Presidente da República
Fernando Henrique Cardoso

Ministro de Estado do Planejamento e Orçamento
Antonio Kandir

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Simon Schwartzman

Diretor de Planejamento e Coordenação
Nuno Duarte da Costa Bittencourt

ÓRGÃOS TÉCNICOS SETORIAIS

Diretoria de Pesquisas
Lenildo Fernandes Silva

Diretoria de Geociências
Trento Natali Filho

Diretoria de Informática
Fernando Elyas Nóbrega Nasser

Centro de Documentação e Disseminação de Informações
David Wu Tai

REVISTA BRASILEIRA DE ESTATÍSTICA

Editor-Responsável
Pedro Luis do Nascimento Silva (IBGE)

Editor de Estatísticas Oficiais
Djalma Galvão Carneiro Pessoa (IBGE)

Editor de Metodologia
Hélio dos Santos Migon (UFRJ)

Editores Associados
Gilberto Alvarenga Paula (USP)
Kaizô Iwakami Beltrão (IBGE)
Lisbeth Kaiserlian Cordani (USP)
Renato Martins Assunção (UFMG)
Wilton de Oliveira Bussab (FGV-SP)

MINISTÉRIO DO PLANEJAMENTO E ORÇAMENTO
INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

ISSN 0034 - 7175

R. bras. Estat., Rio de Janeiro, v.54/57, n. 201/208, p.1-163, jan./dez. 1993/1996

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 três vias (que não serão devolvidas) para:

Pedro Luis Nascimento Silva
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.

Impresso no Centro de Documentação e Disseminação de Informações - CDDI, em meio digital, em 1997

Capa
Pedro Paulo Machado

© IBGE

Revista brasileira de estatística / 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. Divisão de Biblioteca e Acervos Especiais
RJ-IBGE/88-05 Rev.

CDU 31 (05)
PERIÓDICO

Impresso no Brasil/Printend in Brazil

NOTA DO EDITOR

Prezado(a) Leitor(a),

Este volume da Revista Brasileira de Estatística - RBEs - está sendo publicado com referência ao período que vai de 1993 a 1996. Esta medida foi necessária porque a RBEs acumulou um atraso que dificilmente seria eliminado com os mecanismos usuais de produção da revista sem sacrificar a qualidade do material publicado. São várias as razões do atraso, mas as principais foram aguda escassez de artigos e a inexistência de um comitê editorial.

Esses dois problemas estão sendo atacados de frente, visando garantir a continuidade da RBEs, que vem sendo publicada há quase sessenta anos pelo IBGE, bem como propiciar à revista uma nova fase em que seja publicada no prazo e com qualidade. Foi assinado um convênio entre o IBGE e a Associação Brasileira de Estatística - ABE - para publicação conjunta da revista, o qual está sendo implementado a partir de 1997. Já como resultado desse convênio foi nomeado um novo conselho editorial para dirigir a RBEs, o qual ficará responsável pelas decisões sobre política editorial e demais aspectos ligados à publicação da revista. O novo conselho editorial da RBEs, nomeado de comum acordo pelo IBGE e a ABE, é integrado por:

Editor Responsável:
Pedro Luis do Nascimento Silva (IBGE)

Editor de Estatísticas Oficiais:
Djalma Galvão Carneiro Pessoa (IBGE)

Editor de Metodologia:
Hélio dos Santos Migon (UFRJ)

Editores Associados:
Gilberto Alvarenga Paula (USP)
Kaizô Iwakami Beltrão (IBGE)
Lisbeth Kaiserlian Cordani (USP)
Renato Martins Assunção (UFMG)
Wilton de Oliveira Bussab (FGV-SP)

Ao iniciar este novo período da revista é fundamental registrar aqui a contribuição inestimável do Prof. Djalma Galvão Carneiro Pessoa como editor responsável pela revista nesses últimos anos. Sem sua dedicação quase obsessiva, não teríamos uma revista para continuar. Basta também verificar o conteúdo dos volumes publicados desde 1988 para detectar sua contribuição como editor, tendo obtido a colaboração dos principais pesquisadores em Estatística do País como autores e revisores de artigos.

Esta oportunidade serve também para convocar os pesquisadores, praticantes e interessados na Estatística e suas aplicações para que colaborem submetendo seus trabalhos para publicação na RBEs. Brevemente será divulgada a nova política editorial da revista, o que dará maior orientação aos potenciais colaboradores sobre o que publicar nesse veículo.

Contando com sua compreensão e colaboração futura, coloco-me ao seu dispor para o que for necessário. Cordialmente,

Pedro Luis do Nascimento Silva
Editor Responsável pela RBEs

SUMÁRIO

ESTUDOS PARA DEFINIÇÃO DO DESENHO AMOSTRAL DA PESQUISA MENSAL DE COMÉRCIO

Região Metropolitana do Rio de Janeiro

Ana Maria Lima de Farias*
Maria Tereza Serrano Barbosa**

1 INTRODUÇÃO

A Pesquisa Mensal de Comércio - PMC -, em fase de implantação pelo Departamento de Comércio e Serviços - DECSE - do Instituto Brasileiro de Geografia e Estatística - IBGE - , visa ao cálculo, a cada mês, de índices que indiquem o comportamento da receita, dos salários e do pessoal ocupado no comércio varejista.

O âmbito da investigação da PMC, na sua primeira etapa, é o dos estabelecimentos comerciais varejistas localizados na Região Metropolitana do Rio de Janeiro e pretende-se obter os índices para as principais atividades do comércio, estratificadas por porte de acordo com o número de empregados. Assim, para cada uma das dez atividades relacionadas a seguir, além dos índices agregados para a atividade, serão calculados também índices por classe de pessoal ocupado (PO),

* Universidade Federal Fluminense - UFF.
** Universidade do Rio de Janeiro - Uni-Rio.

pensando-se inicialmente nas seguintes classes de PO: [0,10), [10,50) e [50,∞). Levando-se em conta a periodicidade da pesquisa, o tamanho do universo e a necessidade de agilizar, tanto a coleta, quanto a apuração, optou-se por uma pesquisa por amostra probabilística, limitando-se em 1 500 o tamanho máximo da amostra em cada região metropolitana.

O cadastro básico da PMC é constituído pelos estabelecimentos classificados como comerciais na Relação Anual de Informações Sociais de 1992 - RAIS/92. Esse cadastro foi confrontado com o da RAIS/91 e suas classificações detalhadamente criticadas e analisadas, dando origem ao cadastro de informantes da PMC da Região Metropolitana do Rio de Janeiro, com 53 435 sufixos de CGC, distribuídos conforme exibido na Tabela 1. Nesse cadastro, a única variável disponível é o número de empregados, o que faz com que PO seja a variável a ser usada para definir o desenho amostral.

Tabela 1
Distribuição dos estabelecimentos por atividade
PMC-RJ

Atividade		Frequência		Frequência Acumulada	
Código	Nome	Simples	Relativa (%)	Simples	Relativa (%)
01	Super e Hiper Mercados	603	1,13	603	1,13
02	Alimentícios	17.237	32,26	17.840	33,39
03	Lojas de Departamentos	62	0,12	17.902	33,50
04	Farmácias e Drogarias	2.720	5,09	20.622	38,59
05	Vestuário e Têxteis	10.780	20,17	31.402	58,77
06	Outros Pessoais	6.275	11,74	37.677	70,51
07	Móveis e Eletrodomésticos	2.215	4,15	39.892	74,66
08	Automóveis, Motos e Peças	4.985	9,33	44.877	83,98
09	Combustíveis	1.127	2,11	46.004	86,09
10	Material de Construção	7.431	13,91	53.435	100,00

2 DISTRIBUIÇÃO DOS ESTABELECIMENTOS SEGUNDO O PESSOAL OCUPADO

O estudo da distribuição dos estabelecimentos de cada atividade da PMC por classe de PO mostra a homogeneidade existente na maioria das atividades, as principais exceções sendo as atividades 01 (Super e Hiper mercados) e 03 (Lojas de Departamentos). Pode-se observar também a ocorrência de forte assimetria positiva nas distribuições, com a maior parte dos estabelecimentos tendo um pequeno número de empregados e poucos, um grande número de empregados.

Na Tabela A1 do Apêndice 1 temos a distribuição dos estabelecimentos por classe de PO. Analisando essa distribuição, podemos ver que:

- todas as lojas de departamentos têm 50 ou mais empregados;
- com exceção das atividades de supermercados, lojas de departamentos e combustíveis, mais de 70% dos estabelecimentos têm menos de dez empregados e mais de 85% possuem menos de 20 empregados;
- com exceção das atividades de supermercados e lojas de departamentos, mais de 70% dos estabelecimentos possuem menos de 20 empregados; e
- finalmente, ainda com exceção das atividades de supermercados e lojas de departamentos, menos de 2% dos estabelecimentos possuem 100 ou mais empregados.

Essas observações nos levam à conclusão de que a atividade de lojas de departamentos terá que ter um tratamento especial. Com relação às outras nove atividades, a forte assimetria positiva das distribuições faz com que amostras aleatórias simples retiradas dentre os estabelecimentos com 50 ou mais empregados tendam a ter grande variabilidade. Um procedimento aconselhável para reduzir a variabilidade nessas situações

(e, portanto, reduzir o tamanho da amostra) é identificar as observações (estabelecimentos) discrepantes causadoras da assimetria e incluí-las com certeza na amostra, definindo, assim, um estrato certo. A amostra aleatória será retirada, então, do restante da população, que tem distribuição mais homogênea.

3 ESTUDOS PARA DEFINIÇÃO DO DESENHO AMOSTRAL PARTE 1

Com base nos resultados acima, podemos ver que, com exceção da atividade 3, as distribuições dos estabelecimentos segundo PO têm comportamentos bastante semelhantes, o que nos leva à conclusão de que o mesmo desenho amostral deve ser aplicado às nove atividades. Para a atividade 3 estudos à parte serão apresentados. No entanto, as linhas gerais do desenho amostral serão as mesmas para os dois conjuntos de atividades, quais sejam, as atividades serão tratadas como populações independentes. Para cada uma delas serão definidos um estrato certo e um estrato aleatório, de onde será retirada uma amostra aleatória para cada uma das classes de PO. O que vai diferenciar os dois conjuntos de atividades será a forma de definir o estrato certo, que apresentamos a seguir.

3.1 Conjunto das nove atividades

O ponto fundamental para definir o desenho da amostra a ser retirada de cada uma das nove atividades é a determinação do valor de corte para o estrato certo, ou seja, do número de empregados que um estabelecimento precisa ter para pertencer ao estrato certo. Para isso foram adotados dois enfoques, um gerencial e outro estatístico, baseado na metodologia proposta por Hidiroglou (1986).

Definição gerencial do estrato certo

Dentro do enfoque gerencial, todos os estabelecimentos com 100 ou mais empregados foram incluídos no estrato certo. Para a população dos estabelecimentos com menos de 100 empregados, o desenho amostral adotado foi o de amostragem aleatória simples em cada classe de PO, de forma a se ter um coeficiente de variação para o total estimado de pessoas ocupadas em cada uma delas igual a 10% e 15%. No que se segue, o expoente A indica que se está trabalhando com a parte da população não pertencente ao estrato certo. Assim, o tamanho da amostra em cada estrato (atividade x classe de PO) é dado por [Cochran (1977)]

$$n_{hj}^{A} = \frac{(S_{hj}^{A})^2}{CV^2(\bar{Y}_{hj}^{A})^2 + ((S_{hj}^{A})^2 / N_{hj}^{A})} = \frac{(N_{hj}^{A})^2 (S_{hj}^{A})^2}{CV^2 (Y_{hj}^{A})^2 + N_{hj}^{A} (S_{hj}^{A})^2} \qquad (1)$$

onde

n_{hj}^{A} é o tamanho da amostra a ser selecionada para a atividade h e classe de PO j (h=1,...,9 e j=1,2,3)

CV é o coeficiente de variação amostral pré-fixado para o estimador do total Y_{hj} da variável y (pessoal ocupado) para a atividade h e classe de PO j; igual para todos os estratos

N_{hj}^{A} é o tamanho da população na atividade h e classe de PO j

$\bar{Y}_{hj}^{A}$ é a média da variável y para a atividade h na classe de PO j, definida por

$$\bar{Y}_{hj}^{A} = \frac{1}{N_{hj}^{A}} \sum_{i=1}^{N_{hj}^{A}} Y_{hji}$$

Y_{hj}^{A} é o total da variável y para atividade h e classe de PO j definido por

$$Y_{hj}^{A} = \sum_{i=1}^{N_{hj}} Y_{hji}$$

$(S_{hj}^{A})^2$ é a variância da variável y para a atividade h na classe de PO j definida por

$$(S_{hj}^{A})^2 = \frac{1}{N_{hj}^{A} - 1} \sum_{i=1}^{N_{hj}^{A}} (Y_{hji} - \overline{Y}_{hj}^{A})^2$$

Y_{hji} é o valor da variável y para o i-ésimo estabelecimento da atividade h e classe de PO j

Calculado o tamanho da amostra em cada estrato, o tamanho total da amostra é dado por

$$n = \sum_{h=1}^{9} C_h + \sum_{h=1}^{9} \sum_{j=1}^{3} n_{hj}^{A}$$

onde

C_h é o número de estabelecimentos do estrato certo para a atividade h.

Levando em conta os fatos de que o cadastro da PMC não está atualizado e que o tamanho da população em alguns estratos é pequeno, o tamanho mínimo da amostra em cada estrato foi definido como 10, a exemplo do que já é feito na Pesquisa Anual de Comércio - PAC [IBGE (1991)], de modo a contornar o possível problema de não resposta total. Esta escolha se deve aos resultados da distribuição binomial, segundo os quais, utilizando-se uma amostra de tamanho 10 e considerando-se um percentual de não resposta de 30%, a probabilidade de se obter menos de cinco questionários coletados é menor que 5%.

Na Tabela A2 do Apêndice 1 temos os resultados dos estudos feitos para o coeficiente de variação CV igual a 10% e 15%. Trabalhando-se com CV=10%, o tamanho da amostra para essas nove atividades é de 1 346

estabelecimentos, que aumentaria para 1 398 levando em conta o tamanho mínimo de dez estabelecimentos em cada estrato. Com CV=15%, esses valores passam para 913 e 989, respectivamente.

Definição estatística do estrato certo

O método acima, utilizado para definir o estrato certo, trata todas as nove atividades como homogêneas em sua assimetria, o que não é muito razoável, dadas as diferenças existentes nas distribuições (ver Tabela 2). Seria interessante, então, que o estrato certo fosse determinado separadamente para cada uma das nove atividades.

Tabela 2
Distribuição do pessoal ocupado por atividade
PMC-RJ

Atividade		Pessoal Ocupado					
Código	Nome	Total	Total (%)	Médio	Mínimo	Máximo	Coeficiente de Assimetria
01	Super e Hiper Mercados	95.280	19,59	158	0	2.403	4,40
02	Alimentícios	85.587	17,59	5	0	696	20,39
04	Farmácias e Drogarias	28.026	5,76	10	0	640	14,22
05	Vestuário e Têxteis	97.759	20,10	9	0	630	10,77
06	Outros Pessoais	44.123	9,07	7	0	1.507	43,89
07	Móveis e Eletrodomésticos	21.197	4,36	10	0	978	18,81
08	Automóveis, Motos e Peças	43.236	8,89	9	0	357	6,88
09	Combustíveis	19.095	3.93	17	0	342	8,35
10	Material de Construção	52.162	10,72	7	0	541	15,12

Hidiroglou (1986) apresenta um método para dividir uma população em dois estratos, um certo, em que todas as unidades farão parte da amostra, e outro aleatório, de onde uma amostra aleatória simples será retirada. O método consiste em determinar a partição "estrato certo + amostra aleatória" que forneça o menor tamanho da amostra total, fixado o coeficiente de variação desejado para o total estimado. Esse foi o desenho básico adotado na PAC [IBGE (1991)] mas sua aplicação direta na PMC não pode ser feita, uma vez que são necessárias estimativas para cada uma das classes de PO. Mesmo assim, são dados na Tabela 3 os resultados referentes às partições ótimas para cada uma das nove atividades. O algoritmo, programado por Silva (1989), fornece o tamanho ótimo do estrato certo e da amostra aleatória simples a ser retirada do restante da população sem reposição; o valor de corte para o estrato certo é obtido a partir desses resultados. Trabalhando com CV=15%, podemos ver que não existe diferença significativa na definição do estrato certo, a única exceção sendo a atividade de super e hipermercados; o tamanho total da amostra reduz de 913 para 810, o que equivale a uma redução aproximada de 11%.

Tabela 3
Tamanho da amostra e PO para corte por atividade
PMC-RJ
Estrato certo e partição ótima: Hidiroglou (1986)

Atividade		CV=10%				CV=15%			
Código	Nome	PO para Corte	Tamanho da amostra			PO para Corte	Tamanho da amostra		
			Est.Certo	Est.Aleatório	Total		Est.Certo	Est.Aleatório	Total
01	Super e Hiper Mercados	557	24	48	72	761	13	30	43
02	Alimentícios	73	61	222	283	103	28	122	150
04	Farmácias e Drogarias	65	36	109	145	93	19	61	80
05	Vestuário e Têxteis	104	53	154	207	156	21	89	110
06	Outros Pessoais	63	44	123	167	95	21	71	92
07	Móveis e Eletrodomésticos	55	37	75	112	82	21	45	66
08	Automóveis, Motos e Peças	71	117	74	191	99	95	50	145
09	Combustíveis	77	8	44	52	117	2	25	27
10	Material de Construção	68	55	115	170	102	27	70	97
TOTAL			435	964	1399		247	563	810

Alocação proporcional

Para o desenho da PMC, o método de Hidiroglou foi usado para determinar o número de estabelecimentos do estrato certo. Com relação ao estrato aleatório, foram estudados dois desenhos. No primeiro, foi feita uma alocação proporcional nas três classes de PO, usando-se o tamanho da amostra aleatória fornecido pelo método de Hidiroglou. Tal proporcionalidade foi definida em função do número de estabelecimentos (alocação N-proporcional) e em função do total de empregados (alocação Y-proporcional) em cada classe de PO. Para a alocação N-proporcional, o tamanho da amostra em cada classe de PO é dado por

$$n_{hj}^{A} = n_{h}^{A} \frac{N_{hj}^{A}}{N_{h}^{A}}$$

onde

n_{hj}^{A} é o tamanho da amostra aleatória na classe de POj e atividade h

n_{h}^{A} é tamanho total da amostra aleatória para a atividade h, determinado pelo método de Hidiroglou

N_{hj}^{A} é o tamanho da população na classe de PO j e atividade h

N_{h}^{A} é o tamanho da população na atividade h, excluídos os estabelecimentos do estrato certo.

Para a alocação Y-proporcional, o tamanho da amostra em cada classe de PO é dado por

$$n_{hj}^{A} = n_{h}^{A} \frac{Y_{hj}^{A}}{Y_{h}^{A}}$$

onde

Y_{hj}^{A} é total da variável PO na classe de PO j e atividade h

Y_{h}^{A} é o total da variável PO na atividade h, excluídos os estabelecimentos do estrato certo.

Na Tabela A3 do Apêndice 1 são dados os tamanhos das amostras, usando-se o CV igual a 10% e 15%. Cabe ressaltar que as diferenças apresentadas entre os tamanhos totais das amostras para os dois tipos de alocação se devem a problemas de arredondamento. Com relação aos tamanhos das amostras em cada classe de PO, a alocação Y-proporcional, pela sua própria natureza, distribui mais a amostra entre os pequenos, médios e grandes estabelecimentos, o que pode ser mais conveniente, tanto do ponto de vista econômico, quanto do ponto de vista prático, uma vez que o nível de não resposta entre os pequenos estabelecimentos tende a ser maior.

Amostra aleatória simples estratificada

Uma segunda proposta de desenho para a população aleatória foi a de amostragem aleatória estratificada, usando-se o método de Hidiroglou apenas para definir o estrato certo. Assim, para a população não pertencente ao estrato certo, calculou-se o tamanho de uma amostra aleatória simples necessário para se obter um coeficiente de variação igual a 10% ou 15% para o total estimado em cada classe de PO. O tamanho da amostra em cada classe de PO é dado pela fórmula (1). Na Tabela A4 do Apêndice 1 temos os resultados para esse desenho.

Na Tabela 4 temos um resumo comparativo, com relação ao tamanho total da amostra, dos três métodos vistos até o momento. Pode-se ver que a definição gerencial do estrato certo ($PO \geq 100$) superdimensiona a amostra, principalmente em função do estrato certo, onde 272 estabelecimentos são da atividade de super e hipermercados. Com relação aos dois últimos métodos, é natural que o tamanho da amostra para o último seja menor, uma vez que aí foi levada em conta a estratificação da população. Comparando o primeiro e o terceiro métodos, já tinha sido observado que o PO para corte do método de Hidiroglou com

CV=15% era bastante próximo de 100 empregados, com exceção da atividade de super e hipermercados. Assim, não levando em conta essa atividade, os estratos certos para os dois métodos têm tamanhos parecidos: 260 (=532-272) e 234 (=247-13). Com relação à população aleatória, trabalhando-se com CV=15%, não existe diferença substancial entre os tamanhos das amostras: 381 e 404.

Tabela 4
Comparação dos resultados dos 3 desenhos amostrais
PMC-RJ

Método	CV=10%			CV=15%		
	Amostra aleatória	Estrato certo	TOTAL	Amostra aleatória	Estrato certo	TOTAL
Estrato certo: Definição gerencial						
Estrato aleatório: Amostra aleatória estratificada	814	532	1346	381	532	913
Estrato certo: Hidiroglou(1986)						
Estrato aleatório: Alocação proporcional	964	435	1399	563	247	810
Estrato certo: Hidiroglou(1986)						
Estrato aleatório: Amostra aleatória estratificada	830	435	1265	404	247	651

Fonte: tabelas 3,A2,A4

Analisando os resultados específicos para a atividade de super e hipermercados e mercados, podemos observar que, ao trabalharmos com o estrato certo gerencial PO $\geq$ 100, o número de estabelecimentos no estrato certo é muito grande; por outro lado, o PO para corte do estrato certo pelo método de Hidiroglou é muito alto. Dada a importância dessa atividade, foram feitos outros estudos, análogos aos estudos realizados para a atividade de lojas de departamentos, que consistiram em definir gerencialmente diversos cortes para o estrato certo.

3.2 Lojas de Departamentos e Super e Hipermercados

Os estudos realizados para definir o desenho amostral se basearam na idéia de que há necessidade de representação dos grandes estabelecimentos e, assim, é importante definir um estrato certo. Como há uma grande concentração de estabelecimentos na classe de PO de 100 a 500 empregados para essas duas atividades, foram avaliados os seguintes cortes para o estrato certo: PO $\geq$ 300, PO $\geq$ 400 e PO $\geq$ 500. Para o restante da população, manteve-se o desenho de amostragem aleatória estratificada, com CV de 10% e 15%. Como antes, o tamanho da amostra em cada classe de PO é dado pela fórmula (1). Na Tabela A5 do Apêndice 1 temos os resultados pertinentes. Tendo em vista que o tamanho do estrato certo é bastante grande para os dois primeiros casos e também a conveniência de se estabelecer um limite único para todas as Regiões Metropolitanas, é razoável sugerir a adoção do estrato certo definido por PO $\geq$ 500, de forma a evitar um aumento significativo do tamanho desse estrato para Regiões Metropolitanas como São Paulo e Belo Horizonte.

3.3 Análise dos resultados

Em função do tamanho máximo estabelecido para a amostra, podemos ver, a partir dos resultados anteriores, que é impossível adotar coeficiente de variação de 10% para a estimativa do total, considerando-se que o mesmo desenho amostral deve ser aplicado a todas as Regiões Metropolitanas.

Na Tabela 5 temos os tamanhos das amostras segundo os métodos apresentados com CV=15%, levando-se em consideração o tratamento diferenciado para as atividades 1 e 3. O tamanho ajustado é aquele obtido para satisfazer a restrição de que o tamanho mínimo da amostra em cada estrato deve ser 10. Analisando os resultados aí apresentados, podemos ver que não existe diferença entre os métodos de alocação N- ou Y-proporcional, embora o tamanho ajustado para o primeiro seja maior. No entanto, esses dois métodos fornecem tamanhos de amostras bem maiores que os outros.

Tabela 5

Comparação dos resultados dos desenhos amostrais - CV=15%

PMC-RJ

)esenho amostral para conjunto das 8 atividades	Tamanho exato			Tamanho ajustado		
	Amostra aleatória	Estrato certo	TOTAL	Amostra aleatória	Estrato certo	TOTAL
Estrato certo: Definição gerencial PO>=100 Estrato aleatório: Amostra aleatória estratificada	411	294	705	479	294	773
Estrato certo: Hidiroglou(1986) Estrato aleatório: Alocação N-proporcional	598	268	866	677	268	945
Estrato certo: Hidiroglou(1986) Estrato aleatório: Alocação Y-proporcional	599	268	867	629	268	897
Estrato certo: Hidiroglou(1986) Estrato aleatório: Amostra aleatória estratificada	413	268	681	478	268	746

Fonte: tabelas A2,A3,A4,A5

(*) Atividades 1 e 3

Estrato certo: PO>=500

Estrato aleatório: Amostra aleatória estratificada

Embora para a Região Metropolitana do Rio de Janeiro a definição gerencial para o estrato certo tenha fornecido valores de corte próximos àqueles obtidos com o método de Hidiroglou, não podemos garantir que isso se repetirá para as outras regiões; assim, é preferível usar a distribuição da variável para definir esse valor de corte. Sendo assim, dentre os métodos apresentados até o momento, o que melhor se adequa às necessidades da PMC é aquele em que o estrato certo é definido gerencialmente pelo corte $PO \geq 500$ para as atividades 1 e 3 e para as outras atividades, pelo método de Hidiroglou. Em ambos os casos, do restante da população deve ser retirada uma amostra aleatória estratificada de forma a se obter um CV de 15% para a estimativa do total em cada classe de PO. Com essa escolha, o tamanho da amostra para a Região Metropolitana do Rio de Janeiro é de 746 estabelecimentos, sendo 268 pertencentes ao estrato certo e os 478 restantes oriundos de uma amostra aleatória.

4 ESTUDOS PARA DEFINIÇÃO DO DESENHO AMOSTRAL PARTE 2

4.1 O método de Lavallée-Hidiroglou

Nessa segunda parte dos estudos, foi feita a aplicação do método proposto por Lavallée-Hidiroglou (1988) para estratificação de populações assimétricas. Tal método fornece os limites dos estratos, supondo que a população vai ser dividida em L estratos, sendo um estrato certo (*take-all*) e os L-1 restantes, estratos aleatórios (*take-some*), de onde será retirada uma amostra aleatória. O método fornece os limites dos L estratos de modo a minimizar o tamanho total da amostra, dados o coeficiente de variação desejado para a estimativa do total e o esquema de alocação da amostra nos estratos aleatórios.

A solução do problema de minimização do tamanho da amostra é dada a partir de um algoritmo iterativo e no artigo são dados todos os detalhes necessários para a sua programação, que foi desenvolvida por nós em linguagem FORTRAN.

4.2 A aplicação do método de Lavallée-Hidiroglou aos dados da PMC

Um dos parâmetros necessários para a aplicação do algoritmo é o número de estratos. A fim de se ter comparabilidade com os estudos realizados anteriormente, definiu-se como quatro o número de estratos, sendo o último o estrato certo. Vale lembrar que no estudo anterior trabalhamos com os seguintes estratos, a nível de PO:

[0,10) [10,50) [50,*corte*) [*Corte*,∞),

onde *corte* foi definido pelo método de Hidiroglou (1986).

O coeficiente de variação adotado foi de 10% e o esquema de alocação da amostra nos estratos aleatórios foi o de alocação p-proporcional segundo a variável PO [Bankier (1988)] com p=1, o que equivale à alocação proporcional ao número de empregados.

A partir de um conjunto inicial de limites dos estratos, o algoritmo calcula, a cada iteração, novos limites. O critério de convergência adotado estabeleceu em 10^{-6} o valor da diferença máxima entre os limites obtidos em duas iterações consecutivas. No Apêndice 1 apresentamos os resultados pertinentes para as iterações inicial e final, quais sejam: limites dos estratos, número de observações, PO total, peso, média e variância da variável PO em cada estrato. O tamanho total da amostra é dado por

$$n = NW_L + \frac{N[\sum_{h=1}^{L-1}(W_h\sigma_h)^2(W_h\mu_h)^{-p}][\sum_{h=1}^{L-1}(W_h\mu_h)^p]}{Nc^2\mu^2 + \sum W_h\sigma_h^2}$$ onde

N é o tamanho total da população

W_h é o peso do estrato h (h=1,...,L) definido pela razão

$$W_h = \frac{N_h}{N}$$

N_h é o tamanho da população no estrato h

σ_h^2 é a variância da variável PO no estrato h (nos cálculos, estimada pela variância amostral)

μ_h é a média da variável PO no estrato h (estimada pela média amostral)

c é o coeficiente de variação desejado

μ é a média da variável PO na população

É interessante observar que o algoritmo objetiva obter os limites dos estratos de modo a minimizar o tamanho n; como os termos nos somatórios dependem dos limites que se quer obter, a solução é obtida através de um algoritmo iterativo.

Com a alocação p-proporcional adotada, o tamanho da amostra em cada estrato é dado por

$$n_h = \frac{(W_h\mu_h)^p}{\sum_{h=1}^{L-1}(W_h\mu_h)^p}.(n - NW_L) \qquad h = 1,\ldots,L-1$$

Note-se que, com p=1, esse esquema de alocação equivale à alocação proporcional ao PO de cada estrato.

É importante salientar aqui alguns aspectos da aplicação do algoritmo. Como em todo algoritmo iterativo, a escolha dos valores iniciais

(no caso, limites iniciais dos estratos) é importante para garantir, não só a convergência, como também a velocidade de convergência. Para algumas atividades, a escolha inadequada dos limites iniciais provocou problemas de ordem numérica, que foram corrigidos com nova escolha dos valores iniciais.

Um outro ponto importante diz respeito à precisão das estimativas. Como podemos ver dos resultados, o tamanho da amostra na maioria dos estratos é bastante pequeno e estamos trabalhando com CV=10%. Uma tentativa de se trabalhar com CV=15% foi feita, visando novamente a comparação com os resultados anteriores, mas resultou em amostras tão pequenas que não reproduzimos aqui os resultados. Vale a pena ressaltar que os autores, em seu artigo, trabalham com CV igual a 1%, 5% ou 10%.

Com relação à atividade 8 (Automóveis, Motos e Peças), que é a que apresenta o segundo menor coeficiente de assimetria, foi detectado o seguinte problema: trabalhando-se com CV=10%, o algoritmo resultou no valor de corte para o estrato certo superior ao valor máximo de PO para aquela atividade, o que forneceu o tamanho do estrato certo como 0. A solução adotada para contornar esse problema foi aumentar a precisão da estimativa, ou seja, diminuir o CV para 5% (outra possibilidade seria diminuir o número de estratos).

Na Tabela 6, onde apresentamos os limites dos estratos, podemos ver que as atividades 1 (Super e Hipermercados) e 3 (Lojas de Departamentos) continuam apresentando um comportamento diferente das demais. Para as outras atividades, os limites superiores para a classe de médios estabelecimentos, preestabelecido como 50 no estudo anterior, ficam bem abaixo de 50, confirmando o resultado intuído na equipe do DECSE de que, no comércio varejista, não é razoável trabalhar com a classe de PO [10,50) para definir médios estabelecimentos.

Tabela 6
Limites superiores e tamanho da população e da amostra por estrato
Método de estratificação: Lavallée-Hidiroglou(1988) - CV=10%
PMC-RJ

Atividade	Estrato 1			Estrato 2			Estrato 3			Estrato	Total	
	Lim.Sup.	População	Amostra	Lim.Sup	População	Amostra	Lim.Sup	População	Amostra	Certo	População	Amostra
Super e Hiper Mercados	79	273	1	260	224	4	1.107	98	5	8	603	18
Alimentícios	5	12.959	11	26	3.782	23	246	492	13	4	17.237	51
Lojas de departamentos	130	20	0	297	25	1	713	14	1	3	62	5
Farmácias e Drogarias	7	1.632	4	27	836	9	164	248	10	4	2.720	27
Vestuário e Têxteis	7	6.677	6	34	3.627	18	312	473	10	3	10.780	37
Outros Pessoais	6	4.451	8	27	1.596	14	198	221	8	7	6.275	37
Móveis e Eletrodomésticos	6	1.449	4	25	602	8	146	153	7	11	2.215	30
Automóveis,Motos,Peças(*)	4	2.672	9	26	2.043	47	153	218	37	52	4.985	145
Combustíveis	10	307	1	27	656	7	131	163	4	1	1.127	13
Material de Construção	6	4.636	6	28	2.583	19	232	205	8	7	7.431	40
TOTAL GERAL		35.076	50		15.974	150		2.285	103	100	53.435	403

(*) - CV=5%

Analisando a Tabela 6, onde são dados também os tamanhos da população e da amostra em cada estrato, podemos observar que o tamanho total da amostra é bastante pequeno, principalmente quando comparamos com os resultados do estudo anterior. Naquela ocasião, trabalhando com CV=10%, o "melhor" método fornecia o tamanho total exato da amostra como 1 265, sem levar em conta a atividade de Lojas de Departamentos. Com o método agora apresentado, o tamanho da amostra, também sem considerar essa atividade, é de 398. Da diferença de 867 estabelecimentos, 387 aparecem no estrato certo, ou seja, com o método anterior, o corte para o estrato certo era bem menor, provocando um aumento considerável no tamanho da amostra. Na Tabela 7 apresentamos o resumo dos dois métodos, de modo a facilitar a comparação. Os resultados para o primeiro método foram extraídos da Tabela A4 e não consideramos aqui as atividades 3 e 8 por causa do tratamento especial que cada uma deve receber (atividade 3: estrato certo definido gerencialmente; atividade 8: limites dos estratos no segundo método definidos com CV=5%).

