SECRETARIA DE PLANEJAMENTO, ORÇAMENTO E COORDENAÇÃO
FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA- IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

ISSN 0034 - 7175
R. bras. Estat., Rio de Janeiro, v. 52, n. 197/198, p. 1 - 123, jan. / dez. 1991

REVISTA BRASILEIRA DE ESTATÍSTICA

Órgão oficial do IBGE

Publicação semestral, editada pelo IBGE, 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 3 vias (que não serão devolvidas) para:

Djalma G. Pessoa
Editor-Responsável - RBEs

ENCE
Rua André Cavalcante, 106 - Bairro de Fátima
20231-050 - Rio de Janeiro - RJ

- Os artigos submetidos às RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.
- Cada autor receberá, gratuitamente, 20 separatas de seu artigo.

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

Capa
Pedro Paulo Machado

Revista brasileira de estatística / Fundação Instituto Brasileiro de Geografia e Estatística - v.1, n. 1 (jan./mar. 1940) - . - Rio de Janeiro: IBGE, 1940-

v.

Trimestral (1940-1986), semestral (1987-)
Órgão oficial do IBGE.
Continuação de: Revista de economia e estatística.
Índices acumulados de autor e assunto publicados no v. 43 (1940-1979) e v. 50 (1980-1989)

ISSN 0034-7175 - Revista brasileira de estatística

1 - Estatística - Periódicos. I. IBGE

IBGE. CDDI. Departamento de Documentação e Biblioteca
RJ-IBGE/88-05 Rev.

CDU 31 (05)
PERIÓDICO

Impresso no Brasil/Printed in Brazil

SUMÁRIO

ARTIGOS

O CENSO EDUCACIONAL E O MODELO DE FLUXO: O Problema da Repetência

Ruben Klein*

Sergio Costa Ribeiro*

1 INTRODUÇÃO

Durante as últimas décadas uma polêmica se estabeleceu sobre os indicadores educacionais dos Sistemas de Ensino nos países latino-americanos, em geral, e no Brasil em particular.

Vários tabalhos discutem as possíveis razões de discrepâncias observadas em dois pontos básicos.

1 – O porquê da incompatibilidade entre o número de alunos novos que ingressam, por ano, na 1ª série do Sistema de Ensino e as possibilidades demográficas dos países. Esta discrepância se acentuou nas últimas décadas na medida em que o acesso ao ensino fundamental começa a se universalizar nestes países.

2 – As taxas de repetência nas primeiras séries parecem ser sistematicamente menores que a realidade, e as taxas de evasão sistematicamente maiores.

Vários autores (Teixeira de Freitas, 1947, Schiefelbein, 1975, Cuadra, 1989) analisaram estes indicadores e divisaram uma série de métodos alternativos para corrigi -los. Destas pesquisas surge um cenário bastante diferente daquele obtido com a metodologia tradicional. Do quadro que indicava uma evasão precoce da escola como o principal fator da baixa escolaridade da população, passa-se a um quadro em que a repetência é o principal entrave ao aumento da

* Laboratório Nacional de Computação Científica/CNPq

escolaridade, mudando assim completamente o enfoque do problema. De um cenário de dificuldades exógenas à escola, ditadas por fatores sociais externos, para problemas endógenos ao Sistema de Ensino, as altas taxas de repetência.

O caso do Brasil é particularmente grave. As discrepâncias entre os indicadores oficiais e os trabalhos acima referidos são as mais gritantes entre todos os países analisados.

O objetivo deste artigo é explicar o modelo de fluxo de alunos num sistema seriado de ensino, apontar as inconsistências observadas quando se aplica este modelo aos dados coletados pelos censos escolares do Ministério da Educação do Brasil e esclarecer o motivo destas discrepâncias. Mostramos como as diversas informações colhidas pelos censos, da forma como são tratadas conceitualmente e metodologicamente, produzem estes resultados inconsistentes. Fazemos correções na maneira de tratar os dados do Censo Educacional, obtendo resultados mais corretos na aplicação do Modelo de Fluxo. Apresentamos, no Apêndice 1, sugestões de mudança no questionário do Censo Educacional para corrigir os problemas conceituais encontrados.

Neste processo redescobrimos, independentemente, o que já tinha sido detectado por Teixeira de Freitas (1947) com dados da década de 30. Esta descoberta, como veremos adiante, expõe formas diversas de repetência além daquela clássica da reprovação no fim do ano letivo pelas provas ou pela freqüência escolar.

No Brasil, um método alternativo de obtenção dos indicadores educacionais foi desenvolvido utilizando-se como base de dados, não os Censos Educacionais, mas as Pesquisas Nacionais por Amostra de Domicílio – PNADs – e Censos Demográficos do IBGE. Mostramos que a comparação dos indicadores de evasão e repetência estimados por este método, PROFLUXO (Fletcher & Ribeiro, 1986, 1988, 1989 e Klein, 1991a), com os indicadores obtidos com os dados dos Censos Educacionais após correções que podem ser feitas reinterpretando o significado dos dados desses Censos (Klein, 1991b, c, d), torna-se muito semelhante para as primeiras quatro séries do 1º Grau, como também ficam semelhantes às obtidas por outros trabalhos anteriormente mencionados.

A partir da 5ª série estas correções não conseguem explicar corretamente as discrepâncias observadas. É, no entanto, a partir desta série que o Sistema Supletivo de Ensino, no Brasil, começa a interagir com o sistema formal violando várias hipóteses básicas do Modelo de Fluxo, como a duração de um ano para cada série e a entrada de alunos novos de fora do Sistema Regular de Ensino. Infelizmente nem a base de dados do Censo Educacional nem as PNADs e Censos Demográficos, da forma como são construídos seus instrumentos, permitem superar estas dificuldades. No entanto, algumas inferências podem ser feitas sobre a ordem de grandeza deste problema. No apêndice deste trabalho fazemos sugestões para captar estas interações tanto para os instrumentos do Censo Educacional como para os das PNADs e Censos Demográficos.

2 O MODELO DE FLUXO

O Modelo de Fluxo (Thonstad, 1980) é descrito pelas equações 1 e 2.

$$R_{k,t+1} + I_{k+1,t+1} + E_{k,t} = M_{k,t} \tag{1}$$

$$R_{k,t+1} + I_{k,t+1} = M_{k,t+1} \tag{2}$$

onde temos:

$M_{k,t}$ = Número de matriculados na série k no ano $t, k = 1, \cdots, 9$,[1]

$I_{k,t}$ = Número de ingressos novos na série k (promovidos da série $k-1$) no ano $t, k = 1, \cdots, 9$,

$R_{k,t+1}$ = Número de repetentes na série k no ano $t+1$,

$E_{k,t}$ = Número de evadidos entre a série k e a série $k+1$ no ano $t, k = 1, \cdots, 8$,

Lembramos que um aluno evadido pode ser um aluno aprovado ou um aluno não aprovado, o que nos permite adicionar ao modelo a equação 3:

$$E_{k,t} = E_{k,t,a} + E_{k,t,n} \tag{3}$$

onde $E_{k,t,a}(E_{k,t,n})$ é o número de evadidos aprovados (não aprovados) na série k.

Seja $A_{k,t}$, o número de aprovados na série k no ano t.

Então

$$E_{k,t,a} = A_{k,t} - I_{k+1,t+1} \tag{4}$$

$$E_{k,t,n} = M_{k,t} - R_{k,t+1} - I_{k+1,t+1} - E_{k,t,a} \tag{5}$$

Chamamos a atenção que o número de Graduados no 1º Grau no ano t é $A_{8,t}$

As hipóteses básicas do Modelo de Fluxo são representadas pela equação 2 onde um aluno matriculado no ano $(t+1)$ é um aluno novo ou um aluno repetente e pela equação 1 onde este aluno estava matriculado no ano t. Aluno novo na série k, no ano $(t+1)$ é, por definição, um aluno aprovado na série $(k-1)$, se $k \geq 2$, no ano t e que se matricula pela 1ª vez na série k e, portanto, é um aluno *promovido*. Para a 1ª série, um aluno novo é simplesmente um aluno que se matricula pela 1ª vez nesta série. Portanto, um aluno repetente é qualquer aluno que se matriculou no ano t na série k e se *rematriculou* na mesma série no ano $t+1$.

Para incluirmos alunos que vêm de fora do Sistema Regular de Ensino ou que passam pelo menos um ano fora do Sistema (por exemplo, alunos que vêm do Curso Supletivo), temos que modificar o Modelo de Fluxo, incluindo na equação 2 um termo referente aos alunos de fora do Sistema Regular de Ensino.

---

[1] A 9ª série, $k = 9$, representa a 1ª série do 2º Grau.

É comum representarmos as equações 1, 2, e 3 na Tabela 1 de fluxo entre os anos t e $t+1$, para as 8 séries do 1º grau e a 1ª série do 2º Grau (9ª série) do sistema escolar brasileiro.

TABELA 1

Série no ano t	Série no ano t+1												
	1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	E_a	E_n	(E)	Soma
1ª	$R_{1,t+1}$	$I_{2,t+1}$								$E_{1,t,a}$	$E_{1,t,n}$	$(E_{1,t})$	$M_{1,t}$
2ª		$R_{2,t+1}$	$I_{3,t+1}$							$E_{2,t,a}$	$E_{2,t,n}$	$(E_{2,t})$	$M_{2,t}$
3ª			$R_{3,t+1}$	$I_{4,t+1}$						$E_{3,t,a}$	$E_{3,t,n}$	$(E_{3,t})$	$M_{3,t}$
4ª				$R_{4,t+1}$	$I_{5,t+1}$					$E_{4,t,a}$	$E_{4,t,n}$	$(E_{4,t})$	$M_{4,t}$
5ª					$R_{5,t+1}$	$I_{6,t+1}$				$E_{5,t,a}$	$E_{5,t,n}$	$(E_{5,t})$	$M_{5,t}$
6ª						$R_{6,t+1}$	$I_{7,t+1}$			$E_{6,t,a}$	$E_{6,t,n}$	$(E_{6,t})$	$M_{6,t}$
7ª							$R_{7,t+1}$	$I_{8,t+1}$		$E_{7,t,a}$	$E_{7,t,n}$	$(E_{7,t})$	$M_{7,t}$
8ª								$R_{8,t+1}$	$I_{9,t+1}$	$E_{8,t,a}$	$E_{8,t,n}$	$(E_{8,t})$	$M_{8,t}$
9ª									$R_{9,t+1}$	$E_{9,t,a}$	$E_{9,t,n}$	$(E_{9,t})$	$M_{9,t}$
Novos	$I_{1,t+1}$												
Soma	$M_{1,t+1}$	$M_{2,t+1}$	$M_{3,t+1}$	$M_{4,t+1}$	$M_{5,t+1}$	$M_{6,t+1}$	$M_{7,t+1}$	$M_{8,t+1}$	$M_{9,t+1}$				

3 OS DADOS DOS CENSOS EDUCACIONAIS

Os questionários dos Censos Educacionais no Brasil são preenchidos, em princípio, por todas as escolas do País.

Neste trabalho analisaremos principalmente os dados disponíveis (últimos divulgados) dos anos de 1986 e 1987 (Sinopse Estatística de Ensino Regular de 1º Grau e 2º Grau, MEC, 1986 e 1987). Os dados que interessam à nossa análise contidos nestes questionários são os seguintes:

i) Matrícula inicial, no ano do Censo, por idade, nas faixas disponíveis de: menos de 7 anos, 7 anos, 8 anos, $\cdots$, 15 anos e mais de 15 anos, para as séries 1ª à 8ª do 1º Grau e cursos de 1º Grau não seriados (em alguns estados da Federação a 1ª e 2ª séries do 1º Grau são tratadas como um "ciclo básico" com duração de dois anos);

ii) Matrícula inicial de repetentes no ano do Censo;

iii) Movimento e rendimento escolar, no ano anterior ao Censo, por série, incluindo: afastados por transferência e por abandono, aprovados e reprovados;

iv) Movimento escolar, no ano anterior ao Censo, dos alunos em regime não seriado incluindo alunos afastados por abandono, aprovados e reprovados;

v) Concluintes, no ano anterior ao Censo, por idade, nas faixas de menos de 14 anos, 15 anos e mais de 15 anos.

Os manuais de instruções para o preenchimento dos questionários, nos anos de 1986 e 1987, definem os dados da seguinte maneira:

• **Matrícula inicial** – Matrícula no início do ano letivo, por idade (em anos no dia 30 de abril), excluindo os alunos que se matricularam mas nunca compareceram à escola, alunos de classes especiais e de Supletivo.

• **Repetentes** – Define repetentes como "alunos que voltam a freqüentar a mesma série em que estavam matriculados no ano anterior, por não terem obtido a freqüência mínima ou aproveitamento necessário à aprovação".

• **Matrícula de novos** – A matrícula de alunos novos, na série, é obtida pela diferença entre a matrícula inicial e a matrícula de repetentes.

• **Matrícula final** – A matrícula final é obtida pela soma dos *aprovados* e *reprovados* no final do ano anterior.

• **Afastados por Abandono ou com Matrícula Cancelada** – O número de alunos que abandonaram a escola durante o ano letivo anterior tendo sua matrícula cancelada. Para o Brasil como um todo, deveria corresponder à diferença entre a matrícula inicial e final.

• **Afastados por Transferência** – O número de alunos que foram transferidos para outra escola durante o ano letivo anterior. Para o Brasil como um todo, este número não afeta o total de matrículas.

4 A CRÍTICA DOS DADOS DO CENSO EDUCACIONAL

Para permitir a comparatividade dos dados do Censo Educacional, publicados nas "Sinopses Estatísticas do Ensino Regular de 1º e 2º Grau" com os dados das PNADs e indicadores do PROFLUXO foi preciso:

i) Retirar do cômputo geral do País os dados referentes à Região Norte rural, já que nas PNADs não são coletados dados nesta região e situação de domicílio (em 1980, esta população representava cerca de 2% da população nacional).

ii) Na Sinopse de 1986 não estão incluídos os dados referentes ao Estado de Goiás. Estes dados foram imputados pela média aritmética dos dados correspondentes divulgados nas Sinopses de 1985 e 1987.

iii) Os dados referentes a 1ª e 2ª séries do 1º Grau foram corrigidos para levar em conta os alunos registrados no 1º Grau não seriado, em Minas Gerais, em 1986 e 1987. Mesmo sabendo que a implantação do "ciclo básico" neste estado possa ter alterado, do ponto de vista pedagógico, a distribuição dos conteúdos programáticos nos dois primeiros anos verificou-se que dados de matrícula da 3ª a 8ª séries nas sinopses, em 1983, 1985 e 1986, são muito semelhantes em Minas Gerais, o que justifica, do ponto de vista contábil, atribuir 58% da matrícula desse não seriado à 1ª série e o restante à 2ª série (percentagem observada em 1985). Os dados de

aprovados e reprovados em 1986 nas duas primeiras séries em Minas Gerais foram, também por motivos análogos, imputados pela proporção média encontrada nos dados de 1983 e 1985. Procedemos de forma análoga com todos os dados referentes a não seriado nas primeiras séries encontrados nas sinopses estatísticas do MEC, para outras Unidades da Federação.

5 A APLICAÇÃO DO MODELO DE FLUXO

Com o conjunto de dados assim produzidos aplicamos o modelo de fluxo, para o Brasil como um todo, aos dados dos Censos Educacionais de 1986-1987. A Tabela 2 apresenta os resultados obtidos em números absolutos e em proporções da matrícula em 1986.

Observamos nesta Tabela as seguintes inconsistências:

i) Um número de ingressos novos na 1ª série, em 1987, 4 737 279, cerca de 1,36 vezes a coorte[2] de referência de 7 anos de idade, 3 491 821 crianças, o que representa uma total impossibilidade demográfica como explicaremos, em detalhe, mais adiante na Tabela 5;

ii) Um número de evadidos aprovados (aprovados em uma série em 1986 menos ingressos novos ou promovidos à série seguinte em 1987), equação 5, *negativo*, exceto na 1ª série. Este fato pode ser melhor observado na Tabela 3, onde mostramos o número de alunos aprovados, por série, no ano de 1986, o número de alunos novos, na série subseqüente, no ano de 1987 e o número de evadidos aprovados. É fácil ver que o número de alunos novos em todas as séries, com exceção da 1ª, é superior ao número de aprovados no ano anterior, o que não pode ocorrer já que os alunos novos de uma série só podem advir dos alunos aprovados na série anterior no ano anterior. A possibilidade de que estes alunos estivessem ingressando na escola vindos de outro sistema que não o Sistema Regular de Ensino é impossível, pelo menos nas quatro primeiras séries. O número de evadidos aprovados é positivo na 1ª série devido à inclusão, entre os aprovados na 1ª série em 1986, de alunos que, como veremos mais adiante, vão repetir esta série, mesmo tendo sido aprovados.

---

[2]Uma coorte de idade representa o conjunto de pessoas com aquela idade na população considerada.

TABELA 2

Tabela de Fluxo - Dados MEC Não Corrigidos - Brasil

Série em 87

		1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	Ev.Ap.	Ev.NA	(Evadid.)	Matr86
S	1ª	1606413	3536532								124162	1050133	(1174295)	6317240
é	2ª		992625	2927594							-119978	649433	(529455)	4449674
r	3ª			626324	2496053						-94351	438226	(343875)	3466252
i	4ª				407832	2266882					-223967	346186	(122219)	2796933
e	5ª					701621	1661735				-117570	635070	(517500)	2880856
	6ª						411791	1309761			-90129	396685	(306736)	2028288
em	7ª							257582	1071906		-69926	295131	(225205)	1554693
	8ª								139388	1192092	-290534	183037	(-107497)	1223983
86	9ª									206715				1385840
	N	4737279												
	Matr87	6343692	4529157	3553918	2903885	2968503	2073526	1567343	1211294	1398807				

Série em 87

		1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	Ev.Ap.	Ev.NA	(Evadid.)	Matr86
S	1ª	0.2543	0.5598								0.0197	0.1662	(0.1859)	1.000
é	2ª		0.2231	0.6579							-0.0270	0.1460	(0.1190)	1.000
r	3ª			0.1807	0.7201						-0.0272	0.1264	(0.0992)	1.000
i	4ª				0.1458	0.8105					-0.0801	0.1238	(0.0437)	1.000
e	5ª					0.2435	0.5768				-0.0408	0.2204	(0.1796)	1.000
	6ª						0.2030	0.6457			-0.0444	0.1957	(0.1512)	1.000
em	7ª							0.1657	0.6895		-0.0450	0.1898	(0.1449)	1.000
	8ª								0.1139	0.9739	-0.2374	0.1495	(-0.0878)	1.000
86	9ª									0.1492				

TABELA 3

Série	Aprovados em 1986		Ingressos Novos em 1987		Evadidos Aprovados em 1986
1ª	3.660.694	\		//	124.162
2ª	2.807.616	\	3.536.532	//	-119.978
3ª	2.404.702	\	2.927.594	//	-94.351
4ª	2.042.915	\	2.496.053	//	-223.967
5ª	1.544.165	\	2.266.882	//	-117.570
6ª	1.219.632	\	1.661.735	//	-90.129
7ª	1.001.980	\	1.309.761	//	-69.926
8ª	901.558	\	1.071.906	//	-290.534
9ª		\	1.192.092		

Este conceito de evadido aprovado é extremamente importante e mostra inconsistência em *todas* as séries. Mesmo observando somente a evasão total encontramos inconsistências como:

iii) A evasão total entre séries (evasão de alunos aprovados ou não aprovados) pequena da 4ª série para a 5ª série e alta da 5ª série para 6ª série. Isto é estranho devido à história do sistema educacional brasileiro, já que corresponde à passagem do antigo curso primário para o antigo ginásio. Até 1971 (Lei 5 692/71) a obrigatoriedade constitucional de escolarização ia até a 4ª série e sabe-se que não só faltam classes a partir da 5ª série principalmente na área rural e nas periferias urbanas, como, também, devido às altas taxas de repetência nas séries anteriores os alunos atingem esta série com idades "avançadas" onde a opção pela entrada no mercado de trabalho começa a competir com a freqüência à escola formal. Espera-se, portanto, uma evasão alta entre a 4ª e a 5ª séries e não entre a 5ª e 6ª séries.

iv) Uma evasão total entre séries *negativas*, na 8ª série viola as hipóteses de modelo de fluxo, pois indica a presença de alunos novos provenientes de fora do Sistema Regular de Ensino. Ao contrário, deveríamos esperar uma evasão grande na 8ª série por razões análogas às da evasão na 4ª série, já que corresponde ao final do 1º Grau.

Isto, no entanto, não significa que os brasileiros desistem de estudar. O ensino Supletivo, originalmente previsto para atender àqueles que não tinham acesso à escola passa a ter um novo papel, o de recuperar o tempo perdido com o excesso de repetência e obter um diploma, já que o acesso à escola está praticamente universalizado. Não parece ser o trabalho, em si, que impede a escolarização da população, talvez ocorra o contrário, é este mesmo trabalho que permite a sobrevivência dos indivíduos para que possam, através desse "atalho" legal (o sistema Supletivo), continuar seus estudos, os vilões nessa história são, seguramente, as altas taxas de repetência nas primeiras séries.

Para as estimativas do PROFLUXO a aplicação do Modelo de Fluxo está representada na Tabela 4 em proporção de uma geração, em proporção da matrícula e em números absolutos, respectivamente.

TABELA 4

Tabela de Fluxo - Ano 87 - PROFLUXO - BRASIL

Série em 87

		1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	Ev.Ap.	Ev.NA	(Evadid.)	Matr86
S	1ª	0.9790	0.8811								0.0230	0.0314	(0.0544)	1.9145
é	2ª		0.4543	0.8268							0.0473	0.0070	(0.0543)	1.3354
r	3ª			0.3360	0.7578						0.0628	0.0062	(0.0690)	1.1627
i	4ª				0.2262	0.6134					0.1355	0.0089	(0.1444)	0.9840
e	5ª					0.2982	0.5280				0.0781	0.0073	(0.0854)	0.9116
	6ª						0.1496	0.4683			0.0553	0.0044	(0.0597)	0.6776
em	7ª							0.0923	0.4160		0.0522	0.0001	(0.0523)	0.5605
	8ª								0.0916	0.3102	0.1005	0.0052	(0.1058)	0.5076
86	9ª									0.1023	0.0203	0.0037	(0.0240)	0.4125
	N	0.9355												

Série em 87

		1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	Ev.Ap.	Ev.NA	(Evadid.)	Matr86
S	1ª	0.5114	0.4602								0.0120	0.0164	(0.0284)	1.0000
é	2ª		0.3402	0.6191							0.0354	0.0053	(0.0407)	1.0000
r	3ª			0.2890	0.6517						0.0540	0.0053	(0.0593)	1.0000
i	4ª				0.2299	0.6234					0.1377	0.0091	(0.1467)	1.0000
e	5ª					0.3271	0.5792				0.0857	0.0080	(0.0937)	1.0000
	6ª						0.2208	0.6911			0.0816	0.0065	(0.0881)	1.0000
em	7ª							0.1646	0.7421		0.0932	0.0001	(0.0933)	1.0000
	8ª								0.1805	0.6112	0.1981	0.0103	(0.2084)	1.0000
86	9ª									0.2480	0.0492	0.0090	(0.0582)	1.0000
	N	0.4887												

Série em 87

		1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	Ev.Ap.	Ev.NA	(Evadid.)	Matr86
S	1ª	3323775	2991430								78000	106600	(184600)	6500000
é	2ª		1490647	2712945							155121	23224	(178345)	4381937
r	3ª			1049216	2366409						196081	19245	(215326)	3631133
i	4ª				696895	1889538					417392	27584	(444672)	3031166
e	5ª					873203	1545976				228751	21354	(250104)	2669203
	6ª						423014	1324384			156367	12456	(168822)	1916259
em	7ª							263296	1187264		149102	160	(149262)	1599806
	8ª								250934	849833	275463	14322	(289785)	1390525
86	9ª									286284	56790	10389	(67179)	1154278
	N	3176292												

É importante assinalar que o PROFLUXO trabalha com proporções de geração e assume a estabilidade do número de matrículas do ano t para o ano $t+1$.

A tabela em números absolutos é obtida a partir da tabela em proporções da matrícula multiplicando as linhas pela matrícula, coletada pela PNAD no ano t.

6 A MATRÍCULA DE ALUNOS NOVOS NA 1ª SÉRIE

A Tabela 5 mostra, numa série histórica de 1984 a 1987 para o Brasil e para as cinco regiões geográficas do País, os ingressos novos na 1ª série do 1º Grau, segundo os dados do MEC e do

PROFLUXO; as coortes de referência de 7 anos,[3] segundo as PNADs e a relação entre alunos novos e a coorte de referência para estas duas bases de dados.

TABELA 5

Proporções de Ingressos Novos na Primeira Série Em Termos de Uma Coorte de 7 Anos[5]

	Ano	Ingr. MEC	Ingr. MEC corrigidos	Ingr. PRO-FLUXO	Coorte de 7 Anos	MEC /Coorte	Ing. Corr. /Coorte	PRO-FLUXO /Coorte
Brasil	1984	4,955,425	4,237,157	2,825,140	3,188,052	1.55	1.33	0.89
	1985	4,835,184	4,124,360	2,958,925	3,289,308	1.47	1.25	0.90
	1986	4,757,610	3,941,926	3,061,424	3,390,565	1.40	1.16	0.90
	1987	4,737,279	3,888,415	3,176,292	3,491,821	1.36	1.11	0.91
Norte Urbano	1984	220,977	177,854	99,972	107,673	2.05	1.65	0.93
	1985	222,101	168,591	112,345	114,443	1.94	1.47	0.98
	1986	218,770	159,328	119,496	121,213	1.80	1.31	0.99
	1987	215,322	166,358	120,224	127,983	1.68	1.30	0.94
Nordeste	1984	2,318,248	1,893,230	867,242	1,092,893	2.12	1.73	0.79
	1985	2,041,618	1,705,800	896,460	1,116,493	1.83	1.53	0.80
	1986	1,923,633	1,518,370	932,708	1,140,092	1.69	1.33	0.82
	1987	1,901,777	1,468,159	966,260	1,163,692	1.63	1.26	0.83
Sudeste	1984	1,603,593	1,365,628	1,193,678	1,282,278	1.25	1.07	0.93
	1985	1,565,407	1,446,319	1,280,203	1,335,028	1.17	1.08	0.96
	1986	1,619,763	1,457,371	1,316,502	1,387,778	1.17	1.05	0.95
	1987	1,649,536	1,468,520	1,422,658	1,440,527	1.15	1.02	0.99
Sul	1984	584,395	483,287	450,058	472,649	1.24	1.02	0.95
	1985	578,804	497,144	455,330	482,078	1.20	1.03	0.94
	1986	600,390	511,002	466,741	491,507	1.22	1.04	0.95
	1987	576,218	490,267	497,713	500,937	1.15	0.98	0.99
Centro-Oeste	1984	418,191	317,158	210,438	232,381	1.80	1.36	0.91
	1985	427,254	306,506	220,869	240,821	1.77	1.27	0.92
	1986	395,718	295,855	238,129	249,262	1.59	1.19	0.96
	1987	394,426	295,111	237,818	257,703	1.53	1.15	0.92

Nota – Como o PROFLUXO calcula independentemente as proporções de uma geração em cada região, a soma dos ingressos novos das diversas regiões é apenas aproximadamente igual ao total de ingressos novos do Brasil.

Vemos que as matrículas do MEC são bem maiores que uma coorte de idade por vários anos consecutivos o que é *totalmente inconsistente* com a realidade demográfica do País.

Esta inconsistência já foi discutida e denunciada em:

i) M. A. Teixeira de Freitas (1947, 1989), então diretor do Serviço de Estatística do MEC, utilizando dados brasileiros da década de 30.

[3]Os valores dessas coortes de 7 anos foram obtidos por um ajuste linear das estimativas desses valores nas PNADs de 1981 a 1989.

ii) E. Schiefelbein (1975) utilizando dados de 14 países da América Latina inclusive o Brasil com dados da década de 60;

iii) Thonstad (1980, p.34) discute a impossibilidade de taxas de ingressos novos maior que uma coorte de idade de referência por vários anos consecutivos;

iv) E. Cuadra (1989, p.9 e Apêndice) mostra que uma taxa de ingressos novos de 1,3 coortes de 6 anos, em Honduras, é incompatível com a realidade.

Daremos a seguir mais três argumentos para mostrar a inviabilidade de termos por vários anos consecutivos um número de alunos novos na 1ª série maior do que uma coorte de idade de referência.

O primeiro argumento, de natureza mecânica, serve para mostrar, de forma simples, a impossibilidade deste fato. Os dois outros, de natureza probabilística, um assumindo a estabilidade do sistema e o último relaxando esta hipótese de estabilidade, esclarecem definitivamente o problema.

No primeiro argumento supomos que uma fonte alimentando uma caixa-d'água representa os nascimentos das crianças no País a cada ano (o fluxo de nascimentos na população). Na base da caixa-d'água existem pequenas torneiras que representam o acesso à escola de crianças de 5, 6, 7, 8, etc. anos de idade e uma última que representa as crianças que nunca terão acesso à escola. Assumindo uma razoável estabilidade do sistema e do crescimento das coortes, e desprezando a mortalidade, vemos que a soma dos fluxos de água das diversas torneiras não poderá ser maior do que a da fonte. Se isto ocorresse, por vários anos, a caixa se esvaziaria e chegaríamos novamente ao equilíbrio, isto é, voltaríamos a ter uma situação em que a soma do fluxo de todas as torneiras não seria maior do que o da fonte.

Os dois argumentos probabilísticos mostram, sob certas condições de estabilidade do sistema escolar, a impossibilidade deste fato.

Denotemos por:

t, o ano que estamos considerando;

X_u a variável aleatória idade de ingresso na 1ª série para os nascidos no ano u. Observamos que esta variável só está definida para as pessoas que ingressam na 1ª série;

$P(X_u = i)$ a probabilidade que uma pessoa nascida no ano u ingresse na 1ª série com idade i;

$I(u)$ a taxa de cobertura ou participação na 1ª série da coorte nascida no ano u, isto é, a proporção da coorte nascida no ano u que eventualmente ingressa na 1ª série;

$T(u)$ o tamanho da coorte nascida no ano u, descontando a mortalidade;

Ing_t o número total de alunos novos na 1ª série no ano t;

$Ing_t(i)$ o número de alunos novos na 1ª série no ano t com idade i, ou seja, que nasceram no ano $u = t - i$.

Podemos escrever:

$$Ing_t = \sum_i Ing_t(i)$$

e

$$Ing_t(i) \sim P(X_{t-i} = i)I(t-i)T(t-i)$$

Supomos que não entra nenhum aluno com menos de 5 anos, isto é, $P(X_{t-i} = i) = 0$ para $i < 5, T(t-5) \geq T(t-6) \geq T(t-7) \geq \cdots$ e que a taxa de cobertura está crescendo com o tempo, isto é, $I(t-5) \geq I(t-6) \geq I(t-7) \geq \cdots$ Sabemos que $1 > I(t-5)$.

Estas hipóteses são satisfeitas no Brasil até a década de 90.

Logo,

$$Ing_t = \sum_i P(X_{t-i} = i)I(t-i)T(t-i) < I(t-5)T(t-5)\sum_i P(X_{t-i} = i)$$

e

$$\frac{Ing_t}{T(t-5)} < I(t-5)\sum P(X_{t-i} = i) = I(t-5)s, \quad \text{onde} \quad s = \sum_i P(X_{t-i} = i)$$

Supondo que o sistema varia lentamente, a distribuição das variáveis aleatórias varia lentamente e s deve ser próximo de 1. Logo o número de ingressos novos $Ing_t < sI(t-5)T(t-5) \sim I(t-5)T(t-5) < T(t-5)$.

Vemos por este primeiro argumento probabilístico que, sob certa estabilidade do sistema, o número de ingressos novos não pode ser muito maior que uma coorte de referência, sendo provavelmente menor.

Em um segundo argumento, relaxando um pouco a hipótese de estabilidade, permitimos uma certa variação da função de distribuição cumulativa da variável aleatória X_u, qualquer que seja o ano u, da seguinte forma:

$$0,000 \leq P(X_u \leq 5) \leq 0.05$$
$$0,150 \leq P(X_u \leq 6) \leq 0.35$$
$$0,650 \leq P(X_u \leq 7) \leq 0.90$$
$$0,800 \leq P(X_u \leq 8) \leq 0.95$$
$$0,900 \leq P(X_u \leq 9) \leq 0.98$$
$$0,950 \leq P(X_u \leq 10) \leq 1.00$$
$$0,970 \leq P(X_u \leq 11) \leq 1.00$$
$$0,985 \leq P(X_u \leq 12) \leq 1.00$$
$$0,995 \leq P(X_u \leq 13) \leq 1.00$$

Os limites considerados são bastante "generosos", razoáveis e contêm as funções de distribuições cumulativas para o ingresso na 1ª série observados nos últimos 10 anos para o Brasil.

Com estas hipóteses, podemos considerar desprezível o número de alunos que entram com 15 anos ou mais na 1ª série.

A Tabela 6 mostra os ingressos por idade de 5 a 14 anos durante 10 anos, do ano t ao ano $t+9$. Supomos que no ano $t+j, j = 0, \cdots, 9$, o número de ingressos novos é uma percentagem constante x da coorte de 7 anos neste ano ou seja da coorte nascida no ano $u = t+j-7$.

Seja N o total da Tabela 6.

TABELA 6

	t	t+1	t+2	.....	t+9	
5	Ing_t (5)	Ing_{t+1} (5)	Ing_{t+2} (5)	.....	Ing_{t+9} (5)	
6	Ing_t (6)	Ing_{t+1} (6)	Ing_{t+2} (6)	.....	Ing_{t+9} (6)	
7	Ing_t (7)	Ing_{t+1} (7)	Ing_{t+2} (7)	.....	Ing_{t+9} (7)	
8	.	.	.	.....	.	
9	.	.	.	.....	.	
10	.	.	.	.....	.	
11	.	.	.	.....	.	
12	.	.	.	.....	.	
13	.	.	.	.....	.	
14	Ing_t (14)	Ing_{t+1} (14)	Ing_{t+2} (14)	.....	Ing_{t+9} (14)	
	xT(t-7)	xT(t-6)	xT(t-7)		xT(t+2)	N

Então:

$$N = x\sum_{j=0}^{9} T(t+j-7) = \sum_{j=0}^{9}\sum_{i=5}^{14} Ing_{t+j}(i) = \sum_{u=t-14}^{t+4}\ \sum_{t+j-i=u} Ing_{t+j}(i)$$

Observe que o ano de nascimento $u = t+j-i$, $j = 0, \cdots, 9$ e $1 = 5, \cdots, 14$, é constante ao longo de "diagonais". Podemos majorar cada uma das diagonais com o ano de nascimento u constante, lembrando que:

$$Ing_{t+j}(i) \sim P(X_{t+j-i} = i)I(t+j-i)T(t+j-i) = P(X_u = i)I(u)T(u),$$

que $P(X_u \geq i+1) = 1 - P(X_u \leq i)$ e usando a hipótese sobre a distribuição das variáveis aleatórias X_u.

