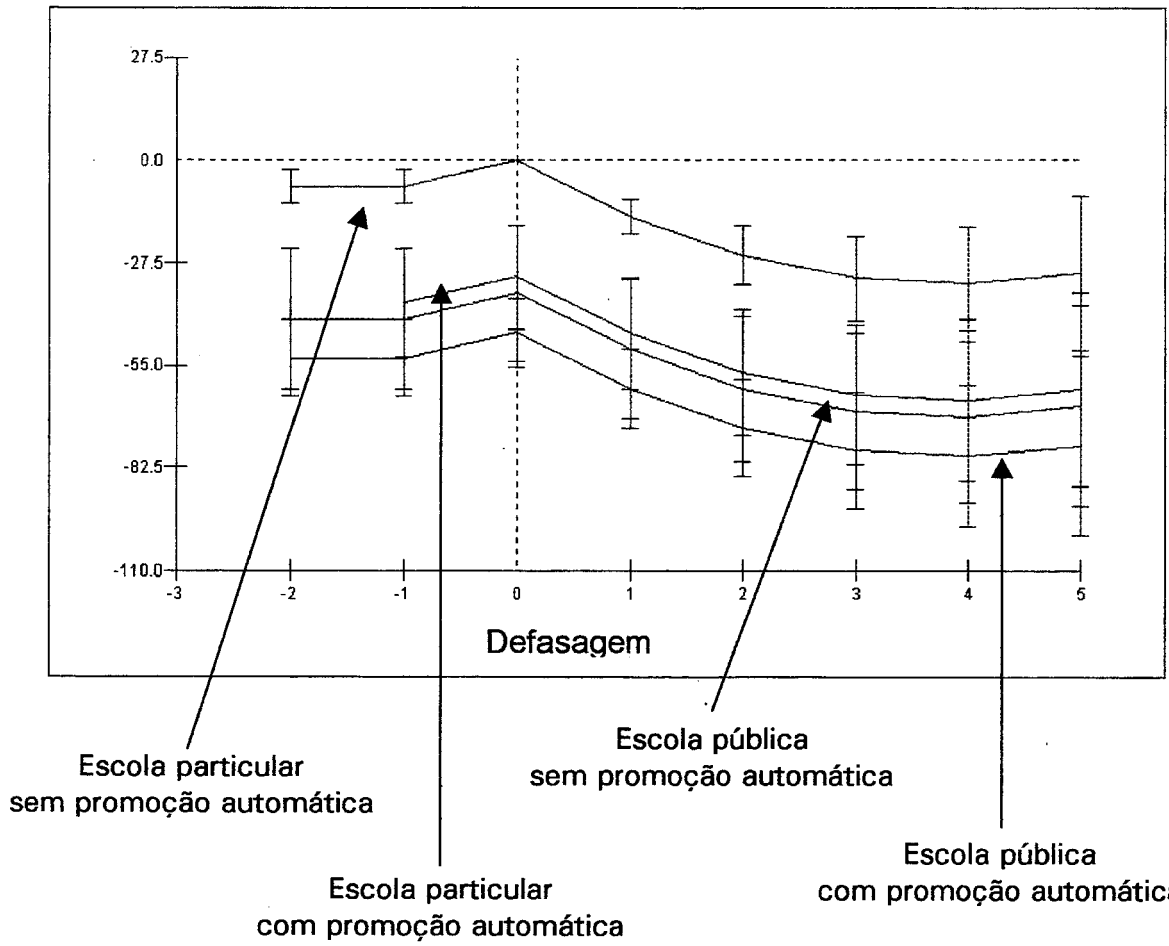
Com relação ao impacto da promoção automática, o modelo 2 é mais elucidativo. O gráfico da Figura 2.6 ilustra os resultados do modelo 2 ajustado para São Paulo. As curvas indicadas na figura são o efeito marginal fixo da variável defasagem idade-série, tipo de administração da escola e sistema de promoção[18] sobre a proficiência. Não há diferença estatisticamente significativa entre o rendimento dos alunos de escolas públicas com e sem promoção automática. Já nas escolas particulares, para os alunos com defasagem menor do que quatro anos, é possível perceber a diferença. O desempenho os alunos em escolas particulares sem promoção automática é superior,

[18] A percentagem de alunos em cada grupo: escola particular sem promoção automática, 26,0%; escola particular com promoção automática, 9,3%; escola pública sem promoção automática, 11,8%; e escola pública com promoção automática, 52,9%.

em média, ao de todos os outros, diferença esta estatisticamente significativa. Além disso, não há diferença entre escolas pública e particular com promoção automática. Os alunos com mais de três anos além da idade adequada, nas escolas particulares sem promoção automática, rendem tanto quanto os outros estudantes em escolas particulares com promoção automática e em escolas públicas sem promoção automática. Existe uma sobreposição dos intervalos de confiança.