Tabela 7
Comparação dos 2 métodos de amostragem
PMC-RJ

Atividade	Estrato 1						Estrato 2						Estrato 3						Estrato Certo				Total		
	Método 1 (1)			Método 2 (2)			Método 1			Método 2			Método 1			Método 2			Método 1		Método 2			Mét.1	Mét.2
	Lim. Inf.	N	n	Lim. Inf.	N	n	Lim. Inf.	N	n	Lim. Inf.	N	n	Lim. Inf.	N	n	Lim. Inf.	N	n	Lim. Inf.	N	Lim. Inf.	N	N	n	n
01	0	22	19	0	273	1	10	173	11	79	224	4	50	384	41	260	98	5	557	24	1.107	8	603	95	18
02	0	14.867	126	0	12.959	11	10	2.239	23	5	3.782	23	50	70	1	26	492	13	73	61	246	4	17.237	211	51
04	0	1.941	51	0	1.632	4	10	702	25	7	836	9	50	41	1	27	248	10	65	36	164	4	2.720	113	27
05	0	7.882	64	0	6.677	6	10	2.690	24	7	3.627	18	50	155	5	34	473	10	104	53	312	3	10.780	146	37
06	0	5.104	71	0	4.451	8	10	1.096	23	6	1.596	14	50	31	1	27	221	8	63	44	198	7	6.275	139	37
07	0	1.693	60	0	1.449	4	10	481	23	6	602	8	50	4	1	25	153	7	55	37	146	11	2.215	121	30
09	0	307	34	0	307	1	10	791	20	10	656	7	50	21	2	27	163	4	77	8	131	1	1.127	64	13
10	0	5.942	72	0	4.636	6	10	1.404	22	6	2.583	19	50	30	1	28	205	8	68	55	232	7	7.431	150	40
Total		37.758	497		32.384	41		9.576	171		13.906	102		736	53		2.053	65		318		45	48.388	1.039	253

(1) - Hidiroglou(1986) + Amostragem aleatória estratificada - CV=10%

(2) - Lavallée-Hidiroglou(1988) - CV=10%

5 O DESENHO DA AMOSTRA DA PMC

Diante da necessidade de se apresentar indicadores também por classe de PO e não só por atividade, a aplicação direta do método acima descrito não é viável, uma vez que os limites dos estratos são diferentes para cada atividade.

Por outro lado, os resultados obtidos nos mostram claramente que a proposta inicial para o desenho da amostra não é razoável. Assim, o desenho final da amostra da PMC será um compromisso entre os dois estudos, conforme explicitado a seguir.

Os estratos por classe de PO serão redefinidos como:

[0,10) [10,20) [20,50) [50,∞),

sendo que no último estrato será incluído o estrato certo.

O corte para o estrato certo será aquele fornecido pelo método de Lavallée-Hidiroglou (Tabela 6) e nos estratos aleatórios será retirada uma amostra aleatória simples sem reposição, trabalhando-se com CV=10%. O tamanho da amostra em cada estrato é dado pela fórmula (1).

Para a atividade 8 (Automóveis, Motos e Peças), o método de Lavallée-Hidiroglou será aplicado com CV=5%. Com relação às atividades 1 (Super e Hipermercados) e 3 (Lojas de Departamentos), continuaremos a dar a elas um tratamento especial, dada a importância econômica de ambas. Para essas atividades, o estrato certo será definido gerencialmente, com o valor de corte igual a 500 empregados (Tabela A5 - Apêndice). Como antes, o tamanho mínimo da amostra em cada estrato será definido como 10. Na Tabela 8 apresentamos os resultados finais para a amostra da PMC para a Região Metropolitana do Rio de Janeiro.

Tabela 8
Amostra da Pesquisa Mensal de Comércio
Região Metropolitana do Rio de Janeiro

Atividade	[0,10)		[10,20)		[20,50)		[50,corte)		Estrato Certo		TOTAL	
	População	Amostra	População	Amostra	População	Amostra	População	Amostra	População	Corte	População	Amostra
Super e Hiper Mercados	22	19	31	10	142	10	380	39	28	500	603	106
Alimentícios	14.867	126	1.579	10	660	10	127	16	4	246	17.237	166
Lojas de Departamentos	0	0	0	0	0	0	56	23	6	500	62	29
Farmácias e Drogarias	1.941	51	397	10	305	10	73	10	4	164	2.720	85
Vestuário e Têxteis	7.882	64	1.732	10	958	10	205	26	3	313	10.780	113
Outros Pessoais	5.104	71	771	10	325	10	68	12	7	199	6.275	110
Móveis e Eletrodomésticos	1.693	60	294	10	187	10	30	10	11	146	2.215	101
Automóveis, Motos, Peças	4.156	85	463	10	227	10	87	10	52	153	4.985	167
Combustíveis	307	34	502	10	289	10	23	10	6	80	1.127	70
Material de Construção	5.942	72	1.040	10	364	10	61	10	24	110	7.431	126
TOTAL GERAL	41.914	582	6.809	90	3.457	90	1.110	166	145		53.435	1.073

APÊNDICE 1
Tabelas Auxiliares

Tabela A1
Distribuição dos estabelecimentos por atividade e classe de PO
PMC-RJ

Atividade e Classe de PO	Frequência		Frequência Acumulada	
	Simples	Relativa (%)	Simples	Relativa (%)
01 - Super e Hiper Mercados				
[0,10)	22	3,65	22	3,65
[10,20)	31	5,14	53	8,79
[20,50)	142	23,55	195	32,34
[50,100)	136	22,55	331	54,89
[100,500)	244	40,46	575	95,36
500 ou +	28	4,64	603	100,00
02 - Alimentícios				
[0,10)	14.867	86,25	14.867	86,25
[10,20)	1.579	9,16	16.446	95,41
[20,50)	660	3,83	17.106	99,24
[50,100)	102	0,59	17.208	99,83
[100,500)	27	0,16	17.235	99,99
500 ou +	2	0,01	17.237	100,00
03 - Lojas de Departamentos				
[0,10)	0	0,00	0	0,00
[10,20)	0	0,00	0	0,00
[20,50)	0	0,00	0	0,00
[50,100)	15	24,19	15	24,19
[100,500)	41	66,13	56	90,32
500 ou +	6	9,68	62	100,00
04 - Farmácias e Drogarias				
[0,10)	1.941	71,36	1.941	71,36
[10,20)	397	14,60	2.338	85,96
[20,50)	305	11,21	2.643	97,17
[50,100)	63	2,32	2.706	99,49
[100,500)	13	0,48	2.719	99,96
500 ou +	1	0,04	2.720	100,00
05 - Vestuário e Têxteis				
[0,10)	7.882	73,12	7.882	73,12
[10,20)	1.732	16,07	9.614	89,18
[20,50)	958	8,89	10.572	98,07
[50,100)	151	1,40	10.723	99,47
[100,500)	56	0,52	10.779	99,99
500 ou +	1	0,01	10.780	100,00
06 - Outros Pessoais				
[0,10)	5.104	81,34	5.104	81,34
[10,20)	771	12,29	5.875	93,63
[20,50)	325	5,18	6.200	98,80
[50,100)	57	0,91	6.257	99,71
[100,500)	17	0,27	6.274	99,98
500 ou +	1	0,02	6.275	100,00

Continua...

Tabela A1 (conclusão)

Atividade e Classe de PO	Frequência		Frequência Acumulada	
	Simples	Relativa (%)	Simples	Relativa (%)
07 - Móveis e Eletrodomésticos				
[0,10)	1.693	76,43	1.693	76,43
[10,20)	294	13,27	1.987	89,71
[20,50)	187	8,44	2.174	98,15
[50,100)	23	1,04	2.197	99,19
[100,500)	16	0,72	2.213	99,91
500 ou +	2	0,09	2.215	100,00
08 - Automóveis, Motos e Peças				
[0,10)	4.156	83,37	4.156	83,37
[10,20)	463	9,29	4.619	92,66
[20,50)	227	4,55	4.846	97,21
[50,100)	45	0,90	4.891	98,11
[100,500)	94	1,89	4.985	100,00
500 ou +	0	0,00	4.985	100,00
09 - Combustíveis (Postos)				
[0,10)	307	27,24	307	27,24
[10,20)	502	44,54	809	71,78
[20,50)	289	25,64	1.098	97,43
[50,100)	26	2,31	1.124	99,73
[100,500)	3	0,27	1.127	100,00
500 ou +	0	0,00	1.127	100,00
10 - Material de Construção				
[0,10)	5.942	79,96	5.942	79,96
[10,20)	1.040	14,00	6.982	93,96
[20,50)	364	4,90	7.346	98,86
[50,100)	58	0,78	7.404	99,64
[100,500)	26	0,35	7.430	99,99
500 ou +	1	0,01	7.431	100,00

Tabela A2
Estatísticas da população e tamanhos das amostras por atividade e classe de PO
PMC-RJ
Estrato certo: PO >= 100 - Estrato aleatório: amostra aleatória estratificada

Atividade e Classe de PO	População			Tamanho da Amostra	
	Nº. obs.	PO		CV=10%	CV=15%
		Média	Variância		
01 -Super e Hiper Mercados					
[0,10)	22	3,23	11,42	19	16
[10,50)	173	29,06	98,52	11	6
[50,100)	136	73,84	238,94	5	2
100 ou +	272	294,63	73.005,21	272	272
Total	603			307	296
02 - Alimentícios					
[0,10)	14.867	2,24	6,37	126	57
[10,50)	2.239	17,87	74,19	23	11
[50,100)	102	68,09	181,23	4	2
100 ou +	29	182,34	17.765,52	29	29
Total	17.237			182	99
04 - Farmácias e Drogarias					
[0,10)	1.941	3,64	6,92	51	23
[10,50)	702	20,75	109,38	25	12
[50,100)	63	64,00	184,84	5	2
100 ou +	14	169,21	19.050,80	14	14
Total	2.720			95	51
05 - Vestuário e Têxteis					
[0,10)	7.882	3,37	7,22	64	29
[10,50)	2.690	19,14	85,26	24	11
[50,100)	151	67,58	204,61	5	2
100 ou +	57	166,93	8.407,17	57	57
Total	10.780			150	99
06 - Outros Pessoais					
[0,10)	5.104	3,14	7,05	71	32
[10,50)	1.096	17,96	72,67	23	10
[50,100)	57	65,37	182,56	4	2
100 ou +	18	260,56	105.095,91	18	18
Total	6.275			116	62

Continua...

Tabela A2 (Conclusão)

Atividade e Classe de PO	População			Tamanho da Amostra	
	Nº. obs.	PO		CV=10%	CV=15%
		Média	Variância		
07 - Móveis e Eletrodomésticos					
[0,10)	1.693	3,37	7,02	60	28
[10,50)	481	19,51	87,96	23	11
[50,100)	23	64,48	137,72	3	2
100 ou +	18	257,33	49.641,65	18	18
Total	2.215			104	59
08 - Automóveis, Motos e Peças					
[0,10)	4.156	2,79	6,68	85	38
[10,50)	690	18,24	78,10	23	11
[50,100)	45	72,22	253,95	5	3
100 ou +	94	168,13	2.875,23	94	94
Total	4.985			207	146
09 - Combustíveis (postos)					
[0,10)	307	5,11	9,70	34	16
[10,50)	791	19,30	73,56	20	9
[50,100)	26	65,31	140,14	3	2
100 ou +	3	187,00	18.075,00	3	3
Total	1.127			60	30
10 - Material de Construção					
[0,10)	5.942	3,21	7,44	72	32
[10,50)	1.404	17,05	63,46	22	10
[50,100)	58	68,02	169,49	4	2
100 ou +	27	192,70	11.893,52	27	27
Total	7.431			125	71
TOTAL GERAL	53.373			1.346	913

Tabela A3
Estatísticas da população e tamanhos das amostras por atividade e classe de PO
PMC-RJ
Estrato certo: Hidiroglou(1986) - Estrato aleatório: Alocação proporcional

	CV = 10%				CV = 15%			
	População		Tamanho da amostra		População		Tamanho da amostra	
	Nº Estab.	PO	N-proporcional	Y-proporcional	Nº Estab.	PO	N-proporcional	Y-proporcional
01 - Super e Hiper Mercados	**PO para corte: 557**				**PO para corte: 761**			
[0,10)	22	71	2	1	22	71	2	1
[10,50) ou [10,corte)	173	5.027	15	4	173	5.027	9	2
[50,corte)	384	67.052	32	45	395	73.936	21	29
Corte ou +	24	23.130	24	24	13	16.246	13	13
Total	603	95.280	73	74	603	95.280	45	45
02 - Alimentícios	**PO para corte: 73**				**PO para corte: 103**			
[0,10)	14.867	33.341	193	96	14.867	33.341	106	51
[10,50) ou [10,corte)	2.239	40.013	29	115	2.239	40.013	16	61
[50,corte)	70	4.211	1	13	103	7.047	1	11
Corte ou +	61	8.022	61	61	28	5.186	28	28
Total	17.237	85.587	284	285	17.237	85.587	151	151
04 - Farmácias e Drogarias	**PO para corte: 65**				**PO para corte: 93**			
[0,10)	1.941	7.061	79	33	1.941	7.061	44	18
[10,50) ou [10,corte)	702	14.564	29	67	702	14.564	16	36
[50,corte)	41	2.277	2	11	58	3.557	2	9
Corte ou +	36	4.124	36	36	19	2.844	19	19
Total	2.720	28.026	146	147	2.720	28.026	81	82
05 - Vestuário e Têxteis	**PO para corte: 104**				**PO para corte: 156**			
[0,10)	7.882	26.554	114	47	7.882	26.554	66	26
[10,50) ou [10,corte)	2.690	51.486	39	90	2.690	51.486	23	50
[50,corte)	155	10.614	3	19	187	14.381	2	14
Corte ou +	53	9.105	53	53	21	5.338	21	21
Total	10.780	97.759	209	209	10.780	97.759	112	111
06 - Outros Pessoais	**PO para corte: 63**				**PO para corte: 95**			
[0,10)	5.104	16.022	101	53	5.104	16.022	58	30
[10,50) ou [10,corte)	1.096	19.685	22	65	1.096	19.685	13	36
[50,corte)	31	1.712	1	6	54	3.433	1	7
Corte ou +	44	6.704	44	44	21	4.983	21	21
Total	6.275	44.123	168	168	6.275	44.123	93	94

Continua...

Tabela A3 (conclusão)

	CV = 10%				CV = 15%			
	População		Tamanho da amostra		População		Tamanho da amostra	
	Nº Estab.	PO	N-proporcional	Y-proporcional	Nº Estab.	PO	N-proporcional	Y-proporcional
07 - Móveis e Eletrodomésticos	**PO para corte: 66**				**PO para corte: 82**			
[0,10)	1.693	5.699	59	28	1.693	5.699	35	16
[10,50) ou [10,corte)	481	9.383	17	47	481	9.383	10	26
[50,corte)	4	202	1	1	20	1.225	1	4
Corte ou +	37	5.913	37	37	21	4.890	21	21
Total	2.215	21.197	114	113	2.215	21.197	67	67
08 - Automóveis,Motos e Peças	**PO para corte: 71**				**PO para corte: 99**			
[0,10)	4.156	11.595	64	34	4.156	11.595	43	22
[10,50) ou [10,corte)	690	12.587	11	37	690	12.587	8	24
[50,corte)	22	1.275	1	4	44	3.151	1	6
Corte ou +	117	17.779	117	117	95	15.903	95	95
Total	4.985	43.236	193	192	4.985	43.236	147	147
09 - Combustíveis (Postos)	**PO para corte: 77**				**PO para corte: 117**			
[0,10)	307	1.568	13	4	307	1.568	7	3
[10,50) ou [10,corte)	791	15.268	32	38	791	15.268	18	21
[50,corte)	21	1.272	1	4	27	1.800	1	3
Corte ou +	8	987	8	8	2	459	2	2
Total	1.127	19.095	54	54	1.127	19.095	28	29
10 - Material de Construção	**PO para corte: 68**				**PO para corte: 102**			
[0,10)	5.942	19.077	93	50	5.942	19.077	57	29
[10,50) ou [10,corte)	1.404	23.937	22	62	1.404	23.937	14	36
[50,corte)	30	1.721	1	5	58	3.945	1	6
Corte ou +	55	7.427	55	55	27	5.203	27	27
Total	7.431	52.162	171	172	7.431	52.162	99	98
TOTAL GERAL	53.373	486.465	1.412	1.414	53.373	486.465	823	824

Tabela A4
Estatísticas da população e tamanhos das amostras por atividade e classe de PO
PMC-RJ
Estrato certo: Hidiroglou(1986) - Estrato aleatório: Amostra Aleatória Estratificada

Atividade e Classe de PO	CV = 10%				CV = 15%			
	População			Tamanho da amostra	População			Tamanho da amostra
	Nº Obs.	PO			Nº Obs.	PO		
		Média	Variância			Média	Variância	
01 - Super e Hiper Mercados	**PO para corte: 557**				**PO para corte: 761**			
[0,10)	22	3,23	11,42	19	22	3,23	11,42	16
[10,50) ou [10,corte)	173	29,06	98,52	11	173	29,06	98,52	6
[50,corte)	384	174,61	13.742,93	41	395	187,18	18.948,84	23
Corte ou +	24	963,75	213.332,80	24	13	1.249,69	213.527,06	13
Total	603			95	603			58
02 - Alimentícios	**PO para corte: 73**				**PO para corte: 103**			
[0,10)	14.867	2,24	6,37	126	14.867	2,24	6,37	57
[10,50) ou [10,corte)	2.239	17,87	74,19	23	2.239	17,87	74,19	11
[50,corte)	70	60,16	32,86	1	103	68,42	190,62	2
Corte ou +	61	131,51	10.705,05	61	28	185,21	18.175,88	28
Total	17.237			211	17.237			98
04 - Farmácias e Drogarias	**PO para corte: 65**				**PO para corte: 93**			
[0,10)	1.941	3,64	6,92	51	1.941	3,64	6,92	23
[10,50) ou [10,corte)	702	20,75	109,38	25	702	20,75	109,38	12
[50,corte)	41	55,54	18,90	1	58	61,33	109,07	2
Corte ou +	36	114,56	9.097,05	36	19	149,68	14.887,56	19
Total	2.720			113	2.720			56
05 - Vestuário e Têxteis	**PO para corte: 104**				**PO para corte: 156**			
[0,10)	7.882	3,37	7,22	64	7.882	3,37	7,22	29
[10,50) ou [10,corte)	2.690	19,14	85,26	24	2.690	19,14	85,26	11
[50,corte)	155	68,48	230,17	5	187	76,90	564,53	5
Corte ou +	53	171,79	8.710,40	53	21	254,19	10.577,66	21
Total	10.780			146	10.780			66
06 - Outros Pessoais	**PO para corte: 63**				**PO para corte: 95**			
[0,10)	5.104	3,14	7,05	71	5.104	3,14	7,05	32
[10,50) ou [10,corte)	1.096	17,96	72,67	23	1.096	17,96	72,67	10
[50,corte)	31	55,23	15,05	1	54	63,57	130,36	2
Corte ou +	44	152,36	49.906,47	44	21	237,29	92.743,41	21
Total	6.275			139	6.275			65

Continua...

Tabela A4 (conclusão)

Atividade e Classe de PO	CV = 10%				CV = 15%			
	População			Tamanho da amostra	População			Tamanho da amostra
	Nº Obs.	PO			Nº Obs.	PO		
		Média	Variância			Média	Variância	
07 - Móveis e Eletrodomésticos	**PO para corte: 55**				**PO para corte: 82**			
[0,10)	1.693	3,37	7,02	60	1.693	3,37	7,02	28
[10,50) ou [10,corte)	481	19,51	87,96	23	481	19,51	87,96	11
[50,corte)	4	50,50	1,00	1	20	61,25	70,30	1
Corte ou +	37	159,81	32.760,05	37	21	232,86	45.974,43	21
Total	2.215			121	2.215			61
08 - Automóveis,Motos e Peças	**PO para corte: 71**				**PO para corte: 99**			
[0,10)	4.156	2,79	6,68	85	4.156	2,79	6,68	38
[10,50) ou [10,corte)	690	18,24	78,10	23	690	18,24	78,10	11
[50,corte)	22	57,95	25,95	1	44	71,61	242,80	3
Corte ou +	117	151,96	3.399,11	117	95	167,40	2.894,94	95
Total	4.985			226	4.985			147
09 - Combustíveis (Postos)	**PO para corte: 77**				**PO para corte: 117**			
[0,10)	307	5,11	9,70	34	307	5,11	9,70	16
[10,50) ou [10,corte)	791	19,30	73,56	20	791	19,30	73,56	9
[50,corte)	21	60,57	40,26	2	27	66,67	184,62	2
Corte ou +	8	123,38	7.975,70	8	2	229,50	25.312,50	2
Total	1.127			64	1.127			29
10 - Material de Construção	**PO para corte: 68**				**PO para corte: 102**			
[0,10)	5.942	3,21	7,44	72	5.942	3,21	7,44	32
[10,50) ou [10,corte)	1.404	17,05	63,46	22	1.404	17,05	63,46	10
[50,corte)	30	57,37	19,21	1	58	68,02	169,49	2
Corte ou +	55	135,04	9.030,70	55	27	192,70	11.893,52	27
Total	7.431			150	7.431			71
TOTAL GERAL	53.373			1.265	53.373			651

Tabela A5
Estatísticas da população e tamanhos das amostras por classe de PO
Atividades: Super e Hiper Mercados e Lojas de Departamentos
Estrato Certo: Definição gerencial
Estrato Aleatório: Amostra Aleatória Estratificada

Atividade e Classe de PO	População			Tamanho da amostra	
	Nº Obs.	PO			
		Média	Variância	CV=10%	CV=15%
Estrato certo: PO >= 300					
01 - Super e Hiper Mercados					
[0,10)	22	3,23	11,42	19	16
[10,50)	173	29,06	98,52	11	6
[50,300)	322	131,82	4.250,85	23	11
300 ou +	86	555,07	126.040,37	86	86
Total	603			139	119
03 - Lojas de Departamentos					
[50,300)	45	147,18	5.090,10	16	9
300 ou +	17	520,12	57.160,11	17	17
Total	62			33	26
TOTAL GERAL	665			172	145
Estrato certo: PO >= 400					
01 - Super e Hiper Mercados					
[0,10)	22	3,23	11,42	19	16
[10,50)	173	29,06	98,52	11	6
[50,400)	355	151,73	7.789,76	31	15
400 ou +	53	685,23	160.653,49	53	53
Total	603			114	90
03 - Lojas de Departamentos					
[50,400)	50	166,88	8.177,86	19	11
400 ou +	12	593,42	63.025,72	12	12
Total	62			31	23
TOTAL GERAL	665			145	113
Estrato certo: PO >= 500					
01 - Super e Hiper Mercados					
[0,10)	22	3,23	11,42	19	16
[10,50)	173	29,06	98,52	11	6
[50,500)	380	171,00	12.627,14	39	19
500 ou +	28	900,11	206.950,62	28	28
Total	603			97	69
03 - Lojas de Departamentos					
[50,500)	56	195,38	14.235,15	23	13
500 ou +	6	754,00	76.107,60	6	6
Total	62			29	19
TOTAL GERAL	665			126	88

APÊNDICE 2

Resultados do algoritmo iterativo para definição dos limites dos estratos ótimos

ATIVIDADE 1: Super e Hiper Mercados				
Número de observações:	603			
Coeficiente de variação:	10%			
Y Total	95.280			
Y Médio	158.009950			
Iteração 0				
Limites superiores dos estratos:	10	20	100	2.404
Número de obs. nos estratos:	22	31	278	272
Total de Y nos estratos:	71	488	14.581	80.140
Peso dos estratos aleatórios:	0.036484	0,051410	0,461028	
Média de Y nos estratos aleatórios:	3,2273	15,7419	52,4496	
Variância de Y nos estratos aleatórios:	11,4221	7,3312	592,3783	
Iteração Final (16)				
Limites superiores dos estratos:	78,6275	259,7373	1.107,0885	2.404,0000
Número de obs. nos estratos:	273	224	98	8
Total de Y nos estratos:	9.951	32.053	41.348	11.928
Peso dos estratos aleatórios:	0,452736	0,371476	0,162521	
Média de Y nos estratos aleatórios:	36,4505	143,0937	421,9184	
Variância de Y nos estratos aleatórios:	397,1529	2.368,4970	22.914,6874	
Tamanho total da amostra:	19			
Tamanho da amostra aleatória:	11			
Tamanho do estrato certo:	8			
ATIVIDADE 2: Alimentícios				
Número de observações:	17.237			
Coeficiente de variação:	10%			
Y Total	85.587			
Y Médio	4,965307			
Iteração 0				
Limites superiores dos estratos:	10	20	100	697
Número de obs. nos estratos:	14.867	1.579	762	29
Total de Y nos estratos:	33.341	20.942	26.016	5.288
Peso dos estratos aleatórios:	0,862505	0,091605	0,044207	
Média de Y nos estratos aleatórios:	2,2426	13,2628	34,1417	
Variância de Y nos estratos aleatórios:	6,3681	7,4689	255,6040	
Iteração Final (4)				
Limites superiores dos estratos:	5,1412	25,5360	245,7227	697,0000
Número de obs. nos estratos:	12.959	3.782	492	4
Total de Y nos estratos:	19.490	41.353	22.891	1.853
Peso dos estratos aleatórios:	0,751813	0,219412	0,028543	
Média de Y nos estratos aleatórios:	1,5040	10,9342	46,5264	
Variância de Y nos estratos aleatórios:	2,8762	22,9846	757,8139	
Tamanho total da amostra:	50			
Tamanho da amostra aleatória:	46			
Tamanho do estrato certo:	4			

ATIVIDADE 3: Lojas de Departamentos

Número de observações:	62			
Coeficiente de variação:	10%			
Y Total	15.465			
Y Médio	249,435484			
Iteração 0				
Limites superiores dos estratos:	100	150	300	1.140
Número de obs. nos estratos:	15	9	21	17
Total de Y nos estratos:	1.043	1.088	4.492	8.842
Peso dos estratos aleatórios:	0,241935	0,145161	0,338710	
Média de Y nos estratos aleatórios:	69,5333	120,8889	213,9048	
Variância de Y nos estratos aleatórios:	279,8380	247,8611	1.395,5904	
Iteração Final (8)				
Limites superiores dos estratos:	130,0212	296,8240	712,7125	1.140,0000
Número de obs. nos estratos:	20	25	14	3
Total de Y nos estratos:	1.586	5.037	5.927	2.915
Peso dos estratos aleatórios:	0,322581	0,403226	0,225806	
Média de Y nos estratos aleatórios:	79,3000	201,4800	423,3571	
Variância de Y nos estratos aleatórios:	510,8526	2.016,3432	5.846,4008	
Tamanho total da amostra:	6			
Tamanho da amostra aleatória:	3			
Tamanho do estrato certo:	3			

ATIVIDADE 4: Farmácias e Drogarias

Número de observações:	2.720			
Coeficiente de variação:	10%			
Y Total	28.026			
Y Médio	10,303676			
Iteração 0				
Limites superiores dos estratos:	10	20	50	641
Número de obs. nos estratos:	1.941	397	305	77
Total de Y nos estratos:	7.061	5.238	9.326	6.401
Peso dos estratos aleatórios:	0,713603	0,145956	0,112132	
Média de Y nos estratos aleatórios:	3,6378	13,1940	30,5770	
Variância de Y nos estratos aleatórios:	6,9218	7,5961	70,8764	
Iteração Final (6)				
Limites superiores dos estratos:	6,8686	26,6549	163,9445	641,0000
Número de obs. nos estratos:	1.632	836	248	4
Total de Y nos estratos:	4.622	10.640	11.606	1.158
Peso dos estratos aleatórios:	0,600000	0,307353	0,091176	
Média de Y nos estratos aleatórios:	2,8321	12,7273	46,7984	
Variância de Y nos estratos aleatórios:	4,0208	28,9722	462,7688	
Tamanho total da amostra:	26			
Tamanho da amostra aleatória:	22			
Tamanho do estrato certo:	4			

ATIVIDADE 5: Vestuário e Têxteis				
Número de observações:	10.780			
Coeficiente de variação:	10%			
Y Total	97.759			
Y Médio	9,068553			
Iteração 0				
Limites superiores dos estratos:	10	20	100	631
Número de obs. nos estratos:	7.882	1.732	1.109	57
Total de Y nos estratos:	26.554	23.301	38.389	9.515
Peso dos estratos aleatórios:	0,731169	0,160668	0,102876	
Média de Y nos estratos aleatórios:	3,3689	13,4532	34,6159	
Variância de Y nos estratos aleatórios:	7.2223	7.3005	252,6700	
Iteração Final (8)				
Limites superiores dos estratos:	6,9886	33,8014	312,3710	631,0000
Número de obs. nos estratos:	6.677	3.627	473	3
Total de Y nos estratos:	17.069	50.171	29.171	1.348
Peso dos estratos aleatórios:	0,619388	0,336456	0,043878	
Média de Y nos estratos aleatórios:	2,5564	13,8326	61,6723	
Variância de Y nos estratos aleatórios:	4.0905	42,8779	1.622,7587	
Tamanho total da amostra:	37			
Tamanho da amostra aleatória:	34			
Tamanho do estrato certo:	3			
ATIVIDADE 6: Outros Pessoais				
Número de observações:	6.275			
Coeficiente de variação:	10%			
Y Total	44.123			
Y Médio	7,031554			
Iteração 0				
Limites superiores dos estratos:	10	20	100	1.508
Número de obs. nos estratos:	5.104	771	382	18
Total de Y nos estratos:	16.022	10.365	13.046	4.690
Peso dos estratos aleatórios:	0,813386	0,122868	0,060876	
Média de Y nos estratos aleatórios:	3,1391	13,4436	34,1518	
Variância de Y nos estratos aleatórios:				
Iteração Final (6)				
Limites superiores dos estratos:	6,4121	26,6880	198,0354	1.508,0000
Número de obs. nos estratos:	4.451	1.596	221	7
Total de Y nos estratos:	10.875	19.401	10.574	3.273
Peso dos estratos aleatórios:	0,709323	0,254343	0,035219	
Média de Y nos estratos aleatórios:	2,4433	12,1560	47,8462	
Variância de Y nos estratos aleatórios:	4,2001	24,5079	624,4398	
Tamanho total da amostra:	36			
Tamanho da amostra aleatória:	29			
Tamanho do estrato certo:	7			

ATIVIDADE 7: Móveis e Eletrodomésticos

Número de observações:	2.215			
Coeficiente de variação:	10%			
Y Total	21.197			
Y Médio	9,569752			
Iteração 0				
Limites superiores dos estratos:	10	20	100	979
Número de obs. nos estratos:	1.693	294	210	18
Total de Y nos estratos:	5.699	3.926	6.940	4.632
Peso dos estratos aleatórios:	0,764334	0,132731	0,094808	
Média de Y nos estratos aleatórios:	3,3662	13,3537	33,0476	
Variância de Y nos estratos aleatórios:	7,0218	7,7447	190,7345	
Iteração Final (3)				
Limites superiores dos estratos:	6,4629	24,5405	145,5926	979,0000
Número de obs. nos estratos:	1.449	602	153	11
Total de Y nos estratos:	3.778	7.227	6.360	3.832
Peso dos estratos aleatórios:	0,654176	0,271783	0,069074	
Média de Y nos estratos aleatórios:	2,6073	12,0050	41,5686	
Variância de Y nos estratos aleatórios:	4,0964	21,7687	443,0230	
Tamanho total da amostra:	30			
Tamanho da amostra aleatória:	19			
Tamanho do estrato certo:	11			

ATIVIDADE 8: Automóveis, Motos e Peças

Número de observações:	4.985			
Coeficiente de variação:	5%			
Y Total	43.236			
Y Médio	8,673220			
Iteração 0				
Limites superiores dos estratos:	5	10	50	358
Número de obs. nos estratos:	3.136	1.020	690	139
Total de Y nos estratos:	4.971	6.624	12.587	19.054
Peso dos estratos aleatórios:	0,629087	0,204614	0,138415	
Média de Y nos estratos aleatórios:	1,5851	6,4941	18,2420	
Variância de Y nos estratos aleatórios:	2,3583	1,7811	78,0995	
Iteração Final (24)				
Limites superiores dos estratos:	3,9809	25,3025	152,9218	358,0000
Número de obs. nos estratos:	2.672	2.043	218	52
Total de Y nos estratos:	3.115	16.658	12.905	10.558
Peso dos estratos aleatórios:	0,536008	0,409829	0,043731	
Média de Y nos estratos aleatórios:	1,1658	8,1537	59,1972	
Variância de Y nos estratos aleatórios:	1,5790	22,6003	1.371,9370	
Tamanho total da amostra:	145			
Tamanho da amostra aleatória:	93			
Tamanho do estrato certo:	52			

ATIVIDADE 9: Combustíveis				
Número de observações:	1.127			
Coeficiente de variação:	10%			
Y Total	19.095			
Y Médio	16,943212			
Iteração 0				
Limites superiores dos estratos:	10	20	100	343
Número de obs. nos estratos:	307	502	315	3
Total de Y nos estratos:	1.568	7.081	9.885	561
Peso dos estratos aleatórios:	0,272405	0,445430	0,279503	
Média de Y nos estratos aleatórios:	5,1075	14,1056	31,3810	
Variância de Y nos estratos aleatórios:	9,7041	8,1066	169,0073	
Iteração Final (5)				
Limites superiores dos estratos:	9,9612	26,9911	130,6102	343,0000
Número de obs. nos estratos:	307	656	163	1
Total de Y nos estratos:	1.568	10.560	6.625	342
Peso dos estratos aleatórios:	0,272405	0,582076	0,144632	
Média de Y nos estratos aleatórios:	5,1075	16,0976	40,6442	
Variância de Y nos estratos aleatórios:	9,7041	20,1340	239,8725	
Tamanho total da amostra:	13			
Tamanho da amostra aleatória:	12			
Tamanho do estrato certo:	1			
ATIVIDADE 10: Material de Construção				
Número de observações:	7.431			
Coeficiente de variação:	10%			
Y Total	52.162			
Y Médio	7,019513			
Iteração 0				
Limites superiores dos estratos:	5	30	100	542
Número de obs. nos estratos:	4.079	3.151	174	27
Total de Y nos estratos:	6.862	31.764	8.333	5.203
Peso dos estratos aleatórios:	0,548917	0,424034	0,023415	
Média de Y nos estratos aleatórios:	1,6823	10,0806	47,8908	
Variância de Y nos estratos aleatórios:	2,5204	28,7148	283,3117	
Iteração Final (6)				
Limites superiores dos estratos:	5,7830	28,4522	232,2178	542,0000
Número de obs. nos estratos:	4.636	2.583	205	7
Total de Y nos estratos:	9.647	28.660	11.496	2.359
Peso dos estratos aleatórios:	0,623873	0,347598	0,027587	
Média de Y nos estratos aleatórios:	2,0809	11,0956	56,0780	
Variância de Y nos estratos aleatórios:	3,3814	26,9076	1.143,9350	
Tamanho total da amostra:	40			
Tamanho da amostra aleatória:	33			
Tamanho do estrato certo:	7			

BIBLIOGRAFIA

BANKIER,. M. D. Power allocations, determining sample sizes for subnational areas. *The American Statistician,* [s. l.], v. 42, n.3, p. 174-177, 1988.