Logo temos, seguindo as diagonais com o ano de nascimento μ constante, do canto inferior esquerdo ao canto superior direito: $N \leq 0.005I(t-14)T(t-14) + 0.015I(t-13)T(t-13) + 0.03I(t-12)T(t-12) + 0.05\ I(t-11)T(t-11) + 0.1I(t-10)T(t-10) + 0.2I(t-9)T(t-9) + 0.35I(t-8)\ T(t-8) + 0.85I(t-7)T(t-7) + I(t-6)T(t-6) + I(t-5)T(t-5) + \ I(t-4)T\ (t-4) + I(t-3)T(t-3) + I(t-2)T(t-2) + I(t-1)T(t-1) + 0.98\ I(t)T(t) + 0.95I(t+1)T\ (t+1) + 0.9I(t+2)T(t+2) + 0.35I(t+3)T(t+3) + 0.05I(t+4)T(t+4)$

Suponhamos $I(u) \leq 0.95$, ou seja, uma taxa de participação maior que a observada até no Brasil, para todo u e supomos que os tamanhos das coortes crescem a uma taxa constante y, isto é, $T(u) = yT(u-1)$.

Então,

$$N = x\sum_{j=0}^{9} T(t+j-7) = xT(t-7)\sum_{j=0}^{9} y^j \leq 0.95T(t-7)[0.005y^{-7} + 0.015y^{-6} + 0.03y^{-5} +$$
$$+ 0.05y^{-4} + 0.1y^{-3} + 0.2y^{-2} + 0.35y^{-1} + 0.85 + y + y^2 + y^3 + y^4 + y^5 +$$
$$+ y^6 + 0.98y^7 + 0.95y^8 + 0.9y^9 + 0.35y^{10} + 0.05y^{11}]$$
$$= 0.95T(t-7)S_a,$$

onde S_a é a soma entre colchetes.

Seja $S_b = \sum_{j=0}^{9} y^j$. Logo $x \leq (0.95S_a)/S_b$, uma cota superior para x.

TABELA 7

y	cota superior de x
1.01	1.026
1.02	1.024
1.03	1.023
1.04	1.021
1.05	1.020
1.06	1.019
1.07	1.019

Vemos na Tabela 7 que mesmo majorando o número total de ingressos novos ao longo dos 10 (dez) anos, a cota superior obtida para x é em torno de 1,02, para valores razoáveis de y. É claro que a taxa de crescimento varia com o ano, mas não deve ser muito diferente das apresentadas na Tabela 5, o que implica que a cota superior também deve estar no intervalo apresentado na Tabela 7. No Brasil, para a década de 80, a coorte de 7 anos, segundo as estimativas das PNADs, cresceu a uma taxa média de 3.4% ao ano, o que significa um valor de $y \sim 1.034$. Vemos, portanto, que a cota superior de x não pode ser superior a 1.023.

O valor x também depende do ano. O argumento é válido, no entanto, para o mínimo dos x. Estas cotas superiores são bem menores que os valores encontrados nos últimos 10 (dez) anos

para o Brasil, ver Tabela 5, o que mostra a impossibilidade demográfica dos valores obtidos pelos Censos Educacionais com a definição de repetente utilizada.

7 AS EXPLICAÇÕES DAS INCONSISTÊNCIAS

Os Censos Educacionais do MEC classificam os alunos, no final do ano letivo, como afastados por abandono ou transferência, aprovados ou reprovados. Os afastados por transferência, em princípio, não interferem no número de alunos do sistema, pois, ao saírem de uma escola, ingressam em outra. No entanto, um aluno afastado por abandono não é considerado nem aprovado nem reprovado e caso se rematricule no ano seguinte, na mesma série, na mesma ou em outra escola, não é considerado um repetente.

Aí está a chave do problema e o principal erro conceitual na definição de repetente pelo MEC.

Nos questionários dos Censos Educacionais pergunta-se o número de alunos matriculados e de repetentes (os alunos novos são calculados por diferença), em cada série, no início do ano letivo. Portanto, um aluno afastado por abandono ao se matricular no ano seguinte na mesma série, na mesma ou em outra escola não será considerado repetente e, conseqüentemente será contado como um aluno novo. Este aluno, ao ser contado na equação 2 do Modelo de Fluxo como aluno novo, viola a hipótese básica do Modelo.

Nestas condições verifica-se que nem todos os alunos novos, considerados pelas escolas ao preencher o questionário do MEC, são realmente novos. Caso tenham abandonado a escola depois da matrícula inicial seria necessário considerá-los como repetentes já que freqüentaram a escola por algum tempo. Ao não considerá-los como repetentes vamos obter um número subestimado de repetentes e superestimados de alunos novos. Esta é a principal causa das inconsistências, mas não a única, encontradas quando se aplica o modelo de fluxo aos dados do MEC.

A Tabela 8 apresenta para o ano de 1986:

i) a matrícula estimada pela PNAD;

ii) a matrícula inicial do Censo Educacional;

iii) os alunos afastados por abandono;

iv) a matrícula final do Censo Educacional.

v) diferença (Dif.) = matrícula inicial menos matrícula final menos afastados por abandono;

vi) diferença dividida pela matrícula inicial;

vii) matrícula PNAD menos matrícula inicial dividida pela matrícula inicial;

viii) matrícula PNAD menos matrícula final dividida pela matrícula final.

TABELA 8

1986	Matrícula PNAD	Matrícula Inicial	Afastados por Abandono	Matrícula Final	MI-MF-AFAB (Dif.)	Dif. por MI	(PNAD-MI) por MI	(PNAD-MF) por MF
1ª S	6.500.000	6.317.240	800.144	5.286.389	230.707	0.04	0.03	0.23
2ª S	4.381.937	4.449.674	394.889	3.866.704	188.081	0.04	-0.02	0.13
3ª S	3.631.133	3.466.252	313.297	3.068.121	84.834	0.02	0.05	0.18
4ª S	3.031.166	2.798.933	242.465	2.480.053	76.415	0.03	0.08	0.22
5ª S	2.669.203	2.880.856	540.193	2.250.627	90.036	0.03	-0.07	0.19
6ª S	1.916.259	2.028.288	322.562	1.659.381	46.345	0.02	-0.06	0.15
7ª S	1.599.806	1.554.693	235.952	1.295.972	22.769	0.01	0.03	0.23
8ª S	1.390.525	1.223.983	154.009	1.054.055	15.919	0.01	0.14	0.32
1ª S2ºG	1.154.278	1.385.840	390.557	1.001.543	-6.260	-0.00	-0.17	0.15

Vemos que o número de alunos afastados por abandono nas diversas séries do 1º Grau é considerável, como mostra a Tabela 8A e não pode ser desprezado.

TABELA 8A

1986	Matrícula Inicial	Afastados por Abandono	Afast./MI
1ª S	6317240	800144	0.13
2ª S	4449674	394889	0.09
3ª S	3466252	313297	0.09
4ª S	2798933	242465	0.09
5ª S	2880856	540193	0.19
6ª S	2028288	322562	0.16
7ª S	1554693	235952	0.15
8ª S	1223983	154009	0.13
1ª S2ºG	1385840	390557	0.28

Até 1986 inclusive, o IBGE, ao fazer a crítica dos dados das PNADs, considerava não fazer sentido um aluno que tinha mais de 7 anos de idade estar matriculado na pré-escola ou classe de alfabetização e automaticamente reclassificava estes alunos como freqüentando a 1ª série do 1º Grau. Portanto, para termos comparatividade com os dados do Censo Educacional, a estimativa da PNAD de 1986 para a 1ª série foi corrigida, utilizando-se os dados extrapolados das PNADs de 1987, 1988 e 1989 onde são registrados corretamente os alunos que declaram estar no pré-escolar até a idade de 9 anos.

A observação mais importante é a constatação de que a matrícula estimada pela PNAD de 1986, coletada entre os meses de setembro a novembro daquele ano, é muito próxima do número de matrículas iniciais do MEC pelo menos até a 7ª série e bem superior à matrícula

final do MEC (veja colunas "PNAD-MI por MI" e "PNAD-MF por MF" na Tabela 8).

O que se depreende desses números é que a maioria dos alunos afastados por abandono, ou seja, com matrícula cancelada, assim o fizeram no final do ano letivo[4]. É possível que estejamos constatando, aqui, uma forma de "repetência branca", onde os alunos são "aconselhados" ou "estimulados" a abandonarem a escola no final do ano letivo, pela certeza do fracasso e preferem a evasão por abandono, ou seja, pelo cancelamento da matrícula que preservará seu histórico escolar. Este comportamento respresentaria uma possibilidade de melhor aproveitamento de conteúdos da série sem o estigma da repetência. Este procedimento resguarda, também, a imagem do professor e da escola.

Certamente, a distribuição da época em que os alunos afastados por abandono (ou seja, que têm sua matrícula cancelada), deixam a escola e as causas deste comportamento merecem ser pesquisadas a fundo. Não temos encontrado, na literatura educacional, trabalhos sobre este tema, o que é surpreendente.

8 A "CORREÇÃO" DOS DADOS DO MEC

Pelas observações acima podemos propor uma correção nos dados dos Censos Educacionais que consiste em considerar os afastados por abandono no ano anterior como repetentes, já que estes alunos ocuparam uma vaga no sistema. Ao considerar todos os afastados por abandono como repetentes no ano seguinte estamos inflacionando os repetentes; no entanto, mesmo assim verificamos que os ingressos novos na 1ª série ainda ficam acima das possibilidades demográficas do País.

É claro que uma parte desses afastados por abandono se evade definitivamente do sistema, porém poderíamos considerar que a taxa de evasão, neste caso, seria pelo menos semelhante às taxas de evadidos oficialmente reprovados que são pequenas, ver Prevor na Tabela 9. Mas não o faremos neste artigo. Cálculos feitos por nós utilizando esta hipótese diminuem as taxas de repetência em, no máximo, 2%, não havendo praticamente alteração na 1ª série.

Chamamos a atenção para a soma dos aprovados com os oficialmente reprovados e afastados por abandono, que deveria ser igual à matrícula inicial nos dados da Sinopse do MEC; por erro na crítica dos dados, apresenta diferenças (ver colunas MI – MF – AFAB e DIF por MI na Tabela 8).

A Tabela 9 apresenta para o Brasil, por série, dados dos Censos Educacionais de:

i) matrículas em 1986 (Matr86) e 1987 (Matr87);

[4]Há suspeitas de que as coortes jovens nas PNADs estão um pouco superestimadas. No entanto, a ordem de grandeza deste problema, a nosso ver, não afeta esta conclusão.

ii) repetentes em 1987 (Repe87);

iii) afastados por abandono em 1986 (Afas86);

iv) aprovados em 1986 (Apro86);

v) reprovados em 1986 (Repr86);

Esta Tabela apresenta ainda os seguintes dados derivados, onde o índice k representa a série:

vi) ingressos novos em 1987 ($Ingn87_k$ = $Matr87_k$ - $Repe87_k$)

vii) evadidos aprovados em 1986 ($Evap86_k$ = $Apro86_k$ - $Ingn87_{k+1}$);

viii) evadidos não aprovados em 1986 ($Evna86_k$ = $Matr86_k$ - $Repe87_k$ - $Ingn87_{k+1}$ - $Evap86_k$);

ix) evadidos oficialmente reprovados ($Evorep_k$ = $Repr86_k$ - $Repe87_k$);

x) proporção de evadidos oficialmente reprovados ($Prevor_k$ = [$Repr86_k$ - $Repe87_k$] / $Repr86_k$);

xi) repetentes corrigidos em 1987 ($Repc87_k$ = $Repe87_k$ + $Afas86_k$);

xii) ingressos novos corrigidos em 1987 ($Ingc87_k$ = $Matr87_k$ - $Repc87_k$);

xiii) evadidos aprovados corrigidos em 1986 ($Evac86_k$ = $Apro86_k$ - $Ingc87_{k+1}$) e

xiv) evadidos não aprovados corrigidos em 1986 ($Evnc86_k$ = $Matr86_k$ - $Repc87_k$ - $Ingc87_{k+1}$ - $Evac86_k$).

TABELA 9

Dados MEC - Brasil

Série	1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª
Matr86	6317240	4449674	3466252	2796933	2880856	2028288	1554693	1223983	1385840
Matr87	6343692	4529157	3553918	2903885	2968503	2073526	1567343	1211294	1398807
Repe87	1606413	992625	626324	407832	701621	411791	257582	139388	206715
Afas 86	848864	430169	313297	242465	540193	322562	235952	154009	390557
Apro86	3660694	2807616	2401702	2042915	1544165	1219632	1001980	901558	746129
Repr86	1625695	1059088	666419	437138	746462	439749	293992	152497	255414
Ingn 87	4377279	3536532	2927594	2496053	2266882	1661735	1309761	1071906	1192092
Evap86	124162	-119978	-94351	-223967	-117570	-90129	-69926	-290534	ND
Evna86	1050133	649433	438226	346186	635070	396865	295131	183037	ND
Evorep	19282	66463	40095	29306	44841	27958	36410	13109	48699
Prevor	0.0119	0.0628	0.0602	0.067	0.0601	0.0636	0.1238	0.086	0.1907
Repc87	2455277	1422794	939621	650297	1241814	734353	493534	293397	597272
Ingc 87	3888415	3106363	2614297	2253588	1726689	1339173	1073809	917897	801535
Evac86	554331	193319	148114	316226	204992	145823	84083	100023	ND
Evnc86	201269	219264	124929	103721	94877	74303	59179	29028	ND

Os números de evadidos aprovados e não aprovados não estão disponíveis para a 9ª série, pois, para calculá-los, é necessário o número de ingressos novos na 2ª série do 2º Grau, não encontrados nas sinopses publicadas.

Observamos, na Tabela 9, que o número de evadidos oficialmente reprovados (Evorep) e a proporção de evadidos oficialmente reprovados (Prevor) são muito baixos indicando que os repetentes em 1987, de fato, são alunos oficialmente reprovados na mesma série em 1986, o que confirma nosso diagnóstico de que somente alunos registrados como reprovados são

considerados repetentes, conforme definição de repetente no Censo Educacional.

A Tabela 10 apresenta o resultado da aplicação do Modelo de Fluxo aos dados corrigidos para os repetentes extraídos da Tabela 9.

TABELA 10

Tabela de Fluxo - Dados MEC Corrigidos para os Afastados por Abandono - Brasil - 1987

Série em 87

		1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	Ev.Ap.	Ev.NA	(Evadid.)	Matr86
S	1ª	2455277	3106363								554331	201269	(755600)	6317240
é	2ª		1422794	2614297							193319	219264	(412583)	4449674
r	3ª			939621	2253588						148114	124929	(273043)	3466252
i	4ª				650297	1726689					316226	103721	(419947)	2796933
e	5ª					1241814	1339173				204992	94877	(299869)	2880856
	6ª						734353	1073809			145823	74303	(220126)	2028288
em	7ª							493534	917897		84083	59179	(143262)	1554693
	8ª								293397	801535	100023	29028	(129051)	1223983
86	9ª									597272				1385840
	N	3888415												
	Matr87	6343692	4529157	3553918	2903885	2968503	2073526	1567343	1211294	1398807				

Série em 87

		1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	Ev.Ap.	Ev.NA	(Evadid.)	Matr86
S	1ª	0.3887	0.4917								0.0877	0.0319	(0.1196)	1.0000
é	2ª		0.3198	0.5875							0.0434	0.0493	(0.0927)	1.0000
r	3ª			0.2711	0.6502						0.0427	0.0360	(0.0788)	1.0000
i	4ª				0.2325	0.6174					0.1131	0.0371	(0.1501)	1.0000
e	5ª					0.4311	0.4649				0.0712	0.0329	(0.1041)	1.0000
	6ª						0.3621	0.5294			0.0719	0.0366	(0.1085)	1.0000
em	7ª							0.3174	0.5904		0.0541	0.0381	(0.0921)	1.0000
	8ª								0.2397	0.6549	0.0817	0.0237	(0.1054)	1.0000
86	9ª									0.4310				1.0000

Observamos que tanto na Tabela 9 como na Tabela de Fluxo 10 que:

i) o número corrigidos de alunos novos cai em todas as séries, mas na 1ª série ainda é bem maior do que as possibilidades demográficas representadas pela coorte de referência de 7 anos. As exceções são as Regiões Sudeste e Sul do País onde parece que esta discrepância é pequena compatível com os erros inerentes aos dados primários;

ii) o número de evadidos aprovados passa a ser positivo como deveria ser;

iii) a evasão entre séries da 4ª para a 5ª passa a ser alta e maior que nas demais séries como deve ocorrer pelos argumentos já discutidos. Esta evasão é, inclusive, maior que a da 1ª série, apesar dos problemas que ainda persistem na matrícula de ingressos novos nesta série;

iv) as taxas de repetência aumentam em *todas* as séries como esperado, mas são excessivamente altas da 5ª a 8ª séries. Acreditamos que estes valores são um pouco maiores do que a realidade, já que estamos supondo que *todos* os alunos afastados por abandono retornam ao sistema no ano seguinte, o que não deve estar ocorrendo nestas séries; e

v) verificamos pela Tabela 9 que, da 3ª a 8ª séries, pelo menos 40% dos alunos afastados por abandono em 1986 têm que retornar à mesma série no ano seguinte para eliminar a

inconsistência observada nas Tabelas 2 e 3 do número negativo de alunos evadidos aprovados da 2ª a 7ª séries.

Já para eliminar esta inconsistência na 8ª série necessitaríamos que pelo menos 75% dos afastados por abandono na 9ª série retornassem a esta série no ano seguinte. O valor alto desta percentagem em relação às outras séries faz supor que parte dos alunos "novos" na 9ª série não tenham sido promovidos da 8ª série no ano anterior, mas, sim, que ingressaram provavelmente na 9ª série após completarem o 1º Grau via Supletivo.

Supondo que a percentagem de 40% observada da 3ª a 8ª séries seja também um valor razoável para a 9ª série, é possível que o excesso de 30 a 35% dos afastados por abandono nesta série (~ 120.000), que são necessários para eliminar a inconsistência, represente uma componente da matrícula inicial da 9ª série de alunos provenientes graduados pelo Supletivo de 1º Grau.

Este número representa cerca de 10% da matrícula inicial na 9ª série. Para levarmos em conta esta suposição no Modelo de Fluxo teríamos que modificar o modelo introduzindo um termo de alunos novos, de fora do Sistema Regular, na equação 2, para a 9ª série.

Isto implicaria que a percentagem de promovidos da 8ª série para 9ª série (Sistema Regular) baixaria em cerca de 10% e conseqüentemente a proporção de evadidos aprovados na 8ª série aumentaria em 10%, chegando, portanto, a 18% da matrícula e os evadidos totais (aprovados + não-aprovados) a 20%. Este valor parece mais realista diante do esperado para a passagem do 1º para 2º Grau e se aproxima do valor obtido pelo PROFLUXO.

As sugestões apresentadas no Apêndice 1, se implementadas, permitirão ao Censo Educacional coletar este dado, número de alunos provenientes do Supletivo ou outras formas de fora do Sistema Regular, não só para a 9ª série, mas para todas as séries.

9 OS ALUNOS "NOVOS" NA 1ª SÉRIE

Concluímos, pelo argumento demográfico, que (ver Tabela 5) o número de alunos novos na 1ª série ainda continua inconsistente com as possibilidades demográficas. Como já estamos incluindo como repetentes todos os alunos afastados por abandono e a taxa de evadidos reprovados, nesta série, é muito baixa (ver o valor de PREVOR = 0,01 da 1ª série na Tabela 9), chegamos à conclusão que deve haver alunos declarados como aprovados que repetem a série. Esta incongruência pode advir de várias causas como, por exemplo, a escola informar um número de aprovados superior ao real por motivos políticos mantendo a informação correta apenas para uso interno, uma subseriação informal da 1ª série ou uma recomendação da escola para o aluno repetir o ano mesmo tendo sido aprovado, pois não está "maduro".

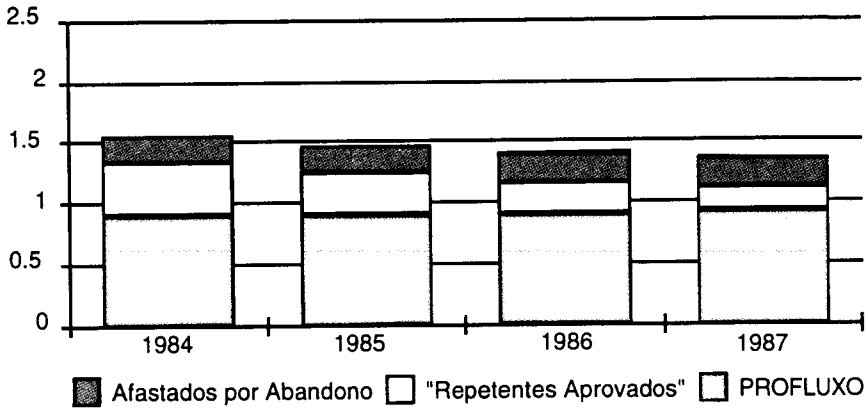
É importante observar que na metodologia do PROFLUXO estes alunos aparecem automaticamente como repetentes. Ao contrário, nos questionários do MEC a única maneira das escolas computarem estes alunos é como novos na série.

A Figura 1 mostra, em proporções da coorte de 7 anos, a evolução destas formas de repetência nos ànos de 1984 a 1987, considerando como alunos realmente novos a taxa de participação na 1ª série calculada pelo PROFLUXO, única taxa compatível com as possibilidades demográficas do Brasil.

FIGURA 1

Composição da Matrícula de Ingressos "Novos" do MEC na 1ª Série, em Proporção da Coorte de 7 Anos

BRASIL



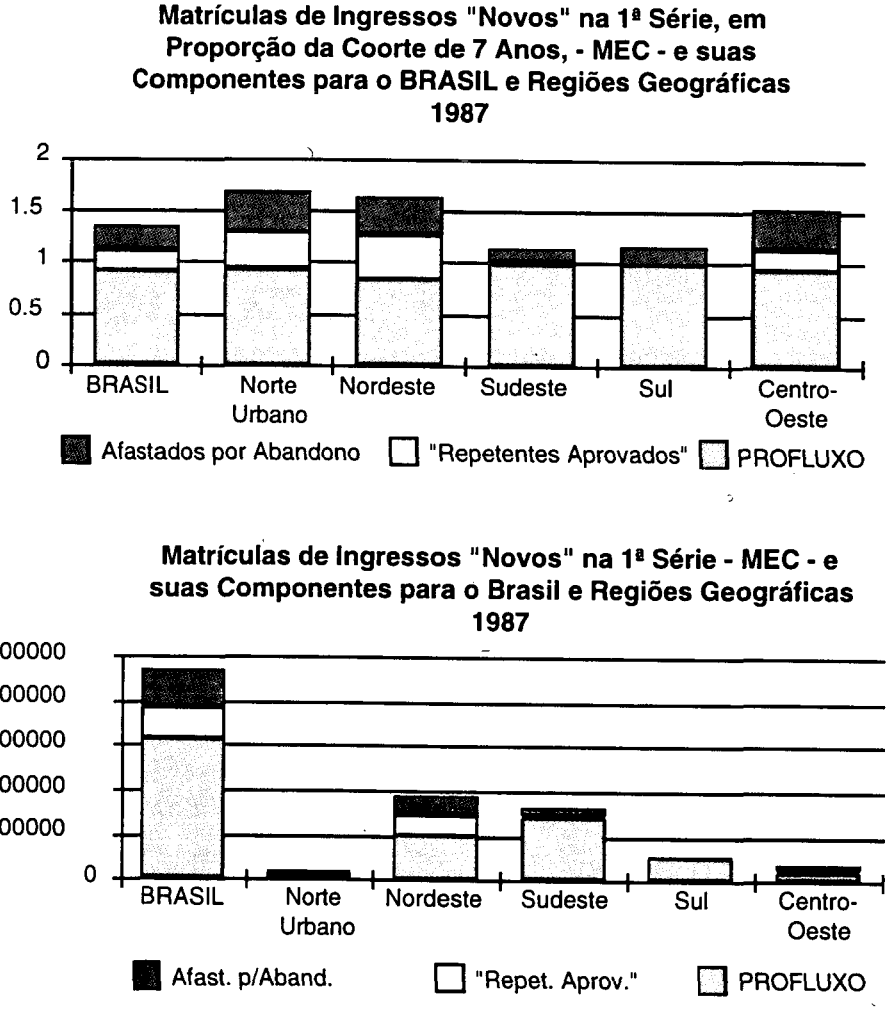
É importante observarmos que a matrícula de "novos" na 1ª série (dados originais do MEC) cai, durante o período analisado (1984-1987), não só em termos relativos como também em números absolutos (ver Tabela 5). O gráfico da Figura 1 mostra ainda que esta queda se dá pela diminuição da matrícula de "repetente aprovados". Já os afastados por abandono correspondem a uma parcela estável dos alunos novos no período. Ao contrário, a matrícula de novos calculada pelo PROFLUXO está crescendo lentamente, no período, como deveríamos esperar.

É forçoso notar que esta queda se dá simultaneamente com o aumento da freqüência à pré-escola observado nos dados das PNADs e dos Censos Educacionais. A idade média de conclusão da 2ª série, no entanto, tem se mantido constante nos últimos anos como mostram as estimativas do PROFLUXO. O que provavelmente está ocorrendo é que a subseriação informal que representa os "repetentes aprovados" está apenas mudando de nome de 1ª série para pré-escola. Nas PNADs a partir de 1987 aparecem dados de alunos matriculados na pré-escola com 7, 8 e 9 anos de idade, o que era automaticamente imputado colocado como 1ª série até aquela PNAD. Isto reforça a hipótese que os "repetentes aprovados" são hoje parte dos matriculados

na pré-escola.

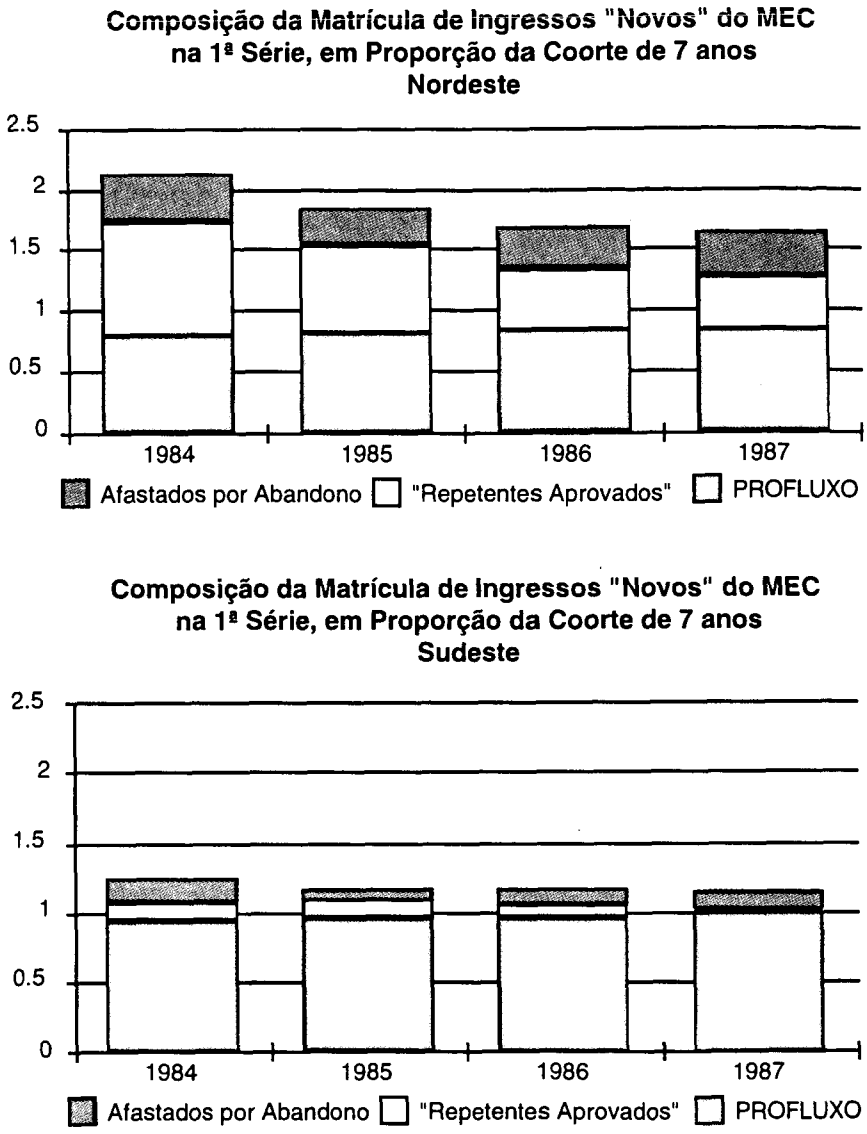
Estes fatos não ocorrem com a mesma importância em todas as Regiões do Brasil. A Figura 2 mostra que os "repetentes aprovados" aparecem principalmente no Norte Urbano, no Nordeste e no Centro-Oeste, sendo desprezível no Sudeste e Sul. Em números absolutos, o Nordeste corresponde à esmagadora maioria dos "repetentes aprovados" no País.

FIGURA 2



A evolução da diminuição dos "repetentes aprovados" ao longo do período pode ser melhor observada na Figura 3 para o Nordeste e Sudeste. Vemos quão importante é analisar os dados ao longo do tempo, através dos diversos Censos e PNADs, pois permite obter informações extremamente relevantes sobre as conseqüências das políticas governamentais na área.

FIGURA 3



Ainda que não possamos, com os dados do MEC, analisar estas ocorrências por nível socioeconômico da população, é impressionante notar que esta "repetência de aprovados" ocorre com maior freqüência nos estados do Nordeste, Norte e Centro-Oeste, onde a pobreza e a má distribuição de renda são mais graves. É forçoso inferir que este fato deve depender do nível socioeconômico da população. Esta forma de repetência representa uma parcela importante da matrícula da 1ª série e deve estar ocorrendo, ainda que com menor intensidade, nas demais séries, na mesma escola ou, o que parece ser comum, no processo de transferência de uma escola para outra.

Podemos estimar a taxa desta forma de "repetência de aprovados" na 1ª série assumindo,

como repetentes, o excesso de alunos novos em relação à taxa de participação[5] de ingressos novos calculada pelo PROFLUXO para 1987, de 0.935 (ver Tabela 4, Novos na 1ª série em Proporção de uma Geração), que representa um limite superior compatível com as possibilidades demográficas do país. Os ingressos novos devem ter, então, um limite superior de Ingct87 = 0.935 x (Coorte de 7 anos)[6]. Assumindo este limite superior como os ingressos novos reais obtemos um limite inferior para os "repetentes aprovados". Este último número será utilizado como os "repetentes aprovados" na 1ª série. Verificamos que estes "repetentes aprovados" representam cerca de 10% da matrícula da 1ª série, em 1987.

Ao fazer esta correção, o número de aprovados na 1ª série diminui e constatamos que o número de evadidos aprovados corrigidos fica negativo. Obviamente isto é absurdo, como já tínhamos observado em todas as demais séries do 1º Grau nos dados originais do MEC. Portanto, o número de ingressos novos na 2ª série corrigido (promovidos da 1ª série) está superestimado, indicando que o mesmo fenômeno de "repetência de aprovados" está ocorrendo nesta série, porém em menor escala. Considerando a proporção de evadidos aprovados na matrícula da 1ª série como sendo da mesma ordem de grandeza obtida pelo PROFLUXO, (1.2%) (ver Tabela 4), podemos estimar o número de "repetentes aprovados" na 2ª série em 2,6% da matrícula nesta série, em 1987.

Estes números corrigidos para 1ª e 2ª séries estão reproduzidos na Tabela 11 e Tabela de Fluxo 12 correspondente.

Onde:

Ingressos novos com correção total Ingct87;

Repetentes com correção total Repct87;

Evadidos aprovados com correção total Evact86;

[5]Taxa de participação de acesso de uma coorte, em uma série, é a proporção da coorte que eventualmente tem acesso, em algum momento, àquela série.

[6]Este valor é um pouco superior à matrícula de novos na 1ª série calculada pelo PROFLUXO, ver Tabela 4, já que a coorte de referência neste modelo é uma média das coortes da distribuição de ingressos novos por idade, que é ligeiramente inferior à coorte de 7 anos. Estamos, portanto, estimando, para os realmente novos na série, o valor mais otimista possível.

TABELA 11

	1ª Série	2ª Série
Ingct87	3.264.853	2.961.325
Repct87	3.078.839	1.567.832
Evact86	75.806	48.281

10 OS PROBLEMAS DA 5ª SÉRIE DO 1º GRAU E DA 1ª SÉRIE DO 2º GRAU

É nestas séries, que correspondem a mudanças do antigo Curso Primário para o antigo Ginásio e do Ginásio para o antigo 2º Ciclo (ou Clássico e Científico), que vamos encontrar problemas adicionais que merecem uma análise detalhada.

Observamos na Tabela 5 que as matrículas iniciais do MEC na 5ª e 9ª séries são bem maiores que as matrículas estimadas pela PNAD. Este fato sugere a existência de um abandono precoce ou dupla contagem nos dados do MEC.

Na 5ª e 9ª séries há mudanças importantes tanto do ponto de vista pedagógico como do ponto de vista de acesso à escola. Sabemos que uma percentagem importante de escolas no Brasil são de apenas quatro séries, o que provoca não só a evasão observada como, em muitos casos, a necessidade de mudanças de escola, com aumento da distância da residência e até de cidade, para os estudantes. Estas peculiaridades aliadas ao fato da idade "avançada" com que, em média, os alunos atingem estas séries devido às altas taxas de repetência nas séries anteriores, podem, talvez, explicar o abandono precoce, antes do final do ano, quando são coletadas as matrículas pelas PNADs. Podem ainda ser a razão de uma dupla contagem de alunos que se matriculam em mais de uma escola e decidem por uma, após a informação sobre a matrícula inicial ser coletada pelo Censo Educacional. A primeira hipótese se fortalece se as escolas realmente estiverem cumprindo as determinações dos manuais de preenchimento de não incluírem na matrícula inicial os alunos que nunca compareceram à escola.

Entretanto, estas hipóteses podem não representar toda a realidade já que existe no Brasil a possibilidade de, após completarem 14 anos de idade, os jovens cursarem ou apenas fazerem os exames de cursos supletivos seriados, modulados ou não seriados. E aqui vamos encontrar dificuldades em detectar os mecanismos que ocorrem, mas é possível que este abandono precoce seja conseqüência da tentativa de parte dos estudantes que, encontrando dificuldades no ensino regular ou devido à idade "avançada", abandonam a escola para prosseguirem seus estudos

via Supletivo e, eventualmente, retornam à escola regular.

Este retorno de alunos provenientes do ensino Supletivo ao Ensino Regular inflaciona a matrícula de novos, violando assim as hipóteses básicas do Modelo de Fluxo. Caso haja registro desse tipo de estudante, é possível introduzir um termo adicional nas equações do Modelo de Fluxo para levar em conta esta situação, como feito anteriormente para a 9ª série.