Quanto a Minas Gerais, os resultados da Tabela 4.3 sugerem que não existe efeito substancial da promoção automática sobre a proficiência dos alunos. Apesar de o efeito ser estatisticamente significativo ao nível de 10% (mas não de 5%), sua magnitude é muito inferior à média ou mesmo ao efeito marginal decorrente da defasagem de um ano a mais.

Figura 2.6 - Modelo preditivo; efeito marginal de defasagem idade-série, tipo de administração da escola e sistema de promoção



Também quisemos investigar se os alunos mais pobres tinham seu rendimento reduzido quando estudavam em escolas com promoção automática. A interação entre o NSE dos alunos e a promoção automática foi testada nos dois modelos, mas não se encontrou efeito estatisticamente significativo. Concluímos, portanto, que não havia evidência de que os alunos mais pobres das escolas com promoção automática tivessem, em média, menor proficiência. Entretanto, a interação entre a variável contextual socioeconômica e a promoção automática foi positiva e estatisticamente significativa.

Tabela 4.3 - Estimativas dos modelos ajustados para Minas Gerais

	Model 1 Estimate (e.p.)	Model 2 Estimate (e.p.)	Model 3 Estimate (e.p.)
Parâmetros Fixos			
Ciências	198,9 (2,5)	235,1 (3,2)	209,1 (4,0)
Geografia	218,7 (2,5)	255,1 (3,2)	229,2 (4,0)
História	205,1 (2,5)	241,2 (3,2)	215,3 (4,0)
Português	196,2 (2,5)	232,2 (3,2)	206,2 (4,0)
Matemática	208,3 (2,5)	244,4 (3,2)	218,4 (4,0)
NSE	6,5 (0,6)	4,6 (0,6)	3,2 (0,7)
Abaixo da idade adequada	-11,7 (2,5)	-11,6 (2,5)	-11,2 (2,5)
Defasagem idade-série	-15,4 (1,5)	-14,1 (1,6)	-13,9 (1,6)
Defasagem idade-série ^2	2,0 (0,3)	2,0 (0,3)	2,0 (0,3)
Pardo/Branco	0,3 (1,4)	0,2 (1,4)	-0,01 (1,4)
Negro/Branco	-10,1 (2,4)	-10,2 (2,4)	-10,1 (2,4)
Amarelo/Branco	2,7 (3,0)	2,5 (3,0)	2,5 (3,0)
Indígena/Branco	2,7 (3,5)	2,7 (3,5)	2,8 (3,5)
Promoção Automática	---	-0,9 (3,6)	-6,1 (3,2)
Promoção Mista	---	5,7 (8,9)	-0,02 (7,6)
Escola Pública	---	-49,8 (4,0)	-4,2 (6,0)
NSE-escola	---	---	17,4 (1,9)
Promoção Automática x Escola Pública	---	(*)	(*)
Defasagem idade-série x Promoção Automática	---	-1,1 (2,5)	-0,8 (1,2)
Parâmetros Aleatórios			
Nível 2 – Escola			
Intercepto	702,8 (86,3)	297,0 (43,4)	187,6 (31,3)
Defasagem	10,3 (5,5)	7,5 (5,0)	8,9 (5,1)
Intercepto x Defasagem	-58,1 (18,3)	-20,4 (12,3)	-20,1 (10,8)
Nível 1 – Aluno			
Intercepto	1578,6 (34,1)	1563,8 (33,8)	1562,3 (33,7)

(*) – Somente as escolas públicas têm promoção automática.

5. Conclusão

A defasagem idade-série é conseqüência da entrada tardia dos alunos no sistema educativo, da repetência e/ou da evasão com posterior reingresso. As políticas de não-repetência, que vêm sendo implementadas no Brasil, visam a resolver os malefícios individual e coletivo do sistema de ensino baseado na promoção seriada. Há, no entanto, a possibilidade de que a defasagem idade-série seja corrigida, mas com perda de qualidade na educação provida à população. A proficiência dos alunos com defasagem idade-série é inferior comparativamente à dos alunos em idade adequada, mas em algumas escolas o desempenho acadêmico daqueles alunos é mais penalizado do que em outras. Além disso, esse efeito não é linear: as perdas seguem um polinômio de 2º grau quando a defasagem é positiva.