COCHRAN, W. G. *Sampling tecnhiques.* 3. ed. New York: J.Wiley, 1977. 428 p.

HIDIROGLOU, M. A. The construction of a self-representing stratum of large units in survey design. *The American Statistician*, [s. l.], v. 40, n. 1, p. 27-31, 1986.

LAVALLÉE, P., HIDIROGLOU, M. A. On the stratification of skewed populations. *Survey Methodology,* [s. l.], v. 14, n. 1, p. 33-43, 1988.

PESQUISA anual de comércio. Rio de Janeiro: IBGE, 1991. 61 p. (Série Relatórios Metodológicos, n. 12).

SILVA, P. L. N. *Macros para seleção de amostras.* Rio de Janeiro: IBGE, 1989. 64 p.

RESUMO

Nesse artigo são apresentados os resultados dos estudos realizados para definir o desenho amostral da Pesquisa Mensal de Comércio. Dada a forte assimetria da distribuição da variável Pessoal Ocupado (PO), utilizada para definir o desenho amostral, técnicas especiais de amostragem foram analisadas. Os dois primeiros métodos usam limites preestabelecidos para os estratos de classe de PO e o método proposto por Hidiroglou (1986) é aplicado para definir o estrato certo. No primeiro método, depois de definidas as unidades pertencentes ao estrato certo, uma amostra aleatória simples estratificada é retirada do restante da população, fixando-se o CV para o total estimado. No segundo, é usado também o tamanho n da amostra obtido pelo método de Hidiroglou para definir uma alocação proporcional dessas n unidades nos estratos restantes (aleatórios) da pesquisa. O terceiro método aplicado foi proposto por Lavallée-Hidiroglou (1988) e consiste em calcular os limites dos estratos, de modo a minimizar o tamanho da amostra necessário para obter a estimativa do total com um determinado coeficiente de variação. Dada a necessidade de se obter estimativas por classe de PO, o desenho amostral da PMC estabelece um compromisso entre os resultados obtidos, usando o valor de corte para o estrato certo obtido pelo último método mas redefinindo os limites das outras classes de PO.

ABSTRACT

The paper describes studies carried out to design the sample for the Monthly Survey of Retail Trade. Special emphasis is given to the definition of the stratification by number of employees, the size measure avaiable for each unit - three approaches were considered envolving either *ad hoc* definition of the strata or a stratification based to deal with size measure with skewed distributions. All three methods define one "take-all" stratum of "large" units, which are included in the sample with certainty. The first approach uses a method proposed by Hidiroglou(1986) to define the cut-off point, above which units are included in the "take-all" stratum, and *ad hoc* limits to the classes of employees, when a stratified simple sample is selected. The second approach follows the same stratum boundaries that the first one, although the overall sample size is computed assuming that simple random sampling is to be selected from one "take-some" stratum comprising units not in the "take-all" stratum [see Hidiroglou (1986)]. In this case, the sample size for the "take-some" stratum is proportionally allocated within the *ad hoc* classes. The third approach [Lavallée-Hidiroglou(1988)] allows more than one "take-some" stratum. The stratum boundaries are computed in such a way that the sample size, needed to achieve a given precision for the estimator of total (of the size variable), is minimized. The stratification finally adopted is a comprise between the cut-off point for the certainty stratum via the third method and the size classes defined ad hoc on the basis of the user's requirement for estimates by classes of employees.

APURAÇÃO DE DADOS NO IBGE: PROBLEMAS E PERSPECTIVAS

Pedro Luis do Nascimento Silva*
Zélia Magalhães Bianchini*
Sonia Albieri (Tradutora)*

APRESENTAÇÃO

Este documento é uma tradução do texto *Data Editing Issues and Strategies at the Brazilian Central Statistical Office*, preparado para subsidiar a apresentação durante o 49º Congresso Bianual do *International Statistical Institute - ISI*, realizado em Florença, no período de 25 de agosto a 2 de setembro de 1993.

A preparação do referido trabalho foi realizada em função do estímulo e honroso convite de Leopold Granquist, do *Statistics Sweden*, organizador do tópico *Data editing strategies* da sessão *Selected Topics for Contributed Papers* do 49º Congresso do ISI.

* IBGE
Título original: *Data Editing Issues and Strategies at the Brazilian Central Statistical Office.*

1 INTRODUÇÃO

A Fundação Instituto Brasileiro de Geografia e Estatística - IBGE - é o órgão responsável pelas estatísticas oficiais no Brasil. Fundado em 1936, atualmente o IBGE é uma organização grande e com uma estrutura complexa, contando com mais de 11 000 empregados. O IBGE é responsável pelos Censos Demográfico e Econômico, inclusive o Agropecuário, bem como por um grande número de pesquisas, tais como:

1. Pesquisa Nacional por Amostra de Domicílios (PNAD) - uma pesquisa anual de múltiplos propósitos;

2. Pesquisa de Orçamentos Familiares (POF) - uma pesquisa por amostra de domicílios, realizada duas vezes, em 74/75 e em 87/88, e pensada para ser realizada aproximadamente a cada cinco anos;

3. Pesquisa Industrial Mensal (PIM) - uma pesquisa mensal por amostra de estabelecimentos industriais para estimar a variação no nível de emprego, salários, vendas e produção no setor industrial;

4. Pesquisa Industrial Anual (PIA) - uma pesquisa anual por amostra de estabelecimentos industriais, para estimar o nível e as variações anuais do setor;

5. Pesquisa Anual do Comércio (PAC) - uma pesquisa anual por amostra de empresas de comércio por atacado e varejo;

6. Sistema Nacional de Índices de Preços ao Consumidor (SNIPC) - um conjunto de índices de preços ao consumidor baseado em uma amostra de cerca de 200 000 observações mensais de preços;

7. Pesquisa Mensal de Emprego (PME) - uma pesquisa por amostra de domicílios sobre emprego e renda, que cobre as seis maiores regiões metropolitanas do País;

8. Pesquisa Anual do Transporte Rodoviário (PATR) - um censo anual das empresas de transporte rodoviário de carga e de passageiros; e

9. Assistência Médico-Sanitária (AMS) - um censo anual de estabelecimentos de saúde, tais como hospitais, etc.

O IBGE é responsável também pelas estatísticas das contas nacionais, assim como por muitas outras pesquisas menores ou especiais, muitas delas realizadas em convênio com órgãos governamentais.

Essas informações gerais têm por objetivo auxiliar na compreensão do diagnóstico e das discussões que serão apresentadas.

Durante o período de 1985 a 1987, o IBGE passou por alterações substanciais em sua estrutura administrativa e em sua forma de operar. As discussões que precederam a reorganização do órgão incluíram uma avaliação dos principais problemas encontrados na realização de suas pesquisas e censos.

Uma das conclusões mais importantes do diagnóstico realizado foi que o processo de apuração da maioria das pesquisas e censos realizados pelo IBGE era uma etapa de trabalho intensivo, fortemente centralizada, extremamente morosa, de alto custo e algumas vezes até mesmo pouco confiável e imprevisível.

Os motivos para um diagnóstico tão severo serão examinados no capítulo 2, que contém uma revisão dos processos de apuração de dados

utilizados no IBGE até a realização dos Censos Econômicos e Agropecuário de 1985. O capítulo 3 apresenta as modificações introduzidas nos procedimentos de apuração desde então, incluindo referências a algumas experiências com procedimentos e estratégias alternativas.

O capítulo 4 contém uma análise resumida da situação atual e das perspectivas futuras, bem como as principais conclusões deste documento. Finalmente, no capítulo 5 são listadas as referências utilizadas.

2 REVISÃO DOS PROCESSOS DE APURAÇÃO UTILIZADOS ATÉ 1985

2.1 Considerações Gerais

Para entender as características do processo de apuração de dados usado no IBGE até a realização dos Censos Econômicos e Agropecuário de 1985, é importante descrever como esse processo está inserido no esquema geral de uma pesquisa[1] típica do IBGE.

A coleta de dados para as muitas pesquisas realizadas pelo IBGE ainda é feita principalmente por entrevistadores ou enumeradores. Para as pesquisas regulares, os entrevistadores são empregados permanentes do órgão. Para os censos e algumas pesquisas especiais, os entrevistadores são contratados na ocasião por tempo determinado.

[1] A partir deste capítulo, a palavra pesquisa será usada para designar tanto uma típica pesquisa por amostra como um censo.

Os entrevistadores coletam os dados e registram as respostas em questionários impressos em papel, para cada pesquisa, ou seja, o método de coleta utilizado é o conhecido como PAPI - *Paper And Pencil Interviewing.*

2.2 Desenho de Questionários e Coleta de Dados

Os problemas de apuração dos dados já começam no estágio de desenho do questionário. No IBGE, o desenho de questionários foi, no passado, uma atividade em que pouca ou nenhuma atenção era dada às necessidades das fases subseqüentes de processamento e apuração dos dados. O desenho de questionários era feito manualmente, por especialistas de cada área, sem nenhuma padronização, não contando nem mesmo com algumas regras básicas ou com instruções para direcionar o trabalho daqueles que executavam as tarefas.

Freqüentemente eram usadas complexas instruções de fluxos dentro do questionário, tornando bastante difícil até mesmo o trabalho de entrevistadores experientes. Além disso, os questionários não continham formas de distinguir zeros estruturais de valores faltantes, especialmente em pesquisas econômicas para as quais essa informação é muito importante.

Os especialistas da área que desenhavam questionários tinham pouca ou nenhuma informação sobre a natureza do processo de apuração, e geralmente deixavam para outras equipes a tarefa de especificar as etapas de apuração, equipes que, algumas vezes, detinham pouco conhecimento sobre o assunto da pesquisa. A etapa de apuração não era considerada uma tarefa "nobre" ou "vital", embora ela fosse responsável por uma grande proporção do tempo e do custo da pesquisa. Além disso,

os especialistas em processamento dos dados tinham pouca chance de influenciar o desenho de questionários.

As instruções para os entrevistadores, geralmente na forma de manuais de pesquisa, eram outra fonte de problemas. Freqüentemente essas instruções envolviam conceitos e definições chaves, os quais afetavam as etapas subseqüentes de processamento dos dados da pesquisa. Novamente, os responsáveis pelo processamento dos dados não participavam dessa etapa. O preparo de tais manuais de instruções, assim como o desenho de questionários, estava sob a responsabilidade dos especialistas da área, excluindo as equipes de processamento dos dados.

Esses manuais de pesquisa freqüentemente requeriam que os entrevistadores ou seus supervisores executassem manualmente uma certa quantidade de regras de crítica do questionário como um todo, enquanto eles ainda estivessem no campo, ou seja, durante a fase de coleta de dados. Esta prática era justificada em função da proximidade com os respondentes. Entretanto, os excessos tanto na quantidade como na complexidade das regras de crítica a serem realizadas manualmente afetavam negativamente não só a duração como também a qualidade da etapa de coleta dos dados. Isto porque os entrevistadores dedicavam uma parte substancial do seu tempo de trabalho às operações de crítica. E até mesmo induzia os entrevistadores a acreditarem que quaisquer erros cometidos durante a coleta dos dados poderiam ser "corrigidos" mais tarde, o que muitas vezes era usado como desculpa pela baixa qualidade das entrevistas.

Outro problema relativo às regras de crítica definidas nos manuais de pesquisa era a ausência de padrões pré-definidos de qualidade a serem atingidos. Isso significa que a decisão de recontatar um respondente no caso de algum erro detectado durante essa operação de crítica manual era deixada inteiramente para o entrevistador ou seu supervisor. Ou seja,

dada uma falha no questionário, o recontato com o informante dependia exclusivamente do julgamento e dos recursos disponíveis para o supervisor que realizava a verificação, e não de um critério pré-definido escrito nos manuais da pesquisa.

2.3 Apuração dos Dados

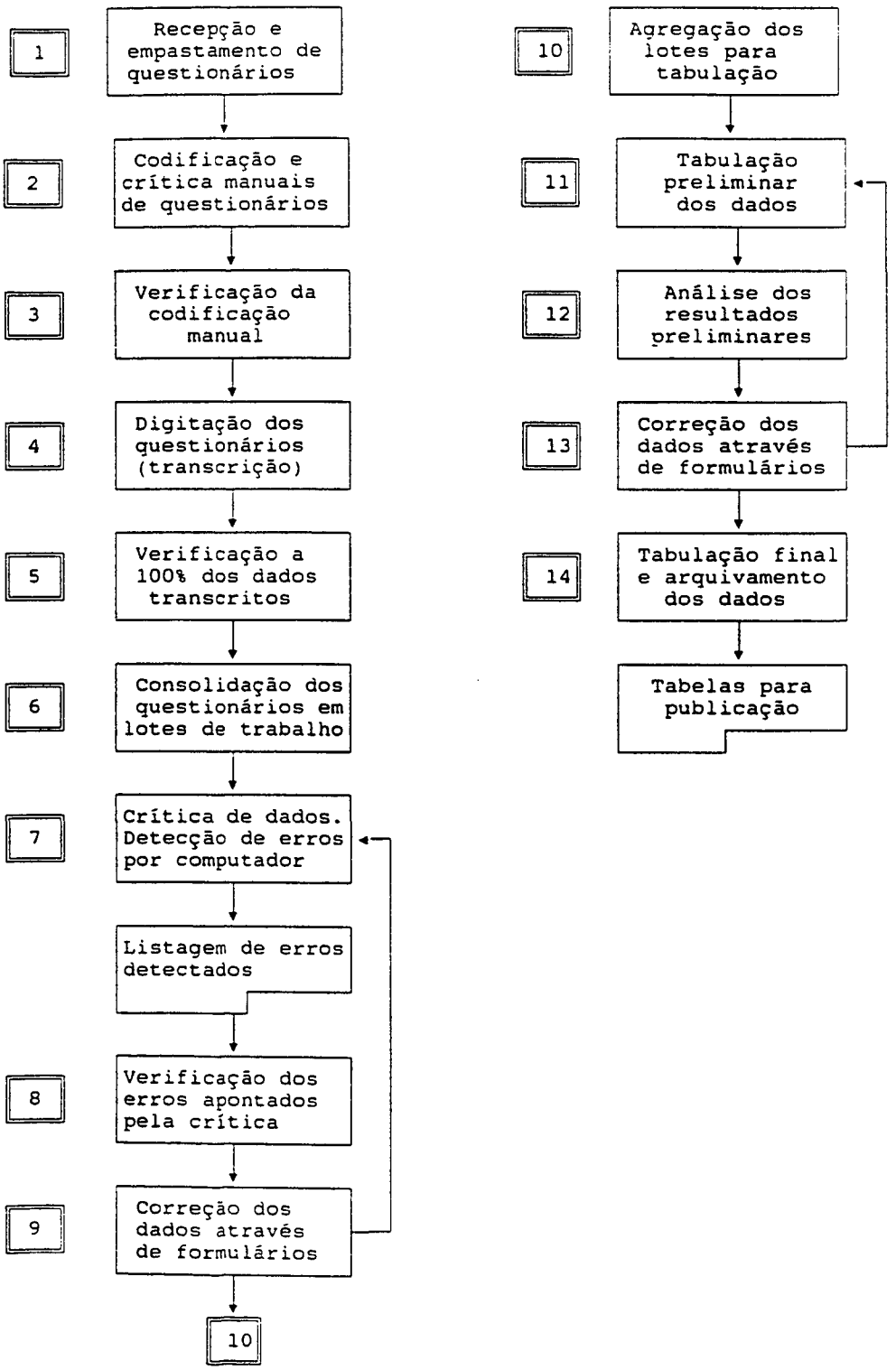
O diagrama apresentado na Figura 1, a seguir, mostra as etapas de apuração realizadas após a coleta dos dados até a obtenção dos resultados finais de uma pesquisa.

A principal característica do esquema de apuração de dados de uma pesquisa usado no IBGE era sua forte centralização. Todas as atividades apresentadas na Figura 1 eram realizadas na sede do Órgão, no Rio de Janeiro, embora mais de 20 escritórios regionais e mais de 500 agências municipais estivessem envolvidas nas atividades de coleta dos dados.

Uma conseqüência imediata dessa centralização das atividades de apuração das pesquisas era a perda de oportunidade de recontatar um informante no caso de ser detectado algum erro no respectivo questionário após ele ter sido enviado ao órgão central. A volta ao informante ou recontato só era efetuada no caso de pesquisas de estabelecimentos econômicos e geralmente a um custo muito alto.

Outra conseqüência negativa era a necessidade de concentrar, no órgão central, um número muito grande de pessoas para realizar as atividades de digitação e de crítica dos questionários, com várias implicações na produtividade e na qualidade do trabalho. Para exemplificar, o número médio de caracteres digitados por hora era cerca de 3 000 por operador no órgão central. Quando foi introduzida a descentralização da atividade de digitação, essa média passou para cerca de 8 000 caracteres/hora por operador, em muitos dos escritórios regionais.

Figura 1
Fluxo das atividades de apuração de uma pesquisa típica



2.3.1 Recepção e Empastamento

Vamos agora analisar cada uma das grandes atividades descritas na Figura 1. Começamos com a recepção e empastamento dos questionários. No passado, as especificações para essa operação eram dadas principalmente pela equipe de apuração dos dados e eram desenvolvidas visando à etapa de digitação dos dados.

Devido às limitações dos sistemas de entrada de dados utilizados, os questionários componentes de uma pasta (lote de trabalho) tinham que ser de um único tipo ou modelo. Por exemplo, nos Censos Econômicos de 1985, isso significava separar os questionários de uma empresa em diferentes pastas em função do processamento dos dados, o que tornava bastante complicada a etapa de crítica e apuração da empresa como um todo, que era realizada mais tarde.

Outra dificuldade era a ausência de facilidades automáticas para a recepção e o empastamento, o que fazia com que essa fosse uma operação bastante sujeita a erros. Por exemplo, apenas recentemente começou a ser usado um sistema de etiquetas pré-impressas a fim de facilitar a identificação de questionários durante a etapa de recepção e empastamento.

2.3.2 Codificação e Crítica Manuais

Após a etapa de empastamento, os questionários passavam para a etapa de codificação manual. Geralmente, essa etapa era realizada sem um controle rígido sobre a qualidade final, embora tenham sido introduzidos procedimentos de controle de qualidade. A respeito, ver

Cabral (1988), Silva, Cabral & Indá (1986), Silva et al. (1986) e IBGE (1983a,b).

Um grande problema relativo a esse esquema de codificação manual residia na equipe que a realizava, equipe essa com pouco ou nenhum conhecimento sobre o assunto da pesquisa. Isto implicava em uma baixa qualidade dos códigos resultantes, o que afetava a etapa posterior de detecção automática de erros nos dados. De fato, a análise dos registros detectados com algum erro apontava que muitos registros falhavam as regras de crítica exatamente pela existência de códigos errados. Por exemplo, um experimento para avaliar a qualidade da codificação no Censo Industrial de 1985 mostrou uma taxa de erro de cerca de 18% dos códigos, como relatado em Silva et al.(1986).

Outra questão relativa à etapa de codificação manual era que em muitas pesquisas ela era realizada simultaneamente com a etapa de crítica manual. O motivo para agir dessa forma não era muito claro. Parece que os responsáveis pela pesquisa tinham um sentimento que apontava para algo do tipo: uma vez que os questionários serão examinados para poderem ser codificados, por que não aproveitar para realizar algumas "verificações de qualidade" ao mesmo tempo?

Atualmente, tem-se a consciência de que essa etapa de crítica manual realizada no órgão central podia introduzir muito mais problemas do que os que eram solucionados. Freqüentemente, um número excessivo de regras de crítica era realizado durante essa etapa, ainda que não conseguissem garantir que o conjunto de dados estivesse livre de erros grosseiros. Muitas das regras de crítica aplicadas manualmente tinham que ser repetidas mais tarde, após a digitação dos dados, resultando em redundância. Além disso, há a considerar a heterogeneidade de tratamento entre os diversos operadores para um mesmo tipo de falha detectada.

Também nessa etapa de crítica manual há que observar a ausência de um controle de qualidade dos resultados. E também que qualquer modificação introduzida nos dados coletados nessa etapa não podia ser detectada mais tarde, após a digitação dos dados.

Outro problema a ser destacado era o atraso que ela provocava no envio dos questionários para a etapa de digitação e demais etapas de apuração, sem qualquer expectativa de melhoria na qualidade dos dados. Outro problema, ainda, residia na impossibilidade de com esse procedimento detectar a existência de erros sistemáticos de coleta, uma vez que não era realizada nenhuma análise estatística dos erros detectados. Isso impedia que a operação de crítica servisse como um guia para o aperfeiçoamento de pesquisas similares futuras, um dos objetivos importantes numa operação de crítica - ver Granquist (1984).

2.3.3 Transcrição dos Dados

O processo de transcrição no IBGE consistia basicamente em digitar os dados dos questionários e gravá-los em fitas magnéticas. Geralmente, o sistema computacional usado era preparado exclusivamente para cada tipo de pesquisa ou de questionário.

A principal preocupação com a operação de digitação dos dados era a velocidade e pouca atenção era dada à qualidade do processo. Não havia nenhuma verificação ou regra de crítica associada ao processo de digitação o que tornava possível a introdução de erros grosseiros durante a sua execução. Um exemplo disso é o fato de que um operador de digitação podia introduzir um código inválido para qualquer campo do questionário sem que o sistema computacional o detectasse. Havia casos de programas de entrada de dados numéricos nos quais podiam ser

introduzidos caracteres não numéricos sem que fossem automaticamente emitidas mensagens de erro ou de alerta.

Um procedimento que era usado com freqüência para "garantir" a qualidade do processo consistia na dupla digitação dos dados, cada uma delas realizada por um operador diferente e independentemente uma da outra. O segundo operador era alertado quando o valor digitado para um determinado campo diferia do valor gravado pela primeira digitação para o mesmo campo. Entretanto, esse operador não tinha como ver qual havia sido o valor digitado. O valor para esse campo era então digitado novamente e, para que esse último fosse definitivamente incorporado ao arquivo de dados, era necessário que coincidisse com qualquer um dos outros dois valores digitados anteriormente. Isso significa que o segundo operador tinha o poder de decisão relativamente a qual valor deveria ser gravado. Convém mencionar que um operador podia realizar ambas as operações: a primeira ou a segunda digitação, essa última denominada verificação.

Fora esse sistema de "verificação a 100%", não era realizado nenhum procedimento para garantir a qualidade do processo de digitação. Como exemplo: o controle dos operadores de digitação não envolvia nenhuma medida da qualidade do seu trabalho, mas somente sua taxa de produção, medida pelo número de caracteres digitados por hora. Não havia nenhum controle da proporção de erros de digitação para cada pesquisa e, muito recentemente, algumas tentativas para a introdução de procedimentos de controle de qualidade como prática rotineira do processo de entrada de dados foram infrutíferas.

Com relação ao sistema de entrada de dados, outro problema a destacar refere-se à centralização de toda a operação no órgão central, no Rio de Janeiro. Durante a apuração dos Censos de 1980, cerca de 3 000 operadores estiveram atuando ao mesmo tempo na tarefa de digitação dos questionários dos censos e das demais pesquisas do IBGE. Os baixos

salários pagos a esses operadores propiciaram a formação de uma forte associação de funcionários, que foi responsável por várias greves ou paralisações e por outras manifestações com os objetivos de melhoria de seus salários ou de "manutenção do emprego", mesmo quando não havia mais trabalho para os quais haviam sido contratados.

Um exemplo do impacto negativo dessa concentração de operadores de digitação é o fato de que o número médio de caracteres digitados por operador que era cerca de 3 000 por hora no órgão central passou para cerca de 8 000 por hora após a descentralização da digitação nos escritórios regionais.

Uma vez que a digitação dos dados de todos os censos e pesquisas era realizada por um único departamento, no órgão central, os dados ficavam distantes dos respectivos departamentos temáticos responsáveis. Isso era um outro problema, pois aquelas pessoas responsáveis pelo planejamento das pesquisas não tinham contato com os problemas encontrados durante a entrada de dados, causados por falhas no desenho dos questionários. Isto implicava, também, que as especificações dadas pela equipe de entrada de dados tinham que ser adotadas pelos especialistas da área que desenhavam questionários, uma vez que não tinham a menor idéia das implicações efetivas de seu trabalho na natureza das atividades de entrada de dados. Como a equipe de entrada de dados não tinha conhecimento sobre o processo de coleta dos dados, muitas de suas sugestões ou demandas eram freqüentemente pouco razoáveis e causavam problemas desnecessários para a equipe de coleta dos dados.

2.3.4 Crítica dos Dados

Nesta seção é feita uma revisão dos procedimentos de crítica adotados no passado pelo IBGE. Para entender o diagnóstico crítico e

negativo anteriormente apresentado, é necessário descrever primeiro como a operação de crítica dos dados era planejada e realizada para uma pesquisa típica do IBGE.

Convém destacar que a etapa de crítica dos dados compreendia uma série de atividades ou operações distintas, a saber:

a) crítica quantitativa - realizada com o objetivo de detectar e corrigir os erros provenientes de regras de crítica de valores possíveis, equações estruturais (por exemplo, totalizações em questionários de pesquisas econômicas), bem como verificações de cobertura, através de cadastros ou outros tipos de controle do número de questionários coletados;

b) crítica qualitativa - realizada com o objetivo de detectar e corrigir os erros provenientes de regras de crítica de consistência, geralmente derivadas da estrutura de rota (fluxo) dos questionários e de regras de aceitação especificadas pelos especialistas temáticos; e

c) crítica de agregados - realizada com o objetivo de comparar os dados agregados da pesquisa com outros similares ou com informações relacionadas, geralmente correspondentes a uma rodada anterior da mesma pesquisa; em alguns casos, procedimentos de detecção de valores suspeitos (*outliers*) eram realizados como parte dessa operação.

Cada uma das operações acima descritas era realizada em cada conjunto de dados (lote de trabalho) de forma cíclica e repetida, tal como mostra o diagrama abaixo.

→ detecção automática → verificação manual → correção manual

As duas primeiras operações (crítica quantitativa e crítica qualitativa) consistem nas denominadas etapas de crítica de microdados - *micro-editing* (ver Granquist(1990)). A última, crítica de agregados, era geralmente realizada após a tabulação dos dados em algum nível de agregação, geralmente considerando nível de desagregação igual ao das tabulações finais de divulgação.

Um primeiro problema, devido à separação da crítica em várias operações distintas, era a existência de duplicação ou de redundância na aplicação de regras de crítica. As alterações (correções) introduzidas nos dados em um dado momento tinham que ser verificadas considerando as regras de crítica usadas nos passos anteriores, a fim de evitar que novas inconsistências fossem introduzidas. Isso acarretava uma complexidade muito grande nos programas de crítica (programas de detecção de erros segundo um conjunto de regras de crítica), principalmente para os estágios mais avançados da crítica.

Um segundo problema, especificamente para a operação de crítica qualitativa descrita acima, era a ausência de metodologia para a análise lógica das regras de crítica especificadas. Isto permitia a aplicação de um conjunto de regras inconsistentes entre si, sem que se percebesse o fato, até que por questões de tempo e prazos se decidisse por encerrar a operação de crítica sem que houvesse sido solucionado o total de falhas apontadas pelo programa de crítica usado.

Outro problema relativo a esse esquema de crítica de dados era a ausência de integração entre as fases de detecção dos erros e correção dos dados. Os operadores que analisavam as falhas apontadas pelo programa de crítica não tinham instruções precisas sobre como corrigir uma determinada falha. Assim, diferentes operadores podiam tentar diferentes "correções" para um mesmo tipo de falha. Essas correções manuais geralmente envolviam uma grande quantidade de imputações, no

sentido de que não era feito novo contato com os informantes do dado original devido à falta de tempo e recursos.

Outro problema ainda era causado pelo tipo de procedimento de correção adotado, o qual não incorporava nenhuma verificação antes que as alterações fossem de fato introduzidas no conjunto de dados. Isso era responsável por situações em que eram detectadas falhas em outras regras de crítica após a introdução indevida de correções comandadas manualmente por um operador.

Todos os problemas acima descritos acarretavam a execução repetida de vários ciclos de crítica sobre um mesmo conjunto de dados, o que tornava essa etapa demasiadamente longa, custosa e trabalhosa.

Entretanto, esses problemas não eram os únicos existentes. A especificação das regras de crítica era geralmente realizada pelos especialistas temáticos, sem a devida assistência de estatísticos ou de especialistas em processamento informatizado de dados. Durante a especificação das regras, esses especialistas temáticos não consideravam o dicionário dos dados (uma descrição dos dados após terem sido codificados e gravados em fitas magnéticas). Habitualmente, eles realizavam seu trabalho apenas considerando o questionário e as instruções de coleta dos dados, na forma de manuais de instrução, desprezando quaisquer transformações que os dados tivessem sofrido para permitir o processamento por computador.

Em muitas situações, as pessoas responsáveis pelo desenho do questionário não participavam do processo de especificação das regras de crítica. Outra questão ainda era que os especialistas temáticos tendiam a dar atenção a cada operação de crítica isoladamente, significando pouca ou nenhuma integração entre as diferentes operações de crítica. A única tentativa de integração que era feita consistia em, durante os estágios

mais avançados da crítica, repetir a aplicação dos programas de crítica usados nas operações anteriores, o que causava a já mencionada redundância das regras de crítica.

Os especialistas temáticos habitualmente usavam a linguagem natural para escrever as regras de crítica. E assim, especialistas temáticos de diferentes áreas tinham diferentes formas de especificar as suas regras de crítica. Isto implicava a necessidade de "tradução" das regras definidas para a linguagem própria de programas de computador, por parte da equipe de processamento dos dados. Essa tradução nem sempre passava por uma verificação por parte dos responsáveis por sua especificação e, portanto, era responsável por muitos erros nos programas de crítica, bem como pelo atraso na sua implantação definitiva.

A ausência de uma linguagem comum entre os especialistas temáticos e a equipe de processamento de dados era um motivo de freqüentes atritos e desentendimentos.

Geralmente os programas de crítica dos dados eram desenvolvidos especificamente para cada pesquisa, usando linguagens tais como COBOL e PL/I. Essa forma de desenvolvimento de sistemas implicava não só altos custos, como também tornava a operação de crítica como um todo em uma operação complexa, demorada e bastante sujeita a erros. A fase de testes desses programas de crítica era bastante difícil. Geralmente era realizada usando conjuntos de dados com erros introduzidos pelos especialistas temáticos, a fim de verificar se os programas seriam capazes de detectá-los. Por motivos óbvios, esse procedimento estava longe de ser ideal e não garantia que os programas ficassem livres de erros, mesmo para conjuntos simples de regras de crítica.

Outro problema relativo às especificações das regras de crítica estava relacionado com o tipo de desenvolvimento de sistemas adotado. Uma vez que o desenvolvimento dos programas de crítica era tão

trabalhoso e demorado, os especialistas temáticos tinham que definir todas as regras de crítica muito antes de terem acesso a um subconjunto dos dados da pesquisa. Isto implicava muitas vezes a definição de limites e de relações entre as variáveis irreais quando analisados, considerando-se os dados. As regras definidas podiam atingir muitos dados coletados ou até mesmo podiam não afetar nenhum dado. Se fossem detectados muitos dados falhos em função de uma mesma regra de crítica, e como era arriscado e custoso realizar alterações nos programas de crítica, as mensagens de erro ou de alerta emitidas pelo programa nesses casos tinham que ser ignoradas pelos operadores de crítica, o que significava perda de tempo e certamente maior complexidade dessa tarefa.

O resultado final da combinação de todos os problemas acima descritos era uma metodologia de crítica trabalhosa, extremamente morosa, de alto custo e algumas vezes até mesmo não muito confiável e imprevisível. De fato, não havia nenhum controle do impacto das operações de crítica nos resultados, nem eram calculados indicadores de qualidade no final da realização de todas essas operações de crítica. Em alguns casos foram encontradas evidências de verificação/correção excessivas (*over-editing*), ou seja, a substituição de dados bons por outros imputados em função do excesso de críticas.

Mesmo após a realização de todas essas operações de crítica, podia ser encontrado algum erro grosseiro até mesmo nos dados já publicados, o que comprometia a confiança do usuário na qualidade do restante das informações da pesquisa.

Apenas em uma situação pode-se dizer que havia padronização no sistema de apuração e crítica dos dados. Era o procedimento adotado para o desenvolvimento de sistemas para o processamento dos dados por computador. Independentemente da natureza e do tamanho da pesquisa, os sistemas eram desenvolvidos da mesma maneira que para um censo:

programas rodando via *batch*, com o controle central fora do departamento temático responsável pela pesquisa, execução de programas sob medida para cada pesquisa, usando computador de grande porte. Essa padronização, longe de trazer algum benefício, implicava em que, para diversas pesquisas, essa fosse uma solução pouco adequada, pouco eficiente.

Um último comentário diz respeito aos problemas específicos da crítica de agregados. Uma vez que essa era a última operação de crítica realizada antes da tabulação final dos resultados, e uma vez que naquele momento geralmente já se contava com um atraso considerável na divulgação dos resultados, ela era executada sob uma forte pressão dos dirigentes para a liberação dos dados. E a análise de resultados agregados é considerada a melhor oportunidade para detectar erros importantes nos dados, como descrito em Granquist (1987,1988).

3 ALGUMAS MUDANÇAS RECENTES NO PROCESSO DE APURAÇÃO DE DADOS

3.1 Considerações Gerais

Vamos agora descrever as principais mudanças introduzidas em diversas etapas do processo de apuração e crítica dos dados das pesquisas do IBGE, como conseqüência do diagnóstico anteriormente descrito. As mudanças serão descritas segundo cada uma das grandes etapas, tal como na seção 2.

Este documento não pretende ser exaustivo na apresentação de todas as mudanças introduzidas. Convém destacar que esta seção considera diversas situações de pesquisa, e que reconhecemos que os

trabalhos de muitas pessoas envolvidas nessas pesquisas forneceram os elementos básicos necessários para a revisão ora apresentada. Sempre que possível será fornecida a referência bibliográfica original da documentação usada para a preparação deste trabalho. Entretanto, as opiniões expressas são de nossa exclusiva responsabilidade.