As sugestões no Apêndice 1, se implementadas, permitirão ao Censo Educacional coletar dados sobre a época do ano que os alunos abandonam a escola em todas as séries, detectando assim o abandono anterior à coleta da PNAD.

11 A COMPARAÇÃO ENTRE AS TABELAS DE FLUXO COM DADOS DO MEC "CORRIGIDOS" COM O PROFLUXO

O PROFLUXO não considera como aluno novo um aluno afastado por abandono, um repetente reprovado ou um "repetente aprovado". Como os alunos repetentes são estimados pela diferença entre os alunos matriculados e alunos novos, em princípio, a hipótese básica do Modelo de Fluxo de que um aluno matriculado ou é novo na série ou repetente é sempre satisfeita.

Chamamos a atenção que, com os atuais itens do questionário das PNADs, há uma forte contaminação do Sistema de Ensino Supletivo nos dados do Sistema Regular. Este fato superestima os promovidos a partir da 5ª série e, portanto, subestima os repetentes.

Observamos na Tabela 5 que as matrículas estimadas pela PNAD na 3ª, 4ª, 7ª e especialmente na 8ª séries são maiores que as matrículas iniciais correspondentes do Censo Educacional. Este fato sugere que os dados das PNADs podem estar contaminados com alunos matriculados no Ensino Supletivo Seriado

Comparando a Tabela de Fluxo 12 para os anos de 1986 e 1987 com os dados do MEC corrigidos inclusive para os "repetentes aprovados" na 1ª e 2ª séries com a Tabela de Fluxo 4 obtida pelos PROFLUXO para 1987, vemos que da 1ª à 4ª série os resultados são semelhantes, mas diferem a partir da 5ª série. Em particular as taxas de repetência para estas últimas séries são bem maiores que as estimadas pelo PROFLUXO.

TABELA 12

Tabela de Fluxo - Dados MEC - Corrigidos Total - Brasil

Série em 87

		1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	Ev.Ap.	Ev.NA	(Evadid.)	Matr86
S	1ª	3078839	2961325								75806	201270	(277076)	6317240
é	2ª		1567832	2614297							48281	219264	(267545)	4449674
r	3ª			939621	2253588						148114	124929	(273043)	3466252
i	4ª				650297	1726689					316226	103721	(419947)	2796933
e	5ª					1241814	1339173				204992	94877	(299869)	2880856
	6ª						734353	1073809			145823	74303	(220126)	2028288
em	7ª							493534	917897		84083	59179	(143262)	1554693
	8ª								293397	801535	100023	29028	(129051)	1223983
86	9ª									597272				1385840
	N	3264853												
	Matr87	6343692	4529157	3553918	2903885	2968503	2073526	1567343	1211294	1398807				

Série em 87

		1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	Ev.Ap.	Ev.NA	(Evadid.)	Matr86
S	1ª	0.4874	0.4688								0.0120	0.0319	(0.0439)	1.0000
é	2ª		0.3523	0.5875							0.0109	0.0493	(0.0601)	1.0000
r	3ª			0.2711	0.6502						0.0427	0.0360	(0.0788)	1.0000
i	4ª				0.2325	0.6174					0.1131	0.0371	(0.1501)	1.0000
e	5ª					0.4311	0.4649				0.0712	0.0329	(0.1041)	1.0000
	6ª						0.3621	0.5294			0.0719	0.0366	(0.1085)	1.0000
em	7ª							0.3174	0.5904		0.0541	0.0381	(0.0921)	1.0000
	8ª								0.2397	0.6549	0.0817	0.0237	(0.1054)	1.0000
86	9ª									0.4310				1.0000

Sabemos que os dados das PNADs são afetados pela interação com o Sistema de Ensino Supletivo. Não só os concluintes do 1º Grau, via Supletivo, não são distinguidos daqueles que terminaram este grau pelo sistema formal (a Lei, para todos os efeitos assim os considera, como também o questionário das PNADs) como mais grave ainda é o fato de que a partir dos quatorze anos de idade os alunos podem, via Supletivo, sair do sistema formal, "recuperar" o tempo perdido com as repetência nas primeiras séries e voltar em seguida ao sistema formal.

Esta parcela de alunos, considerados como repetentes no ano seguinte, e não captada pelo PROFLUXO, ajuda a explicar por que as estimativas das taxas de repetência com os dados corrigidos do MEC são maiores que as do PROFLUXO.

Esta situação é corroborada por alguns resultados do PROFLUXO (Fletcher & Ribeiro, 1986), em regiões urbanas para níveis socioeconômicos médios e baixos da população. A duração média das últimas séries do 1º Grau chega a ser menor que um (1) ano. Este resultado só pode ser conseqüência da impossibilidade, na coleta dos dados das PNADs, de distinguir aqueles que cursam escola regular daqueles que terminam as séries via cursos supletivos com **séries de curta duração. Uma duração média menor que um (1) indica um número significativo** de estudantes completando as séries através desses cursos.

A inclusão de alunos que cursaram séries de curta duração no cômputo dos alunos que cursam cursos regulares viola a hipótese básica do Modelo de Fluxo e no caso peculiar do PROFLUXO superestima a taxa de promoção e por conseqüência subestima a taxa de repetência.

Este fato, de ser possível sair do Sistema Regular, freqüentar o Supletivo e voltar ao Sistema Regular, afeta, também, os dados do Censo Educacional e o cálculo das taxas de repetência.

É, por conseguinte, difícil com as atuais bases de dados disponíveis obter valores precisos sobre as taxas de repetência, promoção e evasão, para as últimas séries do 1º Grau. No entanto, nestas séries, a contaminação do Sistema Supletivo nos dados do MEC é menor do que nos dados das PNADs, já que pelo menos os concluintes via Supletivo não são confundidos nos dados do MEC.

Fica claro, considerando apenas o sistema formal, que a taxa de graduados calculada pelos dados corrigidos do MEC se aproxima mais da realidade do que a taxa calculada pelo PROFLUXO, e que a taxa de participação, estimável somente pelo PROFLUXO, inclui todos os legalmente graduados no 1º Grau.

Nos Apêndices 1 e 2 apresentamos sugestões para o aprimoramento dos questionários do Censo Educacional e da PNAD e Censo Demográfico, onde estes problemas podem ser minimizados.

12 A SIMULAÇÃO DO PROGRESSO DE UMA COORTE

Um exercício interessante é simular o fluxo hipotético de alunos através do 1º Grau, como feito por Thonstad (1980, p.27) em sua Tabela 2.5.

Tomamos uma amostra de 1000 alunos novos na 1ª série, no Brasil, em um determinado ano e supomos a taxa de repetência na série k, r_k, a taxa de promoção para a série k, p_k, a taxa de evasão de aprovados na série k, $e_{k,a}$ e a taxa de graduação, g, fixas ao longo dos anos até que o último aluno deixe o sistema, independente da idade dos alunos e de terem repetências ou não. Usamos taxas de 1987, para o Brasil, calculadas pelos dados originais do MEC (ver Tabela 2) pelos dados corrigidos do MEC (ver Tabela 12) e pelo PROFLUXO (ver Tabela 4).

Tivemos que modificar os dados originais do MEC relativos à 8ª série, pois a taxa de evasão é *negativa*. **Imputamos uma taxa de evasão igual a 0.0 e uma taxa de graduação de 0.85. Para os dados do MEC corrigidos e do PROFLUXO, a taxa de graduação é dada pela soma da taxa** de promoção para a 9ª série com a taxa de evadidos aprovados da 8ª série. Na simulação feita com os dados originais do MEC não pudemos calcular o número de evadidos aprovados, pois as respectivas taxas são negativas a partir da 2ª série.

As equações que regem a simulação são as seguintes:

$$M(1,t) = M(1,t-1)r_1;$$

$$M(k,t) = M(k-1,t-1)p_k + M(k,t-1)r_k, \quad \text{para} \quad k > 1;$$

$$M(k,+) = \sum_t M(k,t);$$

$$M(+,t) = \sum_k M(k,t);$$

$$M(+,+) = \sum_k M(k,+) = \sum_t M(+,t);$$

$$G(t) = M(8,t)g;$$

$$G = \sum_t G(t);$$

$$E(t) = M(+,t) - M(+,t+1) - G(t);$$

$$E = \sum_t E(t);$$

$$Ac(k+1) = M(k,+)p_k;$$

$$E_a(t) = \sum_k M(k,t)e_{k,a}.$$

onde:

$M(k,t)$ é o número de matrículas nas série k no ano t;

$M(k,+)$ é o número de matrículas utilizadas na série k pela coorte de 1000 alunos, ao longo dos anos;

$M(+,t)$ é o número de matrículas em todas as séries no ano t;

$M(+,+)$ é o número total de matrículas produzidas pelos 1000 alunos ao longo dos anos;

$G(t)$ é o número de graduados no primeiro grau no ano t;

G é o número de graduados no primeiro grau;

$E(t)$ é o número de evadidos em todas as séries no ano t;

E é o número de evadidos sem conclusão do primeiro grau;

$Ac(k)$ é o número de alunos que tiveram acesso à série k;

$E_a(t)$ é o número de alunos evadidos aprovados em todas as séries no ano t.

Em cada ano, para podermos escrever a Tabela com números inteiros como Thonstad (1980), arredondamos os valores obtidos para matrículas, etc.

Os resultados estão apresentados nas Tabelas 13, 14 e 15 para os dados originais do MEC, do MEC corrigidos e do PROFLUXO, respectivamente.

TABELA 13

Simulação do Fluxo de Alunos - Taxas MEC Sem Correção, 1987

Série / Ano	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º	13º	14º	15º	16º	17º	18º	19º	20º	21º	Mat	Acesso
1ª	1000	254	65	17	4	1																1341	1000
2ª		560	267	96	31	9	3	1														967	751
3ª			368	242	107	40	13	4	1													775	636
4ª				265	213	108	45	16	5	1												653	558
5ª					215	225	142	71	30	11	3	1										698	529
6ª						124	155	113	64	30	12	4	1									503	403
7ª							80	113	92	57	29	13	5	1								390	325
8ª								55	84	73	48	25	12	5	1							303	269
Matrículas	1000	814	700	620	570	507	438	373	276	172	92	43	18	6	1							**5630**	
Graduados								49	74	65	43	22	11	4	1							**269**	
Evadidos	186	114	80	50	63	69	65	48	30	15	6	3	1	1								**731**	
Evad. Aprov.	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND							ND	
Evad. s/Ap.	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND							ND	

TABELA 14

Simulação do Fluxo de Alunos - Taxas MEC Corrigidas, 1987

Série / Ano	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º	13º	14º	15º	16º	17º	18º	19º	20º	21º	Matr.	Acesso
1ª	1000	487	237	116	57	28	14	7	3	1												1950	1000
2ª		469	393	250	142	77	40	21	11	5	2	1										1411	914
3ª			276	306	230	146	85	47	25	13	6	3	1									1138	829
4ª				179	241	206	143	89	51	28	15	7	4	2								965	740
5ª					111	197	212	180	133	89	56	33	19	11	6	3	1					1051	596
6ª						52	110	138	134	110	81	55	35	22	13	7	4	2	1			764	489
7ª							28	67	94	101	90	71	52	35	23	14	8	5	3	1		592	404
8ª								17	44	66	75	71	59	45	31	21	13	8	5	3	1	459	350
Matrículas	1000	956	906	851	781	706	632	566	495	413	325	241	170	115	73	45	26	15	9	4	1	**8330**	
Graduados								13	32	49	55	52	43	33	23	15	10	6	4	2	1	**338**	
Evadidos	44	50	55	70	75	74	66	58	50	39	29	19	12	9	5	4	1	0	1	1	0	**662**	
Evad.Aprov.	12	11	19	37	48	48	45	39	31	23	17	11	7	5	2	2	0					357	
Evad. n/Ap	32	39	36	33	27	26	21	19	19	16	12	8	5	4	3	2	1	0	1	1	0	305	

TABELA 15

Simulação do Fluxo de Alunos - Taxas PROFLUXO, 1987

Série / Ano	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º	13º	14º	15º	16º	17º	18º	19º	20º	21º	Matr.	Acesso
1ª	1000	511	261	133	68	35	18	9	5	3	2	1										2046	1000
2ª		460	392	253	147	81	44	23	12	6	3	2	1									1424	942
3ª			285	325	251	164	98	56	30	16	8	4	2	1								1240	882
4ª				186	255	222	158	100	59	33	18	9	5	2	1							1048	808
5ª					116	197	203	165	116	75	45	26	14	8	4	2	1					972	653
6ª						67	129	146	128	95	64	40	24	13	8	4	2	1				721	563
7ª							46	97	117	108	83	58	37	23	13	8	4	2	1			597	498
8ª								34	78	101	98	79	57	38	24	14	8	4	2	1		538	443
Matrículas	1000	971	938	897	837	766	696	630	545	437	321	219	140	85	50	28	15	7	3	1		8586	
Graduados								28	63	82	79	64	46	31	19	11	6	3	2	1		435	
Evadidos	29	33	41	60	71	70	66	57	45	34	23	15	9	4	3	2	2	1				565	
Evad.Aprov.	12	22	32	55	65	65	61	53	41	30	19	11	7	4	2	1						480	
Evad. n/Ap.	17	11	9	5	6	5	5	4	4	4	4	4	2	0	1	1	2	1	0	0		85	

As Tabelas 14 e 15 apresentam resultados semelhantes e muito diferentes da Tabela 13. As simulações com as taxas do PROFLUXO e do MEC corrigidos indicam uma grande permanência dos alunos no sistema de ensino enquanto com as taxas originais do MEC observamos um abandono precoce dos alunos do sistema.

Uma série de indicadores importantes podem ser calculados a partir destas simulações.

O número de alunos-ano (*pupil-years*) que uma coorte passa no 1º Grau pode ser calculado:

$$\text{alunos} - \text{ano} = \frac{M(+,+)}{1000} \quad \text{anos},$$

que corresponde ao número médio de anos que os alunos, que tiveram acesso à escola, passam no sistema (Pela População Escolar). Se multiplicarmos este número pela taxa de participação na 1ª série que, em 1987, foi de 0.935 calculada pelo PROFLUXO, teremos o número médio de anos que *toda* a população jovem passa na escola de 1º Grau, incluindo as crianças que não têm acesso à escola (Pela População Total).

Estes números podem ser desagregados para o número médio de anos gastos na escola pelos alunos que se graduam no 1º Grau (Pelos Graduados) e para aqueles que se evadem sem completar o curso (Pelos *Não* Graduados).

A Tabela 16 mostra estes valores.

TABELA 16
Média de *Anos* de Instrução Recebida

	MEC s/correção	MEC c/correção	PROFLUXO
Pela População Escolar	5,6	8,3	8,6
Pelos Não Graduados	4,1	6,3	6,4
Pelos Graduados	9,9	12,2	11,4
Pela População Total	5,3	7,8	8,0

Estes números mostram realidades bem distintas. Enquanto o número médio de anos freqüentados obtidos a partir dos dados originais do MEC indica que faltam matrículas no sistema para universalizar a educação de 1º Grau e que, portanto, faltam escolas e as crianças se evadem precocemente, os dados corrigidos ou obtidos pelo PROFLUXO indicam, ao contrário, que há matrículas suficientes para que, em princípio, *toda* a população complete este grau de ensino.

É possível calcular também o número de alunos-ano de instrução recebida pela população para que se forme um (1) aluno no 1º Grau. Este número é obtido por:

$$\text{Alunos–ano de instrução/Graduado} = \frac{M(+,+)}{G} \quad \text{anos.}$$

A Tabela 17 mostra estes valores:

TABELA 17

Média de *Anos* de Instrução Recebida pela População Escolar

	MEC s/correção	MEC c/correção	PROFLUXO
Por Graduado	20,9	24,6	19,7

Vemos que os números obtidos pelas taxas do MEC sem correção e do PROFLUXO são surpreendentemente semelhantes. Este fato é conseqüência das taxas originais do MEC indicarem uma evasão precoce e um número de graduados menor, enquanto as taxas do PROFLUXO mostram que os alunos freqüentam a escola por mais tempo e um número maior de alunos se graduam. O número maior para as taxas corrigidas do MEC, muito provavelmente, é conseqüência de termos incluído como repetentes todos os afastados por abandono.

De qualquer forma estes números mostram a fantástica ineficiência de nosso sistema de ensino, já que, se não houvesse repetência, bastariam 8 alunos–ano de instrução para graduar cada aluno no 1º Grau.

A última coluna das Tabelas 13, 14 e 15 mostra o número de alunos que tiveram acesso a cada uma das séries do primeiro grau. Estas colunas são bem diferentes. Por construção, na simulação, o acesso à primeira série é de 1000 alunos. Observa–se que, enquanto segundo as taxas derivadas via dados originais do MEC, somente 751 alunos (75%) dos alunos que entraram na 1ª série têm acesso à 2ª série, 914 alunos (91%) têm acesso a esta série pelas taxas corrigidas do MEC e 942 (94%) têm acesso a esta série pelas taxas do PROFLUXO. Em relação à 8ª série, os números de acesso são 269 (27%), 350 (35%) e 437 (44%), respectivamente para as taxas derivadas via taxas originais do MEC, corrigidas do MEC e taxas do PROFLUXO. Para os graduados no 1º Grau, os números são 269 (27%), 338 (34%) e 435 (44%) respectivamente.

As sugestões do Apêndice 2, se implementadas, permitirão que a metodologia do PROFLUXO estime a taxa de graduados pelo Supletivo, por geração, através dos dados coletados pelas PNADs.

13 AS "PIRÂMIDES EDUCACIONAIS"

Durante os últimos 50 anos, as instituições responsáveis pela divulgação dos indicadores educacionais têm publicado uma pirâmide cuja base é a matrícula na 1ª série num ano t e os demais degraus são as matrículas na série k no ano $t + k - 1$, normalizados para que a matrícula na 1ª série seja igual a 1.000.

Estas pirâmides vêm acompanhadas do comentário: "permitem o acompanhamento do fluxo escolar dos efetivos que ingressam no sistema de ensino e o comportamento desses efetivos". Ver Goldenberg (1990 a, b, c), Baquero Miguel (1990). A seguinte explicação é dada: "De cada 1.000 alunos matriculados na 1ª série em 1980, 527 chegam à 2ª série em 1981, 442 à 3ª série em 1982, 370 à 4ª série em 1983, 383 à 5ª série em 1984, 279 à 6ª série em 1985, 212 à 7ª série em 1986 e 173 à 8ª série em 1987.

Estas explicações e interpretações estão *totalmente equivocadas*, como se vê facilmente, pois *mais alunos alcançam a 5ª série do que a 4ª série*, um óbvio absurdo. Levam também a perigosas conclusões erradas como a de que cerca de 40% dos alunos matriculados na 1ª série não alcançam a 2ª série, quando pela simulações do fluxo de alunos apresentadas nas Tabelas 14 e 15 verifica–se que cerca de 91% e 94% dos alunos matriculados na 1ª série alcançam, respectivamente, a 2ª série em até 20 anos, isto é, somente 9% e 6% não alcançam a 2ª série.

É preciso ter em mente que, assumindo as mesmas hipóteses de estabilidade das taxas das

simulações de fluxo, como a cada ano um fluxo semelhante está ocorrendo, a cada ano estas proporções vão compor o acesso às diversas séries, mas *não* são os alunos que entraram na 1ª série no ano t que estão completando o curso no ano $t+7$.

O fato importante é que não é possível acompanhar o fluxo escolar por esta "pirâmide educacional", pois parte dos alunos da 2ª série no ano $t+1$ são repetentes e, portanto, estavam matriculados na 2ª série no ano t e não na 1ª série. E isto vale para todas as séries.

Utilizando–se a simulação do fluxo, com os dados do MEC, com as taxas de tansição de série corrigidas, só uma pequena parcela (2%, segundo a simulação) dos alunos matriculados na 1ª série no ano t chegam à 8ª série no ano $t+7$, *sem nenhuma nova reprovação.* Os outros alunos matriculados na 8ª série no ano $t+7$ estavam matriculados nas outras séries no ano t. De fato, segundo a simulação, cerca de 35% dos alunos matriculados chegam à 8ª série, mas *após 21 anos e não somente 8 anos.*

Torna–se difícil justificar a construção de grandes complexos educacionais com o resultado dessas análises. Não faltam matrículas, faltam eficiência e qualidade no sistema educacional brasileiro.

14 CONCLUSÕES

Os dados coletados pelo Censo Educacional do MEC parecem estar, hoje em dia, razoavelmente corretos segundo as definições encontradas nos manuais de instrução do Censo.

O problema não está na qualidade dos dados mas na definição de repetentes e na metodologia utilizada para o cálculo das taxas de transição de série pelo MEC.

Repetentes para o MEC são apenas os oficialmente reprovados na série k no ano t e que se matriculam na mesma série k no ano $t+1$. Os alunos afastados por abandono (que tiveram sua matrícula cancelada, não são considerados como reprovados (repetentes na definição do MEC) caso se matriculem na mesma série no ano seguinte.

Como conseqüência eles são considerados como alunos "novos" na série provocando as inconsistências detectadas neste texto. Este "erro" conceitual correspondeu a 13% da matrícula da 1ª série, para o Brasil como um todo, em 1987. Este fato ocorre em *todas* as séries do 1º Grau. Teixeira de Freitas (1947, 1989) já havia detectado este problema com dados dos Censos Educacionais da década de 30!

Outra forma de repetência é observada principalmente na 1ª série do 1º Grau, onde, surpreendentemente, um grande número de alunos oficialmente aprovados nesta série repetem a série no ano seguinte. Estes "repetentes aprovados" representam, para o Brasil como um todo, cerca de 10% da matrícula na 1ª série e 2,6% na 2ª série, em 1987. Mostramos, também, que

o número desses "repetentes aprovados" está diminuindo durante os anos analisados (1984–1987) devido, provavelmente, ao aumento da pré–escola, enquanto a proporção dos afastados por abandono permanece estável no período.

As correções aproximadas introduzidas por nós, nos dados do Censo Educacional, validam os resultados do PROFLUXO da 1ª à 4ª série do 1º Grau e sugerem que as taxas de repetência da 5ª à 8ª série são maiores do que se pensava.

Estas correções devem ser utilizadas com cautela, principalmente da 5ª à 8ª série. O correto seria modificar os questionários dos futuros Censos Educacionais para que se possa utilizar corretamente o Modelo do Fluxo.

A existência do Curso Supletivo Seriado acessível a alunos a partir dos 14 anos afeta os dados tanto das PNADs como os do Censo Educacional a partir da 5ª série. Para uma correta aplicação do Modelo de Fluxo no Sistema Regular de Ensino, os alunos advindos do Supletivo têm que ser considerados à parte, pois são alunos que vêm de fora do Sistema Regular de Ensino.

As sugestões apresentadas nos Apêndices 1 e 2 aos questionários dos Censos Educacionais e das PNADs tentam corrigir estes problemas.

Cremos ser importante estudar e questionar o papel do Curso Supletivo e sua interação com o Sistema Regular de Ensino.

Recomendamos fortemente que, além da correção dos questionários do Censo Educacional, o PROFLUXO seja utilizado de forma oficial, pois seu modelo e metodologia são basicamente corretos, embora detalhes técnicos possam ser questionados e aprimorados. O PROFLUXO introduz ainda novos conceitos e informações que os Censos Educacionais não podem fornecer, como, por exemplo, a taxa de participação.

O uso de duas bases de dados diferentes e complementares é extremamente útil e interessante. Desta maneira, é possível verificar se as duas estão de acordo nas informações comuns. Caso contrário, deve–se procurar os erros e corrigi–los. Deve–se também explorar as informações que não são comuns, como, por exemplo, a possibilidade de desagregar os indicadores por características sociais da população que só é possível através do PROFLUXO. Por outro lado, informações sobre professores e várias características das escolas só podem ser obtidas através dos Censos Educacionais.

Foi comparando a matrícula inicial coletada pelo Censo Educacional no início do ano letivo e a matrícula estimada pela PNAD, pesquisa realizada de setembro a novembro, no fim do ano letivo, que nos permitiu concluir que a grande maioria dos alunos afastados por abandono se afastam da escola no final do ano letivo. Achamos importante que os futuros Censos Educacionais investiguem este assunto e também que haja pesquisa para averiguar as causas deste abandono da escola, e dos "repetentes aprovados", pois representam uma parcela importante da matrícula.

Recomendamos que a crítica e apresentação dos dados dos Censos Educacionais sejam melhoradas. Por exemplo, em 1986, não há dados sobre o Estado de Goiás. Todos os totais para o Brasil e Região Centro–Oeste não incluem Goiás. Logo, não se pode comparar os dados de 1986 para o Brasil com outros anos. Mas é possível, baseado na série histórica, estimar valores para o Estado de Goiás e utilizá–los. É claro que o usuário deve ser informado que isto foi feito.

Recomendamos que mais estudos sejam feitos sobre a confiabilidade e precisão dos dados dos Censos Educacionais, das PNADs e das estimativas do PROFLUXO.

O principal resultado deste trabalho é a conclusão de que não há grandes discrepâncias entre os resultados do PROFLUXO e os resultados do Censo Educacional quando o conceito de repetente é corretamente estabelecido. O erro neste conceito que já dura 50 anos no Brasil e provavelmente, também, nos países do chamado Terceiro Mundo pode ser corrigido. Esta correção permitirá aos governos estabelecerem políticas corretas para a melhoria dos sistemas de ensino básico na perspectiva da imperiosa necessidade, no mundo de hoje, de educar *toda* a população do País.

Finalmente, fazemos votos para que não aconteça novamente o ocorrido em 1947, quando Teixeira de Freitas, Diretor do Serviço de Estatística de Educação e Saúde, em seu artigo na RBEs anunciou para o ano seguinte a correção do Censo Educacional, o que nunca foi feito, e o Brasil não perca mais uma oportunidade de corrigir seu sistema de estatísticas educacionais e passe a utilizar Indicadores Educacionais corretos obtidos a partir do Censo Educacional corrigido e do PROFLUXO.

AGRADECIMENTOS

Os autores agradecem a Telma Suaiden Klein, Alberto Sulaiman Sade Jr. e Lino Oliveira Sobral pelo auxílio computacional e a Sônia Olesko pelo auxílio na elaboração gráfica.

Este trabalho foi parcialmente financiado pela United Nations Project Symbol: BRA/90/026 – M.O.D. nº 91–006, através do Ministério da Educação e pelo Convênio: Ford Foundation / NUPES–USP/LNCC, nº 905–0334.

APÊNDICE 1
SUGESTÕES AO QUESTIONÁRIO DO CENSO EDUCACIONAL

Neste apêndice, apresentamos sugestões de modificação ao questionário do Censo Educacional com a finalidade de obter dados aos quais se possa aplicar o Modelo de Fluxo. Achamos também que é importante saber com certeza em que época do ano os alunos afastados por abandono têm a matrícula cancelada ou se afastam da escola. Sugerimos ainda que o questionário modificado seja testado antes que seja aplicado no Censo Educacional.

SUGESTÕES:

i) Na parte da matrícula inicial, registrar por série e idade o número de alunos matriculados um ou dois meses após o início do ano letivo. Chamar a atenção que só os alunos efetivamente freqüentando a escola nesta data devem ser contados. Esta observação não se encontra no Manual de Instrução do questionário, cuja instrução é informar a matrícula no início do ano letivo, sem data de referência. Os Manuais de Instruções dos questionários trazem a recomendação mais fraca para não incluir as pessoas que se matricularam mas nunca compareceram à escola.

O objetivo destas duas novas instruções é eliminar dupla matrícula e/ou alunos que se matricularam na escola mas nunca compareceram, ou compareceram menos de um mês.

ii) Registrar os repetentes iniciais por série e idade da seguinte maneira:

a) o número de alunos repetentes que foram oficialmente reprovados no ano anterior;

b) o número de alunos repetentes que no ano anterior tiveram a matrícula cancelada pela escola (os afastados por abandono); e

c) o número de alunos "repetentes aprovados" (alunos que no ano anterior foram aprovados numa série na mesma ou em outra escola e estão repetindo a série);

iii) Registrar os alunos afastados por abandono, por série e idade, a partir da data do registro da matrícula inicial, por mês ou bimestre. Isto permitirá saber em que época do ano ocorrem os afastamentos.

iv) Registrar o número de alunos por série e idade, provenientes de fora do Sistema Regular de Ensino, por exemplo:

a) alunos provenientes do Ensino Supletivo;

b) alunos que estiveram fora da escola pelo menos 1 (um) ano; e

c) alunos novos provenientes de migração (de fora do município ou do estado ou do país)

v) Registrar o número de alunos aprovados e reprovados por série e idade.

vi) Orientar no Manual de Instrução do questionário para quando houver 1º Grau não seriado, que os alunos devem ser distribuídos pelas diversas séries de acordo com a equivalência entre os módulos do ano seriado e as séries do Sistema Regular de Ensino.

Sugerimos que, em uma amostra de escola e de professores, fossem coletadas informações específicas sobre o professor (como idade, disciplina(s) que leciona, salário, formação, anos de experiência, se costuma fazer curso de reciclagem, em quantas escolas leciona, número de horas que leciona por semana, etc.) e em uma amostra de alunos, informações específicas sobre o aluno (como idade, características socioeconômicas, etc).

APÊNDICE 2
SUGESTÕES AO QUESTIONÁRIO DE MÃO-DE-OBRA DA PNAD

Neste apêndice, apresentamos sugestões ao questionário de Mão-de-Obra da PNAD do IBGE ou ao questionário do Censo Demográfico, com a finalidade de distinguir o Sistema Regular de Ensino Supletivo e assim permitir estimativas mais corretas pelo PROFLUXO do número de ingressos novos e de aprovados no Sistema Regular. Fazemos a sugestão também de tentar captar o acesso à 1ª série do 1º Grau.

SUGESTÕES:

i) Registrar na pergunta sobre qual foi a última série concluída com êxito, se a pessoa terminou a série no Sistema Regular de Ensino ou se terminou no Ensino Supletivo Seriado ou através de Exame Supletivo;
ii) Se a última série concluída com êxito foi no Ensino Supletivo, registrar qual foi a última série concluída com êxito no Sistema Regular de Ensino;
iii) Se o indivíduo está no 2º ou 3º Grau, perguntar se fez Supletivo, qual a última série do Supletivo e qual a última série concluída com êxito no Sistema Regular de Ensino, anterior ao Supletivo;
iv) Chamar a atenção do entrevistador para distinguir série do Sistema Regular de Ensino de série do Sistema Supletivo Seriado;
v) Registrar se a pessoa freqüentou o Sistema Regular de Ensino no ano anterior;
vi) Registrar para quem não está freqüentando a escola e não concluiu nenhuma série com êxito, se freqüentou a escola alguma vez.

Num questionário de suplemento especial de educação da PNAD, cremos ser possível introduzir perguntas mais detalhadas sobre o histórico escolar da pessoa e inclusive obter dados relativos a migrações para tornar mais precisas as análises de regiões e de grandes estados.

BIBLIOGRAFIA

BAQUERO MIGUEL, G. *Ensino regular de 1º grau.* Evolução das 4 primeiras séries. Brasil 1984-87. 1990. MEC. (Série Estudos Estatísticos).

CUADRA, E. Indicators of student flow rates in Honduras: An assessment of an alternative methodology. *Bridges Research Reports Series* n.6. Cambridge, Havard Graduate School of Education, 1989.

FLETCHER, P. R.; COSTA RIBEIRO, S. Projeto fluxo dos alunos de primeiro grau - PROFLUXO. Protótipo de aplicativo para micro computador, 1986.

———. ———. Projeto fluxo dos alunos de primeiro grau - PROFLUXO. Versão preliminar, 1988. (mimeo.).

———. ———. *Modeling education system performance with demographic data.* An introduction to the PROFLUXO model, 1989. (mimeo.).

GOLDENBERG. M. *Ensino de primeiro grau.* Taxa de promoção no processo. MEC, 1990a. (Série Estudos Estatísticos).

———. *Ensino de primeiro grau.* Taxa de sucesso. MEC, 1990b.

———. *Ensino de segundo grau.* Taxa de sucesso. MEC, 1990c. (Série Estudos Estatísticos)

KLEIN, R. *Relatório 1.* Descrição da metodologia do PROFLUXO. United Nations - Project Symbol: BRA/90/026 - M.O.D. NO. 91-006, 1991a.

———. *Relatório 2.* Comparação de estatísticas regionais e do Brasil. United Nations - Project Symbol: BRA/90/026 - M.O.D. NO. 91-006, 1991b.

———. *Relatório 3.* Comparação das metodologias do MEC e do PROFLUXO. United Nations - Project Symbol: BRA/90/026 - M.O.D. NO. 91-006, 1991c.

———. *Relatório 4.* Apresentação dos principais pontos de discordância e explicação. United Nations - Project Symbol: BRA/90/026 - M.O.D. NO. 91-006, 1991d.

SCHIEFELBEIN, E. Repeating: An overlooked problem in Latin American education. *Comparative Education Review*, v.19, n.3, p.468-87.

SINOPSE ESTATÍSTICA DO ENSINO REGULAR DE 1º GRAU, anos de 1978 a 1987. Brasília, MEC.

SINOPSE ESTATÍSTICA DO ENSINO REGULAR DE 2º GRAU, anos de 1978 a 1987. Brasília, MEC.

TEIXEIRA DE FREITAS, M. A. A escolaridade média no ensino primário brasileiro. *Revista Brasileira de Estatística*, Rio de Janeiro, v.8, n.30/31, p.395-474, 1947.

———. A escolaridade média no ensino primário brasileiro. *Revista Brasileira de Estatística*, Rio de Janeiro, v.50, n.194, p.71-160, 1989 (Republicado com comentários de Ribeiro S. C.).

THONSTAD, T. Analysing and projecting school enrolment in developing countries: A manual of methodology. *Statistical Reports and Studies*, nº 24, Paris, UNESCO, 1980.

RESUMO

Neste trabalho analisamos o Censo Educacional do Ministério da Educação e o Modelo de Fluxo para os alunos num sistema seriado de ensino. Mostramos a inconsistência dos resultados quando se aplica o Modelo de Fluxo aos dados deste Censo. Descobrimos que a causa destas inconsistências está no conceito errado de repetente, e que existem mais duas formas de repetência além da reprovação por aproveitamento ou freqüência: os afastados por abandono durante o ano letivo que retornam à mesma série no ano seguinte e os "repetentes aprovados", alunos que, mesmo aprovados pela escola, repetem a série. Após correções introduzidas por nós, baseadas nestas formas de repetência, a aplicação do Modelo de Fluxo produz resultados consistentes e similares às estimativas do PROFLUXO (o mesmo Modelo de Fluxo, cujos termos são estimados através de uma modelagem estatística utilizando, como base de dados, as PNADS do IBGE). Verificamos a contaminação, tanto dos Censos Educacionais como das PNADS, pelo Sistema de Ensino Supletivo. Finalmente, fazemos sugestões para corrigir estas duas bases de dados para que se possa aplicar corretamente o Modelo de Fluxo.