O trabalho descrito neste documento investiga o impacto do regime de organização do ensino (seriado ou com promoção automática) nos resultados escolares dos alunos. Modelos de regressão multinível (nível 1 – aluno, nível 2 – escola) são aplicados aos dados do SAEB - 1999 da Região Sudeste referentes à 4ª série do Ensino Fundamental.

Relativamente às escolas que constituem a amostra, os modelos apresentados sugerem que, nos Estados de São Paulo e Minas Gerais, o regime de promoção automática, pelo menos nas escolas públicas, pode contribuir para a correção da defasagem idade-série sem perda da qualidade na educação. Não foi encontrada evidência de que alunos desfavorecidos socioeconomicamente que estudam em escolas com promoção automática tenham desempenho inferior aos seus colegas. O estudo deve ser refeito com os dados de 2001 e alargado a outras Unidades da Federação que apresentam maiores proporções de alunos com defasagem idade-série.

Os resultados confirmam as evidências relatadas em trabalhos anteriores de que os alunos negros têm rendimento inferior ao de seus colegas. Os autores suspeitam de que essa evidência possa ser atribuída à debilidade do controle da variável socioeconômica utilizada. Investigação adicional está em curso sobre este assunto.

Referências bibliografia

Almeida Júnior (1947). Repetência ou promoção automática? *Revista Brasileira de Estudos Pedagógicos,* 27,65,3-15.

Andrade, D.F., Klein, R. (1999). Métodos estatísticos para avaliação educacional: Teoria da Resposta ao Item. Artigos e Opiniões, *Boletim da ABE*, ano XV, 43

Andrade, D.F., Silva, P.L.N., Bussab, W. (1999). Plano Amostral SAEB-99: Definição do universo a ser investigado. INEP, Relatório Interno

Barbosa, M E Ferrão, Beltrão, K., Farias, M.. Fernandes, C., Santos, 0.(2001). Modelagem do SAEB-99: Modelos multiníveis. Relatório técnico, INEP/MEC.

Barbosa, M.E.Ferrão, Fernandes, C. (2000). A Escola Brasileira az diferença? Uma investigação dos efeitos da escola na proficiência em Matemática dos alunos da 4ª série. Em *Ciclos, Promoção e Avaliação na Educação Brasileira* (ed.) Creso Franco, Porto Alegre, ARTMED.

Beltrão, I. K. e Ferrão, M. E. (2002). Para além das taxas de escolarização. Working paper.

Bryk A., Raudenbush 5. (1992). Hierarchical Linear Models. Sage.

Burstein, L., Linn, R.L. e Capell, F.J. (1978). Analysing multilevel data in the presence of heterogeneous within-class regressions. *Journal of Educational Statistics*, 3, 347-383.

Bussab. W..Andrade, O.F., Silva, P.LN. (1999). Plano Amostral SAEB-99: Definição do plano amostral. Relatório técnico, INEP/MEC.

Cronbach,L.J., Webb, N. (1975). Between class and within class effects in a reported aptitude X treatment interaction: a reanalysis of a study by G.L.Anderson. *Journal of Educational Psycology*, 67, 717-724.

Fernandes, C., Franco, C. (2001). Séries ou ciclos: o que acontece quando os professores escolhem? Em *Avaliação, ciclos e promoção na educação*. (ed.) Creso Franco. Porto Alegre. ARTMED.

Ferrão, M. E, Beltrão, K. (2001). Tracing schools which do not penalize over-age students. Documento apresentado na 27ª Conferência Anual da International Association for Educational Assessment Rio de Janeiro.

Ferrão, M. E., Beltrão, K. I.(2002). Componentes do efeito-escola no Brasil. Working paper.

Ferrão, M. E., Beltrão, K. I, Fernandes, C. (2002). Aprendendo sobre Escola Eficaz – evidências do Saeb-99. INEP/MEC (no prelo).

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, 73: 43-56.

Goldstein H (1995). Multilevel Statistical Models. Edward Arnold.