3.2 Desenho de Questionários e Coleta dos Dados

Apesar do diagnóstico bastante severo relatado na seção 2.2 acerca dos problemas relacionados com as etapas de desenho de questionário e de coleta de dados, essas duas atividades sofreram poucas alterações. A entrevista pessoal com registro de dados em questionários impressos ainda é o método básico de coleta de dados adotado no IBGE.

Entretanto, atualmente há um reconhecimento do impacto dos métodos usados para o desenho de questionários e para a coleta dos dados nas etapas subseqüentes de processamento dos dados das pesquisas. Com relação à etapa de desenho de questionários, algumas ferramentas foram introduzidas de forma a minimizar os problemas detectados.

Para as pesquisas da área econômica, atualmente é prática comum a realização de uma análise do questionário de cada pesquisa nova (ou de um questionário redesenhado de uma pesquisa já existente) por parte de consultores externos, realizada com a finalidade de garantir que as perguntas possam ser respondidas corretamente pelos elementos da população objetivo. Essa nova prática é complementada pela realização de pré-testes dos questionários com uma pequena amostra de informantes. Essas duas medidas propiciaram uma melhoria nas taxas de respostas para perguntas individuais, uma vez que passaram a ser

evitadas aquelas questões que os informantes mostraram dificuldade ou impossibilidade de responder, em função dos respectivos sistemas usados para armazenar as informações.

Para as pesquisas da área econômica, outra medida que causou impacto positivo nas taxas de resposta às pesquisas foi a adoção de diferentes modelos de questionários, com um número menor de questões e questões mais simples, para coletar dados de unidades definidas como "pequenas". Isto foi também uma conseqüência do reconhecimento da existência de diferentes sistemas de armazenamento de informações de acordo com o tamanho das empresas/estabelecimentos investigados, o que levou à adaptação dos questionários de forma a perguntar somente o que é possível obter para um determinado conjunto de informantes. No passado, apesar dos problemas de não resposta observados para questões mais detalhadas, as pesquisas utilizavam um único modelo de questionário, que era aplicado a todos os tipos de informantes de uma mesma pesquisa.

Outro aperfeiçoamento importante foi introduzido no desenho de questionários de pesquisas domiciliares. Para essas pesquisas, o maior desafio geralmente é a alta complexidade da estrutura de rota (fluxo) de preenchimento dos questionários. Freqüentemente são usados fluxogramas (gráficos de rota ou de fluxo) para representar a estrutura de rota de cada questionário de uma nova pesquisa ou de um questionário redesenhado de uma pesquisa já existente. Esses fluxogramas são de fácil entendimento e análise e portanto facilitam a detecção de quaisquer problemas com o fluxo de preenchimento. A introdução de técnicas de análise de fluxograma para representar graficamente a estrutura de rota de questionários considera as idéias apresentadas em Jabine(1985).

Outro conjunto de técnicas baseadas no trabalho de Willenborg(1987) foi usado de maneira experimental por Farias(1990) para analisar a estrutura de rota de questionários de algumas das principais pesquisas

domiciliares realizadas pelo IBGE, quais sejam: a PNAD - Pesquisa Nacional por Amostra de Domicílios; Censo Demográfico de 1991, o questionário da amostra; e a POF - Pesquisa de Orçamentos Familiares de 1987/88. Willenborg fornece medidas objetivas de complexidade e balanceamento de um dado questionário, com base na sua estrutura de rota. Com base nessas medidas e nos fluxogramas correspondentes, foi possível comparar a complexidade de suas estruturas de rota e também sugerir mudanças a fim de obter uma estrutura mais simples, que resultariam em valores menores para as medidas de complexidade.

Essas técnicas seriam de grande valor se consideradas para o desenho de novos questionários. A idéia de gráficos de rota foi facilmente adotada pelos especialistas temáticos, com pequeno suporte da equipe de metodologia. Entretanto, as medidas de complexidade e balanceamento mostraram-se de difícil compreensão e cálculo, o que requeriria, para seu uso, um envolvimento muito intensivo da equipe de metodologia. Por esse motivo, elas ainda não são consideradas de forma regular entre os especialistas temáticos responsáveis pelo desenho de questionários. Espera-se que no futuro, após um maior esclarecimento das idéias contidas nessa técnica e de um treinamento adequado, os especialistas temáticos passem a aceitar a idéia de adotar medidas objetivas de complexidade e de balanceamento para desenhar seus questionários.

Outra alteração importante introduzida na etapa de desenho de questionários foi a utilização de sistemas computacionais específicos para essa etapa. Em algumas situações, sistemas tais como FORMAX(1990) são usados para se obterem protótipos de questionários, os quais podem então ser testados com respondentes da pesquisa na sua "aparência real", de forma muito mais rápida do que anteriormente, quando era necessário esperar a impressão através dos meios convencionais. Embora esse tipo de *software* não forneça instrumentos formais para detectar erros nos questionários, o simples fato de usar sistemas computacionais para

desenvolver protótipos mostrou-se muito útil para a detecção de problemas seja com a estrutura de rota seja com a estrutura de códigos de algumas questões.

Uma outra alteração introduzida nas pesquisas da área econômica foi a impressão de questionários por computador. No passado, todos os itens de identificação de um questionário tinham que ser preenchidos ou pelos respondentes ou pelos entrevistadores, que tinham que copiar as informações necessárias a partir de relatórios impressos, os quais identificavam os informantes selecionados do cadastro da pesquisa.

Entretanto, muitas pesquisas da área econômica coletam dados para um mesmo conjunto de informantes, o que representa não só uma razoável carga de trabalho para os entrevistadores ou para os respondentes, como também uma fonte extra de introdução de erros. Atualmente, cada pesquisa que possui um cadastro (os censos econômicos realizados tiveram pouco suporte de cadastros de informantes) utiliza as informações nele contidas para a identificação dos respondentes, e todas as identificações relevantes são pré-impressas nos questionários da pesquisa. Isso evita a necessidade de repetição (por parte dos respondentes) ou de transcrição manual (por parte dos entrevistadores). Os respondentes apenas atualizam, em campos apropriados, as eventuais mudanças em relação aos dados de identificação pré-impressos.

Essa alteração reduziu bastante as falhas detectadas devido às inconsistências entre as informações do cadastro e os campos de identificação dos questionários da pesquisa, além de ter reduzido também a carga de trabalho dos respondentes ou dos entrevistadores. Outra vantagem foi a redução da necessidade de impressão convencional de questionários, via processos gráficos. Atualmente eles são produzidos por computador, usando modernas impressoras a *laser* ligadas ao computador

de grande porte (*mainframe*), capazes de imprimir um número muito grande de questionários e de trabalhar com tamanhos especiais de papel.

Outra vantagem ainda foi uma maior integração entre a etapa de desenho de questionários e as subseqüentes operações de processamento dos dados, uma vez que a equipe de processamento de dados agora está envolvida no desenho de questionário desde o seu início, e então algumas das falhas anteriormente citadas podem ser evitadas porque os especialistas temáticos inevitavelmente têm que considerar as necessidades da equipe de processamento dos dados.

Infelizmente, essa vantagem somente foi atingida para algumas pesquisas, quais sejam aquelas bem estruturadas e que se baseiam em cadastros. Muitas pesquisas que não se baseiam em cadastros ainda não têm seus questionários sendo produzidos por computador. Espera-se que essa nova tecnologia seja largamente adotada pelo IBGE, uma vez que ficou provado ser muito mais eficiente e menos custosa, quando comparada com os meios tradicionais de impressão de grandes quantidades de questionários.

Vamos mencionar aqui algumas propostas de mudanças que ainda não foram implantadas, mas que já foram objeto de testes formais dentro do IBGE. O teste mais importante refere-se ao uso de métodos de coleta de dados assistida por computador. O documento IBGE(1987) relata uma experiência pequena realizada na coleta de dados do SNIPC - Sistema Nacional de Índices de Preços ao Consumidor, com o uso de microcoletores de dados portáteis ao invés de questionários impressos. Esse experimento foi considerado um grande sucesso, principalmente em função da natureza da pesquisa: o SNIPC é uma pesquisa repetida mensalmente, com atividade de coleta contínua ao longo do ano e concentrada em 11 áreas metropolitanas (as nove Regiões Metropolitanas mais o Distrito Federal e o Município de Goiânia).

A única razão para essa metodologia de coleta não ter substituído o sistema tradicional de coleta reside na necessidade de um alto investimento na compra de equipamentos, para o qual ainda não há recursos financeiros disponíveis. Entretanto, não só o custo decrescente dos microcoletores de dados, como também o aumento da disponibilidade de sistemas computacionais de coleta e processamento de dados para serem usados nesses equipamentos têm tornado essa alternativa bastante atrativa, quando comparada com a tradicional coleta usando questionários impressos em papel. Espera-se para um futuro próximo que o IBGE consiga os recursos necessários a fim de possibilitar a introdução dessa nova tecnologia na realização de suas pesquisas.

É claro que algumas dificuldades são esperadas para a utilização mais abrangente desse método de coleta em pesquisas que não têm um fluxo regular de coleta de dados ou para aquelas em que os respondentes estejam muito espalhados geograficamente, uma vez que para algumas delas poderá não ser economicamente viável a utilização de microcoletores de dados. Outro problema também pode ser esperado em áreas onde a taxa de criminalidade e a violência são altas, uma vez que o simples fato de portar esse tipo de equipamento pode colocar em risco os entrevistadores.

Entretanto, as vantagens esperadas com o uso desse método de coleta de dados superam, e em muito, essas desvantagens citadas. Uma experiência interessante a relatar é a do *Netherlands Statistics* (NCBS), onde essa nova tecnologia foi introduzida em larga escala, usando tanto o sistema CAPI - *Computer Assisted Personal Interview* como o sistema CATI - *Computer Assisted Telephone Interview*. No Brasil, o uso do sistema CATI nunca foi considerado para coleta de dados em função da pouca abrangência do sistema telefônico entre a população. Para as pesquisas da área econômica, para as quais seria possível a coleta de dados por telefone, em função da existência de telefones nas empresas e

nos estabelecimentos, a dificuldade reside no tamanho e no conteúdo dos questionários, geralmente grandes e complexos.

Entre as vantagens da entrevista realizada de forma assistida por computador sobre os métodos tradicionais de coleta de dados, podemos mencionar:

a) aperfeiçoamento do desenho do questionário, uma vez que a integração entre as etapas de coleta e processamento dos dados está embutida no processo;

b) automação da etapa de desenho de questionário;

c) aperfeiçoamento da documentação e padronização dos questionários e das informações a eles relacionadas;

d) necessidade de uma única especificação dos dados;

e) possibilidade de realizar algumas verificações e críticas durante a coleta de dados, tornando possível a eliminação de erros de rota e validade de códigos de respostas junto à fonte das informações, bem como a redução da chance de erros grosseiros através da introdução de questões de verificação/prova dos respondentes; reduz também o número de regras de crítica a serem consideradas mais adiante, assim como a necessidade de um novo contato com o informante após a entrevista ter sido realizada;

f) eliminação da operação de digitação, uma vez que os dados são imediatamente digitados no computador pelos entrevistadores;

g) melhoria da qualidade dos dados coletados, com redução significativa do tempo e do custo da pesquisa;

h) concentração da equipe de coleta de dados na realização efetiva da entrevista e não nas trabalhosas operações de crítica;
i) eliminação da etapa de crítica manual e, portanto, das redundâncias anteriormente citadas; e

j) maior controle do trabalho de campo.

3.3 Recepção e Empastamento

A recepção e empastamento é ainda uma operação predominantemente manual, mas a eficiência e a qualidade foram significativamente melhoradas, devido ao maior uso de questionários com a identificação do informante pré-impressa, seja na forma de questionários impressos por computador seja pela utilização de etiquetas adesivas. Entretanto, ainda há algumas pesquisas que usam o sistema tradicional descrito no capítulo 2.

Para as pesquisas econômicas que se baseiam em cadastros, a disponibilidade de sistemas computacionais modernos para a entrada de dados permitiu o acondicionamento de diferentes tipos de questionários em um mesmo lote de trabalho, reduzindo assim a complexidade dessa tarefa e conseqüentemente sua taxa de erros. Como exemplo podemos citar a PIA - Pesquisa Industrial Anual, que, no período de 1986 a 1990, usou quatro tipos de questionários: um para a empresa como um todo; um para informações locais; um para informações do estabelecimento; e um para pequenas empresas (empresas com um único estabelecimento), que coleta menos informações que os três níveis de organização anteriores.

Uma vez que os questionários de uma dada companhia são preenchidos e enviados de uma só vez, a nova tecnologia de entrada de

dados tornou possível mantê-los juntos durante o processo de formação de lotes de trabalho, evitando assim a necessidade de separação dos questionários e facilitando as etapas subseqüentes de crítica dos dados da companhia como um todo.

É sabido que a etapa de recepção e empastamento foi a que experimentou um dos menores aperfeiçoamentos, e que ela merece muito mais atenção. Um exemplo de uma alternativa atrativa é a automação da recepção e empastamento através do uso de etiquetas de códigos de barra para a identificação de cada questionário da pesquisa e um equipamento de leitura de códigos de barra para verificar o recebimento dos questionários. Entretanto, isso nunca foi tentado, nem mesmo na forma de experimento controlado.

3.4 Entrada de Dados e Codificação

Uma das principais mudanças introduzidas na estratégia global de apuração dos dados das pesquisas foi a eliminação de todas as operações de crítica e codificação manuais realizadas antes da entrada dos dados. Isso foi possível em função do uso de sistemas "inteligentes" de entrada de dados, bem como pelo desenvolvimento e aplicação de sistemas de codificação assistida por computador, o que permitiu que todas as operações de codificação fossem realizadas após os dados terem sido digitados.

Esses dois fatores permitiram a eliminação completa das críticas manuais, as quais eram realizadas juntamente com a codificação manual antes da entrada dos dados, e isso propiciou uma maior agilidade no processamento dos dados, uma vez que os dados agora estão disponíveis no computador muito antes do que ficavam anteriormente. Dessa forma,

as redundâncias nas operações de crítica foram eliminadas e é possível um melhor controle sobre todo o processo de crítica, uma vez que todas as mudanças nos dados podem ser monitoradas facilmente.

O uso de sistemas "inteligentes" de entrada de dados, capazes de lidar com diferentes modelos de questionários em um mesmo lote de trabalho, bem como de aplicar algumas regras de crítica durante o processo de entrada dos dados foi uma das alterações mais significativas na estratégia global de apuração dos dados.

Um bom exemplo foi a operação de entrada de dados do Censo Demográfico de 1991, na qual foram consideradas algumas verificações de identificação e de validade de códigos, evitando assim a possibilidade de introdução de alguma identificação ou código inválidos, como resultado de erros de digitação. A introdução dessas verificações não afetou significativamente a velocidade da operação de entrada de dados, porém auxiliou na redução do número de erros apontados nas etapas subseqüentes de crítica dos dados. Possibilitou também a correção de erros de identificação por meio de etiquetas ou de alocação nos lotes de trabalho, agora detectados na etapa de entrada de dados, erros que no processo anterior poderiam não ser detectados, dificultando sua correção mais tarde.

No caso de pesquisas da área econômica, o fato de juntar em um mesmo lote de trabalho os diferentes modelos de questionários tornou muito mais fácil a etapa seguinte de crítica dos dados para a empresa como um todo, evitando assim a necessidade de procurar em diferentes lotes de trabalho os questionários relativos às diversas unidades de uma mesma empresa.

Outra modificação introduzida no processo de entrada de dados que trouxe melhorias foi a descentralização da operação em 21 escritórios

regionais do IBGE. Isto evitou a concentração de um grande número de trabalhadores de baixos salários em um mesmo local, permitindo assim um melhor gerenciamento desse tipo de equipe. Já foi mencionado também que houve um aumento significativo de produtividade, quando comparadas as taxas de produção das equipes dos escritórios regionais com as da equipe de entrada de dados anteriormente localizada no órgão central.

Entretanto, apesar desse aperfeiçoamento, o problema relativo à metodologia de entrada de dados não foi ainda tratado de forma apropriada, principalmente no que se refere à qualidade do serviço executado. A estratégia adotada para garantir a qualidade permanece a mesma usada anteriormente, qual seja a "verificação a 100%", na qual pouca ou nenhuma atenção é dada para medidas precisas das taxas de erro por operador ou mesmo para a pesquisa como um todo. Esse procedimento foi analisado e criticado por Silva e Bianchini(1992), e um procedimento formal de controle de qualidade por meio de verificação por amostragem foi proposto por Silva, Indá & Lima(1991) para ser usado no Censo Demográfico de 1991, mas não chegou a ser aplicado.

Espera-se que as idéias de controle de qualidade possam ser introduzidas na execução da etapa de entrada de dados, uma vez que ainda há muito a ser aperfeiçoado, tanto em termos de qualidade como em função de análise de custos *versus* benefícios.

Voltando à etapa de codificação das questões em aberto, a principal mudança deu-se através do uso do computador para assistir o operador na atribuição de códigos. O Censo Industrial de 1985 foi a primeira pesquisa do IBGE a usar essa metodologia, tendo sido desenvolvido um sistema de codificação assistida por computador especialmente para ele. Esse sistema consistia de um programa rodando *on-line* que selecionava um código em um arquivo de referência, com base em palavras-chaves

retiradas da descrição do produto que foi informada no questionário. Esse procedimento encontra-se descrito com detalhes em Silva et al.(1986).

O esquema usado para o Censo Industrial de 1985 estava longe de ser o ideal. Ele dependia largamente da habilidade dos operadores de codificação na identificação adequada das palavras-chaves ou combinações delas para subsidiar a procura e a seleção automáticas do código a ser atribuído a cada produto. Entretanto, ele foi o responsável pela busca de novos programas de codificação automática mais sofisticados e eficientes.

Para o Censo Demográfico de 1991, foi desenvolvido um sistema de codificação mais geral, denominado SISCOD, que combina a codificação automática, quando é possível associar um único código às respostas textuais digitadas, com um sistema de codificação assistida por computador (acionado por operadores), para aqueles casos nos quais nenhum ou mais de um código pode ser encontrado para uma dada resposta. Uma descrição do uso bem-sucedido desse sistema, tanto em termos de qualidade como em termos de velocidade, para a codificação dos dados do Censo Demográfico Experimental de Limeira de 1988, pode ser encontrada em Silva(1989a). Silva, Hanono & Barbosa(1993) apresentam uma des-crição do sistema propriamente dito.

Uma desvantagem desse novo sistema de codificação foi a necessidade de digitar a resposta completa dada às questões abertas contidas no questionário da amostra do Censo Demográfico. Mas essa desvantagem foi completamente compensada pela eliminação total de codificação manual, pela alta taxa de codificação automática para a maioria dos campos que necessitava de códigos e pela redução substancial no número de erros devidos a códigos inadequados detectados pelas regras de crítica na etapa posterior de processamento, uma vez que,

por esse sistema, somente códigos válidos podem ser atribuídos seja pelo computador seja pelo operador.

O sistema desenvolvido para o Censo Demográfico é razoavelmente geral, o que permite sua imediata aplicação a outras pesquisas com questões abertas semelhantes, tal como a já mencionada PNAD, que teve seu questionário redesenhado para 1992.

A tarefa principal a ser realizada para a adaptação desse sistema para sua aplicação em uma dada pesquisa é o desenvolvimento de um banco de dados apropriado (um arquivo contendo as descrições textuais possíveis com os códigos correspondentes). Além disso, o desenvolvimento de tal banco de dados tem o efeito positivo de propiciar o aperfeiçoamento da estrutura de códigos para cada questão da pesquisa. Isto contribuirá também para uma maior integração e comparabilidade entre futuras pesquisas, uma vez que uma única estrutura de códigos será usada por todas as pesquisas que investigam questões semelhantes, como é o caso dos códigos de ocupação, investigados tanto no questionário da amostra dos Censos Demográficos como nos questionários das PNADs.

Outra conseqüência positiva da codificação automática é a gravação das descrições textuais ou respostas dadas às questões abertas em arquivos magnéticos, os quais podem ser usados em estudos para o aperfeiçoamento da estrutura de códigos considerada. No esquema anterior de codificação manual isso não era possível, uma vez que as descrições contidas nos questionários não eram digitadas.

Voltando novamente à questão da entrada de dados, outro procedimento testado com sucesso foi a integração entre a operação de crítica de dados com a de entrada de dados, usando *IMPS - Integrated Microcomputer Processing System*, um sistema do *US Bureau of the Census*. Este procedimento consiste em realizar todas as microcríticas

durante a etapa de entrada dos dados, usando uma equipe especializada para realizar essa operação, o que permite a correção de quaisquer erros detectados no questionário quando este está sendo digitado.

Este procedimento mudou a ênfase geral dada à entrada de dados, da velocidade para a qualidade e a integração. O uso experimental do IMPS para processar os cerca de 8 000 questionários do Censo Penitenciário do Rio de Janeiro, realizado em 1988, foi uma boa demonstração das vantagens desse procedimento. A entrada de dados demorou mais tempo que no procedimento anterior, mas o custo e o tempo total de processamento dos dados foram muito menores, uma vez que, após a etapa de entrada dos dados ter terminado, os dados puderam seguir imediatamente para a etapa de tabulação. A equipe envolvida nesse experimento considerou o trabalho muito recompensador, uma vez que esteve com o controle direto e envolvida em todas as etapas de produção dos resultados da pesquisa.

Outros aspectos importantes do experimento acima mencionado foram a proximidade dos técnicos responsáveis pela pesquisa com os dados desde as primeiras etapas do processamento dos dados, e os recursos poderosos de produção e controle de qualidade fornecidos pelo sistema, que em outras situações não estariam disponíveis aos gerentes responsáveis por pesquisas no IBGE. O desenvolvimento de aplicações específicas necessárias para o processamento dos dados foi muito mais fácil e mais rápido, foi realizado principalmente pela equipe de especialistas temáticos do departamento, tendo sido necessária muito pouca assistência externa, evitando assim todos os problemas causados pelo procedimento anterior que usava programas feitos sob medida para o processamento dos dados.

Esse procedimento é visto como o ideal para pesquisas pequenas ou especiais, para as quais o uso do processo descentralizado de entrada de

dados e de crítica não se justifique em função do custo. Um grande problema para a utilização do sistema IMPS como um sistema básico para o processamento das pesquisas no IBGE é o fato que a maioria dos equipamentos instalados nos escritórios regionais trabalha sob o sistema operacional UNIX, e o IMPS foi desenvolvido para máquinas DOS.

Outros procedimentos de entrada de dados também foram examinados no IBGE, visando a sua utilização no Censo Demográfico de 1991. Entre esses, o *FOSDIC - Film Optical Sensing Device for Input to Computers*, desenvolvido e usado pelo *US Bureau of the Census* para o censo americano, leitoras óticas e *scanners* modernas. Entretanto, nenhuma dessas alternativas mostrou-se viável para ser usada no referido censo e, então, o procedimento atual descentralizado de entrada de dados usando operadores de digitação foi adotado.

3.5 Crítica dos Dados

Estamos agora em condições de examinar as principais mudanças na estratégia de crítica e correção de dados usada no IBGE em suas diversas pesquisas, uma vez que já foram apontadas algumas mudanças introduzidas em outras etapas de pesquisa e analisado seu impacto na operação de crítica e correção. Nesta seção estaremos apresentando as principais mudanças, fazendo referências a algumas experiências realizadas com novos procedimentos.

3.5.1 Sistemas Generalizados *Versus* Sistemas sob Medida

Começamos com a idéia de usar sistemas generalizados ao invés de sistemas/programas feitos sob medida para realizar as tarefas de crítica de dados. Como mencionado anteriormente, até 1985, mais especificamente, os Censos Econômicos de 1985, muitas das etapas de uma pesquisa (a crítica de dados aí incluída) eram realizadas usando sistemas feitos sob medida e os problemas com esse tipo de procedimento já foram apontados. Atualmente, é fortemente recomendado que sejam usados sistemas generalizados sempre que possível, ao invés de desenvolver sistemas específicos para cada pesquisa.

Devido a essa mudança de estratégia, nós podemos apontar um número crescente de aplicações bem-sucedidas de sistemas generalizados de crítica de dados, tais como *DIA*[2] - *Deteccion e Imputación Automática de errores para datos cualitativos*, desenvolvido pelo *INE - Instituto Nacional de Estadística da* Espanha, *IMPS - Integrated Microcomputer Processing System*, desenvolvido pelo *US Bureau of the Census* e o sistema CRIPTA[3] - Crítica, Imputação e Tabulação de dados, desenvolvido no próprio IBGE. Simultaneamente, foi observado um aumento no uso de ferramentas de análise de dados baseadas em sistemas estatísticos gerais tal como o *SAS - Statistical Analysis System*, durante as operações de crítica.

A idéia principal em que se baseia o uso de sistemas generalizados para desenvolver as aplicações de crítica de uma pesquisa específica é reduzir o custo desse desenvolvimento, bem como reduzir o tempo de tal tarefa. Antes de adotar essa estratégia de usar sistemas generalizados

[2] Ver Rubio & Criado (1988).
[3] Ver Hanono & Barbosa (1992).

para crítica de dados, foi realizada uma revisão bastante abrangente das alternativas disponíveis. Nessa revisão, foram examinados os seguintes sistemas: *CONCOR - Consistency and Correction*, do *US Bureau of the Census*, na sua versão para computadores de grande porte, OSIRIS IV da Universidade de Michigan, *UNEDIT - United Nations EDIT*, *EDIT78* do *Statistics Sweden*, *CANEDIT* e *SPIDER* do *Statistics Canada.*[4]

Esta revisão forneceu as bases para duas decisões importantes no que se refere ao uso de sistemas generalizados. A primeira, que o IBGE deveria desenvolver seus próprios sistemas de crítica generalizados, para lidar com tarefas básicas de microcrítica. Essa decisão se deu em função de não ter sido encontrado, entre os sistemas de crítica analisados, nenhum sistema suficientemente portátil e amigável que permitisse sua utilização direta pelos especialistas temáticos nem algum que pudesse servir de padrão para o desenvolvimento de aplicações de crítica das pesquisas. A segunda decisão refere-se às questões mais complexas ou especializadas de crítica. Nesses casos, os sistemas generalizados disponíveis no IBGE, com garantia de suporte, deveriam ser usados sempre que possível, evitando assim a necessidade de desenvolvimento de sistemas sob medida.

A primeira dessas decisões levou ao desenvolvimento e aplicação do sistema CRIPTA, no qual agora nos concentramos. CRIPTA é definido pelo seus autores como um "gerador automático de aplicações de crítica e imputações", conforme consta em Hanono & Barbosa(1992). O desenvolvimento do CRIPTA baseia-se na idéia de automação do desenvolvimento de aplicações de crítica para pesquisas específicas, na tentativa de reduzir ou mesmo eliminar a necessidade de intervenção dos especialistas em processamento de dados no desenvolvimento de tais aplicações.

---

[4] Ver Barbosa & Hanono (1988).

Duas das principais características impostas para o desenvolvimento do CRIPTA foram que ele deveria ser portátil e amigável. Ele possibilita que especialistas temáticos com pouco conhecimento de computação descrevam seus dados, por meio de um dicionário padronizado, e também especifiquem as regras de crítica com as quais os dados devem ser verificados. A especificação das regras de crítica é feita usando uma pseudolinguagem em Português, o que facilita bastante a comunicação entre os especialistas da pesquisa e os de processamento de dados, quando necessário.

Algumas outras vantagens do uso do CRIPTA são a imediata verificação da consistência entre as especificações de crítica e a descrição dos dados (dicionário dos dados), a automação do desenvolvimento das aplicações, e o fato de que os especialistas temáticos são os responsáveis diretos pela geração da maioria das aplicações de crítica, evitando assim os demorados ciclos de especificação do sistema, desenvolvimento e testes, citados anteriormente. O sistema CRIPTA permite o uso de rotinas especiais necessárias para a realização de algumas tarefas de apuração e crítica não usuais ou não padronizadas, que podem ser desenvolvidas pela equipe de processamento de dados usando linguagens tais como C ou PL/I. Ele pode lidar com dados categóricos ou numéricos, e permite a utilização de arquivos de dados de quase todos os tipos e de complexidade variada. Além disso, ele pode recuperar dados diretamente de banco de dados relacionais armazenados sob DB2, por exemplo.

Entretanto, há algumas desvantagens que devem ser mencionadas. O sistema CRIPTA não incorpora qualquer forma de análise lógica das regras de crítica, ou seja, ele não garante que só sejam aplicados conjuntos de regras de crítica consistentes[5]. Além disso, apesar de oferecer facilidades para o desenvolvimento de imputação *hot-deck* ou determinística, essas facilidades não são baseadas em metodologia

[5] Por exemplo, no sentido definido por Fellegi & Holt (1976).

estatística de imputação dentre as mais amplamente aceitas. Essas limitações são conseqüência da decisão de não incluir, nos sistemas desenvolvidos pelo IBGE, pelo menos nas versões iniciais, as facilidades e os tópicos já abrangidos por outros sistemas generalizados de crítica e imputação de dados disponíveis no IBGE.

Pelo contrário, a idéia era que o CRIPTA deveria abranger o que não estava disponível em outros sistemas, de tal forma a ser usado em combinação com tais sistemas, em uma dada pesquisa, quando necessário.

Dois exemplos de pesquisas que usaram o sistema CRIPTA para desenvolver parte substancial das aplicações de crítica de que necessitava são a Pesquisa Anual do Comércio - PAC, relativa aos anos de 1988, 1989, 1990 e 1991, e a Pesquisa Nacional por Amostra de Domicílios - PNAD -, em sua edição revisada de 1992.

A política atual do IBGE, no que diz respeito a sistemas de crítica de dados, estabelece claramente que cada nova pesquisa ou cada pesquisa que passe por qualquer redesenvolvimento importante deverá adotar o CRIPTA como o instrumento básico para a geração das aplicações de crítica necessárias. Outros sistemas generalizados poderão ser usados complementarmente ao CRIPTA, mas programas sob medida deverão ser evitados.

Como já mencionado, outros sistemas generalizados de crítica de dados tais como DIA e IMPS são atualmente utilizados na implementação de operações de crítica de dados das pesquisas realizadas pelo IBGE. O desenvolvimento de sistemas generalizados para crítica e imputação de dados, tais como DIA, CANEDIT e, mais recentemente, *Generalized Edit and Imputation System*[6] - *GEIS,* do *Statistics Canada*, e *Design, Analysis*

[6] Ver Kovar & Whitridge (1990).

and Imputation System[7] *- DAISY,* do *Istituto Nazionale di Statistica da Itália, - ISTAT,* só foi possível devido ao desenvolvimento metodológico contido no artigo de Fellegi & Holt(1976). Esse artigo estabeleceu tanto os princípios largamente aceitos que um sistema para crítica e imputação automáticas de dados de pesquisas deve satisfazer, bem como uma metodologia matemática confiável baseada em tais princípios, a qual permitiu o desenvolvimento de todos esses sistemas generalizados.

Entre as principais vantagens dos sistemas generalizados de crítica de dados, destaca-se que eles fornecem uma integração lógica entre a detecção dos erros e a imputação, de forma que os indesejáveis ciclos de crítica de dados podem ser evitados. Além disso, eles preservam tanto quanto possível a distribuição dos dados bons, e entre as estratégias de imputação possíveis escolhem aquela que modifica o menor número de campos em cada registro, garantindo ainda que o registro resultante da imputação passa livremente por todas as regras de crítica especificadas. Esses sistemas têm capacidade para processar grandes pesquisas ou censos, fornecem um modelo explícito de imputação e requerem como entrada um conjunto de regras de crítica, definidas pelos especialistas temáticos.

Quando tais sistemas generalizados são usados, o controle e o acompanhamento da imputação são feitos muito mais facilmente, uma vez que eles automaticamente geram uma série de estatísticas com o objetivo de fornecer aos especialistas temáticos informações completas sobre a aplicação e sobre a extensão da modificação nos dados. Eles permitem que os especialistas temáticos concentrem atenção nas etapas de especificação das regras de crítica e no controle e análise da imputação efetuada, e não mais no desenvolvimento e testes de sistemas para o processamento da pesquisa.

---

[7] Ver Barcaroli (1992a,b).

Algumas desvantagens devem ser mencionadas. Em geral, tais sistemas não são desenvolvidos para serem aplicados em pesquisas pequenas (pelo menos esse é o caso do DIA), e para serem usados requerem computadores de grande porte (*mainframe*), os quais não são tão amigáveis como os microcomputadores disponíveis atualmente. Entretanto, os ganhos em velocidade e o custo reduzido de desenvolvimento das aplicações excedem em muito essas desvantagens em potencial.

Uma aplicação de sistema generalizado que merece um exame mais detalhado é o Censo Demográfico de 1991, que está usando o DIA para a etapa de crítica e imputação automáticas. Isto representa o principal avanço no uso de sistemas generalizados no IBGE.

Deve ser mencionado também que o Censo Demográfico de 1980 foi a primeira pesquisa importante a utilizar imputação automática no IBGE, mas ela foi realizada com programas feitos sob medida. Apesar da utilização de imputação automática ter sido considerada bem-sucedida, o esquema adotado foi profundamente complexo e custoso, porque, para cada regra de crítica especificada pelos especialistas temáticos, uma série de regras de imputação apropriadas tiveram que ser definidas a fim de indicar como um determinado registro com erro deveria ser "corrigido". Essas regras de imputação podiam ser determinísticas ou probabilísticas, e nesse último caso, o método *hot-deck* seqüencial foi usado para selecionar um doador adequado, cujos dados eram usados para substituir os valores considerados inaceitáveis no registro com erro.

Silva et al.(1990) e IBGE(1990) apresentam comparações dos resultados entre a aplicação do DIA e a utilização do esquema adotado no Censo Demográfico de 1980, na crítica e imputação automáticas dos dados do Censo Demográfico Experimental de Limeira de 1988.

O sistema DIA foi considerado também para aplicação na crítica dos dados categóricos do Censo Agropecuário. Uma aplicação experimental a um subconjunto de dados do Censo Agropecuário de 1985, relatado em Farias & Dias(1992), mostrou-se bem sucedida e vantajosa se comparada com o esquema anteriormente adotado.

Além disso, o IBGE obteve uma cópia do GEIS, que pode ser considerado como uma ferramenta complementar em potencial para ser introduzida na crítica e imputação de algumas pesquisas econômicas. Entretanto, uma aplicação experimental do GEIS aos dados do Censo Agropecuário de 1985 foi interrompida quando o Censo Agropecuário de 1991 foi cancelado.