ABSTRACT

In this paper we analyze the Educational Census of the Ministry of Education of Brazil and the Flow for students in a grade school system. We show that the results are inconsistent when the Flow Model is applied to the data of this Census. We have found that the reason for these inconsistencies is the wrong concept of a repeater, and that there are two more forms of repetition besides the flunking out by achievement or class frequency: those that leave the school during the school year and return to the same grade next year and the "repeaters who have passed the grade", students that, having passed the grade, repeat the same grade in the next year. After corrections introduced by us, based on these new forms of repetition, the application of the Flow Model generates consistent results, similar to the estimation by PROFLUXO (the same Flow Model, with its components estimated through a statistical modeling using as data base the Nacional Household Sample Survey (PNAD) annualy realized by the Brazilian Statistical Office (IBGE)). We verify the inability of these two data bases to distinguish between the Formal Grade System and the Supplementary Educational System that exists in Brazil. Finally, we make suggestions to correct both data bases so that the Flow Model can be applied correctly.

BAYESIAN METHODS IN ACCELERATED LIFE TESTS CONSIDERING A LOG–LINEAR MODEL FOR THE BIRNBAUM–SAUNDERS DISTRIBUTION

Jorge Alberto Achcar*

Mariano Martinez Espinosa*

1 INTRODUCTION

In this paper we explore the use of Bayesian methods in accelerated life tests considering an important fatigue model: the Birnbaum-Saunders distribution. The Birnbaum-Saunders distribution was derived (see Birnbaum and Saunders, 1969 a,b) to model times to failure for metals subject to fatigue due to crack growth under cyclic loading (see also, Mann, Schafer and Singpurwalla, 1974, p. 150).

The Birnbaum-Saunders distribution for a random variable T has probability density function given by

$$f(t;\alpha,\beta) = \frac{(t^2-\beta^2)\exp\left\{-\frac{1}{2\alpha^2}\left[\frac{t}{\beta}+\frac{\beta}{t}-2\right]\right\}}{2\sqrt{2\pi}\alpha\beta t^2\left[\left(\frac{t}{\beta}\right)^{1/2}-\left(\frac{\beta}{t}\right)^{1/2}\right]} \tag{1}$$

where $t>0, \alpha>0$ and $\beta>0$.

The parameters α and β are, respectively, shape and scale parameters. The joint maximum likelihood estimators for α and β were derived by Birnbaum and Saunders (1969b). Tests of

*Instituto de Ciências Matemáticas de São Carlos – USP

hypotheses and confidence intervals for α and β were derived by Engelhardt, Bain and Wright (1981).

The mean and variance of T are given by

$$E(T) = \beta\left(1 + \frac{\alpha^2}{2}\right)$$
$$var(T) = (\alpha\beta)^2\left(1 + \frac{5\alpha^2}{4}\right) \tag{2}$$

(see for example, Mann, Schafer and Singpurwalla, 1974, p. 155).

A Bayesian analysis of the Birnbaum-Saunders distribution with density (1) considering noninformative prior densities for the parameters is given by Achcar (1993).

If the random variable T has a Birnbaum-Saunders distribution with parameters α and β, the random variable $Y = ln(T)$ has a sinh-normal distribution with density,

$$f(y;\alpha,\gamma) = \left\{\frac{2}{2\alpha\sqrt{2\pi}}\right\} cosh\left(\frac{y-\gamma}{2}\right) \times$$
$$\times \exp\left\{-\frac{2}{\alpha^2} sinh^2\left(\frac{y-\gamma}{2}\right)\right\} \tag{3}$$

where $-\infty < y < \infty, \alpha$ is the shape parameter and $\gamma = ln(\beta)$ is a location parameter (see Rieck and Nedelman, 1991).

The sinh-normal distribution has some interesting properties (see Rieck, 1989):

a) The distribution is symmetric about the location parameter γ.

b) The distribution is strongly unimodal for $\alpha \leq 2$ and bimodal for $\alpha > 2$.

c) The mean is given by $E(Y) = \gamma$.

d) If Y_α has a sinh-normal distribution with density (3), then $S_\alpha = (Y_\alpha - \gamma)/(.5\alpha\sigma)$ convergences in distribution to the standard normal distribution as α approaches zero.

Rieck and Nedelman (1991), pointed out that $\alpha > 2$ is unusual in practice. Moreover, if $\alpha > 1$, we have a coefficient of variation exceeding 1, which would not be tolerated in industrial applications (new measurements under more controlled conditions would likely be obtained).

Assuming the log-linear model with a sinh-normal distribution for the error proposed by Rieck and Nedelman (1991), we develop a Bayesian analysis considering one stress variable and using a noninformative prior density for the parameters. We also use the Laplace's method for approximation of integrals, to find simple expressions for the marginal posterior densities of interest. We illustrate the proposed method, working with a life data set introduced by Brown and Miller (1978).

2 A LOG-LINEAR MODEL FOR THE BIRNBAUM–SAUNDERS DISTRIBUTION

Let $T_1, T_2, \ldots, T_n$ be independent random variables with the Birnbaum-Saunders distribution with shape parameter α and scale parameter β_i. The distribution of T_i is assumed to depend on a set of p explanatory variables, denoted by

$$\underline{X}_i^{'} = (X_{i1}, X_{i2}, \ldots, X_{ip}) \quad \textbf{as follows:}$$

a) $\beta_i = exp\{\underline{X}_i^{'}\underline{\Theta}\}$, for $i = 1, 2, \ldots, n$, where $\underline{\Theta}^{'} = (\Theta_1, \Theta_2, \ldots, \Theta_p)$ is a vector of unknown parameters to be estimated.

b) The shape parameter is independent of the explanatory vector $\underline{X}_i$.

Birnbaum and Saunders (1969 b) showed that if $c > 0$, then cT_i has a Birnbaum-Saunders distribution with shape parameter $c\beta_i$. Using this fact, and the assumptions mentioned previously, we see that T_i may be expressed as $T_i = \delta_i exp\{\underline{X}_i^{'}\underline{\Theta}\}\delta_i$, where δ_i is distributed according to a Birnbaum-Saunders distribution with shape parameter α and scale parameter 1. Thus, we consider a log-linear model,

$$Y_i = ln(T_i) = \underline{X}_i^{'}\underline{\Theta} + \varepsilon_i \tag{4}$$

where $\varepsilon_i = ln(\delta_i)$ is the error term for the model with a sinh-normal distribution with (3) parameters α and $\gamma = 0$ (see Rieck and Nedelman, 1991).

A special case of the log-linear model (4) applied to accelerated life tests considering only one stress variable $x_i, i = 1, 2, \ldots, n$, is given by

$$Y_i = ln\,(T_i) = a + bx_i + \varepsilon_i \tag{5}$$

Under the model (5), the likelihood function for a, b and α is given by

$$L(a, b, \alpha) = \left(\frac{1}{2\sqrt{2\pi}}\right)^n \left\{\prod_{i=1}^{n} \frac{2}{\alpha} cosh\left(\frac{y_i - a - bx_i}{2}\right)\right\} \times$$

$$\times\, exp\left\{-\sum_{i=1}^{n} \frac{2}{\alpha^2} sinh^2\left(\frac{y_i - a - bX_i}{2}\right)\right\} \tag{6}$$

The log-likelihood function for a, b and α is given by

$$lnL(a,b,\alpha) \propto \sum_{i=1}^{n} lnW_i - \sum_{i=1}^{n} \frac{Z_i^2}{2} \tag{7}$$

where $W_i = \frac{2}{\alpha} cosh\left(\frac{y_i - a - bx_i}{2}\right)$ and

$$Z_i = \frac{2}{\alpha} sinh\left(\frac{y_i - a - bx_i}{2}\right)$$

Differentiating $ln(L)$ with respect to a, b and α, we obtain the likelihood equations

$$\begin{aligned} \frac{\partial lnL}{\partial a} &= \frac{1}{2}\sum_{i=1}^{n}\{Z_i W_i - Z_i/W_i\} = 0 \\ \frac{\partial lnL}{\partial b} &= \frac{1}{2}\sum_{i=1}^{n} x_i \{Z_i W_i - Z_i/W_i\} = 0 \end{aligned} \tag{8}$$

and

$$\frac{\partial lnL}{\partial \alpha} = -\frac{n}{\alpha} + \frac{1}{\alpha}\sum_{i=1}^{n} Z_i^2 = 0$$

An expression for the maximum likelihood estimator of α^2 in terms of the maximum likelihood estimators for a and b may be found by solving the equation $\partial lnL/\partial\alpha = 0$ for α^2 and is given by

$$\hat{\alpha}^2 = \frac{4}{n}\sum_{i=1}^{n} sinh^2\left(\frac{y_i - \hat{a} - \hat{b}x_i}{2}\right) \tag{9}$$

where $\hat{a}$ and $\hat{b}$ are the maximum likelihood estimators for a and b. The maximum likelihood estimators for a and b are obtained by using a numerical procedure (see Rieck and Nedelman, 1991).

For large values of n, hypotheses tests and confidence intervals on a, b and α can be based on the usual asymptotic normality of the maximum likelihood estimators $\hat{a}, \hat{b}$, and $\hat{\alpha}$ given by

$$\left(\hat{a}, \hat{b}, \hat{\alpha}\right) \underset{\sim}{a} N\left\{(a,b,\alpha); I_0^{-1}\right\} \tag{10}$$

where I_0 is the observed information matrix given by

$$I_0 = \begin{pmatrix} -\frac{1}{4}\hat{S}_0 & -\frac{1}{4}\hat{S}_1 & \frac{1}{\hat{\alpha}}\sum_{i=1}^{n}\hat{Z}_i\hat{W}_i \\ & -\frac{1}{4}\hat{S}_2 & \frac{1}{\hat{\alpha}}\sum_{i=1}^{n} x_i\hat{Z}_i\hat{W}_i \\ symmetric & & \frac{1}{\hat{\alpha}^2}\left(3\sum_{i=1}^{n}\hat{Z}_i^2 - n\right) \end{pmatrix} \tag{11}$$

where $\hat{W}_i = \frac{2}{\hat{\alpha}} cosh\left(\frac{y_i - \hat{a} - \hat{b}x_i}{2}\right)$,

$$\hat{Z}_i = \frac{2}{\hat{\alpha}} sinh\left(\frac{y_i - \hat{a} - \hat{b}x_i}{2}\right), \quad \text{and}$$

$$\hat{S}_l = \sum_{i=1}^{n} x_i^l \left(1 - \hat{W}_i^2 - \hat{Z}_i^2 - \frac{\hat{Z}_i^2}{\hat{W}_i^2}\right), \qquad l = 0, 1, 2 .$$

We also could use the Fisher information matrix $I(a, b, \alpha)$ in place of the observed information matrix I_0 in (10). The Fisher information matrix for a, b and α is given by

$$I(a, b, \alpha) = \begin{pmatrix} C_0(\alpha) & C_1(\alpha) & 0 \\ & C_2(\alpha) & 0 \\ symmetric & & \frac{2n}{\alpha^2} \end{pmatrix} \tag{12}$$

where $C_l(\alpha) = \frac{1}{4}\sum_{i=1}^{n} x_i^l \left\{1 + \frac{4}{\alpha^2} + E\left(\frac{Z_i^2}{Z_i^2 + 4/\alpha^2}\right)\right\}$, for $l = 0, 1, 2$.

For small values of $\alpha (0 < \alpha < 1)$, we can consider $C_l(\alpha) \approx \frac{1}{4}\sum_{i=1}^{n} x_i^l \left\{1 + \frac{4}{\alpha^2}\right\}, l = 0, 1, 2$. Therefore, a simplified form for $I(a, b, \alpha)$ is given by

$$I(a, b, \alpha) = \begin{pmatrix} \frac{n}{4\alpha^2}(4 + \alpha^2) & \frac{(4+\alpha^2)}{4\alpha^2}\sum_{i=1}^{n} x_i & 0 \\ & \frac{(4+\alpha^2)}{4\alpha^2}\sum_{i=1}^{n} x_i^2 & 0 \\ symmetric & & \frac{2n}{\alpha^2} \end{pmatrix} \tag{13}$$

3 A BAYESIAN ANALYSIS OF THE LOG-LINEAR MODEL CONSIDERING ONE STRESS VARIABLE

Assuming the log-linear model (5) with one stress variable x, the Jeffreys prior density for a, b and α is given by

$$\pi(a, b, \alpha) \propto \{detI(a, b, \alpha)\}^{1/2} \tag{14}$$

where $I(a, b, \alpha)$ is the Fisher information matrix (12) (see for example, Box and Tiao, 1973).

Considering the approximate simplified form of the Fisher information matrix (13), a noninformative prior density for a, b and α is given by

$$\pi(a, b, \alpha) \propto \frac{(4 + \alpha^2)}{\alpha^3} \tag{15}$$

where $\alpha > 0$ and $-\infty < a, b < \infty$.

We also could consider other priors for a, b and α. If we assume prior independence among the parameters (see for example, Box and Tiao, 1973), we could consider another noninformative prior given by

$$\pi(a, b, \alpha) \propto \frac{1}{\alpha} \tag{16}$$

where $\alpha > 0$ and $-\infty < a, b < \infty$.

Considering the Jeffreys prior density (15), the joint posterior density for a, b and α is given by

$$\pi(a, b, \alpha | Data) \propto \frac{(4+\alpha^2)}{\alpha^{n+3}} \left\{ \prod_{i=1}^{n} cosh\left(\frac{y_i - a - bx_i}{2}\right) \right\} \times$$
$$\times \exp\left\{ -\frac{2}{\alpha^2} \sum_{i=1}^{n} sinh^2\left(\frac{y_i - a - bx_i}{2}\right) \right\} \tag{17}$$

for $\alpha > 0$ and $-\infty < a, b < \infty$.

3.1 Marginal Posterior Density For α

The marginal posterior density for α is obtained by integrating out a and b in the joint posterior density (17). That is,

$$\pi(\alpha | Data) \propto \frac{(4+\alpha^2)}{\alpha^{n+3}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(a, b) e^{-nh(a,b)} da db \,, \tag{18}$$

where $f(a, b) = \prod_{i=1}^{n} cosh\left(\frac{y_i - a - bx_i}{2}\right)$ and

$$nh(a, b) = \frac{2}{\alpha^2} \sum_{i=1}^{n} sinh^2\left(\frac{y_i - a - bx_i}{2}\right)$$

Thus, a Laplace's approximation (see appendix) for the integral in (18), gives an approximate marginal posterior density for α given by

$$\pi(\alpha | Data) \propto \frac{(4+\alpha^2)}{\alpha^{n+1}} exp\left\{ -\frac{2A\left(\hat{a}, \hat{b}\right)}{\alpha^2} \right\} \,, \tag{19}$$

where $\alpha > 0, A\left(\hat{a}, \hat{b}\right) = \sum_{i=1}^{n} sinh^2\left(\frac{y_i - \hat{a} - \hat{b}x_i}{2}\right)$, $\hat{a}$ and $\hat{b}$ are solutions of the equations,

$$\sum_{i=1}^{n} sinh\left(y_i - \hat{a} - \hat{b}x_i\right) = 0$$

and

$$\sum_{i=1}^{n} x_i sinh\left(y_i - \hat{a} - \hat{b}x_i\right) = 0 .$$

Considering the noninformative prior density (16), the Laplace's approximate marginal posterior density for α is given by

$$\pi(\alpha|Data) = \frac{2^{\frac{n}{2}} A^{\frac{n-2}{2}}\left(\hat{a},\hat{b}\right)}{\Gamma\left(\frac{n-2}{2}\right)} \alpha^{-(n-1)} exp\left\{-\frac{2A\left(\hat{a},\hat{b}\right)}{\alpha^2}\right\} \tag{20}$$

where $\alpha > 0$; $\hat{a}$ and $\hat{b}$ are given (19).

The mode of the marginal posterior density (20) is given by

$$\tilde{\alpha}^{*^2} = \frac{4A\left(\hat{a},\hat{b}\right)}{n-1} , \tag{21}$$

where $A\left(\hat{a},\hat{b}\right)$ is given in (19).

3.2 Joint Marginal Posterior Density For a And b

The joint marginal posterior density for a and b is given (from (17)) by

$$\pi(a,b|Data) \propto \left\{\prod_{i=1}^{n} cosh\left(\frac{y_i - a - bx_i}{2}\right)\right\} \int_0^{\infty} f(\alpha) e^{-nh(\alpha)} d\alpha , \tag{22}$$

where $f(\alpha) = \left(4+\alpha^2\right)/\alpha^3$ and

$$nh(\alpha) = nln\alpha + \frac{2}{\alpha^2}\sum_{i=1}^{n} sinh^2\left(\frac{y_i - a - bx_i}{2}\right) .$$

A Laplace's approximate joint marginal posterior density for a and b is given by

$$\pi(a,b|Data) \propto \left\{\prod_{i=1}^{n} cosh\left(\frac{y_i - a - bx_i}{2}\right)\right\} \times$$

$$\times \frac{\left\{1 + \frac{1}{n}\sum_{i=1}^{n} sinh^2\left(\frac{y_i-a-bx_i}{2}\right)\right\}}{\left\{\sum_{i=1}^{n} sinh^2\left(\frac{y_i-a-bx_i}{2}\right)\right\}^{\frac{n+2}{2}}}, \tag{23}$$

where $-\infty < a, b < \infty$.

Considering the prior (16), the joint marginal posterior density for a and b is given by

$$\pi(a,b|Data) \propto \frac{\left\{\prod_{i=1}^{n} cosh\left(\frac{y_i-a-bx_i}{2}\right)\right\}}{\left\{\sum_{i=1}^{n} sinh^2\left(\frac{y_i-a-bx_i}{2}\right)\right\}^{n/2}}, \tag{24}$$

where $-\infty < a, b < \infty$.

4 A BAYESIAN ANALYSIS ASSUMING α KNOWN

Assuming α known, the likelihood function for a and b is given by

$$L(a,b) \propto \left\{\prod_{i=1}^{n} cosh\left(\frac{y_i-a-bx_i}{2}\right)\right\} \times$$
$$\times exp\left\{-\frac{2}{\alpha^2}\sum_{i=1}^{n} sinh^2\left(\frac{y_i-a-bx_i}{2}\right)\right\}. \tag{25}$$

Considering the noninformative prior density for a and b locally uniform, an approximate Laplace's marginal posterior density for a is given by

$$\pi(a|Data) \propto \frac{\left\{\prod_{i=1}^{n} cosh\left(\frac{y_i-a-\hat{b}x_i}{2}\right)\right\}}{\left\{\sum_{i=1}^{n} x_i^2 cosh\left(y_i-a-\hat{b}x_i\right)\right\}^{1/2}} \times$$
$$\times exp\left\{-\frac{2}{\alpha^2}\sum_{i=1}^{n} sinh^2\left(\frac{y_i-a-\hat{b}x_i}{2}\right)\right\}, \tag{26}$$

where $-\infty < a < \infty$, and $\hat{b}$ is given by

$$\sum_{i=1}^{n} x_i e^{y_i-a} e^{-\hat{b}x_i} = \sum_{i=1}^{n} x_i e^{-(y_i-a)} e^{-\hat{b}x_i}.$$

In the same way, we obtain an approximate marginal posterior density for b given by

$$\pi(b|Data) \propto \frac{\left\{\prod_{i=1}^{n} cosh\left(\frac{y_i-\hat{a}-bx_i}{2}\right)\right\}}{\left\{\sum_{i=1}^{n} cosh\left(y_i-\hat{a}-bx_i\right)\right\}^{1/2}} \times$$
$$\times exp\left\{-\frac{2}{\alpha^2}\sum_{i=1}^{n} sinh^2\left(\frac{y_i-\hat{a}-bx_i}{2}\right)\right\}, \tag{27}$$

where $-\infty < b < \infty$ and $\hat{a}$ is given by

$$\hat{a} = \frac{1}{2} ln \left\{ \frac{\sum_{i=1}^{n} e^{y_i - bx_i}}{\sum_{i=1}^{n} e^{-(y_i - bx_i)}} \right\}$$

for each value of b.

4.1 Posterior Density For The Mean Life Time In A Specified Stress Level x^*

Usually, industrial researchers have interest in inferences on the mean life time $\gamma^* = E(Y^*) = a + bx^*$, considering a specified stress level x^*.

Assuming α known, we consider the transformation of variables given by $\gamma^* = a + bx^*$ and b. Considering a locally uniform prior for a and b, the joint posterior density for γ^* and b is given by

$$\pi(\gamma^*, b|Data) \propto \left\{ \prod_{i=1}^{n} cosh \left(\frac{y_i - \gamma^* + b(x^* - x_i)}{2} \right) \right\} \times$$

$$\times exp \left\{ -\frac{2}{\alpha^2} \sum_{i=1}^{n} sinh^2 \left(\frac{y_i - \gamma^* + b(x^* - x_i)}{2} \right) \right\}, \tag{28}$$

where $-\infty < \gamma^* < \infty$.

An approximate Laplace's marginal posterior density for γ^* is given by

$$\pi(\gamma^*|Data) \propto \frac{\left\{ \prod_{i=1}^{n} cosh \left(\frac{y_i - \gamma^* + \hat{b}(x^* - x_i)}{2} \right) \right\}}{\left\{ \sum_{i=1}^{n} (x^* - x_i)^2 cosh \left(y_i - \gamma^* + \hat{b}(x^* - x_i) \right) \right\}^{1/2}} \times$$

$$\times exp \left\{ -\frac{2}{\alpha^2} \sum_{i=1}^{n} sinh^2 \left(\frac{y_i - \gamma^* + \hat{b}(x^* - x_i)}{2} \right) \right\}, \tag{29}$$

where $-\infty < \gamma^* < \infty$, and $\hat{b}$ satisfies

$$x^* \sum_{i=1}^{n} sinh \left[y_i - \gamma^* + \hat{b}(x^* - x_i) \right] =$$

$$= \sum_{i=1}^{n} x_i sinh \left[y_i - \gamma^* + \hat{b}(x^* - x_i) \right].$$

5 PREDICTIVE DENSITY FOR A FUTURE OBSERVATION

Assuming α known, the predictive density for a future observation $Y^*_{(n+1)}$ considering a specified stress level x^*, is given (see for example, Press, 1989) by

$$f\left(y^*_{(n+1)}|Data\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f\left(y^*_{(n+1)}|a,b\right)\pi(a,b|Data)dadb, \tag{30}$$

where

$$f\left(y^*_{(n+1)}|a,b\right) = \frac{1}{\sqrt{2\pi}\alpha} cosh\left(\frac{y^*_{(n+1)} - a - bx^*}{2}\right) \times$$
$$\times \exp\left\{-\frac{2}{\alpha^2} sinh^2\left(\frac{y^*_{(n+1)} - a - bx^*}{2}\right)\right\}$$

and $\pi(a,b|Data)$ is the joint posterior density for a and b. Considering a locally uniform prior for a and b, we have,

$$f\left(y^*_{(n+1)}|Data\right) \propto \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x^*}(a,b)e^{-nh_{x^*}(a,b)}dadb\,, \tag{31}$$

where

$$fx^*(a,b) = cosh\left(\frac{y^*_{(n+1)} - a - bx^*}{2}\right)\prod_{i=1}^{n} cosh\left(\frac{y_i - a - bx_i}{2}\right)$$

and

$$nh_{x^*}(a,b) = \frac{2}{\alpha^2}\left[sinh^2\left(\frac{y^*_{(n+1)} - a - bx^*}{2}\right) + \right.$$
$$\left. + \sum_{i=1}^{n} sinh^2\left(\frac{y_i - a - bx_i}{2}\right)\right].$$

A Laplace's approximation for (31) is given by

$$f\left(y^*_{(n+1)}|Data\right) \propto \frac{cosh\left(\frac{y^*_{(n+1)} - \hat{a} - \hat{b}x^*}{2}\right)\prod_{i=1}^{n} cosh\left(\frac{y_i - \hat{a} - \hat{b}x_i}{2}\right)}{B\left(y^*_{(n+1)}\right)} \times$$
$$\times\ exp\left\{-\frac{2}{\alpha^2}\left[sinh^2\left(\frac{y^*_{(n+1)} - \hat{a} - \hat{b}x^*}{2}\right) + \sum_{i=1}^{n} sinh^2\left(\frac{y_i - \hat{a} - \hat{b}x_i}{2}\right)\right]\right\}, \tag{32}$$

where $y^*_{(n+1)} > 0$ and

$$B\left(y^*_{(n+1)}\right) = \left\{\left[cosh\left(y^*_{(n+1)} - \hat{a} - \hat{b}x^*\right) + V_0\right]\left[x^* cosh\left(y^*_{(n+1)} - \hat{a} - \hat{b}x^*\right) + V_2\right] - \right.$$
$$\left. - \left[x^* cosh\left(y^*_{(n+1)} - \hat{a} - \hat{b}x^*\right) + V_1\right]^2\right\}^{1/2},$$

$$V_l = \sum_{i=1}^{n} x_i^l cosh\left(y_i - \hat{a} - \hat{b}x_i\right), \quad l = 0, 1, 2.$$

6 AN EXAMPLE

In table 1, we have a data set introduced by Brown and Miller (1978). These data represent results of fatigue tests on 1% Cr-Mo-V steel, where cylindrical specimens were subjected to combined torsional and axial loads over constant-amplitude cycles until failure. From empirical laws, it was considered the model $ln(N) = a + bx + error, x = ln(W_c)$, where W_c is the work per cycle and N is the number of cycles to failure of each specimen. This data set was considered and analized by Rieck and Nedelman, 1991.

TABLE 1
BROWN AND MILLER'S BIAXIAL FATIGUE DATA

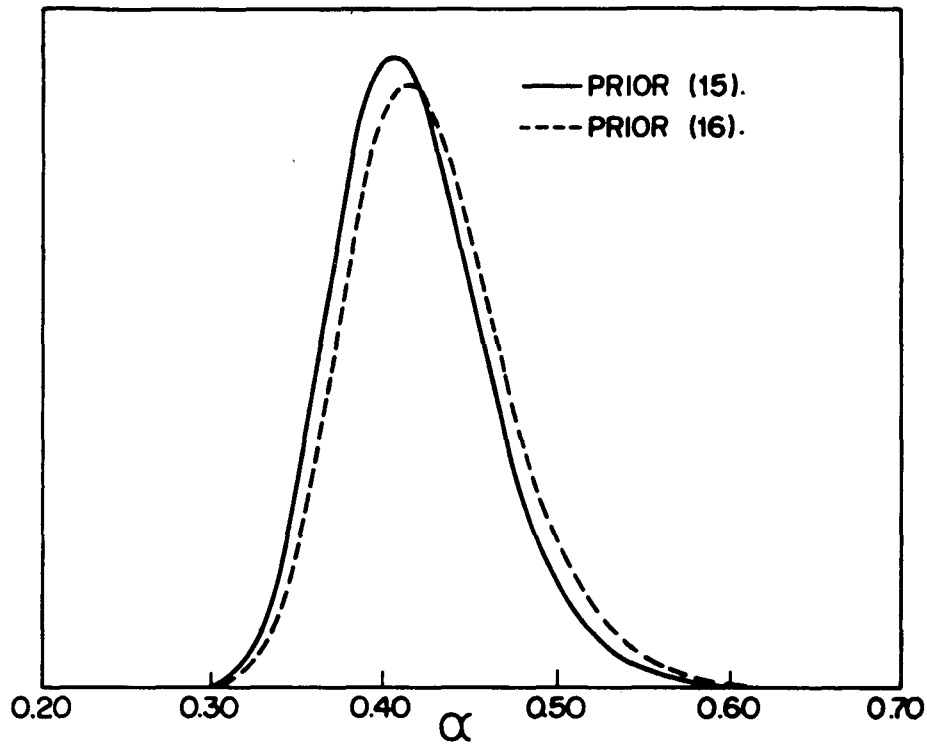
W_c	N	W_c	N	W_c	N	W_c	N
11.5	3280	24.0	804	40.1	750	60.3	283
13.0	5046	24.6	1093	40.1	316	60.5	212
14.3	1563	25.2	1125	43.0	456	62.1	327
15.6	4707	25.5	884	44.1	552	62.8	373
16.0	977	26.3	1300	46.5	355	66.5	125
17.3	2834	27.9	852	47.3	242	67.0	187
19.3	2266	28.3	580	48.7	190	67.1	135
21.1	2208	28.4	1066	52.9	127	67.9	245
21.5	1040	28.6	1114	56.6	185	68.8	137
22.6	700	30.9	386	59.9	255	75.4	200
22.6	1583	31.9	745	60.2	195	100.5	190
24.0	482	34.5	736				

The maximum likelihood estimators for a, b and α (with estimated standard errors considering the observed information matrix (11)) are, respectively, 12.28(0.3890),-1.671 (0.1083) and 0.41 (0.0427). The approximate 90% confidence intervals for a, b and α are, respectively, (11.52,13.04), (-1.88,-1.48) and (0.33,0.49). Considering the simplified form of the Fisher information matrix (13), the 90% approximate confidence intervals for a, b and α are, respectively, (11.63,12.93), (-1.85,-1.49) and (0.34,0.48).

In figure 1, we have the graphs of the marginal posterior densities for α given in (19) and (20), assuming the noninformative priors (15) and (16), respectively. We observe close results considering both noninformative priors (15) and (16) for a, b and α, that is, in practical work, we could use the simplified form of the noninformative prior (16) assuming independence "a priori" of the parameters. An approximate 90% credible interval for α considering the marginal posterior density (19) is given by (0.352,0.498). The mode of the marginal posterior density (19) is given by $\tilde{\alpha}^* = 0.406$.

Figure 1

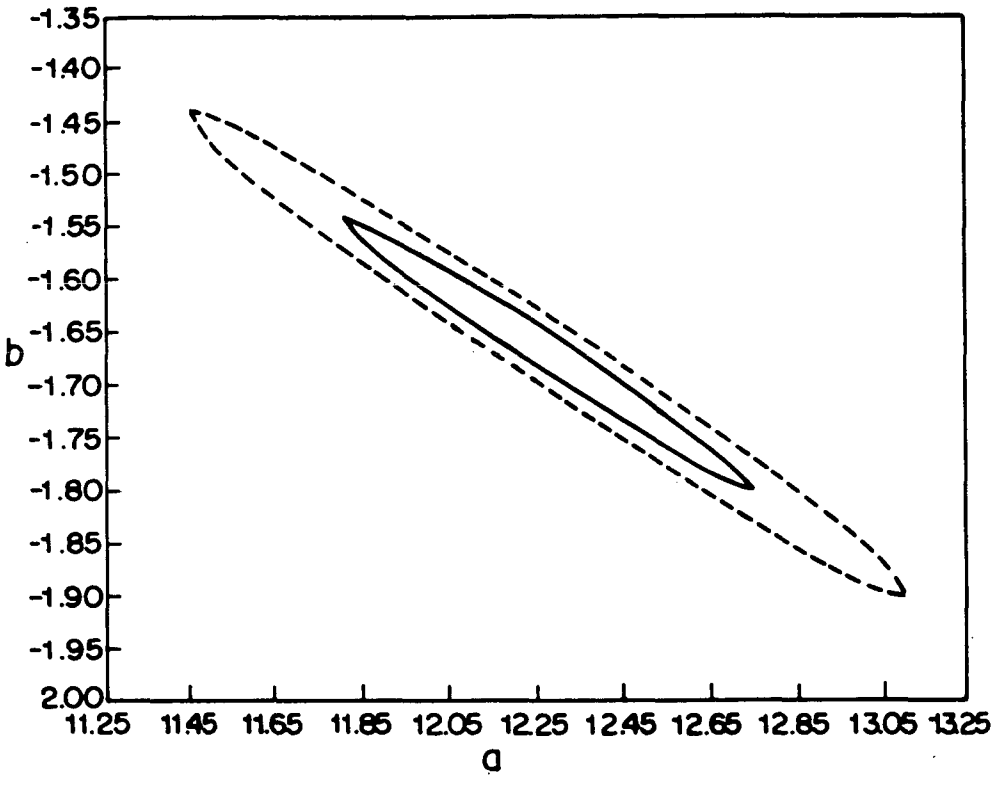
Marginal Posterior Density For α



In figure 2, we have contour plots for the joint marginal posterior density for a and b given in (23). The mode of this marginal posterior density is given by $\tilde{a}^* = 12.28$ and $\tilde{b}^* = -1.671$. From the joint marginal posterior density for a and b, we find approximate 90% credible intervals for a and b, given by (11.65,12.91) and (-1.85,-1.50), respectively.

Figure 2

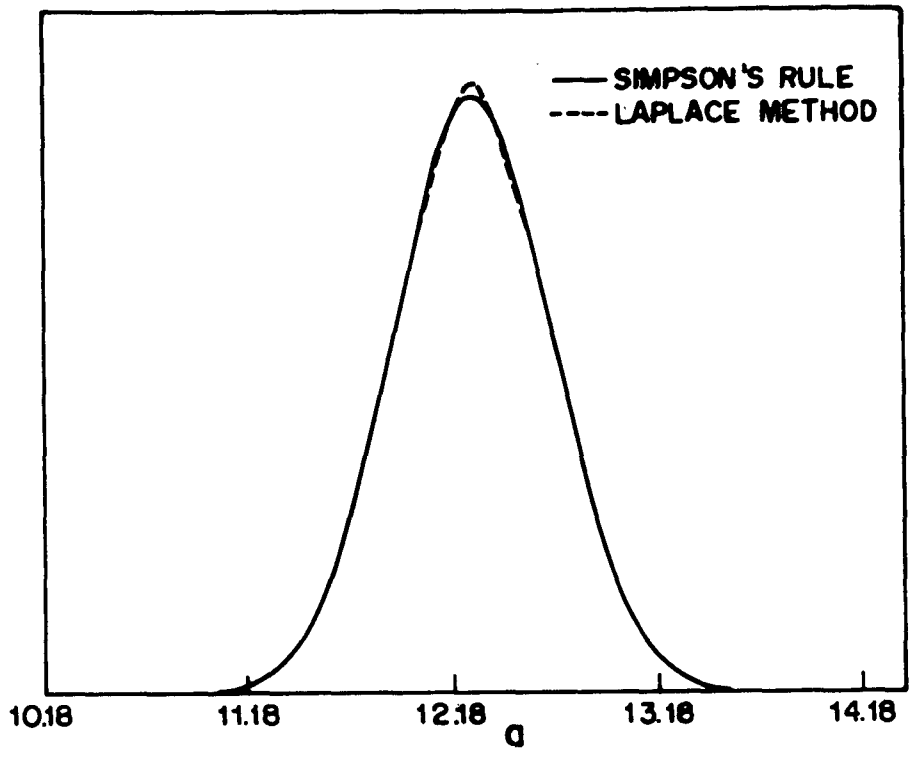
Contour Plots For The Joint Posterior Density For a and b



In figure 3, we have the graph of the approximate marginal posterior density for a (given in (26)), assuming $\alpha = 0.4$ known. We also have in figure 3, the marginal posterior density for a obtained by numerical integration. We observe good accuracy of the Laplace's approximation (see table 2). A 90% approximate credible interval for a considering $\alpha = 0.4$ known, is given by (11.66,12.91). The mode of the marginal posterior density for a (26) is given by $\tilde{a}^* = 12.28$. We observe almost the same results considering α known or unknown.