Hambleton, R.K. (1993). Principles and selected applications of item response theory. Em *Educational Measurement* (ed.) Robert L.Linn, American Council on Education, Oryx Press

Hambleton, R.K.e Swaminathan, H (1985). Item response theory: Principles and aplications. MA:Kluwer Academic Publishers, Boston

INEP (1999). SAEB 97, Primeiros resultados. MEC, Brasília.

Jannuzzi, P.M. (2001). Indicadores Sociais no Brasil. Campinas, Alínea Editora.

Klein, R. and Ribeiro, S.C. (1991). O censo educacional e o modelo de fluxo: o problema da repetência. *Revista Brasileira de Estatística*, 52, 5-45.

Kreft, I. e de Leeuw, Jan(1998). Introducing multilevel modeling. Sage

Longford, N.T. (1985). Mixed linear models and application to schools effectiveness. *Computational statistics quaterly*, 2, 109-117.

Longford N. (1995). Random Coefficient Models. Oxford Ciências Publications.

Mainardes, J. (2001). A organização da escolaridade em ciclos: ainda um desafio para os sistemas de ensino. In *Avaliação, ciclos e promoção na educação*. (Ed.) Creso Franco. Porto Alegre.ARTMED Editora.

Pfeffermann, D., Skinner, C.J., e Holmes, D.J., Goldstein, H. e Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *J.R. Statistical Society* B, 60, 22-40.

Rasbash J., Browne, W., Healy, M., Cameron, B., Charlton, C. (2000). MlwiN. Multilevel Models Project, Institute of Education, University of London.

Schiefelbein, E. (1975). Repeating: an overlooked problem in Latin American Education. *Comparative Education Review*, 19,3, 468-487.

Silva, P.L.N. , Bussab, W., Andrade, D.F. e Freitas, M.P.S.(1999). Plano Amostral SAEB-99: Procedimentos de estimação com a amostra realizada. INEP, Relatório Interno

Teixeira de Freitas M.A. (1947). A escolaridade media no ensino primário brasileiro. *Revista Brasileira de Estatística*, 8, 30/31, 395-474.

Agradecimentos

Os autores agradecem ao INEP pela disponibilização dos dados, aos professores Maria Lígia Barbosa, Ruben Klein e Creso Franco pelos valiosos comentários à versão preliminar deste artigo, ao professor Antônio de Almeida Senna pela revisão do texto e aos avaliadores anônimos as sugestões e comentários que em muito melhoraram o artigo.

Abstract

This paper investigates the impact of the school organisation (automatic promotion) on student performance in the Southeast region of Brazil.

The proportion of overage students in Brazil is 44% at the elementary school level and 55% at the high school level. The existence of overage students can be traced down to three causes: joining the school system late, repetition and return to school after evasion. In order to tone down such figures, aiming at reducing repetition, automatic promotion has been adopted. In the Southeast region of Brazil, The automatic promotion or cycle regime in school organisation is widespread, mainly, in the states of Minas Gerais and São Paulo where more than half the schools adopt such regime.

Our research question is: Does *Automatic Promotion* correct the age-grade gap without loss of quality?

A multilevel model was applied to the Brazilian Educational System Assessment data (some variables were drawn from the Educational Census) and separate models were fitted to São Paulo and Minas Gerais. Results show that the impact of policy on student achievement depends on the state where it has been implemented.

Keywords: Educational assessment, multilevel model, over-age student, automatic-promotion.

REVISTA BRASILEIRA DE ESTATÍSTICA - RBEs

POLÍTICA EDITORIAL

A Revista Brasileira de Estatística - RBEs publica trabalhos relevantes em Estatística Aplicada, não havendo limitação no assunto ou matéria em questão. Como exemplos de áreas de aplicação citamos as áreas de advocacia, ciências físicas e biomédicas, criminologia, demografia, economia, educação, estatísticas governamentais, finanças, indústria, medicina, meio ambiente, negócios, políticas públicas, psicologia e sociologia, entre outras. A RBEs publicará, também, artigos abordando os diversos aspectos de metodologias relevantes para usuários e produtores de estatísticas públicas, incluindo planejamento, avaliação e mensuração de erros em censos e pesquisas, novos desenvolvimentos em metodologia de pesquisa, amostragem e estimação, imputação de dados, disseminação e confiabilidade de dados, uso e combinação de fontes alternativas de informação e integração de dados, métodos e modelos demográfico e econométrico.
Os artigos submetidos devem ser inéditos e não devem ter sido submetidos simultaneamente a qualquer outro periódico.