3.5.2 Métodos Estatísticos ao Invés de "Contabilidade"

Uma outra grande modificação nos procedimentos de crítica e apuração foi particularmente importante para as pesquisas econômicas. No IBGE, a etapa de crítica da maioria das pesquisas econômicas consistia, de modo geral, na aplicação de um conjunto de regras de crítica pré-definidas. Essas regras geralmente representavam a estrutura interna dos dados coletados na forma de "relações contábeis", tais como totais que devem se igualar à soma das parcelas, equações de "fechamento" e outras.

Essas críticas do tipo "relações contábeis" eram aplicadas a todos os questionários coletados em uma dada pesquisa, independentemente do tamanho da empresa ou do estabelecimento. Além disso, para uma regra não satisfeita por um certo registro, o tratamento aplicado para corrigir a inconsistência detectada era o mesmo independentemente da magnitude do erro.

Um bom exemplo desse procedimento é descrito a seguir. Quando, em um dado questionário (registro), era encontrada uma diferença entre um total calculado e a soma de suas correspondentes parcelas, não importava o tamanho da diferença (se 1 ou 1 milhão), o registro recebia o mesmo tratamento: ele era impresso a fim de permitir a verificação manual por operadores, os quais tentavam resolver o problema através de novo contato com o respondente (apenas em um número limitado de casos) ou definindo quais campos necessitavam de correção e os tipos de correção que eles necessitavam. Uma vez que fazer um novo contato com os respondentes era uma operação cara, a decisão tomada pelos operadores era, em geral, a escolha de algum(ns) campo(s) para uma imputação manual, de tal forma que a inconsistência detectada fosse eliminada.

Entretanto, devido à complexidade causada pelo grande número de regras de crítica que eram aplicadas a cada pesquisa, freqüentemente uma correção efetuada provocava que esse mesmo registro fosse novamente apontado para exame por falhar alguma outra regra de crítica. Isto implicava que cada lote de registros (questionários) tinha que ser submetido ao mesmo programa de crítica várias vezes, causando os ciclos de operação anteriormente citados.

Uma vez que o número de tais regras de crítica para uma dada pesquisa era geralmente muito grande, os programas de crítica eram complexos e a proporção de questionários que falhavam pelo menos uma regra de crítica era muito alta. Acrescentando-se a isso a natureza cíclica do procedimento adotado e o fato de que cada pesquisa compreendia ainda um número de operações de crítica separadas (algumas delas redundantes), obtém-se uma explicação para o alto custo e o longo tempo que geralmente era gasto para completar a etapa de apuração e crítica de uma dada pesquisa.

A idéia por trás desse procedimento era que estatísticas econômicas, as quais se baseiam fortemente em registros contábeis das unidades investigadas, devem ser baseadas em registros individuais, os quais devem satisfazer exatamente as relações contábeis desejadas. A idéia pode parecer atrativa, porque, se todos os registros individuais da pesquisa satisfazem todas as regras de crítica, então as medidas agregadas calculadas a partir deles também vão satisfazer as mesmas regras de crítica no nível agregado. Entretanto, isto não implica que todos os erros encontrados em registros individuais devam ser tratados da mesma forma, como era o caso. Por outro lado, pode-se argumentar que as estatísticas econômicas no nível agregado seriam muito pouco afetadas por pequenos erros encontrados em algum registro individual.

Esses dois argumentos forneceram as bases para a introdução de métodos estatísticos para a crítica de pesquisas econômicas. Os métodos estatísticos forneceram critérios objetivos para selecionar o tratamento a ser aplicado em registros falhos em pesquisas econômicas, de tal forma que erros pequenos em registros individuais fossem tratados por meio de métodos menos custosos sem afetar negativamente as estatísticas agregadas.

A primeira aplicação em larga escala no IBGE de métodos estatísticos na etapa de crítica de dados foi realizada nos Censos Econômicos de 1985 (Indústria, Comércio e Serviços) com o objetivo de acelerar o processo. Pinheiro & Assunção(1988) descrevem um método baseado no uso de críticas de razão e em "curvas de rejeição", as quais foram utilizadas em substituição a um grande número de regras de crítica do tipo contábil.

A idéia básica desse procedimento era permitir que fossem aplicados tratamentos diferentes para os erros detectados por uma mesma regra de crítica. Dependendo do tamanho do erro, do tamanho do registro

correspondente e de restrições administrativas definidas pelos responsáveis pelos Censos, cada registro marcado como falho poderia ser submetido ou a uma imputação determinística automática ou ser impresso para uma revisão manual pelos operadores de crítica, os quais deveriam então providenciar o novo contato com alguns dos respondentes.

Em Pinheiro & Assunção(1988), pode ser encontrada uma descrição do procedimento adotado para uma regra de crítica padrão, mais precisamente uma regra que trata de níveis de estoque de bens em um estabelecimento industrial. O mesmo tipo de procedimento foi adotado em substituição a um certo número de equações de crítica similares. Como exemplo, apresentamos a seguir uma dessas equações. Sejam:

A - valor dos bens consumidos durante o ano do censo;
B - valor dos bens comprados (adquiridos) durante o ano do censo;
C - valor dos bens recebidos por transferência durante o ano do censo;
D - valor dos bens em estoque no último dia do ano anterior;
E - valor dos bens em estoque no último dia do ano do censo; e
F - valor dos bens transferidos para fora durante o ano do censo.

Uma equação básica relacionando as variáveis acima descritas estabelece que o estoque final (E) deve se igualar ao estoque inicial (D) mais as entradas (B, C) menos o consumo e as transferências (A, F). Isto é, o que se impõe como uma equação de crítica é:

$$E = D + B + C - A - F$$

ou equivalentemente

$$A = B + C + D - E - F \qquad (1)$$

Em censos e pesquisas anteriores, essa equação foi usada para gerar uma regra que, para qualquer registro para o qual ela não era válida, implicava um relatório impresso por computador e uma revisão manual do

questionário correspondente, independentemente do tamanho da diferença detectada.

A metodologia proposta por Pinheiro & Assunção(1988) consistiu em definir uma razão a partir do conteúdo dos dois lados da equação (1), a saber:

$$R = \frac{A}{B + C + D - E - F} \qquad (2)$$

de tal forma que se a razão for 1 o registro passa pela crítica, e caso contrário o registro é rejeitado.

Entretanto, ao invés de adotar o procedimento anterior de utilizar um único tratamento para os registros que não passaram pela regra de crítica, eles definiram um procedimento bastante interessante para selecionar apenas os registros que necessariamente deveriam passar pela análise dos operadores de crítica. Para tanto, foi definida uma região de aceitação para R mais especificamente para $R^* = |\log(R)|$ uma transformação de R adotada a fim de retirar a assimetria da distribuição e fornecer uma medida relativa de discrepância não-negativa e invariante[8].

É facil notar que $R^* = 0$ quando os dados estão "corretos" e que $R^* > 0$ em caso contrário. O procedimento consistia em calcular um número r tal que se $R^* > r$ então o questionário era marcado para ser analisado pelos operadores de crítica. Entretanto, eles perceberam que apesar de fornecer uma boa medida da discrepância relativa entre os dois lados da equação (1), a razão modificada R^* não incorporava a indicação do tamanho do estabelecimento informante em questão. Por exemplo, uma diferença de 20% podia ser considerada aceitável para ser corrigida

[8] Invariante no que se refere à troca entre o numerador e o denominador.

por meio de imputação para estabelecimentos pequenos, mas podia ser considerada muito grande para estabelecimentos muito grandes.

Então, eles sentiram a necessidade de determinar limites de rejeição diferenciados de acordo com os diferentes tamanhos de estabelecimentos. Isso originou a idéia de "curvas de rejeição", pelas quais o limite de rejeição r_t era então determinado como uma função do tamanho t do estabelecimento informante em questão, sendo t uma variável independente daquelas usadas para definir R^*. A variável usada para medir o tamanho foi então definida como o número de empregados do estabelecimento.

Usando os dados do censo anterior (1980), Pinheiro e Assunção usaram um procedimento engenhoso para estimar os parâmetros necessários para uma especificação detalhada das curvas de rejeição propostas. Foi então possível preparar programas de computador para a crítica e imputação de alguns dos campos do questionário para os quais esse tipo de crítica de razão se aplicava antes mesmo que os dados do censo de 1985 estivessem disponíveis. Isto foi muito importante, pois as pessoas contrárias ao uso de métodos estatísticos na crítica de dados geralmente argumentavam que tais métodos só poderiam ser aplicados após os dados terem sido coletados e estarem disponíveis para análise, o que acarretaria em muito pouco tempo para o desenvolvimento e teste dos programas a serem aplicados em uma operação grande tal como é a dos censos econômicos.

Outro aspecto interessante dessa metodologia é que a proporção esperada de questionários a serem "rejeitados" pelo procedimento para tratamento manual pode ser especificada como um parâmetro. Durante o cálculo das curvas de rejeição para cada regra de crítica, todas as críticas de razão a serem aplicadas foram consideradas simultaneamente de tal forma que puderam ser estimadas taxas gerais de rejeição. Isto deu aos

responsáveis pela pesquisa um grande controle sobre a carga de trabalho esperada na operação de crítica e ao mesmo tempo sobre a qualidade dos dados agregados produzidos. Tal controle não era possível sob o procedimento anteriormente adotado.

Essa aplicação bem-sucedida de idéias estatísticas para a crítica de uma pesquisa grande, como é o caso do Censo Industrial de 1985, com mais de 200 000 informantes, criou a oportunidade para muitos outros desenvolvimentos interessantes e análogos no âmbito das pesquisas econômicas do IBGE. Um exemplo de desenvolvimento posterior a essa metodologia desenvolvida por Pinheiro & Assunção(1988) pode ser encontrada em Pinheiro & Assunção(1989) e Silva & Ribeiro(1990), que descrevem a crítica dos dados a nível de empresas dos Censos Econômicos de 1985.

O novo procedimento sugerido por Pinheiro & Assunção(1989) considerou a quantidade $T_i(\beta)$ como uma estatística de detecção para diferenças entre quaisquer dois valores A_i e B_i, os quais deveriam ser iguais ou pelo menos muito próximos nos registros "bons" de uma pesquisa, onde:

$$T_i(\beta) = S_i^{\beta} \cdot R_i^* \qquad (3)$$

com

$S_i = |A_i - B_i|$;

$R^* = |\log(A_i) - \log(B_i)|$ como anteriormente; e

β = parâmetro a ser fixado para controlar a importância do erro absoluto na estatística de teste.

É fácil notar que S_i mede a discrepância absoluta entre A_i e B_i, enquanto que R_i^* mede a discrepância relativa. Então, $T_i(\beta)$ fornece uma medida combinada das discrepâncias absoluta e relativa entre os valores

A_i e B_i, que tem por objetivo incorporar a medida de tamanho do erro na estatística de detecção. Quando comparada com o procedimento anterior de curvas de rejeição, a equação (3) é um aperfeiçoamento, uma vez que não é necessária nenhuma medida externa de tamanho.

O uso de $T_i(\beta)$ como uma estatística de detecção depende da estimação do parâmetro β, bem como do cálculo dos pontos de corte. Mais uma vez, os autores definiram métodos bastante interessantes para a determinação dessas quantidades, de tal forma que os responsáveis pelo Censo podiam ter controle sobre a proporção esperada de questionários a ser tratada manualmente e ao mesmo tempo sobre a discrepância final esperada entre os totais $A = \sum A_i$ e $A = \sum B_i$. Mais tarde, verificou-se que essa estatística de detecção era muito parecida com a estatística proposta por Hidiroglou & Berthelot(1986) para a crítica de pesquisas econômicas periódicas, embora ela tenha sido desenvolvida de forma independente.

Ao mesmo tempo, outro procedimento para a crítica de dados de pesquisas econômicas utilizando métodos estatísticos estava sendo estudado. Embora os procedimentos desenvolvidos por Pinheiro & Assunção para serem usados nos Censos Econômicos de 1985 fossem considerados muito bem-sucedidos, eles ainda eram bastante restritivos no sentido que dependiam fortemente de informações prévias de alguma pesquisa semelhante para serem implementados. A pesquisa para a qual o processo de crítica estava sendo desenvolvido não contribuía com dados para a definição das equações de crítica (ou para as estatísticas de detecção) e dos limites de rejeição correspondentes. Sentia-se a necessidade de alguma metodologia que permitisse que os dados da pesquisa "falassem por si mesmos", isto é, ter menos regras de crítica ou limites prefixados para a aceitação dos dados.

Silva (1989b) fornece uma descrição detalhada de um procedimento que considera esse objetivo. Baseado em uma adaptação de procedimentos originalmente sugeridos por Little & Smith(1987) e que se encontram analisados em Bustos & Silva(1988), Silva(1989b) desenvolveu um conjunto de rotinas SAS[9], que foi denominado CIDAQ[10]. As rotinas do CIDAQ fornecem uma alternativa aos procedimentos anteriores para a crítica estatística de dados, que considera a natureza multivariada dos dados das pesquisas (econômicas).

A idéia básica da metodologia CIDAQ consiste em substituir a série de críticas de razão anteriormente usadas, do tipo $L \leq R_i \leq U$, por relações multivariadas entre as variáveis envolvidas nas razões, de tal forma que essas relações são estimadas a partir dos dados. Essa metodologia fornece também meios de determinar as regiões de aceitação a partir dos dados, evitando assim a necessidade de pré-definir limites de aceitação (L e U). A metodologia CIDAQ contém ainda um procedimento de imputação automática de dados ausentes ou rejeitados, baseado em métodos de regressão, de tal forma que os registros porventura imputados não são rejeitados por uma nova aplicação da etapa de detecção de erros, se ela for aplicada após a imputação ter sido realizada. Isto implica que a detecção e a correção automática estão totalmente integradas na metodologia CIDAQ.

Entre as possibilidades fornecidas pelas rotinas CIDAQ há métodos gráficos e testes automáticos para detecção de valores suspeitos (*outliers*), um algoritmo robusto para estimar médias e matrizes de covariância de dados multivariados (possivelmente incompletos) e procedimentos iterativos para localização de erros entre os registros rejeitados por terem valores suspeitos. O diagrama da Figura 2 fornece uma rápida descrição

[9] SAS - Statistical Analysis System.
[10] CIDAQ - Crítica e Imputação de Dados Quantitativos.

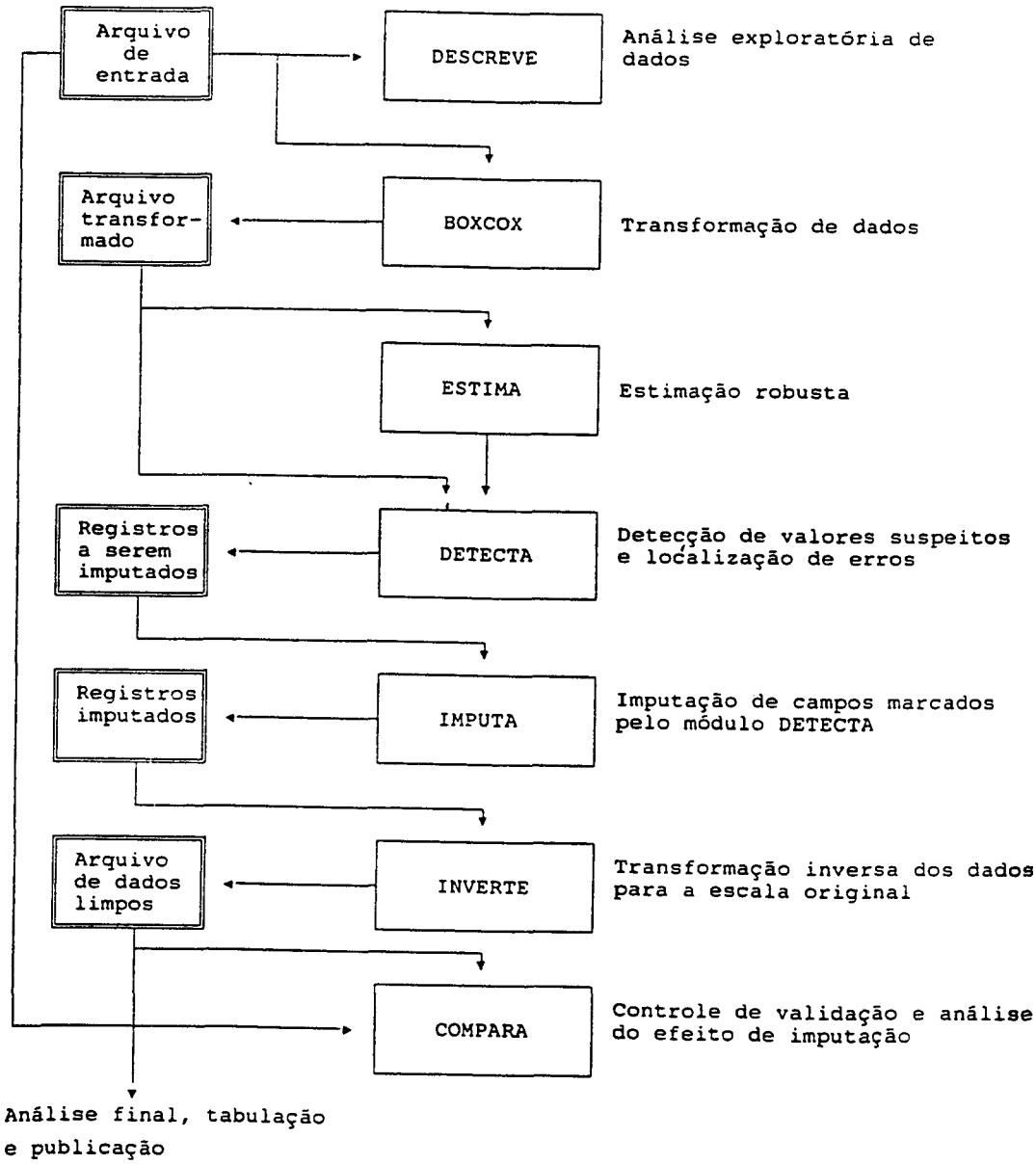
dos módulos CIDAQ e como eles podem ser usados para uma dada operação de crítica dos dados de uma pesquisa.

Um descrição mais detalhada da metodologia CIDAQ está fora dos objetivos deste trabalho, uma vez que ela está baseada em métodos estatísticos razoavelmente sofisticados de análise multivariada e de inferência. Apesar da complexidade apontada, as rotinas CIDAQ são muito simples de serem usadas uma vez que os dados são lidos através do pacote estatístico SAS e mesmo usuários com pouca experiência em computação podem manuseá-las para uma dada operação de crítica de uma pesquisa.

Isto foi de fato observado, uma vez que várias aplicações diferentes foram feitas no IBGE. Nas Pesquisas Anuais do Comércio de 1988 a 1990, a metodologia CIDAQ foi utilizada para a detecção de valores suspeitos e como um procedimento para selecionar questionários para novo contato com o informante, após algumas verificações iniciais de estruturas terem sido realizadas. As variáveis mais importantes da pesquisa foram analisadas usando CIDAQ, em uma operação de crítica que substituiu completamente o procedimento anteriormente usado de críticas de razão.

Figura 2

Breve descrição dos módulos do CIDAQ



A aplicação da metodologia CIDAQ nessas pesquisas foi considerada bem-sucedida tanto pelos especialistas temáticos como pelos estatísticos e responsáveis pelas pesquisas, devido ao baixo custo de implementação e à eficácia na detecção de erros. Entretanto, a imputação automática não foi realizada embora estivesse disponível na metodologia CIDAQ. Os erros detectados usando a metodologia CIDAQ foram ainda tratados pelos operadores e pelo novo contato aos respondentes. Entretanto, uma pequena proporção de registros foram analisados, pois apenas os registros com grandes desvios do padrão esperado foram marcados pela metodologia CIDAQ para revisão. Isto tornou possível para os operadores e especialistas temáticos dedicar atenção especial para os erros relevantes encontrados nos dados.

A metodologia CIDAQ teve uma outra grande aplicação durante a chamada segunda fase do sistema de crítica dos dados da Pesquisa Industrial Anual - PIA para os anos de 1986 a 1990. Santis & Souza(1991) descrevem como a metodologia CIDAQ foi usada para realizar a detecção de valores suspeitos nos dados provenientes da PIA nesses anos. A idéia básica era examinar as variáveis mais importantes da pesquisa para qualquer desvio significante do padrão multivariado esperado. Entretanto, nessa pesquisa, a metodologia CIDAQ fez parte da segunda fase da operação de crítica, que foi realizada somente após a operação de crítica convencional bastante pesada ter sido realizada em cada lote de questionários da pesquisa.

A primeira fase da operação de crítica consistiu de um grande número de críticas aplicadas na forma tradicional cíclica a cada registro individual da pesquisa (microcrítica). Cerca de 500 críticas foram aplicadas aos 4 tipos de questionários usados na pesquisa durante a primeira fase da crítica, o que evidencia um sistema de crítica bastante complexo. Embora essa primeira fase da crítica tivesse sido bastante pesada, erros relevantes

ainda foram detectados pela metodologia CIDAQ e pelas análises de tabulações preliminares durante a segunda fase da operação de crítica.

Motivados pelo atraso significante na publicação dos resultados da pesquisa para alguns dos anos no período 86/91, Santis & Damião desenvolveram um estudo para avaliar o impacto da primeira fase da crítica sobre os dados. Essa análise consistiu na comparação dos dados "sujos", que foram gravados em fitas antes dos dados serem submetidos à operação complexa de crítica, com os correspondentes dados "limpos" que resultaram da primeira fase da operação de crítica. As autoras calcularam diferenças entre totais para algumas das variáveis mais importantes usando os dados sujos e os limpos, ao nível de codificação a 2-dígitos da Classificação Industrial - *Standard Industrial Classification*[11] - *SIC*. Os resultados mostraram forte evidência de que o impacto da crítica pesada foi desprezível na grande maioria dos casos. Esse resultado encorajou as autoras a pensarem numa revisão nos procedimentos de crítica para as pesquisas futuras, de forma que a primeira fase de crítica mais pesada possa ser eliminada ou substituída por um método mais eficaz.

Um terceiro procedimento mereceu atenção para a crítica de dados de pesquisas econômicas no IBGE. Esse novo procedimento baseia-se nas idéias do método *top-down* para macrocríticas propostas por Granquist(1987), embora não totalmente idênticas. O principal exemplo vem da Pesquisa Industrial Mensal - Dados Gerais - PIM-DG, que investiga 8 variáveis relacionadas com emprego, rendimento e valor da produção e que foi redesenhada em 1985/86.

Esta pesquisa tem por objetivo estimar mudanças mensais no total do emprego industrial, dos rendimentos e do valor da produção. Dessa forma, é muito importante analisar as mudanças em estabelecimentos individuais que podem afetar as estimativas gerais de variação a um dado nível de agregação. O método adotado para isso foi calcular o chamado

[11] No Brasil, os resultados são publicados pelo menos ao nível de 2 dígitos da *Standard Industrial Classification - SIC*.

escore de influência para cada respondente, para cada variável da pesquisa. Uma descrição de como esses escores de influência são calculados pode ser encontrada em IBGE(1988).

Os respondentes com os maiores escores de influência para cada variável são listados por um programa de computador para uma análise pelos especialistas temáticos e para eventualmente um novo contato. O sistema usado não é tão sofisticado como o descrito em Granquist(1987), uma vez que ele não fornece facilidades *on-line* para atualização dos registros e recálculo dos escores de influência. Entretanto, mesmo com um programa que é operado na forma tradicional de processamento cíclico dos lotes de questionários, este procedimento é bastante eficaz na produção de conjuntos de dados "limpos" (de qualidade aceitável) após apenas um ou dois ciclos. Ele também é apropriado para pesquisas mensais, devido ao seu ritmo nervoso e pelo fato que somente uma pequena proporção dos dados deve ser analisada.

Uma desvantagem desse procedimento é que as variáveis da pesquisa são tratadas independentemente, o que o torna mais útil para pesquisas com poucas variáveis. Para pesquisas mais complexas com um grande número de variáveis investigadas, esse procedimento pode não ser aplicável. Entretanto, esse fato não tornou impossível sua aplicação bem sucedida aos dados da PIM-DG. Desenvolvimentos futuros desse procedimento no IBGE necessitarão de sistemas mais sofisticados capazes de uma interação *on-line* com os especialistas temáticos que realizam a crítica dos dados.

3.5.3 Outras Mudanças no Processo de Crítica

Muitas outras mudanças foram propostas ou introduzidas no sistema de processamento das pesquisas, as quais podem afetar também a forma como a crítica dos dados é realizada no IBGE. Essa seção tem por objetivo apresentar algumas dessas mudanças.

Atualmente é largamente reconhecido que a forma de processamento em lotes de questionários e a natureza cíclica do processo de crítica são as principais causas do seu alto custo e de sua longa duração. Além dos aperfeiçoamentos metodológicos que foram descritos na seção anterior, foram incorporados também alguns desenvolvimentos na automação do processo.

Atualmente é rotina gravar os dados em discos ao invés de em fitas magnéticas. Tais dados são também gravados sob um sistema gerenciador de banco de dados relacionais (DB2 no órgão central, e TSGBD[12] nas unidades descentralizadas), o que permite um acesso muito mais conveniente aos dados durante o processamento da pesquisa. Isto também significa facilidades *on-line* que podem ser usadas para atualização dos dados durante as operações de crítica. Combinadas com a facilidade fornecida pelo sistema CRIPTA que permite que o mesmo programa de crítica seja executado para cada registro atualizado da pesquisa, isto significa uma forte redução nos custos e no tempo gasto para realizar a mesma crítica para uma dada pesquisa do que antes.

Duas grandes pesquisas realizadas no IBGE, para as quais o sistema de processamento foi redesenvolvido recentemente, adotam atualmente o procedimento descrito acima. A primeira delas é a Pesquisa Industrial Mensal - Produção Física - PIM-PF que investiga mensalmente o volume de produção. Ela envolve a coleta de dados de cerca de 700 produtos industriais de cerca de 6 000 estabelecimentos industriais a cada mês. O sistema de processamento de dados dessa pesquisa foi completamente redesenvolvido, fazendo uso intensivo das facilidades de gravação e recuperação de dados fornecidas pelo DB2.

A outra pesquisa que adotou parcialmente o procedimento acima descrito é o Sistema Nacional de Índices de Preços ao Consumidor - SNIPC que calcula índices de preços mensais para 11 áreas metropolitanas, baseadas em cerca de 200 000 preços coletados e tratados a cada

[12] Tecnocoop Sistema Gerenciador de Banco de Dados.

mês. Os programas de crítica ainda não foram redesenvolvidos em CRIPTA, o que significa que ainda estão sendo usados programas feitos sob medida para a crítica, mas atualmente o cadastro está gravado sob DB2, o que significa que a atualização de registros é feita usando DB2 e as facilidades de interação entre usuário e máquina fornecidas por sua linguagem.

A descentralização do processamento dos dados é outro fator importante por trás das mudanças na operação de crítica de muitas pesquisas. Isto envolve a capacitação dos escritórios regionais com as facilidades necessárias para realizar um número de operações que eram previamente realizadas de forma centralizada. Os escritórios regionais são atualmente responsáveis por todas as atividades de entrada de dados, bem como por uma grande quantidade de microcríticas.

Em algumas pesquisas, a descentralização foi implantada mantendo o mesmo procedimento para crítica de dados que era usado no órgão central. Entretanto, em alguns casos o procedimento de crítica foi modificado a fim de aproveitar ao máximo a descentralização. O Censo Demográfico de 1991 é um bom exemplo disso. Nesse caso, algumas críticas foram realizadas durante a etapa de entrada de dados e todas as microcríticas envolvendo correção manual foram completamente realizadas nos escritórios regionais. A imputação automática foi realizada no órgão central para a "correção" das inconsistências remanescentes e para produzir registros "limpos" para a etapa subseqüente de tabulação e publicação dos resultados.

O procedimento de crítica adotado no Censo Demográfico de 1991 fornece também um exemplo de redução no número de operações de crítica realizadas e de eliminação da natureza cíclica em pelo menos uma das operações realizadas. Essa é outra estratégia para reduzir custos e o tempo de duração das operações de crítica.

Outras iniciativas nesse sentido incluem:

a) a criação de um cadastro central para subsidiar todas as pesquisas econômicas, o que certamente afetará a forma de cada pesquisa lidar com as críticas dos dados de identificação e de classificação;

b) a criação de banco de metadados, para armazenar as descrições dos dados das pesquisas, incluindo dicionários dos dados, classificações, arquivos de códigos e até mesmo especificações de crítica; espera-se que no futuro esse banco de metadados inclua todas as pesquisas realizadas no IBGE, inclusive aquelas realizadas no passado para as quais os microdados estão atualmente disponíveis em meio magnético; e

c) discussões sobre um esquema genérico que identifique os equipamentos (*hardware*) e os sistemas (*software*) necessários para as aplicações relativas ao processamento dos dados das pesquisas; para uma descrição das idéias atualmente em discussão, ver Cabral(1990, 1991) e Anzanello (1993); embora o foco dessa discussão ainda esteja voltado para os equipamentos e sistemas, no futuro ela certamente afetará os procedimentos de crítica, uma vez que atualmente eles são altamente dependentes da automação e no futuro serão ainda mais.

4 PERSPECTIVAS FUTURAS E CONCLUSÕES

No capítulo anterior foram apresentadas diversas mudanças introduzidas na estratégia adotada no IBGE para o processamento de pesquisas, e particularmente para a crítica de dados. Deve-se reconhecer que, em comparação com a situação existente no final dos anos oitenta, foram alcançados avanços significativos. Entretanto, o IBGE ainda está longe de uma solução apropriada para a crítica de dados em geral. Algumas pesquisas que foram redesenhadas recentemente ainda adotam a forma tradicional anterior, que está longe de ser adequada.

Quando se compara a situação da crítica de dados no IBGE com a de outros institutos de pesquisas, tais como *US Bureau of the Census, US National Agricultural Statistics Service, Statistics Canada, Statistics Sweden*, pode-se ver quanto o IBGE ainda tem que fazer para alcançar um padrão similar ao deles.

Pierzchala(1988) analisou a situação das quatro primeiras instituições acima mencionadas, e descreveu seus respectivos procedimentos para o aperfeiçoamento do processo de crítica de dados. Todas essas organizações têm grandes projetos de pesquisa relacionados com esse tema, incluindo investimentos maciços no desenvolvimento de sistemas e no redesenho de pesquisas. Em muitos casos, algum grau de padronização é considerado ou ela já se encontra embutida nos sistemas usados ou em desenvolvimento para o processamento dos dados das pesquisas. Isto é fortemente enfatizado no *Statistics Canada* e no *Netherlands Statistics.*

No *Statistics Canada*, o desenvolvimento e a crescente disponibilidade de sistemas generalizados para a maioria das operações de uma pesquisa, da amostragem à estimação, incluindo coleta de dados e crítica, fornecem um forte incentivo à adoção de ferramentas padronizadas para a crítica de dados quando do planejamento de uma nova pesquisa ou do redesenho de pesquisas atuais. O desenvolvimento do GEIS, o componente para crítica e imputação de dados do sistema generalizado de processamento de pesquisas, foi um grande avanço e seu sucesso na aplicação de um grande número de pesquisas é prova disso.

No *Netherlands Statistics* (NCBS) essa padronização já é uma realidade. Muitas pesquisas adotam atualmente o BLAISE[13] como um sistema para a coleta de dados. A estratégia adotada pelo NCBS está voltada para a crescente automação das operações de pesquisa e

[13] Ver Bethlehem & Keller (1991).

aperfeiçoamento do processo de coleta de dados. Uma vez que grande parte das verificações que devem ser realizadas em cada registro individual pode ser aplicada no momento da coleta dos dados quando o BLAISE é usado no módulo de entrevista assistida por computador, isto reduz bastante ou até mesmo elimina a necessidade de críticas subseqüentes (pelo menos no nível micro).

No IBGE, a padronização ainda não está sendo considerada. O desenvolvimento e a utilização do sistema CRIPTA fornecerá algum grau de padronização, mas o número atual de pesquisas que de fato estão usando esse sistema ainda é pequeno. E uma vez que o sistema CRIPTA não foi desenvolvido para realizar todas as operações de crítica necessárias pela maioria das pesquisas do IBGE, em muitos casos ele terá que ser usado em conjunto com outros sistemas de crítica de dados. Os exemplos descritos neste texto, que mostram como os procedimentos de crítica de dados vêm se desenvolvendo no IBGE, reforçam essa afirmativa e também mostram que a estratégia geral é selecionar o melhor sistema e o melhor método para cada pesquisa para a qual a crítica de dados está sendo planejada.

Este fato vem sendo criticado em função do alto custo e da dependência de sistemas desenvolvidos e mantidos por outros órgãos. Entretanto, o custo de desenvolvimento e manutenção de sistemas próprios como o CRIPTA é alto e no nosso ponto de vista o IBGE não pode ignorar sistemas tais como DIA, IMPS, BLAISE e GEIS que foram cedidos gratuitamente para uso interno. Pode ser difícil providenciar treinamento e suporte adequados para o uso de todos os módulos desses sistemas, mas certamente é bem mais barato e menos arriscado do que desenvolver sistemas sob medida com os mesmos propósitos.

Nesse sentido, nós apoiamos a atual estratégia de uso combinado do sistema CRIPTA com outros sistemas para as aplicações de crítica de

dados de novas pesquisas ou de pesquisas redesenhadas, o que implica que métodos e sistemas diferentes são usados para a crítica de diferentes pesquisas, com o sistema CRIPTA sendo o único componente fixo selecionado para realizar algumas funções básicas de crítica de dados. Esta estratégia pode não ser a ideal, mas ela tem funcionado em algumas pesquisas. Isto é reconhecido como um avanço e mostrou-se mais eficiente que o procedimento padronizado anteriormente usado.

Nós também defendemos a idéia de pesquisas futuras sobre métodos de críticas de dados, não só para aumentar a disponibilidade de novos métodos para a crítica estatística de dados, mas também para a conscientização dos problemas que envolvem a crítica de dados e suas implicações na qualidade geral dos dados divulgados pelo IBGE. Dessa forma, nós achamos bastante oportunas as recentes discussões sobre o assunto que motivaram alguns estudos de avaliação dos procedimentos de crítica adotados em algumas pesquisas do IBGE e exemplos disso podem ser encontrados em Silva(1990a), Silva et al.(1990) e Santis & Damião(1992).