Figure 3

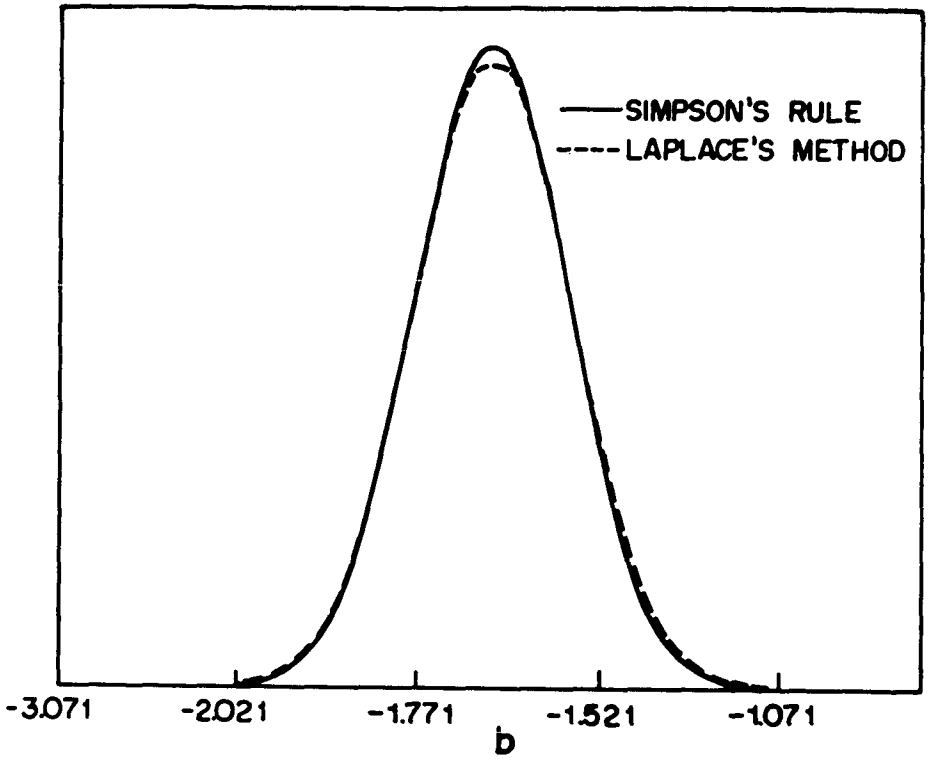
Marginal Posterior Density For a Assuming $\alpha = 0.4$ known



In figure 4, we have the graph of the approximate marginal posterior for b (27), assuming $\alpha = 0.4$ known. We observe good accuracy of the Laplace's approximation (see table 2). A 90% approximate credible interval for b considering $\alpha = 0.4$ known, is given by (-1.85,-1.50).

Figure 4

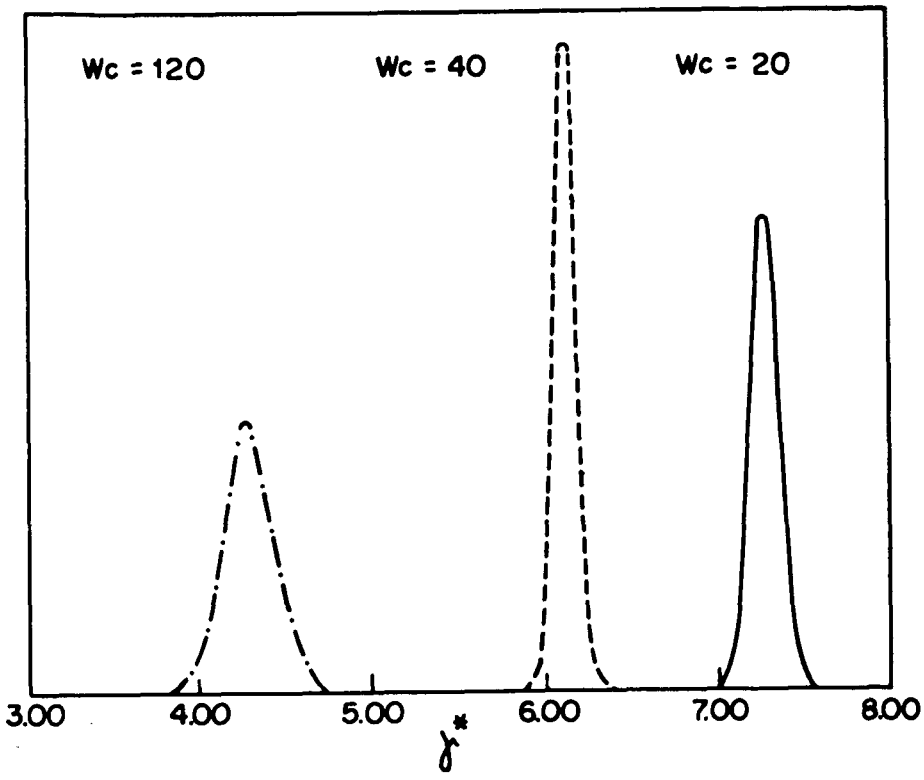
Marginal Posterior Density For b Assuming $\alpha = 0.4$ known



The mode of the marginal posterior density for b (27) is given by $\tilde{b}^* = -1.671$.

Figure 5

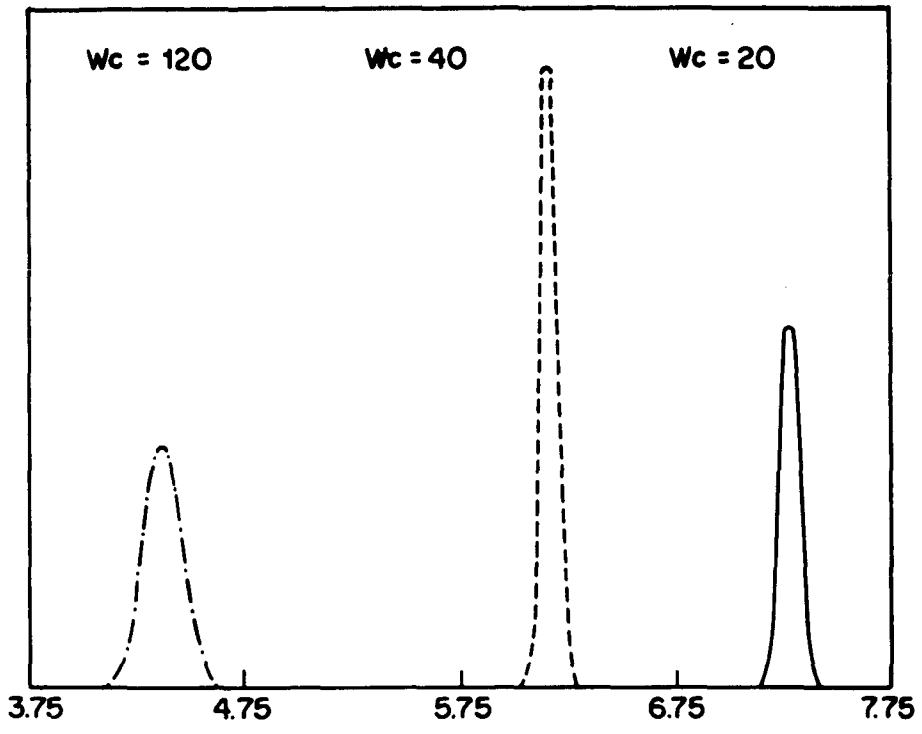
Marginal Posterior Density For The Mean Life Time $\gamma^* = a + bx^*$ considering $W_c = 20$, $W_c = 40$ and $W_c = 120$



In figure 5, we have the graphs of the approximate marginal posterior densities for the mean life time $\gamma^* = a + bx^*$ given in (29), considering $W_c = 20, W_c = 40$ and $W_c = 120$.

Figure 6

Predictive Density For $Y_{(n+1)}$ Considering $W_c = 20$, $W_c = 40$ And $W_c = 120$



In figure 6, we have the graphs of the predictive densities for a future observation $Y_{(n+1)} = ln(N)$ considering $W_c = 20, W_c = 40$ and $W_c = 120$.

TABLE 2

SOME VALUES OF $\pi(a|Data)$ AND $\pi(b|Data)$ CONSIDERING THE LAPLACE'S METHOD AND THE SIMPSON'S RULE (α known)

a	$\pi(a\|\text{Data})$		b	$\pi(b\|\text{Data})$	
	SIMPSON	LAPLACE		SIMPSON	LAPLACE
10.48	0.0000	0.0000	−2.191	0.0000	0.0000
10.68	0.0001	0.0001	−2.151	0.0001	0.0001
10.88	0.0011	0.0011	−2.091	0.0012	0.0016
11.08	0.0069	0.0070	−2.031	0.0104	0.0125
11.28	0.0322	0.0316	−1.971	0.0639	0.0715
11.48	0.1134	0.1154	−1.911	0.2791	0.2960
11.68	0.3012	0.2941	−1.851	0.8750	0.8921
11.88	0.6044	0.6148	−1.791	1.9743	1.9610
12.08	0.9174	0.8928	−1.731	3.2148	3.1523
12.28	1.0538	1.0720	−1.671	3.7840	3.7115
12.48	0.9161	0.8899	−1.611	3.2234	3.2039
12.68	0.6024	0.6128	−1.551	1.9878	2.0284
12.88	0.2993	0.2905	−1.491	0.8871	0.9416
13.08	0.1121	0.1141	−1.431	0.2862	0.3201
13.28	0.0316	0.0307	−1.371	0.0667	0.0796
13.48	0.0067	0.0068	−1.311	0.0112	0.0144
13.68	0.0011	0.0010	−1.251	0.0013	0.0019
13.88	0.0001	0.0001	−1.191	0.0001	0.0002
14.08	0.0000	0.0000	−1.131	0.0000	0.0000

7 OVERALL CONCLUSIONS

Using the Laplace's method for integrals, we obtained simple expressions for the marginal posterior densities of interest considering accelerated life tests with a log-linear model for the Birnbaum-Saunders distribution proposed by Rieck and Nedelman (1991). We observed similar inference results from the classical or Bayesian methodology. These results could be of great practical interest.

One could also consider more than one stress variable in the log–linear model (4) and other priors for the parameters of interest.

APPENDIX
THE LAPLACE'S METHOD

Assuming h is a smooth function of an m-dimensional parametr Θ with $-h$ having a maximum at $\hat{\Theta}$, Laplace's method approximates an integral of the form,

$$I = \int f(\Theta) exp[-nh(\Theta)] d\Theta \tag{A.1}$$

by expanding h and f in a Taylor series about $\hat{\Theta}$ (see for example, Kass, Tierney and Kadane, 1990).

Considering first the case in which Θ is one-dimensional, the Laplace's method gives the approximation,

$$\hat{I} \cong \left(\frac{2\pi}{n}\right)^{1/2} \sigma f\left(\hat{\Theta}\right) exp\left[-nh\left(\hat{\Theta}\right)\right] \tag{A.2}$$

where $\sigma = \left\{h''\left(\hat{\Theta}\right)\right\}^{-1/2}$.

In the multiparameter case, with $\Theta \in R^m$, we have,

$$\hat{I} \cong (2\pi)^{m/2} \left\{det\left(nD^2 h\left(\hat{\Theta}\right)\right)\right\}^{-1/2} f\left(\hat{\Theta}\right) exp\left[-nh\left(\hat{\Theta}\right)\right] \tag{A.3}$$

where $\hat{\Theta}$ maximizes $-h(\Theta)$ and $D^2 h(\Theta)$ is the Hessian matrix of h evaluated at $\hat{\Theta}$.

The accuracy of these approximations are studied by Kass, Tierney and Kadane (1990). A special case of the Laplace's approximation is given for integral of the form $\int e^{nh(\Theta)} d\Theta$ (see Tierney and Kadane, 1986; Tierney, Kass and Kadane, 1989).

BIBLIOGRAPHY

ACHCAR, J. A. Inferences for the Birnbaum-Saunders fatigue life model using Bayesian Methods. *Computational Statistics & Data Analysis*, v.15, n.4, p.367-80, 1993.

BIRNBAUM, Z. W.; SAUNDERS, S. C. A new family of life distribution, *Journal of Applied Probability*, v.6, p. 319-27, 1969.

———. ———. Estimation for a family of life distributions with applications to fatigue, *Journal of Applied Probability*, v.6, p.328-47, 1969

BOX, G. E. P.; TIAO, G. C. *Bayesian inference in statistical analysis*, Reading, MA: Addison-Wesley, 1973

BROWN, M. W.; MILLER, K. J. Biaxial fatigue data, *Report CEMR1/78*, University of Sheffield, Dept. of Mechanical Engineering, 1978.

ENGELHARDT, M.; BAIN, L. J.; WRIGHT, F. T. Inferences on the parameters of the Birnbaum-Saunders fatigue life distribution based on maximum likelihood estimation, *Technometrics*, v.23, p.251-56, 1981.

KASS, R.E.; TIERNEY, L.; KADANE, J.B. The validity of posterior expansions based on Laplace's Method. In: HODGES, J. (ed.) Essays in Honor of George A. Barnard, Amsterdam, North-Holland, 1990.

MANN, N. R.; SCHAFER, R.E.; SINGPURWALLA, N.D. *Methods for statistical analysis of reliability and life data*, New York, John Wiley, 1974.

PRESS, S.J. *Bayesian statistics: principles, models and applications*, New York, John Wiley, 1989.

RIECK, J.R. Statistical analysis for the Birnbaum–Saunders fatigue life distribution. Unpublished Ph.D. thesis, Clemson University, Dept. of Mathematical Sciences, 1989.

———. NEDELMAN, J. R. A Log–linear model for the Birnbaum–Saunders distribution, *Technometrics*, v. 33, n.1, p.51–60, 1991.

TIERNEY, L.; KADANE, J.B. Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association*, v.81, n.393, p.82–86, mar.1986.

———. KASS, R.E.; KADANE, J.B. Full exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, v.84, n.407, p.710–16, sept.1989.

ABSTRACT

In this paper, we consider a Bayesian analysis of a log–linear model for the Birnbaum–Saunders distribution proposed by Rieck and Nedelman (1991). We find simple expressions for the marginal posterior densities and predictive densities of interest considering one explanatory variable and using the Laplace's method for integrals, when we cannot find explicit solutions (for several noninformative priors). In a numerical example, we consider an application to accelerated life tests with a data set introduced by Brown and Miller (1978) in order to compare the Bayesian methodology with the usual standard classical approach based on the asymptotical normal distribution for the maximum likelihood estimators of the parameters of the model.

RESUMO

Neste artigo consideramos uma análise Bayesiana de um modelo log–linear, proposto por Rieck e Nedelman (1991), para a distribuição de Birnbaum–Saunders. Obtemos expressões simples para as densidades marginais *a posteriori* e para as densidades preditivas de interesse, considerando uma variável exploratória e usando o método de Laplace para integrais, quando não foi possível obter soluções explícitas (para várias *a priori* não–informativas). Em um exemplo numérico, consideramos uma aplicação aos testes acelerados de tempo de vida com um conjunto de dados introduzido por Brown e Muller (1978), a fim de comparar a metodologia Bayesiana com a abordagem clássica padrão baseada na distribuição assintótica normal dos estimadores de máxima verossimilhança dos parâmetros do modelo.

ALGUMAS CONSIDERAÇÕES SOBRE A METODOLOGIA BAYESIANA PARA PESQUISAS ELEITORAIS COM APLICAÇÃO ÀS ELEIÇÕES DE 1990 NO ESTADO DO ESPÍRITO SANTO

Gutemberg Hespanha Brasil*

Antonio Fernando Pêgo e Silva*

1 INTRODUÇÃO

O modelo Bayesiano para previsão de resultados eleitorais proposto por Bernardo (1984) utiliza um desenho amostral diferente dos comumente adotados. Através da medida de divergência de Kullback–Leibler, Kullback & Leibler (1951) e Kullback (1968), selecionam–se os locais de votação (municípios, zonas eleitorais, etc.) mais representativos do comportamento político da população sendo considerada (estado, município, etc.). Algumas aplicações anteriores da metodologia são: Brasil, Migon & Souza (1986); Migon, Souza, Brasil & Sant'Anna (1986); Mendonça & Migon (1987) e Souza & Brasil (1989).

Neste trabalho apresentamos um procedimento heurístico para a seleção do número de locais a serem pesquisados (Seção 2). A metodologia também pode ser adotada em outras situações, via uso de outros vetores de dados representativos de uma dada região e suas áreas componentes.

Uma aplicação detalhada da metodologia proposta, juntamente com o modelo Bayesiano para previsão dos resultados finais de uma eleição, foi realizada para as eleições de 1990 ao

* Universidade Federal do Espírito Santo – UFES

Governo/Senado do Estado do Espírito Santo – Brasil. Também verificou–se a estabilidade dos locais selecionados, no período 1982–90, utilizando–se a medida de divergência para cada ano.

A organização do artigo é a seguinte. Na Seção 2 descrevemos a metodologia para seleção dos locais mais representativos, bem como o procedimento proposto para a escolha do número de locais. Na Seção 3 o modelo proposto por Bernardo (1984) é apresentado compactamente. Na Seção 4 aplicamos o modelo às eleições de 1990 (governador e senador) no Estado do Espírito Santo. Na Seção 5 fazemos algumas considerações gerais e, finalmente, apresentamos em três apêndices uma análise Bayesiana do modelo multinomial–Dirichlet, a medida de divergência de Kullback–Leibler e a amostra utilizada na aplicação.

2 SELEÇÃO DE LOCAIS

2.1 Seleção dos Locais mais Representativos

Diferentemente dos métodos tradicionais de planejamento amostral, que supõem que o voto esteja relacionado fortemente a características da população tais como distribuições etárias, de renda e sexo, a metodologia proposta por Bernardo (1984) procura selecionar aqueles locais que sejam mais representativos do comportamento político da população. Uma vez identificados esses locais, uma amostra aleatória de mesma dimensão de eleitores é selecionada em cada um deles.

Assim, deve–se identificar os locais (zonas eleitorais, municípios etc.) mais representativos da população votante, no sentido de que o comportamento do voto assumido pelos eleitores destes locais seja o mais similar possível ao comportamento da população como um todo (o estado, o município, etc.). Isso porque alguns locais podem ter gerado um resultado similar ao de toda a região. Necessita–se, portanto, de uma medida de distância (ou de discriminação) entre as correspondentes distribuições de probabilidade: a de toda a região e a de cada um dos locais.

Na aplicação da Seção 4 os locais serão identificados com os municípios e a região será o estado em questão (Espírito Santo).

A seleção dos locais mais similares é feita com base nos resultados da eleição imediatamente anterior. Aplica–se, então, um critério de minimização das "distâncias" entre esses locais e o universo eleitoral; a partir daí podemos construir uma "hierarquia de similitudes", para selecionar os locais "mais representativos" do universo considerado.

O critério adotado é a medida de divergência direta de Kullback–Leibler (1951), Kullback

(1968). Observe–se que esta medida de divergência é coerente com a abordagem Bayesiana, Aitchinson (1975), Rodrigues (1991). Evidentemente essa abordagem pode ser utilizada em outras situações que não a de pesquisas e/ou previsões de resultados eleitorais. No Apêndice 6.2 apresentamos uma breve descrição da medida de divergência de Kullback–Leibler.

Assume–se implicitamente que existe uma certa estabilidade temporal no comportamento eleitoral, pelo menos de uma eleição para a eleição seguinte. Convém ressaltar que o critério adotado elimina as áreas atípicas a serem pesquisadas.

No Apêndice 6.1 apresentamos os argumentos para a determinação do estimador Bayesiano, $\hat{\theta}_{ij}$, da probabilidade de um eleitor do i–ésimo local (i.e., com características similares às daqueles que vivem na região abrangida pelo local i) votar no j–ésimo candidato (ou partido). Este estimador é:

$$\hat{\theta}_{ij} = \frac{n_{ij} + 1/2}{n_1 + (m+1)/2}, \qquad j = 1, 2, \cdots, m. \tag{1}$$

Quando aplicado ao presente caso, a medida de divergência de Kullback–Leibler fica:

$$D_i = \sum_{j=1}^{m} \theta_j \cdot Ln(\frac{\theta_j}{\hat{\theta}_{ij}}) \qquad i = 1, 2, \cdots, k. \tag{2}$$

onde:

n_{ij} – número de votos para o candidato (ou partido) j no local i (Zona Eleitoral – ZE, município, etc.) na eleição anterior;

m – número de candidatos (ou partidos) existentes;

$n_i = \sum_{j=1}^{m} n_{ij}$ – número de votos no local i (ZE ou município);

θ_j – proporção de votos obtidos pelo candidato j (na área total) na eleição imediatamente anterior; $j = 1, 2, \cdots, m$.

Quanto menor for a medida D_i, mais representativo será o local i (ZE ou município), ou seja, é mais similar ao estado como um todo, no sentido descrito anteriormente. Lembramos que foi adotada a suposição de que esses "locais representativos" de toda a região permanecem os mesmos de uma eleição para a seguinte. Na Seção 4 verificamos essa suposição nos pleitos de 1982, 1986, 1989 e 1990 no Estado do Espírito Santo.

2.2 Um Procedimento para a Seleção do Número de Locais

Através do uso de (2) podemos ordenar os locais em ordem de similitude, i.e., do mais similar ao menos similar a toda região sendo estudada. Restam ainda algumas decisões: quantos locais "mais similares" selecionar e quantas entrevistas realizar em cada local.

Quanto ao número de locais similares a serem selecionados para pesquisa (intenção de voto, "boca de urna", projeção final dos resultados após a apuração dos primeiros votos), adotamos um procedimento heurístico bastante simples.

A medida de divergência, dada por (2), resulta em um valor igual a zero apenas quando as distribuições P e Q são iguais (ver Apêndice 6.2). Assim, quanto mais afastadas de zero forem as medidas D_i, equação (2), menos similares serão os locais. Suponha que o primeiro local mais representativo, com medida D_1, seja "razoavelmente" similar ao todo. Nesse caso, se calcularmos o Erro Quadrático Médio - EQM - dessa distribuição com relação à distribuição global essa quantidade será pequena. Obtemos então, o EQM_1.

Suponha que tomemos médias sucessivas, $\bar{\theta}_{ij}$, entre as distribuições dos locais subseqüentemente similares e realizemos a mesma comparação via EQM, obtendo o EQM_2, EQM_3, etc. Teremos, então, para $i = 1, 2, \cdots, K$: e $j = 1, 2, \cdots, m$:

$$\bar{\theta}_{ij} = \frac{1}{i} \cdot \sum_{l=1}^{i} \hat{\theta}_{lj} \tag{3}$$

$$EQM_i = \frac{1}{m} \cdot \sum_{j=1}^{m} (\theta_j - \bar{\theta}_{ij})^2. \tag{4}$$

Intuitivamente é de se esperar que, enquanto estivermos combinando locais realmente similares, o gráfico EQM versus o número de municípios apresente um comportamento estável ou declinante. De outro lado, quando introduzirmos na média municípios pouco similares, é de se esperar que o EQM apresente comportamento crescente. Desse modo, analisando-se o gráfico mencionado, poderíamos realizar uma escolha do número de locais de modo a que não sejam tão poucos a ponto de termos graus de liberdade insuficientes e nem tantos que tornem a pesquisa muito dispendiosa em termos de custos. Ou seja, deveríamos selecionar o número de locais no entorno do menor EQM_i. Na aplicação da Seção 4 utilizamos este procedimento para selecionar o número de locais.

Cabem alguns comentários adicionais. Na composição da amostra é importante a decisão entre o número de locais similares escolhidos e o número de eleitores a entrevistar em cada local. Esta decisão resulta em um compromisso entre custo e precisão (ver a Seção 3).

Além do mais, como argumentam Mendonça & Migon (1987), considerando-se que os eleitores em cada local selecionado devem apresentar um comportamento heterogêneo para representar o universo, é razoável ter-se "uma maior fração amostral nesse estágio (eleitores), comparativamente à fração amostral do primeiro estágio (locais)".

Por exemplo, Bernardo (1984) menciona que selecionou uma amostra aleatória de 50 eleitores em cada uma das 20 áreas mais representativas da província de Valência na Espanha (aproxi-

madamente 1 800 000 eleitores em 1982), sendo que estes números foram obtidos através de similações com os resultados da eleição imediatamente anterior.

Os resultados do presente trabalho sugerem que seja utilizado o procedimento proposto para a seleção do número de locais mas pouco pode ser dito quanto à quantidade de eleitores a serem entrevistados (no caso de pesquisa de intenção de voto), exceto que deve ser maior do que o número de locais.

3 ANÁLISE BAYESIANA DO MODELO

Descrevemos sucintamente nesta seção o modelo Bayesiano para análise dos resultados, como proposto em Bernardo (1984); outras referências são Mendonça & Migon (1987) e Souza & Brasil (1989), sendo que estes últimos apresentam uma descrição detalhada do algoritimo. Algumas definições necessárias:

θ_{ij}: probabilidade de um eleitor do local de votação i votar no canditado j; $\theta_{ij} > 0$, e $\sum_{j=1}^{m} \theta_{ij} = 1$, $i = 1, 2, \cdots, k$ e $j = 1, 2, \cdots, m$;

n_i: tamanho da amostra no local i; $n_i = [n_{i1} + n_{i2} + \cdots + n_{im}]$;

N: matriz de dimensão $k \times m$, cujo elemento (i, j) é n_{ij};

X_{ij} variável aleatória contínua representando os *log–odds* (razão de chances) associados a θ_{ij}, ou seja:

$$X_{ij} = Log[\frac{\theta_{ij}}{1 - \theta_{ij}}], X_{ij} \in \mathbf{R} \tag{5}$$

Na análise Bayesiana assumimos que o vetor aleatório $\mathbf{n_i} = [n_{i1}, n_{i2}, \cdots, n_{im}]$ tem distribuição multinomial com parâmetros θ_{ij}: $\underline{\theta}_1 = [\theta_{i1}, \theta_{i2}, \cdots, \theta_{im}]$. Assim, o modelo observacional é:

$$[n_{i1}, n_{i2}, \cdots, n_{im}] \approx \text{Multinomial } (\underline{\theta}_i). \tag{6}$$

Do ponto de vista Bayesiano necessitamos especificar uma distribuição *a priori* para os $\theta'_{ij}s$. Para simplificação do trabalho operacional podemos utilizar uma *priori* não–informativa, como descrito no Apêndice 6.1. Daí, *a posteriori* fica:

$$[\theta_{i1}, \theta_{i2}, \cdots, \theta_{im}] \approx \text{Dirichlet } (\mathbf{N} + 1/2). \tag{7}$$

i.e., uma distribuição Dirichlet com parâmetros

$$(n_{ij} + 1/2), \quad j = 1, 2, \cdots, m.$$

O estimador de Bayes para a proporção de eleitores, $\hat{\theta}_{ij}$, é dado pela média da *posteriori*:

$$\hat{\theta}_{ij} = E[\theta_{ij}|\mathbf{N}] = \frac{n_{ij} + 1/2}{n_i + m/2} \tag{8}$$

Podemos fazer uma mudança de escala, e obter os valores X_{ij} na escala da razão de chances, correspondente ao vetor $\mathbf{n}_i = [n_{i1}, n_{i2}, \cdots, n_{im}]$

$$X_{ij} = Log\left[\frac{n_{ij} - 1/2}{n_i - n_{ij} + (m-1)/2}\right]. \tag{9}$$

Note que os $X'_{ij}s$ realmente descrevem, para a amostra observada, a força do candidato j no local i. A vantagem de usar (9) consiste no fato dos $X'_{ij}s$ serem números reais, o que flexibiliza na escolha do modelo para os percentuais de cada candidato em toda a região considerada.

Duas suposições razoáveis precisam ser feitas para completar a análise coerentemente ao plano amostral da Seção 2:

i) o vetor de *log-odds a posteriori*, $\underline{X}_i$ tem distribuição normal multivariada com vetor de médias $\underline{\mu}_i \in \mathbf{R}^m$ e matriz de precisão constante $\mathbf{H}_0 \in \mathbf{R}^{m\times m}$

$$[\mathbf{X}_i|\underline{\mu}_i, \mathbf{H}_0] \approx NM[\underline{\mu}_i, \mathbf{H}_0] \tag{10}$$

A suposição será verdadeira se o tamanho da amostra n_i, em cada local, for o mesmo, e se, para todo $i, \Theta_{ij} \approx \Theta_j$, $j = 1, 2, \cdots, m$. Isso é razoável uma vez que os locais selecionados são os mais similares à região como um todo.

ii) o vetor de médias $\underline{\mu}_i = [\mu_{i1}, \mu_{i2}, \cdots, \mu_{im}]$ é aproximadamente normal em cada local de votação com média comum $\underline{\delta} \in \mathbf{R}^m$ e matriz de precisão comum $\mathbf{H}_1 \in \mathbf{R}^{m\times m}$:

$$[\underline{\mu}_i|\underline{\delta}, \mathbf{H}_1] \approx NM[\underline{\delta}, \mathbf{H}_1] \tag{11}$$

O vetor de médias $\underline{\delta}$ (na escala da razão de chances) contém as quantidades de interesse, i.e., o vetor de preferências de votos para cada candidato em toda a região.

As suposições (i) e (ii) definem um modelo hierárquico como em Lindley & Smith (1972). Combinando (10) e (11), temos a distribuição conjunta dos vetores $\mathbf{X}_i$ e $\underline{\mu}_i$, para cada local de votação

$$P[\mathbf{X}_i, \underline{\mu}_i|\underline{\delta}, \mathbf{H}_0, \mathbf{H}_1] \propto P[\mathbf{X}_i|\underline{\mu}_i, \mathbf{H}_0] \cdot P[\underline{\mu}_i|\underline{\delta}, \mathbf{H}_1] \tag{12}$$

Considerando-se que t é o número de locais escolhidos no plano amostral, pode-se determinar a distribuição de $\mathbf{X}_i$ (centrada no vetor de interesse $\underline{\delta}$), integrando-se fora o parâmetro incômodo (nuisance) $\underline{\mu}_i$ na distribuição conjunta $P[\mathbf{X}_i, \underline{\mu}_i|\underline{\delta}, \mathbf{H}_0, \mathbf{H}_1]$. Obtemos:

$$[\mathbf{X}_i|\underline{\delta}, \mathbf{H}_0, \mathbf{H}_1] \approx NM[\underline{\delta}, \mathbf{H}_0(\mathbf{H}_0 + \mathbf{H}_1)^{-1}\mathbf{H}_1], \tag{13}$$

onde o vetor $\mathbf{X}_i$ tem dimensão m, com média $\underline{\delta} = [\delta_1, \delta_2, \cdots, \delta_m]$, descrevendo o comportamento eleitoral global da região e tem matriz de precisão desconhecida $\mathbf{H} = \mathbf{H}_0(\mathbf{H}_0 + \mathbf{H}_1)^{-1}\mathbf{H}_1] \in \mathbf{R}^{m \times m}$. A prova utiliza resultados encontrados em Lindley & Smith (1972) e Smith (1973).

Desse modo, o modelo tem dois parâmetros desconhecidos: o vetor de médias $\underline{\delta}$ e a matriz de precisão $\mathbf{H}$. Uma estimativa de $\underline{\delta}$ pode ser obtida através de outra análise Bayesiana: uma distribuição para o vetor $\underline{\delta}$ e o estimador de Bayes. Adotando-se uma distribuição *a priori* não-informativa para $\underline{\delta}$ e $\mathbf{H}$,

$$P[\underline{\delta}, \mathbf{H}] \propto |\mathbf{H}|^{-(m+1)/2} \tag{14}$$

determina-se uma distribuição *a posteriori* para $\underline{\delta}$ como:

$$P[\underline{\delta}|\text{DADOS}] \propto P[\mathbf{X}_1|\underline{\delta}, \mathbf{H}] \cdot P[\underline{\delta}, \mathbf{H}]. \tag{15}$$

A distribuição *a posteriori* de referência, $P[\underline{\delta}|\text{DADOS}]$, é uma distribuição T multivariada, com $(t-m)$ graus de liberdade; Bernardo (1979):

$$P[\underline{\delta}|\text{DADOS}] \propto |\mathbf{S} + (\bar{\mathbf{X}} - \underline{\delta})(\bar{\mathbf{X}} - \underline{\delta}^T|^{-t/2} \tag{16}$$

onde $\bar{\mathbf{X}} = [\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_m]$ é a média amostral dos *log-odds*, e,

$$\bar{X}_j = \sum_{i=1}^{t} X_{ij}/t, \qquad j = 1, 2, \cdots, m \tag{17}$$

$$\mathbf{S} = \{S_{i,j}\}; S_{lk} = \sum_{i=1}^{t}(X_{il} - \bar{X}_l)(X_{ik} - \bar{X}_k)/t; \qquad 1, k = 1, 2, \cdots, t. \tag{18}$$

Logo, a média de $[\underline{\delta}|\text{DADOS}]$ é $E[\underline{\delta}|\text{DADOS}] = \bar{\mathbf{X}}$ com matriz de dispersão igual a $\mathbf{S}/(t-m)$. Os intervalos de probabilidade *a posteriori* dos componentes do vetor $\underline{\delta}$ podem ser obtidos a partir das distribuições marginais de (16). Temos:

$$P[\delta_j \in \bar{X}_j \pm h_\alpha \cdot \sqrt{\{s_{jj}/(t-m)\}} = (1-\alpha) \tag{19}$$

onde h_α é o quantil $(1-\alpha/2)$ da distribuição T normalizada com $(t-m)$ graus de liberdade.

Como os $\delta_j's$ estão na escala da razão de chances, utiliza-se a transformação inversa para obter o estimador desejado:

$$\hat{\psi}_j = exp(\bar{X}_j)/[1 + exp(\bar{X}_j)], \qquad j = 1, 2, \cdots, m. \tag{20}$$

onde $\hat{\psi}_j$ representa a proporção de votos a ser obtida pelo candidato j em toda a região considerada. Os intervalos de credibilidade de $(1-\alpha)\%$ podem ser determinados diretamente das distribuições marginais T de Student com $(t-m)$ graus de liberdade, similarmente a (20).

Observe–se que os intervalos de credibilidade dependerão do número de graus de liberdade e da dispersão S_{jj}, i.e., da dispersão dos X_{ij} entre os locais escolhidos; quanto menor a dispersão, mais estreitos serão os intervalos.

4 APLICAÇÃO AO PLEITO DE 1990 – GOVERNO E SENADO DO ESPÍRITO SANTO

4.1 A Seleção dos Locais mais Representativos

Nas eleições de 1990 – Governo, Senado e Câmara Federal – o Estado do Espírito Santo estava dividido em 67 municípios e 48 Zonas Eleitorais – ZE's –, sendo que uma ZE podia abranger mais de um município. Optou–se por trabalhar com a unidade básica município. Cada município possui um número variado de seções (ou urnas) contendo, em média, aproximadamente 270 eleitores. O total de eleitores em todo o estado em 15/11/90 era de 1 423 211. O número de eleitores varia bastante por município; por exemplo, o Município de Vitória (a capital do estado) possuía 160 803 eleitores em 1990, enquanto que o município com menor eleitorado era Divino de São Lourenço com 2 836 eleitores.