O periódico tem como objetivo a apresentação de artigos que permitam fácil assimilação por membros da comunidade em geral. Os artigos devem incluir aplicações práticas como assunto central, com análises estatísticas exaustivas e apresentadas de forma didática. Entretanto, o emprego de métodos inovadores, apesar de ser incentivado, não é essencial para a publicação.

Artigos contendo exposição metodológica são também incentivados, desde que sejam relevantes para a área de aplicação pela qual os mesmos foram motivados, auxiliem na compreensão do problema e contenham interpretação clara das expressões algébricas apresentadas.

A RBEs tem periodicidade semestral e também publica artigos convidados e resenhas de livros, bem como incentiva a submissão de artigos voltados para a educação estatística.

Artigos em espanhol ou inglês só serão publicados caso nenhum dos autores seja brasileiro e nem resida no País.

Todos os artigos submetidos são avaliados quanto à qualidade e à relevância por dois especialistas indicados pelo Comitê Editorial da RBEs.

O processo de avaliação dos artigos submetidos é do tipo 'duplo cego', isto é, os artigos são avaliados sem identificação de autoria e os comentários dos avaliadores também são repassados aos autores sem identificação.

INSTRUÇÃO PARA SUBMISSÃO DE ARTIGOS À RBEs

O processo editorial da RBEs é eletrônico. Os artigos devem ser submetidos via *e-mail* para: helem.ortega@ibge.gov.br.

Após a submissão, o autor correspondente receberá um código para acompanhar o processo de avaliação do artigo. Caso não receba um aviso com este número no prazo de uma semana, fazer contato com a secretaria da revista no endereço:

Revista Brasileira de Estatística
IBGE – Diretoria de Pesquisas - Coordenação de Métodos e Qualidade
Av. República do Chile, nº 500, 10º andar
Centro, Rio de Janeiro – RJ
CEP: 20031-170
Tel.: 55 21 2142-0472
55 21 2142-4549
Fax: 55 21 2142-4802

INSTRUÇÕES PARA PREPARO DOS ORIGINAIS

Os originais entregues para publicação devem obedecer às normas seguintes:

1. Originais processados pelo editor de texto *Word for Windows* são preferidos. Entretanto, serão aceitos também, originais processados em LaTeX desde que sejam encaminhados acompanhados de versões em pdf, conforme descrito no item 3 a seguir;

2. A primeira página do original (folha de rosto) deve conter o título do artigo, seguido do(s) nome(s) completo(s) do(s) autor(es), indicando-se, para cada um, a afiliação e endereço para correspondência. Agradecimentos a colaboradores e instituições, e auxílios recebidos, também, devem figurar nesta página;

3. No caso de a submissão não ser em *Word for Windows*, três arquivos do original devem ser enviados. O primeiro deve conter os originais no processador de texto utilizado (por exemplo, LaTeX). O segundo e terceiro devem ser no formato pdf, sendo um com a primeira página, como descrito no item 2, e outro contendo apenas o título, sem identificação do(s) autor(es) ou outros elementos que possam permitir a identificação da autoria;

4. A segunda página do original deve conter resumos em português e inglês (*abstract*), destacando os pontos relevantes do artigo. Cada resumo deve ser digitado seguindo o mesmo padrão do restante do texto, em um único parágrafo, sem fórmulas, com, no máximo, 150 palavras;

5. O artigo deve ser dividido em seções, numeradas progressivamente, com títulos conciso e apropriado. Todas as seções e subseções devem ser numeradas e receber título apropriado;

6. Tratamentos algébricos exaustivos devem ser evitados ou alocados em apêndices;

7. A citação de referências no texto e a listagem final de referências devem ser feitas de acordo com as normas da ABNT;

8. As tabelas e gráficos devem ser precedidos de títulos que permitam perfeita identificação do conteúdo. Devem ser numeradas seqüencialmente (Tabela 1, Figura 3, etc.) e referidas nos locais de inserção pelos respectivos números. Quando houver tabelas e demonstrações extensas ou outros elementos de suporte, podem ser empregados apêndices. Os apêndices devem ter título e numeração, tais como as demais seções de trabalho; e

9. Gráficos e diagramas para publicação devem ser incluídos nos arquivos com os originais do artigo. Caso tenham que ser enviados em separado, devem ter nomes que facilitem a sua identificação e posicionamento correto no artigo (ex.: Gráfico 1; Figura 3; etc.). É fundamental que não existam erros, quer no desenho, quer nas legendas ou títulos.