Os desafios apresentados para aperfeiçoamentos futuros nas práticas de crítica de dados no IBGE não são pequenos. Ainda há muito a ser feito para que os métodos e sistemas atualmente disponíveis para crítica de dados sejam amplamente utilizados, sem mencionar a necessidade de novos estudos e desenvolvimentos.

Outro desafio refere-se ao treinamento e à conscientização da equipe técnica com relação ao aproveitamento total dos novos recursos disponíveis em termos de sistemas e metodologias, especialmente pelo fato de a literatura ainda não ser facilmente acessível e ser predominantemente na forma de artigos ou relatórios técnicos de pesquisa. Material didático é muito escasso. O livro de Naus(1975) já pode ser considerado ultrapassado. O livro mais recente sobre o assunto é o

livro bastante agradável de Criado & Cabria(1990). A produção de um livro de fácil entendimento sobre o assunto seria muito bem-vinda, um desafio que pode fornecer um campo fértil para a cooperação entre os especialistas estatísticos de diferentes organizações.

Uma iniciativa que contribuiu para esse fim foi o "*Workshop* sobre métodos de crítica e imputação de dados"[14], promovido pelo IBGE na Escola Nacional de Ciências Estatísticas - ENCE -, em 1990, que teve por objetivo a disseminação de resultados recentes de pesquisas, bem como experiências bem-sucedidas com métodos e sistemas alternativos para crítica de dados, e também promoveu contatos e cooperação técnica entre diversos órgãos de estatística.

Um outro desafio é a modelagem de um ambiente de processamento de dados, que considere a necessidade de automação de todas as operações da pesquisa. Em relação à crítica, esse ambiente deve ser flexível o suficiente para acomodar uma variedade de equipamentos, sistemas e métodos de crítica. Ele deve também fornecer algum grau de integração, através de um acesso fácil, bem como uma documentação padronizada dos dados. A fim de ser totalmente aproveitada, as interfaces para a interação humana devem ser amigáveis.

Nós também pensamos que ferramentas para controle e mensuração do impacto das críticas sobre os dados devem ser mais facilmente disponíveis e mais largamente utilizadas. Muitos sistemas generalizados para a crítica e imputação de dados fornecem algumas facilidades com esse objetivo, embora atualmente elas não estejam sendo usadas como deveriam.

Para concluir, notamos que, com objetivo de disseminar dados com maior qualidade e mais rapidamente aos usuários, ainda são necessários

[14] Ver Silva (1990b).

muitos aperfeiçoamentos em nosso processo de crítica de dados das pesquisas. O comprometimento com a realização de estudos e pesquisas sobre o tema e um aumento na cooperação técnica com outras instituições com os mesmos desafios e propósitos são a chave para atingir esse objetivo.

AGRADECIMENTOS

Os autores expressam seus sinceros agradecimentos a todos aqueles que contribuíram para a preparação do presente documento. Agradecimentos especiais a Sonia Albieri, Chris J. Skinner e John G. Kovar pelos comentários e sugestões nas versões preliminares desse documento, quando de sua versão em inglês, e também a Leopold Granquist que estimulou esse trabalho de muitas e distintas formas.

BIBLIOGRAFIA

ANZANELLO, E. *Esquema genérico para desenvolvimento e produção de sistemas para processamento de pesquisas estatísticas*. Rio de Janeiro: IBGE, 1993. Artigo técnico A00I/93.

BARBOSA, D.M.R., HANONO, R.M. Estudo das ferramentas de apuração de dados. *Revista Brasileira de Estatística*, Rio de Janeiro, v. 49, n. 191, p. 85-100, jan./jun. 1988.

BARCAROLI, G. *DAISY (design, analysis and imputation system)*: an integrated system for edit imputation of data in the Italian National Statistical Institute. [S.l.]: Statistical Commission and Economic Commission for Europe,1992a. (Working paper, n.8)

---------- . An integrated system for edit imputation of data in the Italian National Statistical Institute. In: WESTLAKE A. et al. (Ed.) *Proceedings of the Survey and Statistical Computing Conference*. [S.l.]: Elsevier Science, 1992b. p.167--175.

BETHLEHEM, J.G., KELLER, W.J. The Blaise system for integrated survey processing. *Survey Methodology*, [s. l.], v. 17, n. 1, p .43-56, 1991.

BUSTOS, O.H., SILVA, P.L.N. Uso de estimadores robustos para imputación de datos faltantes en encuestas. *Pro Mathematica*, Lima, v. 2, n. 4, p. 3-26, 1988.

CABRAL, M.D.B. *Plano de controle estatístico de qualidade da codificação dos questionários de indústria do Censo Econômico de 1985.* Rio de Janeiro: IBGE, 1988. 80 p. Manuscrito.

----------. *Um esquema para a apuração das pesquisas estatísticas da Fundação IBGE.* Rio de Janeiro: IBGE, Diretoria de Informática, Departamento Técnico,1990. Manuscrito.

---------. *O esquema genérico para o processamento das pesquisas estatísticas do IBGE.* Rio de Janeiro: IBGE, DI/DETEC , 1991. Manuscrito.

COMPARAÇÃO entre o processo de correção automática de dados utilizado no CD-80 e o sistema DIA. Rio de Janeiro: IBGE, Departamento de População, 1990. Manuscrito.

CRIADO, I.V., CABRIA, M.S.B. *Procedimiento de depuración de datos estadísticos.* Vitoria-Gasteiz: EUSTAT - Instituto Vasco de Estadística, 1990. 169p.

FARIAS, A.M.L. *Estrutura de rota de questionários.* Rio de Janeiro: IBGE, 1990. Manuscrito.

---------, DIAS, A.J.R. *Uma aplicação experimental do sistema DIA na depuração dos dados categóricos do questionário CA--2.0l do Censo Agropecuário de 1985.* Rio de Janeiro: IBGE, 1992. 37 p. Manuscrito.

FELLEGI, I.P., HOLT, D. Systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, [s. l.], v. 71, p. 17-35, 1976.

FORMAX : o editor de formulários: manual do usuário - versão 2.5. Campinas: Teles, Tecnologia de Sistemas,1990.

GRANQUIST, L. On the role of editing. *Statistics Tidskrift*, [s. l.], v.2, p. 106-118, 1984.

----------. *Macro-editing* : the top--down method. Stockholm: Statistics Sweden, 1987. 5p. (Report, 04 /09, 1987).

———. *Macro-editing* : the aggregate method. Stockholm: Statistics Sweden, 1988. 8p. (Report, 08/18, 1988).

———. Data editing and quality. *Revista Brasileira de Estatística*, Rio de Janeiro, v. 51, n. 195, p. 21-51, jan./jun. 1990.

HANONO, R.M., BARBOSA, D.M.R. A tool for the automatic generation of data editing and imputation application for surveys processing. In: WESTLAKE A. et al. (Ed.) *Proceedings of the Survey and Statistical Computing Conference*. [S.l.]: Elsevier Science, 1992. p. 449-456.

HIDIROGLOU, M.A., BERTHELOT, J.M. Statistical editing and imputation for periodic business surveys. *Survey Methodology*, [s. l.], v.1, n.1, p. 73-83, 1986.

Resumo gerencial e relatório de viabilidade de implantação de sistema automático de coleta de dados. Rio de Janeiro: IBGE, 1987. Projeto aquisição de dados. Manuscrito.

JABINE, T.B. Flow charts: a tool for developing and understanding survey questionnaires. *Journal of Official Statistics*, [s. l.], v. 1, n. 2, p. 189-208, 1985

KOVAR, J.G., WHITRIDGE, P. Generalized edit and imputation system: overview and applications. *Revista Brasileira de Estatística*, Rio de Janeiro, v. 51, n. 195, p.85-100, jan./jun. 1990.

LITTLE, R.J.A., SMITH, P.J. Editing and imputation for quantitative survey data. *Journal of the American Statistical Association*, [s. l.], v. 82, p. 58-68, 1987.

METODOLOGIA do Censo Demográfico de 1980. Rio de Janeiro: IBGE, 1983a. 477p.

METODOLOGIA do Censo Agropecuário de 1980. Rio de Janeiro: IBGE ,1983b. 384p.

NAUS, J.I. *Data quality control and editing*. New York: M. Dekker, 1975.

PESQUISA Industrial mensal de dados gerais: notas metodológicas. Rio de Janeiro: IBGE, Departamento de Indústria,1988. 2 v. Manuscrito.

PIERZCHALA, M. *A review of the state of the art in automated data editing and imputation*. Washington: NASS, 1988. (Staff report , n. SRB-88-10).

PINHEIRO, J.C.C., ASSUNÇÃO, R.M. Crítica de razões no Censo Industrial de 1985. *Revista Brasileira de Estatística*, Rio de Janeiro, v. 9, n. 191, p. 101-118, jan./jun. 1988.

———. *Metodologia de crítica de equações de fechamento nos Censos Econômicos de 1985.* Rio de Janeiro: IBGE, Escola Nacional de Ciências Estatísticas, 1989. 28 p. (Relatório Técnico 03/89)

RUBIO, E.G., CRIADO, I.V. *Sistema DIA* : sistema de detección e imputación automática de errores para datos cualitativos. Madrid: Instituto Nacional de Estadística , 1988. v. 1: DIA: descripción del sistema.

SANTIS, F.M., DAMIÃO, S.C. *Uma análise do efeito da microcrítica nos dados da Pesquisa Industrial Anual -- PIA 86/91.* Rio de Janeiro: IBGE, Departamento de Indústria, 1992. 36p. Manuscrito.

———, SOUZA, L. *Metodologia do sistema de crítica da Pesquisa Industrial Anual PIA 86/91.* Rio de Janeiro: IBGE, Departamento de Indústria, 1991. Manuscrito.

SILVA, A.C.M.. *Codificação do questionário da amostra do Censo Demográfico experimental de Limeira.* Rio de Janeiro: IBGE, 1989a. 14 p. Manuscrito.

———, HANONO, R.M., BARBOSA, D.M.R. *Sistema gerador de aplicações de codificação assistida* (SISCOD). Rio de Janeiro: IBGE, 1993. 8p. Manuscrito.

SILVA, D.B.N., RIBEIRO, L.R.J. *Crítica de equações de fechamento de empresas no Censo Econômico de 1985.* Rio de Janeiro: IBGE, Departamento de Comércio e Serviços, 1990. 32 p. Manuscrito.

SILVA, P.L.N. *Crítica e imputação de dados quantitativos utilizando o SAS.* Rio de Janeiro: IMPA, 1989b. (Série D, n. 028)

SILVA, P.L.N. *A apuração do Censo Agropecuário de 1985*: uma visão crítica. Rio de Janeiro: IBGE, Divisão de Metodologia, 1990a. 11p. Manuscrito.

———. *Workshop sobre métodos de crítica e imputação de dados.* Rio de Janeiro: IBGE, Divisão de Metodologia, 1990b. 10p. Manuscrito.

———, BIANCHINI, Z.M. *Qualidade de dados e o envolvimento da Divisão de Metodologia no planejamento do Censo Demográfico de 1991.* Versão final. Rio de Janeiro: IBGE, 1992. 27p. Manuscrito.

———, CABRAL, M.D.B., INDÁ, L.B. *Noções de controle estatístico de qualidade para operação de censos*. Rio de Janeiro: IBGE, 1986. Manuscrito.

———, INDÁ, L.B., LIMA, V.M.U.G. *Projeto para o controle de qualidade da transcrição dos dados do questionário básico CD–I.0I no Censo Demográfico de 1991*. Versão preliminar. Rio de Janeiro: IBGE, 1991. 28 p. Manuscrito.

——— et al. *Relatório do teste de viabilidade do uso do sistema "CONPROD":* codificação on–line de produtos no Censo Econômico de 1985. Rio de Janeiro: IBGE, 1986. 74p. Manuscrito.

——— et al. *Uma nova metodologia para correção automática no Censo Demográfico brasileiro*: experimentação e primeiros resultados. Rio de Janeiro: IBGE, 1990. 102 p. (Textos para discussão, n.28).

WILLENBORG, L.C.R.J. The routing structure of questionnaires. In: AUTOMATION in survey processing. Voorburg: Central Bureau voor de Statistiek, 1987. CBS select 4: p. 97-117.

RESUMO

Este documento apresenta uma revisão dos métodos de crítica, dos sistemas computacionais e das estratégias de crítica e processamento dos dados das pesquisas realizadas pelo IBGE, em um passado recente. A partir de um diagnóstico severo dos procedimentos então utilizados, diversas modificações foram introduzidas no processo como resultado de esforços de pesquisa e da realização de experimentos com procedimentos alternativos, os quais são apresentados de forma resumida. Finalmente, é feita uma análise da situação atual e das perspectivas para o futuro.

ABSTRACT

A critical review of the data editing methods, software and strategies formerly used at IBGE, the Brazilian Central Statistical Office, is provided. From the critical diagnosis of editing approach a number of changes resulted, leading to an applied research effort as well as to some experiences with alternative approaches, which are summarized. An analysis of the present situation and future trends is also included.

ANÁLISE BAYESIANA APROXIMADA PARA MODELOS DE CLASSIFICAÇÃO HIERÁRQUICA NÃO-NORMAIS

Jorge Alberto Achcar*
Maria José Pegorin*

1 INTRODUÇÃO

Uma classe de problemas práticos muito importante é dada pelos componentes de variância (ver por exemplo, Box e Tiao, 1973). Quando a variância total do produto observado é grande, temos o interesse em reduzi-la, descobrindo a importância relativa das várias fontes de variação.

Quando existem apenas dois componentes de variância, com J grupos (lotes) de K observações (análises), considere o modelo,

$$y_{jk} = \mu + e_j + e_{jk}, \qquad (1)$$

* Universidade de São Paulo - USP.

onde $j = 1,...,J; k = 1,...,K$, e as variáveis aleatórias e_j e e_{jk} são independentes com

$$E(e_j) = E(e_{jk}) = 0,\ var(e_j) = \sigma_2^2, \text{ e } var(e_{jk}) = \sigma_1^2.$$

Assim a variância total σ_T^2 da (j,k) - ésima observação y_{jk} é dada por $\sigma_1^2 + \sigma_2^2$.

Vamos assumir que a variável aleatória e_{jk} tem distribuição normal $N(0,\sigma_1^2)$ e que o efeito aleatório e_j tem uma distribuição mistura de normais (ver Tiao e Ali, 1971), dada pela densidade

$$p(e_j|\sigma_2^2,\theta,\delta,\lambda) = (1-\theta) f_N(e_j|-\delta\theta\sigma,\sigma^2) + \\ + \theta f_N(e_j|-\delta\theta\sigma+\delta\sigma,\lambda^2\sigma^2), \qquad (2)$$

onde $-\infty < e_j < \infty,\ \sigma > 0,\ -\infty < \delta < \infty,\ \lambda \geq 1,\ 0 \leq \theta \leq 1,$ e $f_N(x|p,q)$ indica a densidade de uma distribuição normal com média p e variância q.

A distribuição em (2) pode ser interpretada como se e_j viesse de uma entre duas populações, um modelo central $N(-\delta\theta\sigma,\sigma^2)$ e um modelo alternativo $N(-\delta\theta\sigma+\delta\sigma,\lambda^2\sigma^2)$, com probabilidades $(1-\theta)$ e θ, respectivamente.

Essas distribuições têm média zero e as expressões para a variância, medidas de assimetria e curtose γ_1 e γ_2 são dadas por

$$var(e_j) = \sigma_2^2 = \sigma^2\{1+\theta(\lambda^2-1)+\delta^2\theta(1-\theta)\},$$

$$\gamma_1 = \frac{\theta(1-\theta)\delta\{\delta^2(1-2\theta)+3(\lambda^2-1)\}}{\{1+\theta(\lambda^2-1)+\theta(1-\theta)\delta^2)^{3/2}}, \qquad (3)$$

$$\gamma_2 = \frac{3\{1+\theta(\lambda^4-1)\}+6\theta(1-\theta)\delta^2\{\theta+(1-\theta)\lambda^2\}+\delta^4\theta(1-\theta)\{\theta^3+(1-\theta)^3\}}{\{1+\theta(\lambda^2-1)+\theta(1-\theta)\delta^2\}^2}$$

Como um caso especial de distribuição (2), considere $\theta = 0{,}05$, isto é, o processo está fora de controle em 5% das vezes. Também, assumindo $|\delta| = \lambda - 1$, isto é $\delta = \phi(\lambda - 1)$, $\phi = -1{,}1$ temos o modelo (ver Tiao e Ali, 1971),

$$p(e_j|\sigma_2^2, \lambda, \phi) = 0{,}95 f_N(e_j|-0{,}05\phi(\lambda-1)\sigma, \sigma^2) + \\ + 0{,}05 f_N(e_j|0{,}95\phi(\lambda-1)\sigma, \lambda^2\sigma^2) \qquad (4)$$

Observe que, se $\lambda = 1$, temos a distribuição normal para e_j (suposição usual). Se $\lambda > 1$, a distribuição é simétrica, mas leptocúrtica para $\phi = 0$, desviada para direita para $\phi = 1$, e desviada para a esquerda para $\phi = -1$.

Para o modelo geral (2), Tiao e Ali (1971) mostram que a distribuição é unimodal para todo θ em $(0,1)$ se e somente se

$$\delta^2 < \frac{27(1-\lambda^{-2})^2}{(1-2\lambda^{-2})(2+\lambda^{-2}-\lambda^{-4}) + 2(1-\lambda^{-2}+\lambda^{-4})^{3/2}}. \qquad (5)$$

Usualmente, inferências para modelos com distribuição mistura de normais apresentam grande dificuldades computacionais (ver por exemplo, Titterington, Smith e Makov, 1985).

Para a análise Bayesiana, o estatístico pode decidir por uma entre muitas estratégias existentes: o uso de métodos numéricos (ver por exemplo, Naylor e Smith, 1982); o uso de métodos de aproximação de integrais (ver por exemplo, Tierney e Kadane, 1986) ou o uso do procedimento de Monte Carlo ou Gibbs sampling (ver por exemplo, Kloek e Van Dijk, 1978; ou Gelfand e Smith, 1990).

Nesse artigo, exploramos o uso de métodos Bayesianos aproximados para modelos de classificação hierárquica, assumindo distribuição mistura de normais com densidade (4), baseados no método de Laplace para aproximação de integrais.

Uma das grandes vantagens do método de Laplace para resolver integrais Bayesianas de interesse está em termos do baixo custo computacional e da facilidade na implementação, a qual não requer nenhuma experiência computacional sofisticada.

Mostramos em um exemplo, considerando um conjunto de dados introduzido por Tiao e Ali (1971), a viabilidade do método proposto.

2 UMA DENSIDADE *A POSTERIORI* CONJUNTA PARA $\sigma_1^2, \sigma_2^2, \phi$ E λ

Considerando o modelo de classificação hierárquica (1) com $\mu = 0$, assumindo uma distribuição normal $N(0,\sigma_1^2)$ para o erro e_{jk} e uma distribuição mistura de normais com densidade (4) para o efeito aleatório e_j, a função de verossimilhança para $\sigma_1^2, \sigma_2^2, \phi$ e λ é (ver Tiao e Ali, 1971) dada por (6)

$$l(\sigma_1^2,\sigma_2^2,\phi,\lambda|y) \propto (\sigma_1^2)^{-\nu_1/2} exp\left\{-\frac{\nu_1 m_1}{2\sigma_1^2}\right\} \times \prod_{j=1}^{J} p(y_{j.}|\sigma_1^2,\sigma_2^2,\phi,\lambda),$$

onde $y_{j.} = K^{-1}\sum_k y_{jk}$, $\nu_1 = J(K-1)$, $\nu_1 m_1 = S_1$,

$$S_1 = \sum_j \sum_k (y_{jk} - y_{j.})^2,\ p(y_{j.}|\sigma_1^2,\sigma_2^2,\phi,\lambda) = A_{1j} + A_{2j},$$

$$A_{1j} = \left(0{,}95 exp\left\{ -\frac{[y_{j.} + 0{,}05\phi(\lambda-1)\sigma]^2}{2(\sigma^2 + \sigma_1^2/K)} \right\} \right) / (\sigma^2 + \sigma_1^2/K)^{1/2},$$

$$A_{2j} = \left(0{,}05 exp\left\{ -\frac{[y_{j.} - 0{,}95\phi(\lambda-1)\sigma]^2}{2(\lambda^2\sigma^2 + \sigma_1^2/K)} \right\} \right) / (\lambda^2\sigma^2 + \sigma_1^2/K)^{1/2},$$

$$\sigma_2^2 = c(\lambda,\phi)\sigma^2, \text{ e } c(\lambda,\phi) = 0{,}95 + 0{,}05\lambda^2 + 0{,}0475\phi^2(\lambda-1)^2,$$

e y é um vetor de dados.

A densidade *a priori* conjunta para $\sigma_1^2, \sigma_2^2, \phi$ e λ pode ser escrita na forma,

$$\pi(\sigma_1^2,\sigma_2^2,\phi, \lambda) = \pi(\sigma_1^2,\sigma_2^2|\phi, \lambda)\pi_0(\phi, \lambda). \quad (7)$$

Considerando uma *priori* não-informativa para σ_1^2 e σ_2^2 dado ϕ e λ, $\pi(\sigma_1^2,\sigma_2^2|\phi, \lambda) \propto (\sigma_1^2)^{-1}(\sigma_1^2 + K\sigma_2^2)^{-1}$ (também considerada por Tiao e Ali, 1971), a densidade *a posteriori* conjunta para $\sigma_1^2,\sigma_2^2,\phi$ e λ é dada por

$$\pi\left(\sigma_1^2,\sigma_2^2,\phi,\lambda|y\right) \propto \pi_0(\phi,\lambda)(\sigma_1^2)^{-1}(\sigma_1^2 + K\sigma_2^2)^{-1} l\left(\sigma_1^2,\sigma_2^2,\phi,\lambda|y\right), \quad (8)$$

onde $\sigma_1^2 > 0, \sigma_2^2 > 0, \lambda \geq 1, \phi = -1, 0, 1$; $\pi_0(\phi,\lambda)$ é uma densidade *a priori* conjunta para ϕ e λ e $l(\sigma_1^2,\sigma_2^2,\phi,\lambda|y)$ é a função de verossimilhança (6).

Para a escolha da *priori* $\pi_0(\phi,\lambda)$, podemos assumir independência *a priori* entre os parâmetros ϕ e λ. Isto é, $\pi_0(\phi,\lambda) = \pi_{01}(\phi)\pi_{02}(\lambda)$. Nesse caso, é possível considerar diferentes escolhas para as densidades *a priori* para ϕ e λ. Como um caso especial, considere $\pi_{01}(\phi) = 1/3$, para $\phi = -1, 0, 1$ e $\pi_{02}(\lambda) = 1/\lambda, \lambda \geq 1$. Podemos também considerar uma densidade *a priori* informativa para λ, uma vez que em muitas aplicações temos opinião *a priori* sobre λ.

3 MÉTODO DE LAPLACE PARA APROXIMAÇÃO DE INTEGRAIS

Considere aproximações para momentos *a posteriori* na forma,

$$E(g(\psi)|y) = \frac{\int g(\psi)\pi(\psi)l(\psi|y)d\psi}{\int \pi(\psi)l(\psi|y)d\psi}, \qquad (9)$$

onde $g(\psi)$ é uma função selecionada de $\psi \in R^m$, $\pi(\psi)$ é uma densidade *a priori*, $l(\psi|y)$ é a função de verossimilhança para ψ; e para densidades *a posteriori* na forma,

$$\pi(\psi_1|y) = \int \pi(\psi_1, \psi_2|y)d\psi_2, \qquad (10)$$

onde $\pi(\psi_1, \psi_2|y)$ é uma densidade *a posteriori* conjunta para

$$\psi = (\psi_1, \psi_2),\ \psi_1 \in R^k \text{ e } \psi_2 \in R^{m-k}$$

O método de aproximação para momentos *a posteriori* introduzido por Tierney e Kadane (1986) é baseado nas aproximações de Laplace para as integrais no numerador e denominador de (9). O método de Laplace para aproximação de integrais é usado para resolver integrais do tipo,

$$I = \int f(\psi)exp\{-nh(\psi)\}d\psi, \qquad (11)$$

onde $-nh(\psi)$ é uma função com máximo em $\hat{\psi}$ e que satisfaz as condições usuais de regularidade (ver por exemplo, Tierney e Kadane, 1986).

Para aproximar integrais da forma (11), o método de Laplace assume uma expansão de h e f em série de Taylor em torno de $\hat{\psi}$ (ver Tierney e Kadane, 1986; ou Tierney, Kass e Kadane, 1989a, 1989b).

Com ψ unidimensional, a aproximação de Laplace para I é dada por

$$\hat{I} \cong \left(\frac{2\pi}{n}\right)^{1/2} \sigma_L f(\hat{\psi}) exp\{-nh(\hat{\psi})\}, \qquad (12)$$

onde $\sigma_L = \{h''(\hat{\psi})\}^{-1/2}$.

No caso m-dimensional,

$$\hat{I} \cong (2\pi)^{m/2} \left\{det(n\sum_h^{-1}(\hat{\psi}))\right\}^{-1/2} f(\hat{\psi}) exp\{-nh(\hat{\psi})\}, \qquad (13)$$

onde $\sum_h^{-1}(\hat{\psi})$ é a matriz Hessiana de h calculada em $\hat{\psi}$, dada por

$$\sum_h^{-1}(\hat{\psi}) = \left(\frac{\partial^2 h}{\partial\psi_i \partial\psi_j}\right)\Bigg|_{\hat{\psi}} : i,j = 1,...,m.$$

Para aproximar o momento *a posteriori* (9), podemos considerar $\pi(\psi)l(\psi|y) = exp\{-nh(\psi)\}$ no numerador e denominador de (9), com f iguais a g e 1, respectivamente. Outra escolha de f que proporciona uma aproximação mais precisa em (9) é dada por $f = 1$ em ambas integrais em (9).

De modo semelhante, calculamos aproximações de Laplace para a densidade *a posteriori* marginal (10).

4 UMA DENSIDADE *A POSTERIORI* MARGINAL APROXIMADA PARA ϕ

A densidade *a posteriori* marginal para ϕ (de (8)) é dada por

$$\pi(\phi|y) \propto \int\int\int \pi_0(\phi,\lambda)(\sigma_1^2)^{-1}(\sigma_1^2 + K\sigma_2^2)^{-1} l(\sigma_1^2,\sigma_2^2,\phi,\lambda|y) d\sigma_1^2 d\sigma_2^2 d\lambda \qquad (14)$$

Considerando,

$$f_{\phi}(\sigma_1^2,\sigma_2^2,\lambda) = \pi_0(\phi,\lambda)(\sigma_1^2)^{-1}(\sigma_1^2 + K\sigma_2^2)^{-1}$$

e

$$-nh_{\phi}(\sigma_1^2,\sigma_2^2,\lambda) = lnl(\sigma_1^2,\sigma_2^2,\phi,\lambda|y)$$

em (11), uma aproximação de Laplace para a densidade *a posteriori* marginal de ϕ é dada por

$$\hat{\pi}(\phi|y) \propto \frac{\pi_0(\phi,\hat{\lambda})(\hat{\sigma}_1^2)^{-1}(\hat{\sigma}_1^2 + K\hat{\sigma}_2^2)^{-1} l(\hat{\sigma}_1^2,\hat{\sigma}_2^2,\phi,\hat{\lambda}|y)}{\left\{det\left(n\sum_{h_{\phi}}^{-1}(\hat{\sigma}_1^2,\hat{\sigma}_2^2,\hat{\lambda})\right)\right\}^{1/2}}, \qquad (15)$$

onde $\phi = -1,0,1$ e $(\hat{\sigma}_1^2,\hat{\sigma}_2^2,\hat{\lambda})$ maximizam $-nh_{\phi}(\sigma_1^2,\sigma_2^2,\lambda)$ para cada valor de ϕ.

Uma vez que temos dificuldade para calcular as derivadas de segunda ordem de $-nh(\sigma_1^2,\sigma_2^2,\lambda)$, podemos considerar o uso de derivadas numéricas localmente em $(\hat{\sigma}_1^2,\hat{\sigma}_2^2,\hat{\lambda})$ para obter um valor aproximado do determinante da matriz Hessiana em (15).

Observe também que a aproximação em (15) é válida para qualquer escolha da densidade *a priori* para os parâmetros σ_1^2,σ_2^2, ϕ e λ.

5 UMA DENSIDADE *A POSTERIORI* MARGINAL APROXIMADA PARA λ COM ϕ CONHECIDO

A densidade *a posteriori* marginal para λ é dada (de (8)) por

$$\pi(\lambda|y) = \sum_{\phi=-1,0,1} \int\int \pi(\sigma_1^2,\sigma_2^2,\phi,\lambda|y)d\sigma_1^2 d\sigma_2^2. \qquad (16)$$

Assumindo ϕ conhecido, consideramos a densidade *a priori*,

$$\pi(\sigma_1^2,\sigma_2^2,\lambda|\phi) = \pi(\sigma_1^2,\sigma_2^2|\phi,\lambda)\pi_0(\lambda|\phi)\,, \tag{17}$$

onde $\pi_0(\lambda|\phi)$ é uma densidade *a priori* para λ dado ϕ .

Com essa *a priori*, a densidade *a posteriori* conjunta para σ_1^2,σ_2^2 e λ é dada por

$$\pi(\sigma_1^2,\sigma_2^2,\lambda|\phi,y) \propto \pi_0(\lambda|\phi)(\sigma_1^2)^{-1}(\sigma_1^2+K\sigma_2^2)^{-1} \times l(\sigma_1^2,\sigma_2^2,\lambda|\phi,y) \tag{18}$$

Com a escolha $f_\lambda(\sigma_1^2,\sigma_2^2) = \pi_0(\lambda|\phi)(\sigma_1^2)^{-1}(\sigma_1^2+K\sigma_2^2)^{-1}$ e $-nh_\lambda(\sigma_1^2,\sigma_2^2) = lnl(\sigma_1^2,\sigma_2^2,\lambda|\phi,y)$ em (11), uma aproximação de Laplace para a densidade *a posteriori* marginal de λ é dada por

$$\hat{\pi}(\lambda|\phi,y) \propto \frac{\pi_0(\lambda|\phi)(\hat{\sigma}_1^2)^{-1}(\hat{\sigma}_1^2+K\hat{\sigma}_2^2)^{-1}l(\hat{\sigma}_1^2,\hat{\sigma}_2^2,\lambda|\phi,y)}{\left\{det\left(n\sum_{h_\lambda}^{-1}(\hat{\sigma}_1^2,\hat{\sigma}_2^2)\right)\right\}^{1/2}}, \tag{19}$$

onde $(\hat{\sigma}_1^2,\hat{\sigma}_2^2)$ maximizam $-nh_\lambda(\sigma_1^2,\sigma_2^2)$ para cada valor de λ.

6 UMA ANÁLISE BAYESIANA COM ϕ E λ CONHECIDOS

Assumindo ϕ e λ conhecidos, a função de verossimilhança para σ_1^2 e σ_2^2 é (ver (6)) dada por

$$l(\sigma_1^2,\sigma_2^2|\phi,\lambda,y) \propto (\sigma_1^2)^{-\nu_1/2}exp\left\{-\frac{\nu_1 m_1}{2\sigma_1^2}\right\}\prod_{j=1}^{J}\left\{A_{1j}+A_{2j}\right\}\,, \tag{20}$$

onde,

$$A_{1j} = \left(0{,}95 exp\left\{ -\frac{\left(y_{j.} + a_1\sigma\right)^2}{2(\sigma^2 + \sigma_1^2 / K)} \right\} \right) / (\sigma^2 + \sigma_1^2 / K)^{1/2},$$

$$A_{2j} = \left(0{,}05 exp\left\{ -\frac{\left(y_{j.} - a_2\sigma\right)^2}{2(\lambda^2\sigma^2 + \sigma_1^2 / K)} \right\} \right) / (\lambda^2\sigma^2 + \sigma_1^2 / K)^{1/2},$$

$$a_1 = 0{,}05\phi(\lambda - 1),\ a_2 = 0{,}95\phi(\lambda - 1),$$

$$\sigma^2 = \sigma_2^2 / b \text{ e}$$

$$b = 0{,}95 + 0{,}05\lambda^2 + 0{,}0475\phi^2(\lambda - 1)^2.$$

A densidade *a posteriori* conjunta para σ_1^2 e σ_2^2 é dada por

$$\pi(\sigma_1^2, \sigma_2^2 | \phi, \lambda, y) \propto \pi(\sigma_1^2, \sigma_2^2 | \phi, \lambda) l(\sigma_1^2, \sigma_2^2 | \phi, \lambda, y), \qquad (21)$$

onde $\sigma_1^2 > 0$, $\sigma_2^2 > 0$; $\pi(\sigma_1^2, \sigma_2^2 | \phi, \lambda)$ é uma densidade *a priori* para (σ_1^2, σ_2^2) com ϕ e λ conhecidos e $l(\sigma_1^2, \sigma_2^2 | \phi, \lambda, y)$ é a função de verossimilhança (20).

6.1 Uma Densidade *a posteriori* Marginal Aproximada para σ_1^2

A densidade *a posteriori* marginal para σ_1^2 pode ser escrita (de (21)) na forma,

$$\pi(\sigma_1^2 | \phi, \lambda, y) \propto \int f_{\sigma_1^2}(\sigma_2^2) e^{-nh_{\sigma_1^2}(\sigma_2^2)} d\sigma_2^2, \qquad (22)$$

onde $f_{\sigma_1^2}(\sigma_2^2) = \pi(\sigma_1^2,\sigma_2^2|\phi,\lambda)$ (uma densidade *a priori* para σ_1^2 e σ_2^2 dado ϕ e λ) e $-nh_{\sigma_1^2}(\sigma_2^2) = lnl(\sigma_1^2,\sigma_2^2|\phi,\lambda,y)$ .Portanto, uma densidade *a posteriori* marginal aproximada para σ_1^2 usando o método de Laplace é dada por

$$\hat{\pi}(\sigma_1^2|\phi,\lambda,y) \propto \frac{\pi(\sigma_1^2,\hat{\sigma}_2^2|\phi,\lambda)l(\sigma_1^2,\hat{\sigma}_2^2|\phi,\lambda,y)}{\left\{-\frac{\partial^2 g(\sigma_1^2,\hat{\sigma}_2^2)}{\partial(\sigma_2^2)^2}\right\}^{1/2}}, \qquad (23)$$

onde $\sigma_1^2 > 0$, $\hat{\sigma}_2^2$ maximiza $-nh_{\sigma_1^2}(\sigma_2^2)$ para cada valor de σ_1^2 e $g(\sigma_1^2,\sigma_2^2) = \sum_{j=1}^{J} ln(A_{1j} + A_{2j})$.