Utilizando o esquema descrito na Seção 2, ordenamos os locais segundo a medida de divergência de Kullback–Leibler para os anos eleitorais de 1982, 1986 e 1990 onde houve eleições ao governo estadual e para 1989 onde tivemos a eleição presidencial. Os resultados estão na Tabela 1. No vetor de resultados finais utilizamos o percentual obtido por cada candidato bem como votos brancos e nulos. Os resultados ficaram praticamente inalterados quando introduzimos as abstenções.

A Tabela 1 foi construída do seguinte modo. Os 25 municípios mais similares ao todo, segundo os resultados de cada ano eleitoral, foram ordenados em ordem decrescente de similitude. Assim, o primeiro município possui maior similaridade de comportamento eleitoral com relação ao estado como um todo.

O ano de 1986, por ser o mais central, foi tomado como referência. Observe–se que na coluna referente a 86 temos a ordenação 1 a 25; a seguir, para os outros anos verificou–se quais posições estes municípios ocuparam. Se o município selecionado pela ordenação não figura entre um dos 25 de 1986, fica identificado com um X. Desse modo, por exemplo, o município de número 6 foi o 6º mais similar em 1986, ficou em 23º em 1982, em 11º em 1989 e em 14º em 1990. De outro lado, o município número 2 praticamente não alterou sua posição ao longo da década. Em nenhum dos anos foi selecionada a capital do estado entre os 25 primeiros municípios mais similares.

TABELA 1

Ordenação dos Municípios Segundo a Medida de Divergência

Mun.	1982(GOVERNADOR)	1986(GOVERNADOR)	1989(PRES.)	1990(GOVERNADOR)
1	01	01	02	09
2	02	02	13	10
3	22	03	07	02
4	16	04	*X*	*X*
5	08	05	*X*	08
6	11	06	01	01
7	07	07	10	04
8	*X*	08	*X*	*X*
9	*X*	09	21	*X*
10	05	10	04	*X*
11	04	11	06	*X*
12	14	12	15	*X*
13	24	13	09	*X*
14	10	14	*X*	06
15	*X*	15	11	13
16	12	16	08	*X*
17	*X*	17	16	05
18	09	18	*X*	*X*
19	19	19	*X*	*X*
20	*X*	20	22	25
21	*X*	21	05	*X*
22	03	22	*X*	*X*
23	06	23	*X*	*X*
24	*X*	24	*X*	16
25	*X*	25	*X*	11

Podemos observar o seguinte:

i) com relação a 82 e 86 – nove municípios são os mesmos entre os 15 primeiros e, 17 entre os 25;

ii) com relação a 86 e 89 – dez municípios são os mesmos entre os 15 primeiros e 15 entre os 25;

iii) com relação a 86 e 90 – oito municípios são os mesmos entre os 15 primeiros e 12 entre os 25;

iv) com relação a 82, 86 e 90 – cinco municípios são os mesmos entre os 15 primeiros e 10 entre os 25;

v) a renovação de municípios entre os 25 primeiros de 86 foi de 52% com relação a 1990; e

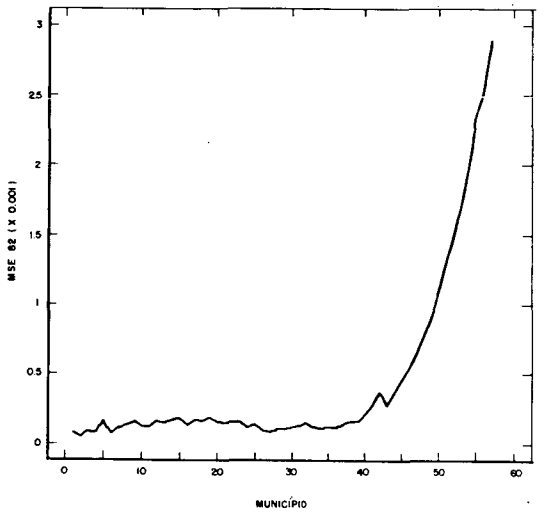
vi) apesar da renovação existente entre cada ano eleitoral pode–se dizer que existe uma certa estabilidade caracterizada pela permanência de vários municípios ao longo do período analisado. Observe–se, contudo, que no ano de 1986 houve um recadastramento eleitoral em todo o estado e que o número de municípios passou de 57 em 1982 para 67 em 1990.

As Figuras 1 a, b, c e d mostram o gráfico do $EQM_i \times i$, sendo i o número de municípios, calculado a partir da equação (4), para os anos de 1982, 1986, 1989 e 1990, respectivamente. Os gráficos apresentam aproximadamente o comportamento esperado, exceto para o ano eleitoral de 1982.

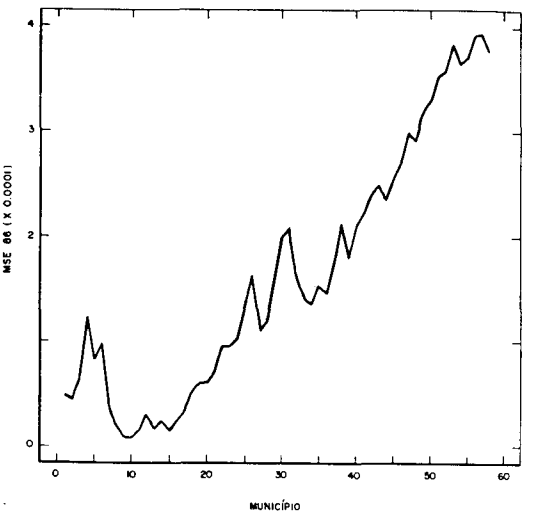
Os gráficos estão em escalas distintas, sendo que a magnitude do EQM é maior para o ano de 1982. Note-se também que, para este ano, obteve-se uma magnitude da medida de divergência superior à dos outros anos para o município mais similar. Para uma melhor visualização, todos os quatro gráficos foram reunidos na Figura 2 em uma mesma escala.

Figura 1 – Erro Quadrático Médio × Número de Municípios

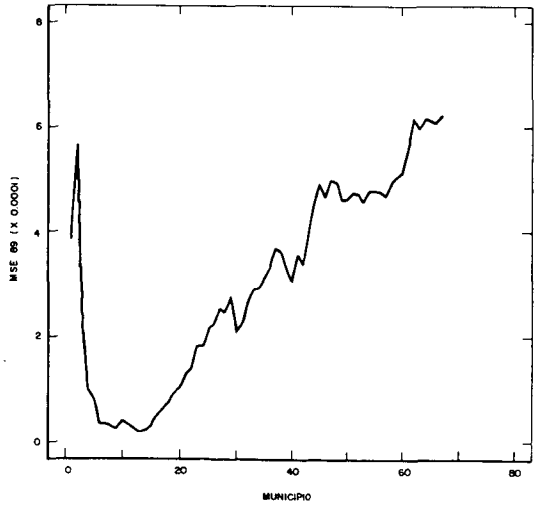
(a) Ano Eleitoral de 1982



(b) Ano Eleitoral de 1986



(c) Ano Eleitoral de 1989



(d) Ano Eleitoral de 1990

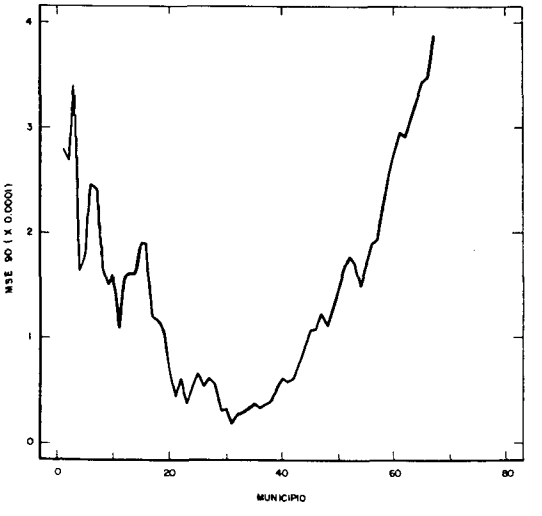
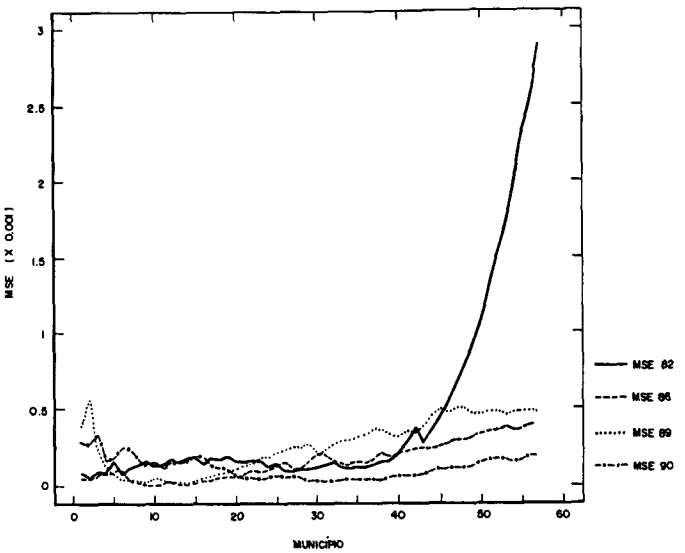


Figura 2 – Erro Quadrático Médio × nº de municípios: 82, 86, 89 e 90



Os gráficos indicam que entre 10 e 25 municípios seriam suficientes para se realizar as previsões dos resultados finais. Particularmente, para o ano–base de 1986, entre 10 e 15 municípios seriam adequados. Veja a aplicação na seção seguinte.

4.2 Previsão dos Resultados Finais

Utilizamos os resultados da apuração pararela realizada por uma emissora de TV (TV Gazeta – ES). A projeção do resultado final foi realizada após apuradas as três primeiras urnas (ou, equivalentemente, aproximadamente os primeiros 900 votos) dos primeiros municípios selecionados segundo o critério de similitude para o ano de 1986 (Base 1986), cujos resultados encontravam–se apurados. Estes resultados eram enviados, via telefone, para uma central de apuração. A projeção foi feita cinco horas após iniciada a apuração.

Um problema ocorreu devido ao fato da apuração não ter–se iniciado simultaneamente em todos os locais. Assim, para os municípios de ordens de similitude 11 e 12, não tínhamos dados disponíveis. Utilizou–se os municípios subseqüentes no ordenamento. O mesmo aconteceu com os municípios de ordens 14,15 e 16. Todos os percentuais foram calculados sobre o total de votantes, i.e., o total de eleitores menos as abstenções (não comparecimento).

As Tabelas 2 a 8 apresentam a projeção dos resultados finais, bem como o intervalo de credibilidade de 90%, tomando–se como base o ano de 1986 (governador), iniciando–se com os 10 municípios mais representativos, acrescendo de um até incluir o 16º mais similar.

TABELA 2

Previsão dos Percentuais Finais

10 Municípios – Base 1986

CANDIDATO	LIMITE INFERIOR	PREVISÃO	LIMITE SUPERIOR	RESULTADO OFICIAL (GOVERNADOR)
ALBUÍNO	20,13	28,10	37,79	28,91
INÁCIO	19,91	25,70	32,52	23,60
ROGÉRIO	6,17	10,11	16,14	10,36
CALMON	2,00	3,03	4,57	3,17
BRANCOS	12,68	17,65	24,06	17,03
NULOS	11,14	15,41	20,95	16,93
EQM		1,31		

Nota – Intervalo de Credibilidade de 90%.

TABELA 3

Previsão dos Percentuais Finais

11 Municípios – Base 1986

CANDIDATO	LIMITE INFERIOR	PREVISÃO	LIMITE SUPERIOR	RESULTADO OFICIAL (GOVERNADOR)
ALBUÍNO	20,84	27,73	35,90	28,91
INÁCIO	20,03	25,12	31,04	23,60
ROGÉRIO	6,78	10,39	15,60	10,36
CALMON	2,16	3,09	4,40	3,17
BRANCOS	13,40	17,75	23,15	17,03
NULOS	11,89	15,92	21,01	16,93
EQM		0,88		

Nota – Intervalo de Credibilidade de 90%.

TABELA 4

Previsão dos Percentuais Finais

12 Municípios – Base 1986

CANDIDATO	LIMITE INFERIOR	PREVISÃO	LIMITE SUPERIOR	RESULTADO OFICIAL (GOVERNADOR)
ALBUÍNO	22,23	29,15	37,24	28,91
INÁCIO	20,36	24,76	29,78	23,60
ROGÉRIO	7,05	10,16	14,43	10,36
CALMON	1,69	2,70	4,29	3,17
BRANCOS	13,76	17,50	22,01	17,03
NULOS	12,26	15,73	19,96	16,93
EQM		0,55		

Nota – Intervalo de Credibilidade de 90%.

TABELA 5

Previsão dos Percentuais Finais

13 Municípios – Base 1986

CANDIDATO	LIMITE INFERIOR	PREVISÃO	LIMITE SUPERIOR	RESULTADO OFICIAL (GOVERNADOR)
ALBUÍNO	23,51	30,06	37,58	28,91
INÁCIO	20,04	24,09	28,68	23,60
ROGÉRIO	7,48	10,27	13,94	10,36
CALMON	1,84	2,78	4,19	3,17
BRANCOS	13,95	17,24	21,13	17,03
NULOS	12,52	15,55	19,17	16,93
EQM		0,61		

Nota – Intervalo de Credibilidade de 90%.

TABELA 6

Previsão dos Percentuais Finais

14 Municípios – Base 1986

CANDIDATO	LIMITE INFERIOR	PREVISÃO	LIMITE SUPERIOR	RESULTADO OFICIAL (GOVERNADOR)
ALBUÍNO	24,03	30,28	37,41	28,91
INÁCIO	20,52	25,27	30,72	23,60
ROGÉRIO	7,49	10,13	13,57	10,36
CALMON	1,63	2,56	3,99	3,17
BRANCOS	13,64	16,85	20,64	17,03
NULOS	11,76	14,91	18,73	16,93
EQM		1,53		

Nota – Intervalo de Credibilidade de 90%.

TABELA 7

Previsão dos Percentuais Finais

15 Municípios – Base 1986

CANDIDATO	LIMITE INFERIOR	PREVISÃO	LIMITE SUPERIOR	RESULTADO OFICIAL (GOVERNADOR)
ALBUÍNO	24,32	29,69	35,72	28,91
INÁCIO	21,40	25,51	30,13	23,60
ROGÉRIO	7,94	10,25	13,14	10,36
CALMON	1,73	2,52	3,67	3,17
BRANCOS	14,31	17,20	20,53	17,03
NULOS	12,14	14,82	17,98	16,93
EQM		1,53		

Nota – Intervalo de Credibilidade de 90%.

TABELA 8

Previsão dos Percentuais Finais

16 Municípios – Base 1986

CANDIDATO	LIMITE INFERIOR	PREVISÃO	LIMITE SUPERIOR	RESULTADO OFICIAL (GOVERNADOR)
ALBUÍNO	24,72	29,62	35,06	28,91
INÁCIO	21,61	25,35	29,52	23,60
ROGÉRIO	8,15	10,27	12,87	10,36
CALMON	1,70	2,42	3,44	3,17
BRANCOS	14,51	17,14	20,15	17,03
NULOS	12,57	15,19	18,25	16,93
EQM		1,20		

Nota – Intervalo de Credibilidade de 90%.

TABELA 9

Previsão dos Percentuais Finais

13 Municípios – Base 1989

CANDIDATO	LIMITE INFERIOR	PREVISÃO	LIMITE SUPERIOR	RESULTADO OFICIAL (GOVERNADOR)
ALBUÍNO	24,38	29,48	35,17	28,91
INÁCIO	20,21	25,14	30,83	23,60
ROGÉRIO	7,27	9,87	13,27	10,36
CALMON	1,92	2,77	3,98	3,17
BRANCOS	13,32	16,60	20,51	17,03
NULOS	12,71	16,13	20,26	16,93
EQM		0,65		

Nota – Intervalo de Credibilidade de 90%.

Resultados para Senador

TABELA 10

Previsão dos Percentuais Finais

13 Municípios – Base 1986

CANDIDATO	LIMITE INFERIOR	PREVISÃO	LIMITE SUPERIOR	RESULTADO OFICIAL (SENADOR)
ÉLCIO ÁLVARES	20,10	24,70	29,98	23,55
RENATO SOARES	8,40	12,70	18,77	12,10
MAGNO PIRES	3,99	5,90	8,64	5,91
JOSÉ MORAES	2,42	4,65	8,78	5,75
JOÃO DALMÁCIO	1,47	2,39	3,85	3,17
BERREDO	1,15	1,59	2,20	2,13
J. AGUIAR	0,58	0,90	1,39	1,29
BRANCOS	25,90	31,87	38,54	29,71
NULOS	11,35	15,30	20,72	16,39
EQM		1,09		

Nota – Intervalo de Credibilidade de 90%.

TABELA 11

Previsão dos Percentuais Finais

14 Municípios – Base 1986

CANDIDATO	LIMITE INFERIOR	PREVISÃO	LIMITE SUPERIOR	RESULTADO OFICIAL (SENADOR)
ÉLCIO ALVARES	20,85	25,74	31,36	23,55
RENATO SOARES	8,93	12,66	17,66	12,10
MAGNO PIRES	4,29	5,98	8,27	5,91
JOSÉ MORAES	2,75	4,83	8,34	5,75
JOÃO DALMÁCIO	1,55	2,35	3,54	3,17
BERREDO	1,20	1,58	2,08	2,13
J. AGUIAR	0,59	0,87	1,28	1,29
BRANCOS	26,45	31,61	37,29	29,71
NULOS	10,43	14,39	19,53	16,39
EQM		1,64		

Nota – Intervalo de Credibilidade de 90%.

Na Tabela 9 temos a previsão tomando–se como base os 13 municípios mais representativos pelo resultado eleitoral de 1989 (presidente). Nas Tabelas 10 e 11 temos a projeção dos resultados para Senador. Em cada uma das Tabelas encontra–se o EQM da previsão, relativamente ao resultado final (oficial).

Com relação aos votos para governador, os menores EQMs ocorrem quando realizamos as previsões com os 11, 12 e 13 municípios mais representativos, sendo que o mínimo ocorre com 12 municípios.

Em ambos os exercícios para governador, os resultados são plenamente satisfatórios quanto à estimativa pontual; entretanto, observe–se que os intervalos de credibilidade são razoavelmente elevados. A base 1989 levou a intervalos mais estreitos. Os comentários da Seção 3 devem ser considerados.

Com relação ao voto para senador, os resultados também são muito bons, tanto com relação à previsão pontual quanto aos intervalos de credibilidade. Uma vez mais são pertinentes os comentários da Seção 3.

5 CONSIDERAÇÕES GERAIS

Descrevemos na Seção 2 uma metodologia para elaboração de um plano amostral baseado na seleção dos "locais de votação mais representativos" de uma região estudada, como proposto por Bernardo (1984). Introduzimos também um procedimento heurístico para escolha do número de locais a serem pesquisados. Na Seção 4 realizamos uma aplicação completa, incluindo a previsão dos resultados finais a partir dos primeiros votos apurados.

Quanto à estabilidade do comportamento eleitoral, relativamente aos locais escolhidos através da medida de divergência de Kullback–Leibler, pode–se verificar que existe uma certa estabilidade no período estudado (1982–90) pelo menos no sentido de que vários municípios permanecem os mesmos entre os 25 primeiros ordenados decrescentemente segundo a medida utilizada. Isso tudo a despeito de ter havido um recadastramento eleitoral no ano de 1986, que causou algum tipo de reordenação na distribuição do eleitorado no estado e da criação de 10 novos municípios no período.

O procedimento sugerido para a escolha do número de locais parece adequado e funcionou muito bem na aplicação particular. Pode–se dizer que existe uma certa "robustez" em toda a metodologia visto que, mesmo não se respeitando inteiramente a ordem de similitude dos locais a partir do 11º(veja o Apêndice 3), em virtude da indisponibilidade de dados, os resultados ainda continuaram razoáveis. Observe–se, contudo, que, na aplicação, os 10 municípios mais similares tiveram dados disponíveis.

Quanto aos aspectos preditivos, em termos globais, os resultados foram plenamente satisfatórios, o que se comprova pela pequena magnitude do erro quadrático médio nas Tabelas 2 a 11. A estimativa pontual foi muito boa em todos os casos, apresentando apenas altos intervalos de probabilidade o que deveu–se principalmente à dispersão dos resultados em cada local selecionado.

Sugere–se maiores estudos para a obtenção de um compromisso entre o número de locais a serem pesquisados e o número de eleitores a serem entrevistados em cada um deles.

6 APÊNDICES

6.1 Análise Bayesiana do Modelo Multinomial–Dirichlet

Neste apêndice apresentamos a análise Bayesiana do modelo Multinomial–Dirichlet; Box &

Tiao (1973) e Berger (1985). Distribuição multinomial – é uma generalização da distribuição binomial. Considere a situação onde tem–se k eventos mutuamente exclusivos, $A_1, A_2, \cdots, A_k$, consistindo uma partição do espaço amostral, i.e., tem–se k resultados possíveis em cada ensaio, ao invés de apenas dois. Suponha que o experimento seja repetido n vezes. Seja n_i o número de vezes que o evento A_i ocorre nas n repetições do experimento e p_i a probabilidade de ocorrência de A_i. Temos então:

$$n_1, n_2, \cdots, n_k \qquad \sum_{i=1}^{k} n_i = n$$

$$p_1, p_2, \cdots, p_k \qquad \sum_{i=1}^{k} p_i = 1, \qquad p_i > 0$$

$$p[n_1, n_2, \cdots, n_k | p_i n] = \frac{n!}{n_1! n_2! \cdots n_k!} \cdot p_1^{n1} \cdot p_2^{n2} \cdots p_k^{nk} = n! \Pi_{i=1}^{k} \frac{p_i^{ni}}{n_i!}$$

$$n_i = 0, 1, 2, \cdots, n; \qquad i = 1, 2, \cdots, k.$$

$$E[n_i] = np_i; \qquad V[n_i] = np_i(1 - p_i), \qquad i = 1, 2, \cdots, k.$$

$$Cov[n_i, n_j] = -np_i p_j, \qquad i = 1, 2, \cdots, k \quad \text{e} \quad i \neq j.$$

As variáveis $n_i's$ são dependentes e, usualmente estabelece–se o valor de uma delas como: $n_{io} = (1 - \sum_{i=1}^{k-1} n_i)$; o mesmo para $p_{io} = (1 - \sum_{i=1}^{k-1} p_i)$.

Distribuição Dirichlet – é uma generalização da distribuição Beta.

Considere a família de distribuições de Dirichlet $(\underline{P}|\underline{\alpha}) \approx D(\alpha)$ com parâmetros $\underline{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_k], \alpha_i > 0$ e $\underline{P} = [p_1, p_2, \cdots, p_k], 0 \leq p_i \leq 1$, com $\sum_{i=1}^{k} p_i = 1$. Definindo–se o parâmetro complementar como $\alpha_0 = \sum_{i=1}^{k} \alpha_i$, a função de densidade Dirichlet fica:

$$P[\underline{P}|\underline{\alpha}] = \frac{\Gamma(\alpha_0)}{\Pi_{i=1}^{k} \Gamma(\alpha_i)} \cdot \Pi_{i=1}^{k} p_i^{(\alpha_i - 1)}$$

$$E[p_i] = \mu = \alpha_i / \alpha_0; \qquad V[p_i] = \frac{(\alpha_0 - \alpha_i) \cdot \alpha_i}{\alpha_0^2 \cdot (\alpha_0 + 1)}$$

$$COV[p_i, p_j] = -\frac{\alpha_i \cdot \alpha_j}{\alpha_0^2 (\alpha_0 + 1)} \qquad i = 1, 2, \cdots, k.$$

A distribuição de Dirichlet é $(k-1)$ dimensional devido à restrição nos $p_i's$. No cálculo das Esperanças Matemáticas com relação a essa densidade, deve–se substituir p_k por $(1 - \sum_{i=1}^{k-1} p_i)$ e integrar sobre $x_1, \cdots, x_{k-1}$.

Para $k = 2$, temos $\Theta \approx D(\alpha_1, \alpha_2)$ com $\alpha_0 = \sum_{i=1}^{2} \alpha_i$.

A distribuição marginal de qualquer p_i é uma Dirichlet de dimensão unitária, ou seja, uma distribuição Beta.

O Modelo Multinomial–Dirichlet se o modelo observacional segue uma multinomial com parâmetro $\underline{P} = (p_1, \cdots, p_k)$, a distribuição Dirichlet é *a priori* natural, assim como no modelo univariado Bernoulli–Beta.

Modelo Observacional – a verossimilhança de uma amostra aleatória de tamanho n de uma distribuição Multinomial é

$$P[n, \cdots, n|p] = n! \Pi_{i=1}^{k} \frac{p_i^{n_i}}{n_i}$$

$$\propto \Pi_{i=1}^{k} p_i^{n_i} \qquad \sum_{i=1}^{k} n_i = n, \quad \sum_{i=1}^{k} p_i = 1, \; p_i > 0.$$

Distribuição *a priori* para $\underline{P}$: $\underline{P} \approx D(\underline{\alpha})$

$$P(p_1, \cdots, p_k|\underline{\alpha}) = \frac{\Gamma(\alpha_0)}{\Pi_{i=1}^{k} \Gamma(\alpha_i)} \cdot \Pi_{i=1}^{k} p_i^{(\alpha_i - 1)}$$

$$\propto \Pi_{i=1}^{k} p_i^{(\alpha_i - 1)}$$

com

$$p_i \geq 0, \alpha_i > 0, \quad \sum_{i=1}^{k} \alpha_i = \alpha_0, \quad \sum_{i=1}^{k} p_i = 1, i = 1, 2, \cdots, k.$$

Distribuição *a posteriori* para $\underline{P}$ seja a amostra observada $\underline{N} = [n_1, \cdots, n_k]$.

$$P[\underline{P}|\underline{N}] = \frac{P[\underline{N}|\underline{P}] \cdot P[\underline{P}]}{\int P[\underline{N}|\underline{P}] \cdot P[\underline{P}] d\underline{P}}$$

$$\propto P[\underline{N}|\underline{P}] \cdot P[\underline{P}]$$

$$\propto \Pi_{i=1}^{k} p_i^{n_i} \cdot \Pi_{i=1}^{k} p_i^{(\alpha_i - 1)}$$

$$\propto \Pi_{i=1}^{k} p_i^{n_i + \alpha_i - 1}$$

Comparando–se com uma Dirichlet, verifica–se que a distribuição *a posteriori* também é da mesma família, logo:

$$(\underline{P}|\underline{N}) \approx D(\underline{\alpha} + \underline{N})$$

ou seja, uma Dirichlet com parâmetros $(\alpha_i + n_i) \quad i = 1, 2, \cdots, k$.

Modelo Multinomial–Dirichlet utilizando uma *priori* Não–Informativa (de referência) – adotando–se a Regra de Jeffreys (ver Box & Tiao, 1973 – pág. 55), pode–se calcular uma *priori* não–informativa para os parâmetros da Multinomial

$$P[\underline{P}] \propto [p_1, p_2, \cdots, p_k]^{-1/2}$$

A *posteriori* para $(\underline{P})$ é calculada do mesmo modo: verossimilhança X *priori*:

$$P[\underline{P}|\underline{N}] \propto \Pi_{i=1}^{k} p_i^{n_i} \cdot \Pi_{i=1}^{k} p_i^{-1/2}$$

$$P[\underline{P}|\underline{N}] \propto \Pi_{i=1}^{k} p_i^{n_i - 1/2}$$

Logo, a distribuição *a posteriori* de referência é também uma Dirichlet: $(\underline{P}|\underline{N}) \approx D(\underline{\alpha}+\underline{N})$, ou seja, uma Dirichlet com parâmetros $\alpha_i = (n_i + 1/2), \quad i = 1, 2, \cdots, k$.

Daí, o Valor Esperado de cada p_i dado os dados $\underline{N}$, é:

$$E[P_i|\underline{N}] = \hat{p}_i = \frac{n_i + 1/2}{\sum_{i=1}^{k}(n_i + 1/2)} = \frac{n_i + 1/2}{n + k/2}$$

visto que $n = (n_1 + n_2 + \cdots + n_k)$.

6.2 A Medida de Divergência de Kullback–Leibler

A medida de divergência de Kullback–Leibler está diretamente relacionada ao conceito de "informação" e este, ao de entropia. Neste apêndice fazemos uma brevé descrição destes conceitos relacionados. Algumas referências são: Berger (1985), Kendall (1973), Kullback & Leibler (1951) e Kullback (1968).

Clausius em 1854 introduziu o conceito de entropia, hoje conhecido como segunda lei da termodinâmica. Em 1877 Boltzman fez a ligação entre o conceito termodinâmico de entropia e o conceito estatístico de desordem. Hartley em 1928 introduziu os fundamentos matemáticos da "teoria da informação" em sistema de comunicações. O livro "The Mathematical Theory of Comunications" de Shannon & Weaver, publicado em 1949, define informação como uma "propriedade estatística de um conjunto de mensagens possíveis, não de uma mensagem individual". A ligação completa entre a teoria estatística e a teoria da informação foi apresentada por Kullback & Leibler (1951) e Kullback (1968). Observe–se que a medida de divergência de $K - L$ é coerente com os argumentos Bayesianos; Aitchinson (1975).

A entropia de uma distribuição P, $H(P)$, é uma medida da incerteza associada a essa distribuição de probabilidade. Assim, se Y tem distribuição contínua $p(Y|\theta)$, a entropia, é dada por:

$$H[p(Y|\theta)] = -E[Ln\{p(Y|\theta)\}] = -\int_Y Ln[p(Y|\theta)] \cdot p(Y|\theta) dy.$$

No caso discreto, onde temos resultados possíveis $Y_1, Y_2, \cdots, Y_n$ com probabilidades $p_1, p_2, \cdots, p_n$ com $p_i > 0$ e $\Sigma p_i = 1$, temos:

$$H[p(Y|\theta)] = -E[Ln\{p_i\}] = -\sum_{i=1}^{n} p_i \cdot Ln(p_i);$$

se $p_i = 0$, então $p_i \cdot Ln(p_i) = 0$.

A informação é exatamente negativo a entropia. Assim, a informação sobre $p(Y|\theta)$, $I[\cdot]$ é:

$$I_p[p(Y|\theta)] = E[Ln\{p(Y|\theta)\}] = \int_Y Ln[p(Y|\theta)] \cdot p(Y|\theta)dy.$$

Enquanto a entropia é uma medida de desorganização, a informação é uma medida de organização de um grupo de mensagens. Vale a esse respeito o comentário de Norbert Wiener: "quanto mais provável é uma mensagem, menos informação ela propicia".

Considere agora a classe de modelos paramétricos: $y \in Y$, $\theta \in \Theta$ e $p(Y|\theta)$ uma função densidade sobre y. Suponha que x seja um conjunto de dados disponíveis descrevendo uma classe de medidas similarmente parametrizadas. Quer–se verificar o ajuste de $q(Y|x)$ sobre $p(Y|\theta)$. Necessita–se então de uma medida global de divergência de $q(Y|x)$ da verdadeira densidade $p(Y|\theta)$. A medida de divergência direta proposta por Kullback–Leibler para problemas de estimação é dada por:

$$D[p(Y|\theta), q(Y|x)] = \int_Y Ln[\frac{p(Y|\theta)}{q(Y|x)}] \cdot p(Y|\theta)dy,$$

onde $D[\cdot] \geq 0$ e, $D[\cdot] = 0$ apenas quando $q(Y|x) = p(Y|\theta)$.

A medida reflete a *perda esperada* entre a utilização de uma distribuição aproximada Q e a distribuição verdadeira P:

$$D[p(Y|\theta), q(Y|x)] = E_p[Ln\{[p(Y|\theta)\}] - E_p[Ln\{[q(Y|x)\}].$$

No caso discreto, se a distribuição verdadeira é $P = [p_1, p_2, \cdots, p_m]$, $p_i > 0$, $\Sigma p_i = 1$; e $Q = [q_1, q_2, \cdots, q_m]$, $q_i > 0$, $\Sigma q_i = 1$ é a distribuição aproximada, a medida de divergência fica:

$$D[Q, P] = \sum_{i=1}^{m} p_i \cdot Ln[\frac{p_i}{q_i}] = E[Ln(P)] - E[Ln(Q)] \geq 0.$$

6.3 Amostra Utilizada na Aplicação (Governo do Estado)

TABELA

Apuração da Votação dos 16 primeiros Municípios com dados disponíveis, segundo a ordem de similitude do pleito de 1986 (exatamente o resultado da apuração das três primeiras urnas)

MUN.	ALBUÍNO	INÁCIO	ROGÉRIO	CALMON	BRANCOS	NULOS	TOTAL
1	121	154	49	21	123	143	611
2	357	205	62	20	230	141	1015
3	117	262	127	27	195	98	826
4	230	280	94	50	146	190	990
5	231	171	92	07	136	61	698
6	270	144	201	32	75	171	893
7	151	304	38	30	129	144	796
8	192	158	32	23	147	93	645
9	341	207	87	21	86	71	813
10	270	151	99	21	148	138	827
13	167	137	93	25	129	150	701
14	271	120	45	03	85	78	602
15	363	147	100	34	124	117	885
19	217	287	55	05	80	54	698
32	283	369	154	26	288	174	1294
34	251	203	92	11	144	190	948

Nota – A coluna Mun. refere–se à ordem de similitude no ano de 1986.

Agradecimentos

Os autores gostariam de agradecer os comentários e sugestões de dois pareceristas desta revista.

BIBLIOGRAFIA

AITCHINSON, J. Goodness of prediction fit, *Biometrika*, v.62, n.3, p.547–54, 1975.

BERGER, J. O. *Statistical decision theory and bayesian analysis*, 2. ed., Springer–Verlag, 1985.

BERNARDO, J. M. Reference posterior distribution for bayesian inference (with discussion), *Journal of the Royal Statistical Society*, série B, v.41, n.2, p.113–47, 1979.

———. Monitoring the 1982 Spanish socialist victory: a bayesian analysis, *Journal of the American Statistical Association*, v.79, n.387, p.510–15, sept. 1984.

BOX, G. E. P.; TIAO, G. C. *Bayesian inference in statistical analysis*, Reading, Mass., Addison–Wesley, 1973.

BRASIL, G. H.; MIGON, H. S.; SOUZA, R. C. *Relatório sobre pesquisas eleitorais no Estado do Rio de Janeiro utilizando a Metodologia Bayesiana*, 1986.(Texto não publicado).

KENDALL, M. G. Entropy, probability and information, *International Statistical Review*, v.41, n.1, p.59–68, 1973.

KULLBACK, S. *Information theory statistics*, New York, Dover press, 1968.

———.; LEIBLER, R. A. On information and sufficiency, *Annals of Mathematical Statistics*, v.22, n.1, p.79–86, mar. 1951.

LINDLEY, D. V.; SMITH, A. F. M. Bayes estimates for the linear model, *Journal of the Royal Statistical Society*, série B, v.34, n.1, p.1–42, 1972.

MENDONÇA, Isabel G. S. Furtado de; MIGON, Hélio S. Pesquisa eleitoral: uma análise bayesiana, *Revista Brasileira de Estatística*, Rio de Janeiro, v.48, n.189/190, p.25–34, 1987.