6.2 Uma Densidade *a posteriori* Marginal Aproximada para σ_2^2

A densidade *a posteriori* marginal para σ_2^2 pode ser escrita (de (21)) na forma,

$$\pi(\sigma_2^2|\phi,\lambda,y) \propto \int f_{\sigma_2^2}(\sigma_1^2)e^{-nh_{\sigma_2^2}(\sigma_1^2)}d\sigma_1^2, \qquad (24)$$

onde $f_{\sigma_2^2}(\sigma_1^2) = \pi(\sigma_1^2,\sigma_2^2|\phi,\lambda)$ e $-nh_{\sigma_2^2}(\sigma_1^2) = lnl(\sigma_1^2,\sigma_2^2|\phi,\lambda,y)$.

Uma densidade *a posteriori* marginal aproximada para σ_2^2 é dada por

$$\hat{\pi}(\sigma_2^2|\phi,\lambda,y) \propto \frac{\pi(\hat{\sigma}_1^2,\sigma_2^2|\phi,\lambda,y)}{\left\{\frac{v_1 m_1}{(\hat{\sigma}_1^2)^3} - \frac{v_1}{2(\hat{\sigma}_1^2)^2} - \frac{\partial^2 g(\hat{\sigma}_1^2,\sigma_2^2)}{\partial(\sigma_1^2)^2}\right\}^{1/2}}, \qquad (25)$$

onde $\sigma_2^2 > 0$, $\hat{\sigma}_1^2$ maximiza $-nh_{\sigma_2^2}(\sigma_1^2)$ para cada valor de σ_2^2.

6.3 Momentos *a posteriori* Aproximados

Podemos também calcular aproximações de Laplace para momentos *a posteriori* da forma,

(26)

$$E\left\{m\left(\sigma_1^2,\sigma_2^2\right)\middle|y\right\}=\frac{\int\int m\left(\sigma_1^2,\sigma_2^2\right)\pi\left(\sigma_1^2,\sigma_2^2\middle|\phi,\lambda\right)l\left(\sigma_1^2,\sigma_2^2\middle|\phi,\lambda,y\right)d\sigma_1^2 d\sigma_2^2}{\int\int \pi\left(\sigma_1^2,\sigma_2^2\middle|\phi,\lambda\right)l\left(\sigma_1^2,\sigma_2^2\middle|\phi,\lambda,y\right)d\sigma_1^2 d\sigma_2^2}$$

Como um caso especial, com a escolha $f(\sigma_1^2,\sigma_2^2)=m(\sigma_1^2,\sigma_2^2)$ para a integral no numerador de (26) e $f(\sigma_1^2,\sigma_2^2)=1$ para a integral no denominador de (26) (f dado em (11)), obtemos a aproximação de Laplace para (26), dada por

$$\hat{E}\left\{m\left(\sigma_1^2,\sigma_2^2\right)\middle|y\right\}\cong m\left(\tilde{\sigma}_1^2,\tilde{\sigma}_2^2\right)\left(1+O\left(n^{-1}\right)\right), \tag{27}$$

onde $(\tilde{\sigma}_1^2,\tilde{\sigma}_2^2)$ é a moda de (21) (ver Tierney, Kass e Kadane, 1989a).

Como um caso especial, temos (de (27)),

$$E(\sigma_2^2/\sigma_1^2|y)\cong\tilde{\sigma}_2^2/\tilde{\sigma}_1^2,\ E(\sigma_1^2|y)\cong\tilde{\sigma}_1^2,\ E(\sigma_2^2|y)\cong\tilde{\sigma}_2^2.$$

Aproximações com maior precisão para (26) são obtidas considerando outras escolhas de $f(\sigma_1^2,\sigma_2^2)$ para as integrais em (26).

6.4 Distribuição Preditiva Aproximada para uma Média de um Grupo Futuro $Y_{(J+1)\cdot}$

Assumindo ϕ e λ conhecidos, a densidade preditiva para uma média de um grupo futuro $y_{(J+1)\cdot}$ é dada por

$$p\left(y_{(J+1)\cdot}\middle|y\right) \propto \int\int p\left(y_{(J+1)\cdot}\middle|\sigma_1^2,\sigma_2^2,\phi,\lambda\right)\pi(\sigma_1^2,\sigma_2^2|\phi,\lambda,y)d\sigma_1^2 d\sigma_2^2, \quad (28)$$

onde $p\left(y_{(J+1)\cdot}\middle|\sigma_1^2,\sigma_2^2,\phi,\lambda\right) = A_{1(J+1)} + A_{2(J+1)}$ e A_{ij} é definido em (20).

Considerando $f(\sigma_1^2,\sigma_2^2) = \pi(\sigma_1^2,\sigma_2^2|\phi,\lambda)$ e $-nh(\sigma_1^2,\sigma_2^2) = lnp(y_{(J+1)\cdot}|\sigma_1^2,\sigma_2^2,\phi,\lambda) + lnl(\sigma_1^2,\sigma_2^2,\phi,\lambda|y)$ em (11), obtemos também uma densidade preditiva aproximada para uma média de um grupo futuro $y_{(J+1)\cdot}$ dada por

$$\hat{p}\left(y_{(J+1)\cdot}\middle|y\right) \propto \frac{p\left(y_{(J+1)\cdot}\middle|\hat{\sigma}_1^2,\hat{\sigma}_2^2,\phi,\lambda\right)\pi(\hat{\sigma}_1^2,\hat{\sigma}_2^2|\phi,\lambda)l(\hat{\sigma}_1^2,\hat{\sigma}_2^2|\phi,\lambda,y)}{\left\{det\left(n\sum_h^{-1}(\hat{\sigma}_1^2,\hat{\sigma}_2^2)\right)\right\}^{1/2}}, \quad (29)$$

onde $-\infty < y_{(J+1)\cdot} < \infty$; $\hat{\sigma}_1^2$ e $\hat{\sigma}_2^2$ maximizam $-nh(\sigma_1^2,\sigma_2^2)$ para cada valor de $y_{(J+1)\cdot}$.

7 UMA ANÁLISE BAYESIANA ASSUMINDO $\lambda = 1$

Assumindo $\lambda = 1$ na densidade mistura de normais (4), ou seja, uma distribuição normal para o efeito aleatório e_j, a função de verossimilhança

para σ_1^2 e σ_2^2 (com $\mu = 0$ no modelo de classificação hierárquica (1)) é dada por

(30)

$$l(\sigma_1^2,\sigma_2^2|y) \propto (\sigma_1^2)^{-\nu_1/2} exp\left\{-\frac{\nu_1 m_1}{2\sigma_1^2}\right\}(\sigma_1^2 + K\sigma_1^2)^{-J/2} exp\left\{-\frac{K\sum y_{j\cdot}^2}{2(\sigma_1^2 + K\sigma_2^2)}\right\}$$

Considerando f iguais à densidade *a priori* $\pi\left(\sigma_1^2,\sigma_2^2\right)$ em (11), obtemos as densidades *a posteriori* marginais aproximadas para σ_1^2 e σ_2^2 dadas por

$$\hat{\pi}\left(\sigma_1^2|y\right) \propto \frac{\pi(\sigma_1^2,\hat{\sigma}_2^2|y)l(\sigma_1^2,\hat{\sigma}_2^2|y)}{\left\{\frac{K^3\sum y_{j\cdot}^2}{(\sigma_1^2 + K\hat{\sigma}_2^2)^3} - \frac{JK^2}{2(\sigma_1^2 + K\hat{\sigma}_2^2)^2}\right\}^{1/2}}, \quad (31)$$

onde $\sigma_1^2 > 0$ e $\hat{\sigma}_2^2$ maximiza $l(\sigma_1^2,\sigma_2^2|y)$ para cada valor fixo de σ_1^2, e

$$\hat{\pi}(\sigma_2^2|y) \propto \frac{\pi(\hat{\sigma}_1^2,\sigma_2^2|y)l(\hat{\sigma}_1^2,\sigma_2^2|y)}{\left\{\frac{\nu_1 m_1}{(\hat{\sigma}_1^2)^3} - \frac{\nu_1}{2(\hat{\sigma}_1^2)^2} - \frac{J}{2(\hat{\sigma}_1^2 + K\sigma_2^2)^2} + \frac{K\sum y_{j\cdot}^2}{(\hat{\sigma}_1^2 + K\sigma_2^2)^3}\right\}^{1/2}}, \quad (32)$$

onde $\sigma_2^2 > 0$ e $\hat{\sigma}_1^2$ maximiza $l(\sigma_1^2,\sigma_2^2|y)$ para cada valor fixo de σ_2^2.

8 UM EXEMPLO

Considere o conjunto de dados simulados da Tabela 1, introduzido por Tiao e Ali (1971), assumindo o modelo de classificação hierárquica (1)

com densidade (4) para o efeito aleatório e_j, onde $\mu = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $\phi = 1$ e $\lambda = 3$.

Tabela 1 - Médias de Grupo $y_{(J+1)\cdot}$ ordenadas (dados simulados com $\mu = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $\phi = 1$ e $\lambda = 3$, $J = 20$ e $K = 3$)

Grupo j	1	2	3	4	5	6
	-3,682	-2,057	-1,780	-1.280	-0,797	-0,671
Grupo j	7	8	9	10	11	12
	-0,646	-0,471	-0,436	-0,401	-0,378	0,000
Grupo j	13	14	15	16	17	18
	0,112	0,791	0,923	1,571	1,712	4,223
Grupo j	19	20				
	6,415	7,072				

Dos dados da Tabela 1, temos $J = 20$ e $K = 3$, $v_1 = J(K-1) = 40$, $JKy_{\cdot\cdot} = \sum_j \sum_k y_{jk} = 0{,}5131$, $m_1 = 1{,}1525$, $v_1 m_1 = S_1 = 46{,}1$ (ver (6)). O logaritmo da função de verossimilhança (6) é dado por

$$lnl(\sigma_1^2, \sigma_2^2, \phi, \lambda | y) \propto -20ln(\sigma_1^2) - \frac{23{,}05}{\sigma_1^2} + \sum_{j=1}^{20} ln\left(A_{1j} + A_{2j}\right), \qquad (33)$$

onde A_{1j} e A_{2j}, $j = 1, \ldots, J$ são dados em (6). Os estimadores de máxima verossimilhança para σ_1^2, σ_2^2, ϕ e λ (ver Tabela 2) são dados por $\hat{\sigma}_1^2 = 1{,}1546$, $\hat{\sigma}_2^2 = 3{,}1355$, $\hat{\phi} = 1$ e $\hat{\lambda} = 3{,}7625$.

Tabela 2 - Estimadores de Máxima Verossimilhança para σ_1^2, σ_2^2 e λ com ϕ Conhecido

ϕ	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}$	$lnl(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \phi, \hat{\lambda} / y)$
1	1,1525	6,6456	1,0001	52,34026
0	1,1525	3,8737	4,0757	-51,93104
1	1,1546	3,1355	3,7625	-50,41669

Usando segundas derivadas numéricas de $\sum_{j=1}^{20} ln(A_{1j} + A_{2j})$ localmente nos estimadores de máxima verossimilhança $\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\phi}$ e $\hat{\lambda}$, obtemos a matriz de informação de Fisher observada dado $\phi = 1$,

$$I = \begin{pmatrix} 15{,}05601 & 0{,}16206 & -0{,}15655 \\ 0{,}16206 & 0{,}38366 & -0{,}06089 \\ -0{,}15655 & -0{,}06089 & 0{,}37574 \end{pmatrix}$$

Considerando a distribuição usual normal limite para os estimadores de máxima verossimilhança $(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\lambda})$ dado $\phi = 1$, obtemos intervalos de confiança 95% para σ_1^2, σ_2^2 e λ dados por

$$0{,}6471 < \sigma_1^2 < 1{,}6617, -0{,}0757 < \sigma_2^2 < 6{,}3467, \quad \text{e}$$
$$0{,}5180 < \lambda < 7{,}0070.$$

Na Tabela 3, temos as densidades *a posteriori* marginais aproximadas (15), considerando diferentes escolhas para a densidade *a priori* $\pi_0(\phi, \lambda)$. Também consideramos o uso de derivadas numéricas para calcular o determinante da matriz Hessiana dada na aproximação de Laplace em (15), para cada valor de ϕ. A moda das densidades *a*

posteriori marginais para ϕ considerando diferentes densidades *a priori* são todas dadas por $\tilde{\phi} = 1$.

Tabela 3 - Densidade *a posteriori* Marginal Aproximada (15) para ϕ Considerando Diferentes Densidades *a Priori* $\pi_0(\phi,\lambda)$

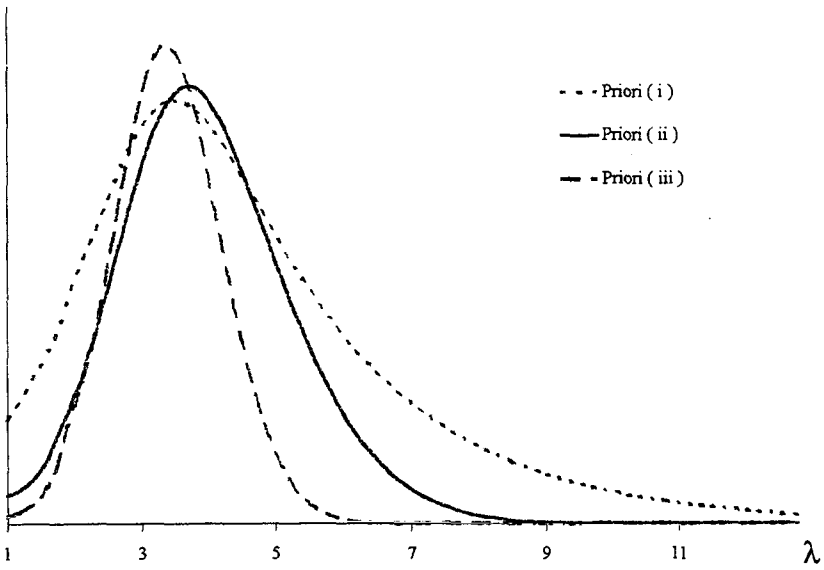
$\pi_0(\phi,\lambda) = \pi_{01}(\phi).\pi_{02}(\lambda)$	ϕ	$\hat{\pi}(\phi / y)$
$\pi_{01}(\phi) = 1/3,\ \phi = -1,0,1$ $\pi_{02}(\lambda) \propto$ *constante*	−1	0,0658
	0	0,2194
	1	0,7148
$\pi_{01}(\phi) = 1/3,\ \phi = -1,0,1$ $\pi_{02}(\lambda) \propto 1/\lambda$	−1	0,2126
	0	0,1738
	1	0,6136
$\pi_{01}(\phi) = 1/3,\ \phi = -1,0,1$ $\pi_{02}(\lambda) \propto exp\left\{-\frac{1}{2}(\lambda-3)^2\right\}$	−1	0,0134
	0	0,1846
	1	0,8020
$\pi_{01}(\phi) = 1/2,\ \phi = 1$ $\pi_{01}(\phi) = 1/4,\ \phi = 0$ $\pi_{01}(\phi) = 1/4,\ \phi = -1$ $\pi_{02}(\lambda) = 1/\lambda$	−1	0,1317
	0	0,1077
	1	0,7606

Assumindo $\phi = 1$ conhecido, temos na Figura 1 o gráfico da densidade *a posteriori* marginal aproximada (19) para λ considerando algumas escolhas diferentes de densidades *a priori* para σ_1^2, σ_2^2 e λ dado $\phi = 1$:

$$
\begin{aligned}
&(i)\ \ \pi_1(\sigma_1^2,\sigma_2^2,\lambda|\phi) \propto \lambda^{-1}(\sigma_1^2)^{-1}(\sigma_1^2+3\sigma_2^2)^{-1},\\
&(ii)\ \ \pi_2(\sigma_1^2,\sigma_2^2,\lambda|\phi) \propto exp\left\{-\tfrac{1}{8}(\lambda-3)^2\right\}(\sigma_1^2)^{-1}(\sigma_1^2+3\sigma_2^2)^{-1},\\
&(iii)\ \ \pi_3(\sigma_1^2,\sigma_2^2,\lambda|\phi) \propto exp\left\{-\tfrac{1}{2}(\lambda-3)^2\right\}(\sigma_1^2)^{-1}(\sigma_1^2+3\sigma_2^2)^{-1}
\end{aligned}
\qquad (34)
$$

Utilizamos também segundas derivadas numéricas para calcular o determinante da matriz Hessiana em (19). A moda da densidade *a posteriori* está em torno de $\tilde{\lambda} \cong 3,5$ considerando todas as densidades *a priori* π_1, π_2, e π_3.

Figura 1 - Densidade *a posteriori* Marginal Aproximada para λ Assumindo $\phi = 1$ Conhecido

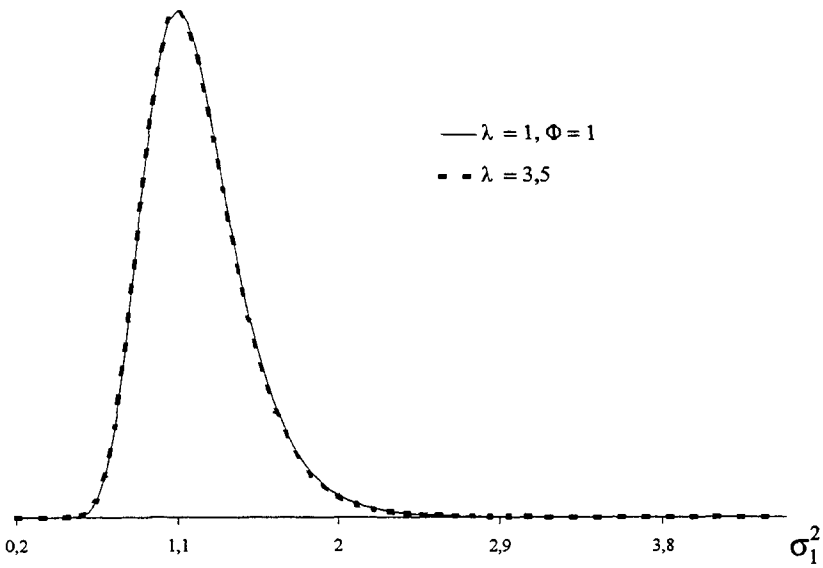


Assumindo $\phi = 1$ e $\lambda = 3,5$ conhecidos, temos na Figura 2 o gráfico da densidade *a posteriori* marginal aproximada (23) para σ_1^2 considerando a densidade *a priori* $\pi(\sigma_1^2, \sigma_2^2 | \phi, \lambda) \propto (\sigma_1^2)^{-1}(\sigma_1^2 + 3\sigma_2^2)^{-1}$, $\sigma_1^2, \sigma_2^2 > 0$.

Temos também, na Figura 2, o gráfico da densidade *a posteriori* marginal aproximada para σ_1^2 (31) assumindo $\lambda = 1$, isto é, uma distribuição normal para o efeito aleatório e_j. Observamos resultados

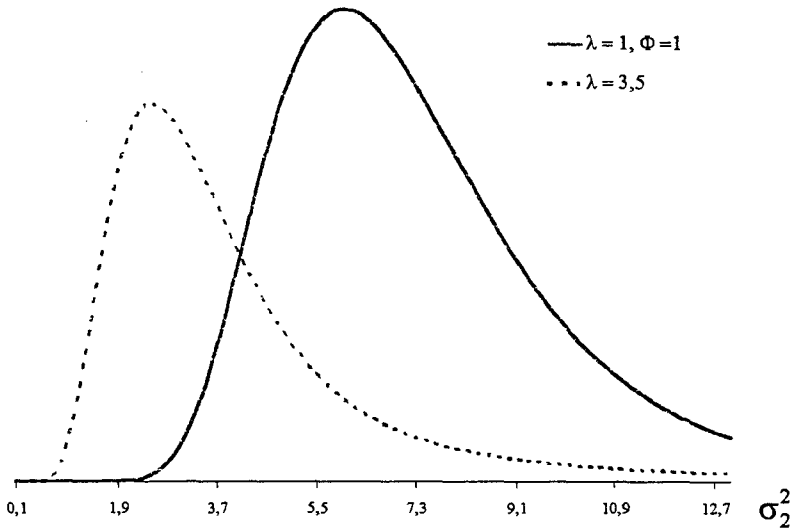
muito próximos para σ_1^2 considerando normalidade ou não-normalidade de $e_j,\ j = 1,...,J$.

Figura 2 - Densidade *a posteriori* Marginal Aproximada para σ_1^2



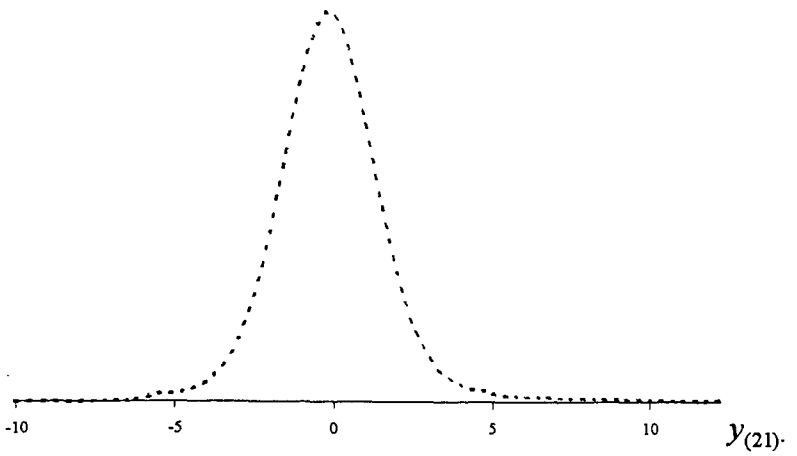
Também, com $\phi = 1$ e $\lambda = 3,5$ conhecidos, e considerando a mesma *a priori* não-informativa para σ_1^2 e σ_2^2, temos na Figura 3 o gráfico da densidade *a posteriori* marginal aproximada (25) para σ_2^2. Temos também, na Figura 3, o gráfico da densidade *a posteriori* marginal aproximada (32) para σ_2^2 considerando $\lambda = 1$. Nesse caso, observamos muita diferença nas inferências resultantes para σ_2^2 considerando normalidade e não-normalidade para os efeitos aleatórios $e_j,\ j = 1,...,J$. É importante salientar que podemos utilizar essas aproximações de Laplace considerando qualquer escolha de opinião *a priori* sobre σ_1^2 e σ_2^2.

Figura 3 - Densidade *a posteriori* Marginal Aproximada para σ_2^2



Na Figura 4, temos o gráfico da densidade preditiva aproximada (29) para uma média de um grupo futuro $y_{(J+1)\cdot}$, considerando o conjunto de dados da Tabela 1.

Figura 4- Densidade Preditiva para uma Média de um Grupo Futuro $y_{(J+1)\cdot}$. Assumindo $\phi = 1$ e $\lambda = 3,5$



9 CONCLUSÕES

O uso de métodos Bayesianos aproximados baseados nas aproximações de Laplace para integrais é uma alternativa satisfatória para análise Bayesiana em modelos de classificação hierárquica com dois componentes de variância, assumindo não-normalidade para os efeitos aleatórios. Resultados similares também são obtidos para modelos com três ou mais componentes de variância.

Podemos também considerar θ desconhecido para introduzir uma análise Bayesiana aproximada similar para modelos com dois componentes de variância, considerando distribuição mistura de normais para o efeito aleatório e_j no modelo (1). Com essa suposição, uma densidade *a posteriori* marginal aproximada para θ é dada por

$$\hat{\pi}(\theta|y) \propto \frac{\pi(\hat{\sigma}_1^2,\hat{\sigma}_2^2,\theta,\hat{\delta},\hat{\lambda})l(\hat{\sigma}_1^2,\hat{\sigma}_2^2,\theta,\hat{\delta},\hat{\lambda}|y)}{\left\{det\left(n\sum_{h_\theta}^{-1}(\hat{\sigma}_1^2,\hat{\sigma}_2^2,\hat{\delta},\hat{\lambda})\right)\right\}^{1/2}}, \tag{35}$$

onde $0 \le \theta \le 1$, $\pi(\sigma_1^2,\sigma_2^2,\theta,\delta,\lambda)$ é uma densidade *a priori* para os parâmetros e $(\hat{\sigma}_1^2,\hat{\sigma}_2^2,\hat{\delta},\hat{\lambda})$ maximizam $-nh_\theta(\sigma_1^2,\sigma_2^2,\delta,\lambda) =$, $lnl(\sigma_1^2,\sigma_2^2,\theta,\delta,\lambda|y)$ o logaritmo da função de verossimilhança (6) com

$$p(y_{j\cdot}|\sigma_1^2,\sigma_2^2,\theta,\delta,\lambda) = (1-\theta)f_N(y_{j\cdot}|-\delta\theta\sigma,\sigma^2+\sigma_1^2/K) + \tag{36}$$
$$+\theta f_N(y_{j\cdot}|-\delta\theta\sigma+\delta\sigma,\lambda^2\sigma^2+\sigma_1^2/K),$$

onde $\sigma_2^2 = \sigma^2\{1+\theta(\lambda^2-1)+\delta^2\theta(1-\theta)\}$, para cada valor de θ.

Do mesmo modo, podemos calcular densidades *a posteriori* marginais aproximadas para os outros parâmetros.

O uso do método de Laplace para aproximação de integrais pode ser justificado comparando-se a densidade *a posteriori* marginal integrada numericamente ou usando o método de Monte Carlo (ver, por exemplo, Kloek e Van Dijk, 1978) com a aproximação de Laplace. Assumindo os

dados da Tabela 1, com $\phi = 1$ e $\lambda = 3{,}5$ conhecidos, temos na Figura 5 o gráfico da densidade *a posteriori* marginal de σ_1^2 considerando o método de Laplace, um procedimento numérico baseado na quadratura Gausiana (Gauss-Hermite com n = 9, raízes da equação polinomial de Hermite) e o procedimento de Monte Carlo. Observamos que os resultados são muito próximos para todos os métodos de integração (ver, também, Tabela 4).

É importante salientar que a precisão da aproximação obtida usualmente depende de boa parametrização e dos dados, especialmente para pequenos tamanhos amostrais (ver, por exemplo, Achcar e Smith, 1990).

Figura 5 - Densidade *a posteriori* Marginal para σ_1^2 com $\phi = 1$ e $\lambda = 3{,}5$ Conhecidos

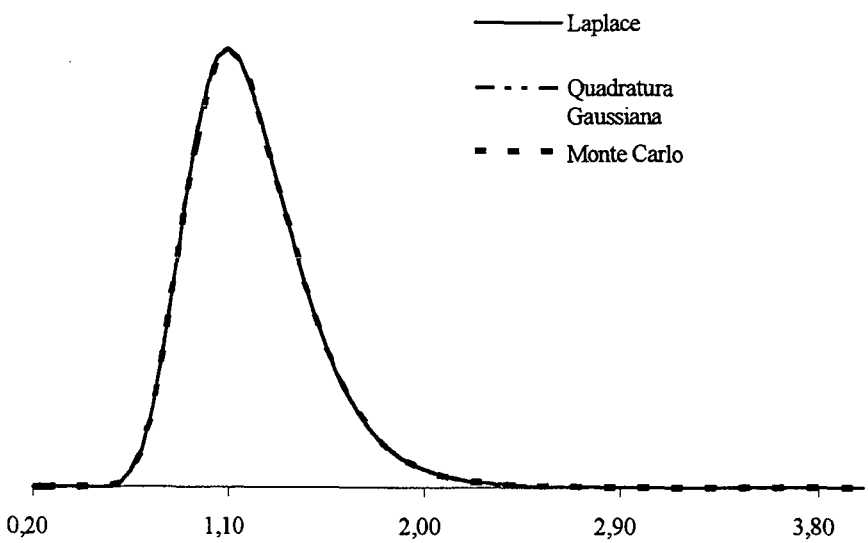


Tabela 4 - Densidade *a posteriori* Marginal para σ_1^2 com $\phi = 1$ e $\lambda = 3{,}5$ Conhecidos

σ_1^2	Laplace	Quadratura Gaussiana	Monte Carlo
0,20	0,000000	0,000000	0,000000
0,30	0,000000	0,000000	0,000000
0,40	0,000000	0,000000	0,000000
0,50	0,000048	0,000048	0,000048
0,60	0,002223	0,002268	0,002252
0,70	0,021107	0,021536	0,021076
0,80	0,078374	0,079903	0,079902
0,90	0,162301	0,165197	0,165061
1,00	0,229981	0,233518	0,229981
1,20	0,232970	0,234881	0,234666
1,40	0,142280	0,142033	0,141824
1,60	0,067468	0,066518	0,066756
1,80	0,028189	0,027383	0,027391
2,00	0,011101	0,010601	0,010483
2,20	0,004277	0,004007	0,003964
2,40	0,001647	0,001511	0,001488
2,60	0,000642	0,000576	0,000575
2,80	0,000255	0,000223	0,000220
3,00	0,000104	0,000088	0,000087
3,20	0,000043	0,000036	0,000035
3,40	0,000018	0,000015	0,000014
3,60	0,000008	0,000006	0,000006
3,80	0,000003	0,000003	0,000003
4,00	0,000003	0,000001	0,000001

AGRADECIMENTOS

Os autores agradecem os comentários feitos pelo revisor, que melhoraram a redação final desse artigo.

BIBLIOGRAFIA

ACHCAR, J.A., SMITH, A.F.M. Aspects of reparametrization in approximate bayesian inference. In: HODGES, J. (Ed.) *Essays in honor of George A. Barnard.* Amesterdam: [s.n.], 1990. p. 431-452 .

BOX, G.E.P., TIAO, G.C. *Bayesian inference in statistical analysis.* New York: Addison-Wesley, 1973.

GELFAND, A.E., SMITH, A.F.M. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association,* [s. l.], v. 85, p. 398-409, 1990.

KLOEK, T., VAN DIJK, H.K. Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica,* [s. l.], v.46, p. 1-19, 1978.

NAYLOR, J.C., SMITH, A.F.M. Applications of a method for the efficient computation of posterior distributions. *Applied Stastistics,* [s. l.], v. 31,p. 214-225, 1982.

TIAO, G.C., ALI, M.M. Effect of non-normality on inferences about variance components. *Technometrics,* [s. l.], v.13, n.3, p. 635-650, 1971.

TIERNEY,L., KADANE, J.B. Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association,* [s. l.], v. 81, p. 82-86, 1986.

——, KASS, R.E., KADANE, J.B. Approximation marginal densities for nonlinear functions. *Biometrika,* [s. l.], v. 76, p. 425-433, 1989a.

——, KASS, R.E., KADANE, J.B. Fully exponential laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association,* [s. l.], v. 84, p. 710-716, 1989b.

TITTERINGTON, D.M., SMITH, A.F.M., MAKOV, U.E. *Statistical analysis of finite mixtures distributions .* New York : J. Wiley, 1985.

RESUMO

Neste artigo, exploramos o uso do método de Laplace (ver, por exemplo, Tierney e Kadane, 1986; JASA, 81, 82-86), para achar quantidades *a posteriori* aproximadas de interesse para modelos de classificação hierárquica com dois componentes de variância, assumindo uma densidade mistura de normais (ver Tiao e Ali, 1971, Technometrics, 13(3), 635-650) para o efeito aleatório. Ilustramos a metodologia proposta considerando um conjunto de dados introduzido por Tiao e Ali (1971).

ABSTRACT

In this paper, we explore the use of Laplace's method (see, for example, Tierney and Kadane, 1986, JASA, 81, 82-86) to find approximate posterior summaries of interest for hierarchical classification models with two variance components assuming a mixture of normal densities (see Tiao and Ali, 1971, Technometrics, 13(3), 635-650) for the random effect. We illustrate the proposed methodology considering a data set introduced by Tiao and Ali (1971).

O MODELO DE REGRESSÃO DE COX COM COVARIÁVEL DEPENDENTE DO TEMPO: UMA APLICAÇÃO ENVOLVENDO PACIENTES INFECTADOS PELO HIV

Enrico A. Colosimo*
Afrânio M. C. Vieira**

1 INTRODUÇÃO

Os estudos na área médica muitas vezes envolvem covariáveis que podem estar relacionadas com o tempo de sobrevivência. Por exemplo, a contagem de CD4 e CD8 ao diagnóstico são duas covariáveis que a literatura médica mostra serem importantes fatores de prognóstico para o tempo até a ocorrência de AIDS em pacientes infectados pelo HIV.

* Departamento de Estatística - UFMG.
** Departamento de Medicina Tropical - UFMG.

Certamente, estas covariáveis devem ser incluídas na análise estatística dos dados.

A forma mais eficiente de acomodar o efeito destas covariáveis é utilizar um modelo de regressão apropriado para dados censurados. Em análise de sobrevivência, existem duas classes de modelos propostas na literatura: os modelos paramétricos e os semiparamétricos. Os modelos paramétricos, também chamados de modelos de tempo de vida acelerado, são mais eficientes, porém menos flexíveis do que os modelos semi-paramétricos. A segunda classe de modelos também chamada simplesmente de modelos de regressão de Cox tem sido muito usada em estudos médicos. Além da flexibilidade, este modelo permite incorporar facilmente covariáveis dependentes do tempo que ocorrem com freqüência na pesquisa clínica. Estes modelos são apresentados em vários livros, entre eles podemos citar Lawless (1982) e Cox e Oakes (1984). O texto em português de Bolfarine, Rodrigues e Achcar (1991) cobre os modelos paramétricos.

Comparações entre as duas classes de modelos podem ser encontradas em Colosimo e Garcia (1991), Wei (1992) e Colosimo et al. (1995).

É difícil não exagerar a importância atual do modelo de regressão de Cox na pesquisa médica. Uma evidência quantitativa deste fato aparece em Stigler (1994). O autor usa citações feitas em periódicos indexados de todas as áreas entre os anos de 1987 e 1989, para quantificar a importância de algumas publicações na literatura estatística. O artigo de Cox (1972), no qual o modelo é apresentado, foi nesse período, o segundo artigo mais citado na literatura estatística, somente ultrapassado pelo artigo de Kaplan e Meier (1959). Isto significa em números, uma média de 600 citações por ano, o que representa aproximadamente 25% das citações anuais ao *Journal of the Royal Statistical Society B*, a revista que publicou o artigo.

Nas seções seguintes, apresentamos o modelo de Cox, sua generalização para incorporar covariáveis dependentes do tempo e a aplicação desta generalização a um conjunto de dados de pacientes infectados pelo HIV.

2 O MODELO DE REGRESSÃO DE COX

A função de risco é a taxa de falha do indivíduo no tempo t condicional a sua sobrevivência até este tempo.O modelo de regressão de Cox pertence a uma classe de modelos com a propriedade de que diferentes indivíduos têm funções de risco proporcionais. Este modelo especifica que a função de risco pode ser escrita como:

$$h(t) = h_0(t)\, g(z,\beta),$$

onde $h_0(t)$ é uma função de risco padrão não especificada, z é o vetor de covariáveis fixas e medidas no início do estudo e β o vetor de parâmetros a serem estimados. A parte paramétrica do modelo é geralmente tomada como $g(z,\beta) = \exp(z'\beta)$. Os dois componentes do modelo têm um efeito multiplicativo na função de risco $h(t)$ devido à presença do componente não-paramétrico $h_0(t)$ no modelo, o método de máxima verossimilhança não pode ser usado para estimar o vetor de parâmetros β. Cox (1975) propôs o método de máxima verossimilhança parcial que vamos apresentar a seguir.

Vamos usar a seguinte notação para escrever a função de verossimilhança parcial: com base em uma amostra de n indivíduos, temos $k \leq n$ falhas distintas nos tempos $t_1 \leq t_2 \ldots \leq t_k$. Uma forma simples de entender a verossimilhança parcial leva em consideração o seguinte argumento condicional: a probabilidade condicional de a *i*-ésima

observação vir a falhar no tempo t_i conhecendo quais observações estão sob risco em t_i é

$$\frac{h_i(t_i)}{\sum_{j\in R(t_i)} h_j(t_i)} = \frac{h_0(t)exp(z_i'\beta)}{\sum_{j\in R(t_i)} h_0(t)exp(z'\beta)} = \frac{exp(z_i'\beta)}{\sum_{j\in R(t_i)} exp(z'\beta)} \quad (1)$$

onde $R(t_i)$ é o conjunto dos índices dos indivíduos sob risco no tempo t_i. Observe que ao condicionarmos a história de falhas e censuras até o tempo t_i, o componente não-paramétrico $h_0(t)$ desaparece de (1).

A função de verossimilhança $L(\beta)$ a ser usada para fazer inferências no modelo é então formada pelo produto de todos os termos representados por (1), associados aos distintos tempos de falha, ou seja,

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(z_i'\beta)}{\sum_{j\in R(t_i)} \exp(z_j'\beta)} = \prod_{i=1}^{n} \left[\frac{\exp(z_i'\beta)}{\sum_{j\in R(t_i)} \exp(z_j'\beta)} \right]^{\delta}. \quad (2)$$

onde δ é o indicador de falha.

O valor de β que maximiza $L(\beta)$ é obtido resolvendo o sistema de equações definido por $U(\beta) = 0$, onde $U(\beta)$ é o vetor escore de primeiras derivadas da função $l(\beta) = log(L(\beta))$. Isto é

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left[z_i - \frac{\sum_{j \in R(t_i)} z_j \exp\left(z_j' \beta\right)}{\sum_{j \in R(t_i)} \exp\left(z_j' \beta\right)} \right] = 0. \qquad (3)$$

A função de verossimilhança parcial (2) assume que os tempos de sobrevivência são contínuos e, conseqüentemente, não pressupõe a possibilidade de empates dos valores observados. Na prática, podem ocorrer empates nos tempos de falhas ou censuras devido à escala de medida. Por exemplo, o tempo não é necessariamente registrado em horas, podendo em alguns estudos ser medido em dias, meses ou até mesmo anos, dependendo da dificuldade em se obter a medida. Da mesma forma, podem ocorrer empates nas censuras. Quando ocorrem empates entre falhas e censuras, usa-se a convenção de que a censura ocorreu após a falha, o que define as observações a serem incluídas no conjunto de risco em cada tempo de falha.

A função de verossimilhança parcial (2) deve ser modificada para incorporar as observações empatadas quando estas estiverem presentes. A aproximação para (2) proposta por Breslow (1972) e Peto (1972) é simples e freqüentemente usada nos pacotes estatísticos comerciais. Vamos considerar s_i o vetor formado pela soma das correspondentes p covariáveis para os indivíduos que falham no tempo t_i; $i = 1, \cdots, k$ é o número de falhas neste mesmo tempo.

A aproximação mencionada anteriormente utiliza a seguinte função de verossimilhança parcial

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp\left(s_i' \beta\right)}{\left[\sum_{j \in R(t_i)} \exp\left(s_j' \beta\right) \right]}.$$

Esta aproximação é adequada quando o número de observações empatadas em qualquer tempo não for grande. Naturalmente, a expressão

acima se reduz a (2) quando não houver empates. Quando o número de empates em qualquer tempo for grande, o modelo de regressão de Cox para dados agrupados deve ser usado (Lawless, 1982, Prentice e Gloeckler, 1978).

3 INCLUSÃO DE COVARIÁVEIS DEPENDENTES DO TEMPO

As covariáveis no modelo de Cox, consideradas na seção anterior, foram medidas no início do estudo ou na origem do tempo. Entretanto, existem outras covariáveis que são monitoradas durante o estudo e seus valores podem mudar ao longo desse período. Por exemplo, pacientes podem mudar de grupo durante o tratamento ou a dose de quimioterapia aplicada em pacientes com câncer pode sofrer alterações durante o curso do tratamento. Se estes valores forem incorporados na análise estatística, resultados mais precisos poderão ser obtidos. Em outros exemplos, a não inclusão destes valores pode acarretar sérios vícios. Um estudo bastante explorado na literatura é o do programa de transplante de coração de Stanford (Crowley & Hu, 1977). Os pacientes eram aceitos no programa quando se tornavam candidatos a um transplante de coração. Quando surgia um doador, os médicos escolhiam, de acordo com alguns critérios, o candidato que iria receber o coração. Alguns pacientes morreram sem receber o transplante. A forma de alocação está fortemente viciada na direção daqueles pacientes com maior tempo de sobrevivência pois somente estes pacientes viveram o suficiente para receber o coração. O uso de uma covariável assumindo o valor zero para aqueles esperando o transplante e um para aqueles com coração novo, serve para minimizar esse vício.

Estas covariáveis que se alteram com o tempo são conhecidas como covariáveis dependentes do tempo. Elas podem ser incorporadas ao modelo de regressão de Cox, generalizando-o como

$$h(t) = h_0(t)\exp\left(z'(t)\beta\right). \qquad (4)$$

Definido desta forma, o modelo (4) não é mais de riscos proporcionais pois a razão das funções de risco no tempo *t* para dois indivíduos *i* e *j*

$$\frac{h_i(t)}{h_j(t)} = \exp\left(z_i'(t)\beta - z_j'(t)\beta\right)$$

depende do tempo. A interpretação dos coeficientes β do modelo devem considerar o tempo *t*.

Cada coeficiente β_l, $l = 1,\ldots,p$ pode ser interpretado como o logaritmo da razão de riscos cujo valor da *l*-ésima covariável no tempo *t* difere de uma unidade, quando as outras covariáveis assumem o mesmo valor neste tempo.

O ajuste do modelo de Cox é obtido estendendo a função de log-verossimilhança parcial. Isto é feito usando

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left[z_i(t_i) - \frac{\sum_{j\in R(t_i)} z_j(t_i)\exp\left(z_j'(t_i)\beta\right)}{\sum_{j\in R(t_i)} \exp\left(z_j'(t_i)\beta\right)} \right] = 0$$

que é uma extensão da expressão (3) considerando covariáveis dependentes do tempo. Propriedades assintóticas dos estimadores de máxima verossimilhança parcial são necessárias para podermos construir intervalos de confiança e testar hipóteses sobre os coeficientes do modelo. Vários autores estudaram estas propriedades (Cox, 1975, Tsiatis, 1981), mas foram Andersen e Gill (1982) que apresentaram as provas mais gerais das propriedades para covariáveis dependentes do tempo, destes estimadores. Eles usaram da relação entre os tempos de falhas e martingales para mostrar que estes estimadores são consistentes e assintoticamente normais sob certas condições de regularidade. Desta

forma, podemos usar as conhecidas estatísticas de Wald e da razão de verossimilhança para fazer inferências no modelo de regressão de Cox.

4 ANÁLISE DOS DADOS DE PACIENTES INFECTADOS PELO HIV

4.1 Descrição dos dados

Neste estudo foram utilizadas informações provenientes de 91 pacientes HIV positivo e 21 HIV negativo, somando assim 112 pacientes estudados. Estes pacientes foram acompanhados no período entre março de1993 a fevereiro de 1995. Somente foram considerados os pacientes que tiveram entrada até julho de 1994. Todos os pacientes incluídos no estudo foram encaminhados ao Centro de Treinamento e Referência em Doenças Infecto-parasitárias (CTR-DIP) da cidade de Belo Horizonte/MG por pertencerem a grupos de comportamento de risco para adquirir o HIV ou por ter um exame HIV positivo. Após a primeira consulta clínica, os pacientes foram encaminhados ao Serviço de Otorrinolaringologia da Universidade Federal de Minas Gerais.

As doenças otorrinolaringológicas (ORL) avaliadas foram definidas com base nos estudos de prevalência destas manifestações na literatura em pacientes infectados pelo HIV. Neste artigo estamos apresentando os resultados para as infecções **candidíase oral** e **sinusite**. A classificação do paciente quanto à infecção pelo HIV seguiu os critérios do CDC (Centers of Disease Control, 1987). Os pacientes foram classificados como:HIV soronegativo, HIV soropositivo assintomático, com ARC (AIDS Related Complex) e com AIDS. Na covariável Grupo de Risco, pacientes HIV soronegativo são aqueles que não possuem o HIV. Pacientes HIV soropositivo assintomáticos são aqueles que possuem o vírus, mas não desenvolveram o quadro clínico de AIDS e que apresentam um perfil

imunológico estável. Pacientes com ARC são aqueles que apresentam baixa imunidade e outros indicadores clínicos que antecedem o quadro clínico de AIDS. Pacientes com AIDS são aqueles que já desenvolveram infecções oportunistas que definem AIDS, segundo os critérios do CDC de 1987. Esta covariável depende do tempo, pois os pacientes mudam de classificação ao longo do estudo. Outras covariáveis neste estudo, como contagem de CD4 e CD8, também são dependentes do tempo. No entanto, elas somente foram medidas no início do estudo.

A cada consulta a classificação do paciente foi reavaliada. Cada paciente foi acompanhado através de consultas trimestrais. A freqüência mediana de consultas foi 4. A resposta de interesse foi o tempo, contado a partir da primeira consulta, até a ocorrência das manifestações ORL. O objetivo foi identificar fatores de risco para cada uma destas manifestações. Os possíveis fatores de risco estão listados na Tabela 1.

Tabela 1: Codificação das covariáveis estudadas

Idade do Paciente	Foi considerada a idade em anos
Sexo do Paciente	1 - Masculino 2 - Feminino
Grupos de Risco	1 - Paciente HIV Soronegativo 2 - Paciente HIV Soropositivo Assintomático 3 - Paciente com ARC 4 - Paciente com AIDS
CD4	Contagem de CD4
CD8	Contagem de CD8
Atividade Sexual	1 - Homossexual 2 - Bissexual 3 - Heterossexual
Uso de Droga Injetável	1 - Sim 2 - Não
Uso de Cocaína por Aspiração	1 - Sim 2 - Não

Para as covariáveis CD4 e CD8 foram registrados 41 valores perdidos, assim como nas covariáveis Atividade Sexual, Uso de Droga e Uso de Cocaína onde também foram registrados 23 valores perdidos.

4.2 Modelagem estatística

Serão apresentados agora os resultados dos ajustes dos modelos de Cox incluindo a covariável Grupo de Risco, que depende do tempo, para cada infecção estudada. Estes são os modelos finais após retirarmos as covariáveis que não afetam as respostas em estudo. Na Tabela 2, está sendo apresentado o modelo para Sinusite.

Tabela 2: Coeficientes estimados do modelo de Cox para Sinusite

Covariável	Coeficiente de Regressão	Erro Padrão	Valor-p	Risco Relativo Estimado (I.C. 95%)
Idade	-0,077	0,031	0,014	0,926 (0,871; 0,984)
HIV soropos. assint.	-0,755	1,000	0,451	0,470 (0,066; 3,338)
com ARC	2,274	0,837	0,007	9,717 (1,884; 50,124)
com AIDS	2,651	0,790	<0,001	14,168 (3,012; 66,646)

Podemos notar que Idade e Grupos de Risco foram identificados como fatores de risco para a ocorrência desta infecção. Verificamos que a cada aumento de 10 anos na idade do paciente o risco de desenvolver Sinusite diminui em exp(0,77)=2,16 vezes, o que indica que pacientes mais jovens estão mais sujeitos a esta infecção. Notamos que o risco de pacientes HIV soropositivo assintomáticos não difere significativamente do grupo HIV soronegativo. Entretanto, no grupo com ARC o risco de se desenvolver Sinusite é 9,7 vezes maior do que o grupo HIV soronegativo. Para o grupo com AIDS, o risco de se desenvolver Sinusite é 14,2 vezes maior em relação ao grupo HIV soronegativo. Por outro lado, a precisão associada a estes dois últimos riscos relativos e estimados é bastante

reduzida como pode ser observado pela grande amplitude de seus respectivos intervalos de confiança.

Tabela 3: Coeficientes estimados do modelo de Cox para Candidíase Oral

Covariável	Coeficiente de Regressão	Erro Padrão	Valor-p	Risco Relativo Estimado (I.C. 95%)
CD4	-0,005	0,002	0,025	0,996 (0,991; 0,999)
CD8	-0,002	0,001	0,019	1,002 (1,000; 1,004)
Droga Injetável	-1,613	0,841	0,055	0,199 (0,038; 1,036

Na Tabela 3, está apresentado o modelo para Candidíase Oral.Com base nos coeficientes de regressão estimados e usando o risco relativo estimado, podemos verificar que, para cada decréscimo em 100 linfócitos CD4, o risco do paciente adquirir Candidíase é 1,58 vezes maior. Já para os linfócitos CD8, o ganho de 100 células acarreta em um risco 1,22 maior de adquirir Candidíase. Para o paciente usuário de drogas injetáveis, o risco de adquirir Candidíase é cerca de 5 vezes maior do que as pessoas que não são usuárias de drogas.

5 CONCLUSÃO

Neste artigo, nós apresentamos um estudo envolvendo dados censurados com a presença de uma covariável dependente do tempo. Usando o modelo de regressão de Cox, foi possível incluir esta covariável na análise dos dados. O ajuste deste modelo foi feito no pacote computacional EGRET. Outros pacotes estatísticos que permitem a

inclusão de covariáveis dependentes do tempo no modelo de regressão de Cox são o SAS (procedimento phreg) e o BMDP (programa 2L).

Os resultados obtidos a partir da análise estatística deste estudo são importantes para explicar a incidência de manifestações ORL em pacientes HIV positivos. A análise feita na Seção 4 é somente parte do estudo. Maiores informações sobre este estudo e a interpretação clínica dos achados na análise dos dados podem ser encontradas em Gonçalves (1995).

AGRADECIMENTOS

Gostaríamos de agradecer os comentários e sugestões da Prof[a.] Emília Sakurai e de um revisor que melhoraram a apresentação deste artigo.

BIBLIOGRAFIA

ANDERSEN, P. K., GILL, R. Cox's regression model for m counting processes: a large sample study. *Ann. Statist.*, [s. l.], v. 10, p. 1100-1200, 1982.

BOLFARINE, H., RODRIGUES, J., ACHCAR, J. A. *Análise de sobrevivência.* 2ª Escola de modelos de regressão. Rio de Janeiro, 1991.

BRESLOW, N. Contribuição à discussão do artigo de D. R. Cox. *Journal of the Royal Statistical Society B*, [s. l.], v. 34, p. 216-217, 1972.

COLOSIMO, E. A., GARCIA, N. C. Testing treatment differences in censored survival data: a small sample study. *Braz. Journal of Probab. and Statist.*, [s. l.], v. 5, p. 135-148, 1991.

-------- et al. Comparação de modelos de sobrevivência aplicados a um estudo de Leucemia de crianças. *Revista Brasileira de Estatística*, 1995. No prelo.

COX, D. R. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B,* [s. l.], v. 34, p. 187-220, 1972.

———. Partial Likelihood. *Biometrika,* [s. l.], v. 62, p. 269-76, 1975.

———, OAKES, D. *Analysis of Survival Data.* London: Chapman and Hall, 1984.

CROWLEY, J., HU, M. Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association,* [s. l.], v. 72, p. 27-36, 1977.

GONÇALVES, D. U. *Incidência, marcadores de prognóstico e fatores de risco relacionados às manifestações otorrinolaringológicas em pacientes infectados pelo HIV.* Belo Horizonte, 1995. Dissertação (Mestrado) - Faculdade de Medicina, Universidade Federal de Minas Gerais, 1995.

HUMAN immuno deficiency virus in the United States: a review of current knowledge. M.M.W.R., Center for Disease Control, [s. l.], v. 36, p. 1-48. 1987. Supplement 6.

KAPLAN, E. L., MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association,* [s. l.], v. 53, p. 457-481, 1958.

LAWLESS, J. F. *Statistical Models and Methods for Lifetime Data.* New York : J. Wiley, 1982.

PETO, R. Contribuição à discussão do artigo de D. R. Cox. *Journal of the Royal Statistical Society B,* [s. l.], v. 34, p. 205-207, 1972.

PRENTICE, R. L., GLOECKER, L. A. Regression analysis of gruped survival data with application to breast cancer data. *Biometrics,* [s. l.], v. 34, p. 57-67, 1978.

STIGLER, S. M. Citation patterns in the Journals of Statistics and Probability. *Statistical Science,* [s. l.], v. 9, p. 94-108, 1994.

TSIATIS, A. A. A large sample study of Cox's regression model. *Ann. Statist.,* [s. l.], v. 9, p. 93-108, 1981.

WEI, L. J. The accelerated failure time model: a useful alternative to the Cox 's regression. *Statistics in Medicine,* [s. l.], v. 11, p. 1871-1879, 1992.

RESUMO

Os métodos estatísticos para analisar dados de tempo de falha censurados têm surgido na literatura com bastante freqüência. Atenção especial é dada ao modelo de Cox, um método flexível para estudar a associação entre covariáveis e taxas de falha. Covariáveis dependentes do tempo aparecem com uma certa freqüência em conjuntos de dados médicos. Uma vantagem do modelo de Cox é poder incorporar com facilidade estas covariáveis na estrutura de um modelo de regressão. Neste artigo, o modelo de Cox com covariável dependente do tempo é usado para analisar um conjunto de dados envolvendo pacientes infectados pelo HIV.

ABSTRACT

In the past years, there has been intense activity in the applied and theoretical statistical literature in methods for analyzing censored failure time data. Special attention has been given to the Cox model which provides a flexible method for exploring the association of covariates with failure rates. Another important issue in survival analysis is the possible presence of time-dependent covariates. They can be easily incorporated in the regression structure of the Cox model. In this paper, the time-dependent Cox model is used for the statistical analysis of a data set of HIV patients.

— ESTIMATION AND INFERENCE IN ECONOMETRICS, R. Davidson e J.G Mackinnon, 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press, 874 pp. ISBN: 0-19-506011-3. US$ 45.00. in Francisco Cribari-Neto, Department of Economics, University of Illinois, 484 Commerce West, 1206 South Sixth St., Champaign/IL, 61820, USA.

Estimation and Inference in Econometrics é o primeiro livro-texto moderno de econometria no sentido que: *(i)* cobre boa parte dos desenvolvimentos recentes sem se prender a uma estrutura arcaica de organização e divisão do material; e *(ii)* reflete talvez a mais importante mudança da econometria na última década: o crescimento da área de testes de hipóteses *vis-à-vis* à de estimação. (Como David Hendry certa vez afirmou, "the three golden rules of econometrics are test, test and test.").

O primeiro capítulo do livro é reservado ao método de mínimos quadrados. Dois aspectos positivos do tratamento apresentado pelos autores são a ênfase na interpretação geométrica do método de mínimos quadrados e um destaque especial ao teorema de Frisch-Waugh-Lovell. Os Capítulos 2 e 3 cobrem os modelos de regressão não-linear, abordando aspectos relacionados à estimação e aos testes. O Capítulo 4 concentra definições e resultados importantes de teoria assintótica. Este capítulo é fundamental para alunos de mestrado e do primeiro ano de doutorado que pretendem ter acesso à literatura mais moderna em econometria. Os

leitores mais interessados e detalhistas podem encontrar maiores detalhes em White (1984).

O Capítulo 5 representa uma aplicação de métodos assintóticos a modelos não-lineares, enquanto o Capítulo 6 é reservado para o modelo de regressão de Gauss-Newton, isto é, modelos que podem ser representados por regressões artificiais. O próximo capítulo discute estimação por variáveis instrumentais. O tratamento neste capítulo é bem claro e auto-contido. Contudo, o leitor sedento por detalhes encontrará um tratamento mais extensivo em Bowden e Turkington (1984). O método de máxima verossimilhança é discutido em detalhes no Capítulo 8, onde os autores introduzem conceitos importantes, como o limite inferior de Cramér-Rao, a igualdade da matriz de informação, consistência e normalidade assintótica do estimador de máxima verossimilhança, log-verossimilhança concen-trada, entre outros. Este capítulo introduz ainda os três princípios clássicos de teste: multiplicador de Lagrange (ou escore),razão de verossimilhança e Wald. Maiores detalhes sobre estes testes podem ser encontrados ao longo do resto do livro ou em Godfrey (1988). O capítulo seguinte relaxa a hipótese de esfericidade dos erros do modelo de regressão no contexto de máxima verossimilhança e mínimos quadrados generalizados. O Capítulo 10 cobre o fenômeno de correlação serial, apresentando e discutindo os modelos auto-regressivos, de médias móveis, ARMA e ARIMA. Testes baseados no modelo de regressão de Gauss-Newton são descritos no Capítulo 11 (teste de Chow, testes de hipóteses não encaixadas, testes de heterocedasticidade, etc.)

Os Capítulos 12 e 13 constituem uma descrição detalhada e razoavelmente rigorosa de vários aspectos relacionados a testes de hipóteses, desde de conceitos básicos (como, por exemplo, os conceitos de tamanho e poder de testes) até resultados tradicionais de equivalência assintótica dos três testes clássicos e o possível conflito de inferência destes testes em amostras finitas. Estes dois capítulos são de alta

importância para qualquer curso sério de econometria. Vários modelos que envolvem transformação da variável dependente são abordados no Capítulo 14. Especial ênfase é dada ao modelo de Box e Cox (1964). O tópico seguinte é a modelagem de variáveis limitadas e qualitativas: modelos probit, logit, tobit, probit ordenado, logit multinomial, etc. Mais uma vez, o leitor sedento por uma abordagem mais detalhada deve consultar referências especializadas, como por exemplo Maddala (1983). Heterocedasticidade e tópicos relacionados (estimação da matriz de covariância, modelos ARCH e GARCH, testes de assimetria e curtose, o teste de matriz de informação, testes de momentos condicionais, etc.) são tratados no Capítulo 16. Uma boa complementação para o tratamento de modelos ARCH e GARCH é Bera (1993). O capítulo seguinte aborda o método dos momentos generalizado, enquanto o Capítulo 18 discute modelos de equações simultâneas.

No Capítulo 19 os autores apresentam modelos de regressão para dados de séries temporais (vetores auto-regressivos, modelos sazonais, etc.). Aqui os autores discutem o conceito de "regressão espúria" introduzido na literatura de econometria por Granger e Newbold (1974) e formalizado por Phillips (1986). Este tópico é de extrema importância, mesmo para aqueles com interesses puramente aplicados. Uma abordagem um pouco mais profunda poderia ser de maior utilidade. O Capítulo 20 é um dos mais importantes do livro: trata de raízes unitárias e co-integração. É uma pena, entretanto, que os autores não tenham reservado um espaço maior para este material. A apresentação dos testes de raízes unitárias e da teoria assintótica para estes testes é demasiadamente superficial. Em particular, a apresentação do teorema do limite central funcional é pouco rigorosa. Leitores sérios são fortemente recomendados a buscar uma complementação em Banerjee *et alii* (1993), Hamilton (1994) e McCabe e Tremayne (1993). Contudo, um aspecto positivo deste capítulo é que os autores incluem uma tabela com os valores críticos *assintóticos* para os testes de raízes unitárias. Estes

valores críticos, ao contrário dos para amostras finitas tabulados for Fuller (1976), não requerem normalidade ou homoscedasticidade. Neste sentido, eles são mais gerais. O Capítulo 21 fecha o livro com uma discussão de métodos de simulação de Monte Carlo. Pode-se criticar o tratamento dado pelos autores aos métodos de bootstrap, alegando que com os avanços computacionais recentes os métodos de bootstrap estão sendo cada vez mais usados em aplicações práticas e assim mereceriam um tratamento mais detalhado em um livro-texto moderno. Leitores interessados devem continuar consultando a literatura especializada, como, por exemplo, Efron e Tibshirani (1993) e Hall (1992). Este capítulo traz, entretanto, uma boa introdução ao assunto e é louvável que um tópico importante como bootstrap seja tratado em um livro-texto de econometria.

Algumas críticas isoladas podem ser feitas a "Estimation and Inference in Econometrics." Por exemplo, pode-se argumentar que os autores dedicam um espaço excessivo ao tratamento de modelos não-lineares, abordando tópicos importantes (como raízes unitárias e co-integração) de forma superficial e ignorando a literatura sobre regressão não-paramétrica. As alternativas ao modelo clássico de regressão linear que se têm consolidado na literatura mais recente de econometria estão mais relacionadas a modelos robustos, não-paramétricos e semi-paramétricos do que a modelos de regressão não-linear simples (ver, *e.g.*, Birkes e Dodge, 1993). Contudo, no geral, "Estimation and Inference in Econometrics" representa uma contribuição importante por incorporar uma vasta literatura que se tem consolidado em econometria na última década em um formato de livro-texto, e por fazê-lo dando uma ênfase especial à parte de testes de hipóteses. Sem dúvida, este livro é substancialmente superior aos outros manuais de econometria no mercado e é extremamente útil não apenas como livro-texto, mas também como um livro de referência para aqueles que fazem pesquisa teórica e aplicada em econometria e estatística.

(Francisco Cribari-Neto - University of Illinois)

BIBLIOGRAFIA

BANERJEE, A. et al. *Co-integration, error correction, and the econometric analysis of non-stationary data.* Oxford: Oxford University Press, 1993.

BERA, A.K. ARCH models: properties, estimation and testing. *Journal of Economic Surveys* , [s. l.], p. 305-366, 1993.

BIRKES, D., DODGE, Y. *Alternative methods of regression.* New York: Wiley, 1993.

BOWDEN, R.J., TURKINGTON, D.A. *Instrumental variables.* New York: Cambridge University Press, 1984.

BOX, G.E.P., COX, D.R. An analysis of transformations. *Journal of the Royal Statistical Society* B, [s. l.], v.26, p. 211-252, 1964.

EFRON, B., TIBSHIRANI, R.J. *An introduction to the bootstrap.* London: Chapman & Hall, 1993.

FULLER, W.A. *Introduction to statistical time series.* New York: Wiley, 1976.

GODFREY, L.G. *Misspecification tests in econometrics*: the lagrange multiplier principle and other approaches. NewYork:Cambridge University Press, 1988.

GRANGER, C.W.J., NEWBOLD, P. Spurious regressions in econometrics. *Journal of Econometrics* , [s. l.], v.2 , p. 111-120 , 1974.

HALL, P. *The bootstrap and edgeworth Expansion.* New York: Springer-Verlag, 1992.

HAMILTON, J. *Time series analysis.* Princeton:Princeton University Press, 1994.

MADDALA, G.S. *Limited-dependent and qualitative variavles in econometrics.* New York: Cambridge University Press, 1983.

McCABE, B., TREMAYNE, A. *Elements of modern asymptotic theory with statistical applications.* Manchester:Manchester University Press, 1993.

PHILLIPS, P.C.B. Understanding spurious regressionsin econometrics. *Journal of Econometrics*, [s. l.], v. 33, p. 311-340, 1986.

WHITE, H. Asymptotic theory for econometricians. New York: Academic Press, 1984.

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; e

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:

Pedro Luis do Nascimento Silva
Editor Responsável - RBEs
ENCE/IBGE
Rua André Cavalcanti, 106
Bairro de Fátima
20231-050 - 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 tais como as demais seções do trabalho.

6. Gráficos e diagramas para publicação devem ser traçados em papel branco, como 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.

SE O ASSUNTO É BRASIL, PROCURE O IBGE

O 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.

Estamos na INTERNET
http://www.ibge.gov.br webmaster@cddi.ibge.gov.br

VOCÊ PODE OBTER AS PESQUISAS, ESTUDOS E LEVANTAMENTOS DO IBGE EM TODO O PAÍS

Rio de Janeiro:

Centro de Documentação e Disseminação de Informações - CDDI
Rua General Canabarro, 706
20271-201 - Maracanã - Rio de Janeiro - RJ
Fax: (021)284-1109
Ligação Direta Gratuita: 0800-218181

Livraria do IBGE
Avenida Franklin Roosevelt, 146 - loja
20021-120 - Castelo - Tel.: (021)220-9147
Avenida Beira Mar, 436 - 2° andar
20021-060 - Castelo - Tel.: (021)210-1250
Fax: (021)220-3543

Norte

RO - Porto Velho - Rua Tenreiro Aranhã, 2643 - Centro
78900-750 - Telefax: (069)221-3658

AC - Rio Branco - Rua Benjamin Constant, 506 - Centro
69900-160 - Tels.: (068)224-1540/1490 - Ramal 6
Fax: (068)224-1382

AM - Manaus - Avenida Ayrão, 667-3° andar - Centro
69025-050 - Telefax: (092)232-1369

RR - Boa Vista - Avenida Getúlio Vargas, 76-E - Centro
69301-031 - Tel.: (095)224-4103 - Ramal 22

PA - Belém - Avenida Gentil Bittencourt, 418 - Batista Campos
66035-340 - Tel.: (091)241-1440 - Fax: (091)223-8553

AP - Macapá - Avenida Cônego Domingos Maltez, 251 - Centro
68900-270 - Tels.: (096)222-3128/3574
Fax: (096)223-2696

TO - Palmas - ACSE 01 - Conjunto 03 - Lote 6/8 - Centro
77100-040 - Tel.: (063)215-1907 - Ramal 308
Fax: (063)215-1829

Nordeste

MA - São Luís - Av. Silva Maia, 131 - Praça Deodoro
65020-570 - Tel.: (098)221-5121 - Fax: (098)232-3226

PI - Teresina - Rua Simplício Mendes, 436 - Centro
64000-110 - Tel.: (086)221-4161 - Fax: (086)221-6308

CE - Fortaleza - Avenida 13 de Maio, 2901 - Benfica
60040-531 - Telefax: (085)243-6941

RN - Natal - Praça Pedro Velho, 161 - Petrópolis
59020-400 - Tels.: (084)211-4681/5310 - Ramal 13
Fax: (084)211-2002 - Telefax: (084)221-3025

PB - João Pessoa - Rua Irineu Pinto, 94 - Centro
68010-100 - Tel.: (083)241-1560 - Ramal 21
Fax: (083)221-4027

PE - Recife - Rua do Hospício,387 - 4° andar - Boa Vista
50050-050 - Tel.: (081)231-0811 - Ramal 215
Fax: (081)231-1033

AL - Maceió -Praça dos Palmares, s/n°
Edifício do INAMPS, 3° andar
57020-000 - Tel.: (082)221-2385 - Fax: (082)326-1754

SE - Aracaju - Rua Riachuelo, 1017 - Térreo - São José
49015-160 - Telefax: (079)222-3122/8197/8198

BA - Salvador - Av. Estados Unidos, 476 - 4° andar - Comércio
Ed. Sesquicentenário - 40013-900 - Tel.: (071)243-9277
Ramais 2005 e 2008 - Telefax: (071)241-2502

Sul

PR - Curitiba - Alameda Dr. Carlos de Carvalho, 625 - Térreo
Centro - 80430-180 - Tel.: (041) 322-5500 - Ramais 253 e 254
Telefax: (041)222-5764

SC - Florianópolis - Rua Victor Meirelles, 170 - Centro
88010-440 - Tel.: (048)224-0733 - Ramais 234 e 256
Telefax: (048)222-0338

RS - Porto Alegre - Avenida Augusto de Carvalho, 1205 - Térreo
Praia de Belas - 90010-390 - Tel.: (051)228-6444
Ramais 211, 213 e 225 - Fax: (051)228-8507
Telefax: (051)228-6444 - Ramal 212

Sudeste

MG - Belo Horizonte - Rua Oliveira, 523 - 1° andar - Cruzeiro
30310-150 - Tel.: (031)223-0554 - Ramais 1112 e 1113
Telefax: (031)223-3381

ES - Vitória - Avenida. dos Navegantes, 675 - 9° andar - Enseada do Suá - 29056-900 - Tel: (027) 325-3857 - Fax: (027) 325-3908

SP - São Paulo - Rua Urussuí, 93 - 3° andar - Itaim Bibi
04542-050 - Tels.: (011)822-2106/0077 - Ramal 281
Fax: (011)822-5264

Centro-Oeste

MS - Campo Grande - Rua Barão do Rio Branco, 1431 - Centro
79002-174 - Tels.: (067)721-1163/1902/1525 - Ramais 32 e 42
Fax: (067)721-1520

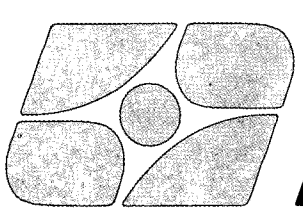
MT - Cuiabá - Avenida Tenente Coronel Duarte, 407 - 1°/ 2° andares - Centro - 78005-750 - Tels: (065)322-2121/2225
Fax: (065)321-3316/623-0573

GO - Goiânia - Avenida Tocantins, 675 - Setor Central
74015-010 - Tel.: (062)223-3121 - Telefax: (062)223-3106

DF - Brasília - SDS - Ed. Venâncio II -Bl H - Quadra 06
1° andar - 70393-900 - Tels.: (061)223-1359/321-7702
Ramal 124 - Fax: (061)226-9106

IBGE possui, ainda, agências localizadas nos principais municípios

ESTAMOS NA INTERNET



IBGE

Instituto Brasileiro de Geografia e Estatística

http://www.ibge.gov.br