MIGON, H. S., BRASIL, G. H. SOUZA, R. C. & SANT'ANNA, A. P. (1986), Relatórios Sobre Pesquisas Eleitorais no Estado do Rio de Janeiro utilizando a Metodologia Bayesiana, (Texto não publicado).

RODRIGUES, J. *The kullback–Leibler approximation of the marginal posterior density: an application to the linear functional model*, Notas do ICMSM, n.92.

SMITH, A. F. M. A general bayesian linear model, *Journal of the Royal Statistical Society*, serie B, v.35, n.1, p.67–73, 1973.

SOUZA, R. C.; BRASIL, G. H. A Bayesian model to forecast an election outcome: an application to the Brasilian states elections of 1986, *Estadistica*, v.41, n.136, p.13–30, 1989.

RESUMO

O modelo Bayesiano para previsão de resultados eleitorais proposto por Bernardo (1984) utiliza um desenho amostral diferente dos comumente adotados. Através da medida de divergência de Kullback–Leibler, selecionam–se os locais de votação (municípios, zonas eleitorais, etc.) mais representativos do comportamento político da população. Neste trabalho apresentamos um procedimento heurístico para a seleção do número de locais a serem pesquisados. A metodologia também pode ser usada em outras situac cões. Uma aplicação detalhada da metodologia proposta, juntamente com o modelo Bayesiano para previsão dos resultados finais de uma eleição, foi realizada para as eleições de 1990 ao Governo/Senado do Estado do Espírito Santo – Brasil. Também verificou–se a permanência dos locais selecionados, no período 1982–90.

ABSTRACT

Bernardo's Bayesian model (1984) to forecast an election outcome makes use of a sample design which is different from those commonly adopted. Utilizing Kullback–Leiber's measure one selects the *polling stations* from electoral zones which are more reepresentative of populations's political behavior. It is presented in this work an heuristic model for choosing the *polling stations* to be considered. This methodology may be applied to other situations such as income distribuition, for example. A practical use of this methodology, used whith the Bayesian model for forecasting election outcomes, was applied en Brazil's state of Espírito Santo for the 1990's elections of Government and Senate. It was verified that the electoral zones that was more representative of this model remainded aproximately the same from the period of 1982 to 1990.

COMPARAÇÃO DE DOIS MODELOS EXPONENCIAIS COM DADOS ACELERADOS: UMA ABORDAGEM BAYESIANA

Francisco Louzada-Neto*

Heleno Bolfarine**

Josemar Rodrigues**

1 INTRODUÇÃO

Um problema de grande interesse dos pesquisadores de indústrias de várias áreas é comparar a durabilidade de dois componentes manufaturados com especificações similares de diferentes fontes produtoras, com tempo e custo de experimentação reduzidos. Um procedimento viável é conduzir um teste de sobrevivência acelerado, no qual consideram-se vários níveis de variáveis físicas que estressam o componente acima dos níveis usuais de funcionamento (ver por exemplo, Nelson, 1990).

Considere T_j, $j = 1, 2$, uma variável aleatória denotando o tempo de sobrevivência de uma unidade, pertencente à j-ésima população, com função densidade de probabilidade dada por,

$$f(t_j, \lambda_j) = \lambda_j e^{-\lambda_j t_j} \tag{1}$$

onde $t_j > 0$ e $\lambda_j > 0$ é a taxa constante de falha.

* Universidade Federal de São Carlos - UFSCar

** Universidade de São Paulo - USP

Suponha que a variável aleatória T_j é afetada por uma variável física de estresse X.

Sob k níveis aleatorizados da variável estresse X, suponha o modelo estresse–resposta geral (ver, por exemplo, Achcar e Louzada–Neto, 1992) dado por

$$\lambda_{ij} = \exp\{-(Z_i + \beta_{0j} + \beta_{1j} X_i)\}, \tag{2}$$

onde β_{0j} e β_{1j} são parâmetros desconhecidos, Z_i é uma função de X_i, para $i = 1, 2, \cdots, k$ e $j = 1, 2$.

Como casos particulares de (2) temos:

a) se $X_i = \log V_i$, $Z_i = 0$, $\beta_{0j} = \log \alpha_j$ e $\beta_{1j} = \beta_j$ temos o modelo de Lei de Potência;

b) se $X_i = 1/V_i$, $Z_i = 0$, $\beta_{0j} = -\alpha_j$ e $\beta_{1j} = \beta_j$ temos o modelo de Arrhenius;

c) se $X_i = 1/V_i$, $Z_i = -\log V_i$, $\beta_{0j} = -\alpha_j$ e $\beta_{1j} = \beta_j$ temos o modelo de Eyring, onde V_i, em (i) representa uma variável de voltagem e em (ii) e (iii) uma variável de temperatura.

Considerando dados sob um esquema de censura de tipo II, isto é , o experimento termina quando r_i falhas são observadas entre n_i unidades em testes no i–ésimo nível de estresse para $j = 1, 2$, temos, $t_{ij1}, t_{ij2} \cdots, t_{ij r_i}$ observações não–censuradas e $n_i - r_i$ observações censuradas iguais a $t_{ij r_i}$ para $i = 1, 2, \cdots, k$.

A função de verossimilhança para β_{0j} e β_{1j}, $j = 1, 2$, considerando–se k níveis aleatorizados de estresse e um esquema de censura de tipo II, é dada por,

$$L_j(\beta_{0j}, \beta_{1j}) \propto \exp\left\{-\beta_{0j} r - \beta_{1j} a - e^{-\beta_{0j}} \sum_{i=1}^{k} A_{ij} e^{-Z_i - \beta_{1j} X_i}\right\}, \tag{3}$$

onde $r = \sum_{i=1}^{k} r_i$ (número total de falhas entre n unidades em teste),

$$a = \sum_{i=1}^{k} r_i X_i \quad \text{e} \quad A_{ij} = \sum_{l=1}^{r_i} t_{ijl} + (n_i - r_i) t_{ij r_i}.$$

Em geral, o interesse dos pesquisadores é obter inferências sobre o tempo médio de sobrevivência sob o nível usual de estresse dado por,

$$\theta_{1j} = \frac{1}{\lambda_{1j}}, \quad j = 1, 2, \tag{4}$$

isto é, de (2) e (4),

$$\theta_{1j} = \exp\{Z_i + \beta_{0j} + \beta_{1j} X_i\}. \tag{5}$$

A função de verossimilhança para θ_{1j} e β_{1j} é dada (de (3)) por,

$$L_j(\theta_{1j}, \beta_{1j}) \propto \theta_{1j}^{-r} \exp\left\{\beta_{1j} r\mathbf{X} - \theta_{1j}^{-1} \sum_{i=1}^{k} A_{ij} e^{-(Z_i - Z_1) + \beta_{1j}(X_1 - X_i)}\right\} \tag{6}$$

onde $\mathbf{X} = \frac{\sum_{i=1}^{k} r_i(X_1 - X_i)}{r}$ (média ponderada dos $X_i's$) e r e A_{ij} são dados em (3).

A matriz de informação de Fisher para θ_{1j} e β_{1j} é dado por,

$$I_j(\theta_{1j}\beta_{1j}) = \begin{bmatrix} \frac{r}{\theta_{1j}^2} & -\frac{r\mathbf{X}}{\theta_{1j}} \\ -\frac{r\mathbf{X}}{\theta_{1j}} & b \end{bmatrix} \tag{7}$$

onde $b = \sum_{i=1}^{k} r_i(X_1 - X_i)^2$.

A partir de matriz de informação de Fisher (7), as inferências sobre θ_{1j} e β_{1j} podem ser obtidas utilizando-se a normalidade assintótica dos Estimadores de Máxima Verossimilhança.

2 REPARAMETRIZAÇÃO ORTOGONAL

Para obtermos facilidades computacionais na obtenção do estimador de β_{1j} , uma parametrização para diagonalizar a matriz de informação de Fisher (7) será obtida resolvendo a equação diferencial (ver por exemplo, Cox e Reid, 1987; Rodrigues, Achcar e Louzada-Neto, 1993), dada por,

$$-I_{\theta_{1j},\beta_{1j}} = I_{\theta_{1j},\theta_{1j}} \frac{\partial \theta_{1j}}{\partial \beta_{1j}} \tag{8}$$

onde $I_{\theta_{1j},\theta_{1j}}$ e $I_{\theta_{1j},\beta_{1j}}$ são elementos da matriz (7).

Uma solução da equação (8) é dada por,

$$\mathbf{X}\theta_{1j} = \log(\theta_{1j}) + c(\psi_{1j}), \tag{9}$$

onde $c(\psi_{1j})$ é uma função arbitrária de (ψ_{1j}).

Considerando $c(\psi_{1j}) = \log(1/\psi_{1j})$ temos,

$$\psi_{1j} = \theta_{1j} e^{-\beta_{1j}\mathbf{X}}. \tag{10}$$

A função de verossimilhança para ψ_{1j} e β_{1j} é dada (de (6) e (10)) por,

$$L_j(\psi_{1j}, \beta_{1j}) \propto \psi_{1j}^{-r} \exp\left\{-\psi_{1j}^{-1} \sum_{i=1}^{k} A_{ij} e^{-(Z_i - Z_1) + \beta_{1j}(X_1 - X_i - \mathbf{X})}\right\}, \tag{11}$$

e a matriz de informação de Fisher para ψ_{1j} e β_{1j} é dada por,

$$I_j(\psi_{1j}, \beta_{1j}) = \begin{bmatrix} \frac{r}{\psi_{1j}^2} & 0 \\ 0 & c \end{bmatrix}, \tag{12}$$

onde $c = \sum_{i=1}^{k} r_i(X_1 - X_i - \mathbf{X})^2$.

Observe que (12) é uma matriz diagonal, isto é, os parâmetros ψ_{1j} e β_{1j} são ortogonais (ver por exemplo, Cox e Ried, 1987).

3 UMA ANÁLISE BAYESIANA VIA REPARAMETRIZAÇÃO ORTOGONAL

A densidade *a priori* não-informativa de Jeffreys para ψ_{1j} e β_{1j} é dada (de (12)) por,

$$\pi_j(\psi_{1j}, \beta_{1j}) \propto \frac{1}{\psi_{1j}}, \tag{13}$$

onde $\psi_{1j} > 0$ e $-\infty < \beta_{1j} < \infty$.

A densidade *a posteriori* conjunta para ψ_{1j} e β_{1j} é dada por,

$$\pi_j(\psi_{1j}, \beta_{1j}/dados) \propto \psi_{1j}^{-(r+1)} \exp\left\{-\psi_{1j}^{-1} \sum_{i=1}^{k} A_{ij} e^{-(Z_i - Z_1) + \beta_{1j}(X_1 - X_i - \mathbf{X})}\right\}, \tag{14}$$

onde $\psi_{1j} > 0$ e $-\infty < \beta_{1j} < \infty$.

A densidade *a posteriori* marginal para β_{1j} é dada por,

$$\pi(\beta_{1j}/dados) \propto \left\{\sum_{i=1}^{k} A_{ij} e^{-(Z_i - Z_1) + \beta_{1j}(X_1 - X_i - \mathbf{X})}\right\}^{-r}, \tag{15}$$

onde $-\infty < \beta_{1j} < \infty$.

A densidade *a posteriori* marginal para ψ_{1j} aproximada pelo método de Laplace (ver, por exemplo, Kass, Tierney e Kadane, 1990) é dada por,

$$\pi(\psi_{1j}/dados) \propto \frac{\psi_{1j}^{-(r+1/2)} \exp\left\{-\psi_{1j}^{-1} \sum_{i=1}^{k} A_{ij} e^{-(Z_i - Z_1) + \hat{\beta}_{1j}(X_1 - X_i - \mathbf{X})}\right\}}{\left\{\sum_{i=1}^{k} A_{ij} (X_1 - X_i - \mathbf{X})^2 e^{-(Z_i - Z_1) + \hat{\beta}_{1j}(X_1 - X_i - \mathbf{X})}\right\}^{1/2}}, \tag{16}$$

onde $\hat{\beta}_{1j}$ é o valor que maximiza

$$-nh_{\psi_{1j}}(\beta_{1j}) = -\psi_{1j}^{-1} \sum_{i=1}^{k} A_{ij} \cdot e^{-(Z_i - Z_1) + \beta_{1j}(X_1 - X_i - \mathbf{X})}.$$

Observe que maximizar $-nh_{\psi_{1j}}(\beta_{1j})$ é equivalente a achar a moda de (15), $\tilde{\beta}_{1j}$, isto é, podemos tomar $\hat{\beta}_{1j} = \tilde{\beta}_{1j}$. Também, é importante observar que após a reparametrização ortogonal obtivemos um estimador para β_{1j} que independe ψ_{1j}. Conseqüentemente obtemos uma simplificação na expressão (16), isto é,

$$\pi(\psi_{1j}/dados) \propto \psi_{1j}^{-(r+1/2)} \exp\left\{-\psi_{1j}^{-1} \sum_{i=1}^{k} A_{ij} e^{-(Z_i - Z_1) + \tilde{\beta}_{1j}(X_1 - X_i - \mathbf{X})}\right\}. \tag{17}$$

Tomando $\beta_{1j} = \tilde{\beta}_{1j}$ em (10) temos

$$\psi_{1j} = \theta_{1j} e^{-\tilde{\beta}_{1j}\mathbf{x}}, \tag{18}$$

função um-a-um de ψ_{1j} e θ_{1j}, e a densidade *a posteriori* para θ_{1j} é dada (de (17)) por

$$\pi(\theta_{1j}/dados) \propto \theta_{1j}^{-(r+1/2)} \exp\left\{-\theta_{1j}^{-1} \sum_{i=1}^{k} A_{ij} e^{-(Z_i - Z_1)+\tilde{\beta}_{1j}(X_1 - X_i)}\right\}. \tag{19}$$

onde $\theta_{1j} > 0$.

A moda da densidade *a posteriori* para θ_{1j} (19) é dada por,

$$\tilde{\theta}_{1j} = \frac{\sum_{i=1}^{k} A_{ij} e^{-(Z_i - Z_1)+\tilde{\beta}_{1j}(X_1 - X_i)}}{r + 1/2}. \tag{20}$$

Observando (de (19)) que

$$\frac{2\sum_{i=1}^{k} A_{ij} e^{-(Z_i - Z_1)+\tilde{\beta}_{1j}(X_1 - X_i)}}{\theta_{1j}}$$

é a densidade de uma Distribuição Qui-Quadrado com $(2r-1)$ graus de liberdade, um intervalo HPD $100(1-\gamma)\%$ para θ_{1j} é dado por,

$$\left(\frac{2\sum_{i=1}^{k} A_{ij} e^{-(Z_i - Z_1)+\tilde{\beta}_{1j}(X_1 - X_i)}}{\chi^2_{2r-1,(1-\gamma/2)}}; \frac{2\sum_{i=1}^{k} A_{ij} e^{-(Z_i - Z_1)+\tilde{\beta}_{1j}(X_1 - X_i)}}{\chi^2_{2r-1,(\gamma/2)}},\right) \tag{21}$$

onde $\chi_{2r-1(\gamma/2)}$ é o quantil da distribuição Qui-Quadrado dado por,

$$P(\chi^2_{2r-1} = \chi^2_{2r-1,(\gamma/2)}) = \frac{\gamma}{2}.$$

É importante salientar que os resultados (20) e (21) são conseqüências diretas da ortogonalização da matriz de informações de Fisher (7).

4 COMPARAÇÃO ENTRE DUAS POPULAÇÕES COM DADOS ACELERADOS

Um procedimento usual para compararmos duas populações, é obtermos inferências sobre $\psi = \theta_{11}/\theta_{12}$ (ver, por exemplo, Box e Tiao, 1973).

Considerando a densidade *a priori* não-informativa de Jeffreys, $\pi(\psi_{11},\beta_{11},\psi_{12},\beta_{12}) \propto (\psi_{11}\psi_{12})^{-1}$, a densidade *a posteriori* conjunta para $\psi_{11},\psi_{12},\beta_{11},\beta_{12}$ é dada (ver procedimento análogo na seção anterior) por,

$$\pi(\psi_{11},\beta_{11},\psi_{12},\beta_{12}/dados) \propto (\psi_{11},\psi_{12})^{-(r+1)}.$$
$$\cdot \exp\left\{-\psi_{11}^{-1}\sum_{i=1}^{k} A_{i1}e^{-(Z_i-Z_1)+\beta_{11}(X_1-X_i-\mathbf{X})}\right\}.$$
$$\cdot \exp\left\{-\psi_{12}^{-1}\sum_{i=1}^{k} A_{i2}e^{-(Z_i-Z_1)+\beta_{12}(X_1-X_i-\mathbf{X})}\right\}, \qquad (22)$$

onde $\psi_{11},\psi_{12} > 0$ e $-\infty < \beta_{11},\beta_{12} < \infty$.

A densidade *a posteriori* conjunta marginal para ψ_{11},ψ_{12} aproximada pelo método de Laplace é dada por,

$$\pi(\psi_{11},\psi_{12}/dados) \propto (\psi_{11},\psi_{12})^{-(r+1/2)}.$$
$$\cdot \exp\left\{-\psi_{11}^{-1}\sum_{i=1}^{k} A_{i1}e^{-(Z_i-Z_1)+\tilde{\beta}_{11}(X_1-X_i-\mathbf{X})}\right\}.$$
$$\cdot \exp\left\{-\psi_{12}^{-1}\sum_{i=1}^{k} A_{i2}e^{-(Z_i-Z_1)+\tilde{\beta}_{12}(X_1-X_i-\mathbf{X})}\right\}, \qquad (23)$$

onde $\tilde{\beta}_{11}$ e $\tilde{\beta}_{12}$ são as modas da *posteriori* (15) para $j = 1,2$.

Considerando a transformação de variáveis $\mu = \psi_{11}/\psi_{12}$ e $\psi_{12} = \psi_{12}$, de (23) temos,

$$\pi(\mu,\psi_{12}/dados) \propto \mu^{-(r+1/2)}\psi_{12}^{-2r}.$$
$$\cdot \exp\left\{-\psi_{12}^{-1}\left[\mu^{-1}\sum_{i=1}^{k} A_{i1}e^{-(Z_i-Z_1)+\tilde{\beta}_{11}(X_1-X_i-\mathbf{X})} + \right.\right.$$
$$\left.\left. + \sum_{i=1}^{k} A_{i2}e^{-(Z_i-Z_1)+\tilde{\beta}_{12}(X_1-X_i-\mathbf{X})}\right]\right\}, \qquad (24)$$

onde $\mu,\psi_{12} > 0$.

A densidade *a posteriori* marginal para μ é dada por

$$\pi(\mu/dados) \propto$$
$$\propto \frac{\mu^{-(r+1/2)}}{\left\{\mu^{-1}\sum_{i=1}^{k} A_{i1}e^{-(Z_i-Z_1)+\tilde{\beta}_{11}(X_1-X_i-\mathbf{X})} + \sum_{i=1}^{k} A_{i2}e^{-(Z_i-Z_1)+\tilde{\beta}_{12}(X_1-X_i-\mathbf{X})}\right\}^{2r-1}} \qquad (25)$$

onde $\mu > 0$.

Como temos interesse em $\psi = \theta_{11}/\theta_{12}$, considerando a transformação de variáveis $\mu = \psi e^{(\tilde{\beta}_{12},\tilde{\beta}_{11})\mathbf{X}}$ (ver (18)), a densidade a posteriori para ψ é dada (de (25)) por,

$$\pi(\psi/dados) \propto$$

$$\frac{\psi^{-(r+1/2)}}{\left\{\psi^{-1}e^{-(\hat{\beta}_{12}-\hat{\beta}_{11})\mathbf{X}}\sum_{i=1}^{k} A_{i1}e^{-(Z_i-Z_1)+\tilde{\beta}_{11}(X_1-X_i-\mathbf{X})} + \sum_{i=1}^{k} A_{i2}e^{-(Z_i-Z_1)+\tilde{\beta}_{12}(X_1-X_i-\mathbf{X})}\right\}^{2r-1}} \tag{26}$$

onde $\psi > 0$

A moda desta densidade *a posteriori* é dada por,

$$\tilde{\psi} = \frac{\sum_{i=1}^{k} A_{i1}e^{-(Z_i-Z_1)+\tilde{\beta}_{11}(X_1-X_i)}}{\sum_{i=1}^{k} A_{i2}e^{-(Z_i-Z_1)+\tilde{\beta}_{12}(X_1-X_i)}} \cdot \left(\frac{r-3/2}{r+1/2}\right), \tag{27}$$

isto é (de (20)).

$$\tilde{\psi} = \frac{\tilde{\theta}_{11}}{\tilde{\theta}_{12}} \cdot \frac{r-3/2}{r+1/2} \tag{28}$$

5 UM EXEMPLO NUMÉRICO

Considere os dados da Tabela 1, gerados assumindo–se uma distribuição Exponencial e o modelo de Lei de Potência como $\alpha_j = 400$ e $\beta_j = 0,70, j = 1,2$, isto é, $\beta_{0j} = 5,9915$, $\beta_{1j} = 0,70$, $X_i = \log V_i$ e $Z_i = 0$ no modelo estresse–resposta geral (2).

TABELA 1

Dados Gerados com $\alpha_j = 400$ e $\beta_j = 0,70$

i	V_i	n_i	r_i	θ_{ij}	AMOSTRA 1	AMOSTRA 2
1	5	10	4	129,65	16,30,41,63	14,27,42,58
2	10	10	6	79,81	15,21,36,39,54,59	17,21,32,45,59,61
3	15	10	7	60,09	18,29,39,42,46,57,60	14,23,26,27,35,42,78
4	20	10	8	49,13	19,20,36,37,41,45,45,57	17,17,20,32,35,40,64,71
5	25	10	9	42,02	8,12,13,14,23,33,42,51,67	10,17,21,28,33,44,53,69,75

As modas da densidade *a posteriori* (15) para as duas amostras são dadas por $\tilde{\beta}_{11} = 0,7526$ e $\tilde{\beta}_{12} = 0,5824$, e as modas *a posteriori* para os tempos médios de sobrevivência sob o nível

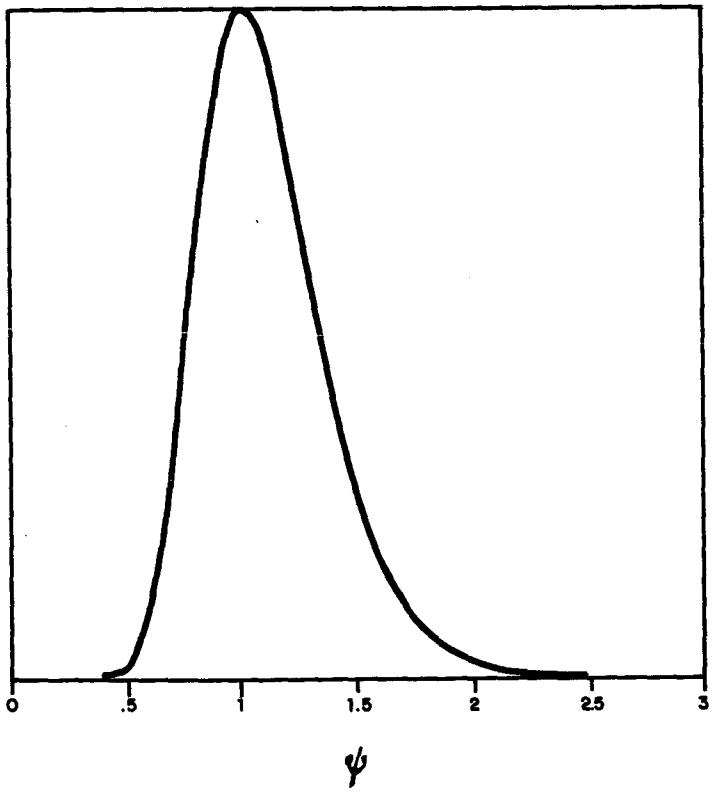
usual, X_1, são dadas (de (20)) por, $\tilde{\theta}_{11} = 135,1634$ e $\tilde{\theta}_{12} = 121,2108$ (próximos do valor de geração dado por 129,65 - ver Tabela 1).

Os intervalos HPD aproximados 95% para θ_{ij}, $j = 1,2$, são dados (de (21)) por, $(102,4588 < \theta_{11} < 203,5550)$ e $(91,8821 < \theta_{12} < 182,5423)$.

Na Figura 1 temos o gráfico da densidade *a posteriori* para o quociente dos tempos médios de sobrevivência ψ (ver (26)). A moda desta densidade *a posteriori* é dada (de (28)) por $\tilde{\psi} = 1,0505$.

Figura 1

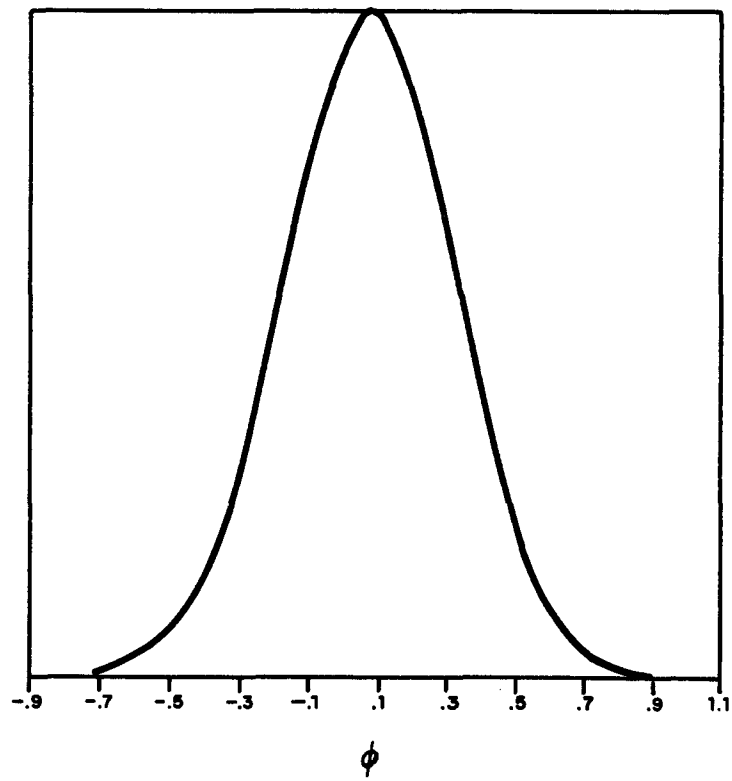
Densidade *a posteriori* para ψ.



Considerando a escala logarítmica ($\phi = \log\psi$), temos na Figura 2 o gráfico da densidade *a posteriori* para ϕ, onde podemos observar uma boa forma simétrica, indicando maior precisão da aproximação de Laplace (ver, por exemplo, Achcar e Louzada-Neto, 1992). Um intervalo HPD aproximado 95% dada ϕ é dado (ver Figura 2) por $-0,4855 < \phi < 0,5587$), isto é, um intervalo HPD aproximado 95% para ψ é dado por $(0,6154 < \psi < 1,7484)$, indicando que os tempos médios de sobrevivência das duas populações são iguais.

Figura 2

Densidade *a posteriori* para $\phi = log\psi$.



6 CONCLUSÕES

Utilizando uma reparametrização ortogonal obtivemos um estimador de β_{1j} que independe de θ_{1j}. Esta estabilidade do estimador foi fundamental para obtermos uma densidade *a posteriori* "fechada" para o parâmetro θ_{1j}, e a moda *a posteriori* do quociente dos tempos médios de sobrevivência sob o nível usual $(\tilde{\psi})$ facilmente interpretável.

BIBLIOGRAFIA

ACHCAR, J. A.; LOUZADA-NETO, F. A Bayesian approach for accelerated life tests considering the Weibull distribution, *Computational Statistics*, n.7, p.355-69, 1992.

BOX, G. E. P.; TIAO, G. C. *Bayesian inference in statistical analysis*. Addison-Wesley, 1977.

COX, D. R.; REID, N. Parameters orthogonality and approximated conditional inference. *Journal of the Royal Statistical Society*, série B, v.49, n.1, p.1-39, 1987.

KASS, R. E.; TIERNEY, L.; KANADE, J. B. The validity of posteriori expansions based on Laplace's method. In: HODGES, J. (ed.) Essays in Honor of George A. Barnard, Amsterdam, North-Holland, p.473-88, 1990.

NELSON, W. *Accelerated testing: statistical models*, test plans, and data analysis. New York, John Wiley, 1990.

RODRIGUES, J.; ACHCAR, J. A.; LOUZADA-NETO, F. A Bayesian analysis of the accelerated life tests via the orthogonal parameters. *Statistics*, n.24, p.353-57, 1993.

RESUMO

Testes de produtos manufaturados sob condições aceleradas, em geral, são mais eficientes que os testes realizados sob condições usuais de funcionamento. Um problema prático, de grande interesse das indústrias, é a comparação da durabilidade média dos componentes sob condições usuais quando os dados são obtidos sob condições aceleradas. Neste artigo, desenvolvemos uma Análise Bayesiana para compararmos duas populações sob condições aceleradas considerando uma distribuição Exponencial para os tempos de sobrevivência. Um modelo estresse-resposta geral com uma variável estresse que inclui alguns dos modelos mais conhecidos em testes acelerados é adotado. Assumindo dados sob um esquema de censura de tipo 11 e uma densidade *a priori* não-informativa de Jeffreys, obtemos a densidade *a posteriori* do quociente dos tempos médios de sobrevivência das duas populações. Através de uma reparametrização ortogonal (Cox e Reid, 1987) obtemos uma expressão, em forma fechada, para a moda *a posteriori* deste quociente. O método de Laplace é utilizado para aproximarmos as densidades a posteriori de interesse quando não conseguimos achar suas soluções analíticas explícitas. A metodologia proposta é ilustrada com um exemplo.

ABSTRACT

Testing manufactured products under accelerated conditions, typically is more efficient than tests performed under the usual stress level conditions. A practical problem of great interestin many industries is the comparison of mean lifetimes of items under the usuable conditions when data is obtained under accelerated conditions. In this paper, we develop a Bayesian analysis for comparing two populations under accelerated conditions considering an exponential distribution for surviving times. A general stress–response model with one stress variable, which includes several well known stress models is adopted. Assuming type II censoring and the Jefreys' non-informative prior, we derive the quocient of the mean survival times under the two populations. Using orthogonal parametrization (Cox and Reid, 1987), a closed form expression is obtained for the posterior mode of this quocient. Laplace's method for approximating integrals is used when explicit analytical expressions for the posterior densities are not possible to obtain. The proposed methodology is illustrated with an example.

SELEÇÃO DE MODELOS DE REGRESSÃO PARA PREDIÇÃO VIA VALIDAÇÃO CRUZADA: UMA APLICAÇÃO EM AVALIAÇÃO DE IMÓVEIS

Emanuel Pimentel Barbosa*

Cláudio P. Bidurin**

1 INTRODUÇÃO

A utilização de metodologia científica baseada em modelos de regressão como instrumento quantitativo útil na abordagem de problemas de avaliação de imóveis, e particularmente no caso de lotes urbanos, é apresentada em Dantas & Cordeiro (1988). Em tal abordagem, os modelos de regressão, modelos lineares generalizados, são utilizados para homogeneizar os dados disponíveis, isto é, possibilitar a comparação de um dado imóvel com outros de um conjunto de referência; nesse caso, as características físicas, de localização, etc., dos imóveis, são utilizadas como covariáveis do modelo de regressão, onde a variável de resposta é o preço do imóvel. O uso de modelos lineares generalizados – McCullagh & Nelder (1983, 89) – e não de modelos lineares normais usuais, deve-se ao fato da variável de resposta, preço, ser estritamente positiva, como é o caso, por exemplo, de uma variável Gama ou Log-Normal.

Neste artigo, é proposto como critério para escolha entre possíveis modelos concorrentes, entre eles os modelos de regressão Gama e Log-Normal, o uso de Validação Cruzada, e não simplesmente medidas de ajuste, como a Deviance, considerada no citado artigo. Também, os

* Universidade Federal de Minas Gerais – UFMG

** Universidade Federal de São Carlos – UFScar

erros padrão das estimativas dos parâmetros do modelo Gama são recalculados separadamente, uma vez que os resultados fornecidos pelo sistema GLIM 3.77, são bastante imprecisos (ver Aitkin pág. 278). Com isso, predições por intervalo omitidas no citado artigo são agora fornecidas.

A organização deste artigo é a seguinte: Na Seção 2 é apresentado sucintamente o problema da avalição de imóveis. Na Seção 3, são introduzidos formalmente os modelos de regressão Box–Cox e modelos lineares generalizados, e em particular, uma análise do modelo Gama, como também uma discussão sobre critérios para comparação desses modelos, com ênfase no critério de Validação Cruzada. Na Seção 4 são apresentados os resultados da análise de um conjunto de dados reais, preços e outras características de um conjunto de lotes urbanos, utilizando–se da metodologia introduzida na seção anterior. Na Seção 5 é apresentada uma discussão sobre a metodologia proposta, além de se sugerir uma extensão dessa metodologia de modo a se possibilitar a inclusão no modelo, caso disponível, de informação subjetiva, de profissionais de avaliação, via uma abordagem Bayesiana.

2 O PROBLEMA DA AVALIAÇÃO DE IMÓVEIS

2.1 Metodologia Usual

Em problemas envolvendo a avaliação de imóveis, como discutido em Dantas & Cordeiro (1988), com o intuito de se avaliar seus valores de compra e venda, geralmente os especialistas nesta área se utilizam de métodos comparativos, recomendados pela ABNT, baseados em um conjunto de imóveis com uma certa homogeneidade de características, com preços conhecidos, para com estes se conseguir determinar os preços de outros imóveis de características similares.

Esta metodologia pode apresentar entretanto alguns problemas sérios, como por exemplo, a questão da homogeneidade do grupo de referência, que em muitos casos pode ser difícil de se obter. Para uma aplicação adequada do método comparativo, é necessário, de alguma forma, homogeneizar previamente os dados do conjunto de referência. Um outro problema é que as comparações são feitas de uma forma subjetiva, com perda de precisão nos resultados, implicando em possíveis variações de critério, quando se muda o especialista que efetua a comparação.

A aplicação de uma metodologia científica poderia facilitar a abordagem do problema da predição dos valores dos imóveis, com a construção de método eficiente de homogeneização do grupo de referência.

2.2 Dados Disponíveis

São disponíveis dados sobre características físicas (Testada e Profundidade), características de localização (Bairro e Grau de Urbanização) e econômicas (Tipo e época do Negócio), de 50 lotes urbanos de 3 bairros da cidade de Recife, PE.

Descrição das características:

– Testada Efetiva (T): Tipo de projeção da frente real do imóvel, em metros.

– Profundidade Equivalente (P): Razão entre a área e a testada efetiva, em metros.

– Natureza do Evento (N): Tipo de venda. Variável Dummy: 0 se foi oferta, 1 se foi negociação.

– Época de Ocorrência (C): Número de meses até a ocorrência do evento, em relação a Dez./79.

– Nível de Urbanização (U): Infra–estrutura existente na vizinhança do imóvel, variando de 3 a 8.

– Localização (R): Variável indicadora: 1– Imóvel do Bairro Iputinga, 2 – Imóvel do Bairro Torre, 3– Imóvel do Bairro Casa Forte.

Uma metodologia científica adequada para proceder à homogeneização do grupo de referência, é a montagem de um modelo estatístico de regressão, para se ajustar o preço dos imóveis, em Cr\$/m^2, em função das características anteriores.

3 MODELOS

3.1 Introdução

Como já evidenciado nas seções anteriores, a construção de modelos de regressão para predição do valor de imóveis, é uma abordagem alternativa adequada para os problemas encontrados nos métodos usuais de avaliação.

Para a construção do modelo, dispomos de um conjunto de 6 variáveis explicativas, que são as características dos terrenos, já citadas anteriormente, e como variável resposta, os preços dos terrenos (V). Sobre a variável resposta, devemos considerar o seguinte:

– Os preços dos terrenos, dados em Cr\$/m^2, são estritamente positivos, $V_i > 0, i = 1, \cdots, 50$. Devido a esta característica, principalmente, o ajuste de modelos de regressão usuais, isto é, modelos lineares normais, é inadequado, sendo possível, entretanto, a utilização de pelo menos duas alternativas distintas: A primeira delas é o uso de transformação nos dados, isto é, de se considerar um modelo de regressão com a transformação BOX–COX nos dados de modo

a se obter aproximadamente normalidade e variância constante. Uma segunda alternativa consiste em se considerar não uma transformação nos dados, mas sim no parâmetro média da distribuição dos preços através de uma função de ligação, como detalhado mais adiante, considerando os dados em sua escala original com distribuição possivelmente na família exponencial, levando ao uso dos chamados modelos lineares generalizados.

3.2 Revisão Teórica

Com o intuito de facilitar a compreensão do leitor não familiarizado com a metodologia proposta, apresentamos uma revisão sucinta dos conceitos relacionados aos modelos que serão utilizados, bem como do critério de seleção baseado na Validação Cruzada.

3.2.1 Modelos de Regressão Box–Cox

A partir do modelo de regressão linear, dado em notação usual por $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i$ para $i = 1, \cdots, n$, de modo a se obter normalidade para os dados y_i, considera–se uma transformação dos mesmos, do tipo Box–Cox, resultando: (ver Zellner, 1971, cap.VI)

$$\frac{y_i^{\lambda} - 1}{\lambda} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i \quad \text{para} \quad i = 1, \cdots, n \tag{3.1}$$

onde: β_i são os coeficientes de regressão e λ é o parâmetro da transformação, com $y_i > 0$. Reescrevendo (3.1) em forma matricial, temos,

$$y^{(\lambda)} = X\beta + e \tag{3.2}$$

$y^{(\lambda)}$: vetor $(n \times 1)$ de valores do tipo $(y_i^{(\lambda)} - 1)/\lambda$.

X: matriz $(n \times k)$ de valores das variáveis explicativas.

β: vetor $(k \times 1)$ dos parâmetros de regressão.

e: vetor $(n \times 1)$ de erros assumidos normais, independentes, com média 0 e variância constante e igual a σ^2.

Dado λ, a partir do máximo da função de verossimilhança do modelo (3.2), temos as estimativas abaixo:

$$\hat{\beta}(\lambda) = (X'X)^{-1} X' y^{(\lambda)}$$

$$\hat{\sigma}^2(\lambda) = \frac{1}{n}(y^{(\lambda)} - X\hat{\beta})'(y^{(\lambda)} - X\hat{\beta})$$

Se λ for conhecido, teremos os estimadores de máxima verossimilhança. Porém , como λ é assumido desconhecido, utilizamos os estimadores acima, para obter o log da função de verossimilhança maximizada, denotada por $L_{max}(\lambda)$, dada por:

$$L_{max}(\lambda) = k + (\lambda - 1)\sum_{i=1}^{n} \log y_i - \frac{n}{2}\log \hat{\sigma}^2(\lambda)$$

onde: k é uma constante.

Então, tomando uma malha de pontos para λ, na prática, $-3 \leq \lambda \leq 3$ aproximadamente, o ponto λ que fizer a função $L_{max}(\lambda)$ atingir seu máximo, será o $\hat{\lambda}$, de máxima verossimilhança procurado, e que indicará a transformação a ser usada.

Como casos particulares importantes, entre outros, temos:

$\lambda = 1$ Caso normal, sem transformação.

$\lambda = 0$ Transformação logarítmica, modelo Log–Normal.

A implementação da transformação Box–Cox será feita via sistema GLIM utilizando–se a macro BOXCOX. (ver Baker & Nelder, 1978).

3.2.2 Modelos Lineares Generalizados

Um modelo linear generalizado (ver McCullagh & Nelder, 1983,89) é caracterizado por 3 componentes básicos:

(a) Uma variável resposta y com distribuição pertencente à família exponencial, tal que $E(y) = \mu$.

(b) Um preditor linear η, que é uma combinação linear das variáveis explicativas x, ou seja, $\eta = X\beta$, onde β é um vetor de coeficientes a serem estimados e X tem suas colunas formadas pelas variáveis explicativas.

(c) Uma função de ligação $g(\bullet)$, monótona e diferenciável, entre η e μ, de forma que $g(\mu) = \eta$.

Alguns casos particulares importantes de variáveis pertencentes à família exponencial, correspondem entre outras, às distribuições Normal, Gama, Poisson e Binomial, sendo de interesse neste trabalho as variáveis Normal e Gama.

Como temos a suposição de que a distribuição dos dados pertence à família exponencial, a função de verossimilhança para os modelos lineares generalizados, pode ser escrita de uma forma geral, como:

$$L(\beta) = \sum_{i=1}^{n} [\phi_i \{y_i\theta_i - b(\theta_i)\} + c(y_i; \phi_i)] \tag{3.3}$$

onde $b(\bullet)$ e $c(\bullet)$ são funções conhecidas, sendo que as duas primeiras derivadas de $b(\theta)$ correspondem respectivamente à média e à função de variância dos dados y; $\phi_i > 0$ é o parâmetro

de escala da distribuição, suposto conhecido e θ_i é o parâmetro canônico da distribuição, que é função de β.

A partir de (3.3), para se encontrar as estimativas de máxima verossimilhança $\hat{\beta}$ do modelo, alguns problemas aparecerão, já que o sistema resultante dado por $(\partial L(\beta))/\partial\beta = 0$ é não-linear. Para resolvê-lo, é necessária a aplicação de métodos numéricos, como o método de Newton-Raphson, ou mais adequadamente, uma modificação deste denominado algorítmo escore, que aplicado ao problema em questão resulta, após alguma álgebra, em um algorítmo iterativo para o cálculo dos $\hat{\beta}$ e que possui a seguinte estrutura:(ver McCullagh & Nelder, 1989).

$$X'W^{(m)}\phi X\beta^{(m+1)} = X'W^{(m)}\phi y^{*(m)} \tag{3.4}$$

onde: (m): m-ésimo passo do algorítmo.

W e ϕ são matrizes diagonais com $W = diag\{w_1 \cdots w_n\}$ com $w_i = (d\mu/d\eta)^2/V$ e $\phi = diag\{\phi_1 \cdots \phi_n\}$ onde ϕ_i; são os parâmetros de escala.

$y^{*(m)}$ é a variável resposta modificada da forma $y^* = X\beta + H(y-\mu)$, onde H é uma matriz diagonal dada por $H = diag\{d\eta_1/d\mu_1 \cdots d\eta_n/d\mu_n\}$.

Este algorítmo é denominado I.R.L.S., "Interactively Reweighted Least Squares", e está implementado no sistema GLIM.

3.3 Modelo Gama

Como já definimos anteriormente, os dados $y_i = V_i$ são estritamente positivos e assimétricos, indicando que o ajuste de um modelo Gama pode ser adequado. Como a distribuição Gama pertence à família exponencial, é possível então, o ajuste de um modelo linear generalizado com distribuição Gama para os dados, supondo a existência de uma função de ligação adequada para o caso.

A implementação do modelo Gama será feita no sistema GLIM 3.77, que se utiliza do algorítmo (3.4) para obter as estimativas dos parâmetros β do modelo. Entretanto, o cálculo dos desvios padrão das estimativas dos parâmetros como fornecido pelo sistema GLIM é bastante impreciso, sendo necessário se efetuar uma correção nos mesmos. Tal limitação do referido *software* se deve ao fato de não haver disponível nesta versão a função matemática digama, necessária para o cálculo dos citados erros padrão, como mostrado mais adiante (para maiores detalhes sobre a limitação do *software* ver Aitkin pág. 278). Esta correção se baseia na relação entre a distribuição Gama e a Exponencial, levando-se em conta a matriz de informação de Fisher. Mostremos então como fica esta correção.

A distribuição Gama é caracterizada usualmente por dois parâmetros, μ e r, onde μ é média e r é o parâmetro de forma da distribuição. A função densidade de probabilidade é então da

forma,

$$f(y,r,\mu) = \frac{r^r}{\Gamma(r)\mu^r} e^{-ry/\mu} y^{r-1} \quad \text{com} \quad y,r,\mu > 0 \tag{3.5}$$

A log–verossimilhança de (3.5) para n observações independentes é então dada por:

$$\log L = nr\log r - n\log\Gamma(r) - r\sum_{i=1}^{n}\log\mu_i - r\sum_{i=1}^{n}\frac{y_i}{\mu_i} + (r-1)\sum_{i=1}^{n}\log y_i \tag{3.6}$$

cuja segunda derivada em relação a β considerando uma função de ligação logarítmica do tipo $\log\mu_i = \beta' x_i$, onde x_i é a i–ésima linha da matriz X definida em 3.2.2, fornece,

$$\frac{\partial^2 \log L}{\partial\beta\partial\beta'} = -r\sum_{i=1}^{n}\frac{y_i X_i X_i}{\mu_i} \tag{3.7}$$

Considerando o valor esperado do lado esquerdo de (3.7) com o sinal trocado, temos a matriz de informação de Fisher, que se relaciona com a matriz de variância–covariância assintótica do estimador $\hat{\beta}$ pela relação,

$$\left[E\left(\frac{-\partial^2 \log L}{\partial\beta\partial\beta'}\right)\right]^{-1} = Var(\hat{\beta})gama \tag{3.8}$$

Partindo de (3.7) e (3.8), podemos concluir que a $Var(\hat{\beta})$ do modelo Gama, será igual a $Var(\hat{\beta})$ do modelo Exponencial, que é fornecida diretamente pelo GLIM, dividida por r. Então nosso problema agora é estimar r, para determinarmos $Var(\hat{\beta})$ para o modelo Gama, que surpreendentemente não é fornecida corretamente pelo sistema GLIM.

Derivando a função de log–verossimilhança dada em (3.6), em relação a r, obtemos o estimador de máxima verossimilhança $\hat{r}$, que é a solução da equação abaixo:

$$n(1+\log r) - n\psi(r) - \sum_{i=1}^{n}\log\hat{\mu}_i - \sum_{i=1}^{n}\frac{y_i}{\hat{\mu}_i} + \sum_{i=1}^{n}\log y_i = 0 \tag{3.9}$$

onde $\psi(\bullet)$ é a função digama, (Abramovitz & Stegun, 1972), que é a derivada do logarítmo da função gama. Temos que nesta equação apenas r é desconhecido, sendo que todas as parcelas que não dependem de r são totalmente determinadas pelo ajuste do modelo. Para o cálculo de r, basta utilizar algum método para encontrar zeros de funções, ou então efetuar uma busca por tentativa e erro. Devemos ainda detalhar a função digama para este cálculo , que é expressa aproximadamente, utilizando expansão em série de potências por: (Abramovitz & Stegun, 1972)

$$\psi(r) \cong \log r - \frac{1}{2r} - \frac{1}{12r^2} + \frac{1}{120r^4} - \frac{1}{252r^6} \tag{3.10}$$

Assim, com as equações (3.9) e (3.10), podemos encontrar o valor $\hat{r}$ de r, e determinar os desvios padrão de $\hat{\beta}$ do modelo Gama, suprindo assim essa deficiência do sistema GLIM, versão 3.77.

3.4 Escolha de Modelos: Validação Cruzada

O problema abordado em Dantas & Cordeiro (1988), e revisado aqui, sugere a montagem de modelos de regressão para predição de preços de imóveis, sendo que no contexto do citado artigo, foi utilizado como critério para seleção de modelos, a performance de ajuste do mesmo, e não a performance de predição.

Entretanto, por coerência à finalidade do modelo que é de predição, um critério de seleção baseado em Validação Cruzada (ver Copas, J.B. (1983), Allen, D. M. (1971), Stone, M. 1974) é claramente mais coerente. Segundo este critério, ao invés de se calcular os resíduos da forma usual, ou seja, cada valor observado menos o valor ajustado, o cálculo se faz da forma: $\epsilon_i = y_i - \hat{y}_{(i)}$, onde $\hat{y}_{(i)}$ representa o valor predito da observação i sem a presença desta no ajuste do modelo.

Para o modelo Log-Normal, o valor da Soma de Quadrados Preditiva ("PRE–diction Sum of Squares"), ver Aitkin, Anderson, Francis & Hide (1989), é dado por:

$$SQP_{LogNormal} = \sum_{i=1}^{n} \left[\frac{(y_i - \hat{y}_{(i)})^2}{n} \right] \tag{3.11}$$

E para o modelo Gama, temos:

$$SQP_{Gama} = 2 \sum_{i=1}^{n} \left[-Log(\frac{y_i}{\hat{y}_{(i)}}) + \frac{(y_i - \hat{y}_{(i)})}{\hat{y}_{(i)}} \right] \tag{3.12}$$

Esta medida é a "versão preditiva" da tradicional deviance para o modelo em questão. Com elas podemos comparar diferentes modelos segundo seu grau de predição. O cálculo de (3.12) será feito fora do sistema GLIM.

4 IMPLEMENTAÇÃO PRÁTICA E RESULTADOS

4.1 Introdução

A metodologia introduzida na seção anterior, subseções 3.2,3.3 e 3.4, é aqui implementada utilizando–se dados reais referentes a preços e outras características de lotes urbanos, como

descrito na Seção 2. Do conjunto original de dados, referentes a 50 lotes urbanos, tomamos arbitrariamente uma subamostra de tamanho 44 e dos 6 restantes, 2 deles foram selecionados aleatoriamente para ilustrar a predição. A questão de se trabalhar com uma subamostra é justificada pelo fato de necessitarmos de dados para se comparar com as predições, não podendo assim todos eles entrarem no ajuste.

Como nosso objetivo principal neste artigo é de sugerir uma metodologia de comparação entre modelos, qualquer subamostra atende a esse propósito.

Apresentamos a seguir, utilizando a notação estabelecida na Seção 2, os dados em questão extraídos do citado artigo:

Lt	T	P	N	C	U	R	V	Lt	T	P	N	C	U	R	V
01	19	030	0	31	6	1	02983	23	12	247	1	26	8	3	07093
02	17	029	1	22	7	1	02695	24	17	021	0	55	5	3	29790
03	36	029	0	32	6	1	03831	25	25	037	0	55	5	3	32258
04	10	020	0	26	4	1	02250	26	24	030	0	44	5	1	04047
05	12	036	1	25	6	2	03588	27	14	026	1	29	5	1	03710
06	11	031	1	20	7	2	01261	28	12	040	0	44	5	1	06250
07	10	030	1	14	8	2	01587	29	13	030	0	44	8	1	10970
08	15	020	0	31	4	2	02667	30	14	028	0	43	4	1	05639
09	10	030	0	31	4	2	03333	31	13	041	0	43	7	1	09259
10	13	036	0	31	6	2	04273	32	10	025	0	44	7	1	10737
11	16	034	0	52	7	2	29994	33	09	022	0	43	4	1	04067
12	09	052	1	27	8	3	04288	34	15	032	0	43	5	1	05208
13	24	039	0	31	6	3	07478	35	15	030	0	32	5	1	03333
14	15	082	0	31	8	3	10339	36	15	064	0	32	5	1	02083
15	36	039	1	03	6	3	02422	37	66	029	0	08	5	1	02712
16	16	135	1	32	8	3	06734	38	15	025	0	31	5	1	05333
17	22	022	1	39	7	3	04019	39	14	024	1	23	7	1	03535
18	40	233	1	50	8	3	16094	40	25	080	1	20	7	1	02306
19	14	084	1	42	8	3	09267	41	18	030	1	44	7	1	13789
20	16	033	0	54	8	3	32567	42	10	028	1	43	6	1	10714
21	17	027	1	45	6	3	09918	43	09	030	0	57	8	2	30476
22	17	030	1	29	8	3	07843	44	14	030	0	57	3	2	28571

Temos também, como foi dito, dois lotes que serão utilizados como referência para se efetuar predições, ou seja, iremos utilizar os modelos construídos aqui para predizer o valor desses dois lotes, e comparar com os valores verdadeiros. Os dados dos dois lotes escolhidos ao acaso são:

Lote	T	P	N	C	U	R	V
A	18	23	0	54	5	3	24154
B	14	30	0	30	5	1	3333

A partir dos dados descritos acima, na subseção 4.2 é identificado um particular modelo de regressão Box–Cox, feito o ajuste do mesmo, via GLIM, e utilizado este modelo para predição (pontual e por intervalo) do valor dos dois imóveis citados. Analogamente, em 4.3 é apresentada a implementação do modelo Gama, e em 4.4 é feita uma comparação dos resultados preditivos obtidos com os dois modelos, dando ênfase ao uso do critério de Validação Cruzada.

4.2 Implementação do Modelo Box–Cox

Como mencionado em 3.1, uma das maneiras de abordar o problema da variável resposta ser estritamente positiva, é a utilização de uma transformação nos dados, a fim de trazer os dados transformados para um espaço real $[-\infty;\infty]$, compatível com as pressuposições de um modelo linear normal. Utilizando–se da metodologia apresentada em 3.2.1 e com o auxílio do sistema GLIM, conseguimos, pela variação de λ, determinar o ponto que faz com que a função $L_{max}(\lambda)$ atinja seu extremo, que no caso do sistema GLIM via macro BOXCOX, corresponde a um mínimo.

Reproduzimos abaixo, uma parte dos valores obtidos segundo este critério, onde podemos indentificar o ponto $\hat{\lambda}$, correspondente ao mínimo:

$L_{max}(\lambda)$	λ	
860.1	−0.6	
850.3	−0.4	
845.0	−0.2	(*)
845.3	0.0	(*)
851.6	0.2	
862.8	0.4	
877.7	0.6	

Vemos claramente que os pontos indicados por (*) representam a região onde está oscilando o mínimo da função $L_{max}(\lambda)$, podendo notar que o ponto $\lambda = 0$, com certeza estará bem próximo ao mínimo procurado. Com isto, na prática podemos tomar $\hat{\lambda}$ como sendo zero, indicando assim a transformação logarítmica. Ajustando então o modelo Log–Normal no sistema GLIM, obtemos as seguintes estimativas:

Deviance = 5.8944

	Estimativa	Desvio Padrão	Parâmetro
1	5.3440000	0.382800	(Intercepto)
2	0.0139400	0.006413	(T)
3	−0.0007378	0.001502	(P)
4	−0.1203000	0.162300	(N)
5	0.0573300	0.005674	(C)
6	0.1241000	0.052780	(U)
7	0.2256000	0.079950	(R)

Como estamos trabalhando com a distribuição Normal, a convergência do algorítmo IRLS foi rápida, apenas 1 passo. A deviance dada acima, é uma quantidade que representa o grau de ajuste do modelo aos dados, que neste caso se equivale à tradicional soma de quadrados de resíduos. Contudo, como utilizaremos o modelo para se efetuar predições, devemos calcular uma quantidade que represente a capacidade do modelo para se efetuar predições, e esta quantidade é a SQP, Soma de Quadrados Preditiva, definida em 3.4.

Utilizando a equação (3.11) definida anteriormente, obtemos o seguinte:

$$SQP_{LogNormal} = 10.10$$

Com os dados dos lotes A e B, vamos utilizar o modelo Log–Normal para fazer a predição dos preços dos mesmos (pontual e por intervalo).

Temos que, para o modelo Log–Normal, a variância de um valor predito é dada aproximadamente por,

$$Var(\hat{V}_i) \cong (e^{\hat{V}_i})^2[x_i Var(\hat{\beta}) x_i' + Var(\epsilon)] \tag{4.1}$$

onde $Var(\epsilon)$ é igual a Deviance dividida pelos correspondentes graus de liberdade.

Então, os valores preditos são:

	Estimativa $\hat{V}_i (D.P.)$	Valor Real	Int. Conf.(95%)
A	21398.38(9184.2)	24154	[3029.98; 39766.78]
B	3239.02(1338.4)	3333	[562.22; 5915.82]

Podemos notar que as estimativas pontuais são boas, contudo, o desvio padrão das estimativas são muito altos, resultando em intervalos grandes.

4.3 Implementação do Modelo Gama

Uma outra opção para abordar o problema dos dados estritamente positivos e com distribuição assimétrica, é construir um modelo linear generalizado, supondo uma distribuição Gama para a variável resposta, como introduzido em 3.2.2 e 3.2.3.

Inicialmente, precisamos escolher uma função de ligação adequada para o modelo Gama. Temos duas escolhas usuais para o caso, a recíproca, e a logarítmica. Apesar da função de ligação logarítmica não ser a canônica, no presente caso envolvendo a variável preço, esta última é a mais indicada. Ajustando então o modelo Gama com a função de ligação logarítmica, encontramos as seguintes estimativas:

Deviance = 5.4921

	Estimativa	Desvio Padrão	Parâmetro
1	5.3610000	0.369500	(Intercepto)
2	0.0157400	0.006190	(T)
3	−0.0007126	0.001450	(P)
4	−0.0553600	0.156700	(N)
5	0.0590400	0.005477	(C)
6	0.1104000	0.050950	(U)
7	0.2316000	0.077170	(R)

Neste caso, o algorítmo IRLS convergiu após 4 iterações. Como comentamos em 3.3, os desvios padrão das estimativas fornecidas pelo GLIM para o modelo Gama são bastante imprecisos, porém, podemos efetuar a correção baseada nos desvios de um ajuste Exponencial, como detalhado na mesma subseção. Efetuando os cálculos, a partir do modelo Gama, a equação dada em (3.9), fica da forma $\log r - \psi(r) = 0.062$, onde $\psi(r)$ é função digama. Variando r, encontramos que $\hat{r} = 8.2275$ implicando que os erros padrão corrigidos do modelo Gama, serão iguais aos correspondentes valores do modelo Exponencial divididos pela raiz quadrada de $\hat{r}$.

Ajustando então o modelo Exponencial, temos que os desvios padrão das estimativas são iguais a:

	Desvio Padrão	Parâmetro
1	0.0001231	(Intercepto)
2	$2.844E-06$	(T)
3	$4.803E-07$	(P)
4	0.0000574	(N)
5	$2.089E-06$	(C)
6	$8.212E-06$	(U)
7	0.0000227	(R)

Portanto, os desvios padrão corrigidos para o modelo Gama serão dados por:

$$DP(Gama) = \frac{DP(Exp)}{\hat{r}^{1/2}} = \frac{DP(Exp)}{2.86836} \tag{4.2}$$

Então temos, para o modelo Gama:

	Desvio Padrão	Parâmetro
1	0.0000429	(Intercepto)
2	0.0000010	(T)
3	0.0000002	(P)
4	0.0000200	(N)
5	0.0000007	(C)
6	0.0000029	(U)
7	0.0000079	(R)

Vale notar que neste caso, a aproximação dos erros padrões fornecidos pelo GLIM são extremamente ruins, o que reforça a necessidade da correção.

Para o cálculo da SQP, utilizando a equação dada em (3.12), temos que:

$$PRESS_{Gama} = 6.36$$

Calculando agora os valores preditos, e seus correspondentes desvios, pela fórmula,

$$Var(\hat{V}_i) \cong (e^{\hat{V}_i})^2[x_i Var(\hat{\beta}) x_i' + Var(\epsilon)] \tag{4.3}$$

onde $Var(\epsilon)$ é igual a Deviance dividida pelos correspondentes graus de liberdade, temos que,

	Estimativa $\hat{V}_i(D.P.)$	Valor Real	Int. Conf.(95%)
A	23456.20(9122.5)	24154	[5211.20; 41701.20]
B	3343.53(1293.8)	3333	[755.93; 5931.13]

Novamente, as estimativas pontuais são boas, no entanto, os desvios são altos.

Vale dizer que os intervalos de confiança a 95% para as quantidades preditas, mesmo com as citadas correções por nós introduzidas, ainda assim são grandes e isto não é causado pelos erros padrões das estimativas dos parâmetros, mas sim, pela não linearidade presente em (4.1) e (4.3).

4.4 Comparação

Pelos resultados obtidos no ajuste dos modelos Log–Normal e Gama, podemos verificar o seguinte:

– Se a escolha do modelo fosse baseada na Deviance, muito provavelmente haveria uma certa dúvida em decidir qual modelo a ser utilizado. Em contra partida, se fosse baseada no critério de Validação Cruzada, a escolha seria imediata. Como o modelo será utilizado para predição é bem coerente utilizar a Validação Cruzada e não a Deviance para a comparação. Podemos resumir este resultado no quadro abaixo:

	Deviance	Validação Cruzada
Log–Normal	5.8944	10.10
Gama	5.4921	6.36

Este resultado se confirmou com as duas predições realizadas, onde o modelo com menor SQP, apresentou melhores predições pontuais e menores desvios das mesmas. Portanto, concluímos que a medida baseada na Validação Cruzada é mais indicada quando o modelo for utilizado para se efetuar predições.

5 DISCUSSÃO FINAL

Neste artigo, foi proposto inicialmente, mostrar a utilização de um critério para comparação de modelos úteis em avaliação de imóveis quando o objetivo é de se fazer predições de preços. Com a ilustração através dos dados sobre imóveis, concluímos que o critério de Validação Cruzada é mais indicado que critérios baseados simplesmente em medidas de aderência como é o caso da deviance. Assim, melhores predições podem ser obtidas utilizando-se o modelo Gama, ao passo que se utilizássemos critérios de ajuste não seria possível distinguir o modelo mais adequado.

Outro ponto que abordamos foi a obtenção de estimativas mais confiáveis dos erros padrões das estimativas para o modelo Gama, já que os resultados fornecidos pelo sistema GLIM são bastantes imprecisos.

Vale notar ainda que, embora a idéia de se considerar uma abordagem científica, baseada em dados e modelos estatísticos, para predição do valor de imóveis (apresentada originalmente em Dantas & Cordeiro, 1988 e por nós reformulada) apresente vantagens concretas em relação à abordagem tradicional, subjetiva, dos avaliadores de imóveis, essas duas metodologias, subjetiva e objetiva, podem ser combinadas dentro de um contexto estatístico mais amplo, que é dos modelos Bayesianos. Concretamente, utilizando-se a informação do especialista em avaliação de imóveis de modo a se definir uma distribuição de probabilidade *a priori* para o valor do imóvel (ou o preço médio μ), e mais a informação amostral considerada neste artigo, via modelos de regressão, uma predição Bayesiana pode ser obtida, via teorema de Bayes. No caso do modelo Log-Normal, por exemplo, tal implementação Bayesiana é facilmente realizada via *software* PcBRAP (Zellner, 1987), e para o modelo Gama, isto também é possível, necessitando-se apenas de *software* adequado.

Agradecimentos

Os autores são gratos ao Prof. Dr. Hélio dos Santos Migon - I. M. & CoPPE / UFRJ e aos avaliadores da RBEs por suas críticas e comentários sobre uma versão preliminar deste artigo, e à FAPESP por auxílio financeiro a um dos autores.

BIBLIOGRAFIA

ABRAMOWITZ, M.; STEGUN, I. *Handbook of mathematical functions*, New York, Dover Publ. 1972.

AITKIN, M.; ANDERSON, D.; FRANCIS, B. Statistical modelling in GLIM, Oxford, Clarendon Press, 1989.

ALLEN, D. M. Mean square error of prediction. *Technometrics*, v.13, p.469–79, 1971.

BAKER, R.; NELDER, J. The GLIM system, release 3, generalized linear interactive modelling, Oxford, NAG, 1978.

BOX, G. E. P.; COX, D. R. An analysis of transformations, *Journal of the Royal Statistical Society*, série B, v.26, n.2, p.211–52, 1964.

COPAS, J. B. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society*, série B, v.45, n.3, p.311–54, 1983.

DANTAS, R. A. *Avaliação de glebas inseridas na malha urbana.* Dissertação de Mestrado. Depto Eng. Civil, UFPE, 1987.

———. CORDEIRO, G. Uma nova metodologia para avaliação de imóveis utilizando modelos lineares generalizados. *Revista Brasileira de Estatística*, Rio de Janeiro, 1988.

McCULLAGH, P.; NELDER, J. A. *Generalized linear models.* London, Chapman and Hall, 1983.

STONE, M. Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society*, série B, v.36, n.2, p.111–47, 1974.

ZELLNER, A. *An introduction to bayesian inference in econometrics*, London, John Wiley, 1971.

———. ABOWD, J.; MOULTON, B. *The bayesian regression analysis package: user' s manual.* Chicago, University of Chicago, 1987.

RESUMO

Metodologia científica para predição do valor de venda de imóveis, lotes urbanos, utilizando-se técnicas de regressão – modelos lineares generalizados e modelos Box–Cox – implementados no sistema GLIM (Baker & Nelder, 1978), como originalmente proposta por Dantas & Cordeiro (1988), e exemplificada para o caso de preços de terrenos em Recife, PE, é aqui revisada criticamente, sendo sugeridas algumas alterações. Neste artigo é proposto como critério para escolha entre possíveis modelos concorrentes, entre eles os modelos de regressão Gama e Log–Normal, o uso de Validação Cruzada, e não simplesmente medidas de ajuste, como a Deviance, considerada no citado artigo. Também, os erros padrão das estimativas dos parâmetros do modelo Gama, o qual apresentou melhor performance preditiva, são recalculados separadamente, uma vez que os resultados fornecidos diretamente pelo sistema GLIM, versão 3.77, são bastante imprecisos, com isso, predições por intervalo, omitidas no citado artigo, são agora fornecidas. Por último, é ressaltada a possibilidade de se estender a metodologia proposta de modo a permitir a introdução, caso disponível, de informação subjetiva de profissionais de avaliação no modelo, via uma abordagem Bayesiana.

ABSTRACT

Cientific methodology for prediction of selling prices of urban plots using regression techniques such as generalized linear models and Box–Cox regression models are implemented through the GLIM system (Baker & Nelder, 1978) as originarily proposed by Dantas & Cordeiro (1988); such methodology is critically reviewed in this paper and some modifications proposed. In this article it is proposed to use Cross–Validation as a criteria for model selection and not simply to use goodness of fit measures as the Deviance. Also, the standard errors of the parameters of the gamma model, which presented the better predictive performance, are explicitly calculated here, since the results produced by the GLIM system are not correct; as a consequence, interval prediction, is now presented. Finally, it is emphasized the possibility of extending the proposed methodology, via a Bayesiana approach, in order to introduce subjective information provided by property assessment professionals.

POLÍTICA EDITORIAL

A RBEs objetiva promover e ampliar o uso de métodos estatísticos (quantitativos) na área das ciências econômicas e sociais através da apresentação, descrição e discussão desses métodos e de suas aplicações, num formato de fácil assimilação pelos membros da comunidade científica. Destina-se também a servir de veículo para troca de idéias entre os especialistas e todos os interessados em análise e desenvolvimento de metodologia estatística.

A RBEs tem periodicidade semestral e publica artigos teóricos e/ou aplicados de métodos estatísticos, com ênfase na análise de fenômenos econômicos e sociais. São também aceitos artigos abordando os diversos aspectos do desenvolvimento metodológico relevantes para órgãos produtores de estatísticas, assim como artigos de revisão do estado da arte em temas específicos.

a) delineamento de pesquisas;
b) avaliação de pesquisas e mensuração de erros;
c) uso e combinação de fontes alternativas de informações;
d) novos desenvolvimentos em metodologia de pesquisa;
e) análise de séries de tempo;
f) estudos demográficos;
g) integração de dados;
h) amostragem e estimação;
i) análise de dados;
j) crítica e imputação de dados;
l) disseminação e confiabilidade de dados;
m) modelos econométricos.

Todos os artigos submetidos serão avaliados pelo Comitê Editorial da RBEs quanto a sua qualidade e relevância, devendo os mesmos serem inéditos. Além disto, não deverão ter sido simultaneamente submetidos a qualquer outro periódico nacional.

A RBEs publicará também resenhas de livros, artigos escritos a convites e ensaios sobre o ensino de Estatística.

INSTRUÇÕES PARA SUBMISSÃO DE ARTIGOS À RBEs

Os artigos remetidos para publicação deverão ser submetidos em 3 vias (que não serão devolvidas) para:

Djalma G. C. Pessoa
Editor Responsável – RBEs
ENCE
Rua André Cavalcanti, 106
Bairro de Fátima
20231 – Rio de Janeiro – RJ

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

– Cada autor receberá, gratuitamente, 20 separatas de seu artigo.

Instruções para preparo de originais:

1. 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 filiação e endereço permanente para correspondência. Agradecimentos a colaboradores, outras instituições e auxílios recebidos devem figurar também nesta página.

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

3. O artigo deve ser dividido em seções numeradas progressivamente, com títulos concisos e apropriados. Subseções, todas, inclusive a primeira, devem ser numeradas e receber título apropriado.

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

5. As tabelas e gráficos devem ser apresentados em folhas separadas, precedidas 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 tal como as demais seções do trabalho.

6. Gráficos e diagramas para publicação devem ser traçados em papel branco, com nitidez e boa qualidade, para permitir que a redução seja feita mantendo qualidade.

Fotocópias não serão aceitas. É fundamental que não existam erros quer no desenho quer nas legendas ou títulos.

7. Serão aceitos originais processados por editores de texto tais como CW, Word, Carta Certa, WP e WS.

Impresso no CDDI/Departamento de Editoração e Gráfica — IBGE, em setembro de 1994 — O.S. 06.01.1.0605/94

SE O ASSUNTO É BRASIL
PROCURE O IBGE

IBGE põe à disposição da sociedade milhares de informações de natureza estatística (demográfica, social e econômica), geográfica, cartográfica, geodésica e ambiental, que permitem conhecer a realidade física, humana, social e econômica do País.

VOCÊ PODE OBTER ESSAS PESQUISAS, ESTUDOS E LEVANTAMENTOS EM TODO O PAÍS

No Rio de Janeiro:
Centro de Documentação e Disseminação de Informações - CDDI

Divisão de Atendimento Integrado - DAT
Biblioteca Isaac Kerstenetzky
Livraria Wilson Távora

Rua General Canabarro, 666
20271-201 - Maracanã - Rio de Janeiro - RJ
Tel.: (021)284-0402
Fax: (021)234-6189

Livraria do IBGE
Avenida Franklin Roosevelt, 146 - loja
20021-120 - Castelo - Tel.:(021)220-9147

Nos Estados procure o
Setor de Documentação e Disseminação de Informações - SDDI
da Divisão de Pesquisa

O IBGE possui, ainda, agências localizadas nos principais municípios